THESIS

GENOMICS AND TRANSCRIPTOMICS OF THE MOLTING GLAND (Y-ORGAN) IN THE BLACKBACK LAND CRAB, *GECARCINUS LATERALIS*

Submitted by

Lindsay Martin

Department of Biology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2016

Master's Committee:

Advisor: Donald L. Mykles

Deborah M. Garrity
Tingting Yao

ABSTRACT

GENOMICS AND TRANSCRIPTOMICS OF THE MOLTING GLAND (Y-ORGAN) IN THE BLACKBACK LAND CRAB, *GECARCINUS LATERALIS*

Molting is required for growth and development in crustaceans. In the blackback land crab *Gecarcinus lateralis,* molting is stimulated by ecdysteroids, hormones produced in the Y-organ (YO). Throughout the molting cycle, the YO demonstrates phenotypic plasticity. The phenotypic plasticity is correlated with the stages of the molt cycle, during which YO ecdysteroid production varies. During intermolt, the longest stage of the molt cycle, the circulating ecdysteroid titers are low and molting is suppressed. In preparation for molting, the YO increases ecdysteroid production during premolt. Circulating ecdysteroids continue to rise, dropping right before the ecdysis and remaining low in the subsequent postmolt period. During the molt cycle, the YO's sensitivity to inhibitory cues also varies, which contributes to ecdysteroid fluctuations.

To better understand how changes in gene expression modulate the YO's phenotypic plasticity, a YO transcriptome from five molt stages was generated. Using over 5.6 million reads from Illumina, 229,278 contigs were assembled to comprise the reference transcriptome. By comparing expression levels of the transcripts between the molt stages, 13,189 unique differentially expressed contigs were identified in *G. lateralis*. Based on differential expression, insect hormone biosynthesis and oxidative phosphorylation pathways were enriched, validating the YO transcriptome identity. Using GO enrichment, *MAP kinase* was identified as a possible candidate gene for regulating YO ecdysteroid synthesis.

To complement and validate the transcriptome, claw muscle genomic DNA was sequenced and assembled using 2.6 million reads. 375,152 scaffolds ≥ 500 bp were built, with an N50 of 1,841 bp. Using k-mer frequencies, the genome size was estimated to be 3.07 Gb, similar to mammalian vertebrates. The median gene size of *G. lateralis* was approximated to be 6,300 bp; the disparity between the median estimate and the N50 prohibited further computational analysis. Genome scaffolds were sufficient in length for manual comparison. Alignment of the transcriptome and genome sequences of the *Rheb* gene showed 100% nucleotide alignment in the open reading frame, and extended the sequence by 7.7 fold, including the identification of four introns. The sequence comparison validated both genome and transcriptome assemblies and extended the gene sequence.

Next-generation sequencing provided us with a global perspective of molecular variations within the YO throughout the molt cycle. We hypothesize variations in gene expression regulate YO phenotypic plasticity by varying ecdysteroid production. YO transitions throughout molting are essential for regulation. YO activation and commitment, both corresponding to increased ecdysteroids, are required to induce ecdysis. YO repression, during which circulating ecdysteroid titers are low, is needed to prevent precocious molting. Identifying changes in gene expression and key regulatory elements correlating with variations in YO phenotype will increase our understanding of molt cycle regulation, which is critical for crustacean development, growth, and repair.

ACKNOWLEGEMENTS

A special thank you to Don. I am grateful for the opportunities you provided and your patience throughout the past decade. I grew both scientifically and personally while in your lab. Thank you for the belief, encouragement, support, and overwhelming patience. You were the voice of scientific reason, and provided guidance when I needed it.

Thank you to Dr. Garrity and Dr. Yao for serving on my committee. Dr. Yao, you were my favorite teacher in graduate school. Your high standards pushed me to be a better scientist.

Thank you to the Mykles' crab lab, especially: Alejandro López-Cerón, Savannah Ciardelli-Mullis, Brigitte Gaudreault, and Genna Frappaolo. Molly and Natalie, thank you for showing me the ropes. An extended thank you to Sunetra Das, who indelibly shaped my academic career. I am honored to be your first pupil. My project is what it is because of you.

Thank you to my family for providing unconditional support and love: Grandpa Jim; Mom and Dad; Megan (Titus); and Jon and Jenn (Breyton, Mara, Dekker). Thanks to Grandma Nancy, whose forward thinking and passion for education paved the way for me. Without family, this would not have been possible. Thank you for your patience and for believing in me.

Thank you to the teachers, friends, and family who are too numerous to be acknowledged within this page. My educational journey was positively influenced by so many. Thank you to everyone who encouraged me to continue my education.

Most of all, thank you to my supportive and loving partner Edwin. You are a source of strength and great encouragement. Your commitment to our relationship contributed greatly to my success and accomplishments. Life with you is my great adventure.

In addition, I would like to thank the following individuals for contributing to this project: Hector Horta for animal collection; Ernie S. and Sharon Chang for hemolymph ELISA; Mykles lab for animal care and comradery; Donald L. Mykles for claw dissection for the genome; Megan Mundron and Natalie L. Pitts for YO transcriptome tissue harvest; David Durica and Sunetra Das for genome and transcriptome library preparation; Jonathan Wren for genome and transcriptome sequencing; Clayton Hallman for R support; Savannah Ciardelli- Mullis for project assistance; Jenn Martin for photo edits; Richard Casey for HPC guidance; Dr. Thomas Hauser for SC15 experience; Wen Zhou for statistical consultations; and Jesse Schafer for extensive HPC support. This project is funded by the National Science foundation, IOS-1257732. Significant computing contributions from OSU HPCC. OSU supported in part through an NSF grant OCI–1126330.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION

Molting in Crustaceans:

Development, growth, and repair are common needs among organisms. The mechanisms used by different species to meet these shared needs are widely diverse. Animals in the Crustacea subphylum experience cyclical regeneration and shedding of a rigid exoskeleton throughout the life cycle (Skinner, 1985). This process, also known as molting, is required to: complete the metamorphosis from zoea to juvenile (Willems, 1982), grow in size from juvenile to and throughout adulthood (Chang and Mykles, 2011), and repair damaged exoskeletons and regenerate lost limbs throughout life (Das, 2015). The blackback land crab, *Gecarcinus lateralis*, exhibits typical molt patterns of crustaceans. *G. lateralis* serves as a model crustacean to better elucidate mechanisms influencing the molting cycle of adult crabs.

Molting is a physiological process with distinct stages originally identified by Drach in 1939. Drach used exoskeletal modifications to classify the molt stages into main groups A-E, with 14 sub-groups (Drach, 1967). Each stage is accompanied by physiological changes in the exoskeleton, which aid in the successful molt of the animal. Skinner and Bliss added to Drach's work by identifying events specific to *G. lateralis* and modifying nomenclature (Bliss and Boyer, 1964; Skinner, 1962). Figure 1 is a compilation of the observations made by Drach, Skinner, and Bliss; further described in the text below. For clarity, molt stage terminology was refined to the following six stages: intermolt, early premolt, mid premolt, late premolt, ecdysis, and postmolt (Chang and Mykles, 2011).

Molt preparation and recovery comprises several molt cycle stages (Fig. 1), with ecdysis being a relatively brief event. Ecdysis, the act of removal from the old exoskeleton, is completed within a few minutes (Kuballa, 2007) to a few hours. The event requires extensive premolt preparation, which constitutes a significant molt stage (Bliss and Boyer, 1964). The premolt period varies from approximately two weeks, to an excess of a month (Kuballa, 2007; Skinner, 1962). After ecdysis the animal is extremely vulnerable due to the soft, uncalcified exoskeleton exposed (Bliss and Boyer, 1964). Recovery in postmolt is also a critical time, and in conjunction with the preparation, comprises one-third of a crustacean's life cycle (Kuballa, 2007). During the entire molt cycle, the animal is undergoing physiological changes, not limited to: exoskeletal restructuring, hormonal fluctuations, and limb regeneration (Das, 2015; Skinner, 1985).

Physiological Changes Indicate Molt Stage Progression:

The exoskeleton is comprised of epicuticle, exocuticle, and endocuticle layers (Fig. 2), which transition throughout the molt cycle (Skinner, 1962). *G. lateralis* spends a majority of time in intermolt (C4), when the exoskeleton is developed and rigid (Bliss and Boyer, 1964; Kuballa, 2007). Throughout intermolt, the innermost layer of the exoskeleton, endocuticle, continues to form and increase in size (Skinner, 1962). Intermolt is the stage when exoskeletal components are most stable and variations in structure are minimal.

When adult *G. lateralis* transitions into early premolt (D0) due to growth or repair needs, restructuring of the exoskeleton occurs. The exoskeleton separates from the underlying epidermis and epidermal cellular hypertrophy occurs (Skinner, 1985). These epidermal cells are responsible for synthesizing two outermost layers of the new exoskeleton, the epicuticle and the endocuticle (Skinner, 1962). To facilitate the construction, part of the existing exoskeleton is catabolized in mid premolt (D1-2) (Bliss and Boyer, 1964; Drach, 1967). The enlarged epidermal cells use the recycled materials to form a new epicuticle and exocuticle below the existing exoskeleton (Skinner, 1985). In late premolt (D3-4) the epidermal cells separate from the endocuticle in preparation for molting (Skinner, 1962). In addition to exoskeletal changes, the entire animal is experiencing an increase in water absorption, weight, and metabolism (Skinner, 1962). Overall, premolt is a time of growth for *G. lateralis* (Bliss and Boyer, 1964).

Ecdysis (E) is the act of molting, when the carapace splits and the animal exits from the old exoskeleton by backward withdrawal (Drach, 1967; Kuballa, 2007). Water engulfed during premolt is used to stretch the new exoskeleton, which is thin and supple (Skinner, 1985). Immediately following ecdysis animals are lighter (Skinner, 1962). In postmolt, weight is rapidly re-gained as muscle and tissue growth occur (Skinner, 1985). The newly expanded carapace width is 1-7% larger, and accommodates an overall weight increase of 6-22% by the end of postmolt (Skinner, 1962). The previously hypertrophied epidermal tissue cells reduce to the original size (A1-2), and a new endocutical begins to form (B1-2) (Skinner, 1962). Calcification and hardening of the exoskeleton marks the end of the postmolt period, as the exoskeleton is now mature (C1-3).

Hormonal fluctuations are also an indication of molt stage progression. Molting is induced by ecdysteroids, a group of steroid hormones produced by the endocrine gland, the Y-organ (Mykles, 2011). In *G. lateralis*, the Y-organs (YOs) are easily accessible in the anterior cephalothorax of the animal (Fig. 3) (Bliss and Boyer, 1964; Chang and Mykles, 2011). The YO synthesizes ecdysteroids from cholesterol, and secretes the molting hormones into the hemolymph (Bliss and Boyer, 1964; Chang et al., 1993; Lachaise et al., 1993; Mykles, 2011). Ecdysteroid hemolymph titers vary, and fluctuations regulate molt stage progression (Chang and Mykles, 2011). *G. lateralis* hemolymph ecdysteroids are quantifiable using a competitive enzyme-linked immunosorbent assay (ELISA), allowing for accurate and consistent molt-staging of animals (Covi et al., 2010).

The presence of the YO is required for the molt cycle (Bliss and Boyer, 1964; Lachaise et al., 1993), but regulation is essential to prevent precocious or continuous molting. To accomplish this, the YO is negatively regulated by hormones produced in the X-organ, XO (Bliss and Boyer, 1964; Chang et al., 1993; Covi et al., 2012; Lachaise et al., 1993). The XO is located in the base of the eyestalks (Fig. 3) and produces two neuropeptide hormones involved in regulating ecdysteroid synthesis: molt inhibiting hormone, MIH, and crustacean hyperglycemic hormone, CHH (Chang et al., 1993; Covi et al., 2012). Experimentally, MIH and CHH demonstrate a negative effect on ecdysteroid secretion in *G. lateralis*, both *in vitro* and *in vivo*, with MIH being identified as the primary inhibitor (Covi et al., 2012). It is hypothesized that MIH and CHH binds to receptors on the YO and inhibit ecdysteroid synthesis through cAMP and cGMP secondary messengers (Covi et al., 2012; Mykles et al., 2010; Nakatsuji et al., 2009). The role and importance of these cyclic nucleotides in suppressing the YO varies by species as well as molt stage (Covi et al., 2009; Mykles et al., 2010).

During the molt cycle, the YO's ecdysteroid secretions and MIH sensitivity are used to identify the molt stage of an animal (Fig. 4). In intermolt, hemolymph ecdysteroid titers remain low and YO steroidogenesis is repressed by MIH (Chang and Mykles, 2011). When *G. lateralis* enters premolt, ecdysteroids linearly increase (McCarthy and Skinner, 1977) and the YO's sensitivity to MIH changes (Chang and Mykles, 2011). Early premolt is initiated by a drop in MIH, accompanied by YO activation and an increase in hemolymph ecdysteroids (Chang and Mykles, 2011). During mid premolt hemolymph titers are continually elevated and the animal transitions into a committed state: experimental introduction of MIH is no longer able to suppress molting (Chang and Mykles, 2011). In late premolt, hemolymph ecdysteroids peak (Chang and Mykles, 2011), followed by a rapid decline just prior to ecdysis (Mykles, 2011). After ecdysis, in postmolt, the YO transitions to a state similar to intermolt, with low ecdysteroid levels (Chang and Mykles, 2011).

In addition to exoskeletal remodeling and hormone fluctuations, limb regenerates also serve as a marker for molt stages (Bliss and Boyer, 1964). Crustaceans have the ability to autotomize, reflexively drop, and regenerate appendages (Mykles, 2001). Epimorphosis is the limb regeneration mode used by crustaceans; it involves the replacement of a walking leg or claw from a mass of cells that undergoes de-differentiation and proliferation (Das, 2015). In *G. lateralis* epimorphosis occurs in two stages: basal growth and proecdysial growth (Das, 2015). Basal growth occurs in intermolt, when hemolymph ecydsteroid titers are low (Das, 2015). During basal growth, cell differentiation occurs; and the regenerate becomes fully differentiated containing all segments and tissues found in a fully developed limb (Das, 2015). Proecdysial growth occurs throughout premolt, and is a time of rapid growth for the limb bud (Bliss and Boyer, 1964; Das, 2015). By the time the animal reaches ecdysis, limb regeneration is complete, and the animal will emerge from the molt with a full set of appendages. By autotomizing multiple limbs, *G. lateralis* can be forced into a precocious molt (Fig. 5) (Skinner and Graham, 1972).

In *G. lateralis,* the increasing size of limb regenerates is compared to carapace width to generate an R-value (Bliss and Boyer, 1964). Increasing R-values track the progression of individuals throughout the molting cycle (Fig. 4, 5) (Bliss and Boyer, 1964). The following R-values correspond to stages in the molt cycle: 0-8 in intermolt, 8-15 in early premolt, 15-19 in mid premolt and >19 in late premolt (MacLea et al., 2012). As R-values are normalized measurements of limb buds, it is a convenient method to quickly stage crabs of varying sizes. The limb regenerates reliance upon molting and regulation by ecdysteroids make them an additional molt stage indicator for *G. lateralis*.

<u>Next Generation Sequencing</u>:

Advances in next generation sequencing, NGS, allow for comprehensive and efficient sequencing of genomic DNA and mRNA (Haas et al., 2013). Researchers are now able to acquire a large amount of data in a relatively short time frame to answer novel, global-scale questions. The applications for NGS are diverse: identifying gene products, classifying regulatory components, researching signaling pathways, and locating disease-causing mutations are a few common areas of study (Grada and Weinbrecht, 2013). Often a database of genomic DNA (genome) or mRNA (transcriptome) can be used to answer a variety of questions or generate new hypotheses. The ability to gain a global perspective and answer a diverse range of questions using a single dataset is the basis for choosing next generation sequencing to study molting in *G. lateralis*.

Experiments involving NGS often start with the same methodology. Tissues are harvested and genomic DNA or cDNA samples are prepared for sequencing (Grada and Weinbrecht, 2013). Libraries are amplified and sequenced using massively parallel sequencers to reduce time and cost (Mardis, 2011). Reads are returned to the researcher, and data analysis begins (Grada and Weinbrecht, 2013). Quality reads are assembled into longer, usable sequences known as scaffolds or contigs. These sequences require annotation, usually performed computationally against related species (Yandell and Ence, 2012). Finding biological significance is dependent upon assembly fidelity and a quality reference database for annotation

(Yandell and Ence, 2012). After basic annotation, the pipeline diverges based on the project. The recent surge in sequencing has led to a rapid development of computational packages designed to address an ever-expanding number of questions being answered through bioinformatics (Yandell and Ence, 2012).

RNA sequencing technology, RNA-seq, provides an opportunity to examine a complete set of transcripts within an organism (Wang et al., 2009). Previous studies involving *G. lateralis* required transcripts to be sequenced individually, with degenerate primers from arthropod orthologs. Using this approach, a total of 59 *G. lateralis* mRNA sequences are available on NCBI. RNA-seq is ideal for increasing this database, because computational parameters, not prior gene knowledge, are used to assemble transcripts on a global scale (Haas et al., 2013; Wang et al., 2009; Willems, 1982). At the same time as assembly, transcripts can be quantified with accuracy comparable to quantitative PCR (Wang et al., 2009). Simultaneously producing and quantifying a global set of transcripts decreases the amount of tissue required, and maximizes the efficiency of the experiment- both in regards to cost and time. Additionally, the large amount of information gleaned can lead itself to new hypotheses previously unconsidered.

Sequencing and assembling a genome is a valuable complement to a transcriptome. A genome can be superimposed onto transcriptome data to provide additional gene information, such as regulatory elements and gene structure (Feuillet et al., 2011). Sequencing both a genome and transcriptome is ideal, as comparisons between the two can improve assembly, annotation, and interpretation of sequence data (Yandell and Ence, 2012). Repetitive sequences, ploidy numbers, and gene duplicates add complications to genome projects (Unamba et al., 2015), however these issues are addressed through constant improvements and innovations in bioinformatics packages.

NGS is popular among biologists, but identifying biological significance can be a challenge. Generating a genome or transcriptome using reads is relatively straightforward, but ensuring the product is accurate and usable is still an issue (Feuillet et al., 2011). Sequences assembled are only as good as the algorithms used to compile them, and annotation is entirely dependent upon the quality of the reference database (Yandell and Ence, 2012). Quality control and verification is required. New statistical models are needed to address questions generated by NGS data. For example, many transcriptome researchers want to identify genetic dependencies and relationships, a complex question requiring new statistical models (Chang, 2016). Although NGS is not new, the availability and standardization of tools is still in its infancy, which may require data generated to be reanalyzed to be substantiated (Yandell and Ence, 2012).

Despite the challenges, NGS was selected to address the question of phenotypic plasticity in *G. lateralis*. The YO in *G. lateralis* plays a critical role in molt regulation through changes in MIH sensitivity, leading to varying levels of ecdysteroid production. How molecular regulation modulates variations in YO physiological states throughout the molt cycle is largely unknown. Mechanistic target of rapamycin, mTOR, and transforming growth factor β, TGF-β are two pathways experimentally identified to play a pivotal role in YO activity and ecdysteroid synthesis (Abuhagr AM, in press 2016). The use of NGS presents the opportunity to validate local studies while gaining a global perspective on coding sequences, regulatory regions, and transcript expression. For this reason, genomic and transcriptomic experiments were designed to study variations in *G. lateralis* YO in various molt stages. Identifying transcriptional mechanisms controlling YO gene expression will create a better understanding of conditions required for crustacean development, growth, and repair.

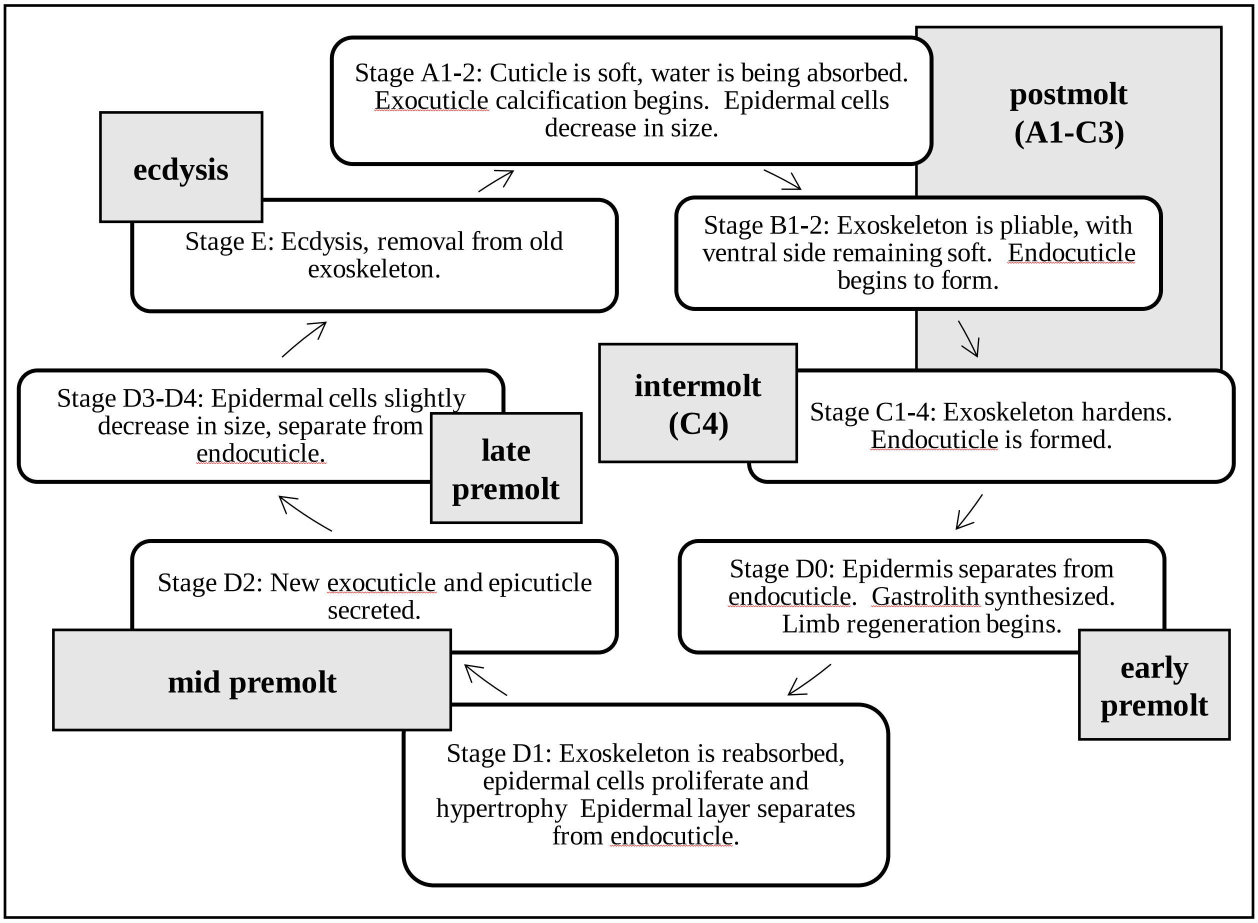


Figure 1: Molt cycle stages and events occurring in *G. lateralis*. Figure information compiled from observations by: Drach 1939, Skinner 1962 and Bliss 1964. Stage A-E identified by Drach. Intermolt-postmolt stage nomenclature coined by Skinner. Location of intermolt-postmolt stages in relation to Drach determined by Mykles 2011.

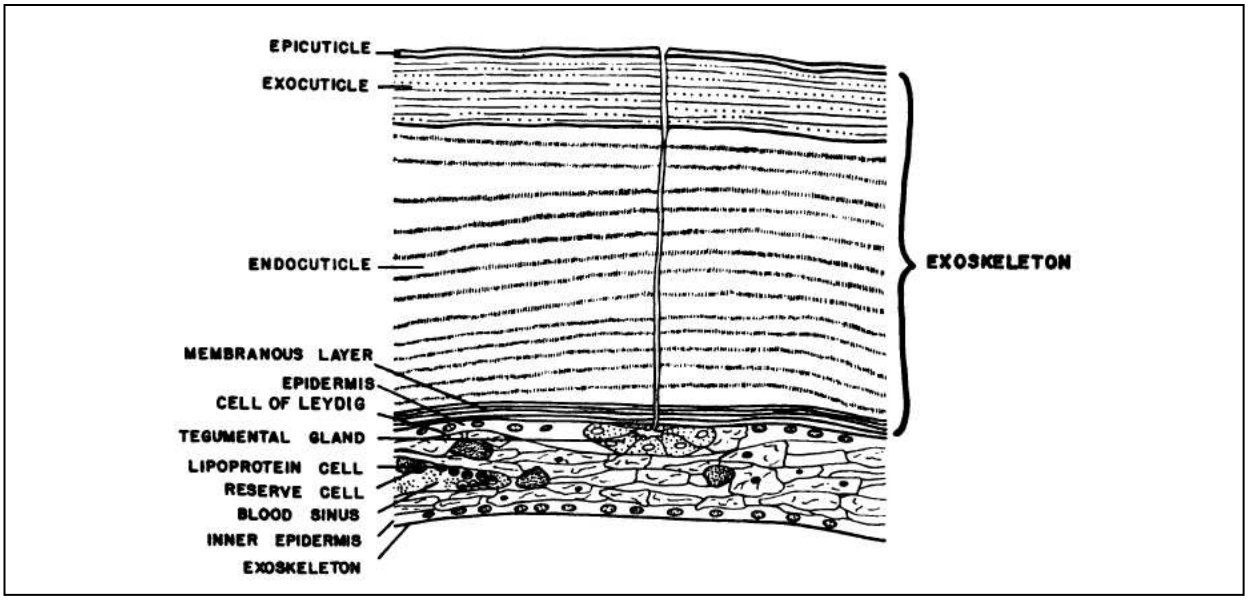
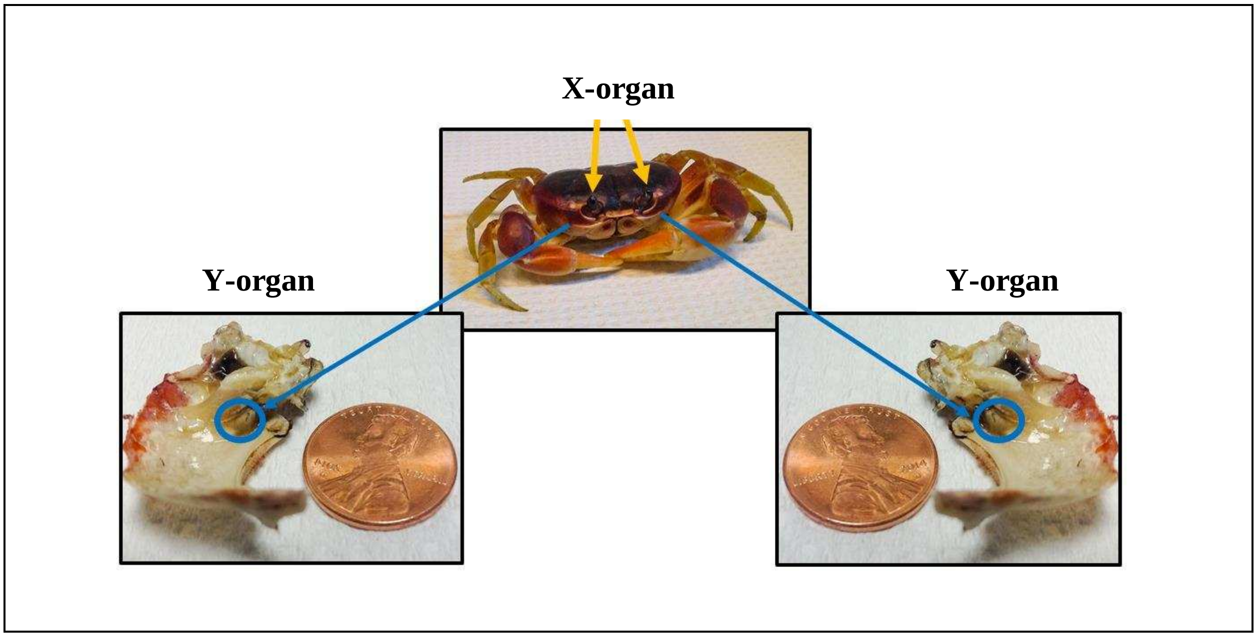
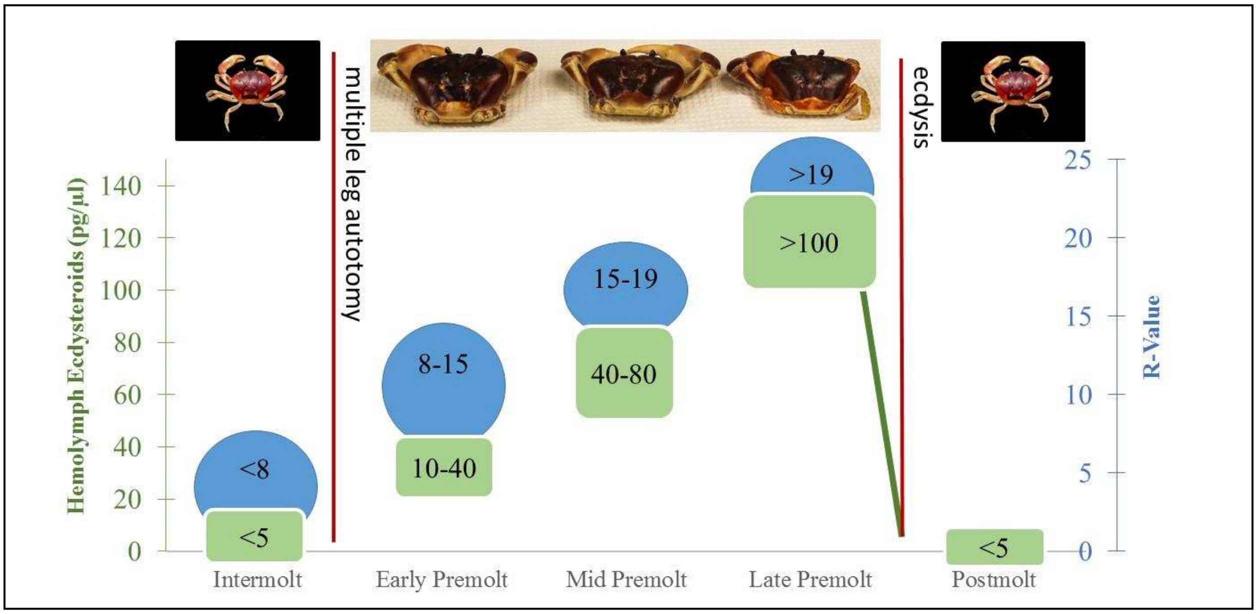


