PHEROMONE HOMOLOGS IDENTIFIED FROM THE CLOACAL GLAND TRANSCRIPTOME OF A MALE *AMBYSTOMA MEXICANUM*

By

Kevin William Hall

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for a degree of

Zoology - Master of Science

2015

ABSTRACT

PHEROMONE HOMOLOGS IDENTIFIED FROM THE CLOACAL GLAND TRANSCRIPTOME OF A MALE *AMBYSTOMA MEXICANUM*

By

Kevin William Hall

Pheromones play an important role in modifying vertebrate behavior, especially during courtship rituals (Wyatt 2014). The courtship behavior within the urodele group often includes female exposure to secretions from the cloacal gland, as well as other scent glands. The first vertebrate proteinaceous pheromone discovered, the decapeptide sodefrin, is a female attracting pheromone secreted by the cloacal glands of male *Cynops pyrrhogaster* (Kikuyama et al. 1995). Other proteinaceous pheromones within the urodele group have been shown to elicit responses from females towards conspecific males (Yamamoto et al. 2000, Kikuyama et al. 2002, Houck et al. 2008). However, the presence and levels of expression of proteinaceous pheromones has yet to be identified within the family Ambystomatidae, which includes several important research models. Therefore, the objective of this research was to identify putative proteinaceous pheromones from male *Ambystoma mexicanum*, as well as their relative expression levels. The results indicate that *A. mexicanum* contains two different forms of sodefrin precursor-like factor (alpha and beta (Janssenswillen et al. 2015)), as well as a putative ortholog of plethodontid modulating factor. The beta form of sodefrin precursor-like factor was amongst the most highly expressed transcripts within the cloacal gland. The ortholog of plethodontid modulating factor was expressed at a level equivalent to the beta sodefrin precursor-like factor. The presence of these highly expressed proteinaceous pheromones may indicate that male *A. mexicanum* use several chemical cues to attract female conspecifics.

This thesis is dedicated to Dr. Amy-Lynn Frankshun and Mr. Mooch Frankshun

ACKNOWLEDGMENTS

I would like to acknowledge Cory Kohn for his assistance with understanding how to code scripts to more efficiently process and analyze data. I would also like to thank Dr. Matthew Scholz for his assistance in understanding of the various modules utilized in the assembly and analysis of the transriptome discussed in this thesis. I would like to acknowledge Dr. Kevin Childs for his time and reassurances about the validity and accuracy of the assembled transcriptome. Additionally, I would like to thank Dr. Heather Eisthen for the axolotl tissues used in this thesis. Finally, I would like to thank Dr. Barry Williams for taking me under his advisement at Michigan State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AFP – Antifreeze protein

EC – Enzyme Commission

GO – Gene ontology

GPI – Glycophosphatidylinositol

KEGG – Kyoto Encyclopedia of Genes and Genomes

ORF – Open reading frame

PPS – Preprosodefrin

PRF – Plethodontid receptivity factor

PMF – Plethodontid modulating factor

SPF – Sodefrin precursor-like factor

TFD – Three-finger domain

TPM – Transcripts per million

INTRODUCTION

Conspecific chemical cues have a well-documented and important role in vertebrate behavior, including mammals (Meredith 1998, Dulac and Torello 2003), reptiles (Mason et al. 1989), fish (Stacey et al. 2003), and amphibians (Rajchard 2005). Within the salamander clade, conspecific chemical cues are involved in the recognition of individuals (Jaeger 1981), territoriality (Mathis 1990), and the modification of sexual receptivity (Houck et al. 2008). The first peptide pheromone discovered was the decapeptide sodefrin, which is released by the cloacal gland of the male salamander *Cynops pyrrhogaster* (Kikuyama et al. 1995). Since then, variants of sodefrin have been identified (Yamamoto et al. 2000, Nakada et al. 2007). In genera that court in water, such as *Cynops* and *Lissotriton*, the male indirectly transfers the pheromone by wafting water currents towards the female's nares via tail fanning movements (Osikowski et al. 2008). Within the family Plethodontidae, which exhibit terrestrial courtship displays, at least three peptide pheromones are secreted from the mental glands located under the chins of males. Direct transfer of pheromones during courtship occurs either when a male's mental gland makes contact with the female's nares or the pheromones are scratched into the female's skin with the male's premaxillary teeth (Palmer et al. 2007b, Palmer et al. 2007a, Houck et al. 2008b, Houck et al. 2008a). Protein pheromones are regulated by androgen and prolactin and are important for salamander courtship (Kikuyama et al. 2002).

Most protein pheromones related to courtship behavior in salamanders are comprised of at least one protein domain belonging to the three-finger domains (TFD) superfamily, which also contains uPar, Ly-6, CD59, and a number of snake toxins. TFDs are so named because they contain at least one tertiary domain structure of three loops extending from a hydrophobic core that is cross-linked by disulfide bridges (Tsetlin 1999). TFD-containing proteins are common across vertebrates and often include a signal peptide for their exportation via the Golgi apparatus and transport vesicles. In addition, the

number of TFDs per protein can vary, and the protein may either be membrane bound via a glycophosphatidylinositol (GPI)-anchor or excreted without a GPI-anchor in soluble proteins (Tsetlin 1999). The TFD has a strongly conserved cysteine motif with a high rate of sequence evolution in the intervening loop regions (Ploug and Ellis 1994). The function of any one TFD is often tissue specific. For example, the long- and short-chain neurotoxins and cardiotoxins found within snake venom are comprised of one TFD and lead to membrane disruption (Ploug and Ellis 1994). However, the complement regulatory protein CD59 has a similar structure to snake toxins, but has a GPI-anchor and functions as an inhibitor of complement-mediated lysis (Kieffer et al. 1994). Preprosodefrin (PPS), sodefrin's precursor protein, is comprised of a signal peptide, a TFD and a low complexity 63-amino acid carboxyl-terminal region. This low complexity region lacks any cysteine disulfide bridges and is cleaved by prohormone convertases to produce the biologically active conspecific female attractant, sodefrin (Iwata et al. 2004).

Since PPS lacks a GPI-anchor, the gene falls within the group of TFD-containing proteins that include the aforementioned snake toxins. Comparative genomic analyses indicate that, within the salamander clade, PPS is a derived homolog of an ancient salamander pheromone, sodefrin precursor-like factor (SPF) (Van Bocxlaer et al. 2014). SPF contains an 18-amino acid signal peptide, a complete TFD, a partial TFD and a short low complexity carboxyl-terminal region (Palmer et al. 2007). Janssenswillen et al. (2015) used molecular evolutionary approaches to split SPF-like genes into two groups, alpha SPF and beta SPF. The alpha SPF contains an amino-terminal eight-cysteine TFD ending with XCXXXXCN, followed by a six-cysteine partial TFD. The beta SPF contains an animo-terminal ten-cysteine TFD ending with CCXXXXCN, followed by an eight-cysteine partial TFD. Furthermore, beta SPF homologs within *Cynops* exhibit an additional 62-base pair (bp) insert between the first complete TFD and the second partial TFD, which comprises PPS. This insert resulted in the loss of the partial TFD by frameshift mutation and the creation of the low complexity tail, from which sodefrin is cleaved. An additional proteinaceous pheromone,

plethodontid modulating factor (PMF), was discovered within the mental gland of *Plethodon shermani* and is comprised of only a twenty-amino acid signal peptide, a complete TFD with a carboxyl CCXXXXCN motif (Palmer et al. 2007). Again, the cysteine motif is highly conserved, yet the intervening composition of amino acids evolved rapidly. It is unknown if the peptide pheromones, SPF, PPS, and PMF, are homologs (Janssenswillen et al. 2015). Lastly, plethodontid receptivity factor (PRF) is an additional peptide pheromone, unique to plethodontid salamanders. PRF lacks the cysteine spacing pattern commonly found within the TFDs of SPF, PMF, or PPS and thus is not part of the same TFD superfamily, but instead is a member of the interleukin-6 cytokine family (Rollmann et al. 1999).

The axolotl, *Ambystoma mexicanum,* is a great model salamander to further develop an understanding of signal mechanisms and potentially uncover receptors involved in conspecific chemical cues (Park et al. 2004). Axolotls are a neotenic salamander, naturally distributed in lakes in central Mexico. Adult axolotls retain juvenile characteristics, such as gills and a dependency on an aquatic lifestyle (Gadow 1903). Sexually mature axolotls do not undergo metamorphosis and thus express paedomorphic traits (Smith 1969). The male axolotl courtship behavior involves nudging the female with his snout and tail motions. The male and female will place their snouts near each other's' cloaca and as the male moves away from the female, she follows the male with her nares proximate to his cloaca. The female nudges the male near the cloaca with her snout. The male leads the female over a deposited spermatophore that consists of a packet of sperm atop a gelatinous base. As the male moves away from the spermatophore, the female follows the male until her cloaca comes into contact with the spermatophore and then she picks up the sperm packet with her cloaca (Salthe 1967). This courtship pattern is similar across many urodele species (Salthe 1967). In addition, conspecific chemical cues produced by glands critical to courtship have been shown to elicit courtship behavior in several members of the urodele group (Kikuyama et al. 1995, Kikuyama et al. 2002, Houck et al. 2008, Osikowski et al. 2008).

Given the female axolotl's attention towards the male cloaca during courtship and the potential for expression of multiple peptide pheromone homologs, I chose to assemble a transcriptome from the male cloacal gland. Transcriptomic data would facilitate the identification and relative expression levels of alpha SPF, beta SPF, PPS, PMF, and PRF orthologs, providing an opportunity to prioritize future studies of pheromones in this species. Specifically, our goals were to address the following three questions: Which homologs are expressed in the male cloacal gland? How many of each gene family are expressed in the male cloacal gland? Are there clear differences in the levels of expression amongst homologs?

In this study, the cloacal gland transcriptome from a male *Axolotl* is assembled to identify which homologs are expressed. The putative contigs from the transcriptome are filtered, mapped, and functionally annotated to determine the number of contigs within each gene family (PPS, PMF, SPF, and PRF) expressed within the male cloacal gland. Expression levels for the filtered contigs are calculated and any clear differences in the levels among the homologs associated with amphibian courtship behavior are identified. The results indicate that A. mexicanum contains two different forms of SPF (alpha and beta (Janssenswillen et al. 2015)), as well as a putative ortholog of PMF. The beta form of SPF was amongst the most highly expressed transcripts within the cloacal gland. The ortholog of PMF was expressed at a level equivalent to the beta SPF. The presence of these highly expressed proteinaceous pheromones may indicate that male A. mexicanum use several chemical cues to attract female conspecifics. Ultimately, a fully annotated transcriptome is completed, providing another important resource for the *A. mexicanum* genomics community.

