

Transcriptomic analysis of ovarian development in parasitic *Ichthyomyzon castaneus* (chestnut lamprey) and non-parasitic *Ichthyomyzon fossor* (northern brook lamprey)

By

Nisha Ajmani

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biological Sciences

University of Manitoba

Winnipeg, Manitoba

Copyright © 2017 by Nisha Ajmani

## Abstract

Lampreys are primitive jawless fishes that diverged over 550 million years ago. As adults, they are either parasitic or non-parasitic. In non-parasitic species, sexual differentiation and oocyte development generally occur earlier than in parasitic species; fecundity is reduced and sexual maturation is accelerated following metamorphosis. The genes controlling ovarian differentiation and maturation in lampreys are poorly understood. This study used RNA-Seq data in the parasitic chestnut lamprey *Ichthyomyzon castaneus* and non-parasitic northern brook lamprey *Ichthyomyzon fossor* to identify suites of genes expressed during different stages of ovarian development that show different developmental trajectories with respect to ovarian differentiation and sexual maturation. For this, reference-guided and *de novo* assembly pipelines were designed for studying a non-model species. To test and explore the relative advantages of the pipelines, expression of insulin superfamily genes was used. This research helps to identify genes involved in lamprey ovarian development and provides insight into evolution of the insulin superfamily in vertebrates.

## **Acknowledgements**

Firstly, I would like to express my sincere gratitude to my advisors, **Prof. Dr. Margaret Docker & Prof. Dr. Sara Good** for their generous guidance, continuous support and motivation throughout my Master's and related research.

I sincerely appreciate all their contributions of time, inestimable support and ideas to make my research experience productive and stimulating. The joy, optimism, and enthusiasm they have for research were inspiring, motivational and contagious. This tenure has been a life changing experience for me and has been a source of friendships and collaboration. Your help in every aspect was imperative to my completion of this project.

I would also like to thank my committee, **Dr. Colin Garroway and Dr. Mike Domaratzki** for their advice, encouragement and insightful comments.

I am really thankful to NSERC Discovery Grant (MFD, SVG), Great Lakes Fishery Commission (MFD), University of Manitoba Faculty of Science Graduate Enhancement of Tri-Council Stipends (MFD), Faculty of Graduate Studies, Faculty of Science and Department of Biological Sciences (travel award) for financially supporting this project.

I have cherished every single day of my research in these last two years and it has given me a positive direction towards life and I feel it is the start of a new beginning. I strongly believe in Kaizen philosophy and have tried my best to improve myself every day. The cooperation, experience, and assistance, of my fellow graduate students were essential for completion of my project and I would like to extend thanks to Biological Sciences, Faculty of Graduate Studies and my lab mates for making my research experience a memorable one. It has been an honor for me to be a part of this wonderful team and project.

My final words go equally to me family and I want to thank them for their tremendous love and support. My loving husband **Adishail**, I promise that all your efforts will gain something awe-inspiring in the near future.

## **Dedication**

I would like to dedicate this M.Sc. degree to my loving Husband, Adishail and would like to rephrase the proverb, “Behind every successful woman there is a man”. Without your love and support this would not have been possible.

I am thankful to my parents especially to my sister, Manisha who encouraged me and were there with me in my thick and thin. I would also like to dedicate this degree to my little baby who is growing inside me and by the time I graduate, you will be here in this beautiful world.

Thank you God for making me what I am today and giving me the courage to face all the challenges with an open heart and smile.

I owe this thesis to all of you.

## Table of Contents

Abstract .....	ii
Acknowledgements .....	iii
Dedication .....	v
List of Tables .....	x
List of Figures .....	xiii
Chapter 1: General introduction.....	1
1.1 Lamprey biology .....	7
1.1.1 Paired species concept.....	7
1.1.2. Morphological differences.....	9
1.1.3. Gonadal development in lampreys and other vertebrates.....	10
1.1.3.1 Ovarian differentiation and sexual maturation in lampreys.....	11
1.2. Gene based studies in lampreys and other vertebrates .....	13
1.2.1. Genes of interest involved in gonadal development in lampreys and other vertebrates .....	15
1.2.1.1. Gonadotropin releasing hormone.....	16
1.2.1.2. Gonadotropin .....	17
1.2.1.3. Growth hormone .....	18
1.2.1.4. The insulin superfamily of peptides and receptors .....	19
1.2.1.5. Insulin like growth factors .....	22
1.3. Lampreys as a model species .....	24
1.4. Different methods for gene expression data analysis.....	26
1.5. Goals and objectives.....	28
1.6. Significance of proposed research.....	33
1.7. References .....	34
1.8. Tables and Figures .....	54
Chapter 2: RNA-Seq pipeline for mapping and counting genes expressed during ovarian development in lampreys: a genome-guided approach.....	62
2.1. Abstract .....	62
2.2. Introduction .....	63
2.3. Material and methods .....	68
2.3.1. Sample collection.....	68

2.3.2. RNA sequencing .....	<b>68</b>
2.4. Pipeline designed for transcriptomic data analysis in the presence of a reference genome .....	<b>70</b>
2.4.1. Mapping and counting the number of times a gene is expressed using a reference genome – Galaxy platform.....	<b>70</b>
2.4.1.1. Gene annotation .....	<b>71</b>
2.5. RNA-Seq analysis tools .....	<b>71</b>
2.5.1. Preparation of raw data .....	<b>72</b>
2.5.2. Removal of adaptor sequences.....	<b>73</b>
2.5.3. Format conversion .....	<b>74</b>
2.5.4. Mapping and assembling the reads .....	<b>74</b>
2.5.5. Counting the number of reads aligned to reference genome .....	<b>76</b>
2.5.6. Differential analysis of count data .....	<b>78</b>
2.5.7. Gene Ontology assignment by GOrilla (Gene Ontology enRIchment anaLysis and visuaLizAtion tool) .....	<b>80</b>
2.5.8. REVIGO (reduce + visualize Gene Ontology) .....	<b>82</b>
2.6. Results .....	<b>83</b>
2.6.1. Assessment by FASTQC .....	<b>83</b>
2.6.2. TopHat mapped percentage and coverage graph .....	<b>83</b>
2.6.3. Low number of genes by HTSeq .....	<b>83</b>
2.6.4. Differential gene expression between species .....	<b>84</b>
2.6.5. Differential gene expression within species .....	<b>84</b>
2.6.6. Orthologs of up- and down-regulated genes .....	<b>85</b>
2.6.7. Annotations of up- and down-regulated genes by GOrilla .....	<b>85</b>
2.6.8. REVIGO analysis.....	<b>86</b>
2.6.9. Up- and down-regulated genes with assigned annotations in lampreys .....	<b>86</b>
2.6.10. Normalized counts for different gonadal stages .....	<b>87</b>
2.6.11. Differential analysis of up- and down-regulated genes of lampreys .....	<b>87</b>
2.6.12. Analysis of insulin family genes during different stages of ovarian development	<b>88</b>
2.7. Discussion .....	<b>88</b>
2.8. Implications from this study.....	<b>97</b>
2.9. References .....	<b>98</b>
2.10. Tables and Figures .....	<b>104</b>

Chapter 3: <i>De novo</i> approach for generating reference transcriptome of lampreys for identifying novel and specific genes across different developmental stages .....	146
3.1. Abstract .....	146
3.2. Introduction .....	148
3.3. Goals of <i>de novo</i> assembly .....	149
3.4. Material and methods .....	152
3.4.1. Sample collection.....	152
3.4.2. RNA Sequencing .....	153
3.4.2.1. Evaluation of raw data .....	154
3.4.2.2. Removal of adaptor sequences.....	154
3.4.2.3. Format conversion .....	155
3.4.2.4. Concatenation of forward and reverse reads.....	155
3.4.2.5. Reference-free transcriptome assembly.....	155
3.4.2.6. Statistics of Trinity Assembled files .....	157
3.4.2.7. Transcript estimation by RSEM.....	157
3.4.2.8. Identification by BLAST (Basic Local Alignment Search Tool) .....	158
3.4.2.9. Specific genes of interest .....	159
3.4.2.10. Putative gene name assignment to Trinity ID's of insulin.....	161
3.4.2.11. RSEM count and BLAST .....	162
3.5. Results .....	162
3.5.1. Trinity statistics.....	162
3.5.2. BLASTX for Trinity ID's .....	163
3.5.3 Combined output of RSEM and BLASTX .....	163
3.5.4. Insulin family genes identified.....	163
3.5.5. Genes reported by genome-guided and <i>de novo</i> assembly .....	164
3.5.6. Novel genes of insulin family identified.....	164
3.5.7. Expression of insulin family genes across different gonadal stages.....	165
3.6. Discussion .....	166
3.7. References .....	171
3.8. Tables and Figures .....	176
Chapter 4: General Discussion.....	185
4.1. Background .....	185
4.2. Relevance of results obtained from both the pipelines .....	190

4.3. Findings of the study .....	<b>190</b>
4.4. Technical difficulties and limitations of the study .....	<b>192</b>
4.5. Directions for future research.....	<b>195</b>
4.6. References .....	<b>196</b>
Appendix 1: Acronyms .....	<b>199</b>
Appendix 2: Glossary .....	<b>201</b>

## List of Tables

Table 1.1. Stages of gonadal development in the lamprey life cycle with complete description and characteristics of each stage which represents different samples of parasitic and non-parasitic lampreys used for this thesis. ....	<b>54</b>
Table 1.2. List of genes selected for this thesis that are known to be involved in ovarian differentiation in other species and their Ensembl ID's reported in zebrafish and lampreys.....	<b>55</b>
Table 1.3. Details on ovarian transcriptomes of chestnut lamprey <i>Ichthyomyzon castaneus</i> and northern brook lamprey <i>I. fossor</i> used in this project. "Gonadal stage characteristics" refers to the characteristics of the gonad at different ovarian development stage.....	<b>56</b>
Table 1.4. List of ten sequenced chordate genomes used for making customized BLAST database for genome-guided assembly .....	<b>57</b>
Table 2.1. Detailed summary of sequenced <i>Petromyzon marinus</i> (sea lamprey) genome obtained from Ensembl Genome browser... ..	<b>104</b>
Table 2.2. FASTQC report on individuals of chestnut lamprey <i>Ichthyomyzon castaneus</i> and northern brook lamprey <i>I. fossor</i> used in this project. "No of reads" refers to the count of the total no of sequences processed. "% GC" refers to the overall GC content of all bases in all sequences. "Peak in Phred distribution" refers to the quality scores assigned to each nucleotide base call and is used for assessing the quality of sequences. ... ..	<b>105</b>
Table 2.3. TopHat report on individuals of chestnut lamprey <i>Ichthyomyzon castaneus</i> and northern brook lamprey <i>I. fossor</i> used in this project, individual information about the samples are included in Table 1.3. "Input reads" refers to the total raw reads entered for mapping. "Left hand reads" refers to the reads aligned to the reverse strand. "Right hand reads" refers to the reads aligned to the forward strand. "Aligned reads" refers to the position of a sequencing read corresponding to the reference genome. "Mapped reads" refers to the total count of reads aligned to a reference sequence to generate transcripts. "Unmapped reads" refers to the total count of reads that does not aligns to a reference sequence. ....	<b>106</b>
Table 2.4. HTSeq report on individuals of chestnut lamprey <i>Ichthyomyzon castaneus</i> and northern brook lamprey <i>I. fossor</i> used in this project, individual information about the samples are included in Table 1.3. "No feature" refers to the reads (or read pairs) which could not be assigned to any feature. "Ambiguous" refers to the reads (or read pairs) which contains more than one feature and were not counted. "Alignment not unique" refers to the reads (or read pairs) with more than one alignment. "Total no of HTSeq counts" refers to the total number of reads reported which includes all 'no feature', 'ambiguous', and 'alignment not unique'. "Total sum of reads counts by HTSeq" refers to the reads which contains feature. "% of genes count" refers to the total sum of genes reported by HTSeq/ Total number of HTSeq counts. "Total non-zero counts" refers to the read (or read pairs) whose numeric value is >0. "Gene count >5" refers to the reads (or read pairs) whose numeric value is >5.....	<b>107</b>

Table 2.5. Details on ovarian transcriptomes of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* used in DESeq2 as an input. Samples from both species were pooled together according to gonadal stage to study the global pattern of gene expression during different stages of ovarian development (Sample numbers C13-03, C20-03, IC1, IC2 and IC3 are chestnut lamprey and M01, NC3, N08 and S11 are northern brook lampreys.....**108**

Table 2.6. Number of genes reported by DESeq2 that were up- and down-regulated in parasitic chestnut lamprey and non-parasitic northern brook lamprey at different stages of ovarian development based on *Padj* values of  $\leq 0.05$ . Samples from both species were pooled together based on three different stages of ovarian development: a) early gonadal stage (stages 1-3), b) mid-gonadal stage (stage 4) and, c) late gonadal stage (early and late vitellogenesis). “Total DESeq2” denotes genes with numeric value  $> 0$ ..... **109**

Table 2.7. Top twenty up-regulated genes reported by DESeq2 in parasitic chestnut lamprey and non-parasitic northern brook lamprey with respect to sea lamprey reference genome at different stages of ovarian development based on *Padj* values of  $\leq 0.05$ . Samples were pooled based on three different stages of ovarian development: a) early gonadal stage (stages 1-3), b) mid-gonadal stage (stage 4) and, c) late gonadal stage (early and late vitellogenesis).....**110**

Table 2.8. Top twenty down-regulated genes reported by DESeq2 in parasitic chestnut lamprey and non-parasitic northern brook lamprey with respect to sea lamprey reference genome at different stages of ovarian development based on *Padj* values of  $\leq 0.05$ . Samples were pooled based on three different stages of ovarian development: a) early gonadal stage (stages 1-3), b) mid-gonadal stage (stage 4) and, c) late gonadal stage (early and late vitellogenesis) .....**111**

Table 2.9 a. Statistics of up regulated genes reported by GOrilla in zebrafish *Danio rerio*. “Genes identified” denotes the total number of input genes. “Duplicated genes” refers to genes which are reported more than once in the database. “Unresolved genes” refers to genes with no information in the database. “Total genes for annotation” refers to genes identified for annotation search. “Genes with GO terms” refers to genes identified with annotation terms whereas “Genes with no GO terms” refers to genes with no associated ontology terms in the database..... **112**

Table 2.9 b. Statistics of down-regulated genes reported by GOrilla in zebrafish *Danio rerio*. “Genes identified” denotes the total number of input genes identified in the Gene Ontology Association (GOA) database. “Unresolved genes” refers to genes with no information in the database. “Total genes for annotation” refers to genes identified for annotation search. “Genes with GO terms” refers to genes identified with annotation terms whereas “Genes with no GO terms” refers to genes with no associated ontology terms in the database. All the genes are converted into sea lamprey orthologs. ....**113**

Table 2.10. The list of top twenty Gene Ontology terms for up-regulated and eight Gene Ontology terms for down-regulated genes reported by GOrilla within GO biological processes. ....**114**

Table 3.1. Details on ovarian transcriptomes of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* pooled together for generating an “Assembled transcript” of

both the species by using Trinity. Sample numbers C13-03, C20-03, IC-1, and, IC-2 are chestnut lamprey and MO-1, NC-3, NO-8 and S11 are northern brook lampreys. “Gonadal Stage” refers to the stage of ovarian development in lampreys.....**176**

Table 3.2. Trinity statistics of the “Assembled Transcripts” of *I. castaneus* and *I. fessor* used in this project. Total trinity ‘genes’ is the genes reported by Trinity after assembling all the contigs. Total trinity ‘transcripts’ is the total assembled transcripts. Percent GC refers to the total GC content. ‘N50’ is the length of the contig above which the assembly contains at least half the total number of bases. Total ‘Assembled bases’ is the total sequences assembled from forward and reverse reads.....**177**

Table 3.3. Comparison of insulin family genes identified by both genome-guided versus *de novo* assembly pipeline in *I. castaneus* and *I. fessor*.....**178**

## List of Figures

- Figure 1.1. A phylogenetic tree showing the relationship between different lineages and their divergence times. On the left, the approximate timing of key radiation events is shown and different species are represented, lancelets by *Branchiostoma floridae* (Florida lancelet), lampreys by *Petromyzon marinus* (sea lamprey), sharks by *Callorhinchus milii* (elephant shark), reptiles by *Gallus gallus* (chicken), and mammals by *Homo sapiens* (human).....**58**
- Figure 1.2. Different stages of the lamprey life cycle, which consists of four stages: (1) Larval Stage or Ammocoetes: At this stage, lampreys are blind and filter feeders, (2) Transformers: Larvae that are in the process of metamorphosis are referred to as transformers, (3) Juveniles: Lampreys that have completed metamorphosis but are not yet sexually mature are referred to as juveniles. They can be either parasitic or non-parasitic and, (4) Adults: Lampreys that have reached sexual maturity are known as adults.....**59**
- Figure 1.3. Different stages (stages 1 through 7) of lamprey metamorphosis where B represents branchiopore, F furrow, L lateral lip of oral hood, P pupil and, T transverse lip of oral hood. ....**60**
- Figure 1.4. Different stages of ovarian differentiation in parasitic and non-parasitic lampreys during the lamprey life cycle which represents different samples of parasitic (C13-03, C20-03, IC1,1C-2 and IC3) and non-parasitic lampreys (N02, N1-10, M01, NC3, N0-8 and S11). These stages were interpreted based on observations and literature, and the images were taken from Spice (2013). In this thesis, “Gonadal stage” refers to the histological characteristics of the gonad during the lamprey life cycle (larva and metamorphosis). Ovarian differentiation begins during stages 2 and 3 and the ovary is differentiated at stage 4.....**61**
- Figure 2.1. Phylogenetic relationships between different parasitic and non-parasitic lamprey species derived from cytochrome b sequence data.....**115**
- Figure 2.2. Transcriptomics data analysis pipeline designed for paired-end reads in the presence of a reference genome. ....**116**
- Figure 2.3 a. The summary report of FASTQC for chestnut lamprey C20-3 with detailed information about the total number of sequences processed and length of the shortest and longest sequence in the set. “%GC” denotes the overall GC content of all bases in all sequences .....**117**
- Figure 2.3 b. The per base sequence quality view provides an overview about the range of quality values across all bases at each position in the FastQ file. The x-axis represents the position of each read and y-axis denotes the quality scores. The background of the graph is divided into three different colors based on the value of quality calls. For very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). ....**117**
- Figure 2.3 c. The per tile sequence quality allows to look at deviation from the average quality for each tile. It generates the quality scores from each tile across all of bases and gives an idea about the quality loss (if any) associated with only one part of the flow cell. Cold colors denote

that the quality is at or above the average for that base and hotter colors indicate that a tile had worse qualities than other tiles for that base. A good plot should be blue in color all over .....	<b>118</b>
Figure 2.3 d. The per sequence quality score report about the sequences that have universally low quality values.....	<b>118</b>
Figure 2.3 e. The per base sequence content about the distribution of each base position for which each DNA bases (A, T, G and C) has been called and if the difference between A and T, or G and C is greater than 10% in any position, the module issues a warning .....	<b>119</b>
Figure 2.3 f. This per sequence GC content measures the GC distribution across the whole length of each sequence when compared to a normal distribution of GC content. If the sum of the deviations from the normal distribution is more than 15% of the reads, a warning is issued but if it is more than 30% it indicates a failure.....	<b>119</b>
Figure 2.3 g. The per base N content module about the percentage of base calls at each position for which an N was called. Generally, an N is substituted when a sequencer is unable to make a base call with sufficient confidence. A warning is issued when N content is >5%.....	<b>120</b>
Figure 2.3 h. The sequence length distribution generates a simple graph showing the distribution of fragment sizes over all sequences and gives a single peak for same size of sequence fragment. A warning is issued if sequences are not of the same length. ....	<b>120</b>
Figure 2.3 i. The sequence duplication levels measure the degree of duplication for every sequence and plots sequences with different degrees of duplication. An error is issued if non-unique sequences are more than 50% of the total sequences. ....	<b>121</b>
Figure 2.3 j. The k mer content module estimates the enrichment pattern about read length, generally it counts the enrichment of every 5-mer within the sequence library. It estimates an expected value based on the base content of the library at which this k mer should have been seen and calculates an observed/expected ratio for that k mer. A warning is raised if any k-mer is enriched more than 3-fold overall enrichment or a 5-fold enrichment at any given base. ....	<b>122</b>
Figure 2.4 a. The coverage graph of mapped reads for sample C13-03, chestnut lamprey generated from UCSC Genome Browser (Lamprey Assembly, WUGSC 7.0/petMar2). ....	<b>123</b>
Figure 2.4 b. The coverage graph of mapped reads for sample C13-03, chestnut lamprey generated from Integrated Genome Browser. “The Accepted Hit File” from TopHat was used as an input which is represented by Y-axis and the length of reads is represented by X-axis.....	<b>124</b>
Figure 2.5 a. The coverage graph of mapped reads for sample N17-1, northern brook lamprey generated from UCSC Genome Browser (Lamprey Assembly, WUGSC 7.0/petMar2). ....	<b>125</b>
Figure 2.5 b. The coverage graph of mapped reads for sample N17, northern brook lamprey generated from Integrated Genome Browser. “The Accepted Hit File” from TopHat was used as an input which is represented by Y-axis and the length of reads is represented by X-axis. ....	<b>126</b>
Figure 2.6. Principal Component Analysis (PCA) plot of pooled samples of chestnut lamprey (parasitic) samples (C13-03, C20-03, IC1, IC2 & IC3) and northern brook (non-parasitic)	

samples (N1-10, N17-1, M01, N08 & S11) at three different time points of ovarian development (early stage (1-3), mid-stage (4) and, late stage (early and late vitellogenesis) from DESeq2. The samples were plotted in the 2-dimensional plane spanned by their first two principal components and is useful for visualizing the overall effect of experimental covariates and batch effects. ....**127**

Figure 2.7. Heatmap of the sample-to sample distances of chestnut lamprey (parasitic) samples (C13-O3, C20-03, IC1, IC2 & IC3) and northern brook lamprey (non-parasitic) samples (N1-10, N17-1, M01, N08 & S11) at three different time points of ovarian development (early stage (1-3), mid-stage (4) and, late stage (early and late vitellogenesis) from DESeq2. ....**128**

Figure 2.8. Dispersion estimate of chestnut lamprey (parasitic) samples (C13-O3, C20-03, IC1, IC2 & IC3) and northern brook lamprey (non-parasitic) samples (N1-10, N17-1, M01, N08 & S11) at three different time points of ovarian development (early stage (1-3), mid-stage (4) and, late stage (early and late vitellogenesis). ....**129**

Figure 2.9. Histogram of chestnut lamprey (parasitic) samples (IC1, IC2 & IC3) and northern brook lamprey (non-parasitic) samples (M01, N08 & S11) during mid-stage (4) and, late stage (early and late vitellogenesis) based on  $p$ -values that are distributed more or less uniformly. ..**130**

Figure 2.10. The MA (Mean difference) plot of chestnut lamprey (parasitic) samples (IC1, IC2 & IC3) and northern brook non-parasitic samples (M01, N08 & S11) during mid-stage (4) and, late stage (early and late vitellogenesis) from DESeq2. It estimates the log<sub>2</sub>-fold change over the log-average mean of normalized counts for all the samples in the DESeq2 data set. If the adjusted  $p$ -value is less than 0.1, points are colored red. ....**131**

Figure 2.11. The “Scatterplot & Table” view output of up-regulated genes of *Danio rerio* from the GOrilla output was given as an input to REVIGO. The “Scatterplot” view estimates the terms which are reduced due to redundancy and are showed as a cluster representative in a two-dimensional space by using semantic similarities. In the lower part, the table view “Lists” different processes related to GO terms. “Black denotes” cluster representatives and other cluster members are represented by gray letters. The parameters are easily customizable, “Bubble color” refers to the user provided  $p$ -value (legend in upper right-hand corner); “Size of the bubble” indicates the frequency of the GO term in the Gene Ontology Association (GOA) database.....**132**

Figure 2.12. The “Interactive graph” view of up-regulated genes from REVIGO. “Bubble color” refers to the user provided  $p$ -value (legend in upper right-hand corner); “Size of the bubble” indicates the frequency of the GO term in the database. Edges are used for linking the highly relevant similar GO terms and line width indicates the degree of similarity.....**133**

Figure 2.13. The “TreeMap” view of up-regulated genes from REVIGO where rectangle is used for representing different clusters. Clusters are joined together into ‘superclusters’ of related terms and are color coded. The size of the rectangle can be adjusted based on  $p$ -value or the size (frequency) of the GO terms in the Gene Ontology Association (GOA) database .....**134**

Figure 2.14. The “Tag Cloud” view of up-regulated genes from REVIGO. It displays words which are based on the relatively similar GO terms in the user-supplied list. Large and dark

letters represent stronger overrepresentation. Underrepresented keywords or terms are not displayed in the Tag Cloud .....135

Figure 2.15. GOrilla analysis output of 2,730 down-regulated genes of *Danio rerio* from the DESeq2 were ranked according to their differential expression based on *Padj* value and was given as an input to GOrilla. The resulting GO terms were enriched and visualized using a DAG graphical representation and were assigned different colors based on their degree of enrichment. The nodes in the graph are clickable and provides additional information about the genes and their enrichment terms (GO).....136

Figure 2.16. The “Scatterplot & Table” view output of down-regulated genes of *Danio rerio* from the GOrilla output was given as input to REVIGO. The “Scatterplot” view gives an idea about the terms which are reduced due to redundancy and are showed as a cluster representative in a two-dimensional space by using semantic similarities. In the lower part, the table view “Lists” different processes related to GO terms. “Black denotes” cluster representatives and other cluster members are represented by gray letters. The parameters are easily customizable, “Bubble color” refers to the user provided *p*-value (legend in upper right-hand corner); “Size of the bubble” indicates the frequency of the GO term in the Gene Ontology Association (GOA) database.....137

Figure 2.17. The “Interactive graph” view of down-regulated genes from REVIGO. “Bubble color” refers to the user provided *p*-value (legend in upper right-hand corner); “Size of the bubble” indicates the frequency of the GO term in the database. Edges are used for linking the highly relevant similar GO terms and line width indicates the degree of similarity.....138