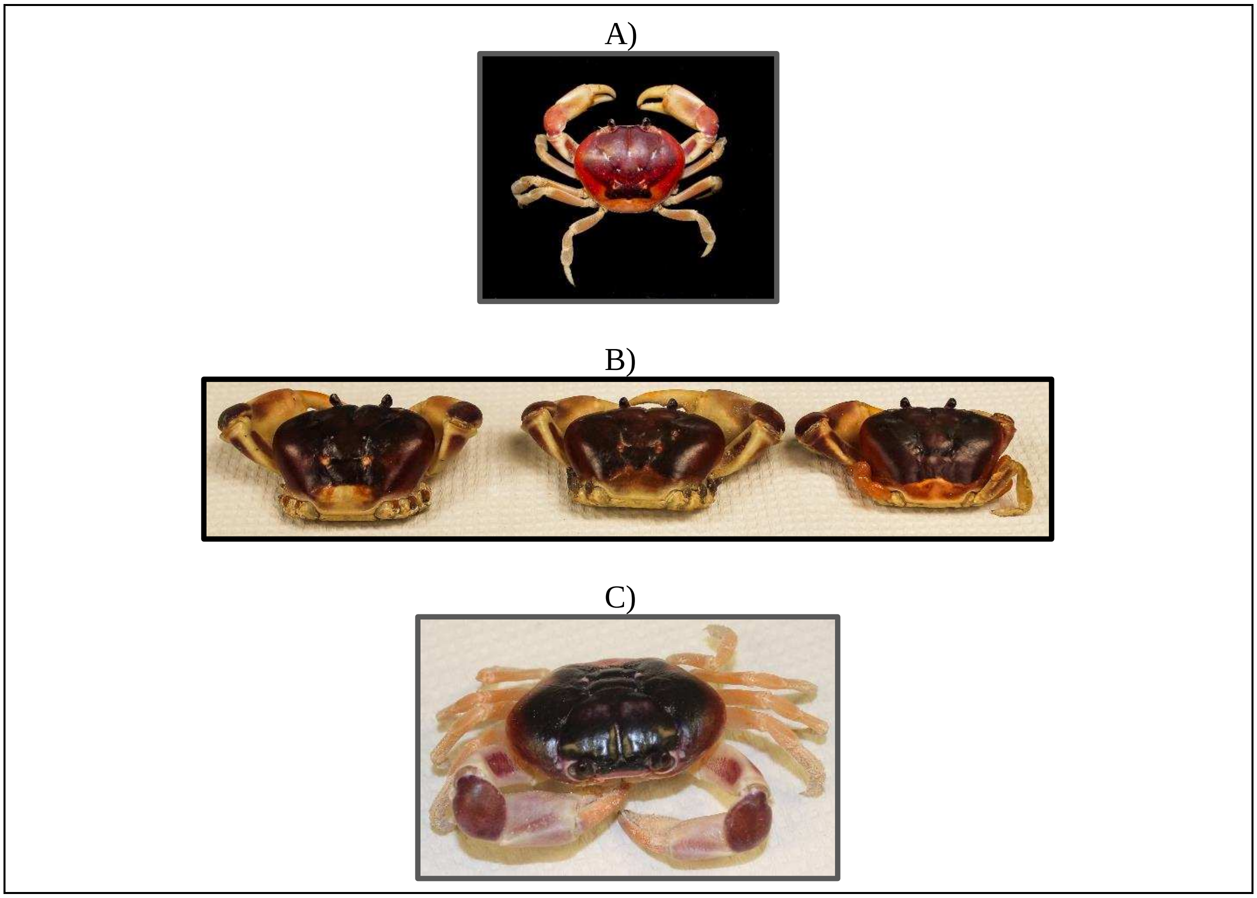
Figure 2: Diagram of intermolt integumentary tissue of *G. lateralis* generated by Skinner, 1962.



<u>Figure 3</u>: Locations of the X-organ and Y-organ in *G. lateralis*. The X-organ is located in the eyestalk ganglia. The Y-organ is located in the anterior cephalothorax.



<u>Figure 4:</u> Molt stage identification of *G. lateralis* in multiple leg autotomy experiment, using R-values and hemolymph ecdysteroid titers. Images of representative individuals are shown above. Note the drop in ecdysteroid titers in late premolt, proceeding ecdysis.



<u>Figure 5</u>: Visualizing the molt cycle stages of *Gecarcinus lateralis*. The images are representative of animals in intermolt, premolt (early, mid, late), and postmolt. (A) The animal in intermolt, with a right walking leg removed for R-value molt stage verification. (B) Animals with multiple leg autotomy progress through early, mid, and late premolt. Molting is precocious. Dorsal view highlights growing limb buds. (C) Post molt animal with full set of regenerated walking legs.

REFERENCES

Abuhagr AM, K.S.M., Megan R. Mudron, Sharon A. Chang, Ernest S. Chang, Donald L. Mykles, in press 2016. Roles of mechanistic target of rapamycin and transforming growth factor-β signaling in the molting gland (Y-organ) of the blackback land crab, *Gecarcinus lateralis*. Comparative Biochemistry and Physiology, Part A.

Bliss, D.E., Boyer, J.R., 1964. Environmental Regulation of Growth in the Decapod Crustacean *Gecarcinus lateralis*. General and Comparative Endocrinology 4, 15-41.

Chang, E.S., Bruce, M.J., Tamone, S.L., 1993. Regulation of Crustacean Molting: A Multi-Hormonal System. American Zoologist 33, 324-329.

Chang, E.S., Mykles, D.L., 2011. Regulation of crustacean molting: a review and our perspectives. General and Comparative Endocrinology 172, 323-330.

Chang, J.Z., Wen; Zhou, Wen-Xin; Wang, Lan, 2016. Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. Biometrics arXiv:1505.04493v3.

Covi, J.A., Bader, B.D., Chang, E.S., Mykles, D.L., 2010. Molt cycle regulation of protein synthesis in skeletal muscle of the blackback land crab, *Gecarcinus lateralis,* and the differential expression of a myostatin-like factor during atrophy induced by molting or unweighting. Journal of Experimental Biology 213, 172-183.

Covi, J.A., Chang, E.S., Mykles, D.L., 2009. Conserved role of cyclic nucleotides in the regulation of ecdysteroidogenesis by the crustacean molting gland. Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology 152, 470-477.

Covi, J.A., Chang, E.S., Mykles, D.L., 2012. Neuropeptide signaling mechanisms in crustacean and insect molting glands. Invertebrate Reproduction & Development 56, 33-49.

Das, S., 2015. Morphological, Molecular, and Hormonal Basis of Limb Regeneration across Pancrustacea. Integrative and Comparative Biology 55, 869-877.

Drach, P.T., Catherine, 1967. Sur la méthode de détermination des stades d'intermue et son application générale aux crustacés (On the method of determining the intermolt stages and its general application to crustaceans). Biologie Marine 18, 595-610.

Feuillet, C., Leach, J.E., Rogers, J., Schnable, P.S., Eversole, K., 2011. Crop genome sequencing: lessons and rationales. Trends in Plant Science 16, 77-88.

Grada, A., Weinbrecht, K., 2013. Next-Generation Sequencing: Methodology and Application. Journal of Investigative Dermatology 133, E1-E4.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., Regev, A., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8, 1494-1512.

Kuballa, A.E., Abigail, 2007. Novel molecular approach to study moulting in crustaceans. Bull. Fish. Res. Agen. 20, 53-57.

Lachaise, F., Leroux, A., Hubert, M., Lafont, R., 1993. The Molting Gland of Crustaceans: Localization, Activity, and Endocrine Control (A Review). Journal of Crustacean Biology 13, 198-234.

MacLea, K.S., Abuhagr, A.M., Pitts, N.L., Covi, J.A., Bader, B.D., Chang, E.S., Mykles, D.L., 2012. Rheb, an activator of target of rapamycin, in the blackback land crab, *Gecarcinus lateralis*: cloning and effects of molting and unweighting on expression in skeletal muscle. Journal of Experimental Biology 215, 590-604.
Mardis, E.R., 2011. A decade's perspective on DNA sequencing technology. Nature 470, 198- 203.
McCarthy, J.F., Skinner, D.M., 1977. Interactions Between Molting and Regeneration. 2. Proecdysial Changes In Serum Ecdysone Titers, Gastrolith Formation, and Limb Regeneration following Molt Induction by Limb Autotomy and/or Eyestalk Removal in the Land Crab, *Gecarcinus lateralis*. General and Comparative Endocrinology 33, 278-292.
Mykles, D.L., 2001. Interactions between Limb Regeneration and Molting in Decapod Crustaceans. American Zoologist 41, 399-406.
Mykles, D.L., 2011. Ecdysteroid metabolism in crustaceans. Journal of Steroid Biochemistry and Molecular Biology 127, 196-203.
Mykles, D.L., Adams, M.E., Gaede, G., Lange, A.B., Marco, H.G., Orchard, I., 2010. Neuropeptide Action in Insects and Crustaceans. Physiological and Biochemical Zoology 83, 836-846.
Nakatsuji, T., Lee, C.-Y., Watson, R.D., 2009. Crustacean molt-inhibiting hormone: Structure, function, and cellular mode of action. Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology 152, 139-148.
Skinner, D.M., 1962. The Structure and Metabolism of a Crustacean Integumentary Tissue During a Molt Cycle. Biological Bullletin 123, 635-647.
Skinner, D.M., 1985. The Biology of Crustacea. Acadmeic Press.
Skinner, D.M., Graham, D.E., 1972. Loss of Limbs as a Stimulus to Ecdysis in Brachyura (True Crabs). Biological Bulletin 143, 222-&.
Unamba, C.I.N., Nag, A., Sharma, R.K., 2015. Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. Frontiers in Plant Science 6, 1-13.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57-63.
Willems, K.A., 1982. Larval Development of the Land Crab Gecarcinus lateralis lateralis (Fréminville, 1835) (Brachyura: Gecarcinidae) Reared in the Laboratory. Journal of Crustacean Biology 2, 180-201.
Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13, 329-342.

CHAPTER TWO: TRANSCRIPTOME

Introduction:

Molting is a cyclical process in crustaceans, required for adult growth and repair. Hormonal regulation of the molt cycle is critical, and ecdysteroids vary as the *G. lateralis* transitions through the molt stages: intermolt, early premolt, mid premolt, late premolt, and postmolt (Chang and Mykles, 2011). To better understand how phenotypic plasticity of the Y-organ (YO) is transitioning between the molt cycle stages, molecular regulators of ecdysteroid synthesis are a target of study.

The YO is the crustacean molting gland homologous to the prothoracic gland in insects (Lachaise et al., 1993; Skinner, 1985). Within this structure, ecdysteroids are synthesized from cholesterol obtained through the animal's diet (Mykles, 2011). The pathway of ecdysteroid synthesis begins with the conversion of cholesterol to the intermediate 5β-diketol, followed by a the conversion of 5β-diketol into secreted ecdysteroids: ecdysone, 25-deoxyecdysone, and 3-Dehydro-25-deoxyecdysone (Mykles, 2011). Ecdysteroids are secreted in the hemolymph, circulatory fluid, and carried throughout the animal. A group of Halloween genes in the biosynthetic pathway are responsible for ecdysteroid synthesis. Shroud, Spook/Spookier/Spookiest, Phantom, Disembodied, and Shadow are all involved in the conversion of cholesterol to ecdysteroids within the YO (Mykles, 2011). These genes serve as markers for ecdysteroid synthesis in the YO. For a full review on ecdysteroid metabolism the reader is directed to Mykles (2011).

The YO is comprised of homogenous cells tightly packed into a structure a few millimeters wide (Skinner, 1985). Within crustaceans the location and appearance of YO can vary widely, and even disappear in those with a terminal molt (Lachaise et al., 1993). *G. lateralis* continues to molt throughout its lifetime, with YO located in the cephalothorax (Skinner, 1985), encased in a cuticular membrane, attached to the exoskeleton. Cells are homogenous and structural changes are reported throughout the molt cycle in crustaceans, but vary widely by species (Skinner, 1985).

Detailed microscopy work has not been completed in the YO of *G. lateralis*, however the YO of a related freshwater crab, *Travancoriana schirnerae,* was examined during premolt (Arath Raghavan Sudha Devi and Sagar, 2015). The premolt YO contained an abundance of ribosomes and mitochondria (Arath Raghavan Sudha Devi and Sagar, 2015). Mitochondria were found concentrated near the nucleus, and contained extensive cristae (Arath Raghavan Sudha Devi and Sagar, 2015). Two sizes, micro and macro mitochondria were identified and both found to be dense (Arath Raghavan Sudha Devi and Sagar, 2015), presumably with cristae. These observations support previous reviews of the YO in other crustaceans, which noted mitochondrial changes. Mitochondria in mid premolt undergo structural remodeling, cristae are numerous (Lachaise et al., 1993). During premolt, mitochondria cristae increased, along with a 15-fold increase in the volume of the entire organelle (Skinner, 1985).

In vivo YO experiments demonstrate the mechanistic target of rapamycin (mTOR) and transforming growth factor-β (TGF-β) pathways are both involved in upregulating ecdysteroid synthesis (Abuhagr AM, in press 2016). Animals induced into a precocious molt by the removal of endogenous MIH were exposed to rapamycin, an inhibitor of mTOR. Results showed a decrease in hemolymph ecdysteroids which persisted for 14 days, until the experiment was terminated (Abuhagr AM, in press 2016). In a similar experiment exposure to SB431542, a TGF-β antagonist, led to a decrease in ecdysteroid synthesis after day 7 (Abuhagr AM, in press

2016). Chemical inhibitors directed at the mTOR and TGF-β pathway resulted in decreased hemolymph ecdysteroids, suggesting an inhibition of ecdysteroid synthesis in the YO. Both mTOR and TGF-β are identified as regulators of the ecdysteroid production in premolt.

Previously a YO transcriptome for *G. lateralis* was constructed with intermolt animals, to gain a global perspective of transcripts present in the YO during a basal state (Das et al., 2016). The intermolt transcriptome was assembled with fidelity, contigs assembled matched the validating Sanger sequences by 99%. The 231K contigs assembled were functionally annotated with Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, utilizing both gene function classes and pathway components to determine biological significance. During intermolt, the YO expressed components of the mTOR and TGFβ signaling pathways (Das et al., 2016), both previously identified to be involved in ecdysteroid synthesis (Abuhagr AM, in press 2016).

This project is aimed at expanding upon the intermolt YO transcriptome, to generate a database with intermolt, early premolt, mid premolt, late premolt, and postmolt animals. Identifying global transcript variation between the molt stages will aid in a better understanding of molecular regulators which drive the YO's phenotypic changes throughout the molt cycle.

Materials and Methods:

Animal Care, Library Preparation, and Sequencing

Gecarcinus lateralis adult males from the Dominican Republic were used for transcriptome sequencing. The animals were acclimated and housed under conditions reported (Covi et al., 2010). Y-organ tissues were harvested from intermolt, early premolt, mid premolt, late premolt, and postmolt animals (10 days after ecdysis).

Y-organs from intermolt animals were dissected from intact individuals, with walking legs and claws remaining on the animal. To collect premolt and postmolt samples, intact animals were induced into a precocious molt using multiple limb autotomy, MLA, with the removal of eight walking legs. R-values were used to stage the animals (Bliss and Boyer, 1964) according to the following values: early premolt 11-15, mid premolt 16-18, late premolt ≥ 19.

A competitive Enzyme Linked Immunosorbent Assay was used to quantify hemolymph ecdysteroid titers from all animals (Abuhagr et al., 2014a). Ecdysteroid titers, in pg/µl, were used to confirm intermolt and premolt animal staging with ranges of: intermolt <5, early premolt 10-40, mid premolt 40-80, late premolt >100. Animals with similar ecdysteroid titers had Y-organs pooled, two to three pairs per library, for sequencing. Each molt stage contained three biological replicates for a total of fifteen libraries. mRNA isolation, library preparation, and sequencing were completed as described in Das et. al 2016. cDNA libraries were sequenced to generate 100 bp paired-end reads at the Oklahoma Medical Research Foundation using the Illumina HiSeq^{TM} 2000 sequencer.

de novo Transcriptome Assembly

Paired-end, 100 bp reads were examined for quality using FastQC 0.11.3 from Babraham Bioinformatics (Andrews, 2010). Reads were edited using Trimmomatic 0.33 (Bolger et al., 2014). Adapters were first removed, followed by the removal of low-quality sequences. Trimmed reads ranged from 36 bp to 100 bp in length, with an average phred quality score ≥ 28 over a 4 nucleotide sliding range. Trimmed paired and unpaired reads were re-examined for quality, to verify trimming, using FastQC and used for transcriptome assembly.

High quality reads were assembled into contigs, also referred to as transcripts, using Trinity r20140413p1 (Grabherr et al., 2011). Contigs generated contained a minimum of 200 bases, with minimum kmer coverage of two. Contigs with 90% nucleotide similarity over eight bases were clustered using CD-HIT-EST 4.6.1 (Fu et al., 2012). This sequence database was referred to as the reference transcriptome.

High-quality reads from the fifteen libraries were aligned to the reference transcriptome using Bowtie 2 2.2.3 (Langmead and Salzberg, 2012). A seed setting of 1 mismatch was allowed during alignment. To estimate contig (transcript) abundance produced by Bowtie 2.0, file conversion was completed using SAMtools 0.1.18 (Li et al., 2009). eXpress 1.5.1 was used to quantify the number of fragments per transcript per library, in the form of raw counts and normalized fragments per kilobase per million reads (FPKM) (Roberts and Pachter, 2013).

Test for Differential Expression with EdgeR

The transcriptome libraries were filtered to increase the statistical power of EdgeR in identifying differentially expressed transcripts. Metagenome Analyzer 5.6.6 identified bacterial sequences from the output of BLAST against NCBI NR database (Huson et al., 2007). These contigs were removed from the transcriptome. Using EdgeR 3.12.0 transcripts with low counts were removed, using the parameter of < 1 count per million reads in < 3 libraries (Robinson et al., 2010). Transcripts containing count values over two million, consistent among biological replicates in at least one molt stage, were also removed. Removal of contigs with low and high expression facilitated the identification of differentially expressed transcripts using EdgeR.

Libraries of biological replicates were examined for similarity based on count values with a multidimensional scaling (MDS) plot in EdgeR and a non-parametric correlation using

Kendall's tau coefficient (Kendall, 1938). Libraries that did not correlate or group with their corresponding biological replicates were removed for downstream analysis.

EdgeR identified differentially expressed transcripts between the molt stages. Five pairwise comparisons were conducted: intermolt to early premolt, early premolt to mid premolt, mid premolt to late premolt, late premolt to postmolt, and postmolt to intermolt. Trimmed mean of M-values (TMM) normalization reduced inherent variabilities between libraries through scaling factors. Dispersions were estimated under a negative binomial model. The square root of the common dispersion identified the transcript variation between biological replicates. To test for differential expression, a quasi-likelihood F-test was run testing pairwise comparisons. Transcripts identified as differentially expressed between molt stages contained an FDR cutoff of ≤ 0.01.

Identify Enriched KEGG Pathways and Gene Ontology Terms

To identify enriched KEGG pathways, transcriptome contigs were annotated using BLASTx 2.2.30 (Altschul et al., 1990) against the *Drosophila melanogaster* KEGG Orthology protein database (Xie et al., 2011) with an e-value cutoff of 1e-5. This database was downloaded from KOBAS 2.0 in March 2016. The *annotate* function in KOBAS 2.0 added KEGG pathway information to contigs annotated in BLASTx using an e-value cutoff of 1e-8. In order to detect enriched KEGG pathways, the *identify* function compared contigs differentially expressed to the entire transcriptome. Enriched KEGG pathways were identified using a hypergeometric test/Fisher's exact test, Benjamini and Hochberg FDR correction, a small term cutoff of 5, and a p-value of ≤ 0.05.

To identify enriched gene ontology (GO) terms, transcriptome nucleotide sequences were annotated against the TrEMBL database, downloaded October 19, 2015, using local BLASTx (Das et al., 2016). BLASTx parameters included ten hits per contig with an e-value of $1e^{-5}$ and was run as described by Das et al. (2016). The output of BLASTx was used for functional annotation by Blast2GO Basic 3.0.8, which added GO terms to annotated contigs (Conesa et al., 2005).

Differentially expressed contigs from five pairwise comparisons were input separately in Blast2GO PRO 3.2.7 for an enrichment analysis, with the complete, functionally-annotated transcriptome serving as background (Conesa et al., 2005). A two-sided Fisher's Exact test with an FDR cutoff of 0.05 determined statistical significance. GO terms were filtered to eliminate redundant, general terms using a Fisher's Exact test with an FDR cutoff of 0.05. Over-enriched GO terms, terms with a higher percentage of differentially expressed sequences in respect to the background, with the two largest number of test group contigs were selected for downstream analysis. Mean expression values (FPKM) of contigs in above GO terms were graphed, and transcripts of interest examined for biological relevance.

Computational Notes: OSU High Performance Computer, RStudio, and perl

The Cowboy sever at Oklahoma State University was used for read trimming, contig assembly, file conversion, read mapping, expression calculations, and BLASTx annotation. The Cowboy cluster consists of 252 compute nodes, each with 32GB 1333 MHz RAM and dual hex-core 2.0 GHz CPUs. Filtering and differential expression was completed in RStudio Version 0.99.486, with R version 3.2.2 (64-bit). Sequence extraction and manipulation for enriched pathways used perl v5.22.1 and perl scripts. FileZilla 3.14.1 used for file transfer.

Results:

Molt Stage Animals and Sequencing

R-values and ecdysteroid titers confirmed that animals were properly staged into intermolt, early premolt, mid premolt, late premolt, and postmolt. Each molt stage contained three biological replicates in which three animals were pooled for each replicate; postmolt being the exception, replicate 2 and 3 each contained two pooled animals. Mean hemolymph ecdysteroid titers (Fig. 6a) and R-values (Fig. 6b) were calculated for each molt stage.

Three biological replicates for each molt stage, a total of fifteen libraries, were sequenced using Illumina. Intermolt, premolt, and postmolt animals were sequenced using the same machine at three different time points. From the fifteen libraries, 616,916,101 100 bp paired-end reads were sequenced. Initial sequencing denoted the beginning of the transcriptome assembly pipeline (Fig. 7).

de novo Transcriptome Assembly

Libraries were examined specifically for adapter content and per base sequence quality. Trimming removed adapters on the 3' end and eliminated reads with average phred score below 28 over a 4 nucleotide sliding range, with a minimum length of 35. The number of high quality reads used for downstream analysis in all fifteen libraries totaled over 568 million (Fig. 7). High quality reads were assembled into ~362K contigs using Trinity. After eliminating redundant contigs, the transcriptome consisted of ~229K contigs (Fig. 7).

As part of the data analysis metrics on the length, mapping, and expression of 229K contigs were reviewed. Contig length ranged from 201 to 21,020 bp, with a 680 bp mean length and 362 bp median length. The N50, a common metric used to assess sequence length, of all contigs was 1,081 bp. Read alignment was used to evaluate assembly quality and expression quantification. Reads mapped back to the reference transcriptome at an average rate of 94.15%. Expression data of contigs in all libraries had a median count (and FPKM) value of 0, indicating a skew in expression. This was expected, as an entire transcriptome will not be constitutively expressed at all time points. Filtering was required to eliminate contigs with low and inconsistent expression, preventing biological inference.

Test for Differential Expression with EdgeR

Following removal of 268 contigs annotated as bacteria, the transcriptome was reduced to 229,011 contigs. Of those, 41,396 transcripts were extracted for downstream analysis using EdgeR filtering parameters. Three transcripts contained disproportionately high expression consistent within all biological replicates (c196747_g1_i2 in intermolt, c241624_g3_i1 and c241624_g1_i1 in postmolt) and were removed from all libraries (Fig. 8a). For differential expression, 41,393 contigs were analyzed using EdgeR. The impact of filtering on individual libraries was measured based on the reads mapped, with postmolt library sizes being greatly reduced (Fig. 8b).

After reviewing MDS plot and Kendall's tau coefficients, postmolt 1 library was removed from the reference transcriptome. The postmolt 1 library had a low Kendall Tau's correlation coefficient, 0.45 and 0.41, when compared to the other biological replicates (Fig. 9a). In addition, postmolt 1 library groups closer to the premolt libraries than other postmolt libraries in the MDS plot, indicating this library contained a higher biological coefficient of variation and was not similar to other postmolt libraries. Therefore, postmolt 1 was removed to improve the differential expression power and detection in EdgeR (Fig. 9b).

Using EdgeR, the common dispersion of the transcriptome was calculated to be 0.2921, with the square root of 0.5405. This suggested that the average amount of variation per gene between biological replicates was 54%. EdgeR was used to identify differentially expressed transcripts between molt stages with an FDR cutoff of ≤ 0.01. A total of 13,189 unique contigs were found to be differentially expressed, representing 5.8% of all transcriptome sequences assembled. Pairwise comparisons revealed differentially expressed contigs among the following molt stages: 11,387 between intermolt and early premolt, 0 between early premolt and mid premolt, 37 between mid premolt and late premolt, 813 between late premolt and posmolt, and 3,210 between intermolt and postmolt (Fig. 10).

Identify Enriched KEGG Pathways and Gene Ontology Terms

The enrichment pipeline, including KEGG pathway analysis, is outlined in Figure 11. Enriched KEGG pathways contain a significant number of pathway components in the differentially expressed data across all molt cycle stages. The results from KOBAS 2.0 showed two enriched KEGG pathways: insect hormone biosynthesis (KO id: dme00981) and oxidative phosphorylation (KO id: dme00190). Insect hormone biosynthesis contained 37% components of the pathway, of which five genes were enriched (Fig. 12). In the oxidative phosphorylation pathway 25% of the components were identified. Enriched genes included subunits of NADH dehydrogenase, succinate dehydrogenase, cytochrome c reductase, cytochrome c oxidase, F-type ATPase, and V-type ATPase (Fig. 13).