MATERIALS AND METHODS

Animal maintenance and tissue collection

The *A. mexicanum* was obtained from the Ambystoma Genetic Stock Center (University of Kentucky, Lexington, KY). The animal was kept in recirculated, filtered Holtfreter's solution (60 mmol l^{-1} NaCl, 2.4 mmol l^{-1} $NaHCO_3$, 0.67 mmol l^{-1} KCl, 0.81 mmol l^{-1} $MgSO_4$, and 0.68 mmol l^{-1} $CaCl_2$ in dH_2O [pH 7.5]) and maintained at 20°C. The photoperiod was altered each month to match that of the animal's native Mexico City, MX habitat. The axolotl was fed commercial salmon chow (Rangen, Buhl, Idaho, USA) three times each week. The experimental conditions were approved by and carried out in accordance with the Michigan State University's Institutional Animal Care and Use Committee recommendations. The axolotl was sacrificed by decapitation and the cloacal gland was dissected out at the morphological boundaries that define glandular tissue from the surrounding non-glandular tissue. The entire gland was immediately frozen in liquid nitrogen, ground into a powder with a mortar and pestle, and stored at -80^{o}C until RNA extraction.

RNA-Seq preparation and sequencing

Total RNA was extracted from the axolotl cloacal gland tissue using RNeasy Mini Kit reagents (Qiagen, USA) according to manufacturer instructions. Residual DNA was removed using DNase I digestion and cDNA was constructed using the Promega cDNA kit. Briefly, total RNA (1 µg) was reverse-transcribed at 42^{o}C into DNA using 5 U of reverse transcriptase, 1x RT buffer , 1 mM of deoxyribonucleotide triphosphate, 10 ng/µl of random primer, and 20 U RNase inhibitor in a 20µL reaction. RNA quality and quantity were assessed with an Agilent 2100 BioAnalyzer. Approximately 20 µg of cDNA was used for sequence analysis via Illumina HiSeq™ 2000 platform at the Michigan State University Research

Technology Support Facility (East Lansing, MI) in order to generate 50 bp single-end reads. The raw data from Illumina HiSeq were deposited in the NCBI Short Read Archive (SRA) database.

RNA-Seq data trimming, read filtering and digital normalization

Trimmomatic v.0.32 (Bolger et al. 2014) was used to remove the single end adapters from the Illumina (FASTQ) data. Low quality reads or bases were removed using FastQ Quality Control Software v.1.3 (FaQCs) (Lo and Chain 2014). FaQCs was used to perform a low-quality filtering process where reads with a phred quality score lower than 20 were discarded. Bases with a phred quality score less than 20 were trimmed from the 5' and 3' ends of the reads. Reads with lengths that dropped below 35 bp were discarded. All retained reads comprised the 'filtered reads' database.

The filtered reads were digitally normalized to facilitate rapid transcriptome assembly. The program khmer v.1.3 and the short read sequence utility screed v.0.7 (Crusoe et al. 2014) executed the python script 'normalize-by-median.py', which reduced the number of redundant reads to twenty reads per k-mer. The k-mer length used to determine the maximum number of reads with repetitive sequences was twenty, based on recommendations from Brown et al. (2012).

RNA-Seq transcriptome assembly and contig filtering

Transcriptome assembly was carried out using the program Trinity v. r20140413p1 (Grabherr et al. 2011). The assembly parameters were set to --min_contig_length 200, --JM 48G, --glue_factor 0.01, and --min_iso_ratio 0.1, based upon the recommendations from Li et al. (2014). The program TransDecoder v.2.0.1 (Haas et al. 2013) translated predicted protein sequences from the assembled transcriptome by identifying open reading frames (ORFs) coding for at least 125 amino acids.

Putative proteins within contigs were used as queries to search three protein databases using BLAST+ v.2.2.30 (Altschul et al. 1990): *Xenopus tropicalis* Protein (ftp://ftp.xenbase.org/pub/Genomics/Sequences/xtropProtein.fasta), UniProtKB Swiss-Prot (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz), and Uniref – Uniref90 (ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref90/uniref90.fasta.gz). All three databases were downloaded on March 20, 2015. Contigs that had at least one match with an e-value of at least 10^{-5} from the BLASTx or BLASTp searches were retained for further analyses and comprised the 'filtered transcriptome assembly'. Bowtie2 v.2.2.3 was used to align the filtered reads to the filtered transcriptome assembly as a measure of assembly quality (Langmead and Salzberg 2012). To facilitate further discovery of potential contigs, the reads that Bowtie2 could not align to the filtered transcriptome were reassembled, and putative ORFs were identified as previously described. The filtered contigs from this unaligned read assembly were added to the existing data and duplicate contigs were removed, resulting in the final assembly.

Transcript expression and functional annotation

Functional annotation was carried out using Interproscan v.5.11-51.0, BLASTx or BLASTp to generate xml formatted files of all matches and their corresponding contigs from the final assembly, which were subsequently analyzed using BLAST2GO v.3.0.10 (Conesa et al. 2005). The top matches from BLASTx or BLASTp were retained from the *Xenopus* genome database. If there was no match to the *Xenopus* database, then the best match from the Refseq database was retained. This matching process was continued using databases in the following order: UniProtKB Swiss-Prot, Uniref90, and NR protein databases. BLAST2GO was then used to map and annotate the sequences with the associated GO terms describing biological processes, cellular components, and molecular functions. The sequences with

corresponding Enzyme Commission (EC) numbers obtained from BLAST2GO were mapped to the Kyoto Encyclopedia of Genes And Genomes (KEGG) to determine metabolic pathways that correspond with putative gene functions (Kanehisa et al. 2004). RSEM-eval from the DETONATE software package (Li et al. 2014) was used to estimate the abundance of transcripts by mapping the filtered reads back to the final assembly. Expression levels were calculated as Transcripts Per Million (TPM) with 95% confidence intervals. Descriptions of all custom scripts used for these analyses are included in the supplemental information.

RESULTS

Illumina HiSeq reads, read filtering, read assembly, assembly filtering

The workflow for bioinformatics analyses are illustrated in Figure 1. After removal of the sequence adapters, the Illumina single-end sequencing produced 187,617,671 reads of 50 bp each, totaling 9.38 gigabases (Gb) (Table 1). FaQCs filtered out 3,824,501 (2.04 %) reads with a quality score less than 20, resulting in 183,793,170 reads that were trimmed and filtered (97.96 % of the total raw reads). The final filtered and trimmed FastQ file contained 89.27% of the total reads with a quality score greater than or equal to 30. The GC content of the filtered data was 53.98%. The mean read length (± standard deviation) after filtering was 49.42 ± 2.02 bp. Redundant reads were removed through digital normalization resulting in 19,074,717 retained reads with a mean length of 49.56 ± 1.67 bp. The Trinity assembly of all of the filtered reads resulted in 17,953 contigs and a total length of 9,653,916 bp with a GC content of 45.56%. The N50 was 696 bp with a mean length of 537.73 ± 930.9 bp. The final assembly resulted in 7,952 contigs and a total length of 5,377,213 bp with a GC content of 47.48%. The N50 was 1,143 bp with a mean length of 676.21 ± 1246.3 bp.

Assembly analysis

The distribution of sequence lengths from the contigs is displayed in Figure 2. Of the 7,952 contigs obtained, 5,465 (68.7%), 2,000 (25.2%) and 487 (6.1%) were distributed between 200-599 bp, 600-1999 bp, and greater than or equal to 2000 bp, respectively. The longest transcript was 36,996 bp and sixteen transcripts were longer than 11,000 bp. Quality was measured by the percent of reads that Bowtie2 aligned to the final assembly. Bowtie2 aligned 90.2% of the filtered raw reads to the final assembly. TransDecoder identified putative proteins of at least 125 amino acids from the filtered transcripts, resulting in a predicted 3,168 (39.8%) contigs with ORFs meeting the minimum length criteria.

The top-hit species distribution generated from BLASTx and BLASTp against the *X. tropicalis*, RefSeq, UniprotKB, UniRef90, and NR protein databases is displayed in Figure 3. As a result of the filtering process, all of the 7,952 transcripts had at least one hit against *X. tropicalis*, UniprotKB or Uniref90 protein databases, which corresponds to 38% of the approximately 21,000 known protein-coding sequences in the *X. tropicalis* genome (Hellsten et al. 2010). Of the 7,952 contigs, 6,095 had a best match within the *X. tropicalis* database; however, forty genes received their only match (or statistically stronger match) within the *X. laevis* database.

A total of 5,261 (66.2%) contigs received at least one gene ontology (GO) term. A total of 17,706 GO terms, across fifteen GO levels, was applied to the sequences and were classified into one of three level-one GO terms: biological process, cellular component and molecular function (Figure 4). Biological processes represented 49%, cellular components represented 21%, and molecular functions represented 30% of the GO terms. Within the biological processes, “metabolic process” and “cellular process” represented the most abundant level-two GO annotations, occurring 1,242 and 1,219 times, respectively. Among the cellular component category, “cell” was the most abundant GO annotation, while “binding” and “catalytic activity” were the most abundant GO annotations in the molecular function category. The level-two GO classifications of the contigs are presented in Figure 5. Finally, 1,504 contigs were grouped into 99 KEGG pathways. The most highly represented pathways were ‘purine metabolism’ (302 contigs, 20.1%), ‘thiamine metabolism’ (244 contigs, 16.2%), and ‘biosynthesis of antibiotics’ (108 contigs, 7.2%).

The homologs of PPS, SPF, and PMF were identified by their characteristic domain structure, as well as the characteristic pattern of spacing among cysteines within their three finger domain(s) (Figure 7). These patterns, as well as significant BLAST scores and Interproscan matches, were used to identify all putative protein pheromone homologs within the final assembly (Table 3). An alignment of the putative

homolog to SPF is shown in Figure 7. A PPS homolog was not found when using strict criteria, which included a signal peptide, a complete TFD, and a low complexity region at the carboxyl-terminus. Ten SPF-like homologs were identified that contained a signal peptide at the amino-terminus, followed by one full TFD, a partial TFD, and a short low complexity region. Three PMF-like homologs were characterized by a signal peptide, one TFD, and a very short or nonexistent low complexity region. Five homologs to a previously discovered axolotl antifreeze protein (AFP) also were identified, but the sequence exhibited a pattern also consistent with PMF-like homologs (Zhang, 2013). The homologs' Gene IDs, domain structure, TPM and putative homolog are summarized in Table 3.