Figure 2.18. The “TreeMap” view of down-regulated genes from REVIGO where rectangle is used for representing different clusters. Clusters are joined together into ‘superclusters’ of related terms and are color coded. The size of the rectangle can be adjusted based on *p*-value or the size (frequency) of the GO terms in the Gene Ontology Association (GOA) database.....139

Figure 2.19. The “Tag Cloud” view of down-regulated genes from REVIGO. It displays words which are based on the relatively similar GO terms in the user-supplied list. Large and dark letters represent stronger overrepresentation. Underrepresented keywords or terms are not displayed in the Tag Cloud. ....140

Figure 2.20. Graphs of genes like *oxytocin receptor*, *zona pellucida glycoprotein 3a*, *neuropeptide Y receptor Y8b* and *Wilms tumor 1 associated protein* reported by genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts. All these genes (a, b, c and d) showed an increased expression in gonadal stages 2, 3 & 4 in northern brook lamprey... ..141

Figure 2.21. Graphs of genes like *thyroid stimulating hormone receptor*, *adrenoceptor alpha* and *estrogen receptor 2*, reported by genome-guided assembly, where the x-axis represented the

gonadal stages and y-axis represented the HTSeq gene counts. All these genes (a, b and c) show an increased expression in gonadal stage 4 in parasitic chestnut lamprey. ....	<b>142</b>
Figure 2.22. Graphs of genes like <i>calcitonin receptor</i> and <i>forkhead box l2</i> , <i>SRY (sex determining region Y)-box 2</i> , reported by genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts. Both are expressed in gonadal stages 2, 3 and 4 in parasitic chestnut lamprey.....	<b>143</b>
Figure 2.23. Graphs of insulin family genes ( <i>RXFP1</i> , <i>RXFP3</i> , <i>INSR-like</i> and <i>Igf1R-L3</i> ) reported by genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts. All these genes (a, b, c, d and, e) show an increased expression in gonadal stages 2, 3 & 4 in northern brook lamprey .....	<b>144</b>
Figure 2.24. Graphs of insulin family genes ( <i>igf2</i> , <i>Igf1R-L1</i> or <i>Insrr-L1</i> and <i>Ins-LI</i> ) reported by genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts. In (a), <i>igf2</i> gene was expressed in parasitic chestnut lamprey during gonadal stages 2 and 3 but in gonadal stage 4 there was a sharp decrease in gene expression in parasitic lamprey, in (b), <i>Igf1R</i> or <i>Insrr-L1</i> was highly expressed in non-parasitic northern brook lamprey during gonadal stages 2, 3 and 4 as compared to the parasitic lamprey and, in (c) <i>Ins-LI</i> was only expressed in parasitic lamprey during gonadal stage 3.....	<b>145</b>
Figure 3.1. Transcriptomics data analysis pipeline also known as <i>de novo</i> assembly for paired-end reads in the absence of a reference genome by using customized BLAST database.....	<b>179</b>
Figure 3.2. Plots of insulin family genes ( <i>igf1</i> , <i>igf2</i> and <i>igf1r</i> or <i>insrr</i> ) reported in chestnut lamprey (parasitic) samples, where the x-axis represented the gonadal stages and y-axis represented the RSEM count. Genes, <i>igf1</i> and <i>igf1r</i> or <i>insrr</i> are reported as novel genes and are not annotated in the sea lamprey genome. ....	<b>180</b>
Figure 3.3. Plots of insulin family genes ( <i>igf1</i> , <i>igf2</i> , <i>insrr</i> , <i>relaxin-like</i> , <i>RXFP1</i> and <i>RXFP-2 like</i> ) expressed in northern brook lamprey (non-parasitic) samples, where the x-axis represented the gonadal stages and y-axis represented the RSEM count. Genes <i>igf1</i> , <i>insrr</i> , <i>relaxin-like</i> and <i>RXFP2-like</i> are reported as novel genes and are not annotated in sea lamprey.....	<b>181</b>
Figure 3.4. Plots of insulin family genes ( <i>igf1r</i> or <i>insrr</i> ) reported in northern brook lamprey (non-parasitic) samples, where the x-axis represented the gonadal stages and y-axis represented the RSEM count.....	<b>182</b>
Figure 3.5. Plots of insulin family genes ( <i>igf1</i> , <i>igf2</i> and <i>igf1r</i> ) reported in both species of lamprey, chestnut lamprey and northern brook lamprey, where the x-axis represented the gonadal stages and y-axis represented the RSEM count. Gene, <i>igf1</i> is reported as novel genes and is not annotated in sea lamprey.....	<b>183</b>
Figure 3.6. Plots of relaxin family peptides receptors – <i>RXFP1</i> and <i>RXFP-2 like</i> novel reported only northern brook lamprey (non-parasitic) samples, where the x-axis represented the gonadal stages and y-axis represented the RSEM count. <i>RXFP-2 like</i> is reported as a novel gene (i.e., not annotated in the sea lamprey genome) and is expressed in all stages of ovarian development... <b>184</b>	

## Chapter 1

### General introduction

Lampreys are primitive jawless fishes (agnathans) of the class Petromyzontida. Agnathans (lampreys and hagfishes; class Myxini) diverged from the rest of the vertebrate lineage (i.e., the lineage that gave rise to the gnathostomes) approximately 550 million years ago (Figure 1.1), prior to the evolution of jaws (Docker *et al.*, 2015). Although originally a topic of debate, it is now thought that lampreys diverged after the two rounds (2R) of whole genome duplication (WGD) that occurred in early vertebrate evolution (Smith *et al.*, 2013) although a recent study by Smith and Keinath (2015) argued that the duplication pattern of genes in lamprey is more compatible with the hypothesis that the lineage has undergone a single round (1R) of WGD followed by several evolutionarily independent segmental duplications rather than a second WGD *per se*. This second hypothesis requires further sequencing and bioinformatic analyses before it can be adequately resolved, but in either case, agnathans (lampreys and hagfishes) continue to be important models of the early vertebrate genome. There are an estimated 41- 44 extant species of lampreys; 18 of them are classified as parasitic and 23-26 as non-parasitic (Potter *et al.*, 2015).

Completion of the sequencing of the sea lamprey *Petromyzon marinus* genome in 2013 was an important milestone in comparative vertebrate evolutionary genomics (Smith *et al.*, 2013) and analyses of the sea lamprey genome has already provided important insights into early vertebrate evolution (Smith *et al.*, 2013). However, identifying **orthologous** (note that terms that are bolded on first use within each chapter are defined in the glossary in Appendix 2) genes between lampreys and other vertebrate taxa has been quite difficult, in part, because it appears

that lampreys have had significant lineage-specific changes, but also because the annotation of the lamprey genome is still far from complete and complex patterns of gene loss and retention further complicate determining orthologous versus **paralogous** relationships of lamprey genes to those in later diverging vertebrates (Smith *et al.*, 2013). Early vertebrate evolution was marked by rapid changes in anatomy and physiology, changes that were presumably accelerated by the greatly expanded genomic toolkit afforded by 2R. Lampreys are one of the earliest diverging taxa following 2R and thus represent one of the first evolutionary experiments in how to organize the new vertebrate body plan, which included greatly expanded roles for the nervous, immune and endocrine systems (Docker *et al.*, 2015). In some cases, lampreys appear to represent the “primitive” state in vertebrate evolution, while in other cases (such as the adaptive immune system) they evolved their own unique system (Docker *et al.*, 2015). In either case, lampreys are an excellent model for understanding many aspects of early vertebrate evolution because of their position at the base of vertebrate phylogeny (Docker *et al.*, 2015) and are a great source to provide deep insights about their divergence pattern from other vertebrates after whole genome duplication (WGD) ~550 Mya (Smith & Keinath, 2015).

The overarching goal of this research was to use whole-genome RNA-Seq transcriptomic data from the ovarian samples of parasitic chestnut lamprey *Ichthyomyzon castaneus* and non-parasitic northern brook lamprey *I. fossor* to identify suites of genes expressed during ovarian development. Since these species belong to the same genus and sample sizes per ovarian stage were relatively small, samples were pooled together for finding the genes expressed at different stages of ovarian development. For this, the genome of sea lamprey was used as a **reference** to align and map the reads because the genus *Ichthyomyzon* is closely related to the sea lamprey (Figure 2.1). Hence, 12 ovarian transcriptomes collected during different stages of the lamprey

life cycle (larval stage and metamorphosis) available from previous studies in the Docker laboratory (e.g., Spice, 2013, Table 1.3) were analyzed in an effort to understand changes in gene expression over time and relate these changes to developmental processes. The primary hypothesis was that differences in the timing of oocyte development and in potential fecundity in parasitic versus non-parasitic lampreys (below) must be associated with differences in the timing of gene expression. Histologically the process of ovarian differentiation is well studied in lampreys (Section 1.1.3.1), but the genes involved are poorly understood (Spice *et al.*, 2014; McCauley *et al.*, 2015). Thus, a major goal of this thesis was to identify genes that were up-regulated and down-regulated in the ovary during different developmental stages in lampreys in the hopes that this will contribute to our understanding of the genetic control of ovarian development in lampreys.

Additionally, since RNA sequence analysis is less well developed for non-model organisms, a second goal of the thesis was to develop an RNA sequence analysis pipeline that is suitable for non-model taxa that facilitates identification of both **novel** (i.e., those not currently identified in the sea lamprey genome) and unannotated genes expressed during ovarian development. For this, two different pipelines were designed which can be used in the presence and absence of a reference genome, which was useful in this study because some of the genes in the sea lamprey reference genome are still uncharacterized. The insulin superfamily genes were used to test the strengths and weaknesses of the pipelines designed because of a) *a priori* reasons that several members of the insulin superfamily of genes play roles in ovarian development, and b) the evolutionary history of the family has been sufficiently studied in deuterostomes to suggest that the lamprey genome contains these genes. Since some of these genes were annotated while others were unannotated but expected to be expressed in reproductive tissues, the two

pipelines (i.e., with and without a reference genome and annotation) can be compared and examined for the identification of family members (Table 1.2). For this, the genomic location and sequence information for the insulin superfamily family - relaxin family peptides and their receptors (Yegorov & Good, 2012; Yegorov *et al.*, 2014) and insulin and insulin-like growth factor (INS/IGF) and their RTKs (receptor tyrosine kinases) in lampreys (unpublished data) were obtained from Dr. Sara Good's laboratory at the University of Winnipeg. All these factors made these genes ideal for study.

To address these research objectives, I have developed two pipelines (Chapters 2 & 3). One of them employed the use of the sea lamprey reference genome, and the second one was used to create a *de novo* assembly of ovarian messenger RNA (mRNA) transcripts sampled in each species/time point followed by putative identification of these transcripts using a custom-made BLAST (Altschul *et al.*, 1990) database. For the genome-guided approach, the RNA-Seq reads from *I. fossor* and *I. castaneus* were mapped to the sea lamprey genome and those reads that aligned to an annotated region were counted using the program HTSeq (Anders *et al.*, 2015). Subsequently, the raw counts were normalized (with respect to the total number of mapped reads) and samples from either species in the same developmental stage were pooled and a test of differential expression of genes across stages was employed using the program DESeq2 (Love *et al.*, 2014). The list of genes generated from DESeq2 were then summarized to identify the functions of the differentially expressed (both up- and down-regulated) genes. All the differentially expressed genes of sea lamprey (**Ensembl ID's**) were converted into their putative orthologs in zebrafish *Danio rerio* since many of the genes annotated in the lamprey genome were not characterized or linked to gene functions; zebrafish are one of the oldest teleost fishes for which a well-annotated genome exists. Using this set of orthologous ID's from zebrafish,

**putative gene** ontology terms were assigned using the program GOrilla (Eden *et al.*, 2009) to the genes found to be up- and down-regulated in chestnut and northern brook lamprey samples across ovarian gonadal stages.

The goal of this pipeline was to identify the suite of annotated genes which were up- and down-regulated in chestnut and northern brook lamprey ovaries across developmental stages. Although the pipeline was commonly used for RNA-Seq studies, it has the limitation of informing about changes in expression of genes that successfully mapped to the sea lamprey genome and were also annotated on the lamprey genome. At present, there are only 13,114 genes annotated on the lamprey genome, a number that probably represents only about one-half of the genes present in the genome (based on the size of most vertebrate genomes being between 20 and 27k genes). Moreover, annotation of genomes is typically performed with a mixture of both *ab initio* gene prediction programs and confirmation of expected genes using mRNA libraries (often obtained from RNA-sequencing). In the case of lamprey, there has been limited *ab initio* gene prediction performed on the genome and an ovarian transcriptome was not used to annotate the genome. Thus, it is likely that there were hundreds of transcripts present in our samples which were not annotated.

To overcome this problem, a second pipeline was employed to perform a *de novo* assembly, in which raw reads of both parasitic and non-parasitic lampreys was assembled to generate a *de novo* transcript assemblies using the program Trinity and these assembled transcripts were used to calculate normalized counts of these reads using the program RSEM (RNA-Seq by Expectation Maximization) for all samples across developmental stages. The program RSEM provided normalized counts of *de novo* assembled transcripts, while the counts provided in the first pipeline (HTSeq) were estimated counts of genes (including multiple splice

variants) annotated in the lamprey genome, and these counts were then normalized and compared using additional approaches. Hence, the transcripts identified through *de novo* approach were used as an input for BLASTX to identify the gene names associated with these unique **Trinity ID's** (these ID's are assigned to the assembled transcripts by the program Trinity). For this purpose, a customized BLAST database comprising sequences of 10 annotated chordate species was used for querying the assembled transcript file.

Although the *de novo* assembly has the ability to identify new, unannotated genes, in the lamprey genome, it was challenging to assemble and identify the *de novo* reads obtained from the *de novo* assembly. In part this is because the reads contain 5' and 3' untranslated regions (UTRs) which are often longer than the genes themselves and which aligned to non-coding segments of the genome. Hence, to assign tentative IDs to the Trinity transcripts, the homology between these transcripts with the genes contained in a custom-made database consisting of all annotated genes in 10 chordate genomes was compared. This allowed me to create a database of tentative ID's to all the *de novo* assembled transcripts and search for specific genes of interest hypothesized to be in the lamprey genome but yet not annotated.

This thesis aims to identify new genes which are not present in the sea lamprey database and update the existing knowledge about different genes expressed in the lamprey ovary at different developmental stages (Section 1.2.1). In this chapter, I provide a brief review of general lamprey biology (Section 1.1), with a focus on the differences in ovarian development observed in parasitic versus non-parasitic species (Section 1.1.3.1). During larval stage, parasitic and non-parasitic lampreys are generally not distinguishable especially in closely related "paired species" (Section 1.1.1) following metamorphosis, non-parasitic lampreys undergo rapid sexual maturation relative to parasitic species (Section 1.1.2). I also review RNA-Seq based studies that

have identified genes that play an important role in ovarian differentiation in lampreys and other vertebrates (Section 1.2). This thesis builds on this research and has identified different genes that are up- and down-regulated in lampreys during different stages of ovarian development by using RNA-Seq data. Some of these genes have not yet been identified in the sea lamprey and this information will be useful for future investigations.

## **1.1. Lamprey biology**

All lampreys spend a prolonged larval stage as filter feeders, as blind “ammocoetes” living in the bottom of streams and rivers of fresh water for several years (Dawson *et al.*, 2015) but, following metamorphosis, lampreys can exhibit one of two adult life histories (Docker, 2009). They are either parasitic (at sea or in fresh water), when they may feed on commercially-important fishes (e.g., sea lamprey in the Great Lakes), or they are non-parasitic and do not feed following metamorphosis (e.g., northern brook lamprey). The non-parasitic “brook” lampreys remain in their natal streams and undergo rapid sexual maturation after metamorphosis, spawning and dying within 6–10 months (Docker, 2009). On the other hand, after one or more years of living parasitically, the parasitic species leave their hosts and then also undergo sexual maturation, spawn and die (Potter *et al.*, 2015).

### **1.1.1. Paired species concept**

In seven of the ten lamprey genera, there are closely related parasitic and non-parasitic lampreys that are generally indistinguishable as larvae and are called “paired species” (e.g., European brook and river lampreys, *Lampetra planeri* and *L. fluviatilis*, respectively). The larvae of these paired species are often morphologically indistinguishable and diverge only during

metamorphosis (Docker, 2009). During completion of metamorphosis, the gonad is well developed and maturation is rapid in non-parasitic lampreys whereas in parasitic lampreys the gonad remains immature during the feeding phase (Docker, 2009). Hence, following metamorphosis, non-parasitic species undergo sexual maturation within several months while the process may take several years in parasitic species (Hardisty & Potter, 1971; Hardisty, 2006; Docker, 2009).

Silver lamprey *Ichthyomyzon unicuspis* and northern brook lamprey, which are parasitic and non-parasitic lampreys, respectively, found in Manitoba, the Great Lakes, and elsewhere in northeastern North America, are one such example of paired species. Not all parasitic and non-parasitic lampreys form species pairs, but paired species are phylogenetically closely related, and it is usually assumed that non-parasitic species are derived from their parasitic counterparts (Zanandrea, 1959; Vladykov & Kott, 1979; Docker, 2009). It is believed that non-parasitic lampreys have evolved through delayed sexual maturation relative to metamorphosis (Docker, 2009). Non-parasitic lampreys, especially females, are larger at the time of metamorphosis than females in parasitic taxa, which may enable them to undergo sexual maturation without benefit of the parasitic feeding phase (Docker, 2009). In general, a term known as heterochrony is used for defining this change in the timing of the life cycle which can occur due to an environmental disturbance, experimental induction or mutation in the genes which results in new traits (Manzon *et al.*, 2015). Some authors have suggested that life history type can be determined by some important factors such as body condition (in lampreys, Pavlov *et al.*, 2007) or size at maturity (e.g., in salmonids; Garant *et al.*, 2003). Relatively little is known about the divergent history types between parasitic and non-parasitic ecotypes (Docker, 2009). Several studies have found little genetic differentiation between lamprey paired species (Schreiber & Engelhorn 1998;

Docker *et al.*, 1999; Espanhol *et al.*, 2007; Blank *et al.*, 2008; Pereira *et al.*, 2010; Bracken *et al.*, 2015; Rougemont *et al.*, 2015); however, some studies have identified genetic divergence between paired species. Mateus *et al.* (2013) found evidence of strong genome-wide divergence between European river *Lampetra fluviatilis* and brook lampreys *L. planeri* sampled sympatrically in a river in Portugal and concluded that the two species are taxonomically divergent but maintain low levels of ongoing gene flow. They used restriction site associated DNA sequencing (RAD-Seq) and identified 166 fixed genetic differences between the species (Mateus *et al.*, 2013). Recently, Rougemont *et al.* (2016) also used RAD-Seq and similarly refuted the notion that feeding type is plastic (i.e., not genetically determined) and also showed ongoing gene flow in this species pair.

### **1.1.2. Morphological differences**

Morphologically, it is very difficult to distinguish the larvae of parasitic and non-parasitic lampreys (Hardisty, 2006), but following metamorphosis, parasitic species have a functional open gut, well developed oral cavity, and needle like, keratinized and more frequently replaced teeth while, on the other hand, the non-parasitic species have a weakly developed alimentary canal, and blunt teeth that are replaced once or not at all (Hardisty & Potter, 1971; Vladykov & Kott, 1979; Youson, 1980; Hardisty, 2006). Parasitic adults are larger (often much larger) than non-parasitic adults (Docker, 2009). For example, parasitic Pacific lamprey *Entosphenus tridentatus* reaches a maximum adult size of 850 mm (Orlov *et al.*, 2008), whereas non-parasitic Pit-Klamath brook lamprey *E. lethophagus* has a maximum adult size of 116 to 142 mm (Hubbs, 1971).

Metamorphosis usually lasts for three to four months and typically begins in the summer (Manzon *et al.*, 2015). Several endogenous and exogenous (environmental) factors such as larval size and condition, the accumulation of sufficient lipid reserves for the non-trophic metamorphic phase, thyroid hormones, and a rise in water temperature influence the timing of metamorphosis (Manzon *et al.*, 2015). At metamorphosis, certain morphological changes occur in the a) appearance, shape and size of the eye; b) buccal cavity; c) growth and differentiation of the fins; d) body coloration; and e) shape of branchiopores and branchial region. These five broad changes are used to distinguish lampreys at each of seven different stages of metamorphosis (Manzon *et al.*, 2015; Figure 1.3).

### **1.1.3. Gonadal development in lampreys and other vertebrates**

Generally, three important events in gonadal development are: sex determination, sex differentiation, and sexual maturation. a) **Sex determination** refers to the potential of gonads to develop into an ovary or a testis (Sandra & Norma 2010); b) **sex differentiation** is the process by which the phenotypic sex of an individual is produced (Piferrer & Guiguen, 2008) and can be further divided into ovarian differentiation and testicular differentiation; and c) **sexual maturation** is the final stage where changes occur in the gonad (ovary or testes) prior to reproduction. The timing of sex differentiation in lampreys varies from species to species.

Sex differentiation and oocyte development occur earlier in non-parasitic species than in parasitic species and, given that oocytes are elaborated at a smaller body size, non-parasitic lampreys may be less able to produce large numbers of oocytes relative to parasitic lampreys (Hardisty, 1960). Spice & Docker (2014) found that during ovarian differentiation when the ovaries are developing, non-parasitic lampreys had higher expression of cytochrome c oxidase

subunit III during cystic and growth stages as compared to chestnut lampreys. They particularly used *coIII* as a marker to study energy consumption and suggested that chestnut lamprey requires higher energy consumption to produce more oocytes during ovarian differentiation and this gene helps in inhibiting apoptosis. The expression of insulin-like growth factor 1 receptor (*igf1r*) was higher during oocyte growth stages in chestnut lamprey relative to northern brook lamprey. However, Beamish & Thomas (1983) suggested that differences in genetic and/or genetic and environmental factors influence sexual maturation and ovarian differentiation in parasitic and non-parasitic taxa.

#### **1.1.3.1. Ovarian differentiation and sexual maturation in lampreys**

In lampreys, during the larval phase when the gonads are undifferentiated, ovarian differentiation occurs in the spring of the first, second, or third year following hatch (Hardisty, 1969, 1970). Non-parasitic lampreys generally undergo ovarian differentiation in the first year following hatch, while parasitic lampreys generally undergo ovarian differentiation in the second year (Hardisty, 1970); the large anadromous sea lamprey delays ovarian differentiation to the third year following hatching (Hardisty 1969; Barker & Beamish 2000); larger body size at ovarian differentiation is generally correlated with higher adult fecundity (Vladykov, 1951). Ovarian and testicular differentiation are separated by several years in lampreys. Testicular differentiation is delayed for several years and does not occur until metamorphosis (Hardisty, 1971), although in many teleost species, testicular differentiation occurs slightly later than ovarian differentiation (Patino & Takashima 1995; Meijide *et al.*, 2005; Sandra & Norma 2010).

At metamorphosis, differences in life history type (i.e., parasitic versus non-parasitic) become evident externally. After metamorphosis, non-parasitic lampreys stop feeding and

undergo sexual maturation, reproduce and then die (Docker, 2009). Several changes occur in the gonads during this process and these changes have been categorized into six different gonadal stages (Figure 1.4, Table 1.1). During gastrulation, primordial germ cells first appear and divide mitotically in the gonad prior to differentiation to form protogonia or deuteroogonia. These cysts of germ cells are surrounded by thin fibrous tissue and follicle cells (Lewis & McMillan 1965; Hardisty, 1971). In Far Eastern brook lamprey *Lethenteron reissneri*, germ cell cysts are not reported frequently which suggests that in this species, oocytes develop directly from protogonia (Fukayama & Takahashi, 1983). However, in some other species, these cysts of germ cells enter meiotic prophase and remain in the meiotic phase throughout larval development (Hardisty, 1971). Cysts of germ cells enlarge and become basophilic after entering the cytoplasmic growth phase and are gradually broken up by follicle cells (Lewis & McMillan, 1965). Ovarian differentiation is considered to occur during stages 2 and 3 and the differentiated ovary occurs in stage 4 (Hardisty, 1965). Reduced fecundity in non-parasitic species may also be achieved through oocyte degeneration or atresia at metamorphosis (Hardisty, 1961; Hughes & Potter 1969; Beamish & Thomas 1983). Beamish & Thomas (1983) found that body size alone may not be the only factor affecting oocyte production. Non-parasitic species could be of the same size as parasitic species at ovarian differentiation and still produce fewer oocytes through some other mechanism. They suggested that fecundity could be reduced in southern brook lamprey *Ichthyomyzon gagei* relative to the paired chestnut lamprey through oocyte atresia at metamorphosis (Beamish & Thomas, 1983), as has been documented in another lamprey species (e.g., European brook lamprey and Far Eastern brook lamprey; Hardisty 1971; Fukayama & Takahashi, 1983).

The ovaries are comparatively stable between ovarian differentiation and metamorphosis, although some cytoplasmic growth and/or atresia may occur (Hardisty, 1971). Generally, ovarian differentiation in lampreys can be considered complete when the oocyte stops growing (Hardisty, 1971) and final oocyte growth and vitellogenesis in lampreys occur after metamorphosis (Hardisty, 1971). During early stages of metamorphosis, oocyte diameter varies from 43.1  $\mu\text{m}$  in pouched lamprey *Geotria australis* (Hughes & Potter, 1969) to 200  $\mu\text{m}$  in southern brook lamprey (Beamish & Thomas, 1983).

Non-parasitic species undergo vitellogenesis rapidly after metamorphosis; however, in parasitic species, vitellogenesis proceeds much more slowly (Hardisty, 1971; Hardisty & Potter 1971). Hence, sexual maturation occurs early in non-parasitic taxa compared to the parasitic taxa (Figure 1.2); this suggests that there may be different genes that are up-regulated and down-regulated at different time points in parasitic chestnut lamprey versus non-parasitic northern brook lamprey during and after metamorphosis which may help to understand different trajectories of the lamprey life cycle and shed light on the genes involved in different stages of ovarian development such as ovarian differentiation, vitellogenesis, oocyte maturation and sexual maturation (Table 1.1, Figure 1.4).