The enrichment pipeline, with GO term analysis, is outlined in Figure 11. BLASTx against the TrEMBL database resulted in significant hits for 53,668 contigs, which comprised 23% of the entire transcriptome. Blast2GO analysis associated 9,806 unique GO terms to 44,273 transcripts. This annotated transcriptome was used to identify enriched GO terms specific to molt stage comparisons. Following GO term reduction, the following number of terms were extracted in each pairwise comparison: 4 in intermolt to early premolt (Fig. 14), 4 in mid to late premolt (Fig. 15), 8 in late premolt to postmolt (Fig. 16), and 5 in intermolt to postmolt (Fig. 17).

Between the molt stage comparisons, over-enriched GO terms are represented in Table 1. Over-enriched GO terms are defined as terms with a higher representation in the test set (molt stage comparison) when compared to the reference. The seven GO terms which contained the highest number of contigs include: cellular protein modification process (GO:0006464), chitin binding (GO:0008061)/chitin metabolic process (GO:0006030), structural molecule activity (GO:0005198), energy derivation by oxidation of organic compounds (GO:0015980), ribosome (GO:0005840), and translation (GO:0006412). Mean expression data (FPKM) of contigs assigned to a particular GO term were graphed (Fig. 18-21).

When the expression data from contigs within enriched GO terms was graphed, distinct expression profiles emerged. Expression profiles were best illustrated in the GO term cellular protein modification process, as this term contains the most contigs (Fig. 18a). Expression levels were expected to rise or fall between intermolt and early premolt, as all contigs included in this analysis were differentially expressed at this molt stage transition. Throughout the rest of the molt cycle stages however, random expression was expected, as contigs were not further clustered or grouped. However, two clear expression trends were identified: FPKM values reached a maximum at intermolt and decreased throughout premolt (Fig 18c), or contigs expression remained elevated throughout intermolt and decreased at postmolt (Fig 18b). These findings were surprising, as contigs annotated within this broad GO term were not expected to share patterns of expression across the molt stage. Similar trends are evident in other pairwise comparisons, especially in regards to postmolt. In other pairwise comparisons with postmolt, a majority of contigs show a downregulation of expression in postmolt (Fig. 20, 21).

In addition to expression profiles, pairwise comparisons identified contigs of interest for future studies. Within the GO term cellular protein modification, 5% of contigs in the test group were upregulated in from intermolt to early premolt (Fig. 18b). Mitogen-activated protein kinase ERK-A was one of the upregulated transcripts (c251373_g1_i1) of specific interest. The MAP kinase signaling pathway is reviewed to be the dominant pathway regulating ecdysteroid synthesis in *Drosophila melanogaster* (Covi et al., 2012). The kinase correlated with changing ecdysteroid levels throughout midpremolt (Fig. 22), indicating this may be a supportive pathway in *G. lateralis*, a trend found in other arthropod insects (Covi et al., 2012).

Two chitin binding proteins of interest were found based on mid premolt to late premolt comparisons. Peritrophin (c240827_g2_i1), a protein with a chitin-binding domain, is shown to produce a matrix in insects and protect from parasites, bacteria, and viruses (Du et al., 2006). Peritrophin has also been identified in various shrimp tissues and is hypothesized to aid in immune defense (Du et al., 2006). The second contig of interest, a gastrolith protein (c205464_g1_i1), is a product typically found in the gastrolith of the animal, a temporary deposit in the stomach epithelium used to store calcium and minerals (McCarthy and Skinner, 1977). The significant differential expression of these two contigs were examples of candidates for future investigation. The YO transcriptome can be used as a hypothesis-generating dataset.

Discussion:

The assembly and analysis of this YO transcriptome, containing all molt stages, is the continuation of a previous transcriptome project focused on intermolt animals (Das et al., 2016).

Utilizing a similar assembly and annotation pipeline, we generated a transcriptome containing expression data for five molt stages. This project provides the opportunity to compare transcriptome-wide expression levels between molt stages. It is the initial step in understanding global transcriptional changes in the YO correlating with molt cycle progression.

Before examining transcriptional changes, the YO data were validated using enriched KEGG pathway annotation. Twelve components of the insect hormone biosynthesis pathway were confirmed to be represented within the assembled YO transcriptome, with five differentially expressed components: Neverland, Phantom, Disembodied, Shadow, and CYP18A1 (Fig 12). All five components were specific to the molting hormone portion of the KEGG pathway, and were enzymes in the crustacean ecdysteroid biosynthetic pathway (Mykles, 2011). As the YO is the primary site for ecdysteroid synthesis in Crustacea (Lachaise et al., 1993), the presence of these particular transcripts validated that identity of the YO reference transcriptome.

The second enriched KEGG pathway also served to validate the YO transcriptome. The enriched oxidative phosphorylation KEGG pathway included multiple enzymes used in energy metabolism occurring within the mitochondria (Fig. 13). As ecdysteroid synthesis requires ATP, the enrichment of energy metabolism components was expected. The enrichment of mitochondrial components aligned with previous histological studies, which identified a large number of dense mitochondria (Arath Raghavan Sudha Devi, 2015), presumably due to large metabolic needs. The presence of an enriched oxidative phosphorylation pathway further verified the transcriptome contained valid YO contigs by identifying changes in metabolic components contributing to ATP production located in the mitochondria, required for ecdysteroid synthesis.

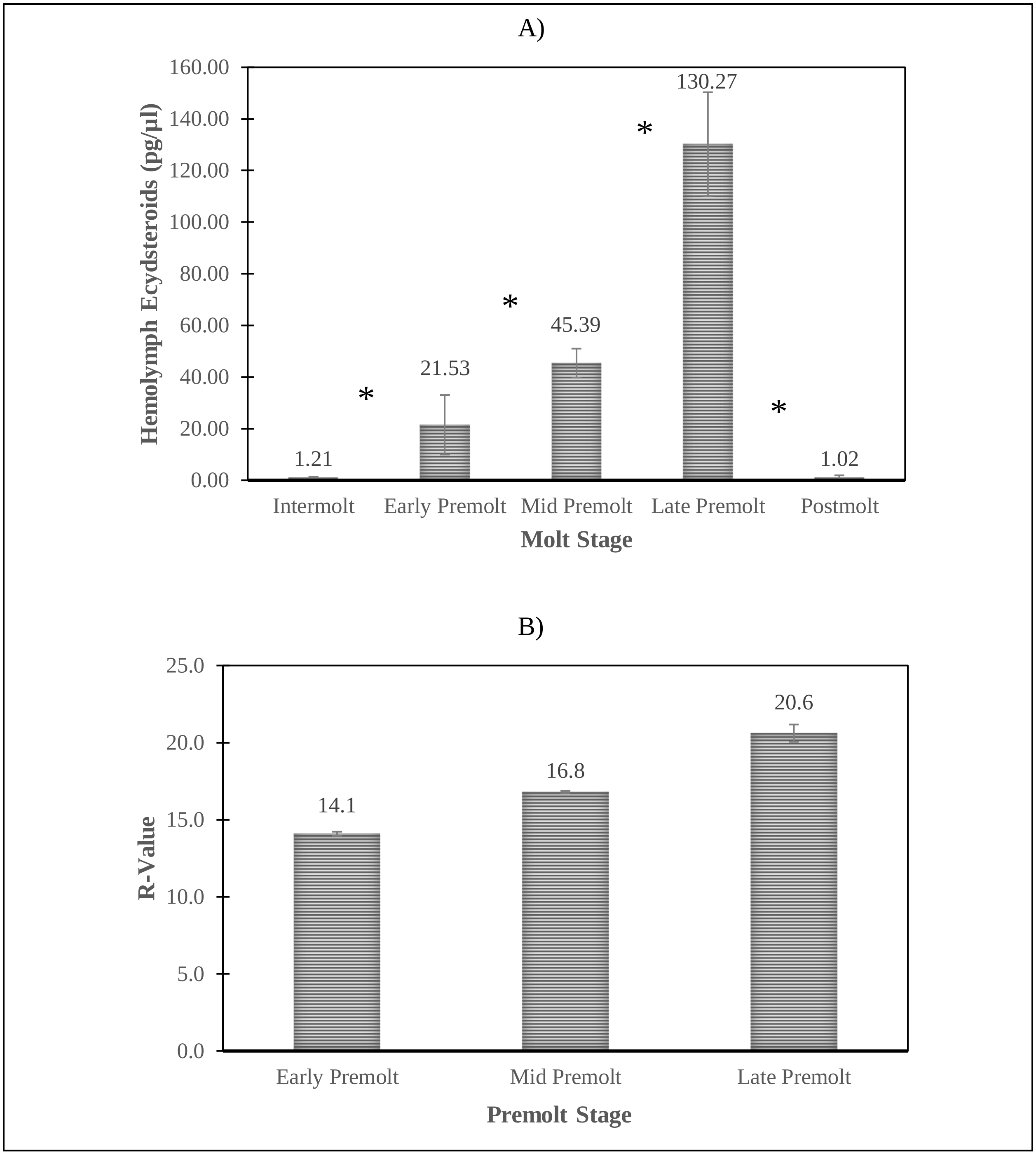
The identification of distinct expression profiles was a finding previously unobservable, due to the lack of global data. This suggests trends among gene products with similar functions, which may contribute to the phenotypic plasticity in the YO. This is especially evident in the postmolt stage. Over 50 contigs peaking in postmolt were not present in other molt stages. The contigs upregulated in postmolt were mostly bacterial sequences, and detectable because of the global downregulation of transcripts in the postmolt state. This indicates postmolt is a distinct molt stage, identifiable by transcriptional repression in the YO. The postmolt repression, taken together with the trends observed in other pairwise comparisons, indicate temporal trends in contig levels. We propose commonalities in expression profiles could serve as transcriptional markers for molt stage transitions and that a large number of gene networks may contribute to YO transitions throughout the molt cycle.

The YO transcriptome provides a perspective on large-scale changes in global transcription, but there are limitations in data analysis. Most notably, the GO terms and KEGG pathways are not specific to *G. lateralis*. Annotation is based on the assumption of orthologs between species, which need to be manually verified for contigs of interest. Even with a well annotated database in related species like *Drosophila*, assigning biological significance to *G. lateralis* transcripts and pathways can be a challenge. Function is not always analogous between the arthropods even if annotation is correct.

Another limitation is the inability to equally detect all variations in transcript expression. Contigs with large variation were identifiable as differentially expressed, however those with minimal changes will be left undetected. The biological variation between replicates is 54.05%, making it challenging to identify genes with inherently lower variability. This statistical limitation could be why regulatory pathways previously identified in YO were not identified as enriched.

Previous studies demonstrate the mechanistic target of rapamycin (mTOR) pathway is involved in regulating ecdysteroid synthesis (Abuhagr et al., 2014b). Expression of transcripts in the mTOR pathway positively correlate with increasing ecdysteroids (Abuhagr et al., 2014b). In addition, *in vitro* inhibition of mTOR using rapamycin decreased YO ecdysteroid secretions (Abuhagr et al., 2014b). mTOR and corresponding pathway components were identified within the transcriptome by manual curation, however they were not identified in the enrichment assay. We hypothesize the mTOR pathway, which was previously identified as a regulator of ecdysteroid synthesis, was not enriched within the transcriptome due to incomplete annotation and stringent parameters for differential expression and enrichment.

Despite the limitations, the assembly of the YO transcriptome and quantification of expression provides a global picture of transcriptional changes occurring throughout the molt cycle. Annotation with *Drosophila melanogaster* KEGG Orthology protein database produced enriched KEGG pathways which validated the transcriptome. Pairwise comparisons of TrEMBL annotated contigs generated lists of enriched GO terms, which were used to identify expression profiles and contigs of interest for future study. This demonstrates the usability of the transcriptome to investigate novel questions that would otherwise be overlooked due to a lack of preliminary data. Further investigation of specific transcripts will contribute to the understanding of regulatory factors controlling the variation in YO phenotype and regulating crustacean molting.



<u>Figure 6</u>: Ecdysteroid titers and R-values for transcriptome libraries. (A) Mean ecdysteroid titers of libraries. Error bars represent ± SEM. Molt stages include three biological replicates. ANOVA used to identify statistically significant groups at the .05 level, denoted by asterisk. (B) Mean R-values of premolt libraries. Error bars represent ± SEM.

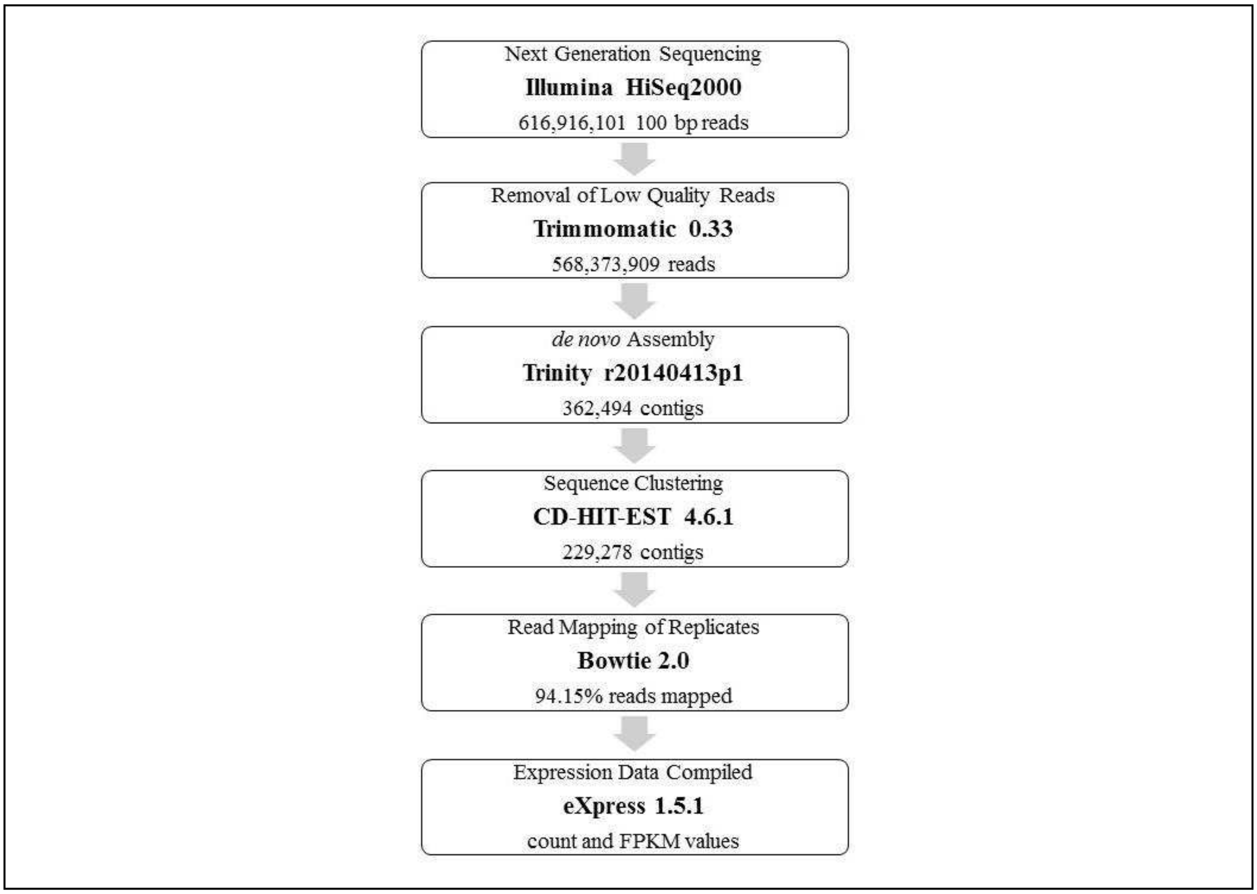
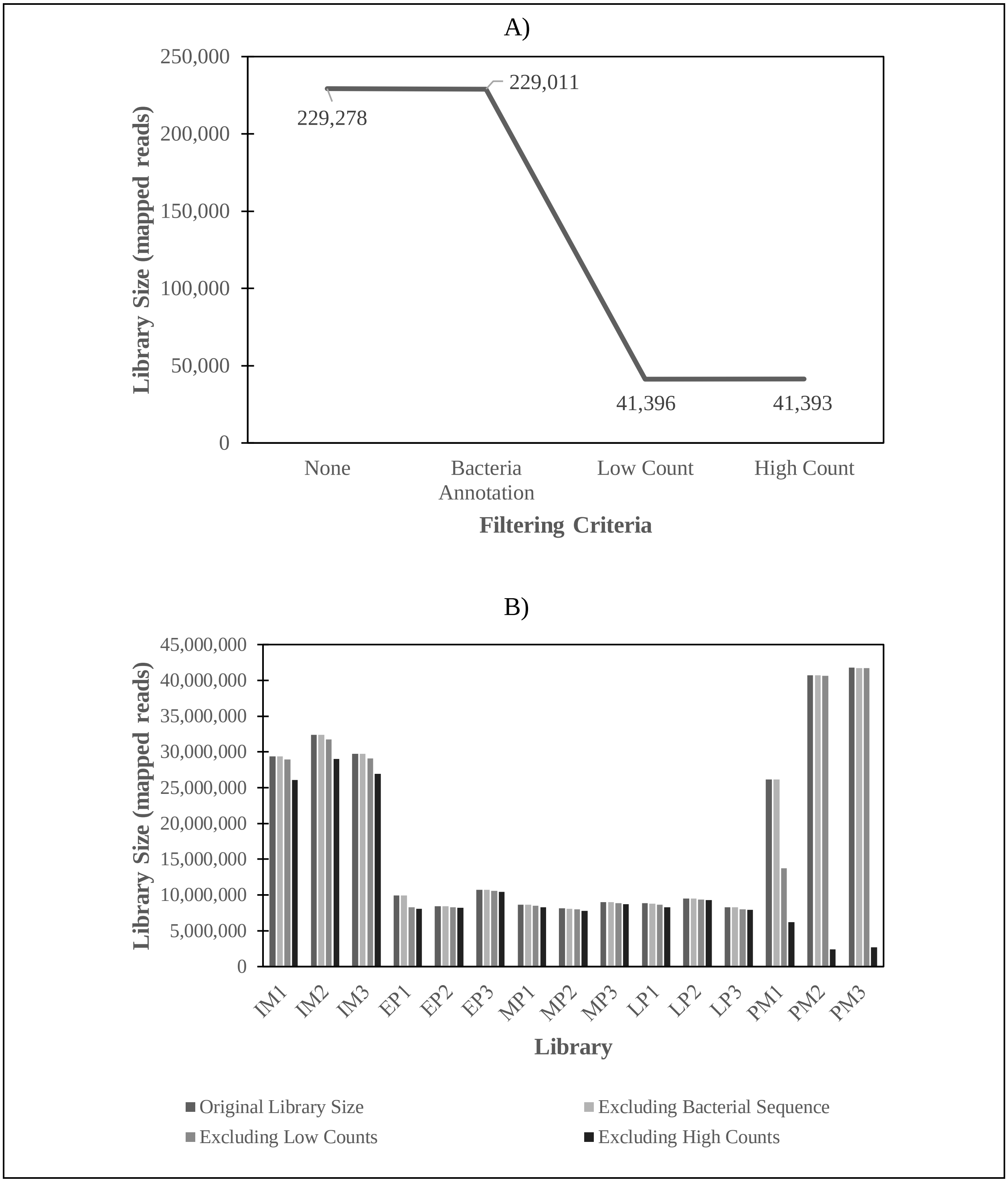


Figure 7: *G. lateralis* YO transcriptome *de novo* assembly pipeline. Platforms and versions are indicated in bold.



<u>Figure 8</u>: Contig number and library sizes after filtering. A) Contigs in transcriptome were removed based on annotation and low or high representation. B) The effect of filtering on individual library size. IM= intermolt, EP= early premolt, MP= mid premolt, LP= late premolt.

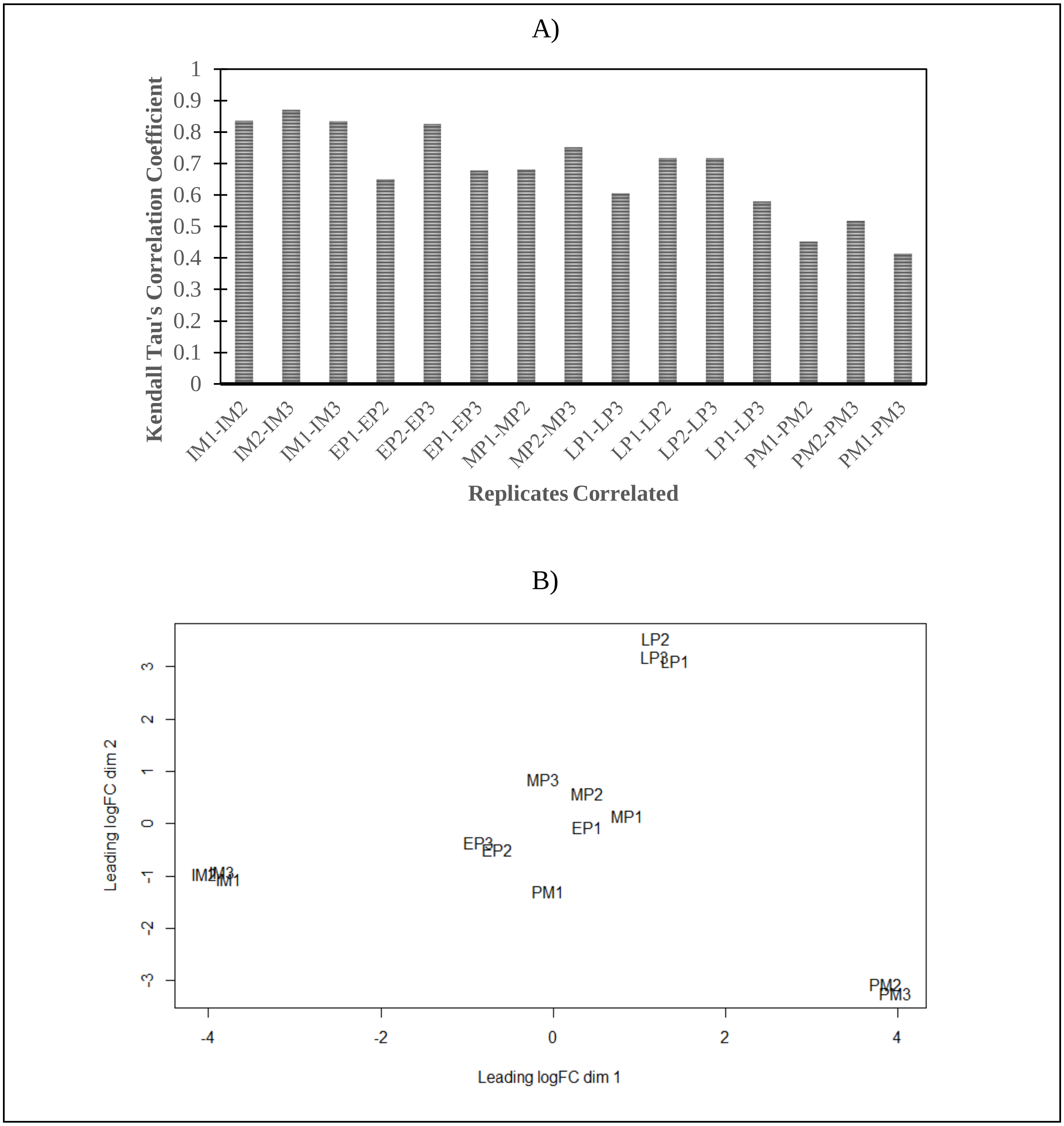


Figure 9: Comparison of biological replicates. A) Similarities among biological replicates evaluated using Kendall Tau's correlation coefficient. B) Multidimensional scaling plot (MDS) groups biological replicates together based on commonalities.

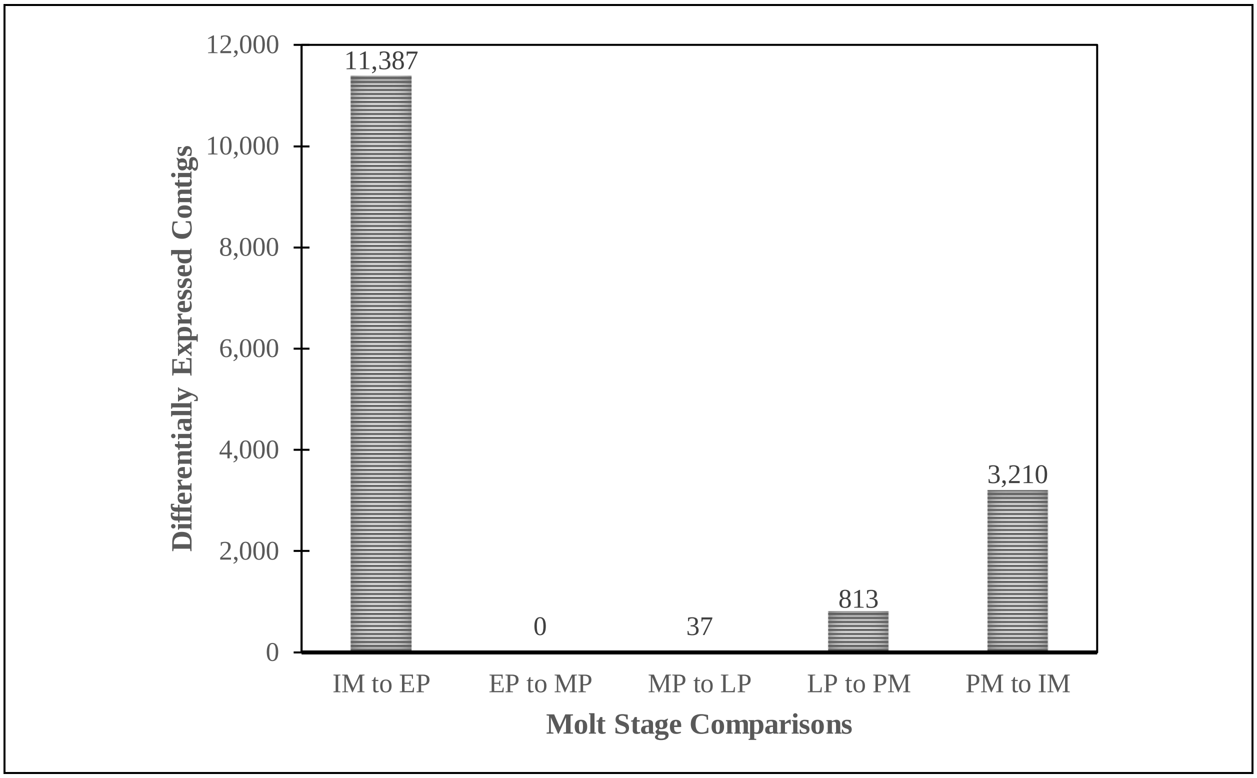
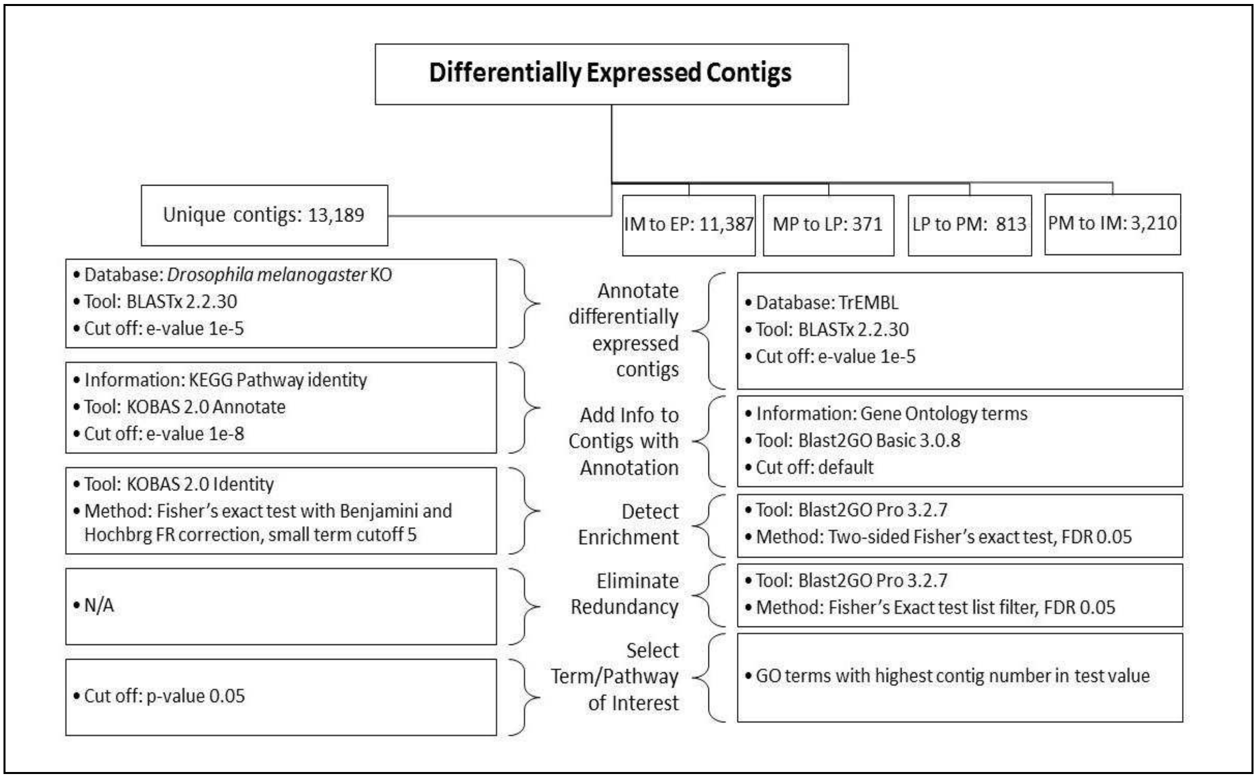


Figure 10: The number of differentially expressed contigs, at the ≤ 0.01 significance level, between molt stage comparisons. A total of 13,189 unique contigs were differentially expressed.