The number of contigs with at least two TPM, the minimal TPM considered biologically active, was 6,637 (Figure 6). Nineteen contigs exhibited an expression level greater than 1,000 TPM (See supplemental Table S1 for a complete list). Beta SPF was the 17th most highly expressed transcript at 1476.9 ± 6.4 (mean ± 95% confidence interval) TPM. A putative homolog of PMF was the 18th most highly expressed transcript at 1336.2 ± 15.0 TPM. A number of other cysteine-rich secretory proteins were highly expressed, but the spacing among cysteines was not consistent with a canonical three-finger domain pattern. Additionally, only one ribosomal protein was expressed at a higher level than beta SPF.

DISCUSSION

Ilumina single-end sequencing and assembly

With the development of next generation sequencing, data concerning the transcriptomes of non-model organisms have grown exponentially. Yet, most non-model organisms lack a well annotated reference genome. At present, there is no single standardized method agreed upon to assess the quality of a *de novo* transcriptome assembly; thus, different metrics are utilized. In this study, approximately 187 million single-end reads (9.38 Gb) were generated and assembled into 17,953 contigs. From this assembly, a total of 7,952 contigs (44.3%) had a BLASTx or BLASTp match, with a mean contig length of 676.2 ± 1246.3 bp. These contigs comprised the final assembly and their mean length is comparable to that of previous studies using Trinity as a *de novo* transcriptome assembler (Wheat and Vogel 2011, Li et al. 2015). The final assembly had an N50 score of 1,143 bp and a total of 1,147 contigs had lengths greater than 1,000 bp (14.4% of the filtered assembly), comparable to other salamander transcriptome assemblies (Li et al. 2015). According to Ramskold et al. (2009), approximately 60-70% of genes in the genome are expected to be expressed in any particular cell type. Because approximately 38% of genes known from the *Xenopus tropicalis* genome were detected, approximately half of the genes expressed in the male axolotl cloacal gland were identified. Alternatively, this tissue may express an unusually small number of genes, indicative of a highly specialized tissue. Further independent transcriptomic analyses will be required to distinguish between these competing hypotheses.

The contigs without BLAST matches may have been derived from untranslated regions, non-coding RNAs, sequences not containing a known protein domain, or may potentially represent genes unique to axolotls. These unmatched contigs could provide a potential resource for identifying novel genes expressed in the cloacal gland of male axolotls. Of particular interest is the large number of reads that mapped back onto the final assembly. Read alignment is an important determinant of assembly quality;

the higher the percentage of read alignment after transcript filtering, the better the assembly. Bowtie2 mapped greater than 90% of the filtered raw reads onto the final assembly, indicating that most reads were used to assemble the contigs with BLAST hits. This high percentage, in conjunction with the percentage of matched contigs, contig length, and N50 score, provides a high level of confidence that the assembly is accurate and consistent with comparable *de novo* transcriptome assemblies.

Functional annotation of contigs

Because axolotls lack a well annotated reference genome, it is difficult to predict the potential functions encoded by the cloacal gland transcripts; therefore, various protein databases were used to help identify putative functions based on homology. The results indicate that 6,095 contigs (76.6%) have a putative homolog within the *Xenopus* protein database. The remaining 1,857 contigs (23.4%) have a putative homolog within the RefSeq, UniProtKB Swiss-Prot, Uniref90, or NR protein databases. The 5,261 contigs (66.2%) that had at least one GO term is considered a large proportion of contigs to receive a GO term (Li et al. 2015). These GO terms comprised a wide array of GO categories. The breakdown of many GO categories is comparable to other transcriptome assemblies (Wheat and Vogel 2011, Li et al. 2015), except for the relatively high number of contigs (108 contigs [7.2%]) annotated within the "biosynthesis of antibiotics" category. One possibility is that these transcripts are expressed as a mechanism to control bacteria living within the cloaca or on the skin of axolotls, assisting in maintaining the microbiome of the host organism or preventing infection from invasive microorganisms (Rawls et al. 2006).

Pheromone homologs in the male *A. mexicanum* cloacal gland

The cloacal gland plays a primary role during axolotl courtship, similar to the cloacal gland of *C. pyrrhogaster* or the mental gland of *P. shermani* (Iwata et al. 2004, Nakada et al. 2007, Palmer et al.

2007, Houck et al. 2008). Given that proteinaceous pheromones are present in organs related to courtship in closely related species (Zhang et al. 2008, Janssenswillen et al. 2015), it is not surprising to find putative homologs identified based upon their primary sequence and domain structures. The orthology amongst PPS, SPF, and PMF is unclear; however Janssenswillen et al. (2015) showed that PPS and SPF are paralogs following a gene duplication of SPF into alpha SPF and beta SPF. In this study, putative homologs to both alpha and beta SPF, as well as PMF, were identified in axolotls; however, no putative homologs to PPS were identified. In addition, putative androgen and prolactin receptors in the cloacal gland were identified, which were shown to regulate synthesis of proteinaceous pheromones in *Cynops* (Kikuyama et al. 2002). Although alpha SPF has been previously identified within *A. mexicanum* (Janssenswillen et al. 2015), this is the first time a beta SPF has been identified in the species. Additionally, this is the first time a homolog to PMF has been identified outside of the plethodontids. One homolog to PRF was identified by Interproscan, based upon the PRF domain, but the contig was much shorter than and poorly aligned to PRF and thus was considered non-orthologous to PRF.

The SPF-like contigs contained a signal peptide at the amino-terminus, followed by a complete TFD and a partial TFD. The partial TFD of the SPF-like contigs contained eight of the twelve cysteines found within the anterior TFD and notably lacked the final CCXXXXCN motif. Whether the partial TFD forms the disulfide bridges typically found within the complete TFD is unknown. The complexity of the carboxyl-terminus of the SPF-like contigs has not been determined and may contain the canonical disulfide bridges or may have low complexity similar to PPS. Based upon this domain structure and the cysteine pattern, a total of four complete and six partial SPF-like contigs were identified. Of the ten SPF-like contigs, nine closely followed the beta SPF structure. One SPF-like contig followed the alpha SPF structure, except its posterior cysteine motif contained only five cysteines instead of the canonical six cysteines (Figure 7).

The PMF-like homologs contained a signal peptide and a complete TFD and then terminated shortly following the CCXXXXCN motif at the end of the TFD on the carboxyl-terminus. The PMF-like homologs contained more amino acids between the cysteines and were longer than the canonical PMF. Additionally, the spacing between the cysteines varied from SPF. This domain structure and cysteine pattern was consistent with one complete PMF-like contig and two partial PMF-like contigs. One partial PMF-like contig extended 25 amino acids beyond the CCXXXXCN motif. AFP follows a domain structure similar to PMF (Zhang et al. 2013) and has a moderate level of expression. AFP should be considered for conspecific chemical communication similar to PMF. One complete AFP-like contig was identified as well as four partial AFP-like contigs.

Peptide pheromones are amongst the most abundantly expressed genes in the male *A. mexicanum* cloacal gland

The program RSEM-eval revealed that peptide pheromones are among the most highly expressed genes within the male axolotl cloacal gland (Table 2). The top twenty most highly expressed genes included three forms of collagen (ranked 1, 3 and 4), two cell cycle checkpoint associated proteins (ranked 2 and 16), two ribosomal proteins (ranked 5 and 6), two forms of plasminogen (ranked 7 and 10), two forms of mucin (ranked 8 and 9), several uncharacterized proteins (ranked 11, 12, and 13), two cysteine-rich proteins (ranked 14 and 19), two forms of ferritin heavy chains (ranked 15 and 20), one complete SPF-like contig (ranked 17), and the complete PMF-like contig (ranked 18). Many of these most highly expressed genes make sense within the context of an axolotl cloaca. Collagen is among the most common proteins to be expressed in vertebrate cells (Di Lullo et al. 2002) and the salamander cloaca excretes mucus for the spermatophore (Stebbins and Cohen 1995). Generally, ribosomal proteins are expressed at very high levels relative to other proteins in most tissues (Amaldi et al. 1995). Since the level of expression of this SPF-like contig (1477 TPMs) and this PMF-like contig (1336 TPMs) exceed most

ribosomal proteins, these two pheromone proteins should be considered highly expressed. Additionally, a contig with a TPM greater than 100 was considered highly expressed and thus these two contigs would be considered very highly expressed based upon their TPM. The other complete and partial SPF-like contigs ranged in expression from 2 TPMs to 235 TPMs with three partial SPF-like contigs exceeding 100 TPMs. The two partial PMF-like contigs were expressed at 2 TPMs. The complete AFP contig and two of the four partial AFP contigs were each expressed around 32 TPMs. This study provides a first glimpse in determining relative expression levels, in a genome wide manner, by a relatively unbiased method (Li et al. 2014). These levels of expression provide a starting point for prioritizing which putative pheromones should be tested functionally.

FUTURE DIRECTIONS

The two most highly expressed putative protein pheromones from our final assembly, beta SPF-like and PMF-like, should be the focus of future experiments. These genes can be further studied by using a combination of proteomics, neurophysiology, and behavioral assays to determine the functional significance of each gene independently and in combination. The proteomic assay would determine the amount of protein that is actually translated from these contigs, as well as potentially identify disulfide bridge formation within the partial TFD within the SPF-like contig. These genes also could be synthesized within a cell culture containing protein disulfide isomerase family member 6. This specific protein disulfide isomerase may form the appropriate disulfide bridges within the TFD given that eight contigs of protein disulfide isomerase family member 6 were identified within the final assembly with an expression level in the range of 3 TPMs to 27 TPMs. These synthesized protein pheromones could be utilized in electrophysiological experiments to identify the pheromone receptor for a neurophysiological response. In addition, behavioral assays could determine if the presence of the synthesized protein pheromones could elicit a courtship activity, such as the hula display, from axolotl females (Park et al. 2004).

APPENDICES

APPENDIX A

Tables and Figures

Table 1. Summary of sequencing data and assembly of the male *A. mexicanum* cloacal gland transcriptome

Category	Raw Data	Filtered Raw Data	Digital Normalization	Assembly	Final Assembly
Total Reads	1.87×10^8	1.83×10^8 (98% of raw reads)	1.9×10^7 (10% of raw reads)		
Total Bases (Gb)*	9.4	9.1	0.9		
Mean read length (bp[§] ± S.D.[¥])	50.0 ± 0.0	49.4 ± 2.0	49.6 ± 1.7		
GC content	54%	53.98%	46.97%	45.56%	47.48%
Contigs > 200 bp				17,953	7,952
N50				696 bp	1,143 bp
Average length ± S.D.				537.7 ± 930.9 bp	676.2 ± 1246.3 bp
Total length				9,653,916 bp	5,377,213 bp

* gigabases (Gb), [§] base pairs (bp), [¥] standard deviation (S.D.)