## **1.2. Gene based studies in lampreys and other vertebrates**

Studies of sex differentiation in other fish have used quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR) to look at expression of individual candidate genes (e.g., Baron *et al.*, 2005; Jorgensen *et al.*, 2008; Berbejillo *et al.*, 2012). Recently, Spice *et al.* (2014) used qRT-PCR to study gene expression in parasitic chestnut lamprey and non-parasitic northern brook lamprey during ovarian differentiation. They targeted eight genes (17 $\beta$ -hydroxysteroid

dehydrogenase, germ cell-less, estrogen receptor b, insulin-like growth factor 1 receptor, daz-associated protein 1, cytochrome c oxidase subunit III, Wilms' tumour suppressor protein 1, and dehydrocholesterol reductase 7) to assess if the genes were differentially expressed before, during, and/or after ovarian differentiation in chestnut and northern brook lampreys. This was the first study in lampreys to identify genes that were involved in ovarian differentiation and shed light on differences in development and gene expression in parasitic and non-parasitic lampreys.

For example, Spice *et al.* (2014) found that insulin-like growth factor 1 receptor (*igf1r*) was highly expressed in chestnut lamprey during the oocyte stage, which is consistent with its known role in promoting growth of many cells including the ovary via its role in inhibiting apoptosis (Reinecke, 2010). On the other hand, Spice *et al.* (2014) found higher expression of cytochrome c oxidase subunit III (*coIII*) in northern brook lamprey during the early presumptive male, cystic, and oocyte growth stages suggesting that it may act as an inducer of apoptosis. Wilms' tumor suppressor protein 1 (*WT1*), which plays an important role in testicular development, was highly expressed in the presumptive male stage and was found to correlate with whether gonads differentiated into testis or ovaries (Spice *et al.*, 2014), which is consistent with its role promoting testicular development (Nachtigal *et al.*, 1998) in mammals and other fish species (Scharnhorst *et al.*, 2001; Sandra & Norma, 2010). Genes like 17 $\beta$ -hydroxysteroid dehydrogenase (*hsd17  $\beta$* ) and daz-associated protein-1 (*dazap1*) were expressed in late stages of ovarian differentiation and in differentiated females (Spice *et al.*, 2014), which suggests that these genes may play an important role in ovarian development in lampreys and help maintain differentiated oocytes.

Previous studies on diverse fish species have identified genes that have conserved roles in sexual differentiation in teleost fishes (Piferrer *et al.*, 2012). Teleosts diverged ~350 Mya from

Semionotiformes (gars) and Amiiformes (bowfin) (Hoegg *et al.*, 2004; Amores *et al.*, 2011) which collectively diverged from the osteichthyan ancestor ~450 Mya (Postlethwait *et al.*, 2000; Jaillon *et al.*, 2004; Naruse *et al.*, 2004; Woods *et al.*, 2005; Kohn *et al.*, 2006). Thus, although the suite and function of genes in teleosts can serve as a guide for the putative identification of genes in lampreys, given that lampreys diverged from the common vertebrate ancestor ~550 Mya, the process of sexual differentiation may be quite different between teleost fish and lower vertebrates such as lampreys (reviewed in Piferrer *et al.*, 2012). Given this, in this thesis, I take the following approaches to find genes influencing ovarian development in lampreys a) on the one hand, I will look for evidence of genes known to be involved in ovarian differentiation in lampreys using a comparative approach; and b) I will look for evidence of any (*de novo* or annotated) genes exhibiting high or highly variable expression during different stages of lamprey ovarian development to assess the novelty and identity of such genes.

### **1.2.1. Genes of interest involved in gonadal development in lampreys and other vertebrates**

Lamprey reproduction is controlled by the hypothalamic-pituitary (HP) system that emerged prior to or during the evolution of vertebrates (Sower, 2015) due to the two rounds (2R) of whole genome duplication (WGD). Reproduction in vertebrates is controlled by a hierarchically organized endocrine system. All fishes initiate reproduction after reaching a certain age and size (Taranger *et al.*, 2010) and need energy reservoirs for reproduction (Thorpe *et al.*, 1990; Campbell *et al.*, 2006; Shearer *et al.*, 2006). The growth and energetic metabolism of fish are mediated by growth-related hormones (Reindl & Sheridan, 2012). Fasting or reduced feeding decreases growth and energy storage leading to reduced fecundity and incidence of sexual maturation (Bromage *et al.*, 1992; Karlsten *et al.*, 1995; Silverstein *et al.*, 1998). In

vertebrates, growth and reproduction are two major physiological processes that are closely related and controlled by the somatotrophic axis (Zakes & Demska-Zakes, 1996; Gomez *et al.*, 1999; Papadaki *et al.*, 2005) consisting of growth hormone (GH) also called co-gonadotropin (Hull and Harvey, 2002) and insulin-like growth factors (IGF-I and IGF-II). In fishes, ovarian differentiation generally increases steroid production, as ovarian somatic cells synthesize estrogens critical for oogonial proliferation and differentiation into oocytes (Suzuki *et al.*, 2004; Guiguen *et al.*, 2010; Lubzens *et al.*, 2010). Some genes that have been shown to play an important role in vertebrate gonadal development, particularly ovarian development (in one or more species), are as follows: *insulin family genes*, *gonadotropin releasing hormones*, *growth hormones*.

#### **1.2.1.1. Gonadotropin-releasing hormone (GnRH)**

Gonadotropin-releasing hormone is a decapeptide released from the hypothalamus of the brain that controls reproductive processes via the pituitary-gonadal axis (Sower, 2003). GnRHs are highly conserved in vertebrates. Growing evidence reveals almost all vertebrates synthesize at least two isoforms of GnRH in the brain (Dubois *et al.*, 2002; Silver *et al.*, 2004; Guilgur *et al.*, 2006; Kah *et al.*, 2007; Kavanaugh *et al.*, 2008; Okubo & Nagahama, 2008). GnRH-II is generally found in the hypothalamus and peripheral tissues (Gorbman & Sower, 2003; Kah *et al.*, 2007). It stimulates gonadotropin function by controlling hormones secreted from pituitary (Trinh *et al.*, 2007). In the sea lamprey (family Petromyzontidae), two forms of GnRH known as lamprey GnRH-I and lamprey GnRH-III have been sequenced (Sherwood *et al.*, 1986; Sower *et al.*, 1993). Both GnRH-I and -III stimulate steroidogenesis and ovulation in adult sea lamprey. In the two Southern Hemisphere families of lampreys, Geotriidae and Mordaciidae, gonadotropin-

releasing hormone GnRH-like molecules have also been identified in the brain (Sower *et al.*, 2000).

Lampreys are among the most ancient of vertebrates to show functional roles for multiple forms of gonadotropin-releasing hormone (Sower, 2003). Gnathostomes generally have one or two GnRHs that act as hypothalamic hormones, two pituitary gonadotropins (GTHs) - luteinizing hormone (LH or lutropin) and follicle stimulating hormone (FSH or follitropin), and one gonadal FSH receptor and one LH receptor (i.e., one receptor for each of these gonadotropins). However, although two GTHs (LH and FSH) were known from all taxonomic groups of gnathostomes by the 1990s (Suzuki *et al.*, 1988; Kawauchi *et al.*, 1989; Quérat *et al.*, 2000, 2004), only one GTH has been identified in the agnathans, in both lampreys (Sower *et al.*, 2006) and hagfishes (Uchida *et al.*, 2010, 2013). Thus, lampreys have three hypothalamic GnRHs (GnRH-I, GnRH-II, and GnRH-III), three GnRH receptors (GnRH-R-1, -2 and -3) involved in regulating reproduction (Sower *et al.*, 2009; Joseph *et al.*, 2012), but only one pituitary gonadotropin-type hormone (GTH), one gonadal glycoprotein hormone (GpH) receptor, and one thyroidal GpH receptor (Sower, 2015).

#### **1.2.1.2. Gonadotropin**

In response to GnRH, a major class of glycoprotein hormones known as gonadotropins (GTHs) are released from the pituitary gland (Ogiwara *et al.*, 2013) and influence steroidogenesis and gametogenesis in almost all vertebrates, including fish (Shi *et al.*, 2015). The gonadotropins, follicle-stimulating hormone (FSH) and luteinizing hormone (LH), play a vital role in gnathostome reproductive function (Evans *et al.*, 1922; Greep *et al.*, 1941). In jawed vertebrates, both FSH and LH consist of a common subunit (GpA1) and unique subunits (Gp-1, -

2, and -3) (Sower, 2015), while no such molecule has been identified in lampreys and hagfishes. As discussed above (Section 1.2.1.1), only one GTH has been identified in the jawless vertebrates (Sower, 2015).

The two gonadotropins in teleost fishes, initially termed GTH-I and GTH-II, are now referred to as FSH and LH (Li & Ford, 1998; Querat *et al.*, 2000). In fish, FSH is involved in the control of puberty and gametogenesis whereas LH plays an important role at final gonadal maturation, ovulation, or spermiation (Swanson *et al.*, 1991; Kazeto *et al.*, 2012; Shi *et al.*, 2015). Lampreys, however, do not have LH or FSH, but they do have gonadotropin- $\beta$ , a precursor that appears to have undergone gene duplication to diverge into LH and FSH prior to the divergence of cartilaginous fishes (Sower *et al.*, 2006; Roch *et al.*, 2011; Sower, 2015).

### **1.2.1.3. Growth hormone**

Growth hormone (GH) is primarily produced in the pituitary and exerts its actions in various tissues to regulate growth and metabolism in mammals (Daughaday, 1989). It works either directly on somatic cells or indirectly through its key mediator IGF-I, which is produced primarily in the liver but also in various tissues (Duan *et al.*, 1993; Ohlsson *et al.*, 2009; Courtland *et al.*, 2011). In teleosts, both growth hormone and prolactin (PRL) play an important role in reproductive processes, particularly gonadal steroidogenesis. In goldfish *Carassius auratus*, GH enhances the effects of gonadotropins on ovarian steroidogenesis (Van der Kraak *et al.*, 1990) and stimulates steroidogenesis in spotted seatrout *Cynoscion nebulosus* (Singh & Thomas, 1993). It promotes body growth (teleosts, Donaldson *et al.*, 1979), follicle growth (mouse *Mus musculus*, Liu *et al.*, 1998), oocyte maturation (cow *Bos taurus*, Bevers & Izadyar, 2002) and ovulation (rabbit *Oryctolagus cuniculus*, Yoshimura *et al.*, 1994).

#### **1.2.1.4. The insulin superfamily of peptides and receptors**

The insulin superfamily of peptides consists of two groups of related peptides: a) insulin (INS) and insulin like growth factor (IGF); and b) relaxin (RLN) and insulin-like peptides (INSL). All the insulin peptides share a similar three-dimensional protein structure in their mature form (Sherwood, 2004). INS and IGF hormones are involved in carbohydrate and fat metabolism, as well as growth and development and signal primarily via receptor tyrosine kinases (RTKs) (Yegorov & Good, 2012). Most mammals harbor a single Ins and two Igf (Igf1 and Igf2) genes; Ins signals via the insulin receptor (Insr), while Igf1 and Igf2 activate Igf1 receptor (Igf1r), although binding is promiscuous and Ins can bind Igf1R and Igf1 and Igf2 can bind Insr (Yegorov & Good, 2012). In addition to these well-characterized ligands and receptors, an additional Igf ligand, called either IGf1b or Igf3, has been identified in multiple teleost species, and a third receptor, called insulin-receptor related-receptor (Insrr) is present in most tetrapods including humans *Homo sapiens* (Shier & Watt, 1989; Hänze *et al.*, 1999). Although these two genes were originally thought to be specific to the teleost or tetrapod lineages respectively, recent phylogenomic analyses in the Good laboratory suggests that both genes originated during 2R and therefore may be present in lampreys (Good *et al.*, in prep). Unlike Igf1 and Igf2 which are produced predominantly in the liver, teleost Igf3 has been found to be expressed predominantly in the testis and been shown to stimulate spermatogenesis following FSH secretion (Nobrega *et al.*, 2015). On the other hand, despite having strong sequence homology to both Insr and Igf1R, Insrr is an orphan receptor and does not, apparently, bind any members of the insulin superfamily examined so far (Dissen *et al.*, 2006). Insrr (or Irr) is expressed in a variety of tissues in humans including kidney, heart, skeletal muscle, liver, pancreas and ovary. However, of interest for the current study, Insrr is expressed in thecal-

interstitial cells in the mammalian ovary, and exhibits a peak in expression prior to ovulation, indicating that it plays a role in late oocyte development (Dissen *et al.*, 2006).

On the other hand, the relaxin and insulin-like (RLN/INSL) peptides mediate a broad variety of primarily reproductive and neuroendocrine functions and have played an important role in mammalian evolution (Wilkinson & Bathgate, 2007; Bathgate *et al.*, 2013); they signal via two very different classes of G-protein coupled receptors (GPCR) known as relaxin family peptide receptors (RXFP) (Bathgate *et al.*, 2013). One class of the relaxin family peptide receptors (RXFPs), consisting of RXFP1 and RXFP2, is closely related to glycoprotein hormone receptors (FSH and LH); they are activated by RLN and INSL3 and are involved in reproductive processes (Halls *et al.*, 2006). The other class of RXFPs, consisting of RXFP3 and RXFP4, are related to small peptide (e.g., angiotensin and somatostatin) receptors; they are activated by RLN3 and INSL5 and are involved in neuroendocrine processes (Bathgate *et al.*, 2006). While the role of INS/IGF peptides is quite well studied in vertebrates, the relaxin/insulin-like peptides are newly characterized hormones, and some of these peptides have also been found to play roles in reproduction, as suggested by the fact that they exert their action through glycoprotein hormone receptors (Bathgate *et al.*, 2006).

Yegorov & Good (2012) reconstructed the evolutionary history of RLN/INSL peptides and their two classes of GPCR receptors throughout vertebrate history and demonstrated that there was a single RLN/INSL molecule, a single RXFP3/4 ancestral molecule and one RXFP1/2 receptor in the vertebrate ancestor. They predicted that, following 2R, there were a total of four RLN/INSL peptides (namely, RLN3, RLN, INSL3 and INSL5), four RXFP 3/4 type receptors (RXFP3-1, 3-2, 3-3 and RXFP3-4) and three RXFP1/2 type receptors (RXFP1, RXFP2, and RXFP2-like) (Yegorov & Good, 2012). This evolutionary model has been confirmed in later

diverging vertebrates, such as spotted gar *Lepisosteus oculatus* and coelacanth *Latimeria chalumnae*, but could not be confirmed in lamprey (Yegorov *et al.*, 2014). Given this scenario, the lamprey genome should contain four RLN/INSL peptides and seven RXFP receptors barring no gene loss. The current annotation of the sea lamprey genome includes no RLN/INSL peptides, but 3-4 of the receptors. This is owing to the relatively poor assembly of the current sea lamprey genome and to the difficulty of data mining and confirming the sequence of small fast evolving peptides.

Unpublished work in the Good laboratory also provides insight into the early evolution of the INS/IGF peptides and their RTKs. Bioinformatics data mining and analysis of early insulin like peptide (ILPs) and the insulin-related receptors suggests that the vertebrate ancestor harbored single INS and IGF genes, and that following 2R, a single INS and at least two but probably three IGF genes were retained, giving rise to INS/IGF2 (which are linked), IGF1 and IGF3 in the post 2R vertebrate genome. Additionally, analysis suggests that there was a single insulin related RTK receptor in the vertebrate ancestor, but this receptor diversified during 2R to give rise to three receptors in the post 2R genome (INSR, INSRR, and IGF1R). This suggests that the lamprey genome could contain up to four INS/IGF ligands, and three receptors for these genes. The current annotation of the sea lamprey genome includes only two ligands and two receptors.

The insulin superfamily of genes is known to play diverse roles in vertebrate reproduction and biology (Yegorov *et al.*, 2014). Some of these genes could be expressed during ovarian differentiation and analysis of the lamprey ovarian transcriptome may yield confirmation of some of the unannotated genes.

#### 1.2.1.5. Insulin like growth factors (IGFs)

Insulin like growth factors (IGFs) are important molecules that affect a wide diversity of processes related to vertebrate growth, development, differentiation, metabolism and reproduction (Duan *et al.*, 1997; Nakamura *et al.*, 1998; Lu *et al.*, 2005; Wood *et al.*, 2005). Insulin-like growth factors (i.e., IGF peptides, IGF receptors and IGF binding proteins) help in regulation of ovarian physiology in teleosts (Reinecke, 2010). Most vertebrates possess two copies of IGFs, IGF1 and IGF2, while tilapia *Oreochromis niloticus* and zebrafish harbor a gonad specific IGF called *igf3*, which has been reported in the follicular layer of full grown oocytes (Wang *et al.*, 2008, Nelson & Van Der Kraak, 2010, Li *et al.*, 2011). IGF1 shows the most conserved role across vertebrates. It is primarily expressed in the liver in both mammals and fishes (Reinecke *et al.*, 1997; Ohlsson *et al.*, 2009; Reinecke, 2010; Reindl & Sheridan, 2012). IGF1 travels through the blood by binding to one or more of the essential IGF-binding-proteins where upon it locates its receptor (IGF1R or INSR) on diverse tissues, thereby exerting a central role in regulating growth, differentiation, and reproduction (Moriyama *et al.*, 2000). *In vitro* and *in vivo* studies suggest that IGFs induce the gonads to activate germ cell proliferation, oocyte maturation, and hence act as mediators of GH action in several tissues and regulate development and somatic growth (Moriyama *et al.*, 2000). Insulin-like growth factor 1 is associated with increased growth and fecundity but reduced lifespan, and it also stimulates the production of sex steroids (Dantzer & Swanson, 2012). The physiological roles of IGF2 in fish is still not clear (Reinecke *et al.*, 2005). In mammals, IGF2 has been co-opted to play a role in fetal development (Callan *et al.*, 2009) and it has recruited an unrelated receptor, IGF2 (a mannose-6-phosphate receptor). Spice *et al.* (2014) observed increased expression of *igf1r* in the oocyte growth stage in chestnut lamprey as compared to northern brook lamprey; chestnut lamprey larvae elaborate

more oocytes than do northern brook lamprey and this gene has been shown to regulate growth and reproduction by promoting cell division and differentiation and inhibits apoptosis (Reinecke, 2010). This suggests that the *igf1r* gene is involved in the growth of oocytes.

IGF3 is another newly discovered member of the insulin superfamily. The first studies on IGF3 suggested that it was a product of the teleost-specific WGD event that took place in the ancestor of teleosts ~320 Mya (Hurley *et al.*, 2007; Santini *et al.*, 2009). Using ancestral genome reconstruction methods developed in the Good laboratory, we hypothesize that Igf3 arose during 2R (Good *et al.*, in prep). Unlike IGF1 and IGF2 which are primarily expressed in the liver, IGF3 in fish is predominantly expressed in adult gonads (Wang *et al.*, 2008; Zou *et al.*, 2009). In zebrafish ovary, Igf3 mediated gonadotropin action in the final stages of oocyte development (Li *et al.*, 2015), while it stimulated spermatogenesis in male zebrafish (Nobrega *et al.*, 2015). Thus, if the gene is present in lampreys, as I predict, it may be identified in the transcriptomic analyses conducted in the current study. In human and several other studied mammals, both RXFP1 and RXFP2 with their respective ligands, RLN and INSL3, have been shown to be closely involved in reproduction (Halls *et al.*, 2006). Relaxin and RXFP1 are widely involved in pregnancy in mammals, where they are involved in relaxation of the pubic symphysis during labor or germ cell survival (Kawamura *et al.*, 2004; Anand-Ivell & Ivell, 2014), while INSL3 and RXFP2 are involved in spermatogenesis and testicular descent (Nef & Parada, 1999; Zimmermann *et al.*, 1999; Feng *et al.*, 2009).

Comparative evolutionary analyses in teleosts indicate that the ancestral Rln gene is very divergent from the RLN gene in mammals and is not hypothesized to play a predominant role in reproduction. This finding is supported by work in European eel *Anguilla anguilla* (Donizetti *et al.*, 2009) and zebrafish (Hu *et al.*, 2011), showing that fish RLN is primarily expressed in brain

not gonad. On the other hand, Good-Avila *et al.* (2009) showed that INSL3 is highly expressed in fish testis, and Bogerd *et al.* (2016) have recently shown that INSL3 appears to play conserved roles in spermatogenesis in both mammals and teleosts. On the other hand, other Rln/Insl ligands such as Rln, Rln3 and Insl5 all exhibit expression in fish ovaries (Good-Avila *et al.*, 2009). Although all four ligands are present in elephant shark *Callorhinchus milii*, none of the ligands have been adequately verified in lampreys. I hypothesize that lampreys harbor orthologs of both RLN and INSL3 and that INSL3 is likely to be expressed by gonad, probably more in testis than ovary. I further hypothesize that either of RXFP1 or RXFP2 could be expressed in ovary, since both receptors are preferentially expressed in reproductive organs. In addition to RLN/INSL3, I expect lampreys to have orthologs of RLN3 and INSL5. While these ligands primarily signal via somatostatin-like receptors, RXFP3 and RXFP4, RLN3 is known to activate the glycoprotein receptor RXFP1 and RXFP4 and RLN activates RXFP1 and RXFP2 (Halls *et al.*, 2005). Thus, while I do not expect to find high expression of the RXFP3 or RXFP4 receptors in lamprey gonad, either RLN3 or INSL5 could be identified. It should be noted that these hypotheses are novel and data mining of the lamprey genome and transcriptome is a powerful means to test hypotheses about the early evolution of any gene family that was greatly diversified during 2R, including the insulin superfamily.

### **1.3. Lampreys as a model species**

Lampreys have been studied as a model for the ancestral vertebrate condition including the evolution of hypothalamus-pituitary-gonadal axis and sexual maturation (Takio *et al.*, 2007; Bajoghli *et al.*, 2009; Sower *et al.*, 2009). However, in lampreys, the hypothalamic-pituitary-gonadal (HPG) and hypothalamic-pituitary-thyroid (HPT) systems overlap (Sower *et al.*, 2009),

which probably reflects the ancestral vertebrate condition but is different from that in later diverging vertebrates. The lamprey hypothalamic-pituitary system (Sower, 2015) controls the release of various peptides involved in reproduction: three hypothalamic GnRHs, one pituitary gonadotropin-type hormone, and one gonadal glycoprotein receptor (Sower *et al.*, 2009; Section 1.2.1.1). To make effective comparisons between lampreys and other vertebrates, a fuller understanding of ovarian differentiation and sexual maturation in lampreys is necessary and, as such, the existence of well-characterized histological stages of ovarian differentiation in lampreys is helpful. The two lamprey species selected for my study (congeneric parasitic and non-parasitic species) show differences in ovarian development that permit effective comparison. Most notably, as discussed above (Section 1.1.3.1), relative to the parasitic chestnut lamprey, the non-parasitic northern brook lamprey undergoes sexual differentiation and oocyte development at a smaller size/younger age, exhibits reduced fecundity, and undergoes an accelerated sexual maturation (including vitellogenesis) following metamorphosis. This will allow me to shed light on the genes controlling ovarian differentiation and sexual maturation in lampreys, which, until recently, has eluded study. The primary reason for this is that lampreys have a large and quite divergent genome from other vertebrates, and it was not until the publication of a full draft of the sea lamprey genome in 2013 (Smith *et al.*, 2013) that avenues were opened to understand many aspects of lamprey biology (Docker *et al.*, 2015; McCauley *et al.*, 2015). However, the lamprey genome is still not fully annotated, and it appears that a transcriptome of the ovary was not used to annotate the current lamprey genome (Smith *et al.*, 2013). Supporting this, Ensembl currently has gene ID numbers for 13,114 lamprey genes, but this represents only about half of all the lamprey genes because the lamprey genome contains almost 26,000 genes (Smith *et al.*, 2013). Thus, my thesis will help to identify the genes that have not been reported in sea lamprey

genome by using two different approaches: genome-guided and *de novo* assembly. Additionally, the pipelines designed for this project are easily reproducible and can be used for any non-model organism. All the data generated in this project can be used for updating the present list of known genes in lampreys.

#### **1.4. Different methods for gene expression data analysis**

There are multiple methods that may be used to analyze gene expression (i.e., the relative abundance of messenger RNA for a particular gene). The choice of method is dependent on the question being asked, the genomic resources available for the species, and financial considerations (Huestis & Marshall 2009). Next generation sequencing (NGS) techniques are currently a popular method of transcriptomic analysis (RNA-Seq), not only for model species, such as zebrafish (Collins *et al.*, 2012) but also for non-model organisms (Ekblom & Galindo 2011), including alternative life history types of other fish species (Jeukens *et al.*, 2010). It is also used for commercially important fish species, such as channel catfish *Ictalurus punctatus* (Liu *et al.*, 2012), European sea bass *Dicentrarchus labrax* (Sarropoulou *et al.*, 2012), and rainbow trout *Oncorhynchus mykiss* (Palstra *et al.*, 2013), and gives a descriptive representation of the transcriptome at a lower price relative to high throughput Sanger sequencing (Wang *et al.*, 2009).

RNA-Seq is assumed to provide a reliable estimate of absolute gene expression levels (Fu *et al.*, 2009). It produces thousands or millions of short sequence reads that can be assembled into a transcriptome by using different genome assembly software; this provides both sequence information and quantification of gene expression (Ekblom & Galindo, 2011). Sequencing an entire eukaryote genome is often still difficult and costly; however, sequencing the transcriptome

of an organism is a more manageable task and is often more informative (Hudson, 2008). Also, it does not require prior knowledge of the genes involved in particular processes or their sequence. It can provide data on hundreds of thousands of single nucleotide polymorphisms (SNPs) spread densely across a genome, thus helping to identify signatures of selection, produce a high density genetic map, and assemble genomes. For example, experiments using RNA-Seq from liver tissue determined that adults of the dwarf form of lake whitefish *Coregonus clupeaformis* overexpressed 51 genes related to energy metabolism, whereas the adults of the normal form overexpressed 116 genes related to protein synthesis; these differences correspond to metabolic and growth differences observed between the life history types (Jeukens *et al.*, 2010).