<u>Figure 11</u>: Enrichment pipeline. From a dataset of all differentially expressed contigs enriched KEGG pathways were identified using KOBAS 2.0. From pairwise comparisons enriched GO terms were identified using Blast2GO.

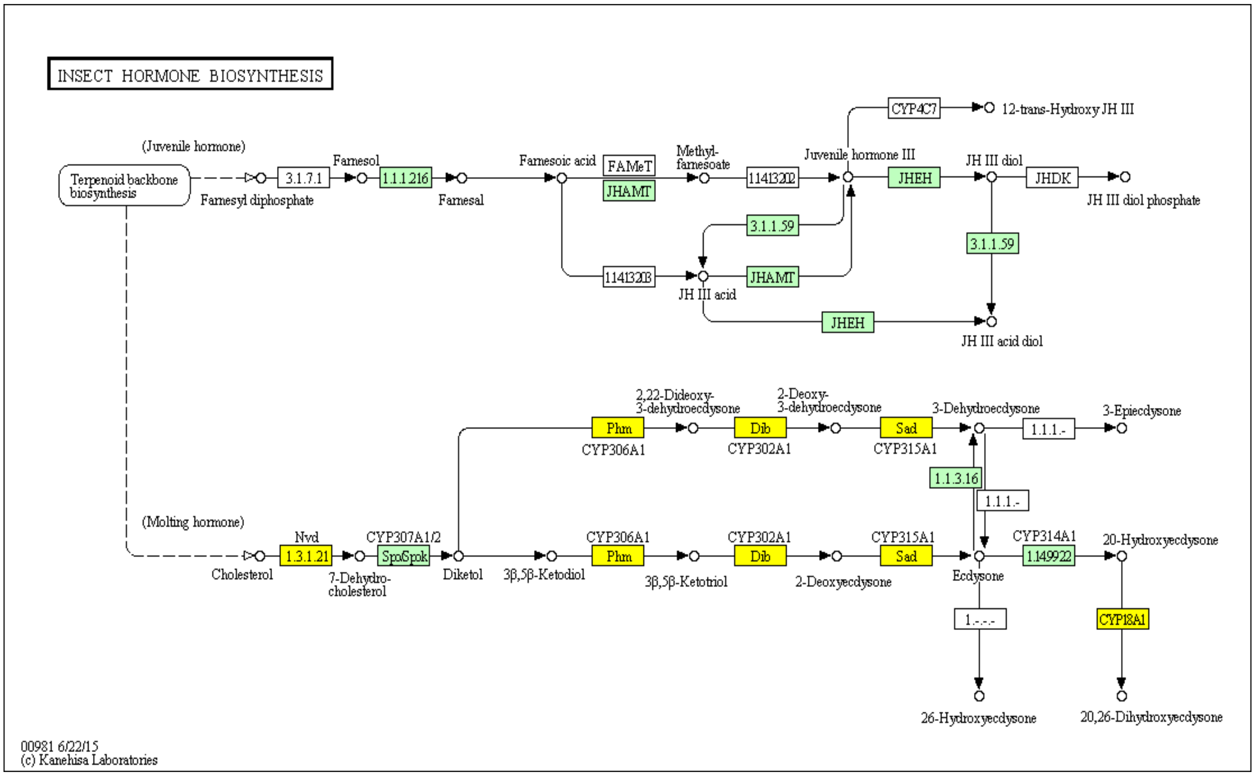


Figure 12: Enriched KEGG insect hormone biosynthesis pathway (KO id: dme00981). Yellow components indicated enrichment in differentially expressed dataset. Green components were present in the reference transcriptome dataset.

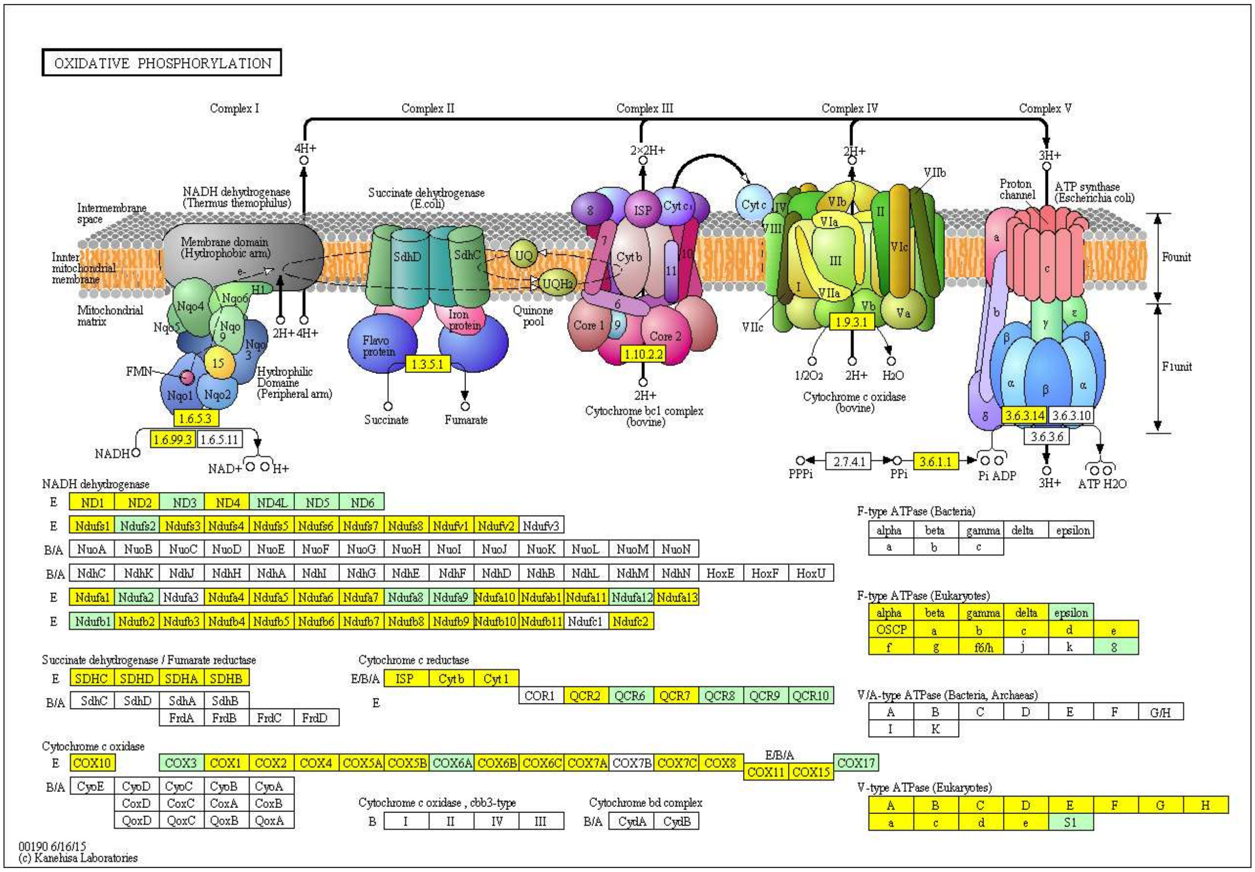


Figure 13: Enriched KEGG oxidative phosphorylation (KO id: dme00190). Yellow components indicated enrichment in differentially expressed dataset. Green components were present in the reference transcriptome dataset.

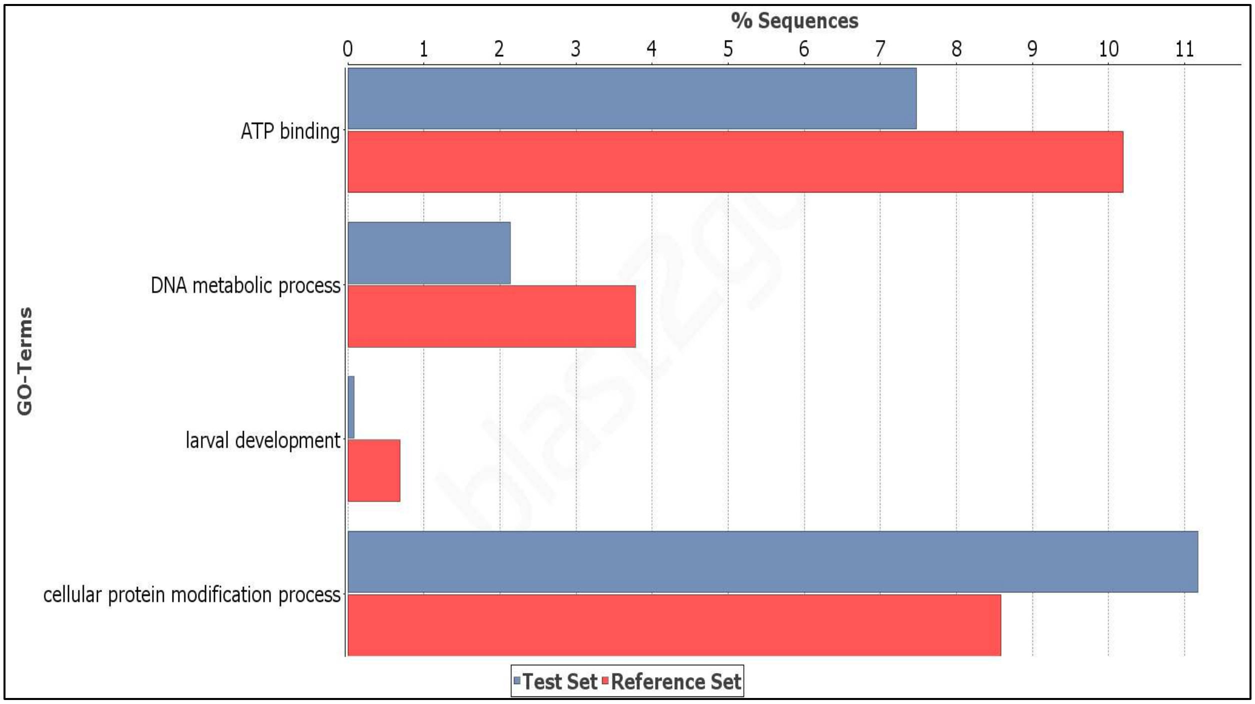


Figure 14: Enrichment of gene ontology terms, at the ≤ 0.05 significance level, between intermolt to early premolt. Test set refers to differentially expressed contigs, reference set includes reference transcriptome contigs.

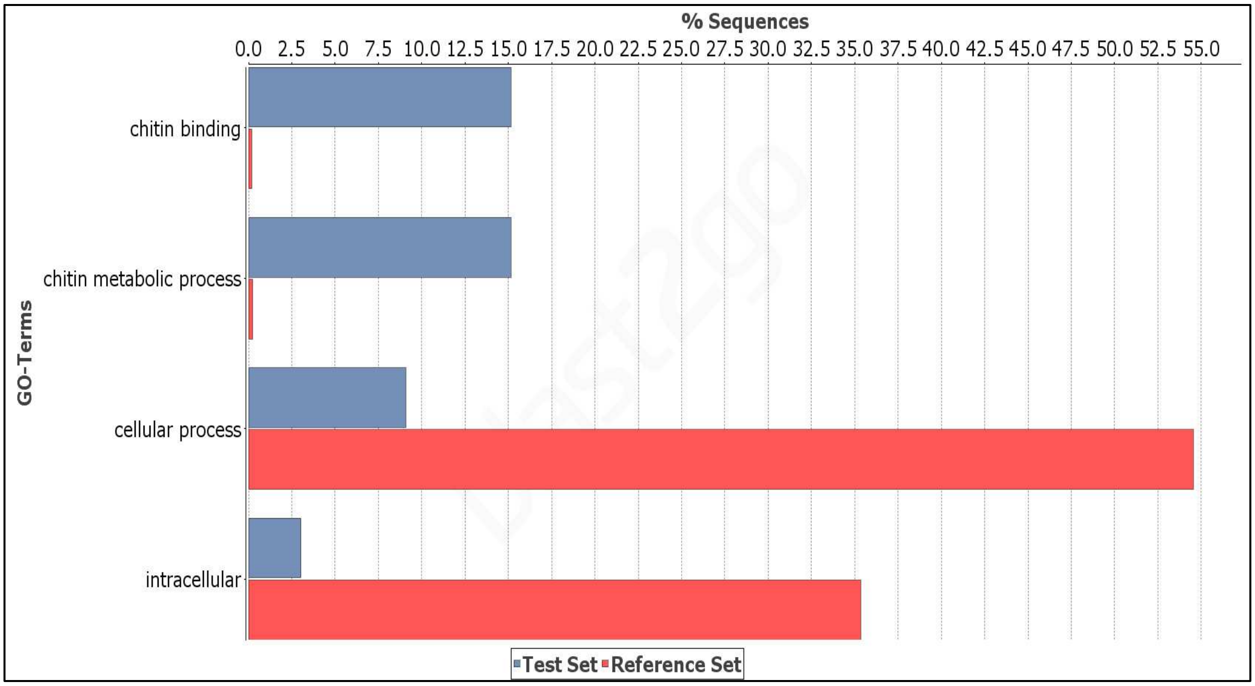


Figure 15: Enrichment of gene ontology terms, at the ≤ 0.05 significance level, between mid to late premolt. Test set refers to differentially expressed contigs, reference set includes reference transcriptome contigs.

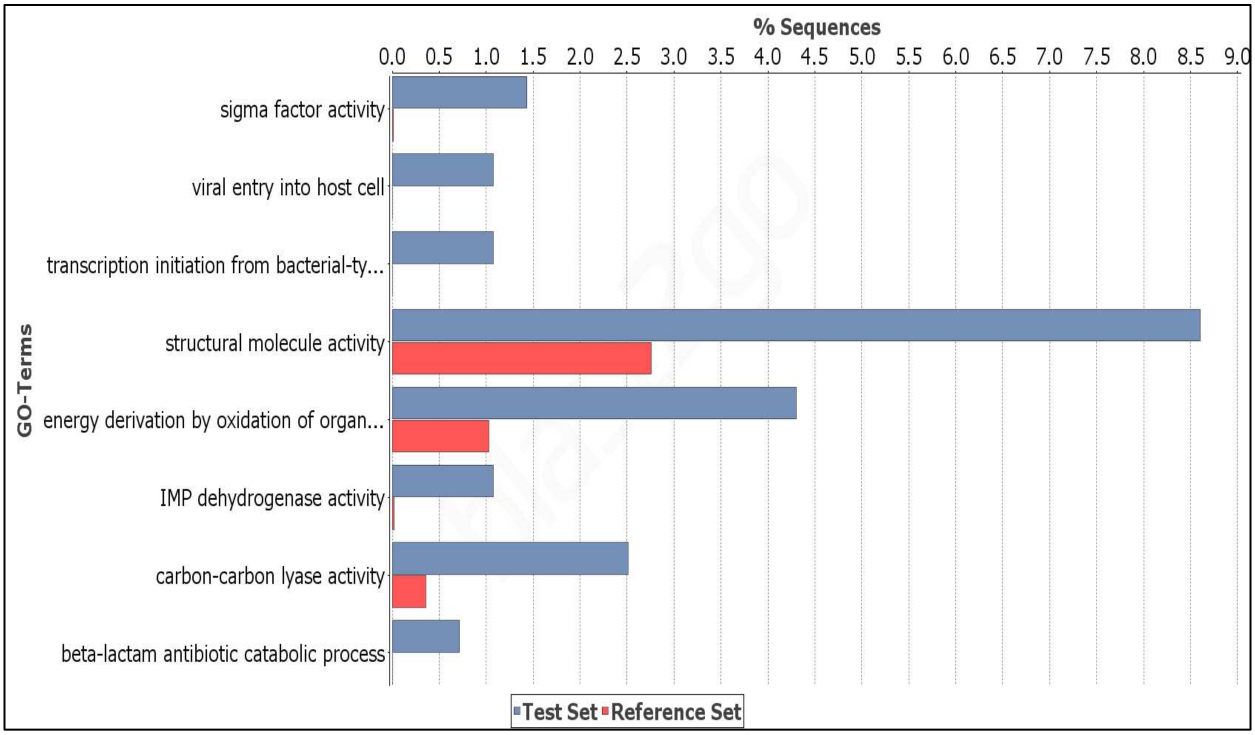


Figure 16: Enrichment of gene ontology terms, at the ≤ 0.05 significance level, between late premolt to postmolt. Test set refers to differentially expressed contigs, reference set includes reference transcriptome contigs.

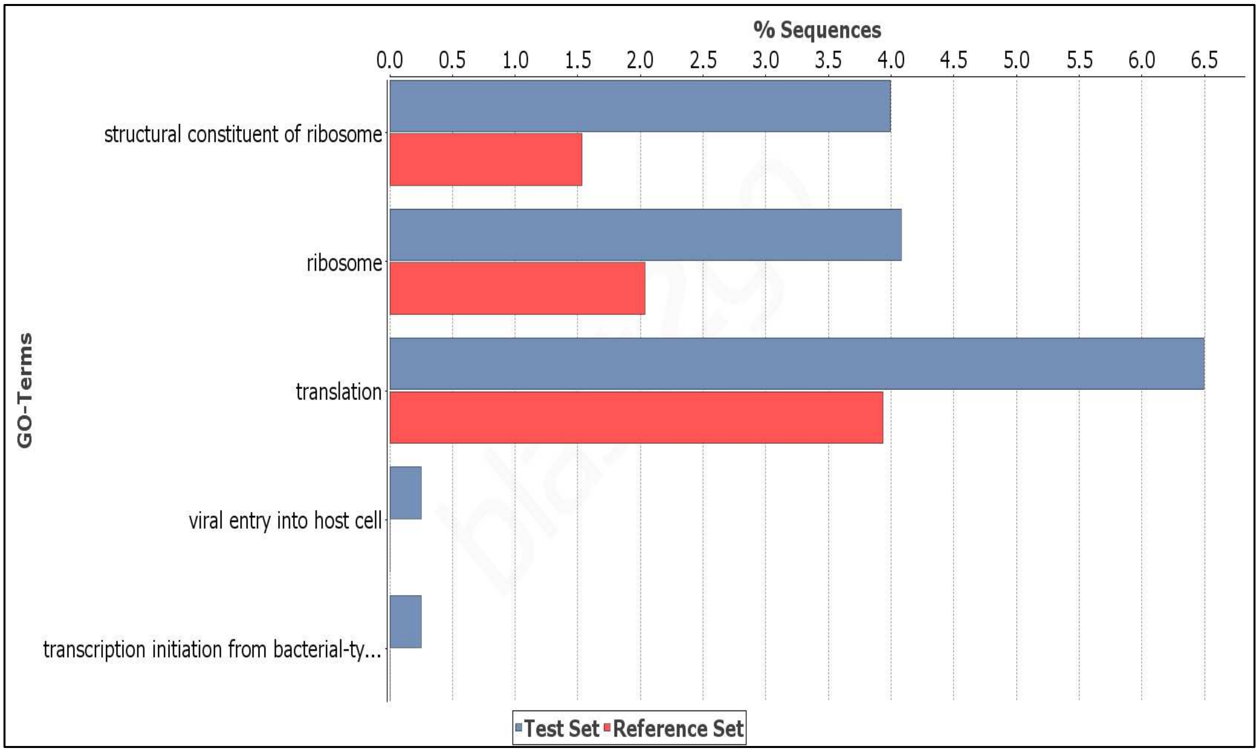
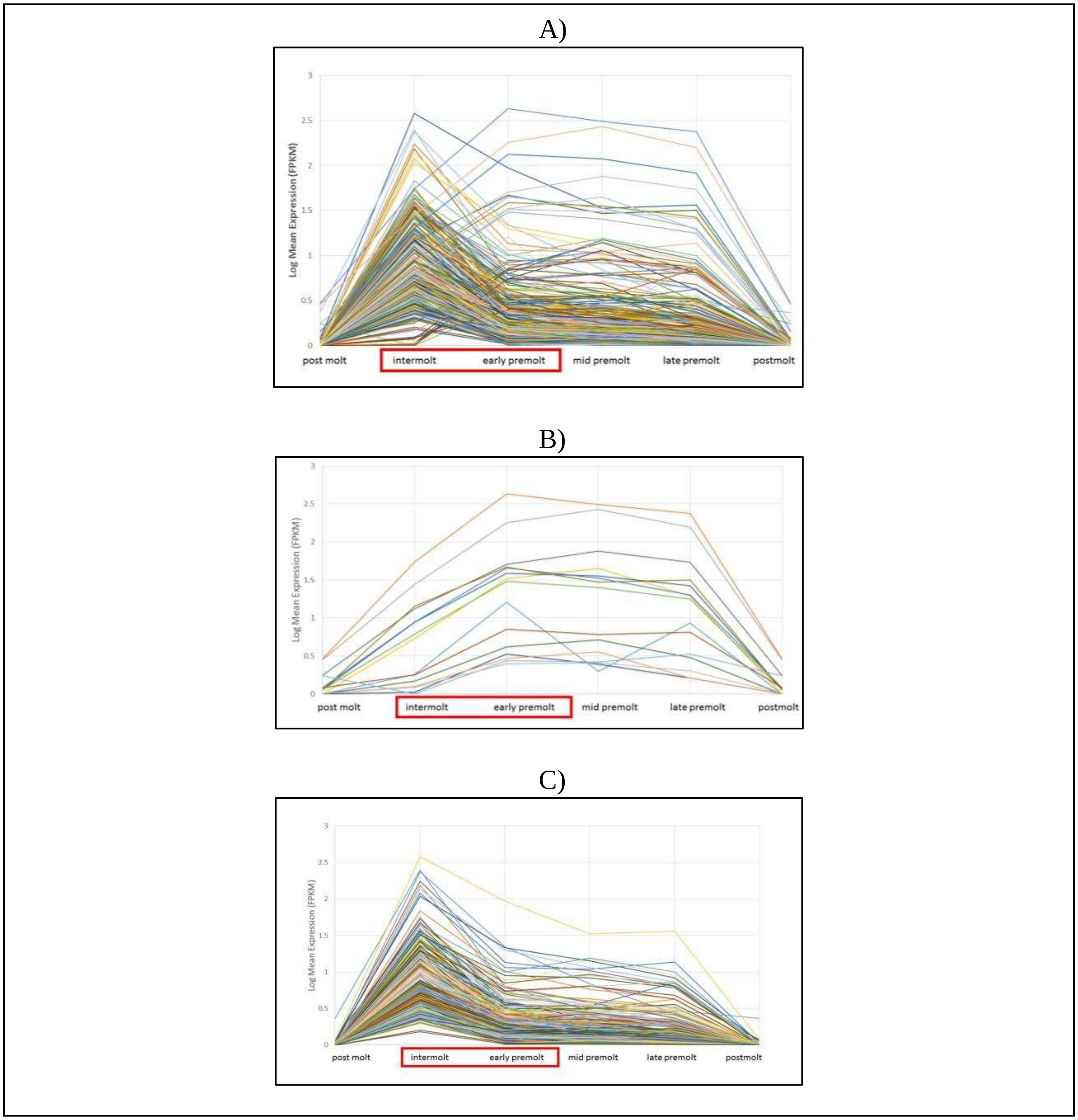


Figure 17: Enrichment of gene ontology terms, at the ≤ 0.05 significance level, between postmolt to intermolt. Test set refers to differentially expressed contigs, reference set includes reference transcriptome contigs.



<u>Figure 18</u>: Log mean expression of contigs contained in enriched GO term: cellular protein modification process (GO:0006464). Enrichment of GO term occurring in intermolt to early premolt transition. Contigs graphed were differentially expressed between molt stages in red box. A) All contigs. B) Contigs upregulated in early premolt. C) Contigs upregulated in intermolt.