Table 2. BLAST matches of the most highly expressed contigs with a transcript per million (TPM) score greater than or equal to 1000 in the male *A. mexicanum* cloacal gland transcriptome, listed from most highly expressed to least highly expressed. The putative homologs to SPF and PMF are bolded with their BLAST description in parentheses.

Transcript ID	TPM	E-value	Species	Protein Description[§]
c3405_g1_i2	307839.2	3.9×10^{-58}	*Rattus norvegicus*	Collagen alpha-1 chain-like
c3412_g1_i2	284003.8	4.1×10^{-20}	*Xenopus tropicalis*	Chk1 checkpoint homolog
unc7258_g1_i1	81920.0	6.9×10^{-6}	*Cavia porcellus*	Collagen alpha-1 chain-like
c3405_g1_i1	81032.7	5.4×10^{-15}	*Rattus norvegicus*	Proline-rich protein HaeIII subfamily 1-like
unc2297_g1_i1	67435.5	1.6×10^{-20}	*Manacus vitellinus*	40S ribosomal protein S29
c11288_g1_i1	29833.9	1.6×10^{-12}	*Sorex araneus*	FAM228b-like protein
unc2267_g1_i1	9441.2	9.0×10^{-8}	*Epinephelus bruneus*	Plasminogen
c3422_g1_i1	6545.2	1.2×10^{-30}	*Xenopus tropicalis*	Integumentary mucin -like
c2358_g1_i1	4353.1	2.7×10^{-180}	*Xenopus tropicalis*	Mucin-5B- partial
unc2554_g1_i1	3832.7	4.1×10^{-20}	*Epinephelus bruneus*	Plasminogen
c4389_g1_i1	3677.0	1.1×10^{-15}	*Sarcophilus harrisii*	Uncharacterized protein
unc757_g1_i1	3632.4	5.0×10^{-6}	*Clostridium papyrosolvens*	Uncharacterized protein
unc2538_g1_i1	3110.5	2.8×10^{-7}	*Strigamia maritima*	Uncharacterized protein
c1055_g1_i1	1774.6	6.6×10^{-21}	*Elephantulus edwardii*	A-kinase anchor protein 13-like
c2453_g1_i1	1670.1	2.4×10^{-114}	*Xenopus tropicalis*	Ferritin, heavy polypeptide 1
unc3333_g1_i1	1586.6	1.6×10^{-20}	*Loa loa*	Senescence-associated protein
c3350_g1_i1	**1476.9**	**4.6×10^{-11}**	***Xenopus tropicalis***	**SPF-like (Phospholipase A2 inhibitor subunit gamma B-like)**
c2860_g1_i1	**1336.2**	**3.8×10^{-27}**	***Cynops pyrrhogaster***	**PMF-like (Sodefrin)**
c2847_g1_i1	1332.5	3.9×10^{-52}	*Xenopus tropicalis*	Cysteine-rich secretory protein 3
c1304_g1_i1	1307.1	2.7×10^{-113}	*Xenopus tropicalis*	Ferritin, heavy polypeptide 1

[§]Raw descriptions from BLAST match

Table 3. Putative full length homologs of alpha SPF, beta SPF, PMF, and AFP.

Transcript ID	Domain Structure	TPM	Putative Homolog
c3350_g1_i1		1476.9	Beta SPF
c2860_g1_i1		1336.2	PMF
c7051_g1_i1		34.6	Alpha SPF
c883_g1_i1		32.1	AFP
c11097_g1_i1		8.6	Beta SPF

 Signal peptide

 Full Three-finger domain

 Partial three-finger domain

 Low complexity region

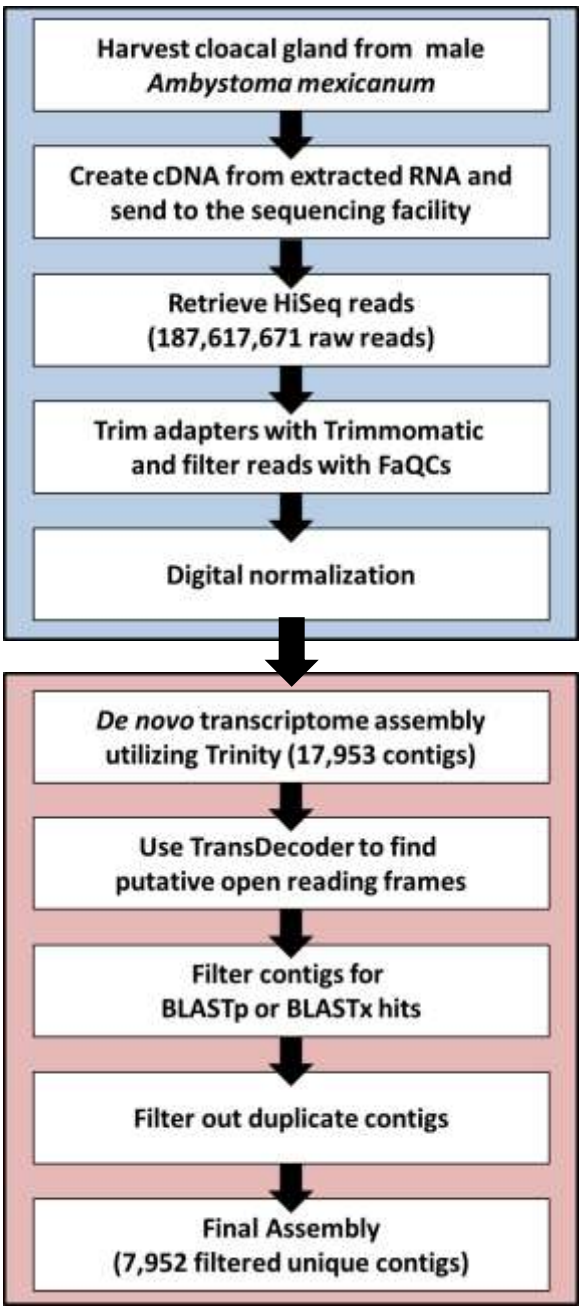
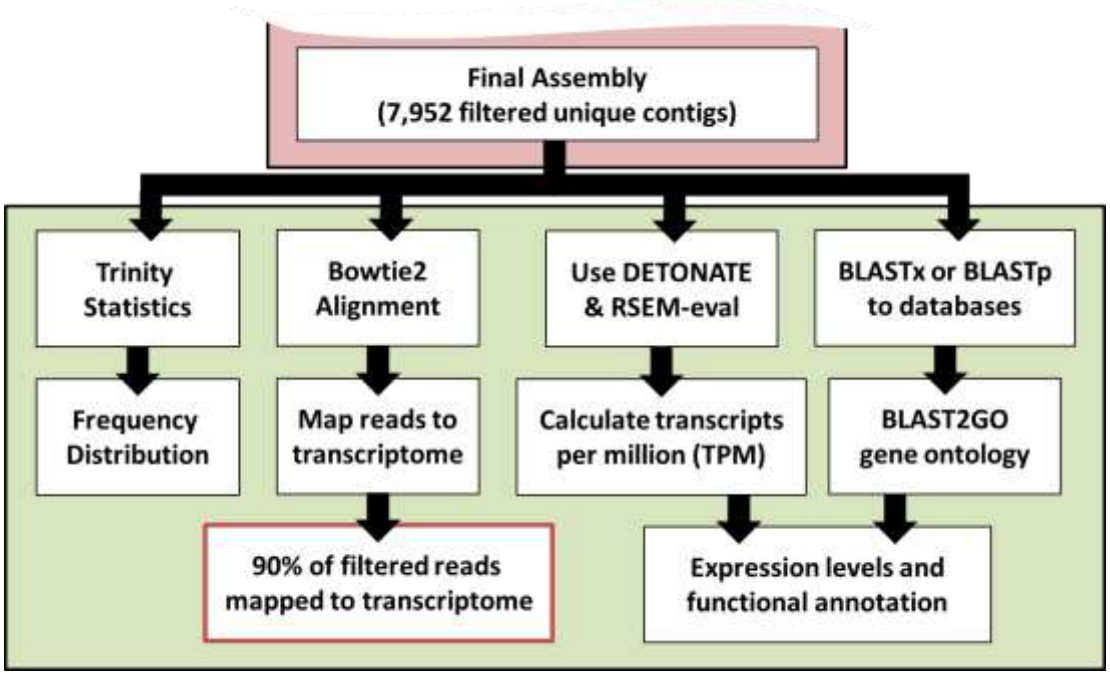


Figure 1. Pictorial representation of the process to assemble a transcriptome. First, RNA was extracted from the cloacal gland and HiSeq Illumina reads were created. The sequence reads were parsed and filtered for quality and removal of adaptor sequences (blue). Next, *de novo* assembly was generated and the transcripts were filtered based upon BLAST hits and redundant contigs were removed (red). Reads were mapped back to contigs, genes were annotated, and gene ontology was applied using BLAST and BLAST2GO (green). Finally, an analysis of the assembly and the quantity and distribution of transcripts was performed.