Goetz *et al.* (2010) used RNA-Seq to examine differences in liver gene expression between lean and siscowet forms of lake trout *Salvelinus namaycush* and found differences in groups of genes involved with immunity, lipid synthesis, metabolism, and transport. In the Midas cichlid *Amphilophus citrinellus*, thick- and thin-lipped forms have evolved repeatedly in lakes in Nicaragua. RNA-Seq was used to examine gene expression in the lips of these forms in four lakes, and revealed parallel differences in gene expression between forms, with greater divergence in gene expression in forms from older lakes (Manousaki *et al.*, 2013). In sea lamprey, transcriptome sequencing has been used to sequence mRNA from sea lamprey brain, intestine, liver, kidney, olfactory tissue and embryonic tissue (Smith *et al.*, 2013), as well as male sea lamprey rope, gill, and muscle tissue (Chung-Davidson *et al.*, 2013). Recently Chung-Davidson *et al.* (2015) studied liver transcriptomes of sea lamprey during the late larval, metamorphic, and emerging juvenile (i.e., early parasitic) stages.

Several studies of sex differentiation in other fish species have used results from RNA-Seq to perform qRT-PCR (rather than global transcriptomic analysis) for measuring the

expression of individual candidate genes (e.g., Baron *et al.*, 2005; Jorgensen *et al.*, 2008; Berbejillo *et al.*, 2012). Similar rapid changes in life history have been accompanied by gene expression changes in anadromous and landlocked forms of rainbow trout.

## 1.5. Goals and objectives

This M.Sc. thesis is an ovarian transcriptomics-based study in lampreys with a particular focus on identifying the suite of genes expressed at different gonadal stages. To this end, I have used two approaches for the analysis of the lamprey *Ichthyomyzon* spp. transcriptomes. In the first approach, I used the current assembly and **annotation** of the sea lamprey genome as a template for identifying and counting expression of the two *Ichthyomyzon* spp. (*Ichthyomyzon castaneus* and *Ichthyomyzon fossor*) transcriptomes. In this approach, transcriptomic reads that show high similarity to the genes in the sea lamprey genome were counted, normalized, and then the samples from both the species were pooled to look for differences in gene expression across different developmental stages (Section 1.1 & 1.1.3.1, Figure 2.2). In the second approach, I performed a *de novo* assembly of the *I. castaneus* and *I. fossor* transcriptomes separately by merging reads from multiple gonadal stages and then this *de novo* assembled transcriptome was used as a reference to align and count the number of transcripts expressed at different stages of ovarian development in both species. Since this second approach is done without the aid of an annotated reference and, as such, should identify novel genes (i.e., genes not currently identified in the sea lamprey genome), the putative identity of the *de novo* assembled transcripts were obtained by searching for similarity between the assembled transcripts against a custom database including 10 focal and well-characterized chordate genomes (Table 1.4). The second pipeline is

expected to identify genes that are expressed in the lamprey ovary but not identified in the current annotation of the sea lamprey genome.

To test the ability of the second pipeline to identify novel genes, I used the genes belonging to the insulin superfamily as test cases. Based on prior bioinformatics analyses, the number of genes for these families expected in the 2R lamprey genome is known, but not all the genes have been identified. I looked for evidence of expression of novel genes belonging to this gene family from pipeline 2 (Chapter 3). Thus, I test the ability of pipelines 1 and 2 to identify annotated genes belonging to the insulin superfamily to identify the strengths and weaknesses of both approaches. Additionally, I compared the genes identified and their counts from both pipelines 1 and 2.

***Objective 1:***

The first objective was to describe the genes involved in ovarian differentiation during six gonadal stages in lampreys. The primary hypothesis was that differences in oocyte development in parasitic and non-parasitic lampreys (Section 1.1.3.1) will be associated with differences in the timing of gene expression. In early gonadal stages (1, 2 and 3) when the ovaries are undifferentiated, there will be few genes associated with ovarian differentiation and maturation while in later gonadal stages (4, 5 and 6), genes associated with sexual maturation and vitellogenesis are expected to be highly expressed. I predict that global and specific patterns of gene expression will be the same for samples collected from the same gonadal stage in both parasitic and non-parasitic species of lampreys.

### ***Test of hypothesis:***

To address this hypothesis, whole transcriptomic samples of lamprey ovaries from both parasitic and non-parasitic species were pooled into three different groups a) early, b) mid and, c) late stages based on the stage of ovarian development, and the genes expressed at each stage were identified by mapping them onto the reference sea lamprey genome. The number of copies of all transcripts (genes) putatively identified at each gonadal stage was counted, genes that are up- or down-regulated during gonadal development identified, and Gene Ontology terms were assigned to the up- and down-regulated genes and the functional classes of the differentially expressed genes were identified (Chapter 2).

### ***Caveats:***

An important limitation of this approach was that reads that did not map to the reference genome or to unannotated regions of the lamprey genome on Ensembl were not included in this pipeline. ‘Mapping’ in simple terms, refers to the process of aligning short reads to a reference sequence to identify genes and transcripts.

### ***Objective 2:***

While pipeline 1 described above was useful for identifying differences in the expression of known (annotated) genes during ovarian differentiation, I hypothesized that some genes that were expressed during ovarian differentiation have not been identified in the lamprey genome. Thus, the second objective was to identify novel genes (i.e., those not currently annotated in the sea lamprey genome) expressed during ovarian development. For this second approach, I generated a *de novo* reference ovarian transcriptome from parasitic and non-parasitic species

separately, and then used this reference transcriptome to map *de novo* assembled transcripts from each sample of parasitic and non-parasitic taxa separately. The read-out from this first step was a list of *de novo* assembled transcripts from each sample/species. To assign tentative ID's to these transcripts, I queried the entire set of transcripts against a custom database of 10 fully sequenced chordate genomes. To find novel genes (i.e., those not currently identified in the sea lamprey genome), the list of putative genes identified by this method was compared to the list of known sea lamprey genes obtained in the first approach.

### ***Objective 3:***

My third objective was to develop a list of genes known to be involved in sexual maturation and ovarian development in other vertebrates and determine if they are expressed during ovarian development in lampreys; although many of these genes are not clearly annotated in the sea lamprey genome, I looked for evidence of their expression during ovarian development in the two *Ichthyomyzon* species. Furthermore, the number of gene copies and their sequence may not be known. One of the main roles of transcriptomic data is to provide support for *ab initio* gene predictions and confirm gene annotations. Thus, I looked for evidence for expression of each of the genes listed in Table 1.2 in the transcriptomic data, and obtained normalized counts of their expression in chestnut and northern brook lampreys during different stages of ovarian development. I aimed to provide sequence confirmation (and potentially annotation) and comparison of the lamprey sequence to that found in other vertebrates. This allowed me to test the two pipelines, a reference-guided and a reference-free assembly (Chapters 2 & 3), and to provide information about the role of these known genes in the early vertebrate genome.

### ***Test of hypothesis:***

In some cases, homologs (putative orthologs and paralogs) of the genes listed in Table 1.2 have been identified as novel and were not found in the Ensembl annotation file of the sea lamprey genome. For these genes, I identified when and how many times each gene was expressed at different gonadal stages of chestnut and northern brook lampreys using HTSeq (Section 2.5.6). For those genes that were not identified by HTSeq, I used a different pipeline and executed BLAST (Basic Local Alignment Search Tool) to find regions of similarity between sequences. In short, I created a database of later diverging species such as zebrafish, frog *Xenopus tropicalis*, chicken *Gallus gallus* and human *Homo sapiens* and BLASTed it against the *de novo* assembled *Ichthyomyzon* transcripts. If the gene was present in the sample, it gave a hit and hence provided evidence that these genes are expressed during lamprey ovarian development and maturation.

### ***Objective 4:***

Additionally, since RNA-Seq analysis is less well developed for non-model organisms, another goal of this thesis was to develop an RNA-Seq analysis pipeline that is suitable for non-model taxa and can have significance beyond the study of lamprey gonadal development. One of the biggest challenges of transcriptomic data analysis for non-model species is the availability of a reference genome. If a well-annotated reference genome of an organism is available, it becomes easier to perform RNA-Seq analysis, but well-annotated reference genomes are often not available and it is not yet clear how closely related the reference species should be for obtaining optimal results. In my project, I had to choose between two publicly available genomes, that of sea lamprey or Arctic lamprey *Lethenteron camtschaticum* for mapping the

reads. The selection of a reference genome depends on different parameters such as coverage, distance between the reference species and the species of interest, choice of mapping software, and possible contamination during sequencing and library preparation. The sea lamprey genome was preferred over Arctic lamprey because it gave better mapping percentage compared to the Arctic lamprey. In this thesis, a pipeline was designed for a non-model organism with a reference genome (Chapter 2), but at the same time it was observed that the mapping percentage was still low (12-45%), so another pipeline was designed in which the reference genome was not used (Chapter 3). By *de novo* assembly of the two species of lampreys, chestnut and northern brook lamprey, assembled transcriptomes of both the species were formed and used for further mapping. It was observed that 20-30% of the reads were not assembled into **contigs**. The ‘Assembled Transcriptome’ was used further for identifying novel genes. This clearly suggests that transcriptomic data can be analyzed in both ways, with a reference genome (genome-guided) and without a reference genome (*de novo* assembly).

## **1.6. Significance of proposed research**

This research has helped to identify and annotate genes involved in ovarian differentiation and sexual maturation in lampreys, which will help shed light on the evolution of these genes and their functions in vertebrates. Even more broadly, the different pipelines generated during this project can be further used to address specific questions for other non-model organisms in terms of sample size, mapping percentage, presence or absence of a reference genome. These pipelines will be of general use to studies on other non-model organisms.

## 1.7. References

- Altschul S, Gish W, Miller W *et al.* (1990) A basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Amores A, Catchen J, Ferrara A *et al.* (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799–808.
- Anand-Ivell R. & Ivell R (2014) Regulation of the reproductive cycle and early pregnancy by relaxin family peptides. *Molecular Cell Endocrinology*, **382**, 472–479.
- Anders S, Pyl PT, Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
- Bajoghli B, Aghaallaei N, Hess I *et al.* (2009) Evolution of genetic networks underlying the emergence of thymopoiesis in vertebrates. *Cell*, **138**, 186–97.
- Barker LA & Beamish FWH (2000) Gonadogenesis in landlocked and anadromous forms of the sea lamprey, *Petromyzon marinus*. *Environmental Biology of Fishes*, **59**, 229–234.
- Baron D, Houlgatte R, Fostier A *et al.* (2005) Large-scale temporal gene expression profiling during gonadal differentiation and early gametogenesis in rainbow trout. *Biology of Reproduction*, **73**, 959–966.
- Bathgate RAD, Halls ML, van der Westhuizen, ET *et al.* (2013) Relaxin family peptides and their receptors. *Physiological Reviews*, **93**, 405–480.
- Beamish FWH & Thomas EJ (1983) Potential and actual fecundity of the “paired” lampreys, *Ichthyomyzon gagei* and *I. castaneus*. *Copeia*, **1983**, 367–374.

- Berbejillo J, Martinez-Bengochea A, Bedo G *et al.* (2012) Expression and phylogeny of candidate genes for sex differentiation in a primitive fish species, the Siberian sturgeon, *Acipenser baerii*. *Molecular Reproduction and Development*, **79**, 504–516.
- Bevers M & Izadyar F (2002) Role of growth hormone and growth hormone receptor in oocyte maturation. *Molecular Cell Endocrinology*, **197**, 173- 78.
- Blank M, Jürss K, Bastrop R (2008) A mitochondrial multigene approach contributing to the systematics of the brook and river lampreys and the phylogenetic position of *Eudontomyzon mariae*. *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 2780–2790.
- Bracken FSA, Hoelzel AR, Hume JB *et al.* (2015) Contrasting population genetic structure among freshwater-resident and anadromous lampreys: the role of demographic history, differential dispersal and anthropogenic barriers to movement. *Molecular Ecology*, **24**, 1188–1204.
- Bromage N, Jones J, Randall C *et al.* (1992) Broodstock management, fecundity, egg quality and the timing of egg production in the rainbow trout (*Oncorhynchus mykiss*). *Aquaculture*, **100**, 141-166.
- Brown WM, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Science*, **76(4)**, 1967-1971.
- Callan AC & Milne E (2009) Involvement of the IGF system in fetal growth and childhood cancer: an overview of potential mechanisms. *Cancer Causes Control*, **20**, 1783-1798.
- Campbell B, Dickey J, Beckman B *et al.* (2006) Previtellogenic oocyte growth in salmon: relationships among body growth, plasma insulin-like growth factor-1, estradiol-17b, follicle-stimulating hormone and expression of ovarian genes for insulin- like growth

- factors, steroidogenic-acute regulatory protein and receptors for gonadotropins, growth hormone, and somatolactin. *Biology of Reproduction*, **75**, 34–44.
- Chung-Davidson YW, Priess MC, Yeh CY *et al.* (2013) A thermogenic secondary sexual character in male sea lamprey. *The Journal of Experimental Biology*, **216**, 2702–2712.
- Chung-Davidson YW, Yeh CY, Bussy U *et al.* (2015) Hsp90 and hepatobiliary transformation during sea lamprey metamorphosis. *BMC Developmental Biology*, **15** (1), 47.
- Collins JE, White S, Searle SM *et al.* (2012) Incorporating RNA-Seq data into the zebrafish Ensembl genebuild. *Genome Research*, **22**, 2067–2078.
- Courtland HW, Sun H, Beth-On M *et al.* (2011) Growth hormone mediates pubertal skeletal development independent of hepatic IGF-1 production. *Bone and Mineral Research*, **26**, 761–768.
- Crespo D, Assis LH, Bogerd J *et al.* (2016) Expression profiling identifies Setoli and Leydig cells genes as FSH targets in adult zebrafish testis. *Molecular Cell Endocrinology*, **437**, 237–251.
- Dantzer B & Swanson EM (2012) Mediation of vertebrate life histories via insulin-like growth factor-1. *Biological Reviews of the Cambridge Philosophical Society*, **87**, 414–429.
- Daughaday WHA (1989) Personal history of the origin of the somatomedin hypothesis and recent challenges to its validity. *Perspectives in Biology and Medicine*, **32**, 194–211.
- Dawson HA, Quintella BR, Almeida PR *et al.* (2015) The ecology of larval and metamorphosing lampreys. In: *Lampreys: Biology Conservation and Control* (ed. Docker MF) *Fish & Fisheries Series*, Springer, **37**, 75-138.

- Dissen, GA, Garcia-Rudaz C, Tapia V *et al.* (2006) Expression of the insulin receptor-related receptor is induced by the preovulatory surge of luteinizing hormone in thecal-interstitial cells of the rat ovary. *Endocrinology*, **147**, 155–165.
- Docker MF, Hume JB, Clemens BJ (2015) Introduction: A surfeit of lampreys. *Lampreys: Biology, Conservation and Control, Springer Netherlands*, **1**, 1-34.
- Docker MF, Youson JH, Beamish RJ *et al.* (1999) Phylogeny of the lamprey genus *Lampetra* inferred from mitochondrial cytochrome b and ND3 gene sequences. *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 2340–2349.
- Docker MF (2009) A review of the evolution of nonparasitism in lampreys and an update of the paired species concept. *American Fisheries Society Symposium*, **72**, 71–114.
- Donaldson EM, Fagerlund UHM, Higgs DA *et al.* (1979) Hormonal enhancement of growth. In: *Fish physiology* (Eds: W.S. Hoar, D.J. Randall and J.R. Brett). *Academic Press, New York*, **8**, 455-597.
- Donizetti A, Fiengo M, Minucci S *et al.* (2009) Duplicated zebrafish relaxin-3 gene shows a different expression pattern from that of the co-orthologue gene. *Development Growth and Differentiation*, **51**, 715–722.
- Duan C, Duguay SJ, Plisetskaya EM (1993) Insulin-like growth factor I (IGF-I) mRNA expression in coho salmon, *Oncorhynchus kisutch*: tissue distribution and effects of growth hormone/prolactin family proteins. *Fish Physiology and Biochemistry*, **11**, 371–379.
- Duan, C (1997) The insulin-like growth factor system and its biological actions in fish. *American Zoologist*, **37**, 491–503.

- Dubois EA, Zandbergen MA, Peute J *et al.* (2002) Evolutionary development of three gonadotropin-releasing hormone (GnRH) systems in vertebrates. *Brain Research Bulletin*, **57**, 413–41.
- Eden E, Navon R, Steinfeld I *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.
- Ekblom R & Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.
- Espanhol R, Almeida PR, Alves MJ (2007) Evolutionary history of lamprey paired species *Lampetra fluviatilis* (L.) and *Lampetra planeri* (Bloch) as inferred from mitochondrial DNA variation. *Molecular Ecology*, **16**, 1909–1924.
- Evans HM & Long JA (1992) Characteristic effects upon growth, oestrus and ovulation induced by the intraperitoneal administration of fresh anterior hypophyseal substance. *Proceedings of the National Academy of Science*, **8**, 38–910.
- Feng S, Ferlin A, Truong A *et al.* (2009) INSL3/RXFP2 signaling in testicular descent. *Endocrinology*, **1160**, 197–204.
- Fu X, Fu N, Guo S *et al.* (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, **10**, 161.
- Fukayama S & Takahashi H (1983) Sex differentiation and development of the gonad in the sand lamprey, *Lampetra reissneri*. *Bulletin of the Faculty of Fisheries Hokkaido University*, **34**, 279–290.
- Garant D, Dodson JJ, Bernatchez L (2003) Differential reproductive success and heritability of alternative reproductive tactics in wild Atlantic salmon (*Salmo salar* L.). *Evolution*, **57**, 1133–1141.

- Goetz F, Rosauer D, Sitar S *et al.* (2010) A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (*Salvelinus namaycush*). *Molecular Ecology*, **19**, 176–196.
- Gomez JM, Weil C, Ollitrault M *et al.* (1999) Growth hormone (GH) and gonadotropin subunit gene expression and pituitary and plasma changes during spermatogenesis and oogenesis in rainbow trout (*Oncorhynchus mykiss*). *General Comparative Endocrinology*, **113**, 413–428.
- Good-Avila SV, Yegorov S, Harron S *et al.* (2009) Relaxin gene family in teleosts: phylogeny, syntenic mapping, selective constraint, and expression analysis. *BMC Evolutionary Biology*, **9**, 293.
- Gorbman A & Sower SA (2003) Evolution of the role of GnRH in animal (metazoan) biology. *General Comparative Endocrinology*, **134**, 207–213.
- Greep RO, Van Dyke HB, Chow HB (1941) Use of anterior lobe of prostate gland in the assay of metakentrin. *Proceedings of the Society for Experimental Biology and Medicine*, **46**, 644.10.3181.
- Guiguen Y, Fostier A, Piferrer F *et al.* (2010) Ovarian aromatase and estrogen: a pivotal role for gonadal sex differentiation and sex change in fish. *General Comparative Endocrinology*, **165**, 352–366.
- Guilgur LG, Moncaut NP, Canario AV *et al.* (2006) Evolution of GnRH ligands and receptors in gnathostomata. *Comparative Biochemistry and Physiology Part A Molecular and Integrative Physiology*, **144**, 272–283.

- Hänze J, Berthold A, Klammt J *et al.* (1999) Cloning and sequencing of the complete cDNA encoding the human insulin receptor related receptor. *Hormone Metabolism Research*, **31**, 77–79.
- Halls ML, Bathgate RAD, Summers RJ (2006) Relaxin family peptide receptors RXFP1 and RXFP2 modulate cAMP signaling by distinct mechanisms. *Molecular Pharmacology*, **70**, 214–226.
- Halls ML, Bond CP, Sudo S *et al.* (2005) Multiple binding sites revealed by interaction of relaxin family peptides with native and chimeric relaxin family peptide receptors 1 and 2 (LGR7 and LGR8). *Pharmacology and Experimental Therapeutics*, **313** (2), 677–687.
- Hardisty MW & Potter IC (1971) Paired species. In: *The Biology of Lampreys, volume 1* (eds Hardisty MW, Potter IC), pp 249–277. Academic Press, New York.
- Hardisty MW (1960) Development of the gonads in parasitic and non-parasitic lampreys. *Nature*, **187**, 341–342.
- Hardisty MW (1961) Oocyte numbers as a diagnostic character for the identification of ammocoete species. *Nature*, **191**, 1215–1216.
- Hardisty MW (1965) Sex differentiation and gonadogenesis in lampreys. II. The ammocoete gonads of the landlocked sea lamprey, *Journal of Zoology*, **146**, 346–387.
- Hardisty MW (1969) A comparison of gonadal development in the ammocoetes of the landlocked and anadromous forms of the sea lamprey *Petromyzon marinus* L. *Journal of Fish Biology*, **1**, 153–166.
- Hardisty MW (1970) The relationship of gonadal development to the life cycles of the paired species of lamprey, *Lampetra fluviatilis* (L.) and *Lampetra planeri* (Bloch). *Journal of Fish Biology*, **2**, 173–181.

- Hardisty MW (1971) Gonadogenesis, sex differentiation and gametogenesis. In: *The Biology of Lampreys, volume 1* (eds Hardisty MW, Potter IC), pp. 317–324. Academic Press, New York.
- Hardisty MW (2006) *Lampreys: life without jaws*. Forrest Text, Ceredigion, UK.
- Hoegg S, Brinkmann H, Taylor JS *et al.* (2004). Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Molecular Evolution*, **59**, 190–203.
- Hu GB, Kusakabe M, Takei Y (2011) Localization of diversified relaxin gene transcripts in the brain of eels. *General and Comparative Endocrinology*, **172**, 430–439.
- Hubbs CL (1971) *Lampetra (Entosphenus) lethophaga*, a new species, the nonparasitic derivative of the Pacific lamprey. *Transactions of the San Diego Society of Natural History*, **16**, 125–164.
- Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.
- Huestis DL & Marshall JL (2009) From gene expression to phenotype in insects: Non-microarray approaches for transcriptome analysis. *BioScience*, **59**, 373–384.
- Hughes RL & Potter IC (1969) Studies on gametogenesis and fecundity in the lampreys *Mordacia praecox* and *M. mordax* (Petromyzonidae). *Australian Journal of Zoology*, **17**, 447–464.
- Hull KS & Harvey S (2001) Growth hormone: roles in female reproduction. *Endocrinology*, **168**, 1–23.
- Hurley IA, Mueller RL, Dunn KA *et al.* (2007) A new time-scale for ray-finned fish evolution. *Proceedings of the Royal Society of Edinburgh Biology*, **274**, 489–498.

- Jaillon O, Aury JM, Brunet F *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–57.
- Jeukens J, Renaut S, St-Cyr J *et al.* (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology*, **19**, 5389–5403.
- Jørgensen A, Morthorst JE, Andersen O *et al.* (2008) Expression profiles for six zebrafish genes during gonadal sex differentiation. *Reproductive Biology and Endocrinology*, **6**, 25.
- Joseph NT, Aquilina-Beck A, MacDonald C *et al.* (2012) Molecular cloning and pharmacological characterization of two novel GnRH receptors in the lamprey (*Petromyzon marinus*). *Endocrinology*, **153**(7), 3345–3356.
- Kah O, Lethimonier C, Somoza G *et al.* (2007) GnRH and GnRH receptors in metazoa: A historical, comparative, and evolutive perspective. *General Comparative Endocrinology*, **153**, 346–364.
- Karlsen Ø, Holm JC, Kjesbu OS (1995) Effects of periodic starvation on reproductive investment in first-time spawning Atlantic cod (*Gadus morhua* L.). *Aquaculture*, **133**(2), 159-170.
- Kavanaugh SI, Nozaki M, Sower SA (2008) Origins of gonadotropin-releasing hormone (GnRH) in vertebrates: identification of a novel GnRH in a basal vertebrate, the sea lamprey. *Endocrinology*, **149**, 3860–3869.
- Kawamura K, Kumagai J, Sudo S *et al.* (2004) Paracrine regulation of mammalian oocyte maturation and male germ cell survival. *Proceedings of the National Academy of Sciences*, **101**, 7323–7328.

- Kawauchi H, Suzuki K, Itho H *et al.* (1989) The duality of teleost gonadotropins. *Fish Physiology and Biochemistry*, **7**, 29–38.
- Kazeto Y, Kohara M, Tosaka R *et al.* (2012) Molecular characterization and gene expression of Japanese eel (*Anguilla japonica*) gonadotropin receptors. *Zoological Science*, **29**, 204–211.
- Kohn M, Hogel J, Vogel W *et al.* (2006) Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends in Genetics*, **22**, 203–210.
- Kucheryavyi A, Savvaitova K, Pavlov D *et al.* (2007) Variations of life history strategy of the arctic lamprey *Lethenteron camtschaticum* from the Utkholok River (Western Kamchatka). *Journal of Ichthyology*, **47**, 37–52.
- Lewis JC & McMillan DB (1965) The development of the ovary of the sea lamprey (*Petromyzon marinus* L.). *Journal of Morphology*, **117**, 425–441.
- Li J, Chu L, Sun X *et al.* (2015) IGFs mediate the action Of LH on oocyte maturation in zebrafish. *Molecular Endocrinology*, **29**, 373–383.
- Li J, Liu Z, Wang D *et al.* (2011) Insulin-like growth factor 3 is involved in the oocyte maturation in zebrafish. *Biology of Reproduction*, **84**, 473–486.
- Li MD & Ford JJ (1998) A comprehensive evolutionary analysis based on nucleotide and amino acid sequences of the  $\alpha$ - and  $\beta$ -subunits of glycoprotein hormone gene family. *Journal of Endocrinology*, **156**, 529–542.
- Liu S, Zhang Y, Zhou Z *et al.* (2012) Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. *BMC Genomics*, **13**, 595.

- Liu X, Andoh K, Yokota H *et al.* (1998) Effects of growth hormone, activin, and follistatin on the development of preantral follicle from immature female mice. *Endocrinology*, **5**, 2342–2347.
- Lu C, Lam HN, Menon RK (2005) New members of the insulin family: regulators of metabolism, growth and now reproduction. *Pediatric Research*, **57**, 70R–73R.
- Lubzens E, Young G, Bobe J *et al.* (2010) Oogenesis in teleosts: how fish eggs are formed. *General Comparative Endocrinology*, **165**, 367–389.
- Manousaki T, Hull PM, Kusche H *et al.* (2013) Parsing parallel evolution: ecological divergence and differential gene expression in the adaptive radiations of thick-lipped Midas cichlid fishes from Nicaragua. *Molecular Ecology*, **22**, 650–669.
- Manzon RG, Youson JH, Holmes JA (2015) Lamprey metamorphosis. In: Docker MF, editor. *Lampreys: Biology, Conservation, and Control*, volume **1**. Springer, 139-214.
- Mateus CS, Stange M, Berner D *et al.* (2013) Strong genome-wide divergence between sympatric European river and brook lampreys. *Current Biology*, **23**, R649-R650.
- McCauley DW, Docker MF, Whyard S *et al.* (2015) Lampreys as diverse model organisms in the genomics era. *BioScience*, **65**, 1046–1056.
- Meijide FJ, Lo Nostro FL, Guerrero GA (2005) Gonadal development and sex differentiation in the cichlid fish *Cichlasoma dimerus* (Teleostei, Perciformes): a light- and electron-microscopic study. *Journal of Morphology*, **264**, 191–210.
- Moriyama S, Ayson FG, Kawauchi H (2000) Growth regulation by insulin-like growth factor-I in fish. *Bioscience, Biotechnology, and Biochemistry*, **64**, 1553–1562.