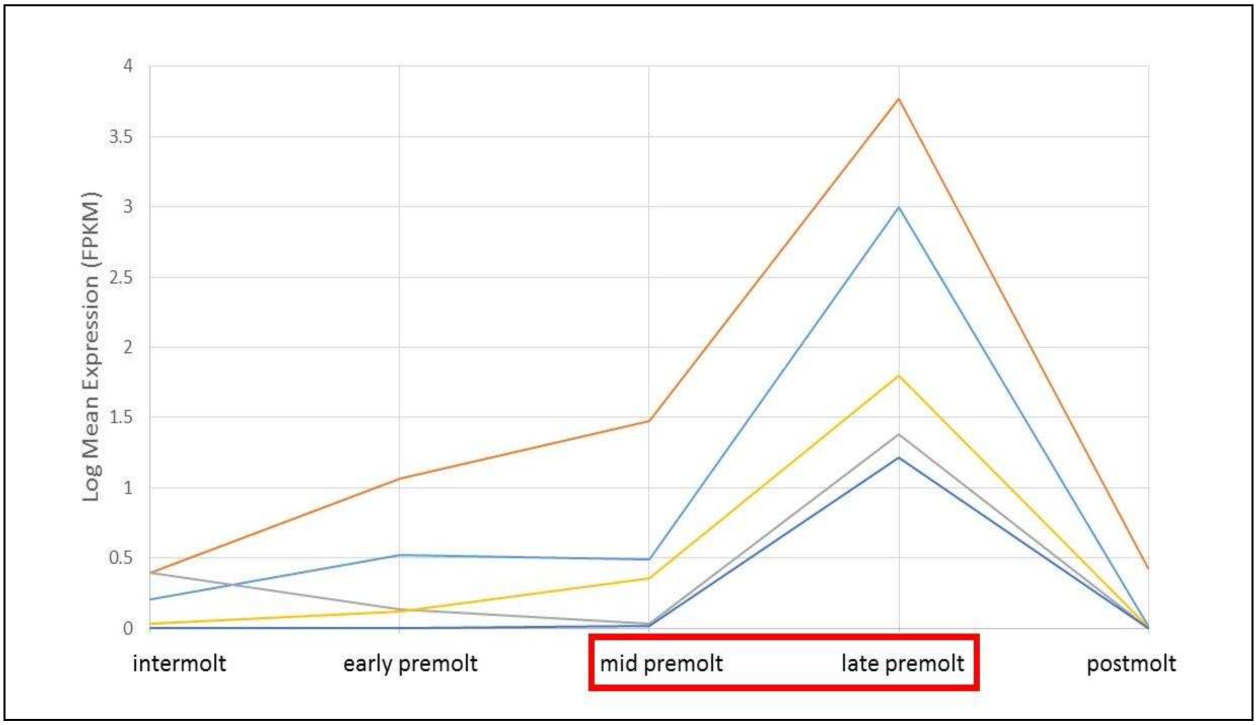
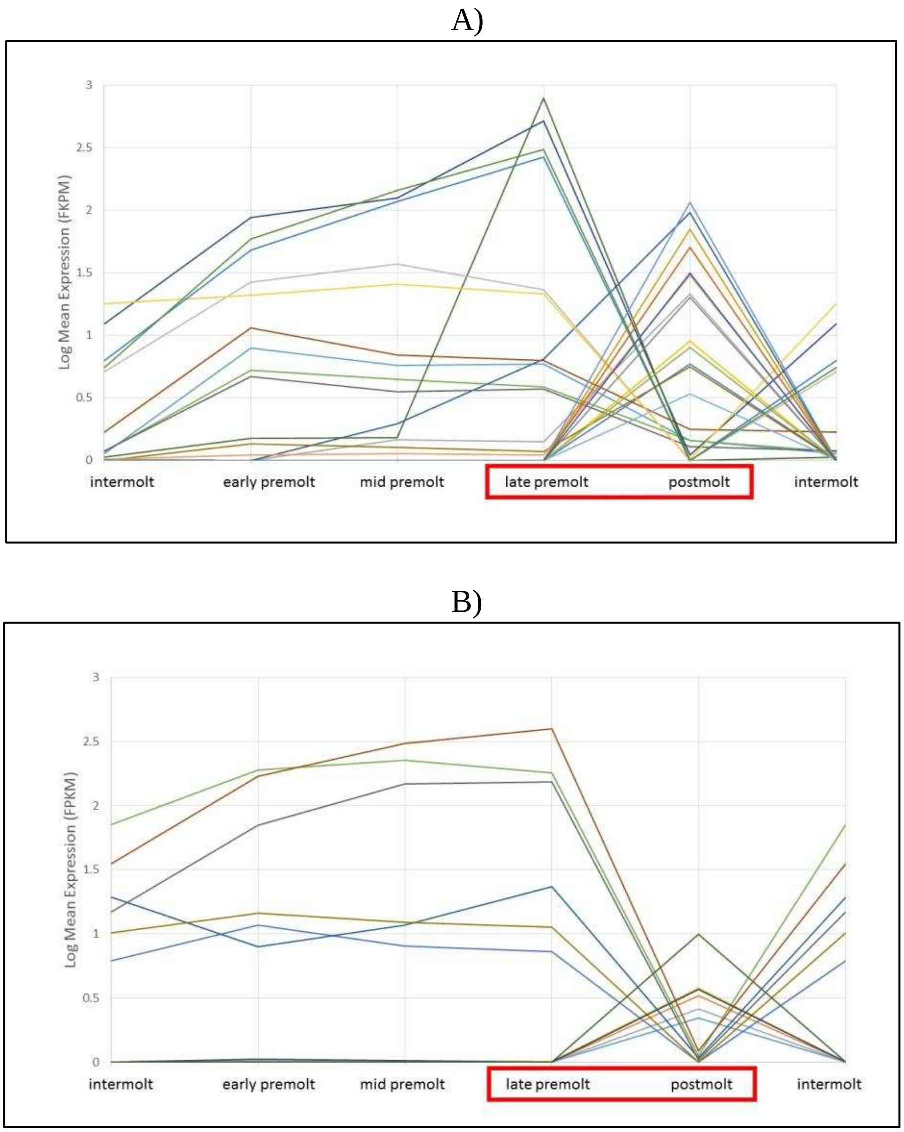
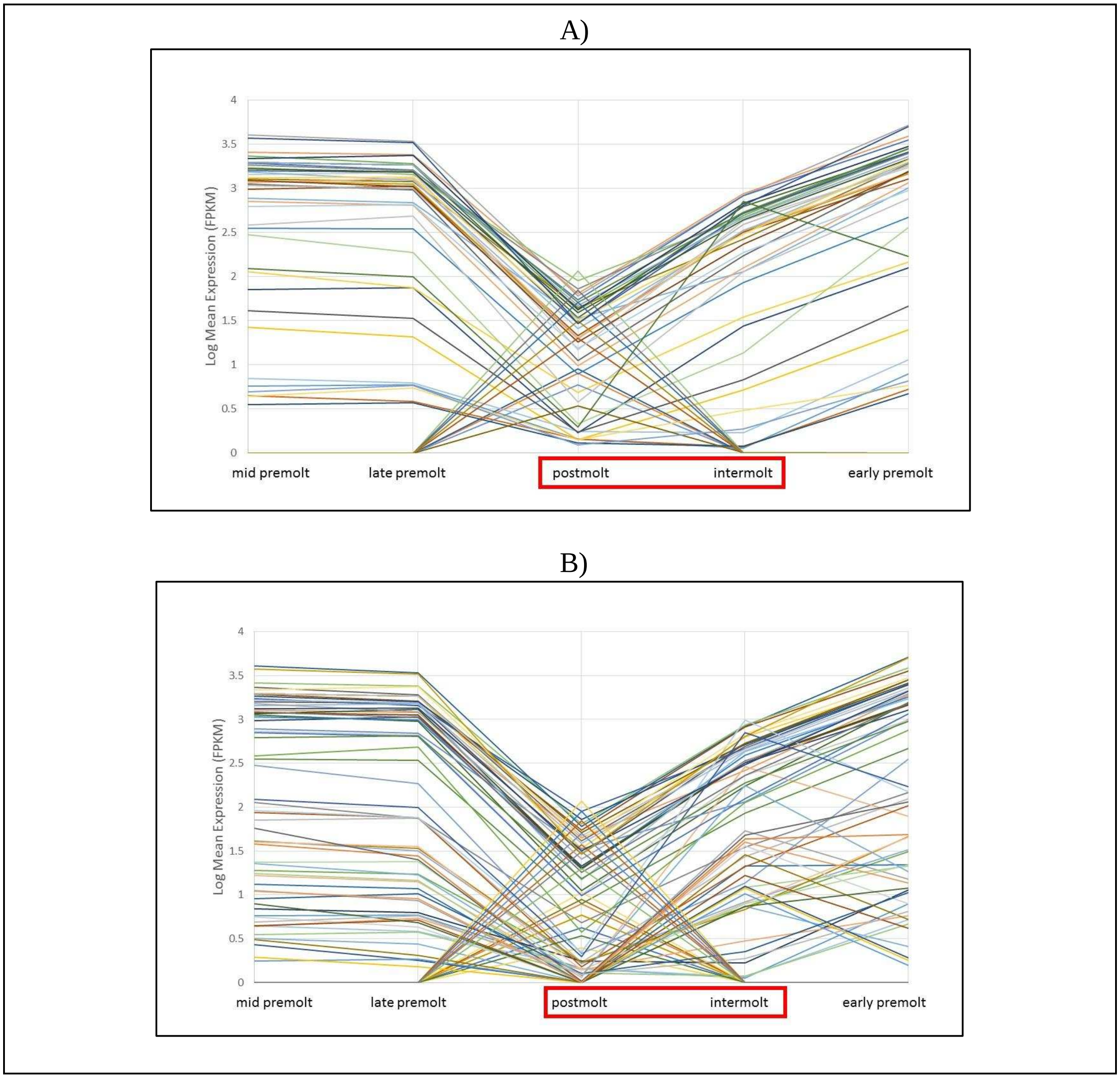


Figure 19: Log mean expression of contigs contained in enriched GO term: chitin binding (GO:0008061) and chitin metabolic process (GO:0006030). Contigs were identical for both terms. Contigs graphed were differentially expressed between molt stages in red box.



<u>Figure 20</u>: Log mean expression of contigs contained in enriched GO term: A) structural molecule activity (GO:0005198) and B) energy derivation by oxidation of organic compounds (GO:0015980). Contigs graphed were differentially expressed between molt stages in red box.



<u>Figure 21</u>: Log mean expression of contigs contained in enriched GO term: A) ribosome (GO:0005840) and B) translation (GO:0006412). Contigs graphed were differentially expressed between molt stages in red box.

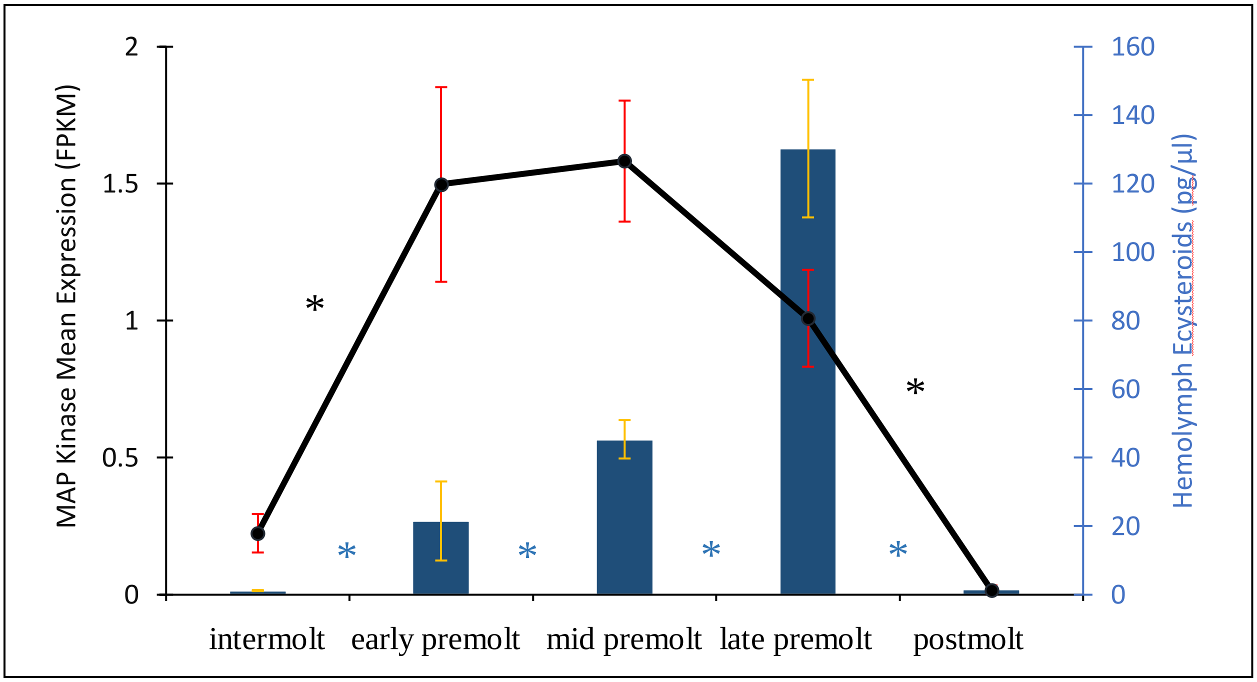


Figure 22: MAP kinase (c251373_g1_i1) and hemolymph ecdysteroid titers compared throughout the molt cycle. Error bars represent ± SEM. ANOVA used to identify statistically significant groups, when compared to the molt cycle stages at the .05 level. Statistical significance between molt cycle stages and MAP Kinase denoted by a black asterisk. Statistical significance between molt cycle stages and hemolymph ecdysteroids denoted by a blue asterisk.

Table 1: Summary of over enriched GO terms between molt stage transitions, when compared to reference transcriptome. GO terms highlighted in red were selected for additional analysis.

Transition	Contigs Differentially Expressed	Reduced GO Terms	Over Enriched GO Terms	Over Enriched Go Terms		
				GO Number	GO Term Name	Number of Contigs in Test Group
Intermolt to Early Premolt	11,387	4	1	GO:0006464	cellular protein modification process	272
Early Premolt to Mid Premolt	0	N/A				
Mid Premolt to Late Premolt	371	4	2	GO:0008061	chitin binding	5
				GO:0006030	chitin metabolic process	5
Late Premolt to Postmolt	813	8	8	GO:0016987	sigma factor activity	4
				GO:0046718	viral entry into host cell	3
				GO:0001123	transcription initiation form bacterial-type RNA polymerase promoter	3
				GO:0005198	structural molecule activity	24
				GO:0015980	energy derivation by oxidation of organic compounds	12
				GO:0003938	IMP dehydrogenase activity	3
				GO:0016830	carbon-carbon lyase activity	7
				GO:0030655	beta-lactam antibiotic catabolic process	2
Postmolt to Intermolt	3,210	5	5	GO:0003735	structural constituent of ribosome	48
				GO:0005840	ribosome	49
				GO:0006412	translation	78
				GO:0046718	viral entry into host cell	3
				GO:0001123	transcription initation from bacterial-type RNA polymerase promoter	3

REFERENCES

Abuhagr, A.M., Blindert, J.L., Nimitkul, S., Zander, I.A., LaBere, S.M., Chang, S.A., MacLea, K.S., Chang, E.S., Mykles, D.L., 2014. Molt regulation in green and red color morphs of the crab *Carcinus maenas*: gene expression of molt-inhibiting hormone signaling components. Journal of Experimental Biology 217, 1830-1830.

Abuhagr AM, K.S.M., Megan R. Mudron, Sharon A. Chang, Ernest S. Chang, Donald L. Mykles, in press 2016. Roles of mechanistic target of rapamycin and transforming growth factor-β signaling in the molting gland (Y-organ) of the blackback land crab, *Gecarcinus lateralis*. Comparative Biochemistry and Physiology, Part A.

Abuhagr, A.M., MacLea, K.S., Chang, E.S., Mykles, D.L., 2014b. Mechanistic target of rapamycin (mTOR) signaling genes in decapod crustaceans: Cloning and tissue expression of mTOR, Akt, Rheb, and p70 S6 kinase in the green crab, *Carcinus maenas*, and blackback land crab, *Gecarcinus lateralis*. Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology 168, 25-39.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403-410.

Andrews, S., 2010. "FastQC: A quality control tool for high throughput sequence data.", http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arath Raghavan Sudha Devi, M.K.S., Sagar, B.K.C., 2015. Light and electron microscopic studies on the Y organ of the freshwater crab *Travancoriana schirnerae*. Journal of Microscopy and Ultrastructure 3, 161-168.

Bliss, D.E., Boyer, J.R., 1964. Environmental Regulation of Growth in the Decapod Crustacean *Gecarcinus lateralis*. General and Comparative Endocrinology 4, 15-41.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Chang, E.S., Mykles, D.L., 2011. Regulation of crustacean molting: a review and our perspectives. General and Comparative Endocrinology 172, 323-330.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674-3676.

Covi, J.A., Bader, B.D., Chang, E.S., Mykles, D.L., 2010. Molt cycle regulation of protein synthesis in skeletal muscle of the blackback land crab, Gecarcinus lateralis, and the differential expression of a myostatin-like factor during atrophy induced by molting or unweighting. Journal of Experimental Biology 213, 172-183.

Covi, J.A., Chang, E.S., Mykles, D.L., 2009. Conserved role of cyclic nucleotides in the regulation of ecdysteroidogenesis by the crustacean molting gland. Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology 152, 470-477.

Covi, J.A., Chang, E.S., Mykles, D.L., 2012. Neuropeptide signaling mechanisms in crustacean and insect molting glands. Invertebrate Reproduction & Development 56, 33-49.

Das, S., Pitts, N.L., Mudron, M.R., Durica, D.S., Mykles, D.L., 2016. Transcriptome analysis of the molting gland (Y-organ) from the blackback land crab, Gecarcinus lateralis. Comparative Biochemistry and Physiology D-Genomics & Proteomics 17, 26-40.

Du, X.J., Wang, J.X., Liu, N., Zhao, X.F., Li, F.H., Xiang, J.H., 2006. Identification and molecular characterization of a peritrophin-like protein from fleshy prawn (Fenneropenaeus chinensis). Molecular Immunology 43, 1633-1644.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics 28, 3150-3152.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29, 644-U130.

Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Research 17, 377-386.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81-93.

Lachaise, F., Leroux, A., Hubert, M., Lafont, R., 1993. The Molting Gland of Crustaceans: Localization, Activity, And Endocrine Control (A Review). Journal of Crustacean Biology 13, 198-234.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-U354.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data, P., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

McCarthy, J.F., Skinner, D.M., 1977. Interactions Between Molting and Regeneration. 2. Proecdysial Changes in Serum Ecdysone Titers, Gastrolith Formation, and Limb Regeneration following Molt Induction by Limb Autotomy and/or Eyestalk Removal in the Land Crab, *Gecarcinus lateralis*. General and Comparative Endocrinology 33, 278-292.

Mykles, D.L., 2011. Ecdysteroid metabolism in crustaceans. Journal of Steroid Biochemistry and Molecular Biology 127, 196-203.

Roberts, A., Pachter, L., 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods 10, 71-U99.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Skinner, D.M., 1985. The Biology of Crustacea. Acadmeic Press.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.-Y., Wei, L., 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Research 39, W316-W322.

CHAPTER THREE: GENOME

Introduction:

In 2001 the draft human genome was completed (Lander et al., 2001). The draft was an international, collaborative effort, which required eleven years to sequence, (Lander et al., 2001) four additional years to finalize, and 100 million dollars to complete (Wetterstrand, 2016). Researchers assembled 2.85 billion nucleotides and produced a draft genome with 99% euchromatic coverage identifying 30,000 to 40,000 protein-coding genes (Collins et al., 2004). The success of the Human Genome Project demonstrated the feasibility of generating genome assemblies for complex eukaryotes. Genomes are now being used to address more complex and diverse questions. In 2015, the 1000 Genomes Project was completed; sequences from 2,504 people were used to identify differences between individuals to better understand genetic variation (Altshuler et al., 2015).

The increase in genome assemblies and applications is evident among all species. As of April 2016, NCBI hosts 2,980 eukaryotic genome databases, ranging from *Acanthamoeba astronyxis* to *Zymoseptoria tritici*. This is a massive increase from 2013, when NCBI contained only 644 genome assemblies (Ellegren, 2014). Although projects are biased toward mammals and crop plants (Ellegren, 2014), the diversity in species sequenced is growing, including the subphylum *Crustacea*. Recent crustacean genomes assembled include the water flea, *Daphnia pulex* (Colbourne et al., 2011); cherry shrimp, *Neocaridina denticulate* (Kenny et al., 2014); and the Chinese mitten crab, *Eriocheir sinensis* (Song et al., 2016).

Largely driving the boom in genomic projects is the decrease in sequencing price. Fifteen years after the first human genome was completed, sequencing costs have dropped to just over 1,000 dollars (Wetterstrand, 2016). The development of, and continued innovations in, massively parallel sequencers make genome sequencing more affordable and accessible. Researchers with the means to sequence genomic data usually lack the man-power of the Human Genome Project, making draft genomes dependent upon computational tools for assembly and annotation. Growing in popularity are *ab initio* tools, which operate independently of biological evidence by using mathematical predictions (Yandell and Ence, 2012). As a result, quality tools for genome assembly and annotation, not sequencing, are the current limiting factors in high quality genome production (Yandell and Ence, 2012). In 2010, the giant panda genome, *Ailuropoda melanoleura*, was assembled using only next-generation sequencing (NGS) methods and relied heavily on *ab initio* tools, in conjunction with evidence-based methods, for analysis (Li et al., 2010). The affordability of sequencing, along with the improvement of bioinformatics packages have contributed to the increase in genome projects.

The *G. lateralis* draft genome project was driven largely by sequence affordability (Wetterstrand, 2016) and the development of *de novo* tools (Yandell and Ence, 2012). The purpose was to sequence and construct a draft genome for *G. lateralis* to complement the Y-organ (YO) transcriptome discussed in Chapter 2. A quality genome is needed to computationally validate and annotate transcriptome sequences. Global comparison of genome and transcriptome information is a powerful tool in understanding transcriptional changes occurring in the molting gland by combining annotated transcripts with regulatory, noncoding regions and expression data. This project is part of a larger goal to understand how changes in YO gene expression modulate molting in *G. lateralis.*

Materials and Methods:

Animal Care, Library Preparation, and Sequencing

One adult male *Gecarcinus lateralis* from the Dominican Republic was used for genome sequencing. The animal was housed in a communal cage under conditions described by Covi et al. (2010). Claw muscle was dissected and genomic DNA extracted. The library was prepared for Next Generation Sequencing using the Nextera Mate Pair Sample Preparation Kit from Illumina, and sequenced with Illumina HiSeq 2000 at the Oklahoma Medical Research Foundation. The library was divided into two technical replicates, each in different lanes, in the same sequencing reaction.

Genome Assembly:

The Illumina HiSeq 2000 sequencer was used to produce raw paired-end 100-bp reads. Technical replicates from two sequencing reactions, nicknamed L01 and L08, were examined for quality and adapter presence using Babraham Bioinformatics' FastQC (Andrews, 2010). Trimmomatic 0.33 was used to remove Nextera Paired-End adapters from L08 and TruSeq2 Paired-End adapters from L01 (Bolger et al., 2014). Both libraries were trimmed using a phred score, a quality measurement assigned to each base sequenced, with a cutoff of an average phred quality score ≥ 20 over a four-nucleotide sliding range, with a minimum length of 35 nucleotides. The resulting paired-end trimmed reads were used for genome assembly.

Paired-end reads from technical replicate L08 and L01 were assembled using Ray 2.3.1 (Boisvert et al., 2010). Three separate draft genomes were generated, using the individual replicates separately and after pooling reads from both reactions. The default settings in Ray were used, with an increased minimum k-mer, fixed sequence length used in assembling reads, of 37 nucleotides. The seed parameters used required at least one read coverage and a length of 100 nucleotides. Contigs had a minimum length of 100 nucleotides before being combined into scaffolds, which had a minimum length of 147 nucleotides. Assembly metrics were extracted from Ray (Boisvert et al., 2010).

As another measurement of assembly quality, reads were aligned to the draft genome and tallied using Bowtie 2.2.6 (Langmead and Salzberg, 2012), in conjunction with Samtools 1.3 for formatting (Li et al., 2009). Tablet 1.15.09.01 was used to visualize scaffolds with corresponding contigs and reads (Milne et al., 2013). Mapping and assembly metrics were used to identify the optimal assembly to serve as the *G. lateralis* draft genome.

Estimate of Genome Size

The size of the *G. lateralis* genome was calculated using an equation for non-model organisms with *de novo* genome assemblies (Li et al., 2010; Song et al., 2016); the total sequence length is divided by read sequence depth to estimate genome size (Liu, 2013). Total sequence length was calculated by multiplying the average read length by the total number of reads. Read sequence depth required identifying a peak k-mer frequency. Forward and reverse paired-end reads were concatenated into one file, in which standard 17-bp long k-mers (Li et al., 2010; Song et al., 2016) were identified and counted using Jellyfish 2.1.4 (Marcais and Kingsford, 2011). A histogram of k-mers identified the highest k-mer frequency (M). This value, along with mean read length (L), and k-mer length (K), were used to calculate sequence depth (N) with the equation: $M = N*(L-K+1)/L$ (Li et al., 2010). The total sequence length was divided by sequence depth.

Local Comparison of Genome and Transcriptome

A local draft genome BLAST database comprised of scaffolds was generated using prfectBLAST2.0 (Santiago-Sotelo and Humberto Ramirez-Prado, 2012). To verify the assemblies of the draft genome and reference transcriptome, representative transcriptome contigs were queried against the draft genome. Local BLASTn searches against the draft genome in prfectBLAST identified and extracted scaffolds of interest. Reference transcriptome contigs and draft genome scaffolds were manually compared and aligned in prfectBLAST as verification.

Computational Notes: OSU High Performance Computer and perl

The Cowboy server at Oklahoma State University was used for read trimming, genome assembly, file conversion, and read mapping. The Cowboy cluster consisted of 252 compute nodes, each with 32GB 1333 MHz RAM and dual hex-core 2.0 GHz CPUs. File manipulation for manual comparison used perl v5.22.1 and perl scripts.

Results:

Genome Assembly

Sequencing of technical replicates L01 and L08 produced 407K 100-bp paired-end reads. Replicate L01 contained 13% more reads than L08 (Table 2). When compared in FastQC, the replicates appeared comparable in sequence quality, GC content, and adapter presence (Table 2). After trimming both replicates, adapters were eliminated and sequence quality improved, verified by an additional FastQC analysis. A total of 260K high-quality reads (Table 2), averaging 87 bp in length, were used for assembly.

The draft genome assembly pipeline (Fig. 23) was used to generate three assemblies: L01, L08, and pooled. Assembly variation is evident in the scaffold metrics. Scaffolds at least 500 bp in length were selected to assess genome assembly due to usability of extended scaffolds (Table 3). The pooled assembly, with reads from L01 and L08, generated more scaffolds with a higher N50 (Table 3), indicating the number of reads and genome quality were correlated. This was confirmed when comparing individual assemblies; L01 contained more raw reads than L08 (Table 2) and produced a larger number of longer scaffolds than L08 (Table 3).

The need for a large quantity of reads to assemble a quality draft genome was also evident in the rate of read alignment to the genome scaffolds. This read mapping is not to be used for quantification, as it is for the transcriptome, rather it serves as an additional measure of assembly quality. Again, the pooled assembly had the highest percentage of reads mapped back to the assembly, 75.87%, than either replicate separately (Table 3). The higher mapping rate and more favorable metrics resulted in the pooled assembly being used as the *G. lateralis* draft genome and for downstream analysis.

Estimate Genome Size

To calculate genome size, a histogram identifying the highest k-mer frequency was generated (Fig. 24). With a standard 17-mer parameter, the frequency peak was identified at 12, meaning 17 bp long k-mers were most commonly found in the draft genome 12 times (Fig. 24). The most common 17-mer depth, 12, is found 2.3 million times in the draft genome (Fig. 24). Average read length (86.66 bp) and number of reads (520K) were used in the estimation equation (Fig. 25). Sequencing depth was estimated at 14.72, indicating each base was read and reported approximately 15 times. The genome size of *G. lateralis* was calculated at 3.07 Gb (Fig. 25).

Compare Genome and Transcriptome Sequence

The transcriptome contig for Rheb, Ras homolog enriched in brain (c280169_g1_i1), was queried against the draft genome database using a local BLASTn search in prfectBLAST2.0. The scaffolds that aligned with Rheb (scaffold-14687, scaffold-25268, scaffold-20402, and scaffold-628615) were used to elongate the sequence (Fig. 26). Manual alignment of the scaffolds identified four introns in the open reading frame and extended the 5' and 3' sequences by 6,170 and 1,584 bp, respectively (Fig. 26). Sequences were identical on the nucleotide level except for one nucleotide discrepancy in the 3' UTR (Fig. 27). Overall, the Rheb sequence length was increased 7.7-fold, from 1,914 nucleotides of the cDNA obtained by RT-PCR and RACE (MacLea et al., 2012) to 14,813 nucleotides. There were two instances when scaffolds contained overlapping identical sequence but were not combined in the Ray assembler: scaffold-628615 and scaffold-25268 have a 607 bp overlap; and scaffold-20402 and scaffold-14687 overlap by 1766 bp (Fig. 26). Manually comparing the genome and transcriptome demonstrated all four scaffolds align with Rheb and extend the non-coding sequence. The high sequence similarity between the Rheb scaffold and contigs served to validate both genome and transcriptome assemblies.

Discussion:

This preliminary assembly of a draft genome of *G. lateralis* is the first step towards generating a genomic database aimed at understanding gene structure. The potential for global alignment between the *de novo* assembled transcriptome and genome is evident in the local annotation of the Rheb gene. When compared to the Rheb cDNA, the genome ORF sequence was identical to the transcript. By combining multiple scaffolds together, we significantly extended the Rheb gene sequence and identified four introns containing the canonical GU-AG splice site sequence (Fig. 26, 27) (Pagani and Baralle, 2004). Validation of the Rheb transcriptome sequence and the increase in non-coding information represent two main goals of the genome project.