Figure 1 (cont'd)



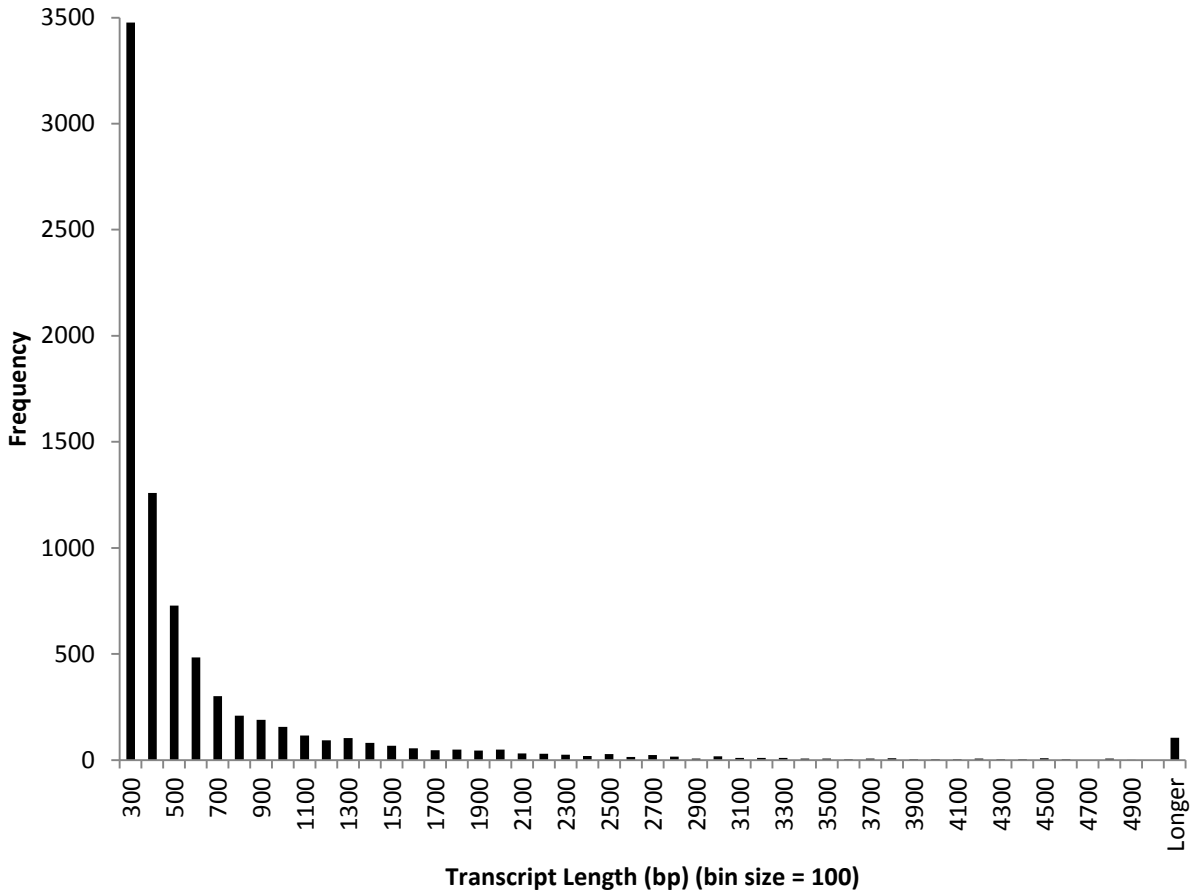


Figure 2. Distribution of sequence lengths from the final transcriptome assembly. The x-axis indicates sequence sizes from 200 bp to ≥ 5000 bp with a bin size of 100; the y-axis indicates the number of transcripts for a given sequence length bin. Most (60%) of the final assembly sequences range between 200 399 nucleotides in length; however, more than 6% of transcripts are longer than 2000 bp.

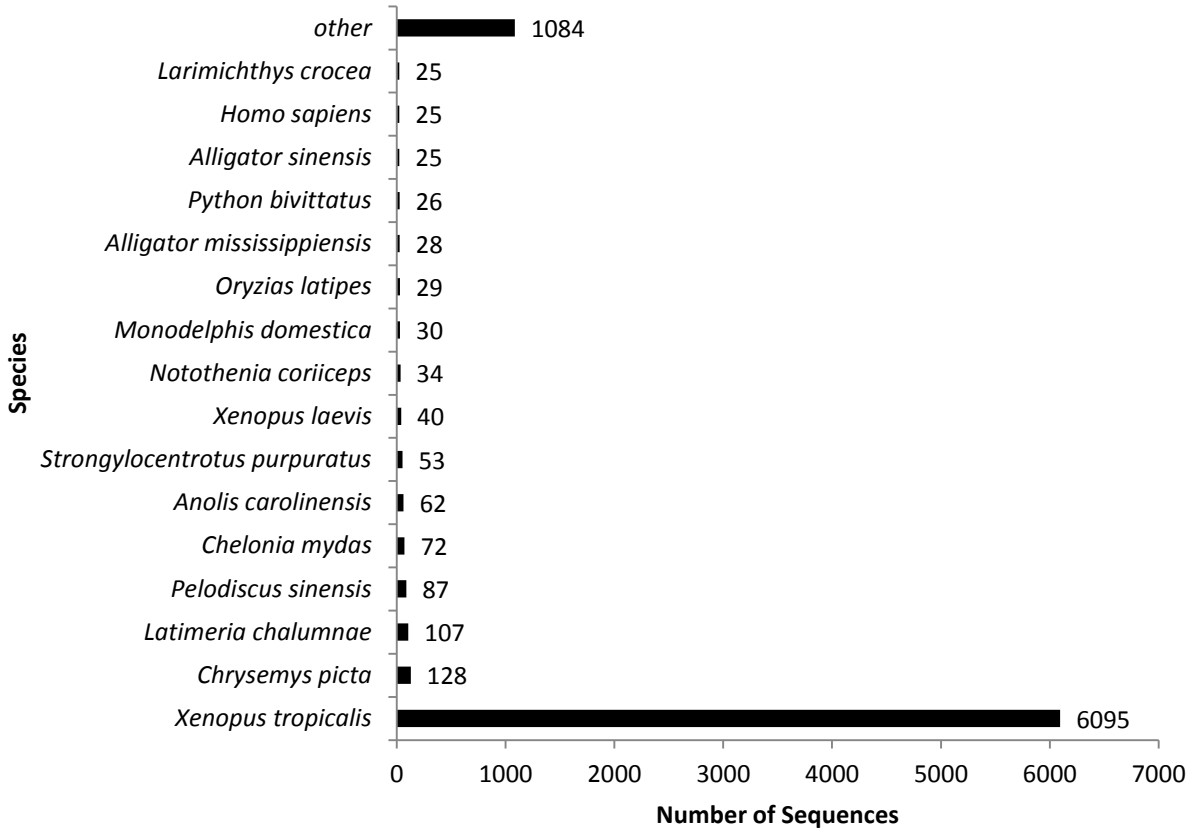


Figure 3. BLASTx or BLASTp top-hit species distribution generated from matches to the *X. tropicalis*, UniProtKB, and UniRef90 protein databases.

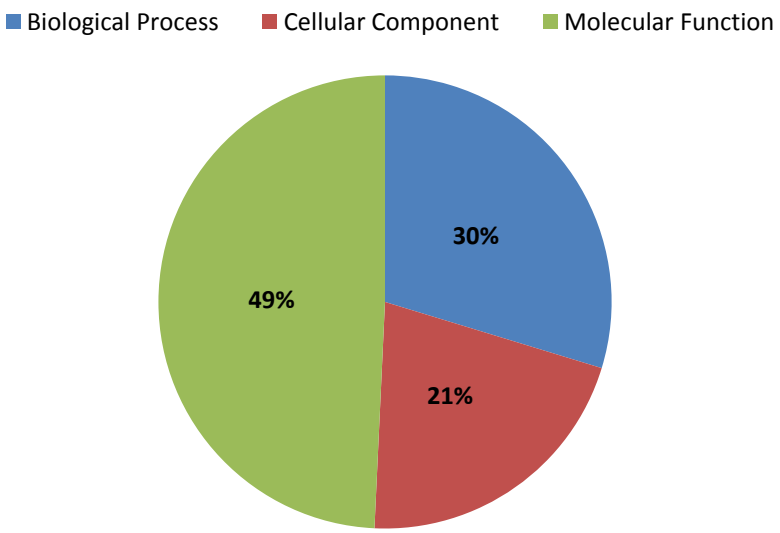


Figure 4. Distribution of GO annotations among biological processes, cellular components and molecular function.

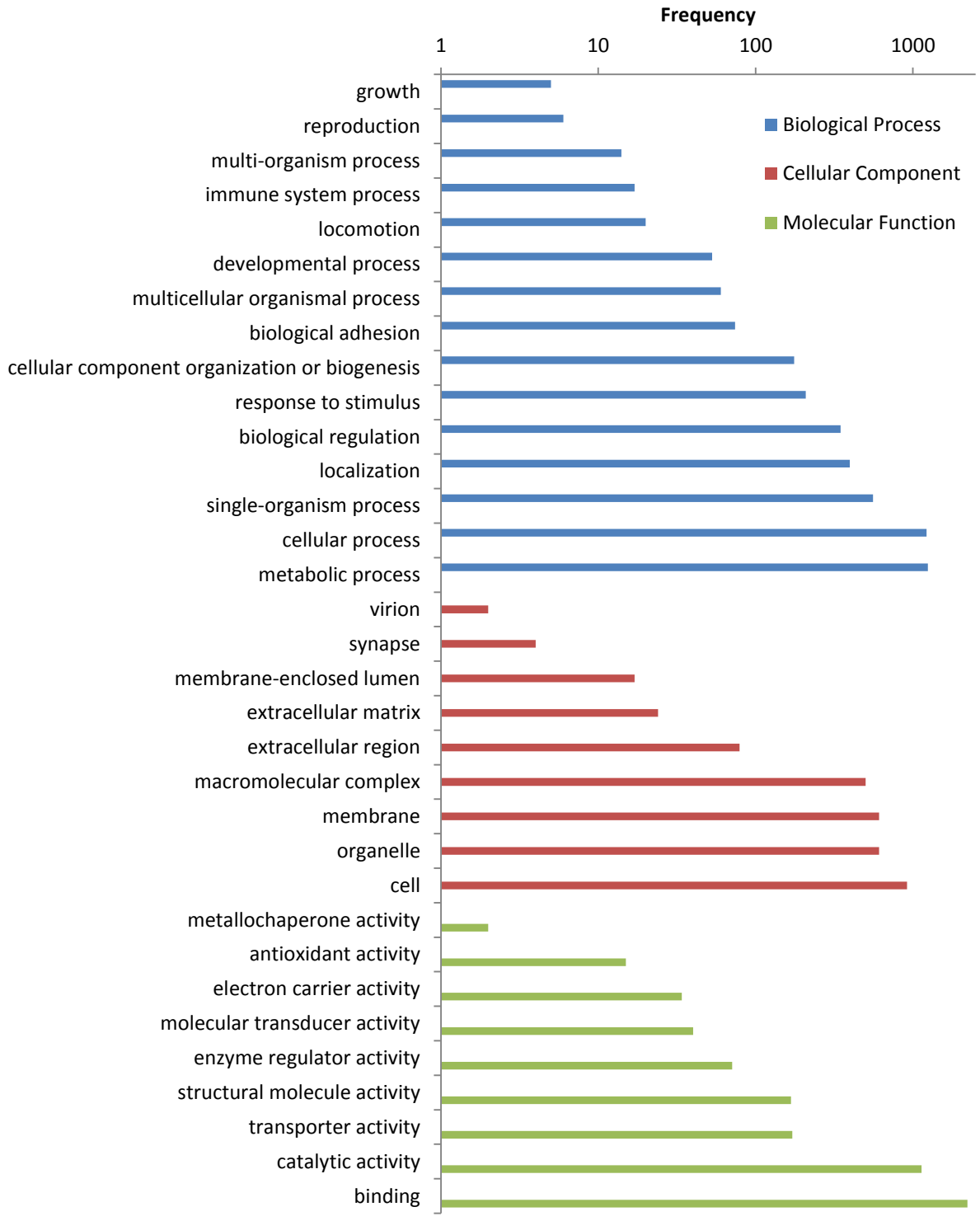


Figure 5. Level two GO classifications of final assembly contigs. The results are summarized in three main categories: biological process (blue), cellular component (red) and molecular function (green). The x-axis indicates the number of genes in a category on a log scale. The y-axis on the left indicates the specific category of genes in that main category.