- Nachtigal MW, Hirokawa, Y, Enyeart-VanHouten *et al.* (1998) Wilms' tumor 1 and Dax-1 modulate the orphan nuclear receptor SF-1 in sex-specific gene expression. *Cell*, **93**, 445–454.
- Nakamura M, Kobayashi T, Chang XT *et al.* (1998) Gonadal sex differentiation in fish. *Experimental Zoology*, **281**, 362–372.
- Naruse K, Tanaka M, Mita K *et al.* (2004) A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Research*, **14**, 820–828.
- Nef S & Parada LF (1999) Cryptorchidism in mice mutant for *Insl3*. *Nature Genetics*, **22**, 295–299.
- Nelson SN & Van Der Kraak G (2010) Characterization and regulation of the insulin-like growth factor (IGF) system in the zebrafish (*Danio rerio*) ovary. *General and Comparative Endocrinology*, **168**, 111–120.
- Nóbrega RH, de Souza Morais RD, Crespo D *et al.* (2015) Fsh stimulates spermatogonial proliferation and differentiation in zebrafish via *Igf3*. *Endocrinology*, **156**, 3804–3817.
- Ogiwara K, Fujimori C, Rajapakse S *et al.* (2013) Characterization of Luteinizing hormone and luteinizing hormone receptor and their indispensable role in the ovulatory process of the Medaka. *PLoS ONE*, **8**, 1–14
- Ohlsson C, Mohan S, Sjögren K *et al.* (2009) The role of liver-derived insulin-like growth factor-I. *Endocrine Reviews*, **30(5)**, 494–535.
- Okubo K & Nagahama Y (2008) Structural and functional evolution of gonadotropin releasing hormone in vertebrates. *Acta Physiologica*, **193**, 3–15.

- Orlov AM, Savinyh VF, Pelenev DV (2008) Features of the spatial distribution and size structure of the Pacific lamprey *Lampetra tridentata* in the North Pacific. *Russian Journal of Marine Biology*, **34**, 276–287.
- Palstra AP, Beltran S, Burgerhout E *et al.* (2013) Deep RNA sequencing of the skeletal muscle transcriptome in swimming fish, *PloS ONE*, **8**, 53171.
- Papadaki M, Piferrer F, Zanuy S *et al.* (2005) Growth, sex differentiation, and gonadal and plasma levels of sex steroids in male- and female-dominant populations of *Dicentrarchus labrax* L. obtained through repeated size grading. *Journal of Fish Biology*, **66**, 938–956.
- Patino R & Takashima F (1995) Gonads. In: *An atlas of fish histology, normal and pathological features* (eds Takashima F, Hibiya T), pp. 128–153.
- Pereira AM, Robalo JJ, Freyhof J *et al.* (2010) Phylogeographical analysis reveals multiple conservation units in brook lampreys *Lampetra planeri* of Portuguese streams. *Journal of Freshwater Ecology*, **77**, 311–434.
- Piferrer F & Guiguen Y (2008) Fish gonadogenesis. Part II: Molecular biology and genomics of sex differentiation. *Reviews in Fisheries Science*, **16**, 35–55.
- Piferrer F, Ribas L, Díaz N (2012) Genomic approaches to study genetic and environmental influences on fish sex determination and differentiation. *Marine Biotechnology*, **14**, 591–604.
- Postlethwait JH, Woods IG, Ngo-Hazelett P *et al.* (2000) Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Research*, **10**, 1890–1902.
- Potter IC, Gill HS, Renaud CB *et al.* (2015) The taxonomy, phylogeny, and distribution of lampreys. Docker MF, ed. *Lampreys: Biology, Conservation, and Control*, volume **1**. Springer, 35–37.

- Quérat B, Arai Y, Henry A *et al.* (2004) Pituitary glycoprotein hormone  $\beta$  subunits in the Australian lungfish and estimation of the relative evolution rate of these subunits within vertebrates. *Biology of Reproduction*, **70**, 356–363.
- Quérat B, Sellouk A, Salmon C (2000) Phylogenetic analysis of the vertebrate glycoprotein hormone family including new sequences of sturgeon (*Acipenser baeri*)  $\beta$ -subunits of the two gonadotropins and the thyroid-stimulating hormone. *Biology of Reproduction*, **63**, 222–228.
- Reindl KM & Sheridan MA (2012) Peripheral regulation of the growth hormone-insulin-like growth factor system in fish and other vertebrates. *Comparative Biochemistry and Physiology Part A*, **163**, 231–245.
- Reinecke M (2010) Insulin-like growth factors and fish reproduction. *Biology of Reproduction*, **82**, 656–661.
- Reinecke M, Björnsson BT, Dickhoff WW *et al.* (2005) Growth hormone and insulin-like growth factors in fish: where we are and where to go. *General Comparative Endocrinology*, **142**, 20–24.
- Reinecke M, Schmid A, Ermatinger R *et al.* (1997) Insulin-like growth factor I in the teleost *Oreochromis mossambicus* the tilapia: gene sequence, tissue expression, and cellular localization. *Endocrinology*, **138**, 3613–3619.
- Roch GJ, Busby ER, Sherwood NM (2011) Evolution of GnRH: diving deeper. *General Comparative Endocrinology*, **171**, 1–16.
- Rougemont Q, Gagnaire PA, Perrier C *et al.* (2016) Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. *Molecular Ecology*, **26**, 142–162.

- Rougemont Q, Gaigher A, Lasne E *et al.* (2015) Low reproductive isolation and highly variable levels of gene flow reveal limited progress towards speciation between European river and brook lampreys. *Journal of Evolutionary Biology*, **28**, 2248–2263.
- Sandra GE & Norma MM (2010). Sexual determination and differentiation in teleost fish. *Reviews in Fish Biology and Fisheries*, **20**, 101–121.
- Santini F, Harmon L, Carnevale G *et al.* (2009) Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evolutionary Biology*, **9**, 194.
- Sarropoulou E, Galindo-Villegas J, García-Alcázar A *et al.* (2012) Characterization of European sea bass transcripts by RNA Seq after oral vaccine against *V. anguillarum*. *Marine Biotechnology*, **14**, 634–642.
- Scharnhorst V, van der Eb AJ, Jochemsen, AG (2001) WT1 proteins: functions in growth and differentiation. *Gene*, **273**, 141–161.
- Schreiber A &, Engelhorn R (1998) Population genetics of a cyclostome species pair, river lamprey *Lampetra fluviatilis* (L.) and brook lamprey *Lampetra planeri* (Bloch). *Journal of Zoological Systematics and Evolutionary Research*, **36**, 85–99.
- Shearer K, Parkins P, Gadberry B *et al.* (2006) Effects of growth rate/body size and a low lipid diet on the incidence of early sexual maturation in juvenile male spring Chinook salmon (*Oncorhynchus tshawytscha*). *Aquaculture*, **252**, 545-556.
- Sherwood NM, Sower SA, Marshak DR *et al.* (1986) Primary structure of gonadotropin-releasing hormone from lamprey brain. *Journal of Biological Chemistry*, **261**, 4812–4819.

- Sherwood OD (2004) Relaxin's physiological roles and other diverse actions. *Endocrine Reviews*, **25**, 205–234.
- Shi B, Liu X, Xu Y *et al.* (2015) Molecular characterization of three gonadotropin subunits and their expression patterns during ovarian maturation in *Cynoglossus semilaevis*. *International Journal of Molecular Science*, **16**, 2767–2793.
- Shier P & Watt VM (1989) Primary structure of a putative receptor for a ligand of the insulin family. *Journal of Biological Chemistry*, **264**, 14605–14608.
- Silver MR, Kawauchi H, Nozaki M *et al.* (2004) Cloning and analysis of the lamprey GnRH-III cDNA from eight species of lamprey representing the three families of *Petromyzoniformes*. *General Comparative Endocrinology*, **139**, 85–94.
- Silverstein JT, Shearer KD, Dickhoff WW *et al.* (1998) Effects of growth and fatness on sexual development of chinook salmon (*Oncorhynchus tshawytscha*) parr. *Canadian Journal of Fisheries and Aquatic Sciences*, **55**, 2376–2382.
- Singh H & Tomas P (1993) Mechanism of stimulatory action of growth hormone on ovarian steroidogenesis in spotted seatrout, *Cynoscion nebulosus*. *General Comparative Endocrinology*, **89**, 341–353.
- Smith JJ & Keinath, MC (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Research*, **25**, 1081–1090.
- Smith JJ, Kuraku S, Holt C *et al.* (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genetics*, **45**, 415–421.
- Sower SA (2003) The endocrinology of reproduction in lampreys and applications for male lamprey sterilization. *Journal of Great Lakes Research*, **29**, 50–65.

- Sower SA (2015) The Reproductive Hypothalamic-Pituitary Axis in Lampreys. In: Docker MF, Editor. *Lampreys: Biology, Conservation and Control*, volume **1**. Springer, 305-373.
- Sower SA, Chiang YC, Lovas S *et al.* (1993) Primary structure and biological activity of a third gonadotropin-releasing hormone from lamprey brain. *Endocrinology*, **132**, 1125–1131.
- Sower SA, Freamat M, Kavanaugh SI (2009) The origins of the vertebrate hypothalamic-pituitary-gonadal (HPG) and hypothalamic-pituitary-thyroid (HPT) endocrine systems: new insights from lampreys. *General and Comparative Endocrinology*, **161**, 20–29.
- Sower SA, McGregor AJ, Materne OL *et al.* (2000) Evidence for lamprey GnRH-I and-III-like molecules in the brains of the southern hemisphere lampreys *Geotria australis* and *Mordacia mordax*. *General Comparative Endocrinology*, **120**, 168–175.
- Sower SA, Moriyama S, Kasahara M *et al.* (2006) Identification of sea lamprey GTH beta-like cDNA and its evolutionary implications. *General Comparative Endocrinology*, **148**, 22–32.
- Spice EK (2013) Ovarian Differentiation in an Ancient Vertebrate: Timing, Candidate Gene Expression, and Global Gene Expression in Parasitic and Non-parasitic Lampreys (M.Sc. thesis). University of Manitoba, Winnipeg, Canada.
- Spice EK, Whyard S, Docker MF (2014) Gene expression during ovarian differentiation in parasitic and non-parasitic lampreys: Implications for fecundity and life history types. *General and Comparative Endocrinology*, **208**, 116–125.
- Suzuki A, Tanaka M, Shibata N (2004) Expression of aromatase mRNA and effect of aromatase inhibitor during ovarian development in the medaka, *Oryzias latipes*. *Journal of Experimental Zoology*, **301**, 266–273.

- Suzuki K, Kawauchi H, Nagahama Y (1988) Isolation and characterization of subunits from two distinct salmon gonadotropins. *General Comparative Endocrinology*, **71**, 302–306.
- Swanson P (1991) Salmon gonadotropins: reconciling old and new ideas. *Proceedings of International Symposium Reproductive Physiology of Fish. Fish Symposium*, **91**, Sheffield.
- Takio Y, Kuraku S, Murakami Y *et al.* (2007) Hox gene expression patterns in *Lethenteron japonicum* embryos - insights into the evolution of the vertebrate Hox code. *Developmental Biology*, **308**, 606–620.
- Taranger GL, Carrillo M, Schulz RW *et al.* (2010) Control of puberty in farmed fish. *General Comparative Endocrinology*, **165**, 483–515.
- Thorpe JE, Talbot C, Miles MS *et al.* (1990) Control of maturation in cultured Atlantic salmon, *Salmo salar*, in pumped seawater tanks, by restricting food intake. *Aquaculture*, **86**, 315–326.
- Trinh LA, McCutchen MD, Bonner-Fraser M *et al.* (2007) Fluorescent *in situ* hybridization employing the conventional NBT/BCIP chromogenic strain. *Biotechniques*, **42**, 756–759.
- Uchida K, Moriyama S, Chiba H *et al.* (2010) Evolutionary origin of a functional gonadotropin in the pituitary of the most primitive vertebrate, hagfish. *Proceedings of the National Academy of Sciences*, **107**, 15832–15837.
- Uchida K, Moriyama S, Sower SA *et al.* (2013) Glycoprotein hormone in the pituitary of hagfish and its evolutionary implications. *Fish Physiology and Biochemistry*, **39**, 75–83.
- Van Der Kraak G, Rosenblumm PM, Peter RE (1990) Growth hormone-dependent potentiation of gonadotropin stimulated steroid production by ovarian follicles of the goldfish. *General Comparative Endocrinology*, **79**, 233–239.

- Vladykov VD (1951) Fecundity of Quebec lampreys. *Canadian Fish Culturist*, **10**, 1–14.
- Vladykov VD, Kott E (1979) Satellite species among the holarctic lampreys (Petromyzonidae). *Canadian Journal of Zoology*, **57**, 860–867.
- Wang DS, Jiao B, Hu C *et al.* (2008) Discovery of a gonad-specific IGF subtype in teleost. *Biochemical and Biophysical Research Communication*, **367**, 336–341.
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, **10**, 57–63.
- Wilkinson TN & Bathgate RA (2007) The evolution of the relaxin peptide family and their receptors. *Advances in Experimental Medicine and Biology*, **612**, 1–13.
- Wood AW, Duan C, Bern HA (2005) Insulin-like growth factor signaling in fish. *International Review of Cytology*, **243**, 215–285.
- Woods IG, Wilson C, Friedlander B *et al.* (2005) The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Research*, **15**, 1307–1314.
- Yegorov S, Good S (2012) Using paleogenomics to study the evolution of gene families: origin and duplication history of the relaxin family hormones and their receptors. *PLoS ONE*, **7(3)**, e32923.
- Yegorov S, Bogerd J, Good SV (2014) The Relaxin family peptide receptors and their ligands: new developments and paradigms in the evolution from jawless fish to mammals. *General Comparative Endocrinology*, **209**, 93–105.
- Yoshimura Y, Tischkau SA, Bahr JM (1994) Destruction of the germinal disc region of an immature preovulatory follicle suppress follicular maturation and ovulation. *Biology of Reproduction*, **51**, 229–233.

- Youson JH (1980) Morphology and physiology of lamprey metamorphosis. *Canadian Journal of Fisheries and Aquatic Sciences*, **37**, 1687–1710.
- Zakes Z & Demska-Zakes K (1996) Effect of diets on growth and reproductive development of juvenile pikeperch, *Stizostedion lucioperca* (L.), reared under intensive culture conditions. *Aquaculture Research*, **27**, 841–845.
- Zanandrea SJG (1959) Speciation among lampreys. *Nature*, **184**, 380.
- Zimmermann S, Steding G, Emmen JM *et al.* (1999) Targeted disruption of the *Ins13* gene causes bilateral cryptorchidism. *Molecular Endocrinology*, **13**, 681–691.
- Zou S, Kamei H, Modi Z *et al.* (2009) Zebrafish IGF genes: gene duplication, conservation and divergence, and novel roles in midline and notochord development. *PLoS ONE*, **4(9)**, e7026.

## 1.8. Tables and Figures

Table 1.1. Stages of gonadal development in the lamprey life cycle with complete description and characteristics of each stage. All the stages listed below are used in Figure 1.4, which represents different samples of parasitic and non-parasitic lampreys used for this thesis.

Stage	Stage name	Description	Characteristics based on inferences
1.	Undifferentiated Stage	Primordial germ cells closed within a cluster of follicular cells, no germ cell cysts are visible.	Prior to ovarian differentiation
2.	Cystic Stage- Oocyte differentiation	Transformation of oogonia into oocytes (formation of ovarian follicle.	Ovarian differentiation begins
3.	Growth Stages-Primary oocyte growth and Secondary stage	In primary oocyte growth - Ovarian follicles consists only of a basement membrane and granulosa cells that is confined to the basal or external pole of the oocyte.	Ovarian differentiation begins
4.	Differentiated females Secondary stage (cortical alveoli stage)	Secondary stage consists of basophilic cytoplasm, rich in RNA and phospholipids. Shows no differentiated ergastoplasm and have reduced smooth endoplasmic reticulum. Mitochondria localized around the nucleus and placed near the apical pole of oocyte	Differentiated ovary
5.	Vitellogenesis (early and late vitellogenesis)	Onset of final growth phase marked by reduction in basophilia and migration of cellular organelles towards the periphery of the cytoplasm, chorion is a single layer between the granulosa cells and the surface of the oocyte. The follicle is surrounded by a theca interna.	Different stages of vitellogenesis (early, mid and late stages)
6.	Oocyte maturation	Oocytes are liberated from the follicles. The theca in the animal pole region becomes elevated to form conical projection	Oocyte maturation

Table 1.2. List of genes selected for this thesis that are known to be involved in ovarian differentiation in other vertebrate species and their Ensembl ID's reported in zebrafish and sea lamprey (for gonadal stages description of lampreys, Table 1.1).

<b>Genes</b>	<b>Gonadal stage reported in other species inferred based on my observation</b>	<b>Gene reported in zebrafish in Ensembl browser</b>	<b>Orthologs / Paralogs reported in lamprey in Ensembl browser</b>	<b>Reference to the gene reported in the literature of another organism</b>
Insulin super family genes and their receptors	Probably involved in ovarian development in lampreys (ovarian differentiation stages, 3 & 4 and oocyte maturation, stage 5)	ENSDARG00000094132	ENSPMAG00000002950	Dissen <i>et al.</i> , 2006;
<b>Ligands:</b>		ENSDARG00000058058	ENSPMAG00000008570	Good-Avila <i>et al.</i> , 2009;
Ins, igf1, igf2 and igf3, rln, rln3, insl3, insl5 and insl5a		ENSDARG00000018643	ENSPMAG00000005896	Yegorov <i>et al.</i> , 2014;
<b>Receptors:</b>		ENSDARG00000033307	ENSPMAG00000006274	Li <i>et al.</i> , 2015;
INSR, IGF1R, INSRR, RXFP1, RXFP2, Rxfp2-like, Rxfp3-1, RXFP3-2, RXFP3-3 and RXFP3-4		ENSDARG00000034434	ENSPMAG00000008971	Nobrega <i>et al.</i> , 2015;
		ENSDARG00000071524	ENSPMAG00000006421	Bogerd <i>et al.</i> , 2016;
		ENSDARG00000035350	ENSPMAG00000000636	Hend Al Nafea <i>et al.</i> , (in prep)
		ENSDARG00000069028	ENSPMAG00000001344	
		ENSDARG00000062111	ENSPMAG00000010113	
		ENSDARG00000069246		
		ENSDARG00000059348		
		ENSDARG00000090071		
		ENSDARG00000019660		
		ENSDARG00000032820		
		ENSDARG00000068731		
		ENSDARG00000057410		
		ENSDARG00000022739		
		ENSDARG00000061846		
		ENSDARG00000070780		
		ENSDARG00000103762		

Table 1.3. Details on ovarian transcriptomes of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* used in this project. “Gonadal stage characteristics” refers to the characteristics of the gonad at different ovarian development stage (Table 1.1 & Figure 1.4); ovarian differentiation occurs in stages 2 and 3, and the differentiated ovary occurs in stage 4.

<b>Samples</b>	<b>Species</b>	<b>Length (mm)</b>	<b>Life cycle</b>	<b>Gonadal stage characteristics</b>	<b>Collection date</b>
C13-03	<i>I. castaneus</i>	41	larva	cystic stage, oocyte differentiation	April, 2012
C20-03	<i>I. castaneus</i>	57	larva	cystic stage, oocyte differentiation	June, 2012
IC3	<i>I. castaneus</i>	122	metamorphosis (5/6)	differentiated female, secondary growth stage	July, 2011
IC1	<i>I. castaneus</i>	136	metamorphosis (2/3)	differentiated female, secondary growth stage	July, 2011
IC2	<i>I. castaneus</i>	133	metamorphosis (2/3)	differentiated female, secondary growth stage	July, 2011
N02	<i>I. fossor</i>	97	larva	undifferentiated stage, prior to oocyte differentiation	Summer, 2010
N1-10	<i>I. fossor</i>	59	larva	cystic stage, oocyte differentiation	May, 2011
N17-1	<i>I. fossor</i>	67	larva	primary growth stage of the oocyte	June, 2012
M01	<i>I. fossor</i>	97	larva	differentiated female, secondary growth stage	April, 2012
NC3	<i>I. fossor</i>	106	metamorphosis (4/5)	early vitellogenesis, vitellogenic oocytes	July, 2011
N08	<i>I. fossor</i>	99	metamorphosis (6/7)	late vitellogenesis, vitellogenic oocytes	Summer, 2010
S11	<i>I. fossor</i>	100	metamorphosis (7/8)	oocyte maturation	Summer, 2010

Table 1.4. List of 10 annotated chordate genomes (arranged phylogenetically) used for making customized BLAST database for genome-guided assembly.

**List of species**

---

*Branchiostoma floridae*

*Petromyzon marinus*

*Callorhinchus milli*

*Lepisosteus oculatus*

*Danio rerio*

*Latimeria chalumnae*

*Xenopus tropicalis*

*Gallus gallus*

*Mus musculus*

*Homo sapiens*

Figure 1.1. A phylogenetic tree showing the relationship between major chordate lineages and their divergence times. On the left, the approximate timing of key radiation events is shown and different species are represented, lancelets by *Branchiostoma floridae* (Florida lancelet), lampreys by *Petromyzon marinus* (sea lamprey), sharks by *Callorhynchus milii* (elephant shark), reptiles by *Gallus gallus* (chicken), and mammals by *Homo sapiens* (human). Republished with permission of author, from “The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications” in Genome Research by Smith & Keinath (2015), 25, 1081–1090.

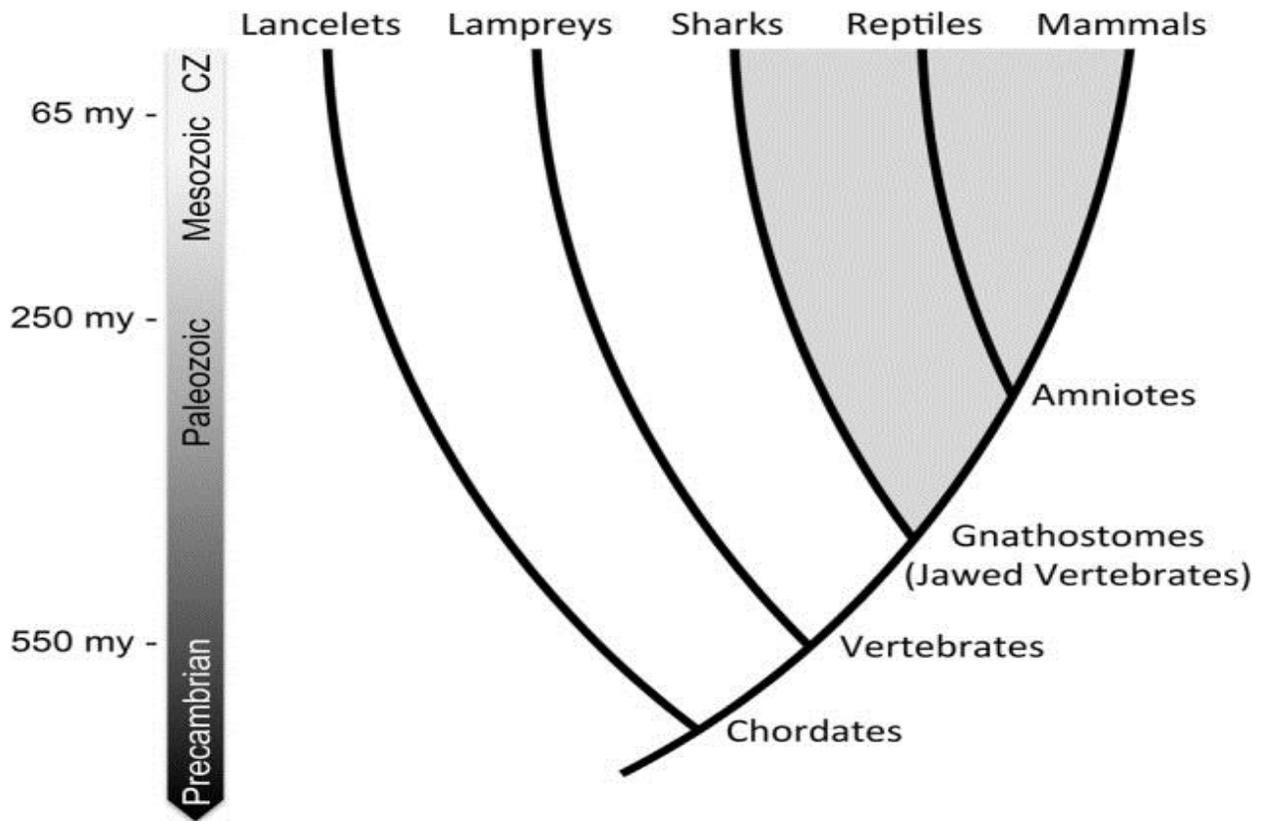


Figure 1.2. Different stages of the lamprey life cycle, which consists of four stages: (1) larval stage or ammocoetes: At this stage, lampreys are blind and filter feeders, (2) transformers: larvae that are in the process of metamorphosis are referred to as transformers, (3) juveniles: lampreys that have completed metamorphosis but are not yet sexually mature are referred to as juveniles; they can be either parasitic or non-parasitic; and (4) adults: lampreys that have reached sexual maturity are known as adults.