Using the scaffolds generated in the genome assembly, we can approximate the genome size of *G. lateralis* and compare it to other crustaceans. Based on the 17 k-mer frequency, we calculated the genome to be 3.07 Gb. This value is similar to *N. denticulata* approximations, which used the C-value of a related shrimp to estimate genome size to be 3 Gb (Kenny et al., 2014). When the crustacean genome sizes were superimposed onto a figure generated by Yandell *et al.* (2012), the genome sizes of crustaceans clusters in the same category as humans and other vertebrate mammals (Fig. 28). *E. sinensis* reports a more modest genome size of 1.66 Gb, but it is still larger than other invertebrates, such as *C. elegans* and *D. melanogaster* (Fig. 28). The only crustacean with a reported genome size below 1 Gb is the *D. pulex* (Colbourne et al., 2011), with an estimated genome size of 200 Mb. The small genome size of *D. pulex* is closer to that of *D. melanogaster* (Fig. 28) than other crustaceans, possibly due to decreased intron size (Colbourne et al., 2011). Based on the genome sizes of other crustaceans, *G. lateralis* appears to be closer in size to mammals than *Drosophila.*

While the draft genome was used to complement transcriptome sequence, as well as estimate genome size, there are limitations in regards to global usability. To take full advantage of a genome, *ab initio* analysis is required, including computational gene prediction. This form of prediction takes advantage of the entire genome by identifying potential genes throughout all sequences, rather than manual annotation. Computational gene prediction relies upon contiguous scaffolds, with the recommendation of an N50 scaffold length equivalent or larger than median gene size (Yandell and Ence, 2012). A genome meeting this standard is estimated to have 50% of the genes contained on a single scaffold (Yandell and Ence, 2012), increasing the success of *ab initio* gene identification. Based on the gene size (Fig. 28), we estimate that the N50 scaffold length for *G. lateralis* must be at least five times longer than the 1,819-bp N50 length obtained in the draft genome. The inadequate N50 value, in conjunction with previous evidence of non-contiguous scaffolds in Rheb, supports the need for more sequencing to improve the assembly. The *E. sinensis* genome sequenced ~83% more reads than *G. lateralis* (Song et al., 2016). Inadequate read numbers are attributed to producing shorter scaffolds, with a recommendation for additional sequencing (Yandell and Ence, 2012). Generating longer reads with additional sequencers, such as PacBio, would complement the current set of reads.

Future work will also involve adjusting parameters of the current assembler, Ray. Ray was selected because of its parallel processing ability. Assemblies can be produced within a reasonable time frame, allowing for multiple iterations (Boisvert et al., 2010). Ray generated useable, verifiable scaffolds in the draft genome, therefore optimization will likely improve scaffold length. An alternative to improve *G. lateralis* assembly is to utilize a program shown to be successful in crustaceans. The draft genome of *E. sinensis* was assembled using Platanus, with an N50 of 224 Kb (Kajitani et al., 2014; Song et al., 2016). In addition to optimization, we are interested in using Platanus to increase the N50 of *G. lateralis* scaffolds.

A major accomplishment from this project is generating a useable *de novo* draft genome from a limited number of reads. For *G. lateralis*, the first assembled draft is complete. With only two technical replicates, we generated scaffolds that can be utilized to extend and validate transcriptome sequences. Reads were also used to calculate genome size. Future work will involve additional sequencing and assembly optimization to generate a computationally workable draft genome, with global utility to enhance transcriptome data.

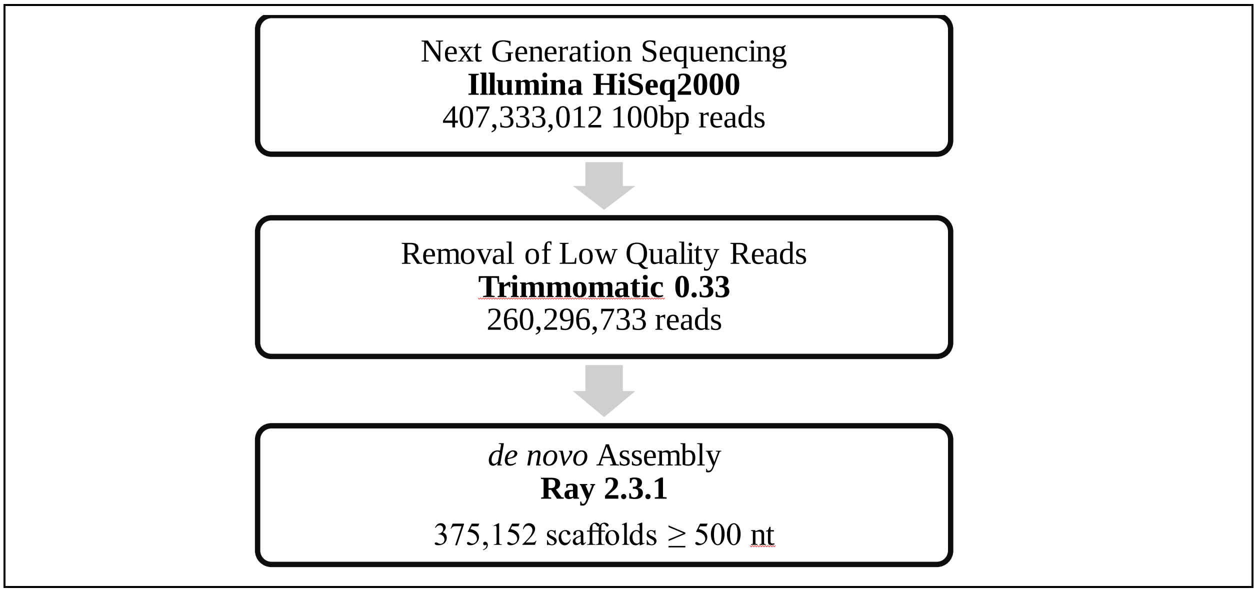
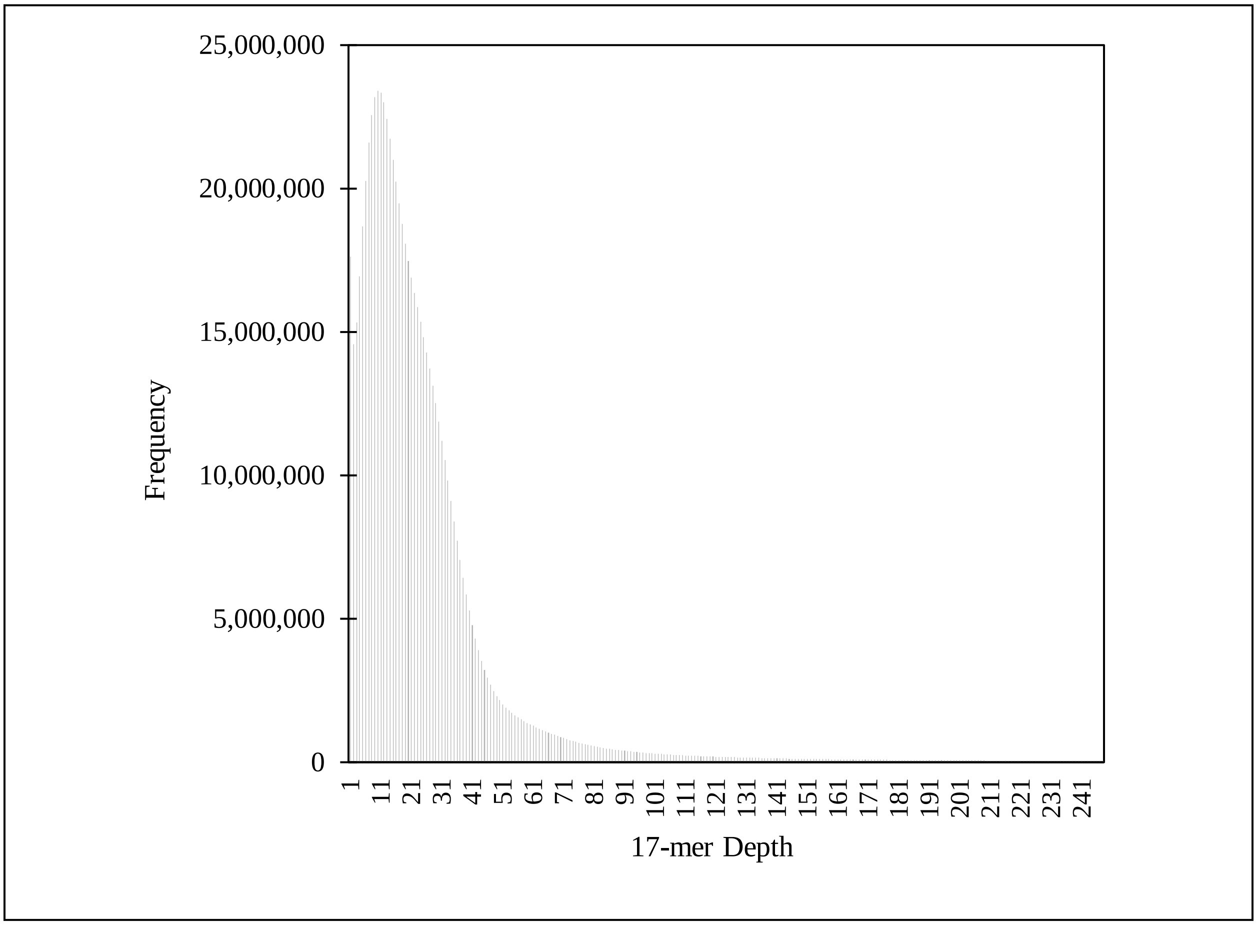


Figure 24: Draft genome assembly pipeline. Only Pooled metrics reported, L01 and L08 used same pipeline. Sequencing platform and packages indicated in bold.



<u>Figure 25</u>: Histogram required to approximate genome size. Histogram of 17-mer depth. Depth graph data began at 3. The most common 17-mer depth, 12, was found 23,404,080 times in the draft genome.

A)

Parameter:	Abbreviation:	Value:
Number of Reads (trimmed)	R	520,593,466
Average Read Length	L	86.6599
K-mer Length	K	17
Highest k-mer Frequency	M	12

B)

G= total sequence length/read sequence depth

total sequence length = R*L
total sequence length = 45,114,577,700

read sequence depth = N
M= N*(L-K+1)/L
12=N*(86.6599-17+1)/86.6599
N= 14.72

G= 45,114,577,700/14.7172
G= 3,065,432,127
G= 3.07 Gb

<u>Figure 26</u>: A) Values required to approximate genome size. Highest k-mer frequency calculated in Figure 2. B) Genome calculation, including equation and output.

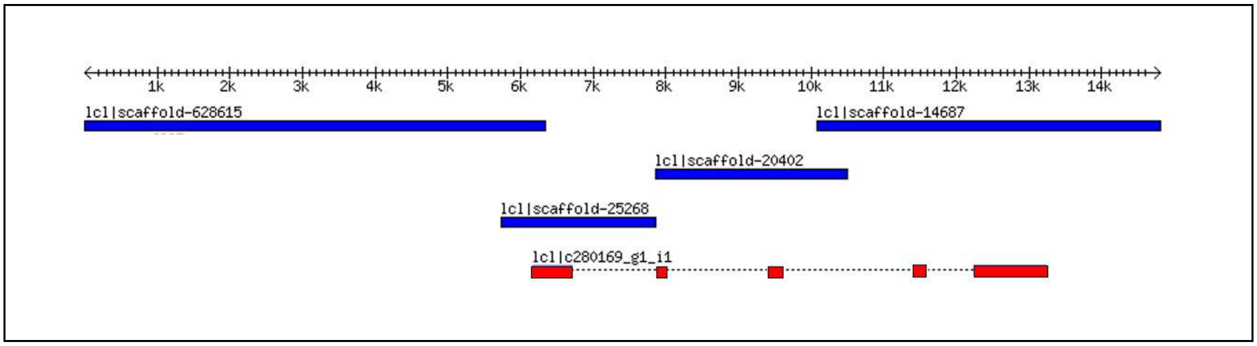


Figure 27: Rheb, Ras homolog enriched in brain, sequence alignment. Alignment includes reference transcriptome contig for Rheb and four scaffolds from draft genome. Introns are denoted by the dotted line in the reference transcriptome. Scaffold-628615 and scaffold-25268 have a 607 bp overlap; scaffold-20402 and scaffold-14687 overlap by 1766 bp.

AGTATGAGGCCAATAAACTATACATGCATTCAAATTTCTTTTTACTTCATCCAAGATCACTACAAAACACTCGCACCATCCCTAATCTCCACCACTACTTATCTGTA
CCCCAGGCAAAATCAAAATGGATGTGCCAAACTGGTGTATACTAACCATAGCCAAGAAAGAGTGACATTTCTAGTACCTGGTGCAATGCAACAGAGCCCTAGAA
AATAGTCTGAAGGCTGCAATACCTGAGATGCAATGAGCAACTTTCCAAAACAGGCTTACAGCCTGCTTGAGCCACAAAACTGCAGTGAAATACAACTTAGCAAA
AATGAGGCAAAAGAATGTCAAATGACATTGAATTACTAATTCAAAATGACACAGAGCAAAAGAATCCTTTATTATACTGTGCTTCACTGAAAAGTAATATAGCAG
GGTGATATAATGGGAGGAGAGGTGTTGGAAGCACTTCCATCAATAATACTTTGAAGTGAAGTACATTTAAGCCATTTAATGTCATATGACATTCTAACTCACACC
TTGCAGAGGTGTCAGTATTTCACTGTAGTTATGACTCAACTAGGCTACTGGGCTCAAAGCCAGCTGACACAGCTGTGGAGGGTCACCCACTGATAACAATAAAGT
TGACTAAGCAGAGATTGGTGCATTCAAACCTGTATTTATTCATCTCAGCAATCCTAATATGCTATAAATGGCCCCAGTCATGAAAATAAGAAGTAATATGATAC
AGCTGTATCAGTGGTTACTTTCTAATACATGAAAGCACAGTAGAAGCAAAAGATGTCAGTTTAACAAAACAGAAACTTTAAATTTTCTTGAGATGGCTCTCTTAA
AGTAGAAAATCAGCAGGAGCAGAAATGTCTGAGGCCATGAACCATGTGACATCATATTCAGAGATCACCAGTGTACTGTTTACTGTTTGATTATTAACACAAACT
GCAATTTCAAAATTTAGCCTACTGCCCAGTGGATGTGAGGAAATTTTTTCAGCAATGGCATATGTGTTTTCAAGGTGTGAACCACAAAGGAGTGAATATTTGAAA
AGCTAACATTCCGTTCTGCACCCTAATCCACAAAAGACGCTGAAATTTGATATGTAACAAAGCATAACATCTAGCTGTTATACTTCTTATAGAGATGATATGTAAA
CTTGTGGTTCCATTGTAGTTAGGCACAATAAAAATCAATTATTTAATTCAATTCAGCTGAAGTGGAATGGCATGGTACACGGACTCTAATATTTGAGTGTTTTTTT
GGTTGTTTGCACTGTAACAGCAAGATAGGCTACAAAACTTGTACTCATCACACTATGTAGGTGCAATATAAGATAATGGCCATGACTCTGGCCTTCTATAGCATAC
TAGGGTAACTGAGATGAGTGCAATGCAGATCTGAATGAACCAACCTCTGTTCAGTCAACTTTGGTTTTATCAATGCACCATAGCTATCTAACAATTGTCCCGTGTT
CTATTGAATTGTACCAGGTAACAAAAATGTCACTCTTTGTTAGGTGAATGATCAAACAAAAGATTTTCTAACAAAATAAAAAAATTATATAAATAATGTACTGGC
ATGGTCAGCATCAATCTCTTAAGTTGCCAATGTGCAGCCTGTCCCAGCTCGGCTAGTGACTGGTCCTAACTTCTCCCAGGATACATACCCTTCCACACTTGGTGAT
TATGCAGCTTGGTTTCACTAATCATACTCAACCCTAAGCTGACCCAACCCCTCTACTCTACAATAATATGAATATCTCAATCGTATCTATCACCTGACACCACCCTC
TACTCAGCCCTACCAGACCAAAGTGTCTCCAGCAGCAGCTTCATTACTGTGACCTGTTTCTACTTCATGATAAGCTGTGATATACAAGTACTGCCATATATCAATC
TAATAACTAAATGCTCAATAATGTAAACAATGCAAGCACAATCATTCTTGCTATTCCTTCTACAATCAGAGATACTGAAAGAGCTGGCAACTCCCATCCACAGCA
TCTTTTCAGTTTTCATTTCATTTACCTATATTATTTAGAAAGTGCTACGTACTGGCCTTGCTACAAGTCATGTATATATGACTGACTGGTCTGTTAACTTGATGCAAC
AGAAGTGGTTATTACTTCCCAGCAAATGCTAGCACAGCACAGGCTGATATGGGTTAGTGATGAGCTATGCTGAATAAATTTTTGTTAATTAAAACTGAAGCTGAC
TCTAAAAGGTGGTACTTATAGTGAACAGTAACAGAAGAAAAAATCAGCAACAGTTTTATGGAGGGTAAATATAAATGAAGGTAAACTGCCAGCTTCCTTCTGAA
AGTCTGTTTTGTTTGGGTGAAAAAACAACTTCCAGTAATGCATAGTGCCCCCGTGTAACTGTAAGCATGCCTAGCTATGCATCTGCAGGCCTGGGTTTGATTCCGG
CCAGGGCAGCCAGTGCACAGCCCACCCAACTGTTCATCCACCCTTTTTGAGTTGATCAATAAATGAGTACCTGGGGAAATCTGGAGAAGGTAAAATGTGGTAACC
TAGATGTCATACCTGCCATGTTTCCCGTAGTAATGGGTTCTTACCCACTACTGGCTCCAGGGACAAAGAGACAGATGAGCGCCAATGCAACATGCAGCTTAGCAT
GTGCTCCCCACTTAACCCTTACTGTAACATATTTTCATAAGCTATGTTTTACCATGCAGCATTTTAGTAATTTTAATGATATATAACAGCTTCTCAGGGGAGTAAAA
ATAACTGCCAATAATAAGACAATTACTGGAGACTCCTAAGCAAGAAAAGCTGGGTGATGGGATGGCACATGGCAACAGACAACATTCAAAAAACTTTCTGTCAA
CTTTCTCAGTAAGATTCTAGAACAGGTGTGGATTGGAACGGTTAGAAGAAAATATCAAGAAAATTCAGATCATATCTAGGTAGCAATAACAAATATACTCTCAAA
AATTCTTTTCCATATATCACTAAGTTTAATCTGATCATTATAAAGTCAATAATATATGGACAATGTACATGAGCATTTTTAGCCCAAGATTTTCTCTTGAGCCTATT
TTATCTTGCATAATATAAGCTTTTCAACAAAGACCCATCCACAACTATCCTGGTATCCACTGCTGAAGGCTGATAAAGTATGGTGTGTTTTAAACATGGACAGCAG
TGAAGATGAAACAATCACATTCAACTATGAATCAGCAAAAGCCAATCCGTTGCTCCATTATAAAAATTACATTTGTACATGCACCCACAATACCTACTGTACAAT
GGGATGACTTGTACCTTACCTGCATACAGCCACACAGCTTGACAATCTGCTTGCTGTCCAGTACAAAAGTTCTTCAATACCGAGGTTTTACAATGATCTCTTCAAG
CTGTACAACAACCCTTGCTACCGCCAATGTAATGGGTTGATGTTGGGATGCCTCGTTTTACGCTCTATTCAGGGTGGCCAGTAAGACAAAGATGGGTGACTGTAA
ATAAGAGGTCCTTCCAGAATCTAAAATAAAGACAATTCAATCATTACTGTTCCTGTTTTATGGTGAGTGAAAAGTCATTTGGCGGCAAGGGAACAACCAAGAGAT
TTCAGGTTTATCTGTGAGGAATGCCGTGTAGTACGCCTATAAATCCTGCATGATCCTTCTTACATAGACACATTATTAATTAGTCCAATTTCTTCTCATCTCTTTACT
GTCAAATCCCGAGATGTAGTCTCTCTCCACTTTCTTGAAATAGAAAGTGAGAGAAATGAGTACTGAAATGGTAAGTCACCAACTAATGTTGAATAACAAGGAGAG
AATAAAGAAATAAATAAGCAGGCAAATAACATGTTAGTGAAGAACAAAGAGTTTGAGGAATCTGTTGATGCAGTTCACACATTAAAGGATAAGGTGAATACAAC
TAAAGAAAAGAATATAATAATGGGCAGACAAAACACTGACACAAGAGGGTTTGAAAATGTAAACATCAGACAAGAAGTCAAGGAATTAATAAGATAAAATACA
AAATTACTTACAGGAGACAGTTGACATGAACAAAGCAATAATAAAAAAGGCAAGTAAGAGGAAAAAAGAAATAAAAAAATCTGAGACAAAGCTGAAATAAGT
ACTACCATGAAATTCATCTATAAATATAGTAACAGAAGGAGCATCATAAAAAGATACTCAGGAATTTTACAGAGTAAGAAAGTATGAGTGCAGAAAGAACAGGC
CAATAAAAGTAACAAAGATCAAATATAATGAAAGATGTAATGAGAAATTTTAAAAAAACATAATAATGTCAGTAAAATGATCATTGCAAACATGCCTGTGTTGC
TCTCCTACCTGAGCTGATTAAGTACAGGATATGCTGGCTCTTTTCTCTTAATAAAATTTTGAGTGAGGCTATTTACCGAGCAGCAGCTCGTCTAATAAGTCAATCA
TGTGTGATTACTTACTAAGCAGCTCAGCAAATAAGCCAGTCTTGTCAAGCTACTTATGCCATTAATGCAACCTAATGGATAGTAACCCATCATCCTGTAATTAGTT
AAGAACTCTTTATAAACTAACAGAACTCCATCATAACCCAATGGATAGTAATCTCTCCCTCCTAAGAATAGTAACTTATCACCGCTAAAATAGAAACTTATCAGT
GCTTAGTTAGTAACCTACCGCCTTCCAGTCAGCTCAATAAATAGTAATGGAAGTATATAATATTTACCAAGCTGCTGCTCAGTAAAAAGCCACTCTGCAAAATGTT
CTCTTTAAGGGTTGTATCTTGTCACCAGAGTTGGCAAACAATACAAAAATTATTGCTATCCCCCTGATAATGGTTGTAACACATCATTACATCACGAGGGCACAGC
ATTGTGATGATTTCTGCCAGGACTTACCGATTATATTCATGTCAGGGACACCTTCAGGTAAATTTTTACATTGTTCCCTGTCTCACAATTCATGAAGAGAATATTTT
CTTCTAGTAGATTTAGGTGTCTTTTGAACGAATTTGAAAACTGTTCCTGTCACAAATCAGTAGTTTTCCTTACCCCCCTCCCCTCATCAAAAAACTAAAGATTTGCG
ACAAAATCAGTTAGCACCTAACCAAACCTAACCCGTGAAATGTCAACACGTATGGGTGTTACGCTCTCCCTCCCAGCCCCCCGTGACGCCATGAGGCGGCTGCAA
GAAACTCACTAACTTTCCAAAAATCGCTAGCGGCGTTTCTAGCAGTTTGCGACGAAAATATATAGGGTTCAAAATGCATAGAATAATGAGATCTAACCGATGAGA
AGTGTGTGAAGTACCGCTGTGAGACGGGGAACAGTGTAAAATAATACCCACCTTCACTCCTCCAGATGCAATTCTGACGGTGTCTAGTAAGGCATGGCGGCACAT
GAAAATAATCTATATATCTGTAACCTTTCAACGAGACAGCTCCGCCCCCATCGACCCCCCGAACTCCAAAAACGCTCTACAGTACAGAAGAAAAGTGATCAGGA
AAATGAAACACTGGCTGATGTGTCCAAAAATGGCGTACGTGTTAAGAATCTGTACCTAAATATCAACCAACACTTGCAAACACAACCTTCCCTACCCACCTGGAC
TTGGTCGTGAGCTCCACCATTTGGTTCGCCCGCCTGCTGCCACCCACGCACAAGGCCCCACACACAGCAGTATTCCTCTTACACTTACTTGAATTACCTTCACACC
CCAACACACAATTAAAGTGACCCTAGAGCGATCCATTACAAGTAAAAGTAATGTGTATCGACCTTTCTCCCTCCCGTGGTGCTGTGGAAACACTGCCGCTGCGCC
ACCCGTCAGCCGCCAGACCCAGGTGACAACAACAGCATCAGCTGATCGTGGGTACATTTGATAAAAATAATATAACCCTCCCCACGTACACTTATACAACTTACA
ACTCAAATTCAGCTCACTAAATCATTTCAATGCATGACTTCACAATAAATCACCTGTATATCCCTTTACGGCATAAAACAAACACACAGCATCTCTGGCAAGTATT
GCGTAAATTATGCAACTGGGGGACATCCTATCACTAGGCAGCAAAAAACGTTAACAAGGTAGTCTTCTAGTGCGCCCTCAGTGTTAACTGGAGTATACACGTGCA
CTGTAAGGGGAAAACAGGTAAAATTAGATGTAGTAGTAGATGTAGATGGGGAGCTATAGTCTTTGCAACGGCAGCTCCTAAAGTGTTCGTATTACTGCCACAGTT
CTTGTCTTCCTAATTATGTCTTTATCCGCATTTTATAACAGATATTTTGTTGCACTGAGCTTTGTTTTCTTGATTCTGCTTCCTCCAATACTCCTCATCAGATGTTCAT
AATTGTCGATCCAGTTCTTACAACACTAACCCTAATACTTGAAAGAATTGAGTTGTGTACTTTCTGGTAAGTCAAAAGTGAAGAAAAGCTGAGGAAATAGGCGTT
ATCTATATTTTTTAACTAACGGTGAGGAGAGATCTCGCCTAGTGGTGTTGTTGCCTCCTCCACAGGGCAGCTGGCGCAGAGTGACGCCATCACCGCCACCAAAAC
AAACCCACACCTCCACCTGTGTTTCCGGGCACCTCAGGCGCACCCAGGGGCCCCAGGTGTGTTGCGAGCTACTCTTATGGACCACTGAGCATCTCTCCACGCAGC
CATGCCTCCCAAGACGAGAAAAGTGGCCGTCATGGGCTACAGAAGCGTGGGTAAGTGTTGCCGTCTCAAGGCTAAGTGAAGATTACCACCAATGCGTATCACTG
CTCGGTAGGGAGATGTTATTTTTCCCTTCTTTTGTCTTTTCTCGGTCCTATCCATTGCGACTACGAATAAATGGCTTGCAGGATAAAGTAATGCCAGCCTCGAGAGG
GAAATAAAGCACTAAATCTCTGAACCTGATGAGATATATTAACAAAACCAGATCCTCACTTGTCCCAGTCCTGTTGCTTACATCAGTGGGTCCCCCTCCCCACAGG
AAGAAAAAAAAGTCCTTTGTCCAGTGAATCTTCTGTGACAAAGGTTCCTTAGTGTTCCCACCCTTGCCAAGCCTGCCTGAGTGAAGGTGTCATGCTGCTTCACCCT
GCCAACACATGCCACAGGAATACACAGATAAATTCTTTGTCTGACTAACACGGGGCATACTGAGTTGAAGGTCACACCCTTGAATTTGTTTGAGCCATTACAAGT
AGAAGGAAGGTTATTTATCACTGCCAGTTCTATTATTGGTTTATGATAAGGGAAAATAAAAGTTAGACATATCAGTTGGTTATTTCTGGACATATGATAAGCAAG
GTGGCTGAAAATTAGTGCCTCCTCCAATGCCTTTTGGCAAGTTTTCTATTCCTGATGATCTACTCAGCATGTACTACACTATATTCAATACATGAAGAAGCTGAATT
GTGTATTTTCTAGTTATTACAGGGACAGGAAAGAAAGTCTGTGGAAAAGAGAACTAGAAATCTGTAGCATGATTTTTTTCTCAGATATACCCAACAGTTGAGATA
TAAGGTAACAAATCAAAAGTAACATATGGGAAAGGAAATGGCCATATTTGATATTATAAAAGTGTTTAGGTGATATTGTTAGTGATACAAATACTCTGAGATTTT
GAAACATTTTGTCATTTGAGTCTTGGAATATTCATAACCATTGTATCTAAGGGAGCATTGAAAAGCTCACAGGTTTAAGAATCGGGTGGTATGGCTCAGGGAATG
TCTTGATAAATCTGAAATTTCCCATTTGATTTAGTTTTCTGCTTTACAGCAAATGGAACAGATTGGTGCACGAGCCTTTCATGACTGCTTTTCTTAATGGATGCCAA
CCAGGAATGAGAAAACATAACAAAAGATTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC….CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC
TCTCTTGCACAGGGAAGTCATCTCTATGCATTCAGTTTGTTGATGGCCAGTTTGTGGACAGCTATGATCCCACCATTGAAAACACCTTCACAAAAAAACTCAAGGT
AACTGTTCTCCACTTTTACCATTTTTGCATCACTTTACAGGCCTTGATGGGAGTGAGGCAGATATTGCTTTAAGTATATTTCATGGTTTGAGGATTCTGTCAATATA
TAAATGAAAATAAGAATTATGTGTTTTCCATCTCGGGTAAAAACTGTTATTACAAGTATATAACAACTGTCAGAATATAGAAGGATGATGATGATGATGATAGCA
AGAATAATATTTAAGGAGAGCACCAACAATAATAATTATAATAATATTTAATGAGAACACCAACAGTAATGATGATAATAATAATAATAGTTTGTTGACTTACAG
GAACTACATCCACATGCAAAATGCTTGTCTGAAGCCTTGCAGCCTTAAGGAAAGACACCCATTTTCCAATGAGCACACTCCCCTGGCTACTCTCATTGCACACAG
CATTGCTTCTCATGTAAAAGTTCAATAAATACAAGTCCATGGACATTCACATATGCTGTGGAAGTGGGTGTGGCTCCATTGATAATATTGATGATGCTGCAATGAC
AGCACACTCACCTCTCTGACCAATTCCTTGCCATTTCATAACTTTTGGATTTTTTAACATGATGCAACTTCACACCACAGAGGCTTATCTCTAAAGCAAGCAGCTG
ACAATTTTTATATTAAATTGTTGCTAATGCATTAATAGGGTTTAGAACTTTAAAGCATAATTCATGGAGAAAATTTGTCACTTGTATGTTAACCTAAGAGCAATAG
GTGACTAATTTAGAGATGCAAATGAAATATAAATCATCAGCTGCTTGCTTAAAGATAAGTCAAGTGTGATGTTAAGATGCACCATGTTCAAGAAATCTTAGTTGG
GAAAGAGCAAGAATTGATCAGGGAGGTGAGAGTATTGTCAACAGAGTATTGTGAGTATTATTAGAGGAGCCACATGCACATTCACAGCATAATTGAATGTCTGT
GGACTTGCATGTATTAATTTTACCCAGGAAGCAATACTGTGTGCAGAGAGTAACCAGAGGAATGTGCTCATGGGAAAACAGACATCCATCTCTGAGGTTGTGAGG