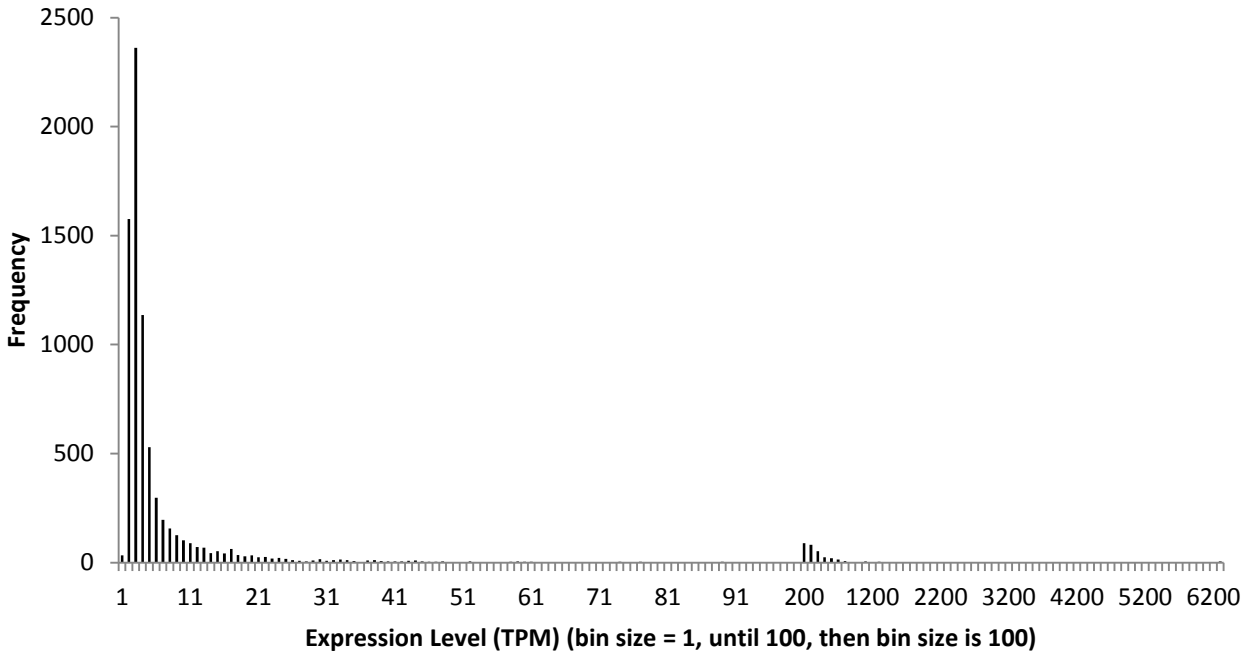


Figure 6. Frequency distribution of expression levels of putative genes expressed in transcripts per million (TPM). The x-axis indicates expression levels from 0 TPM to 100 TPM with a bin size of 1 and from 101 to ≥ 6200 with a bin size of 100; the y-axis indicates the number of transcripts for a given TPM bin.

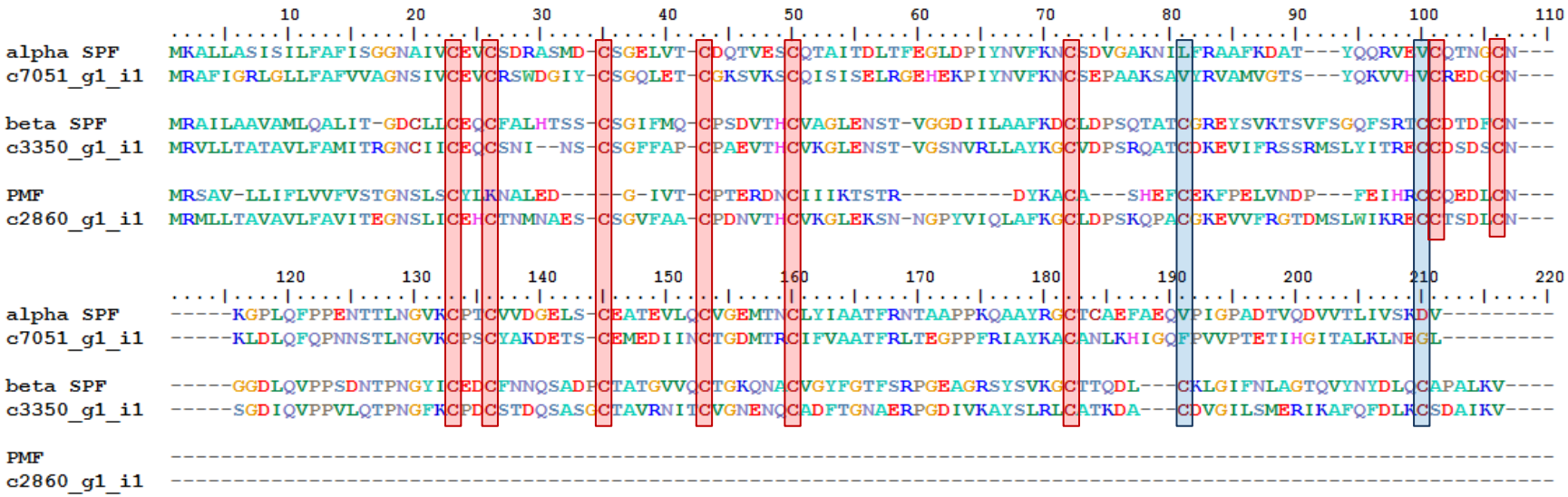


Figure 7. Comparison of deduced amino acid sequences for alpha SPF, beta SPF, and PMF aligned using ClustalW Multiple Alignment. The alpha SPF example is from *Notophthalmus viridescens* and the best homolog was contig c7051_g1_i1 which was expressed at 35 TPM. The beta SPF example is from *Lissotriton vulgaris* and the best homolog was contig c3350_g1_i1 which was expressed at 1477 TPM. The PMF example is from *Plethodon shermani* and the best homolog was contig c2860_g1_i1 which was expressed at 1336 TPM.

APPENDIX B

Code

```
java -jar $TRIM/trimmomatic SE -threads 4
/path/WilliamsEisthen_ATCACG_L001_R1_001.fastq
/path/WEH.fastq.trimmed
ILLUMINACLIP:/path/illuminaClipping.fa:2:30:10
```

Code 1: Trimmomatic (v0.32) used four threads that required an additional core for a total request for ppn=5. Trimmomatic then took the original fastq file (WilliamsEisthen_ATCACG_L001_R1_001.fastq) and the file containing the Illumina adapters (illuminaClipping.fa:2:30:10) and removed the adapters from each sequence, creating a file of fastq sequences without the Illumina adapters (WEH.fastq.trimmed). The 'WEH' is for Williams (PI), Eisthen (collaborator) and Hall. This command took less than ten minutes to execute using two gigabytes of memory in the process.

```
FaQCs.pl -u /path/WEH.fastq.trimmed –q 20 –avg_q 20 -d /path/FaQCs_output_WEH
-threads 8 -prefix WEH -min_L 35
```

Code 2: FaQCs (v1.3) is a perl script that analyzes the unpaired reads (-u WEH.fastq.trimmed) for reads (-avg_q 20) and bases (-q 20) with a target number of 20 as quality level for trimming. The number of threads (-threads) (ppn) used was eight and the prefix (-prefix) added to each of the output files was 'WEH'. The trimmed reads are set to have a minimum length of at least 35 (-min_L 35). Three files were created by this command: a pdf of graphics and reports (WEH_qc_report.pdf), a summary of the stats (WEH.stats.txt), and a FastQ file of the reads (WEH.unpaired.trimmed.fastq). FaQCs combines various features of comparable applications into a single, user-friendly process, where the original data and trimmed summaries are stated within a portable document format (pdf) containing an assortment of graphics and reports, enabling a simple assessment of data quality control. This command took less than two hours to execute using less than one gigabyte of memory in the process.

Digital normalization removes highly redundant (high-coverage) reads from the billions of generated reads from the sequencer and normalizes average coverage. This reduces the total number of reads to be assembled and any sampling variation, as well as, the sequencing errors contained within removed reads (Brown et al. 2012). A k-mer is a nucleotide polymer of 'k' length.

```
normalize-by-median.py -C 20 -k 20 -N 4 -x 2e9 /path/WEH.unpaired.trimmed.fastq
-o /path/WEH.fq.keep
```

Code 3: Digital normalization utilized screed (v0.7) (Brown et al. 2012) and khmer (v1.3) (Crusoe et al. 2014) to take the filtered reads from outputted by FaQCs from Code 2 and limits the number of redundant reads to twenty by discarding sequences based on whether or not their median k-mer abundance lies above the specified cutoff (-C 20). A k-mer is a nucleotide sequence of 'k' length. The k-mer length specified here is twenty (-k 20). Four k-mer counting tables (-N 4) were used, each with a hash table size of $2x10^9$ (-x 2e9). The WEH.fq.keep file was created containing the reads from WEH.unpaired.trimmed.fastq that were not discarded due to overly high coverage. This command took less than two hours to execute using approximately eight gigabytes of memory in the process.