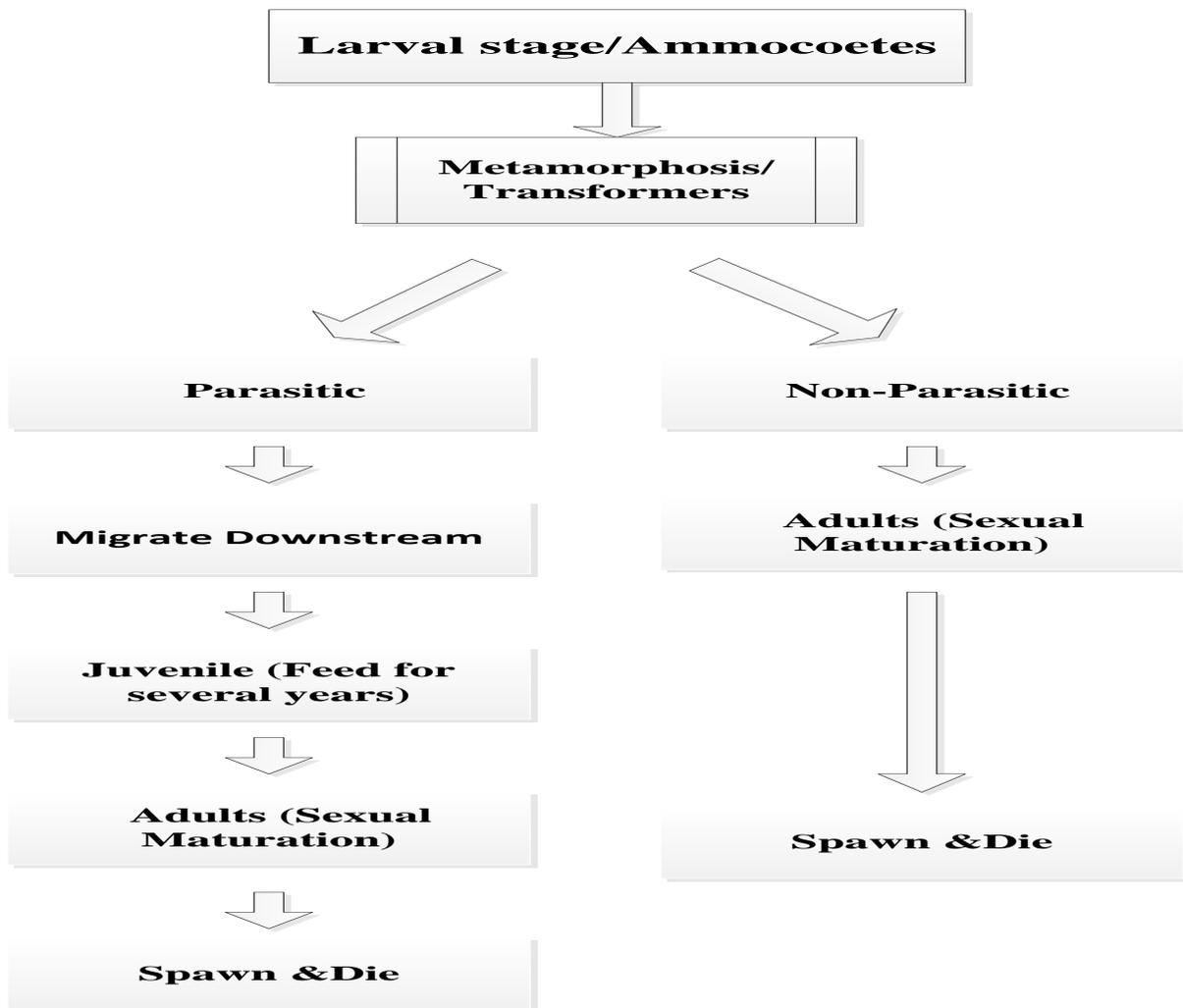


Figure 1.3. Different stages (stages 1 through 7) of lamprey metamorphosis where B represents branchiopore, *F* furrow, *L* lateral lip of oral hood, *P* pupil and, *T* transverse lip of oral hood. Republished with permission of NRC Research Press from “A description of the stages in the metamorphosis of the anadromous sea lamprey, *Petromyzon marinus* L” in Canadian Journal of Zoology by Youson & Potter (1979), 57(9), 1808-1817; with permission conveyed through Copyright Clearance Center, Inc.

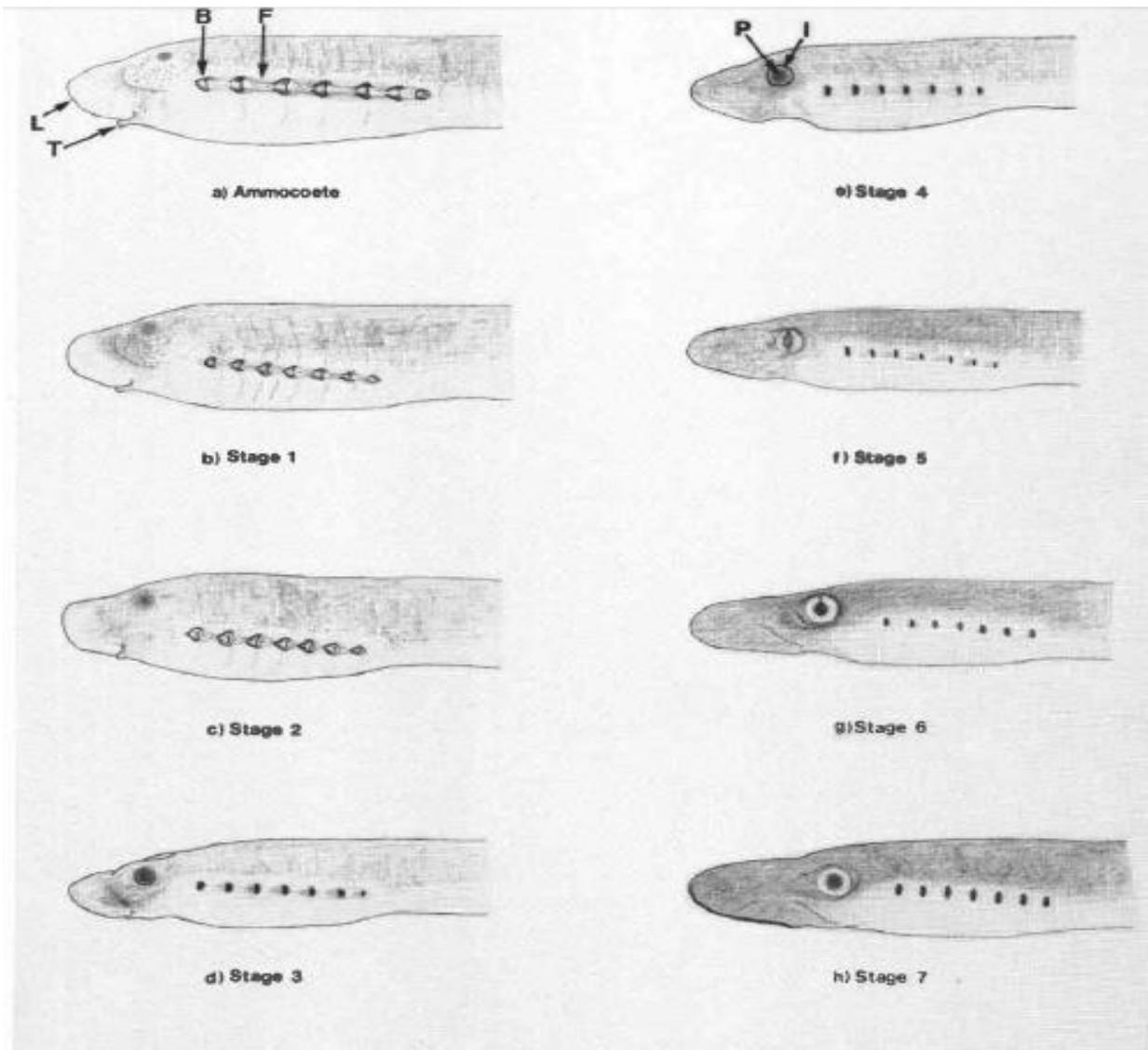
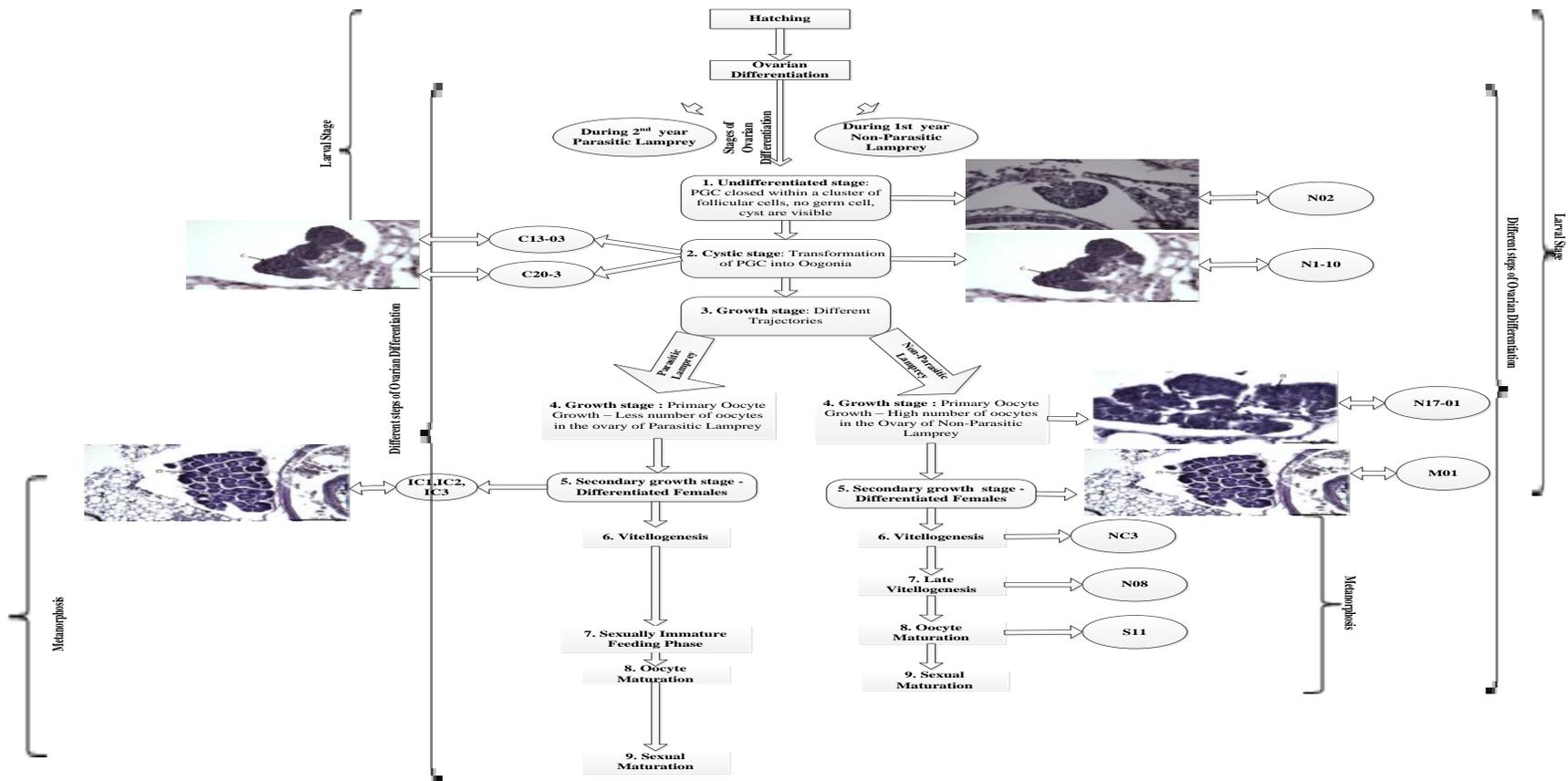


Figure 1.4. Different stages of ovarian differentiation in parasitic and non-parasitic lampreys during the lamprey life cycle which represents different samples of parasitic (C13-03, C20-03, IC1, 1C-2 and IC3) and non-parasitic lampreys (N02, N1-10, M01, NC3, N0-8 and S11). These stages are interpreted based on observations and literature, and the images are taken from Spice (2013). In this thesis, “Gonadal stage” refers to the histological characteristics of the gonad during the lamprey life cycle (larva and metamorphosis). Ovarian differentiation begins during stages 2 and 3 and the ovary is differentiated at stage 4.



## Chapter 2

### RNA-Seq pipeline for mapping and counting genes expressed during ovarian development in lampreys: a genome-guided approach

#### 2.1. Abstract

In this chapter, I present an RNA-Seq pipeline for **mapping** (note that terms that are bolded on first use within each chapter are defined in the glossary in Appendix 2) transcriptomic reads to a reference genome. The pipeline is designed to a) identify genes that map to those annotated in the reference genome; b) count reads using a gene-based counting method; c) characterize the functional ontology of those genes; and d) look for changes in expression of genes based on the associated **Gene Ontology** terms. The primary disadvantage of this pipeline is that reads that do not map to the genes annotated in the reference genome are discarded. These reads primarily fall into two categories a) those reads that do not map to the reference genome because the divergence between the sequenced and reference genomes exceeds the threshold accepted by the mapping wrapper; or b) reads that map to unannotated segments of the reference genome. Normally, the number of reads falling into case b) is considered small, but in the case of sea lamprey, the genome is not fully annotated and many genes are not annotated. To address this, I have employed a second pipeline (discussed in Chapter 3) to help identify both novel and unannotated genes that is useful for non-model species. All the software used in the pipelines (Chapters 2 & 3) are publicly available and run as a stand-alone program or with a Graphical User Interface.

## 2.2. Introduction

The **transcriptome** is the collection of all transcripts present in a cell, a population of cells (e.g., tissue or organ), or an organism at a developmental stage or functional form. Understanding the transcriptome provides an opportunity for interpreting the functional elements of the genome and provides insight into various developmental and disease mechanisms (Wang *et al.*, 2009). Transcriptomic analyses have been essential for understanding gene structure, alternative splicing and post-transcriptional modifications, as well as to characterize differences in gene expression across tissues, stages of development and environments. Thus, transcriptomic studies are essential for functional annotation of genomes as well as for understanding functional expression of the genome in different tissues, stages and/or conditions (Ruan *et al.*, 2004; Wang *et al.*, 2009; Costa *et al.*, 2010).

Recent progress in high-throughput DNA sequencing now allows highly efficient whole transcriptome shotgun sequencing (WTSS) popularly called RNA-Seq, which holds the promise of a more informative and accurate view of the transcriptome at lower costs compared with high-throughput Sanger sequencing (Wang *et al.*, 2009). The abundance of RNA-Seq data has resulted in the generation of new techniques for both mapping and quantifying the transcriptome. “**Mapping**” refers to the total count of reads aligned to a reference genome to identify transcripts. Thus, in simple terms, mapping refers to the process of aligning short reads to a reference sequence, whether the reference is a complete genome, transcriptome, or *de novo* assembly. For well-annotated genomes, transcripts can be easily identified by mapping RNA-Seq reads onto the reference genome of a studied organism. However, for organisms without a reference genome, the reads are assembled *de novo* into **contigs** and mapped to the assembled transcriptome (Conesa *et al.*, 2016). Thus, after sequencing, the raw reads can either be **aligned**

to a reference genome or reference transcripts or the transcripts can be assembled *de novo* without the genomic sequences (Trapnell *et al.*, 2010) depending on the organism and research goals.

RNA-Seq data consists of thousands or millions of short sequence reads of molecules from which transcripts and ultimately the transcriptome is assembled, providing both sequence information and quantification of transcript/gene expression (Ekblom & Galindo, 2011). Sequencing an entire eukaryote genome is often still difficult and costly; however, sequencing the transcriptome of an organism is a more manageable task. RNA-Seq is a suitable technology even for non-model organisms because it does not require prior knowledge of the genes involved in processes of interest or their sequence. For example, RNA sequencing (RNA-Seq) of lake whitefish *Coregonus clupeaformis* determined that adults of the dwarf form overexpressed 51 genes related to energy metabolism, whereas the adults of the normal form overexpressed 116 genes related to protein synthesis; these differences correspond to metabolic and growth differences between the life history types (Jeukens *et al.*, 2010). Goetz *et al.* (2010) also used RNA-Seq in lean and siscowet forms of lake trout *Salvelinus namaycush* to examine differences in gene expression in the liver. Groups of genes involved with immunity, lipid synthesis, metabolism, and transport showed differential expression. In the Midas cichlid *Amphilophus citrinellus*, RNA-Seq was used to examine gene expression in the thick- and thin-lipped forms in Nicaraguan lakes and revealed parallel differences in gene expression between forms, with greater divergence in gene expression in forms from older lakes (Manousaki *et al.*, 2013). Preliminary *de novo* transcriptome analysis of three northern brook lamprey *Ichthyomyzon fossor* (N1-10, N17-1 & M01) and three chestnut lamprey *I. castaneus* (C13-3, C20-3 & IC3) by Spice (2013) identified several genes related to developmental processes that were differentially

expressed in the parasitic chestnut lamprey relative to the non-parasitic northern brook lamprey before, during, and after ovarian differentiation.

Several studies of life history types in other fishes have employed RNA-Seq data analyses to identify candidate genes of interest and then performed quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR) to perform more detailed and confirmatory support for changes in expression of individual candidate genes. Spice *et al.* (2014) used qRT-PCR to study expression of a subset of genes, but in a larger number of northern brook and chestnut lampreys during ovarian differentiation than was possible using RNA-Seq. They targeted eight genes (17 $\beta$ -hydroxysteroid dehydrogenase, germ cell-less, estrogen receptor b, insulin-like growth factor 1 receptor, daz-associated protein 1, cytochrome c oxidase subunit III, Wilms' tumour suppressor protein 1, and dehydrocholesterol reductase 7) to assess if the genes were differentially expressed at three different time points (before, during, and/or after ovarian differentiation). Insulin-like growth factor 1 receptor (*igf1r*) was highly expressed in chestnut lamprey during the oocyte stage (Spice *et al.*, 2014); *igf1r* has been shown to help promote growth of many cells including reproductive tissues such as the ovary via its role in inhibiting apoptosis (Reinecke, 2010). The expression of cytochrome c oxidase subunit III (*coIII*) was higher in northern brook lamprey during the early presumptive male, cystic, and oocyte growth stages (Spice *et al.*, 2014), suggesting that it may act as an inducer of apoptosis.

In the preliminary RNA-Seq study by Spice (2013), *de novo* assembled transcriptomes were generated for one individual from each chestnut and northern brook lamprey and was BLASTed to identify the quality of assembled contigs by using a non-redundant database and was used further for identifying the differentially expressed genes and were assigned Gene ontology terms by using BLAST2GO. At the time of analysis, the sea lamprey genome was not

published, making it difficult to identify many of the genes that were differentially expressed during ovarian differentiation and, moreover, only larval samples were used for comparative analysis.

Thus, to carry this work forward, I analyzed the same ovarian transcriptomic data from chestnut and northern brook lamprey, six from Spice (2013) plus another five transcriptomes from the ovaries of metamorphosing chestnut and northern brook lampreys (Docker unpublished; Table 1.3) to identify genes involved in ovarian development in lampreys that show different developmental trajectories with respect to ovarian differentiation and sexual maturation during and after metamorphosis. Northern brook lamprey stop feeding and undergo vitellogenesis rapidly after metamorphosis; however, in parasitic species like the chestnut lamprey, vitellogenesis proceeds much more slowly (Hardisty, 1971; Hardisty & Potter 1971). Hence, **sexual maturation** occurs early in non-parasitic taxa compared to the parasitic taxa (Figures 1.2 & 1.4); this suggests that there may be different genes that are regulated at different time points in chestnut lamprey versus non-parasitic northern brook lamprey which may help to understand different trajectories of the lamprey life cycle (Section 1.1, Figure 1.4). Thus, to shed light on this biological question, a genome-guided pipeline was designed for mapping and assembling the reads using the sea lamprey *Petromyzon marinus* as a **reference genome**. Both *Ichthyomyzon* and *Petromyzon* belong to the Northern Hemisphere lamprey family Petromyzontidae, and they diverged from each other approximately 7-8 Mya (Brown *et al.*, 1979).

For this analysis, reads from northern brook and chestnut lamprey were trimmed and cleaned, and paired-end reads were then aligned (mapped) to the complete sequence of the sea lamprey genome. Only those reads that mapped to genes annotated in the Ensembl sea lamprey genome (both **novel** and putatively identified genes) were retained and counted using the

program HTSeq, which provides estimates of gene (rather than transcript) counts for each **Ensembl ID** available for sea lamprey. Using this list of gene counts, the gene expression values were first normalized to take into account sequencing depth (which differed between samples), and then samples from the same stage (across species) were pooled to obtain at least three samples per gonadal stage. Then, differences in the log<sub>2</sub>-fold change expression of genes were compared using a Wald test as implemented in the program DESeq2. This generated list of genes that were significantly up- and down-regulated between stages. To assess whether up and down-regulated genes were enriched for specific gene ontology terms and/or pathways, the gene ontology terms associated with the up- and down-regulated genes was obtained and used to find up and down-regulated processes and pathways.

The major goal of this genome-guided pipeline was to identify the suite of genes that are expressed in lampreys during ovarian development by using a reference genome of sea lamprey. It was observed that approximately 13,000 genes are still not annotated or identified. However, with the help of this pipeline, some novel genes (i.e., genes which have **Ensembl ID's** but no gene name) were identified. This list of new genes will help to shed light on different processes of ovarian development, particularly ovarian differentiation and sexual maturation. The resources generated in this pipeline will contribute to our understanding of the genetic control of ovarian development in lampreys. There are some technical difficulties and limitations of this pipeline, most notably, that it only takes the mapped reads of the species and counts only genes that are annotated in the reference genome. Since a major proportion of the reads lie in the unmapped region, they were not utilized for identifying genes that are not yet annotated in the sea lamprey genome. Hence, for this, a *de novo* transcriptome assembly pipeline was designed (discussed in Chapter 3).

## **2.3. Material and methods**

### **2.3.1. Sample collection**

Two species of lampreys, chestnut lamprey and northern brook lamprey (Table 1.3), were collected using a backpack electroshocker (Smith-Root LR-24) per collection permit protocol (for 2011, SCP 05-11 and SECT 73 SARA C&A 11-012; for 2012, SCP 15-12 and SECT 73 SARA C&A 12-009, Spice *et al.*, 2014). All chestnut lamprey was collected from the Rat River in St. Malo, Manitoba, whereas northern brook lamprey was collected from two different rivers, one sample was collected from the Birch River near Prawda, Manitoba, and the other two were collected from McKinnon Creek near Sault Ste. Marie, Ontario. The individuals were sacrificed according to the Animal Use Protocol F11-019. The gonad was removed from each individual and put in liquid nitrogen, and then subsequently stored at -80°C. Qiagen RNeasy Mini Kit was used for extracting the total RNA from each gonad by following the manufacturer's instructions. To remove the contamination from the genomic DNA, a DNase digestion step was incorporated along with the RNase-free DNase Set (Qiagen). Total concentration of RNA was measured and the RNA was stored at -80°C (Spice *et al.*, 2014).

### **2.3.2. RNA sequencing**

In brief, RNA-Seq is a process in which messenger RNA (mRNA) is reverse-transcribed to complementary DNA (cDNA) fragments wherein adaptors are attached to one or both ends of these fragments. By using a high-throughput sequencing technology, cDNA is then sequenced from one or both ends (Hudson, 2008; Wang *et al.*, 2009; Tucker *et al.*, 2009; Ekblom & Galindo 2011). Samples were sent for RNA sequencing to the Oklahoma Medical Research Foundation (Oklahoma City, Oklahoma) and Hussman Institute for Human Genomics (Miami, Florida). For

sample numbers IC3, M01, and, N1-10 (Spice, 2013), IC1, IC2, NC3, and, S11 that were sent to the Oklahoma Medical Research Foundation, messenger RNA (mRNA) was isolated from ribosomal RNA and small modulating RNA using a poly-T bead step. For the samples C13-3, C20-3, and, N17-1 (Spice, 2013) that were sent to the Hussman Institute for Human Genomics, mRNA was isolated using a poly-A step and non-normalized libraries were prepared using the Illumina TruSeq DNA Kit and Epicentre ScriptSeq Kit. Sequencing was performed in both forward and reverse directions using 75-100 base pair (bp) paired-end sequences. Truseq sequence adaptors were used on both ends on an Illumina Hi-Seq 2000 for paired-end reads. The adaptors were 60 bp long and the insert size, which is normally the stretch of sequence between the paired-end adaptors, was ~300 bp (e.g., 2x100 bp reads + 100 bp unsequenced middle piece). Hence, the inner mate distance, which is the gap left between the two fragments after sequencing, was calculated. The total RNA sequence fragment size was 200-1000 bp which, after subtracting the adaptors from both each end, left RNA fragments of 80-880 bp, of which 75-100 bp was sequenced from each end. Thus, the “inner mate distance = insert size – fragment size,” which comes out to be 220 bp (see below).

Total RNA fragment size	(200 to 1000) =====
Fragment size (Total RNA fragment size – adaptors both ends)	(200-60-60 = 80); (1000-60-60 = 880) =====
Insert size	(300 bp) =====
PE reads	R1----->                      <-----R2
Inner mate distance	300-80= 220 and 300-800= -500=====

## **2.4. Pipeline designed for transcriptomic data analysis in the presence of a reference genome**

To determine what suite of genes are expressed at different ovarian developmental stages of lampreys, I have taken all the RNA reads, removed adaptor sequences and removed low-quality sequences and then mapped these reads to the publicly available sea lamprey genome (Smith *et al.*, 2013). For this thesis, the annotation of sea lamprey *Petromyzon marinus*, version 87.7 was downloaded from the Ensembl Genome browser.

### **2.4.1. Mapping and counting the number of times a gene is expressed using a reference genome - Galaxy platform**

To reduce the complexity of transcriptomic data analysis, a user-friendly web based open source platform known as Galaxy (Giardine *et al.*, 2015) was used for performing several components of the RNA-Seq data analysis. Galaxy allows the user to create and manage reproducible workflows that can be used with multiple kinds of data sets and the workflow can be easily generated and shared with other users and published. It is a platform in which different tools of data analysis and common bioinformatics software such as BLAST, PHYLIP, Bowtie, and TopHat are made available and can be accessed through the Galaxy interface along with many file conversion formats, statistics and data manipulation. The Galaxy Project uses a web interface to cloud computing resources to bring command-line-driven tools such as TopHat, Trimmomatic, HTSeq, and Trinity to users without UNIX skills through the web and the computing cloud. It can also be run on a local server install.