CAGTCAGACATGCATTCTCTGTGTGGCTGGAATTCCCATAAACCAGCAAATGTTGTTGTTAATATTATATTACTGGTAGTAGTAGTTGTAGTAGTAGTACTGAAAT
AAATACTTAGAGTATGCTGGCCAGTTTATATGCACCAATCCAATTAATACTCAAGTCAAGAAGAAAACAGGTCTGGAAAATTGAGGTGAAATGTACAAATTTCAG
CAATATTTCCTTTTTTTCTCTCAGCTCTTTTTTTTCTGATACAAGTACTAATTGGAAGAGTGCATACAAAGTGGCACACATACCTTAAGCTTGAGACCAACCTAAGC
ATGATGTGGCTTTTTCAGGTGCGAGGGCAGGAATATGGCCTGGAGCTGGTGGACACGGCAGGTCAGGATGAGTATAGCATCTTCCCAGCCCAATACTCCATGAAC
ATCCACGGCTATGTCCTGGTCTACTCCATCACCTCGGAAAAGTCCTTCGAGGTAGCCCAGGTCATCTATGACAAGATTCTCGACATGATGGGCAAAGTCACGTAA
GTTCTTGCACCCGTGCCAGAGGGAGTCCATACAAACACAAAATTACCCAGTGCTGTGAAACACACACTGCATGTTTGATATATTGAGCATTGTAGTCTGTTAGCAT
TTTACATGAATTTAGGTCCGTGGAAAGGTTAAATCAATAAATTTTTATTAAAATAGAATTTAATATTCCCTTGAGAAATATTGCTAATAGGTGTAGTTGCTGTATG
TGAACAGGTAAAAACCTACAGTAACAAAAAATTAAAAGTCACTGAAATCCATTTTATGAATATAGTAGTACAAAAAGCAACTAGGCTTAGTACTTTTCTTGATTT
TGTCAGCATTATTATTTATGGTGCACAGTCTGCTCCCTTTTAATTGCCTAAAATGTTTTATGGCAACTGTTATGTTACCCCGATCCCAACACAGTAATACTAGAAGT
GTTAGTAACCGGAGCGCAGAAGTTATACTTAAACTGTATCTGGCACTGGTTAGACCCCACCTTGATTATGCTGTTCAGTTTTGGTGTCCATATTGTAGAAAAGATA
TAGATTCCCCTGAATCTGTACAAAGAAGAATGACTAAATGATCCAAGGACTTAGAAACCTACCGTACAGAGAAAGACTACAGAAGCTTAATTTACACTTTTTAGA
AAGGTGTAAACTGCGAGGAGACTTAATTGAAGTTTTAAATGAGTTAAGGGGATCATTAAGGGTGATTTAGATAAAATACTGATACTGAAAACAGATGTTAGAAC
AAGTAGCAATGGGTATAGATTGGATAAATTCAGATTTAGGAAGGAAATAGGTAAAAACCGGTTCACGAATAGGGTTGTGGATGAGTGGAACAAATTAAGCAGAT
ATGTAGTCAATGCTACCACGATTGAAAGTTTTAAATGGAAGTTAGACAGATTTATGGATGGGGAAGGGGGTTGGTGAAAACTCCCTCTCAGGAGCTACCAAGTGT
AGGCTGTCTGGCCTCTTGTAGTCTCCTTATGTCCTTGTGTATCCTTGTGTATATTTTAAAGTTGTCTTGAGGTTGATTAGCTTCTCTCATTTGGAAGCATGTGAGTTA
GTGCTATTCCTTACTAAACACCAGTTTGCTGAGGTGCATGCACTATATACTCAGAGAATATTTCACTAGGTACATTAAACATGATATTTGAAGTAGGCAGAATGTG
AGACTGACTGCATACACCATGGCAAGCTAGTGTTTAGCAAGAAACAGTGCTATTCCATGTGCTCCTAAATATGAAAAGCTGATAATCTTCATGACAACTTGAGGT
GTTACTATGTTAGAATCAGGTTGACAAAGCAGCCTCCGTAGACATTGAAGGCAGTTAAAAGGGAGCATGTTGTGCACTGTGTGATAATGCTGACATTACCAAGAA
AAGTGCAAGTTGTTTTAGGGAAATTTTACTTGTAAAATGCGTTTTTTTAGCTTTTCATGCTGCTGCTGCTGGAGGCACAGCATATGCACAACAGTGCCACATGTCT
CATGGTGGAACAACAGTGTTGCATCCAATAGTGTTGCAGTATTTTTGTAAATGGAAGTAAAGTTACATTATGGATGATATGTGCAATTTCAAATTTGTATAAATTA
AATTTCTGAAAATGCAGGGACTGCATTTTAATTACACAAAGGCAGAAAAGGTCCTGAATATGTAATTTTCTTTCTGCACTTCAGTCCTAGTAAGTTAGTGTTTCTT
CATGGTATTGACAAAGAGTTATCACGACAAGTTGTATTGTTCTAACTGTTTAGGACTGATTCACACTAACCTCTGTGGTAGAACACAGTCAGCTTTCACCTGCAGA
CTGGTAAAGATGCTTATCATGATTTTCGTATGCATGCGTGTAGAGAGTGTCCTTGCAGAGACATGACTTCATTTGGTGGTTCCAGAGTTCCTGTGGTGTTGGTGGG
CAACAAGAATGACTTGCACCTGGAGCGTGTGGTGAGCACCGACCAGGGGCGCCGCGTGGCAGACAACTGGAAGGCTGTGTTTCTTGAGACAAGTGCCAAGGAGC
ATGAGGTAAGGCTGTGCTCCTCTGTGTTGGATCCTCAAGGAACTTTGCTCACTCCTTTAACTCAGCTCAGCTGACTTTCTGTGGAACTTGTCATCGCTGATTTTTAT
ATTTAGCATGCTAGGATGAGACCTAATATATTTTATGTCCTCAGAGGATACTGTCCATTCTTTCAACTGTTAATGTGGTGGTTGGTGCATTCATGGCTCATTTTTAG
CCAACTTGTACATCTTTTGCTGTAAAGATTCTTTTATCTAAATCTGAAATGACTATGAGGTGGTTATGTTGATTATCAGAATACTCAGTCTTGCTTTTGTAGTCTTGT
GACGTTCAACAGCTTGCACCTTTTCAGATAAAGACTGTCATTGGTAAATAAAACTATGAAAGCTTTGTATACCACTGATAGTCTTTGAGAAGGTGAAGGGTGCCA
AATGTCCCAAGAAAACTATCAAAGCAAGATATTACCCACCTGCAATTTGAAACATAGGGTTGAGGCATTCTGGTAATTAATGTTGAATGGAGAAAACCAACCAA
ACTGCTGCCAGGTCAACACTAGACACTTCTGGTGATATCATACTAGAATTTACAGTTTCTTGTGCTTGATGCTATGCTTCTGTTTGTGAACCTAATGCTTGTCTTAC
AGGCAGTGAGTGACATCTTCACTCGAGCCATCCTGGAGATTGAGAAGGCTGATGGGAACCTGCCCTCCGGTAACGGCTGTAGTATTTCATGAAGCCTCTGTGATA
TAGCCAGAACCTTTATTGCCACCACCTCTCTGACAACCGATTTGGATCTTGAAAACAAGACTTTGTACATGGCTTATTCTCTTCACGGGCAACAGGATCCAGAAAT
TTGTGTTTTCTTCTGTTGTATCAGTTCTTTATGGCCTTGCCTGTTGTGAGTATGAGCCAGCCCACTGGACCCATGCAGTACCCTCCTAGTCTGTTGTGGGATAATGG
TCAGTACTGTTGGCAGTGGGTGTGGCTAACAATTCAAGGCTCAGGTATATACATAGACAAAAACATTAGGTGGATTCGTATTAGGGAACACAGCTCATCCACTGC
CTGAAATGGTGGATTTTGTACACTAAAGCCTGTCTGTGAGGGTCTGGCTTTGAGGGACAGAAAGACGTGTGTGTGGTTAGGGAATCACGATAATCAAATACCTAC
TTCATGGTCATGGCCTTGCAACACATCAGTGGAGGGACCAGTTACCCACCACTCCTTGTGTTTAGGAAAAGTTGTAGTAGCTCTTATTTACTTTTTTCTTCTCTAAA
TTGTTGGTCACAATTTCAGAGGAAAACTTTTTTATGAAATCCATCAGTGTTGGGTACCATAATGATGACCATATTGGAATTCCACTTAAGTTTGTAGAAAATATTT
ACATGTGGGGGTGACATATTTGTTAACAGATAATTTACATTTTAAGTATGTGATGAAGATTACACAGTTGTGGAAGCTGAGCTAACAGTGAAATAGCTGGCTTA
TAGAATATAAGGCAATAAATTCGTAGAAAAATACCATGTGCATCTCACTCTTCTCCACAGTGTAGTCAGGATTCTCAATACCACTATTTGGTGTTCTTAAAAACTG
CATCATGACAAGGATTGGTGTACAGTAAGTCAGCAATAAAAGGTATTTTAACTCCCAGGGTTTTTGGGCCTAGACATGATGACTGTTCCCTATTTGTTGTGTTACT
TGCTAGCAAAATACACACCATGACATGCACTATCAGTACACACAAAGAATGAATGTTAGTATTGTTCTCTAAAATATAATTACAAACATTACTGTAGCTTCAAGT
ACAAAAATATATATCTTAATGTTGAAAACATAATAAACTTTCAGTGAACTGCTATGTGCAATTAGTAGATGCAGTGGCATAGCTAGGGGTCTCCTTATGAAGGGA
CCCCATGAGGGGGGCCATGTTTTCTTTGAGGGGGAGTCCCACTGGACATTTGTAATGAGTGAATCAATGTCTGTTTCTCAGAATTTCCTAGAGAAAGAAAAAAAG
CTTATAGGTGTAGAGGTAGCATGGGTTTTCATATGTGGGAGTATCCGGCGCTGGGGATGACATCGCTAATTGGAAATTAAATGGTGATGTGAAATGGGCAGTAAG
GGGATACGGGAAGGCTAGAACTAACAGAAGCAACAGAAGTCCTTTTAATGAAGATGGTGTTAGGTGTGAACGTATGACGAGGGTTCTTTTATGTAGCATTTTCTA
ATGGTGGTGATGCCCAGGTTAAATGGAAGTGAAATGAGGATGTGACATTAGCAAGAGGAAGTGTCAAGGGATGAACACAGGTTATAATGAGACATAGGTGTGG
ACGAATAACAAACGTTGTCTTATTTGGCAATATCTAAACGCTAAACGGAAGTGAATTGAGGAGGTGACATGATCAAGAAGGGGCATCAGAGCAGATGAAATAAC
TAAACTAAAAACATACGCGGAGTGTACAGTTATGAAAGATGGTGACACATAAGAATGAGGAAACAGAACACAAGCAGAACCTAATACACATTGAAGGGACTTC
CCGCAAGTACTCGTGTGAAAGTCGAATTTTTGAACGATTTTTTCGGCCAAGTTTAAAAGAGTCAACTTTTTACAAGGATACTGCAATTCTAAGAGCTTTTAATTAT
ATTCCTCAGTCGAACATATCATAGCCTTAGGGAATAGGTGGGGTTGTAATTAGTAACATGATATCGACTGTTACACGAGTAGGCTATACACGGTACGATAAACTG
TAATAAATATAAGAATAATGAAATAAAGCACCCACGGAACCTATACGTACAATGAGCCTCATAAAAAATATATCATAGTTAACAATAAAGAACAAAGCGGCAG
CAGAACCCATACATTGACCCACAACAGGAAAAAAAATGTCTTAGGATGAGATCAAATCATGCCCACGTGAGATCCGATCCTTAGTGAGATAAACTGGGAGGGAC
AGGTGACCAGACACGTGTAAAAAGAGAAAGCGAGGAAGGAGGGAAGGTTACAAGGGACATGGAGGGCCTTGAAGAGGGGAGGAGGCGGAAGATGAGAAAGA
CAAGAGAAGCGGAGACAAGTAGGAGGAAAGCTAGGCAAAGAAGAGAGAGGGAGAGAGGAGTTAAAGGGCGAGGGTCGCAATGAAACAAACAGAAGGAGGAG
AATACGAAGCGAGAGATAGATAGATAGATAGATAGATAGATAGATAGATAGAGAGAG

Figure 28: Rheb, Ras homolog enriched in brain, sequence. Yellow boxes represents ORF, gray boxes indicate intron sequences. Red letter denotes nucleotide discrepancy, thymine in transcriptome and adenine in genome.

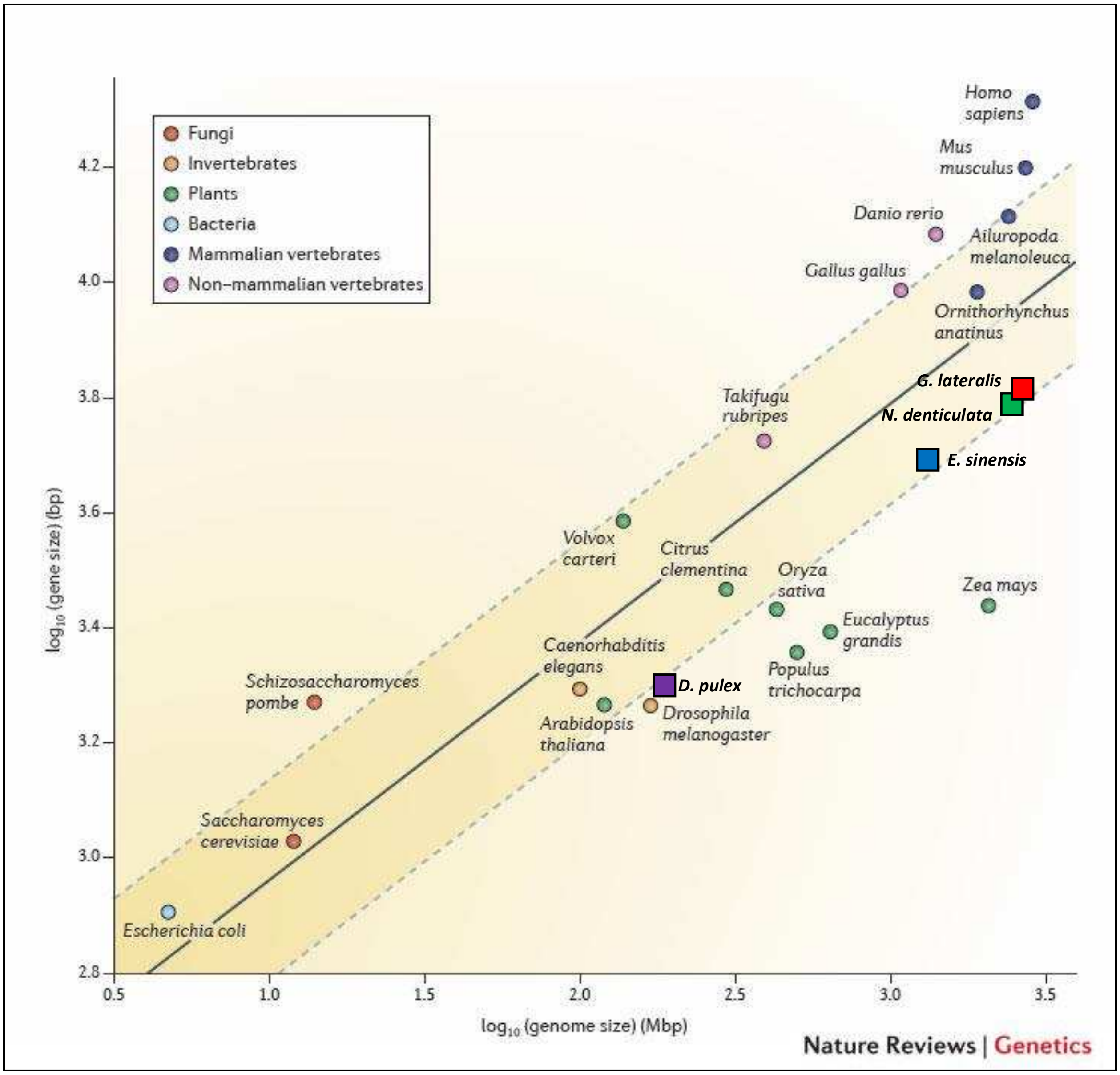


Figure 29: Comparison of genes and genome size with a modified figure generated by Yandell et al. 2012. A direct linear relationship between genome size and median gene size was denoted with the solid line. Crustacean species were superimposed onto the graph, estimation based on genome size with conservative gene size.

Table 2: Metrics on genomic read data from replicates L01 and L08, and the Pooled library. Percent GC, sequence quality, and adapter presence reported after trimming. The difference between the number of reads sequenced in L01 and L08 was 13%.

Metric:	**L08: No Index**	**L01: TGA**	**Pooled:**
Raw Reads from Illumina	188,858,188	218,474,824	407,333,012
Trimmed Paired-End Reads	116,621,615	143,675,118	260,296,733
% GC	42.50%	42.50%	42.00%
Sequence Quality	Reads contain a phred score >20 throughout all base pairs		
Adapter Presence	Nextera Adapters absent	Illumina Adapter absent	Illumina and Nextera Adapters absent

Table 3: Metrics on draft genome assemblies from replicates L01 and L08, and the Pooled library. The information was used to select the Pooled assembly as the draft genome for *G. lateralis*, to be used for downstream analysis.

Metric:	L08:	L01:	Pooled:
Scaffolds Assembled	132,406	231,961	375,152
Scaffold N50 (bp)	873	1867	1841
Scaffold Median Length (bp)	704	852	1011
Longest Scaffold (bp)	169,115	43,460	289,650
Read Alignment Rate	49.48%	62.46%	75.87%

Table 4: Comparison of crustacean genome assemblies to *G. lateralis* draft genome.

Year	Organism	Sequencer	Raw Read Length	N50*	Genome Size	Coverage	Estimated Number of Genes:
	G. lateralis	Illumina Hi Seq 2000	100 bp	1,011	3.07 Gb	14.7	TBD
2011	*D. pulex*	Paired-end Shotgun Sanger Sequencing on ABI and MegaBACE 4000	Not reported	83	200 Mb	8.7	30,907
2014	*N. denticulata*	Illumina Hi Seq 2000	100 bp	400	3Gb	12	3750 (annotated via *D. pulex*)
2016	*E. sinensis*	Illumina HiSeq 2000, short and long insert	500 - 800 bp	224,000	1.66 Gb	Not reported	7,549 - 14,436 (*ab inito*, two methods)

*N50 value is reflective of author reporting, but parameters for scaffolds included varies

REFERENCES

Abuhagr, A.M., Blindert, J.L., Nimitkul, S., Zander, I.A., LaBere, S.M., Chang, S.A., MacLea, K.S., Chang, E.S., Mykles, D.L., 2014a. Molt regulation in green and red color morphs of the crab *Carcinus maenas*: gene expression of molt-inhibiting hormone signaling components (vol 217, pg 796, 2014). Journal of Experimental Biology 217, 1830-1830.

Abuhagr AM, K.S.M., Megan R. Mudron, Sharon A. Chang, Ernest S. Chang, Donald L. Mykles, in press 2016. Roles of mechanistic target of rapamycin and transforming growth factor-β signalingin the molting gland (Y-organ) of the blackback land crab, *Gecarcinus lateralis*. Comparative Biochemistry and Physiology, Part A.

Abuhagr, A.M., MacLea, K.S., Chang, E.S., Mykles, D.L., 2014b. Mechanistic target of rapamycin (mTOR) signaling genes in decapod crustaceans: Cloning and tissue expression of mTOR, Akt, Rheb, and p70 S6 kinase in the green crab, *Carcinus maenas*, and blackback land crab, *Gecarcinus lateralis*. Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology 168, 25-39.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215, 403-410.

Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Gupta, N., Gharani, N., Toji, L.H., Gerry, N.P., Resch, A.M., Barker, J., Clarke, L., Gil, L., Hunt, S.E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W.M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R.E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Fulton, L., Fulton, R., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C.J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C.L., Kong, Y., Marcketta, A., Yu, F., Antunes, L., Bainbridge, M., Sabo, A., Huang, Z., Coin, L.J.M., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Garrison, E.P., Kural, D., Lee, W.-P., Leong, W.F., Stromberg, M., Ward, A.N., Wu, J., Zhang, M., Daly, M.J., DePristo, M.A., Handsaker, R.E., Banks, E., Bhatia, G., del Angel, G., Genovese, G., Li, H., Kashin, S., McCarroll, S.A., Nemesh, J.C., Poplin, R.E., Yoon, S.C., Lihm, J., Makarov, V., Gottipati, S., Keinan, A., Rodriguez-Flores, J.L., Rausch, T., Fritz, M.H., Stuetz, A.M., Beal, K., Datta, A., Herrero, J., Ritchie, G.R.S., Zerbino, D., Sabeti, P.C., Shlyakhter, I., Schaffner, S.F., Vitti, J., Cooper, D.N., Ball, E.V., Stenson, P.D., Barnes, B., Bauer, M., Cheetham, R.K., Cox, A., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E.E., Batzer, M.A., Konkel, M.K., Walker, J.A., MacArthur, D.G., Lek, M., Herwig, R., Ding, L., Koboldt, D.C., Larson, D., Ye, K., Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Willems, T.F., Simpson, J.T., Shriver, M.D., Rosenfeld, J.A., Bustamante, C.D., Montgomery, S.B., De La Vega, F.M., Byrnes, J.K., Carroll, A.W., DeGorter, M.K., Lacroute, P., Maples, B.K., Martin, A.R., Moreno-Estrada, A., Shringarpure, S.S., Zakharia, F., Halperin, E., Baran, Y., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F.C.L., Craig, D.W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A.A., Sinari, S.A., Squire, K., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, K., Burchard, E.G., Hernandez, R.D., Gignoux, C.R., Haussler, D., Katzman, S.J., Kent, W.J., Howie, B., Ruiz-Linares, A., Dermitzakis, E.T., Devine, S.E., Goncalo, R.A., Kang, H.M., Kidd, J.M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Wing, M.K., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Y., Shi, X., Quitadamo, A., Lunter, G., Marchini, J.L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D.K., Oleksyk, T.K., Fu, Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Browning, B.L., Browning, S.R., Hormozdiari, F., Sudmant, P.H., Khurana, E., Tyler-Smith, C., Albers, C.A., Ayub, Q., Chen, Y., Colonna, V., Jostins, L., Walter, K., Xue, Y., Gerstein, M.B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y., Harmanci, A.O., Jin, M., Lee, D., Liu, J., Mu, X.J., Zhang, J., Zhang, Y., Del Angel, G., Hartl, C., Shakir, K., Degenhardt, J., Meiers, S., Raeder, B., Casale, F.P., Stegle, O., Lameijer, E.-W., Hall, I., Bafna, V., Michaelson, J., Gardner, E.J., Mills, R.E., Dayama, G., Chen, K., Fan, X., Chong, Z., Chen, T., Chaisson, M.J., Huddleston, J., Malig, M., Nelson, B.J., Parrish, N.F., Blackburne, B., Lindsay, S.J., Ning, Z., Zhang, Y., Lam, H., Sisu, C., Challis, D., Evani, U.S., Lu, J., Nagaswamy, U., Yu, J., Li, W., Abecasis, G.R., Habegger, L., Yu, H., Cunningham, F., Dunham, I., Lage, K., Jespersen, J.B., Horn, H., Kim, D., Desalle, R., Narechania, A., Sayres, M.A.W., Mendez, F.L., Poznik, G.D., Underhill, P.A., Coin, L., Mittelman, D., Banerjee, R., Cerezo, M., Fitzgerald, T., Louzada, S., Massaia, A., Ritchie, G.R., Yang, F., Kalra, D., Hale, W., Dan, X., Barnes, K.C., Beiswanger, C., Cai, H., Cao, H., Henn, B., Jones, D., Kaye, J.S., Kent, A., Kerasidou, A., Mathias, R., Ossorio, P.N., Parker, M., Rotimi, C.N., Royal, C.D., Sandoval, K., Su, Y., Tian, Z., Tishkoff, S., Via, M., Wang, Y., Yang, H., Yang, L., Zhu, J., Bodmer, W., Bedoya, G., Cai, Z., Gao, Y., Chu, J., Peltonen, L., Garcia-Montero, A., Orfao, A., Dutil, J., Martinez-Cruzado, J.C., Mathias, R.A., Hennis, A., Watson, H., McKenzie, C., Qadri, F., LaRocque, R., Deng, X., Asogun, D., Folarin, O., Happi, C., Omoniwa, O., Stremlau, M., Tariyal, R., Jallow, M., Joof, F.S., Corrah, T., Rockett, K., Kwiatkowski, D., Kooner, J., Tran Tinh, H., Dunstan, S.J., Nguyen Thuy, H., Fonnie, R., Garry, R., Kanneh, L., Moses, L., Schieffelin, J., Grant, D.S., Gallo, C., Poletti, G., Saleheen, D., Rasheed, A., Brook, L.D., Felsenfeld, A., McEwen, J.E., Vaydylevich, Y., Duncanson, A., Dunn, M., Schloss, J.A., Brooks, L.D., Genomes Project, C., 2015. A global reference for human genetic variation. Nature 526, 68-+.