Trinity has the ability to be broken down into five individual stages. This compartmentalization enabled faster assembly due to shorter wait times on the high performance computing cluster without impacting the quality of the transcriptome assembly (Code 4). Trinity has three main components, “Inchworm”, “Chrysalis”, and “Butterfly”. The Inchworm component assembles the filtered reads by extending the reads through k-mer space to create a set of contigs, where each read is used no more than once. Next, the Chrysalis component groups contigs that share at least (k-1)-mer bases, and if reads span the junction between contigs, it then builds individual de Bruijn graphs for each group. Finally, the Butterfly component takes the de Bruijn graphs from Chrysalis and removes unlikely paths resulting in more compact graphs. It then reconciles the de Bruijn graph with reads and outputs one linear sequence for each alternatively spliced transcript represented in the de Bruijn graph (Grabherr et al. 2011). Each stage of Trinity required the initial filtered single-end reads and output directory of where to save files created while Trinity ran. The minimum isomer ratio parameter setting means that ten percent of one inchworm contig must align to another inchworm contig in order for the two contigs to be put together as one larger contig. The low glue factor parameter setting means fewer reads are required to support the joining of inchworm contigs, resulting in a greater likelihood that contigs are discovered (Li et al. 2014). The minimum contig length parameter sets the minimum length of contigs in the final Trinity.fasta file to 200 bp. The final parameter specifies how much memory the subprogram Jellyfish should utilize. In order to compartmentalize each stage, a separate parameter was added to each coding sequence. These parameters would indicate where Trinity should end. With each subsequent stage, Trinity would check to see what files are present and automatically pick up the program where it left off from previous stage. The first stage generates the kmer-catalog using the program ‘Jellyfish’ and stopped prior to Inchworm beginning (Code 4a). Upon completion of the kmer-catalog, Inchworm was executed and stopped prior to running Chrysalis (Code 4b). The third step began to run the Chrysalis component by clustering the Inchworm contigs and mapping reads, but stopping before graph quantification (Code 4c).

The fourth step finished Chrysalis by creating the de Bruijn graphs and stopping before Butterfly (Code 4d). The fifth and final step excludes any --no_ options and runs Butterfly to generate the final Trinity.fasta file (Code 4e). At this point, a Fasta file of a set of putative contigs was created and renamed 'WEH.fasta'.

(a) Trinity.pl --seqType fq --single /path/WEH.fq.keep --glue_factor 0.01 --min_iso_ratio 0.1 --min_contig_length 200 --JM 48G --output /path/trin_WEH --no_run_inchworm
(b) Trinity.pl --seqType fq --single /path/WEH.fq.keep --glue_factor 0.01 --min_iso_ratio 0.1 --min_contig_length 200 --JM 48G --output /path/trin_WEH --no_run_chrysalis
(c) Trinity.pl --seqType fq --single /path/WEH.fq.keep --glue_factor 0.01 --min_iso_ratio 0.1 --min_contig_length 200 --JM 48G --output /path/trin_WEH --no_run_quantifygraph
(d) Trinity.pl --seqType fq --single /path/WEH.fq.keep --glue_factor 0.01 --min_iso_ratio 0.1 --min_contig_length 200 --JM 48G --output /path/trin_WEH --no_run_butterfly
(e) Trinity.pl --seqType fq --single /path/WEH.fq.keep --glue_factor 0.01 --min_iso_ratio 0.1 --min_contig_length 200 --JM 48G --output /path/trin_WEH

Code 4: The ability of trinity to be compartmentalized enables faster shorter wait times on high-performance computing clusters. For each step, the glue factor was set to 0.01 (--glue_factor 0.01), the minimum isomer ratio was set to 0.1 (--min_iso_ratio 0.1), the minimum contig length was set to 200 base pairs (--min_contig_length 200) and the jellyfish memory was set to 48 gigabytes (--JM 48G). The first stage required the initial filtered single-end reads (--single /path/WEH.fq.keep) and output directory (--output /path/trin_WEH) to generate the kmer-catalog using the program Jellyfish and stopped prior to Inchworm beginning (--no_run_inchworm) (a). This first step took less than ten minutes to execute using less than forty gigabytes of memory. Upon completion of the kmer-catalog, Inchworm was executed and stopped prior to running Chrysalis (--no_run_chrysalis) (b). This second step took approximately an hour to execute, using seventeen gigabytes of memory. The third step began to run the Chrysalis component by clustering the Inchworm contigs and mapping reads, but stopping before graph quantification (--no_run_quantifygraph) (c), taking approximately 45 minutes in the process and utilizing less than four gigabytes of memory. The fourth step finished Chrysalis by creating the de Bruijn graphs and stopping before Butterfly (--no_run_butterfly) (d). This step took less than ten minutes and less than one gigabyte of memory. The fifth step excludes any --no_ options and runs Butterfly to generate the final Trinity.fasta file (e). This step took less than two hours to execute utilizing less than five gigabytes of memory in the process.

TransDecoder is packaged with the Trinity program and has several components to maximize the likelihood of identifying putative ORFs (Code 5).

a) TransDecoder.LongOrfs -t /path/WEH.fasta -m 125
b) blastp -query /path/WEH.fasta.transdecoder_dir/longest_orfs.pep -db /path/uniprot_sprot.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 8 -out /path/WEH.fasta.transdecoder_dir/blastp_WEH_uniprot.outfmt6
c) hmmscan --cpu 8 --domtblout /path/WEH.fasta.transdecoder_dir/pfam_WEH.domtblout /path/Pfam-AB.hmm /path/WEH.fasta.transdecoder_dir/longest_orfs.pep
d) TransDecoder.Predict -t /path/WEH.fasta --retain_pfam_hits /path/WEH.fasta.transdecoder_dir/pfam_WEH.domtblout --retain_blastp_hits /path/WEH.fasta.transdecoder_dir/blastp_WEH_uniprot.outfmt6

Code 5: In order to identify the maximum number of putative protein coding regions, (a) TransDecoder first searches for the longest ORFs in the list of contigs (-t WEH.fasta) that are at least 125 amino acids in length (-m 125) and outputs them into longest_orfs.pep. (b) BLASTp (Altschul et al. 1990) broke this list of putative proteins (-query longest_orfs.pep) into eight individual threads for eight computer cores (-num_threads 8) to search for alignments within the UniProKB Swiss-Prot database (-db uniprot_sprot.fasta) and returns the top one target sequence (-max_target_seqs 1) with an e-value greater than 10^{-5} (-evalue 1e-5) in a tabular output format (-outfmt 6 -out blastp_WEH_uniprot.outfmt6). (c) Next, the list of putative proteins (longest_orfs.pep) is searched for protein families found within the pfam database (Pfam-AB.hmm) to create a tabular output of putative pfam domains. (d) Finally, TransDecoder runs predictions about which contigs (-t WEH.fasta) are proteins using the information from the longest ORFs, BLASTp results (--retain_blastp_hits blastp_WEH_uniprot.outfmt6) and identified pfam domains (--retain_pfam_hits pfam_WEH.domtblout). This process required about six hours to execute and utilized less than two gigabytes of memory in the process.

The Basic Local Alignment Search Tool (BLAST) allows for the searching of databases for sequences to which the putative contigs align (Altschul et al. 1990). BLASTx takes nucleotide sequences and translates them into all six frames and then takes those protein sequences to search for alignments against a protein database. BLASTp takes amino acid sequences and searches for alignments against a protein database.

a) blastx -query /path/WEH.fasta -db /path/xtropProtein.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 8 -out /path/blastx_xenprot_WEH.outfmt6
b) blastx -query /path/WEH.fasta -db /path/uniref90.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 8 -out /path/blastx_uniref90_WEH.outfmt6
c) blastx -query /path/WEH.fasta -db /path/uniprot_sprot.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 8 -out /path/blastx_sprot_WEH.outfmt6
d) blastp -query /path/WEH.fasta.transdecoder.pep -db /path/xtropProtein.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 8 -out /path/blastp_xenprot_WEH.outfmt6
e) blastp -query /path/WEH.fasta.transdecoder.pep -db /path/uniref90.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 8 -out /path/blastp_uniref90_WEH.outfmt6
f) blastp -query /path/WEH.fasta.transdecoder.pep -db /path/uniprot_sprot.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -num_threads 8 -out /path/blastp_sprot_WEH.outfmt6

Code 6: BLASTx (a-c) or BLASTp (d-f) divided this list of contigs (-query WEH.fasta) into eight individual threads for eight computer cores (-num_threads 8) to search for alignments within the *X. tropicalis* protein database (a), UniRef90 protein database (b), UniProKB Swiss-Protein database (c) and returns the top one target sequence (-max_target_seqs 1) with at least an e-value greater than 10^{-5} (-evalue 1e-5) in a tabular output format (-outfmt 6). The time and memory requirements for BLAST to function depend upon the size of the query and the size of the database. When querying the *X. tropicalis* protein database, the time and memory requirements are much lower because the database contains only proteins from *X. tropicalis*, finishing within an hour and utilizing less than two gigabytes of memory, while the UniRef90 database may require up to half a day to complete and fifteen gigabytes of memory.

a) bowtie2-build /path/WEH_filtered.fasta /path/WEH_filtered
b) bowtie2 -p 7 --end-to-end -x /path/WEH_filtered -U /path/WEH. unpaired.trimmed.fastq -S /path/ WEH_filtered_align.sam --un /path/ --al /path/

Code 7: Bowtie2 takes reads and aligns the reads against a reference genome. In this case, the filtered fasta transcriptome (WEH_filtered.fasta) sequences are used as the reference genome. Bowtie2-build creates a Bowtie index from the transcriptome sequences, outputting six files with an assigned basename (WEH_filtered), which are required to align the filtered reads to the reference transcriptome (a). The Bowtie2 program utilizes seven cores (-p 7, but note you will need one more core available) to align the single-ended filtered reads (-U WEH. unpaired.trimmed.fastq) from FaQCs.pl (Code 2) to the index files of the reference genome (-x WEH_filtered) (b). Bowtie2 aligns the reads end to end, preventing any soft-clipping or trimming of the reads (--end-to-end). Bowtie2 can also create a sequence alignment/map file (-S WEH_filtered_align.sam) and parses the unaligned reads (--un) from the aligned reads (--al). This process took less than an hour to execute and less than one gigabyte in memory.

a) rsem-eval-estimate-transcript-length-distribution /path/WEH_filtered_total.fasta /path/para_filt_tot_WEH.eval
b) rsem-eval-calculate-score -p 8 --output-bam --calc-ci --bowtie2 --overlap-size 5 --transcript-length-parameters /path/para_filt_tot_WEH.eval /path/ WEH.unpaired.trimmed.fastq /path/WEH_filtered_total.fasta filt_tot_WEH_Score 50

Code 8: RSEM-eval first calculates the average transcript length (676.2) and the standard deviation (1246.3) of the multi-fasta complete filtered transcriptome (WEH_filtered_total.fasta) and outputs the calculations to a parameter file (para_filt_tot_WEH.eval) (a). Then RSEM-eval utilizes eight threads (-p 8) and Bowtie2 (--bowtie2) to estimate the level of expression for each transcript by mapping the filtered reads (WEH.unpaired.trimmed.fastq) to the complete filtered transcriptome (WEH_filtered_total.fasta) when the read length is specified (50). Each outputted file and directory has the assigned prefix (filt_tot_WEH_Score). RSEM-eval can also generate a BAM file (--output-bam) and calculate the confidence intervals (--calc-ci) for the Transcripts Per Million (TPM). This process required about five hours and sixteen gigabytes of memory to execute.