The **genome (FASTA file) and annotation (GFF file)** for the species of interest were uploaded. For my project, I used the genome assembly of sea lamprey *Petromyzon marinus* provided by the

lamprey consortium available on Ensembl Genome browser. The database version 87.7 of the genome/annotation was used. The genome was sequenced to a total of 5.0X genome coverage and the size is 1.92 gb. The details about the summary and statistics are given in Table 2.1.

#### **2.4.1.1. Gene annotation**

After uploading the genome file, the file was sorted (because some of the downstream tools may only recognize the file format, if it's sorted properly). For this, a few simple steps were used:

- a) The 'FASTA-to-Tabular' was run on the genome file under the tool tab "Convert Formats".
- b) 'Sort' tool option was used (in the 'filter and sort' section) on the converted tabular file.
- c) The sorted tabular file was converted back to FASTA format using the 'Tabular-to-FASTA' tool. Once the files were sorted, several assessments were done for RNA-Seq analysis.

#### **2.5. RNA-Seq analysis tools**

All the RNA-Seq analysis software in Galaxy are put together under the common tab 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: RNA Analysis' and, 'NGS: SAMtools'. It also contains some important Phenotypic Association packages such as g: Profiler, DAVID which allows performance of different kinds of orthology searches, gene id conversion tools, etc. The first step in RNA-Seq analysis was to prepare and edit the raw reads obtained from Illumina for downstream analysis.

### **2.5.1. Preparation of raw data**

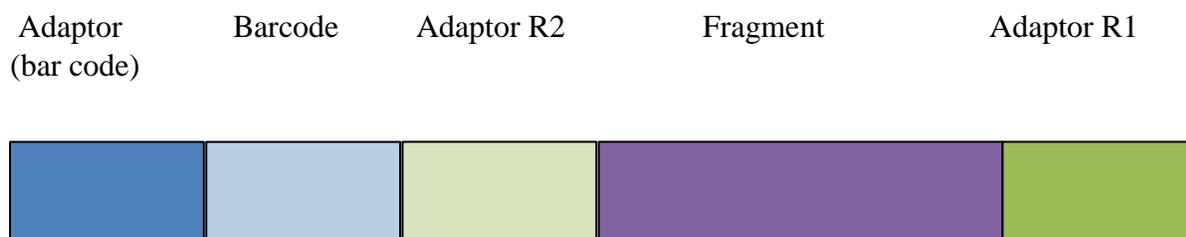
The raw RNA-Seq paired-end (PE) reads for both chestnut and northern brook lampreys were obtained from Illumina Hi-seq. Before analyzing the raw reads or sequences, it is very important to perform quality control checks on the raw reads to ensure that the data is free from any biases which may affect the ability to draw important conclusions. Generally, a quality control report (QC) is generated from the sequencers as part of the analysis process, but this only identifies the problems generated by the sequencer itself. Hence, a program known as FASTQC (Andrews, 2010) was used to evaluate the reads from each species (both in forward and reverse strand) based on quality score.

This program provides a quality control (QC) check report on the raw sequence data generated from high throughput sequencing pipelines which are used to identify problems or errors which originate either in the sequencer or in the starting library material. There could be several reasons for this: poor sequencing quality or bad base calling, contamination during sequencing the library, or a lot of repetitive sequences in the genome. Hence, FASTQC provides a modular set of analyses, which gives a quick view of the data and the quality of the sample. This allows the user to set a quality of read threshold and then only reads above this threshold are accepted for analysis. For example, it provides an idea about the quality of the sample and allows the user to check whether the raw reads are of good or bad quality and whether they need any further refinement or trimming. Thus, it gives basic summary statistics of the raw reads, per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences and overrepresented content.

FastQC is a very user-friendly program and can be run in one of two modes. It can either be run as a stand-alone interactive application for the analysis of a small numbers of FastQ files, or it can be run in a non-interactive mode for processing a larger number of files or datasets (Figures 2.3-2.3 j, Table 2.2).

### 2.5.2. Removal of adaptor sequences

After evaluating the quality of raw reads, the next step is to remove the adaptor sequences from the paired-end reads by using a tool known as Trimmomatic (Bolger *et al.*, 2014). During sample library preparation, the RNA is fragmented and size-selected. Because the size selection captures a range of fragment sizes (200-1000 bp) and 75-100 bp of sequence was obtained from both the right and left hand adaptors, fragments that are smaller than ~100 bp could have the entire fragment sequenced plus part of the adaptor sequence (as shown below where R1 and R2 are adaptors). As such, for small fragments, the sequence reads may contain flanking adaptor sequences that need to be removed prior to transcript assembly. In addition to the two adaptor sequences, barcode and barcode adaptor sequences are also ligated onto the fragment, to allow differentiation of samples:



To remove the adaptor sequences, software programs are used to either a) identify and remove overly represented DNA sequences or, preferably, b) remove the adaptor sequences given as input to the program. For this process, I used the program Trimmomatic, which is a user-friendly

tool designed for trimming adaptor sequences for both paired-end (PE) and single-end (SE) reads. Trimmomatic can be run either through the Galaxy user interface, or DOS/Linux command line. I used the command line version. Trimmomatic produces four output files for paired-end reads a) a file containing all paired (both left and right) sequence reads and, b) another file containing all the unpaired reads in separate files (both left and right). Only the former two files will be used for the remainder of the analysis. The unpaired read files were deleted and then the paired read files (both right and left) were again used to check the quality report by using FASTQC. The output generated from FASTQC provided a quality report about the trimmed files generated from Trimmomatic.

### **2.5.3. Format conversion**

The files obtained from Trimmomatic were “groomed” using the tool, FASTQ Groomer (Blankenberg *et al.*, 2010). This tool is available under the “Next generation (NGS): quality check (QC) and manipulation” section within Galaxy. The FASTQ groomer tool is used to support and convert the FASTQ data into other FASTQ variant file formats for downstream analyses with different software packages in Galaxy.

### **2.5.4. Mapping and assembling the reads**

Once the input data is ready, the next steps in the pipeline are to map the reads to the reference genome and estimate the number of times each transcript or gene is expressed. The two major objectives of mapping in RNA-Seq experiments are a) to identify novel transcripts which are currently not identified in both the samples and reference genome. For this research objective, a

*de novo* pipeline was designed which is discussed in Chapter 3) and b) to determine transcript abundance based on the coverage depth in the mapping.

There are numerous programs available that have been developed to map reads to a reference sequence but programs usually vary in algorithms and speed. For **mapping**, there were two reference genomes available for lampreys on UCSC and Ensembl, the sea lamprey *Petromyzon marinus*, and the Arctic lamprey *Lethenteron camtschaticum* (also known by its older name, the Japanese lamprey *Lethenteron japonicum*). After preliminary analysis using both sea lamprey and Arctic lamprey as reference genomes, the reference genome of the sea lamprey was selected for mapping based on the higher percentage of mapped reads. The percentage of *Ichthyomyzon* spp. reads that mapped to the sea lamprey genome was 30-40% as compared to only 15% using the Arctic lamprey as a reference (Table 2.3). This finding is consistent with phylogenetic reconstructions that demonstrate that the genus *Ichthyomyzon* is more closely related to *Petromyzon* (Potter *et al.*, 2015, Figure 2.1).

After the selection of a reference genome, the paired-end reads were mapped to the reference genome using the program TopHat (Trapnell *et al.*, 2009). TopHat was run for both the forward and reverse reads for all samples; it performs alignment and *ab initio* splice variant estimation based on a suite of alignment parameters chosen by the user. There are many parameters that can be modified depending on the experimental design. Some of the default parameters should be modified by the user, such as the estimated mean inner distance between mate pairs, while other parameters are more difficult to estimate and the default values have been optimized for human transcriptomic data and for some model species. For this analysis, I experimented with changing the maximum number of mismatched bases between the transcriptomic and reference sequences, and settled on a mismatch value of 4, which is slightly

higher than the default value of 2, to account for divergence between the *Ichthyomyzon* and *Petromyzon* genomes. Thus, the mapping was performed using paired-end reads, a mean inner distance between mate pairs of 220 bp, and a final alignment mismatch score of 4 via the Galaxy interface. Once the TopHat run was completed, five output files were generated:

- a) alignment summary (it gives the number and percentage (%) of reads successfully aligned to the genome)
- b) insertions (a list of all insertions relative to the reference genome)
- c) deletions (a list of all deletions relative to the reference genome)
- d) splice junctions (a list of splice sites, intron-exon boundaries)
- e) accepted hits (a list of all the read alignments in BAM format (compressed binary version of the Sequence Alignment Map (SAM) format). It contains all the relevant information about the percentage of mapped reads with respect to the reference genome and annotation. This is the file that was carried forward to the next step for further analysis) and,
- f) unmapped hits (a list of all the read alignments in BAM format which were unmapped with respect to the reference genome).

This approach is also known as a genome-guided assembly wherein a reference genome is used to guide the assembly process. This works well for species with sequenced genomes or for those that are very closely related to a species with a sequenced genome.

### **2.5.5. Counting the number of reads aligned to reference genome**

Once the assembly of transcripts is completed, the next step is to use the program HTSeq (Anders *et al.*, 2015), which counts of how many times each gene is estimated to have been

expressed based on user-defined methods of dealing with overlapping reads. The “Accepted Hit File” generated from TopHat is used as input to the program.

HTSeq was built on a Python framework that can be run as a stand-alone script from a command shell, or through the Galaxy user-interface. It offers three different modes of counting genes based on criteria for dealing with overlapping reads: a) reads can be combined by union; b) intersection-strict; and c) intersection-nonempty. If none of these fit the user’s needs, users with a reasonable knowledge of Python can write their own scripts. The HTSeq program has inbuilt parsers for reference sequences (FASTA), short reads (FASTQ) and short-read alignments (the SAM/BAM format and some other formats), as well as for genomic features, annotation and score data (GFF/GTF, VCF, BED and Wiggle). The script *htseq-count* is widely used for RNA-Seq analysis and it allows processing of paired-end data directly without sorting the SAM/BAM file by read name first. For calculating the counts, *htseq-count union* script was used. HTSeq counts the aligned reads that had overlapping exons by taking the SAM/BAM file (generated from TopHat) and a GTF or GFF file (reference genome of sea lamprey downloaded from Ensembl) as gene models. It estimates the number of times each gene is expressed, and uses a genome **annotation** file (the sea lamprey annotation file) to reference the mapped reads to a specific genomic location and thus estimates the total number of times a gene is expressed in the sample and output the unique **Ensembl ID** associated with it.

For RNA-Seq data, the term ‘Features’ used in this program refers to the genes, where each gene is considered as the union of all its exons or each exon can also be considered as a feature. If the read contains more than one feature, then a read or read pair is counted as ‘Ambiguous’ (hence, not counted for any feature). Because the script is designed for gene-based differential analysis, it counts only the reads mapping unambiguously to a single gene,

whereas the reads which overlap and align to multiple positions with more than one gene are disposed of and counted as ‘No feature’. This is one of the biggest limitations of this program; because the reads with ‘No feature’ have no Ensembl ID’s, they may (probably) contain **novel** genes. However, because they are not present in the current annotation, they are not counted. Unfortunately, these reads are thrown away by this program and not put into a separate file. Thus, HTSeq is a program that is only useful for counting known annotations.

Hence, the output generated from HTSeq is a list of all the lamprey associated Ensembl gene ID’s and the number of times each gene is estimated to be expressed in that sample (thus, many read or read pairs had no expression (Table 2.4). All the Ensembl gene ID’s were then converted into their associated gene name by using a package known as BioMart present at Ensembl (Zhang *et al.*, 2011). These counts files can be used further for differential gene expression analyses by using DESeq2 (Love *et al.*, 2014) or edgeR (Robinson *et al.*, 2010).

### **2.5.6. Differential analysis of count data**

The count data generated from HTSeq for both the chestnut and northern brook lampreys are in tabular format, which provides important information about the number of sequence fragments that have been assigned to each gene. One of the main goals of the RNA-Seq data analysis is to find the genes that are differentially expressed across groups of samples. This accounts for measuring all the specificities of count data such as non-normality and variance dependence on the mean, but other than that, one of the biggest challenges of high throughput sequencing experiments is the small number of samples with two or three replicates per condition. Due to smaller sample size, statistical methods often lack power, which leads to high uncertainty for calculating within-group variance. In such a condition, this weakness can be

overcome by pooling information across genes. So, all the HTSeq count files generated for both parasitic and non-parasitic lamprey samples were divided into three groups based on the stage of ovarian development:

- a) early gonadal stage,
- b) mid-gonadal stage and,
- c) late gonadal stage (Table 2.5)

Based on this information, samples from chestnut and northern brook lampreys were pooled together and the count tables that were generated from the HTSeq data for each stage were grouped as input to the program DESeq2 (Anders & Huber, 2010). DESeq2 is available as a Bioconductor package in R but can also be run through Galaxy. One of the important features of this tool is that it can handle different experimental designs. The tool allows selection of a multi-factorial experiment with multiple levels of each factor from which to test for differential expression of genes across factors. Hence, one can input several other secondary factors that might influence experimental conditions, but the final output is based on the changes in genes due to the primary factor in the presence of secondary factors. Differential log<sub>2</sub>-fold changes are calculated on primary factor level 1 versus factor level 2. Hence, log<sub>2</sub>-fold change between two groups - 'Treatment' and 'Untreated,' is the fold change of input factor 1 'Treated' samples versus input factor 2 'Untreated'. So, it is very important to input the correct order of factor levels and the values correspond to up or down regulations of genes in treated samples. The output generated from this program is a tabular file which contains seven different columns, namely gene identifiers, mean normalized counts, averaged over all samples from both conditions, the logarithm (to base 2) of the fold change, standard error estimate for the log<sub>2</sub>-fold change estimate, Wald statistic, *p*-value for the statistical significance of this change and *p*-value

adjusted for multiple testing with the Benjamini-Hochberg procedure which controls false discovery rate (FDR) and visualized results as PDF output. This PDF output file gives valuable information about the rate of variance across different samples, their sample-to-sample distances, dispersion estimates and fold change.

Based on the order of input, this program gives an extensive list of genes which are up- and down-regulated and normalized counts of all the genes expressed across different developmental stages

### **2.5.7. Gene Ontology assignment by GOrilla (Gene Ontology enRiChment anaLysis and visuaLizAtion tool)**

The list of Ensembl ID's generated from DESeq2 was converted into a list of **orthologous** genes of zebrafish with the aid of the program g: Profiler (<http://biit.cs.ut.ee/gprofiler/>). It is a web based tool widely used for performing different kinds of bioinformatics analyses such as functional profiling, orthology search, gene id convertor and expression similarity search. **Gene Ontology** is a major resource for understanding differences in gene enrichment and is widely used to provide a structure that organizes genes into biologically related groups according to three categories of gene annotations: Biological Process, Cellular Component and, Molecular Function (du Plessis *et al.*, 2011). It classifies genes and their products into different categories in which they are acting, and provides a vocabulary to describe a gene and its products. Hence, the GO database is a hierarchically ordered set of terms for describing genes where each annotation or ontology forms a directed graph in which each node is associated with a GO identifier (or GO term). If a gene is annotated with any of the GO terms it is also annotated with all ancestral GO terms and thus provides an opportunity to group large-

scale gene-expression data into biologically related groups. In lampreys, many of the genes have no associated Gene Ontology (GO) term. Therefore, the lamprey Ensembl ID's were converted into zebrafish orthologous Ensembl ID's (Table 2.6) and was then used as input to the program GOrilla (Eden *et al.*, 2009), a web based tool that identifies the global pathways linking a set of genes and provides a means to visualize the GO terms associated with the genes. It is a very user-friendly tool and is publicly available at: <http://cbl-GOrilla.cs.technion.ac.il>. The user can simply input the list of gene names in any of the formats such as gene symbol, protein RefSeq, Uniprot, Unigene or Ensembl without providing the explicit target and background sets. In the case of genomic data, this is particularly useful because genomic data may be naturally represented as a ranked list of genes by level of expression or of differential expression.

GOrilla currently supports the following organisms: human, mouse, rat *Rattus norvegicus*, yeast *Saccharomyces cerevisiae*, zebrafish, fruit fly *Drosophila melanogaster*, roundworm *Caenorhabditis elegans* and mouseear cress *Arabidopsis thaliana*. It uses a statistical approach to assign GO terms to top ranked genes based on '*p*-value', 'FDR *q*-value' and, 'Enrichment'. It computes an exact *p*-value for the observed enrichment by taking threshold multiple testing into account without the need for simulations. 'FDR *q*-value' is the corrected *p*-value used for ranking genes using the Benjamini & Hochberg (1995) method. "Enrichment =  $(b/n) / (B/N)$ ", is computed on four factors (N, B, n, b) where 'N' refers to the total number of genes; 'B' is the estimate of total number of genes associated with a GO term; 'n' is the number of genes provided by the user and; 'b' is the number of intersecting genes.

The output of the enrichment analysis is visualized as a hierarchical structure, providing a clear view of the relations between enriched GO terms. The results can be easily visualized with a built-in easy to use graphical representation of the GO hierarchy, emphasizing the enriched

nodes. Then, the Gene Ontology terms associated with up- and down-regulated genes of zebrafish were then converted back to lamprey orthologs by g: Profiler for identifying the genes that were up- and down-regulated in lampreys. This pipeline provided an estimate of genes that are up- and down-regulated in lampreys and was particularly useful for assigning Gene Ontology terms because in lampreys many genes do not have a Gene Ontology term assigned. Hence, this pipeline (Figure 2.2) was used to assign GO terms in lampreys. It can also be further used to group genes related to specific processes (for example, genes related to reproduction can be put under the category “Reproduction”). Hence, the associated GO term related to lampreys can be used to update the existing information (Tables 2.9 a, 2.9 b & 2.10).

#### **2.5.8. REVIGO (reduce + visualize Gene Ontology)**

The annotations of up- and down-regulated genes generated by GOrilla consist of multiple repetitive and co-occurring GO terms. For inferring the actual relationship between different GO terms from GOrilla network map, a freely available web based server REVIGO was used. It summarizes and visualizes the data generated from GOrilla by using a clustering process like hierarchical clustering methods such as the neighbor joining approach (Saitou & Nei, 1987) and forms groups of highly similar GO terms based on the *p*-values, enrichments or user input values alongside the GO terms. Hence, it takes the GO list generated from GOrilla and summarizes the results by removing the functional redundancies. The remaining GO terms are then visualized in two-dimensional plots, interactive graphs, tree maps and tag clouds via the Web browser. The graphs can also be exported as a XGMML file, or opened in the stand-alone Cytoscape program (Cline *et al.*, 2007) via Java Web Start in an offline mode. Hence, it allows generation of high resolution, publication-quality images.

## **2.6. Results**

### **2.6.1. Assessment by FASTQC**

All the samples of chestnut and northern brook lamprey samples were analyzed by FASTQC to assess the quality of raw reads. The summary statistics generated “Pass Result” for all the samples of lampreys and indicated that all the samples had good quality scores (Figures 2.3a – 2.3j).

### **2.6.2. TopHat mapped percentage and coverage graph**

Samples of chestnut and northern brook lampreys were mapped to the reference genome of sea lamprey; the TopHat mapped percentage was in the range of 12-45%. Sample numbers IC3, N1-10 and NC3 showed the lowest mapping percentage of 12.1, 15.5 and 17.6% (Table 2.3). This indicates that many reads did not map to the reference genome. A cursory examination of the unmapped hits from the TopHat unmapped reads file indicated that a substantial fraction of the unmapped reads was bacterial in origin (especially from the one company that didn't select poly-A tails, Section 2.3.2). Despite this, the cursory examination of the contigs generated from the mapped reads indicated that the samples had a good coverage when mapped to a sea lamprey genome reference (Figures 2.4 a, b, & 2.5 a, b).

### **2.6.3. Low number of genes by HTSeq**

The mapped “Accepted Hit” BAM files of all the samples was then used as input to HTSeq to count the number of genes present in all the samples with respect to the reference genome. The maximum number of Ensembl ID's reported by HTSeq for all chestnut and northern brook lampreys was 13,114, out of which only 6,913 had Ensembl ID's with an

associated gene name; the remaining 6,201 had Ensembl ID's with no gene name and are reported as "novel protein coding gene" (table available on request). Sample number N02 and C20-3 had the lowest percentage of HTSeq gene counts, indicating that only 7.11 and 15.57% of the genes had Ensembl ID's. All other samples reported gene counts in the range of ~22-45% (Table 2.4).

#### **2.6.4. Differential gene expression between species**

All the samples of chestnut and northern brook lampreys were pooled into three different groups: a) early stage, b) mid-stage, and, c) late stage based on the stage of ovarian development (Table 2.5) to form a set of replicates. The DESeq2 analysis identified 2,864 and 1,738 up- and down-regulated lamprey genes, respectively, with positive and negative log<sub>2</sub>-fold changes at an adjusted *P*<sub>adj</sub> value <0.05 between the three different stages of ovarian development (Tables 2.6, 2.7 & 2.8). It was observed that during the early gonadal stage, IC1 and IC2 of chestnut lamprey and M01 of northern brook lamprey showed similar values in the principal component analysis (PCA) plot but interestingly IC3 at metamorphic stage 5/6 of chestnut lamprey clustered itself with other samples of northern brook lamprey (NC3, N08 & S11) of late gonadal stage. The histogram based on *p*-values showed that IC3 was not uniformly distributed with other samples IC1, IC2 and M01 of the same gonadal stage (Figures 2.6-2.10).

#### **2.6.5. Differential gene expression within species**

The PCA plot showed that parasitic chestnut lamprey IC3 of gonadal stage 4 showed differential expression pattern within parasitic species IC1 and IC2 of same gonadal stage (Figures 2.6, 2.7 & 2.8).

### **2.6.6. Orthologs of up- and down-regulated genes**

DESeq2 reported that 2,864 genes were up-regulated and 1,738 genes were down-regulated in lampreys at three different time points of ovarian development. These up- and down-regulated genes were converted into orthologs of zebrafish by the program g:Profiler. This program reported that some of the lamprey genes have more than one orthologs in zebrafish, hence 5,401 and 2,730 orthologs of up- and down-regulated genes of zebrafish was generated for 2,864 up- and 1,738 down-regulated lamprey genes (tables available upon request). The list of up- and down-regulated orthologs were used as input to GOrilla. The database assigned Gene Ontology terms to 2,609 up- and 2,730 down-regulated orthologs of zebrafish, roughly half of the genes had no Gene Ontology terms in the Gene Ontology Association (GOA) database and were discarded (Tables 2.9 a & 2.9 b).

### **2.6.7. Annotations of up- and down-regulated genes by GOrilla**

GOrilla provided 67 annotations to 2,609 up-regulated genes and 8 annotations to 2,730 down-regulated genes of zebrafish (Table 2.10). The 'Scatterplot view and Table' output of GOrilla generated a list of GO terms based on the '*p*-value', FDR '*q*-value' and 'Enrichment of genes'. Processes related to biological regulation, reproduction, neurological system, sensory perception and cellular recognition were found to be significantly expressed in up-regulated genes of zebrafish (Figures 2.11-2.14). On the other hand, down-regulated genes were mostly involved in processes such as detection of stimulus, biological regulation and signaling pathways (Figures 2.15-2.19). Description about *p*-value, FDR *q*-value and Enrichment is given in section 2.5.7.

### **2.6.8. REVIGO analysis**

GOrilla annotations for up- and down-regulated zebrafish genes were visualized in REVIGO to find the list of co-occurring related GO terms. Processes such as egg coat formation, G protein coupled receptor signaling pathways, glucuronate metabolism, membrane raft organization, biological regulation, multicellular organismal processes, localization within membrane and response to stimulus were found to be significantly activated in up-regulated genes. The ‘Interactive graph’, ‘Treemap’ and ‘Tag cloud view’ of REVIGO suggested that all these processes have interrelated co-occurring GO terms mostly related to “Reproduction and Metabolism” (Figures 2.12 & 2.13), whereas down-regulated genes had co-occurring terms related to basic cellular processes such as detection of stimulus and response to stimulus. (Figures 2.17 & 2.18).

### **2.6.9. Up- and down-regulated genes with assigned annotations in lampreys**

One of the main goals of this pipeline was to assign and find the GO terms of lamprey genes expressed during ovarian development. It was observed that GOrilla provided 67 annotations to 2,609 up-regulated genes and 8 annotations to 2,730 down-regulated genes of zebrafish (Figures 2.11–2.14 & 2.15-2.19); these genes were converted back to orthologs of lampreys to find the associated Gene Ontology terms (Table 2.10, full table available upon request). In lampreys, 1,552 up-regulated and 750 down-regulated Ensembl ID’s were reported with Gene Ontology terms; only some of these Ensembl ID’s had gene names (tables available upon request).

### **2.6.10. Normalized counts for different gonadal stages**

Different comparisons were made between chestnut and northern brook lampreys from early, mid and late gonadal stages (Table 2.5) to obtain the normalized counts of all the genes across different gonadal stages in lampreys. These counts were a list of Ensembl ID's reported in lampreys and can help to design future studies based on the list of genes (tables available upon request).

### **2.6.11. Differential analysis of up- and down-regulated genes of lampreys**

Some genes like *zona pellucida glycoprotein 3a*, *adrenoceptor alpha*, *oxytocin receptor*, *calcitonin receptor*, *forkhead box l2*, *SRY (sex determining region Y)-box 2* and *neuropeptide Y receptor Y8b* were found to be up-regulated in lampreys whereas genes like *estrogen receptor 2*, *thyroid stimulating hormone receptor*, *wilms tumor 1a* and *Wilms tumor 1 associated protein* were down-regulated during ovarian development (tables available upon request). The HTSeq count of these genes were obtained for chestnut and northern brook lampreys for identifying the genes that were differentially expressed during different stages of ovarian development. It was not possible to plot samples from late stages of ovarian development because, for chestnut lamprey, there were no samples available which were undergoing vitellogenesis and sexual maturation (Table 1.3). Thus, only the pre-vitellogenic stages (2-4) were used for comparative analysis. Histograms were generated by Prism GraphPad where the x-axis represented different gonadal stages and the y-axis represented the HTSeq counts for both the species (Figures 2.20-2.22).