Andrews, S., 2010. "FastQC: A quality control tool for high throughput sequence data.",

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
Arath Raghavan Sudha Devi, M.K.S., Sagar, B.K.C., 2015. Light and electron microscopic studies on the Y organ of the freshwater crab *Travancoriana schirnerae*. Journal of Microscopy and Ultrastructure 3, 161-168.
Bliss, D.E., Boyer, J.R., 1964. Environmental Regulation of Growth in the Decapod Crustacean *Gecarcinus lateralis*. General and Comparative Endocrinology 4, 15-41.
Boisvert, S., Laviolette, F., Corbeil, J., 2010. Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. Journal of Computational Biology 17, 1519-1533.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.
Chang, E.S., Bruce, M.J., Tamone, S.L., 1993. Regulation of Crustacean Molting: A Multi-Hormonal System. American Zoologist 33, 324-329.
Chang, E.S., Mykles, D.L., 2011. Regulation of Crustacean Molting: a review and our perspectives. General and Comparative Endocrinology 172, 323-330.
Chang, J.Z., Wen; Zhou, Wen-Xin; Wang, Lan, 2016. Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. Biometrics arXiv:1505.04493v3.
Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., Bauer, D.J., Caceres, C.E., Carmel, L., Casola, C., Choi, J.-H., Detter, J.C., Dong, Q., Dusheyko, S., Eads, B.D., Froehlich, T., Geiler-Samerotte, K.A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E.V., Kueltz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J.R., Muller, J., Pangilinan, J., Patwardhan, R.P., Pitluck, S., Pritham, E.J., Rechtsteiner, A., Rho, M., Rogozin, I.B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y.I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J.R., Andrews, J., Crease, T.J., Tang, H., Lucas, S.M., Robertson, H.M., Bork, P., Koonin, E.V., Zdobnov, E.M., Grigoriev, I.V., Lynch, M., Boore, J.L., 2011. The Ecoresponsive Genome of *Daphnia pulex*. Science 331, 555-561.
Collins, F.S., Lander, E.S., Rogers, J., Waterston, R.H., Int Human Genome Sequencing, C., 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.
Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674-3676.
Covi, J.A., Bader, B.D., Chang, E.S., Mykles, D.L., 2010. Molt cycle regulation of protein synthesis in skeletal muscle of the blackback land crab, *Gecarcinus lateralis*, and the differential expression of a myostatin-like factor during atrophy induced by molting or unweighting. Journal of Experimental Biology 213, 172-183.
Covi, J.A., Chang, E.S., Mykles, D.L., 2009. Conserved role of cyclic nucleotides in the regulation of ecdysteroidogenesis by the crustacean molting gland. Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology 152, 470-477.
Covi, J.A., Chang, E.S., Mykles, D.L., 2012. Neuropeptide signaling mechanisms in crustacean and insect molting glands. Invertebrate Reproduction & Development 56, 33-49.
Das, S., 2015. Morphological, Molecular, and Hormonal Basis of Limb Regeneration across Pancrustacea. Integrative and Comparative Biology 55, 869-877.

Das, S., Pitts, N.L., Mudron, M.R., Durica, D.S., Mykles, D.L., 2016. Transcriptome analysis of the molting gland (Y-organ) from the blackback land crab, *Gecarcinus lateralis*. Comparative Biochemistry and Physiology D-Genomics & Proteomics 17, 26-40.
Drach, P.T., Catherine, 1967. Sur la méthode de détermination des stades d'intermue et son application générale aux crustacés. On the method of determining the intermolt stages and its general application to crustaceans. Biologie Marine 18, 595-610.
Du, X.J., Wang, J.X., Liu, N., Zhao, X.F., Li, F.H., Xiang, J.H., 2006. Identification and molecular characterization of a peritrophin-like protein from fleshy prawn (*Fenneropenaeus chinensis*). Molecular Immunology 43, 1633-1644.
Ellegren, H., 2014. Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution 29, 51-63.
Feuillet, C., Leach, J.E., Rogers, J., Schnable, P.S., Eversole, K., 2011. Crop genome sequencing: lessons and rationales. Trends in Plant Science 16, 77-88.
Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics 28, 3150-3152.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29, 644-U130.
Grada, A., Weinbrecht, K., 2013. Next-Generation Sequencing: Methodology and Application. Journal of Investigative Dermatology 133, E1-E4.
Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., Regev, A., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8, 1494-1512.
Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Research 17, 377-386.
Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., Itoh, T., 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research 24, 1384-1395.
Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81-93.
Kenny, N.J., Sin, Y.W., Shen, X., Zhe, Q., Wang, W., Chan, T.F., Tobe, S.S., Shimeld, S.M., Chu, K.H., Hui, J.H.L., 2014. Genomic Sequence and Experimental Tractability of a New Decapod Shrimp Model, Neocaridina denticulata. Marine Drugs 12, 1419-1437.
Kuballa, A.E., Abigail, 2007. Novel molecular approach to study moulting in crustaceans. Bull. Fish. Res. Agen. 20, 53-57.
Lachaise, F., Leroux, A., Hubert, M., Lafont, R., 1993. The Molting Gland of Crustaceans: Localization, Activity, And Endocrine Control (A Review). Journal of Crustacean Biology 13, 198-234.
Lander, E.S., Int Human Genome Sequencing, C., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P.,

McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H.M., Yu, J., Wang, J., Huang, G.Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S.Z., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H.Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W.H., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J.R., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., Int Human Genome Sequencing, C., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-U354.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data, P., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C.C., Lam, T.T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-W., Yiu, S.-M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.K.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., Wang, J., 2010. The sequence and *de novo* assembly of the giant panda genome. Nature 463, 311-317.

Liu, B., Yujian, S., Yuan, J.,Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., Fan, W. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Unpublished.

MacLea, K.S., Abuhagr, A.M., Pitts, N.L., Covi, J.A., Bader, B.D., Chang, E.S., Mykles, D.L., 2012. Rheb, an activator of target of rapamycin, in the blackback land crab, *Gecarcinus lateralis*: cloning and effects of molting and unweighting on expression in skeletal muscle. Journal of Experimental Biology 215, 590-604.

Marcais, G., Kingsford, C., 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27, 764-770.

Mardis, E.R., 2011. A decade's perspective on DNA sequencing technology. Nature 470, 198-203.

McCarthy, J.F., Skinner, D.M., 1977. Interactions Between Molting and Regeneration 2. Pro-Ecdysial Changes in Serum Ecdysone Titers, Gastrolith Formation, and Limb Regeneration Following Molt Induction by Limb Autotomy and/or Eyestalk Removal in the Land Crab, *Gecarcinus lateralis*. General and Comparative Endocrinology 33, 278-292.

Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shaw, P.D., Marshall, D., 2013. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics 14, 193-202.

Mykles, D.L., 2001. Interactions between Limb Regeneration and Molting in Decapod Crustaceans. American Zoologist 41, 399-406.

Mykles, D.L., 2011. Ecdysteroid metabolism in crustaceans. Journal of Steroid Biochemistry and Molecular Biology 127, 196-203.

Mykles, D.L., Adams, M.E., Gaede, G., Lange, A.B., Marco, H.G., Orchard, I., 2010. Neuropeptide Action in Insects and Crustaceans. Physiological and Biochemical Zoology 83, 836-846.

Nakatsuji, T., Lee, C.-Y., Watson, R.D., 2009. Crustacean molt-inhibiting hormone: Structure, function, and cellular mode of action. Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology 152, 139-148.

Pagani, F., Baralle, F.E., 2004. Genomic variants in exons and introns: identifying the splicing spoilers. Nature Reviews Genetics 5, 389-U382.

Roberts, A., Pachter, L., 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods 10, 71-U99.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Santiago-Sotelo, P., Humberto Ramirez-Prado, J., 2012. prfectBLAST: a platform-independent portable front end for the command terminal BLAST plus stand-alone suite. Biotechniques 53, 299-300.
Skinner, D.M., 1962. The Structure and Metabolism of a Crustacean Integumentary Tissue During a Molt Cycle. Biol. Bull. 123, 635-647.
Skinner, D.M., 1985. The Biology of Crustacea. Academic Press.
Skinner, D.M., Graham, D.E., 1972. Loss of Limbs as a Stimulus to Ecdysis in Brachyura (True Crabs). Biological Bulletin 143, 222-&.
Song, L., Bian, C., Luo, Y., Wang, L., You, X., Li, J., Qiu, Y., Ma, X., Zhu, Z., Ma, L., Wang, Z., Lei, Y., Qiang, J., Li, H., Yu, J., Wong, A., Xu, J., Shi, Q., Xu, P., 2016. Draft genome of the Chinese mitten crab, *Eriocheir sinensis*. Gigascience 5.
Unamba, C.I.N., Nag, A., Sharma, R.K., 2015. Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. Frontiers in Plant Science 6.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57-63.
Wetterstrand, K., 2016. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), Available at: www.genome.gov/sequencingcosts.
Willems, K.A., 1982. Larval Development of the Land Crab *Gecarcinus lateralis lateralis* (Fréminville, 1835)(Brachyura: Gecarcinidae) Reared in the Laboratory. Journal of Crustacean Biology 2, 180-201.
Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.-Y., Wei, L., 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Research 39, W316-W322.
Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13, 329-342.

APPENDIX: SCRIPTS FOR TRANSCRIPTOME

###Prefix before each script###

```
#!/bin/bash
#PBS -q batch
#PBS -l nodes=1:ppn=12
#PBS -l walltime=120:00:00
#PBS -N BamIndexSorted
#PBS -j oe

cd $PBS_O_WORKDIR
export NCORES=12
```

###Trimmomatic###

```
java -jar /opt/trimmomatic/0.33/trimmomatic-0.33.jar PE -threads 2 -phred33
/scratch/lamartin/ESrawData/In1/In1_ACAGTG_L005_R1_001.fastq.gz
/scratch/lamartin/ESrawData/In1/In1_ACAGTG_L005_R2_001.fastq.gz
/scratch/lamartin/trimfiles/In1/L005/In1_ACAGTG_L005_R1_forward_paired.fq.gz
/scratch/lamartin/trimfiles/In1/L005/In1_ACAGTG_L005_R1_forward_unpaired.fq.gz
/scratch/lamartin/trimfiles/In1/L005/In1_ACAGTG_L005_R2_reverse_paired.fq.gz
/scratch/lamartin/trimfiles/In1/L005/In1_ACAGTG_L005_R2_reverse_unpaired.fq.gz
ILLUMINACLIP:/opt/trimmomatic/0.33/adapters/TruSeq3-PE.fa:2:30:10
ILLUMINACLIP:/opt/trimmomatic/0.33/adapters/NexteraPE-PE.fa:2:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:28 MINLEN:36
```

###Trinity and Normalization###

```
Trinity --output Trinity_MLA --seqType fq --JM 81G --left
/scratch/lamartin/trimfiles/Forwardtrimfiles/MLA_left.fq --right
/scratch/lamartin/trimfiles/Reversetrimfiles/MLA_right.fq --CPU $NCORES --
min_contig_length 200 --SS_lib_type RF --full_cleanup --min_kmer_cov 2

#Normalization Job

# run trinity
Trinity --normalize_max_read_cov 20  --output Trinity_MLA_Norm --seqType fq --JM 80G --
left /scratch/lamartin/trimfiles/Forwardtrimfiles/MLA_left.fq --right
/scratch/lamartin/trimfiles/Reversetrimfiles/MLA_right.fq --CPU $NCORES --
min_contig_length 200 --SS_lib_type RF --full_cleanup --min_kmer_cov 2
```

###CD-HIT-EST###

```
module load cd-hit
export OMP_NUM_THREADS=12

time cd-hit-est -T 12 -i MLA_Trinity.fasta -o MLA_cdhitest-02_04.fasta -c 0.9 -n 8 -b 200 -M 30000
```

###Bowtie###

```
# run bowtie2
time bowtie2 -p $NCORES -x /home/lamartin/MLA_mapping/MLA_cdhitest-10_02.bt2 -1 /scratch/lamartin/Bowtie_Prep/DA1/DA1_ACTTGA_L005_R1_forward_paired.fq.gz -2 /scratch/lamartin/Bowtie_Prep/DA1/DA1_ACTTGA_L005_R2_reverse_paired.fq.gz -U /scratch/lamartin/Bowtie_Prep/DA1/DA1_ACTTGA_L005_unpaired.fq.gz -N 1 -S /scratch/lamartin/MLA_mapping/DA1_L005.sam
```

###SAMtools###

```
/opt/samtools/1.3/gcc/samtools view -b -S -o MLA.bam MLA.sam
/opt/samtools/1.3/gcc/samtools sort MLA.bam -o MLA.sorted
/opt/samtools/1.3/gcc/samtools index MLA.sorted
```

###eXpress###

```
/opt/express/1.5.1/prebuilt/express --rf-stranded -o /scratch/lamartin/DIRECTORIES_TEMPLATE/DA1/L005 /home/lamartin/ESA_cdhitest-10_02.fasta /scratch/lamartin/ESA_mapping/DA1_L005.bam
/opt/express/1.5.1/prebuilt/express --rf-stranded -o /scratch/lamartin/DIRECTORIES_TEMPLATE/DA1/L006 /home/lamartin/ESA_cdhitest-10_02.fasta /scratch/lamartin/ESA_mapping/DA1_L006.bam
/opt/express/1.5.1/prebuilt/express --rf-stranded -o /scratch/lamartin/DIRECTORIES_TEMPLATE/DA1/L007 /home/lamartin/ESA_cdhitest-10_02.fasta /scratch/lamartin/ESA_mapping/DA1_L007.bam
```

###EdgeR for Use in R Studio###

```
##EdgeR script for differential expression##

#open package
library(edgeR)

#read in data and establish groups/molt stages
```

```
x <- read.csv("2016_2_18_MLAfilt_cpm1_g3_no_mill_noPM1.csv", row.names=1)
group <- factor(c("Aim", "Aim", "Aim",
                "Bep", "Bep", "Bep",
                "Cmp", "Cmp", "Cmp",
                "Dlp", "Dlp", "Dlp",
                "Epm", "Epm"))

#make DGEList to store data and calculate library size
y <- DGEList(counts=x, group=group)
y$samples

#model-based normalization, TMM (2.7.6)
y <- calcNormFactors(y)
y$samples

#visualize library size
names = c("IM1", "IM2", "IM3",
        "EP1", "EP2", "EP3",
        "MP1", "MP2", "MP3",
        "LP1", "LP2", "LP3",
        "PM2", "PM3")
barplot(y$samples$lib.size*1e-6, ylim=c(0,30), names=names, main="Size of MLA Libraries",
xlab="Library", ylab="Library size (millions)")

#visualize library similarities
plotMDS(y)

#mean difference plot, 4.4.4 (should be 0)
#need to complete for all molt stages
par(mfrow=c(3,1))
plotMD(cpm(y, log=TRUE), column=1)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=2)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=3)
abline(h=0, col="red", lty=2, lwd=2)

par(mfrow=c(3,1))
plotMD(cpm(y, log=TRUE), column=4)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=5)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=6)
abline(h=0, col="red", lty=2, lwd=2)

par(mfrow=c(3,1))
```

```
plotMD(cpm(y, log=TRUE), column=7)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=8)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=9)
abline(h=0, col="red", lty=2, lwd=2)

par(mfrow=c(3,1))
plotMD(cpm(y, log=TRUE), column=10)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=11)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=12)
abline(h=0, col="red", lty=2, lwd=2)

par(mfrow=c(3,1))
plotMD(cpm(y, log=TRUE), column=13)
abline(h=0, col="red", lty=2, lwd=2)
plotMD(cpm(y, log=TRUE), column=14)
abline(h=0, col="red", lty=2, lwd=2)

#####
#establish experimental design (3.2.3)
design <- model.matrix(~0+group, data=y$samples)
colnames(design) <- c("Aim", "Bep", "Cmp", "Dlp", "Epm")
design

#####

#estimate dispersion (spread, scatter, distribution, 2.9.1)
#given a generalized linear model, using DGEList
y <- estimateDisp(y, design, robust = TRUE)

#plot square root of dispersion under negative binomial model
par(mfrow=c(1,1))
plotBCV(y)
y$common.dispersion

# Use quasi-likelihood F-test for differential expression
#reflects uncertaintiny in estimating gene dispersion
#provides robust and reliable error rate control, replicate
# number is small (2.10.3)
fit <- glmQLFit(y, design, robust = TRUE)
head(fit$coefficients)
plotQLDisp(fit)
```

```
#pull out specific contrast for differentail expression
#QL F-test tests for significant diff expression in each gene
my.contrasts <- makeContrasts(Bep_vs_Aim=Bep-Aim,
                         Cmp_vs_Bep=Cmp-Bep,
                         Dlp_vs_Cmp=Dlp-Cmp,
                         Epm_vs_Dlp=Epm-Dlp,
                         Epm_vs_Aim=Epm-Aim,
                         levels=design)
qlf <- glmQLFTest(fit, contrast=my.contrasts[,"Bep_vs_Aim"])
topTags((qlf))

#view bio rep similarities in psuedo count
top <- rownames(topTags(qlf))
cpm(y)[top,]

#export document
printout <- topTags(qlf, n=Inf)
write.csv(printout, file="2_EPtoIM.csv")
```

APPENDIX: SCRIPTS FOR GENOME

#The following scripts were used in genome assembly. Ray had increased nodes of 40, with wall time extended by OSU HPCC technician.

###Prefix before each script###

```
#!/bin/bash
#PBS -q batch
#PBS -l nodes=1:ppn=12
#PBS -l walltime=120:00:00
#PBS -N BamIndexSorted
#PBS -j oe

cd $PBS_O_WORKDIR
```

###Trimmomatic###

#Necessary to select appropriate adapters

```
java -jar /opt/trimmomatic/0.33/trimmomatic-0.33.jar PE -threads 2 -phred33 \
/scratch/lamartin/genomic_data/G._lateralis_gDNA_NoIndex_L008_R1_001.fastq \
/scratch/lamartin/genomic_data/G._lateralis_gDNA_NoIndex_L008_R2_001.fastq \
/scratch/lamartin/genomic_data/NoIndex_L008_forward_paired.fq.gz \
/scratch/lamartin/genomic_data/NoIndex_L008_forward_unpaired.fq.gz \
/scratch/lamartin/genomic_data/NoIndex_L008_reverse_paired.fq.gz \
/scratch/lamartin/genomic_data/NoIndex_L008_reverse_unpaired.fq.gz \
ILLUMINACLIP:/opt/trimmomatic/0.33/adapters/NexteraPE-PE.fa:2:30:10 \
SLIDINGWINDOW:4:20 MINLEN:35
```

###Concatenate Reads###

#Repeat for forward and reverse, only needed for Paired Assembly

```
cat
/scratch/lamartin/genomic_data/trim/Correct_Attempt/No_Index/Ray/NoIndex_L008_forward_p
aired.fq \
/scratch/lamartin/genomic_data/trim/Correct_Attempt/TGA/Ray/TGAAGAGA_L001_forward_
paired.fq > All_foward_paired.fq
```

###Ray Assembly###

```
module load ray/2.3.1

export DATA=/scratch/lamartin/genomic_data/trim/Correct_Attempt/No_Index/Ray
K=37
NP=`cat $PBS_NODEFILE | wc -l`
mpirun -np ${NP} Ray -k $K -p  $DATA/All_forward_paired.fq $DATA/All_reverse_paired.fq
-o ray$K
```

###Bowtie Build Index###

```
/opt/bowtie2/2.2.6/prebuilt/bowtie2-build
/scratch/lamartin/genomic_data/trim/Correct_Attempt/No_Index/Ray/ray37/Scaffolds.fasta \
/scratch/lamartin/genomic_data/trim/Correct_Attempt/No_Index/Ray
```

###Bowtie Mapping###

```
/opt/bowtie2/2.2.6/prebuilt/bowtie2 -x Ray -1 NoIndex_L008_forward_paired.fq -2
NoIndex_L008_reverse_paired.fq -S NoIndex.sam
```

###Sam to Bam Conversion###

```
/opt/samtools/1.3/gcc/samtools view -b -S -o ALL.bam ALL.sam
```

###Sorting Bam File###

```
/opt/samtools/1.3/gcc/samtools sort NoIndex.bam -o NoIndex.sorted
```

###Index Sorted Bam File###

```
/opt/samtools/1.3/gcc/samtools index NoIndex.sorted
```

###Convert Sorted Bam File###
#Ran interactively

```
cp NoIndex.sorted NoIndex.sorted.bam
```

<u>**APPENDIX: PERL SCRIPTS USED FOR TRANSCRIPTOME AND GENOME**</u>

#split_fasta.pl written by Paul Stothard, Canadian Bioinformatics Help Desk; provided by Das. Script used to divide .fasta files to run jobs in parallel.

#extract_sequences_pl.txt author unknown; provided by Das. Script used to extract specific sequences out of a .fasta file.

<u>###split_fasta.pl###</u>

```
#!/usr/bin/perl
#split_fasta.pl version 1.0
#This script accepts a file consisting of multiple FASTA formatted sequence records.
#It splits the file into multiple new files, each consisting of a subset of the original records.
#
#There are three command line options:
#
#-i input file.
#-o output file prefix. This script will append numbers to this prefix name so that each created
file is unique.
#-n the number of sequences to place in each output file.
#
#Example usage:
#
#perl split_fasta.pl -i sample_in.txt -o new_sequences -n 100
#
#Written by Paul Stothard, Canadian Bioinformatics Help Desk.
#
#stothard@ualberta.ca

use strict;
use warnings;

#Command line processing.
use Getopt::Long;

my $inputFile;
my $outputFile;
my $numberToCopy;

Getopt::Long::Configure ('bundling');
GetOptions ('i|input_file=s' => \$inputFile,
		'o|output_file_prefix=s' => \$outputFile,
		'n|number=i' => \$numberToCopy);
```

```
if(!defined($inputFile)) {
   die ("Usage: split_fasta.pl -i <input file> -o <output file> -n <number of sequences to write
per file>\n");
}

if(!defined($outputFile)) {
   die ("Usage: split_fasta.pl -i <input file> -o <output file> -n <number of sequences to write
per file>\n");
}

if(!defined($numberToCopy)) {
   die ("Usage: split_fasta.pl -i <input file> -o <output file> -n <number of sequences to write
per file>\n");
}

if ($numberToCopy <= 0) {
   die ("-n value must be greater than 0.\n");
}

#count the number of sequences in the file
#read each record from the input file

my $seqCount = 0;
my $fileCount = 0;
my $seqThisFile = 0;

open (OUTFILE, ">" . $outputFile . "_" . $fileCount) or die ("Cannot open file for output: $!");

open (SEQFILE, $inputFile) or die( "Cannot open file : $!" );
$/ = ">";

while (my $sequenceEntry = <SEQFILE>) {

   if ($sequenceEntry =~ m/^\s*>/){
         next;
   }

   my $sequenceTitle = "";
   if ($sequenceEntry =~ m/^([^\n]+)/){
         $sequenceTitle = $1;
   }
   else {
         $sequenceTitle = "No title was found!";
   }
```

```
    $sequenceEntry =~ s/^[^\n]+//;
    $sequenceEntry =~ s/[^A-Za-z]//g;

    #write record to file
    print (OUTFILE ">$sequenceTitle\n");
    print (OUTFILE "$sequenceEntry\n");
    $seqCount++;
    $seqThisFile++;

    if ($seqThisFile == $numberToCopy) {
        $fileCount++;
        $seqThisFile = 0;
        close (OUTFILE) or die( "Cannot close file : $!");
        open (OUTFILE, ">" . $outputFile . "_" . $fileCount) or die ("Cannot open file for
output: $!");
    }

}#end of while loop

close (SEQFILE) or die( "Cannot close file : $!");

close (OUTFILE) or die( "Cannot close file : $!");
```

###extract_sequences_pl.txt###

```
if (scalar(@ARGV) != 3)
        {
        print "\nThis script extracts fasta sequences using a list of names\n";
        print "The user should provide a file that contains the list of sequence names\n";
        print "and the fasta sequence file\, in addition to the output filename\n\n";
        print "\n\nUSAGE\: perl extract_sequence_pl \[list of sequence name\] \[fasta file\]
\[output file\]\n\n";
        exit;
        }
open (FILE1, "$ARGV[1]");
while ($line = <FILE1>)
        {
        chomp $line;
        if ($line =~ /\>/)
                {
                $line =~s /\>//g;
                ($ky, $c2)  = split (/\s+/, $line, 2);
                $myHash{$ky} = "\>$line\n";
                next;
                }
```

```
		$myHash{$ky} .= "$line\n";
		}
close (FILE1);

open (FILE3, ">$ARGV[2]");

open (FILE1, "$ARGV[0]");
while ($line = <FILE1>)
		{
		chomp $line;
		$line =~s /\>//g;
		if (exists($myHash{$line}))
				{
				print FILE3 "$myHash{$line}";
				next;
				}
		else
				{
				print "$line \.\.\.\. could not be extracted\n";
				}
		}
close (FILE1);
close (FILE3);

print "\n\nGenerated \"$ARGV[2]\"\n\n";
```

GLOSSARY OF BIOINFORMATICS TERMS

The descriptions below are not comprehensive, but are intended to provide additional clarification within the context of this project.

- Annotation: add biological meaning to contigs/transcripts in the transcriptome based on sequence comparisons to other species
  - Identity: provide name of gene product to unknown sequence
  - Functional: provide biological context to an identified gene product; includes KEGG pathways and GO terms
- Contig (genome): individual sequence, used to generate longer segments known as scaffolds
- Contig (transcriptome): individual sequence within the reference transcriptome; possible gene product; has a unique identifying number (also referred to as a transcript)
- Count: quantification of contig expression in the transcriptome, not normalized
- FPKM: normalized quantification of contig expression in the transcriptome, normalized based on library size and contig length
- Genome/draft genome: a database of assembled genomic sequences
- k-mer: a subunit of a read (length denoted by k) that is used to align reads into longer sequences in genome/transcriptome assembly, can also be quantified to estimate genome size
- N50: common metric for measuring scaffold length and assessing genome quality; larger N50s indicate higher quality genome assembly; calculation gives increased consideration to longer scaffolds
- phred score: quality score used to assess reads, higher phred scores indicate more accurate sequencing
- Read: the output of Illumina sequencer, used to assemble genome/transcriptome
- Scaffold: individual genomic sequences within the reference genome; each has a unique identifying number
- Transcript: individual sequence within the reference transcriptome; possible gene product; has a unique identifying number (also referred to as a contig)
- Transcriptome/Reference transcriptome: a database of assembled cDNA sequences

GLOSSARY OF BIOINFORMATICS TOOLS AND PACKAGES

The descriptions of tools and packages below are not comprehensive, but are intended to provide additional clarification within the context of this project. In addition to a description, packages that are not standard or have many alternatives contain a brief justification for their use.

- BLAST2GO Basic 3.0.8/BLAST2GO PRO 3.2.7: identify enriched GO terms; software provides clear visuals and graphics of enriched GO terms based the three major GO subcategories; interface is convenient with results obtained in the PRO version at an accelerated rate
- BLASTx: annotate transcriptome contigs (identity)
- Bowtie 2 2.2.3: align reads back to the reference transcriptome or draft genome
- CD-HIT-EST 4.6.1: cluster transcriptome contigs with nucleotide similarity; used to eliminate redundancy in the transcriptome in the absence of a reference genome
- Edge R 3.12.0: test for differential expression between molt cycle stages; package is well-cited in transcriptome projects, contains clear manual with diverse examples; and is a compilation of versatile tools (multiple options for library normalization, visualization, and statistical analysis)
- eXpress 1.5.1: quantify count and FPKM expression of contigs in transcriptome
- FastQC 0.11.3: view raw and trimmed reads
- Illumina HiSeqTM 2000: sample sequencer, produced 100 bp paired-end reads
- Jellyfish 2.1.4: quantify k-mers with a set of genomic reads; simple command line with straightforward manual; part of Trinity packages
- KOBAS 2.0: identify enriched KEGG pathways; online program contains multiple databases for comparative analysis; easy to use interface with multiple options; output includes both enriched and non-enriched (yet present) components in various KEGG pathways
- prfectBLAST: extract genome scaffolds or transcriptome contigs using unique identifying number; allows for local blast searches of genome/transcriptome database; program contained all desired functions for working with unpublished databases
- Ray 2.3.1: assemble trimmed reads into contigs and subsequent scaffolds (genome); recommended due to the small size of trimmed reads and for parallel processing, allowing for multiple iterations with minimal computational requirements
- SAMtools 0.1.18: file converter; required to quantify expression levels following Bowtie
- Tablet 1.15.09.01: visualize genomic scaffolds; simple interface and quickly provides metrics and a visual for individual scaffolds
- Trimmomatic 0.33: trim or remove raw reads based on quality
- Trinity r20140413p1: assemble trimmed reads into contigs (transcriptome)