a) interproscan.sh -d /path/iprscan/ -f XML -goterms
-i /path/WEH_unique_total.xx.fasta -t n -T temp_xx –iprlookup

b) blastx -query /path/WEH_unique_total.xx.fasta -db /path/xenopus/xtropProtein_mod.fasta
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/xenopus_outfmt5/blastx_xenopus_WEH_xx.outfmt5.xml

c) blastx -query /path/WEH_unique_total.xx.fasta -db /path/uniprot/uniprot_sprot_mod.fasta
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/uniprot_outfmt5/blastx_uniprot_WEH_xx.outfmt5.xml

d) blastx -query /path/WEH_unique_total.xx.fasta -db /path/uniref90/uniref90_mod.fasta
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/uniref90_outfmt5/blastx_uniref90_WEH.xx.outfmt5.xml

e) blastx -query /path/WEH_unique_total.xx.fasta -db /path/nr/nr.fasta
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/nr_outfmt5/blastx_nr_WEH.xx.outfmt5.xml

f) blastx -query /path/WEH_unique_total.xx.fasta -db /path/refseq_protein/refseq_protein
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/refseq_protein_outfmt5/blastx_refseq_WEH.xx.outfmt5.xml

g) blastp -query /path/WEH_unique_total.xx.fasta.transdecoder.pep
-db /path/xenopus/xtropProtein_mod.fasta
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/xenopus_outfmt5/blastp_xenopus_WEH_xx.outfmt5.xml

h) blastp -query /path/WEH_unique_total.xx.fasta.transdecoder.pep -db
/path/uniprot/uniprot_sprot_mod.fasta
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/uniprot_outfmt5/blastp_uniprot_WEH_xx.outfmt5.xml

i) blastp -query /path/WEH_unique_total.xx.fasta.transdecoder.pep
-db /path/uniref90/uniref90_mod.fasta
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/uniref90_outfmt5/blastp_uniref90_WEH.xx.outfmt5.xml

j) blastp -query /path/WEH_unique_total.xx.fasta.transdecoder.pep
-db /path/nr/nr.fasta
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/nr_outfmt5/blastp_nr_WEH.xx.outfmt5.xml

k) blastp -query /path/WEH_unique_total.xx.fasta.transdecoder.pep
-db /path/refseq_protein/refseq_protein
-word_size 3 -show_gis -num_alignments 20 -max_hsps 20 -outfmt 5 -evalue 1e-5 -num_threads 8
-out /path/refseq_protein_outfmt5/blastp_refseq_WEH.xx.outfmt5.xml

Code 9: Each nucleotide sequence was run through interproscan (a) and BLASTx (b-f). Each peptide sequence was run through BLASTp (g-k). The term ‘xx’ refers to the numeric value assigned to the fasta sequence file by ‘split_fasta.pl’. The amount of time required depended on the length of the sequence. For most sequences, the time was less than one hour, however

there were a number of sequences that required more than one, but less than six hours. All sequences needed at least fifteen gigabytes of memory due to the size of the protein databases utilized.

BIBLIOGRAPHY

BIBLIOGRAPHY

Altschul, S. F., et al. (1990). "Basic local alignment search tool." Journal of molecular biology **215**(3): 403-410.

Amaldi, F., et al. (1995). "Structure and expression of ribosomal protein genes in Xenopus laevis." Biochemistry and cell biology **73**(11-12): 969-977.

Bolger, A. M., et al. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics: btu170.

Brown, C. T., et al. (2012). "A reference-free algorithm for computational normalization of shotgun sequencing data." arXiv preprint arXiv:1203.4802.

Conesa, A., et al. (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Crusoe, M., et al. (2014). "The khmer software package: enabling efficient sequence analysis." URL http://dx. doi. org/10.6084/m9. figshare **979190**.

Di Lullo, G. A., et al. (2002). "Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen." J Biol Chem **277**(6): 4223-4231.

Dulac, C. and A. T. Torello (2003). "Molecular detection of pheromone signals in mammals: from genes to behaviour." Nat Rev Neurosci **4**(7): 551-562.

Gadow, H. (1903). "The mexican axolotl." Nature **67**(1736): 330-332.

Grabherr, M. G., et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol **29**(7): 644-652.

Haas, B. J., et al. (2013). "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." Nat Protoc **8**(8): 1494-1512.

Hellsten, U., et al. (2010). "The genome of the Western clawed frog Xenopus tropicalis." Science **328**(5978): 633-636.

Houck, L. D., et al. (2008). "A recombinant courtship pheromone affects sexual receptivity in a plethodontid salamander." Chemical senses **33**(7): 623-631.

Houck, L. D., et al. (2008). A candidate vertebrate pheromone, SPF, increases female receptivity in a salamander. Chemical Signals in Vertebrates 11, Springer**:** 213-221.

Iwata, T., et al. (2004). "Processing of multiple forms of preprosodefrin in the abdominal gland of the red-bellied newt Cynops pyrrhogaster: regional and individual differences in preprosodefrin gene expression." Peptides **25**(9): 1537-1543.

Jaeger, R. G. (1981). "Dear enemy recognition and the costs of aggression between salamanders." American Naturalist: 962-974.

Janssenswillen, S., et al. (2015). "Origin and diversification of a salamander sex pheromone system." Mol Biol Evol **32**(2): 472-480.

Kanehisa, M., et al. (2004). "The KEGG resource for deciphering the genome." Nucleic acids research **32**(suppl 1): D277-D280.

Kieffer, B., et al. (1994). "Three-dimensional solution structure of the extracellular region of the complement regulatory protein CD59, a new cell-surface protein domain related to snake venom neurotoxins." Biochemistry **33**(15): 4471-4482.

Kikuyama, S., et al. (1995). "Sodefrin: a female-attracting peptide pheromone in newt cloacal glands." Science **267**(5204): 1643-1645.

Kikuyama, S., et al. (2002). "Peptide and protein pheromones in amphibians." Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology **132**(1): 69-74.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Li, B., et al. (2014). "Evaluation of de novo transcriptome assemblies from RNA-Seq data." Genome Biol **15**(12): 553.

Li, F., et al. (2015). "RNA-Seq Analysis and Gene Discovery of Andrias davidianus Using Illumina Short Read Sequencing." PLoS One **10**(4): e0123730.

Lo, C.-C. and P. S. Chain (2014). "Rapid evaluation and quality control of next generation sequencing data with FaQCs." BMC bioinformatics **15**(1): 366.

Mason, R. T., et al. (1989). "Sex pheromones in snakes." Science **245**(4915): 290-293.

Mathis, A. (1990). "Territoriality in a terrestrial salamander: the influence of resource quality and body size." Behaviour **112**(3): 162-175.

Meredith, M. (1998). "Vomeronasal, Olfactory, Hormonal Convergence in the Brain: Cooperation or Coincidence? a." Annals of the New York Academy of Sciences **855**(1): 349-361.

Nakada, T., et al. (2007). "Isolation, characterization and bioactivity of a region-specific pheromone,[Val 8] sodefrin from the newt Cynops pyrrhogaster." Peptides **28**(4): 774-780.

Osikowski, A., et al. (2008). "Asymmetric Female Preferences for Courtship Pheromones in the Abdominal Glands of the Smooth Newt (Lissotriton vulgaris) and Montandon's Newt (L. montandoni)(Salamandridae)." Zoological science **25**(6): 587-592.

Palmer, C. A., et al. (2007). "Plethodontid modulating factor, a hypervariable salamander courtship pheromone in the three-finger protein superfamily." FEBS Journal **274**(9): 2300-2310.

Palmer, C. A., et al. (2007). "EVOLUTIONARY REPLACEMENT OF COMPONENTS IN A SALAMANDER PHEROMONE SIGNALING COMPLEX: MORE EVIDENCE FOR PHENOTYPIC-MOLECULAR DECOUPLING." Evolution **61**(1): 202-215.

Park, D., et al. (2004). "Discrimination of conspecific sex and reproductive condition using chemical cues in axolotls (Ambystoma mexicanum)." J Comp Physiol A Neuroethol Sens Neural Behav Physiol **190**(5): 415-427.

Ploug, M. and V. Ellis (1994). "Structure—function relationships in the receptor for urokinase-type plasminogen activator Comparison to other members of the Ly-6 family and snake venom α-neurotoxins." FEBS letters **349**(2): 163-168.

Rajchard, J. (2005). "Sex pheromones in amphibians: a review." Veterinary Medicine–Czech **50**(9): 385-389.

Ramskold, D., et al. (2009). "An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data." PLoS Comput Biol **5**(12): e1000598.

Rawls, J. F., et al. (2006). "Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection." Cell **127**(2): 423-433.

Rollmann, S. M., et al. (1999). "Proteinaceous pheromone affecting female receptivity in a terrestrial salamander." Science **285**(5435): 1907-1909.

Salthe, S. N. (1967). "Courtship patterns and the phylogeny of the urodeles." Copeia: 100-117.

Smith, H. M. (1969). "The Mexican axolotl: some misconceptions and problems." BioScience **19**(7): 593-615.

Stacey, N., et al. (2003). "Hormonally derived sex pheromones in fish: exogenous cues and signals from gonad to brain." Can J Physiol Pharmacol **81**(4): 329-341.

Stebbins, R. C. and N. W. Cohen (1995). A natural history of amphibians, Princeton Univ Press Princeton.

Tsetlin, V. (1999). "Snake venom α-neurotoxins and other 'three-finger'proteins." European Journal of Biochemistry **264**(2): 281-286.

Van Bocxlaer, I., et al. (2014). "Ancient pheromone blend as an alternative for copulation in internally fertilizing salamanders." PeerJ PrePrints **2**.

Wheat, C. W. and H. Vogel (2011). Transcriptome sequencing goals, assembly, and assessment. Molecular Methods for Evolutionary Genetics, Springer**:** 129-144.

Wyatt, T. D. (2014). Pheromones and animal behavior: chemical signals and signatures, Cambridge University Press.

Yamamoto, K., et al. (2000). "Silefrin, a sodefrin-like pheromone in the abdominal gland of the sword-tailed newt, Cynops ensicauda." FEBS letters **472**(2): 267-270.

Zhang, P., et al. (2008). "Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes." Mol Phylogenet Evol **49**(2): 586-597.

Zhang, S., et al. (2013). "Molecular cloning, sequence analysis and homology modeling of the first caudata amphibian antifreeze-like protein in axolotl (Ambystoma mexicanum)." Zoolog Sci **30**(8): 658-662.