### **2.6.12. Analysis of insulin family genes during different stages of ovarian development**

To test the functionality of the reference-guided pipeline, Ensembl ID's of genes belonging to the insulin family were obtained for both chestnut and northern brook lampreys. It was reported that eleven Ensembl ID's belong to insulin family genes (tables available upon request), but only three (ENSPMAG00000001344, rxfp3; ENSPMAG00000006421, rxfp3-3 and ENSPMAG00000010113, rxfp3-1) have gene names in the current sea lamprey genome, ENSPMAG0000000636 is reported as novel rxfp1 but is almost completely annotated. The remaining seven Ensembl ID's (ENSPMAG00000008971, ENSPMAG00000002950, ENSPMAG00000007101, ENSPMAG00000006274, ENSPMAG00000005896, ENSPMAG00000007436, and ENSPMAG00000008570) have no gene name and are unannotated. Based on previous genomic locations and bioinformatics data mining, putative gene names of these Ensembl ID's and their HTSeq count was obtained (table available upon request). These genes were plotted against each other (chestnut versus northern brook lampreys) across different gonadal stages for comparative analysis and differential gene expression (Figures 2.23 & 2.24; tables available upon request).

### **2.7. Discussion**

This study was designed to identify suites of genes expressed during ovarian development in lampreys. During the lamprey life cycle, non-parasitic lampreys generally undergo ovarian differentiation in the first year following hatch and stop feeding after metamorphosis; hence, they are smaller at the time of ovarian differentiation and at sexual maturity and exhibit lower larval and adult fecundity (Neave *et al.*, 2007). On the other hand, in parasitic lampreys, ovarian differentiation generally happens in the second or third year

following hatching (Hardisty, 1969, 1970; Barker & Beamish, 2000), individuals achieve larger adult size and are more fecund (Vladykov, 1951; Section 1.1.3.1). This suggests that differential regulation of ovarian development and **sexual maturation** occurs in the two species. A previous study by Spice *et al.* (2014) employed RNA-Seq data analysis of the gonads in larval stages of *I. castaneus* and *I. fossor*; they identified differentially expressed genes such as 17 $\beta$ -hydroxysteroid dehydrogenase (*hsd17 $\beta$* ) and daz-associated protein-1 (*dazap1*) in the late stages of ovarian differentiation and, in differentiated females, insulin-like growth factor 1 receptor (*igf1r*) was highly expressed in chestnut lamprey during the oocyte stage whereas cytochrome c oxidase subunit III (*coIII*) was highly expressed in northern brook lamprey during the early presumptive male, cystic, and oocyte growth stages. This was the first study in lampreys to identify genes that were involved in ovarian differentiation and shed light on differences in development and gene expression in parasitic and non-parasitic lampreys.

At the time of transcriptomic analysis, a reference genome was not available and Spice (2013) employed a reference-free inference. Here, I present a re-analysis of the gene expression in lamprey gonads during larval stages, and present results of the gene expression during post-larval metamorphosing individuals in both parasitic and non-parasitic taxa using both a reference-guided analysis (Chapter 2) and reference-independent approaches (Chapter 3).

In this chapter, a reference-guided pipeline was used to map the raw reads to the sea lamprey genome. There were two reference genomes available, the sea lamprey genome and the Arctic lamprey genome, but final selection was made based on the phylogenetic relatedness (Figure 1.1), and the mapping percentage was higher (30-40%) for the sea lamprey compared to the Arctic lamprey genome (15%). However, Smith *et al.* (2013) calculated that the lamprey genome has 26,046 genes spread across 84 pairs of chromosomes, but the annotation of the sea

lamprey genome available at Ensembl has 13,114 Ensembl ID's. Using this reference, it was observed that 12-45% of the reads mapped to the genome, while 55-88% of the genes did not map. There could be several reasons for this a) selection of the reference genome, b) sequencing protocol, c) contamination in sequence library, d) reference genome, e) short read length, f) software for analysis and g) selection of parameters. Thus, keeping all these points in mind a comprehensive examination was performed and samples of both chestnut and northern brook lamprey were assembled *de novo* using the DNA lasergene NextGenSeq Module, and assembled contigs were BLASTed to identify the quality of assembled **contigs**. The results indicated that a substantial fraction of these *de novo* assembled contigs were bacterial in origin and therefore were not mapped to the reference genome by TopHat. Another possible cause for the low mapping percentage, could be due to the poor assembly of the sea lamprey genome. In an effort to understand the low mapping percentage, I analyzed some of the contigs of chestnut and northern brook lamprey on the UCSC Genome Browser and found that many transcripts had portions spread across two contigs of the sea lamprey genome, and these fragments have been assigned different Ensembl ID's. Although there are an estimated 84 chromosomes in lamprey, the genome still exists in contigs; indeed, the contigs are still not assembled into large chromosomal sections. This fragmentation of the genome may also have reduced the mapping percentage. The poor assembly of the genome is mirrored by the incomplete annotation of the genome as available on Ensembl. As mentioned, the annotation file (GFF) of lamprey on Ensembl contains 13,114 Ensembl ID's, of which only 10,415 are identified as being protein coding genes (Table 2.1). Since sea lamprey has an estimated 26,046 genes (Smith *et al.*, 2013), this indicates that the annotation file available on the Ensembl Genome Browser is not updated. These are some of the problems I identified which partially explain the low percentage of reads

that mapped to either annotated genes or the genome. Despite these problems, I was able to perform an analysis of the expression of genes in lamprey ovaries across gonadal stages. Using the ~6,900 genes that mapped to annotated genes in the sea lamprey genome, I compared gene expression across gonadal stages, identified differentially expressed genes in early, mid- and late gonadal stages and assigned Gene Ontology terms to those genes that were found to be up- and down-regulated.

A major goal of this chapter was to identify the expression pattern of genes when different samples of chestnut and northern brook lamprey were pooled together based on their stage of ovarian development. Following mapping and counting of the genes, a differential expression analysis was carried out using DESeq2, and a Principal Component Analysis (PCA) plot was generated of the overall pattern of gene expression from all samples (based on where dots cluster together) as shown in the graphs 2.6 & 2.7. In the early gonadal stages, larval samples of chestnut (C13-03 and C20-3) and northern brook (N1-10 and N17-1) lampreys showed no observable differences between species when pooled together based on the stage of ovarian development (blue dots cluster together on the PCA graph, Figures 2.6 & 2.7). This suggests that gene expression was the same for all sizes of larvae and across all stages of gonadal development in the early stages. However, in the mid-gonadal stage, there were four samples, all in gonadal stage 4, three from the parasitic species (IC1 & IC2 & IC3) and one sample from the non-parasitic species (M01). Ovaries from two of the parasitic species (IC1 & IC2) and the non-parasitic species (M01) exhibited very similar gene expression (green dots); however interestingly, the third parasitic sample (IC3) had an overall gene expression pattern similar to northern brook samples in the vitellogenesis stage. There were three samples of non-parasitic lamprey in vitellogenesis stage (red dots); all of these samples were in late metamorphosis (stage

5-8) and exhibited a similar expression profile, similar to that of sample IC3 in gonadal stage 4 (green dot). Although the sample IC3 was in gonadal stage 4, it was in late metamorphosis (stage 5-6), as compared to the other two chestnut lamprey samples which were in early metamorphosis (stage 2-3). The similarity in gene expression profiles between IC3 and the three vitellogenic northern brook lamprey samples is surprising, suggesting that during late metamorphosis in chestnut lamprey, gene pathways related to sexual maturation may be turned on well in advance of histological evidence of vitellogenesis. Externally, during metamorphosis stages 5-6, teeth and lingual laminae are clearly visible in the enlarged oral disc, the eye becomes large, and some adult pattern of body coloration is observed (Manzon *et al.*, 2015). However, the ovary in chestnut lamprey is still sexually immature and spawning will not occur for another approximately 1.8 years. The similarity in gene expression profiles between this one late-stage metamorphosing chestnut lamprey and the three vitellogenic northern brook lamprey was one of the most interesting result obtained in this chapter. The sample was indeed a chestnut lamprey (i.e., had sequence homology with the other samples of chestnut lamprey), but should be studied further since this pattern was observed in only a single individual.

This pipeline also identified different genes that were up- and down-regulated in lampreys during ovarian development. The next step was to identify the Gene Ontology terms associated with these up- and down-regulated genes. In lampreys, since many of the genes are not annotated and have no associated Gene Ontology (GO) term, a different route was suggested in this pipeline for assigning Gene Ontology terms. All the up- and down-regulated genes were converted into zebrafish orthologs and were used as an input to GOrilla and REVIGO. These tools are computationally very fast and, using this approach, all the up-regulated and down-regulated genes were assigned GO terms and were visualized in two-dimensional plots,

interactive graphs, tree maps and tag clouds. The up-regulated genes had Gene Ontology terms related to cellular and reproductive processes such as regulation of biological and reproductive processes, positive regulation of multicellular organisms, sperm egg recognition, cell-cell recognition, egg coat formation and signal transduction pathway (Figures 2.11-2.13); down-regulated genes had Gene Ontology terms related to G protein receptor signaling pathway, detection of stimulus involved in sensory perception, regulation of ARF protein signal transduction etc. (Figures 2.15-2.18). This suggests that up-regulated genes were involved in regulating different processes related to ovarian differentiation and sexual maturation in lampreys. This approach led to the identification of different genes and provided an overview of co-occurring or related GO terms found to be up- or down-regulated over gonadal stage. Based on the annotations obtained from GOrilla, the HTSeq counts of up- and down-regulated genes of lampreys were obtained and it was observed that genes such as *zona pellucida glycoprotein 3a* (ZPG3) which helps in growing oocytes and has been detected in different fish (Bobbe *et al.*, 2008; Lubzen *et al.*, 2010), *oxytocin receptor* (OXTR), *neuropeptide Y receptor Y8b* (NPYR) known to play diverse roles in regulating satiety, neuroendocrine axes, vasoconstriction, and cardiovascular remodeling (Pedrazzini *et al.*, 2003) were significantly expressed in northern brook lamprey during gonadal stage 4 as compared to the chestnut lamprey (Figure 2.20 a, b, c & d). On the contrary, genes like *adrenoceptor alpha* (ADRA2A) which is a G protein coupled receptor known to regulate embryogenesis and adulthood (Héctor *et al.*, 2017), *estrogen receptor 2* (ER2) which plays an important role in steroid hormone signaling (Baker, 1997, 2003; Escriva *et al.*, 1997; Escriva, 2000; Thornton *et al.*, 2003), oogenesis, vitellogenesis (Gustafsson, 2003, Heldring *et al.*, 2007, Hess, 2003; Nilsson *et al.*, 2001) and is important for male and female development (Piferrer & Guiguen 2008), and *thyroid stimulating hormone receptor* were

up-regulated in chestnut lamprey during gonadal stage 4; it was observed that these genes showed minimal expression during early gonadal stages (2 & 3), but at gonadal stage 4 there was a significant increase in expression. Previous studies reported that, during ovarian maturation in European sea bass *Dicentrarchus labrax* and channel catfish *Ictalurus punctatus*, TSH-R expression level increased (Rocha *et al.*, 2007; Goto-Kajeto *et al.*, 2009). These genes, were mostly expressed in chestnut lamprey during gonadal stage 4 (Figure 2.21). This is quite interesting and is correlated with the previous result obtained from the PCA plot where one parasitic lamprey in gonadal stage 4 (green dot, Figure 2.6) clustered with the non-parasitic lampreys undergoing vitellogenesis. This suggests that during gonadal stage 4, these genes help chestnut lamprey to prepare for vitellogenesis.

Another goal of this pipeline was to identify and count the insulin superfamily genes expressed during different stages of gonadal development in the parasitic and non-parasitic lampreys using both the reference-guided assembly pipeline (this Chapter) and the *de novo* pipeline (Chapter 3). Of the ~6,900 genes identified in the HTSeq analyses, more than 6,000 of these were reported as “novel protein coding” in Ensembl, and had no assigned ID. However, based on previous analyses, we have identified several members of the insulin superfamily of genes on Ensembl. Specifically, for the relaxin family of peptides and receptors, three Ensembl ID’s (ENSPMAG00000001344, ENSPMAG00000006421, and, ENSPMAG00000010113) pertain to members of the relaxin family peptide receptors, rxfp3-3 and rxfp3-1, while one of the Ensembl ID’s (ENSPMAG00000001344) had no gene name, but through data mining, the **putative** name of this Ensembl ID is rxfp3. There are two Ensembl ID’s (ENSPMAG00000001344 & ENSPMAG00000010113) reported for the same gene rxfp3. Also, one (ENSPMAG00000000636) is reported as novel, but is likely rxfp1 (Good *et al.*, 2014). The

rxfp3 receptors are related to other small peptide hormones (e.g., angiotensin and somatostatin) receptors; in vertebrates rxfp3's are activated by rln3 and insl5, and are predominantly involved in neuroendocrine processes, although they also show some expression in gonads (Good *et al.*, 2014). On the other hand, rxfp1 is closely related to other glycoprotein hormone receptors, such as luteinizing hormone and follicle-stimulating hormone receptors (lhr and fshr); in mammals, it is activated by the peptide hormone rln and plays diverse roles in partuition, reproduction and cartilage remodulation (Bathgate *et al.*, 2013). For the insulin and insulin like growth factor (ins and igf) peptides and receptors, data mining and previous genomic locations had identified seven Ensembl ID's associated with the family (ENSPMAG00000008971, ENSPMAG00000002950, ENSPMAG00000007101, ENSPMAG00000006274, ENSPMAG00000005896, ENSPMAG00000007436, and ENSPMAG00000008570). None of these genes have associated gene names in the current version of the sea lamprey genome, only Ensembl ID's. The putative gene names of some of the Ensembl ID's were obtained by looking at the **orthologous** and **paralogous** genes in other taxa related to the sequence on Ensembl and by examination of their respective gene trees available at Ensembl which are created automatically by the Ensembl Genome Browser based on sequence similarity (not gene name). Using these approaches, ENSPMAG00000002950 was found to be an ortholog of IGF2 (insulin growth factor 2); ENSPMAG00000007101, ENSPMAG00000008570 and ENSPMAG00000006274 are orthologs of igf1r or insrr; ENSPMAG00000008971 is insulin (ins); and ENSPMAG00000005896 is the insulin receptor (insr). While the ortholog/paralog of ENSPMAG00000007436 could not be assessed; it appears to be either a segment of the above insr gene or a novel insr-like gene. The HTSeq counts of these Ensembl ID's were obtained for both parasitic and non-parasitic lampreys and histograms were plotted by GraphPad.

The histograms of *rxfp1*(novel), *rxfp3-1* and *igf1r/insrr* indicated that they are mostly expressed in northern brook lamprey during gonadal stages 2, 3 and 4 as compared to the parasitic lamprey (Figure 2.23 a, b, c and, d & 2.24 a and b). It was not possible to compare the gene counts of both species during late stages of gonadal development because there were no samples of chestnut lamprey undergoing vitellogenesis. However, the gene count of northern brook lamprey undergoing vitellogenesis is quite high, which suggests that these genes are expressed in all stages and may be involved in vitellogenesis and sexual maturation in northern brook lamprey (tables available upon request). However, *igf2* or *igf1* was found to be highly expressed in the parasitic chestnut lamprey during gonadal stages 2 and 3 but, in gonadal stage 4, there was a sharp decrease in gene expression in chestnut lamprey (Figure 2.24 b); through subsequent data mining, it was learned that this gene is probably *igf1*. Insulin was only found to be expressed in chestnut lamprey during gonadal stage 3 (Figure 2.24 c); the putative name of this gene is *INS-L1*. All these results suggest that insulin family genes play a vital role in ovarian differentiation and maturation in lampreys. Future investigations on these genes are required to confirm their roles in specific gonadal stages, and to confirm the full nomenclature of the lamprey insulin superfamily genes.

There were some technical difficulties reported in the present study. Firstly, the low mapping percentage was a matter of concern and future studies should be designed by taking samples from different tissues (brain and gut) and the mapping percentage should be compared across different tissues. Secondly, this study focused only on mapped reads; hence, a major proportion of reads were unmapped and discarded. The mapped read file was used further for counting genes by HTSeq; half of the reads were recognized as “No feature” and were not counted by the program, thus it was not possible to recover the uncounted reads. The differential

expression by DESeq2 was estimated on only half of the mapped reads. The sample size used for performing differential analysis was not sufficient to capture the species-specific differences in gene expression, so a pooled global analysis was performed. There were no samples of chestnut lamprey undergoing vitellogenesis; hence comparative analysis was done until gonadal stage 4, so future studies should be designed by taking a good sample size from both the species of lampreys at different gonadal stages.

## **2.8. Implications from this study**

This study generated a list of up-regulated and down-regulated genes known to be involved in ovarian differentiation as well as normalized counts of all annotated genes across different stages of ovarian development in lampreys. This extensive gene list can be used for targeting specific genes of interest and for performing qPCR. Spice *et al.* (2014) used qRT-PCR to study gene specific expression in parasitic chestnut lamprey *Ichthyomyzon castaneus* and non-parasitic northern brook lamprey *Ichthyomyzon fossor* during ovarian differentiation. Having a complete list of the annotated genes involved in ovarian differentiation is helpful for designing future studies, as well as having informed us about the overall pathways that are turned on or off during different gonadal stages. The reference-guided pipeline has successfully identified insulin family genes; this clearly suggests that this approach can be used for updating the annotations of sea lamprey. Future studies can be designed for extracting the sequence of insulin family genes in the database and potentially be used to construct gene phylogenies with the goal of contributing new knowledge about genes involved in ovarian differentiation in lampreys. The process of assigning Gene Ontology in this pipeline can be used successfully for species whose

genomes have not been extensively annotated. The present pipeline has worked well for lampreys and can be used on any non-model organism.

## 2.9. References

- Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.
- Anders S, Pyl PT, Huber W (2015) HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Baker ME (1997) Steroid receptor phylogeny and vertebrate origins. *Molecular and Cellular Endocrinology*, **135**, 101-107.
- Baker ME (2003) Evolution of adrenal and sex steroid action in vertebrates: A ligand-based mechanism for complexity. *BioEssays*, **25**, 396-400.
- Barker LA & Beamish FWH (2000) Gonadogenesis in landlocked and anadromous forms of the sea lamprey, *Petromyzon marinus*. *Environmental Biology of Fishes*, **59**, 229– 234.
- Blankenberg D, Gordon A, Von Kuster G *et al.* (2010) Manipulation of FASTQ data with Galaxy, *Bioinformatics*, **26**, 1783-1785.
- Bobe J, Nguyen T, Mahé S *et al.* (2008) *In silico* identification and molecular characterization of genes predominantly expressed in the fish oocyte, *BMC Genomics*, **9**, 499.
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

- Brown WM, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Science*, **76**(4), 1967-1971.
- Cline MS, Smoot M, Cerami E *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, **2**, 2366–2382.
- Conesa A, Madrigal P, Tarazona S *et al.* (2016) A survey of best practices for RNA-Seq data analysis. *Genome Biology*, **17**, 13.
- Costa V, Angelini C, De Feis I *et al.* (2010) Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedical and Biotechnology*, **2010**, 853916.
- Docker MF, Youson JH, Beamish RJ *et al.* (1999) Phylogeny of the lamprey genus *Lampetra* inferred from mitochondrial cytochrome b and ND3 gene sequences. *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 2340–2349.
- Eden E, Navon R, Steinfeld I *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.
- Eklom R & Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.
- Giardine B, Riemer C, Hardison RC *et al.* (2015) Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, **15**, 1451–1455.
- Goetz F, Rosauer D, Sitar S *et al.* (2010) A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (*Salvelinus namaycush*). *Molecular Ecology*, **19**, 176–196.
- Goto-Kazeto R, Kazeto Y, Trant JM (2009) Molecular cloning, characterization and expression of thyroid-stimulating hormone receptor in channel catfish. *General Comparative Endocrinology*, **161**, 313-319.

- Gustafsson JA (2003) What pharmacologists can learn from recent advances in estrogen signaling. *Trends in Pharmacological Sciences*, **24**, 479–485.
- Hardisty MW & Potter IC (1971) Paired species. In: *The Biology of Lampreys, volume 1* (eds Hardisty MW, Potter IC), pp 249–277. Academic Press, New York.
- Hardisty MW (1969) A comparison of gonadal development in the ammocoetes of the landlocked and anadromous forms of the sea lamprey *Petromyzon marinus* L. *Journal of Fish Biology*, **1**, 153–166.
- Hardisty MW (1970) The relationship of gonadal development to the life cycles of the paired species of lamprey, *Lampetra fluviatilis* (L.) and *Lampetra planeri* (Bloch). *Journal of Fish Biology*, **2**, 173–181.
- Hardisty MW (1971) Gonadogenesis, sex differentiation and gametogenesis. In: *The Biology of Lampreys, volume 1* (eds Hardisty MW, Potter IC), pp. 317–324. Academic Press, New York.
- Cespedes HA, Zavala K, Opazo J (2017) Evolution of the  $\alpha$ 2-adrenoreceptors in vertebrates: ADRA2D is absent in mammals and crocodiles. *bioRxiv*, **106526**, doi: <https://doi.org/10.1101/106526>.
- Heldring N, Pike A, Andersson S *et al.* (2007) Estrogen receptors: how do they signal and what are their targets. *Physiology Review*, **87**, 905–931.
- Hess RA (2003) Estrogen in the adult male reproductive tract: a review. *Reproduction Biology and Endocrinology*, **1**, 52.
- Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

- Jeukens J, Renaut S, St-Cyr J *et al.* (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology*, **19**, 5389–5403.
- Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, **15**, 550.
- Lubzens E, Young G, Bobe J *et al.* (2010) Oogenesis in teleosts: how fish eggs are formed. *General Comparative Endocrinology*, **165**, 367–389.
- Manousaki T, Hull PM, Kusche H *et al.* (2013) Parsing parallel evolution: ecological divergence and differential gene expression in the adaptive radiations of thick-lipped Midas cichlid fishes from Nicaragua. *Molecular Ecology*, **22**, 650–669.
- Manzon RG, Youson JH, Holmes JA (2015) Lamprey metamorphosis. In: Docker MF, editor. *Lampreys: Biology, Conservation, and Control*, volume **1**. Springer, 139-214.
- Neave FB, Mandrak NE, Docker MF *et al.* (2007) An attempt to differentiate sympatric *Ichthyomyzon ammocoetes* using meristic, morphological, pigmentation, and gonad analyses. *Canadian Journal of Zoology*, **85**, 549–560.
- Nilsson S, Mäkelä S, Treuter E *et al.* (2001) Mechanisms of estrogen action. *Physiology Review*, **81**, 1535–1565.
- Pedrazzini T, Pralong F, Grouzmann E (2003) Neuropeptide Y: The universal soldier. *Cell and Molecular Life Sciences*, **60**, 350–377.
- Piferrer F & Guiguen Y (2008) Fish gonadogenesis. Part II: Molecular biology and genomics of sex differentiation. *Reviews in Fisheries Science*, **16**, 35–55.

- Potter IC, Gill HS, Renaud CB *et al.* (2015) The taxonomy, phylogeny, and distribution of lampreys. Docker MF, ed. *Lampreys: Biology, Conservation, and Control*, volume **1**. Springer, 35-37.
- Reinecke M (2010) Insulin-like growth factors and fish reproduction. *Biology of Reproduction*, **82**, 656–661.
- Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Rocha A, Gómez A, Zanuy S *et al.* (2007) Molecular characterization of two sea bass gonadotropin receptors: cDNA cloning, expression analysis, and functional activity. *Molecular and Cellular Endocrinology*, **272**, 63-76.
- Ruan Y, Le Ber P, Ng HH *et al.* (2004) Interrogating the transcriptome. *Trends in Biotechnology*, **22**, 23–30.
- Saitou N & Nei M (1987) The Neighbor-Joining method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406-425.
- Smith JJ, Kuraku S, Holt C *et al.* (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genetics*, **45**, 415–421.
- Spice EK (2013) Ovarian Differentiation in an Ancient Vertebrate: Timing, Candidate Gene Expression, and Global Gene Expression in Parasitic and Non-parasitic Lampreys (M.Sc. thesis). University of Manitoba.
- Spice EK, Whyard S, Docker MF (2014) Gene expression during ovarian differentiation in parasitic and non-parasitic lampreys: Implications for fecundity and life history types. *General and Comparative Endocrinology*, **208**, 116–125.

- Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science*, **301**, 1714–1717.
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
- Trapnell C, Williams BA, Pertea G *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–5.
- Tucker T, Marra M, Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. *American Journal of Human Genetics*, **85**, 142–154.
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, **10**, 57–63.
- Zhang J, Haider S, Baran J *et al.* (2011) BioMart: a data federation framework for large collaborative projects, *Database*, **2011**, bar038.

## 2.10. Tables and Figure

Table 2.1. Detailed summary of sequenced *Petromyzon marinus* (sea lamprey) genome obtained from Ensembl Genome browser.

Assembly	Pmarinus_7.0, Jan 2011
Database version	87.7
Bp	647,368,134
Golden Path Length	885,550,958
Genebuild by	Ensembl
Genebuild method	Full genebuild
Genebuild started	Feb 2011
Genebuild released	Sep 2011
Genebuild last updated/patched	Apr 2013
Gene counts	
Coding genes	10,415
Non coding genes	2,652
Small non coding genes	2,623
Misc non coding genes	29
Pseudogenes	47
Gene transcripts	14,141
Genscan gene predictions	34,895

Table 2.2. FASTQC report on individuals of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* used in this project. “No of reads” refers to the count of the total no of sequences processed. “% GC” refers to the overall GC content of all bases in all sequences. “Peak in Phred distribution” refers to the quality scores assigned to each nucleotide base call and is used for assessing the quality of sequences.

<b>Samples</b>	<b>Species</b>	<b>No of reads</b>	<b>% GC</b>	<b>Peak in Phred distribution</b>
C13-03	<i>I. castaneus</i>	16,458,931	46	37
C20-03	<i>I. castaneus</i>	5,249,442	48	37
IC3	<i>I. castaneus</i>	96,717,768	54	36
IC1	<i>I. castaneus</i>	56,683,988	53	36
IC2	<i>I. castaneus</i>	53,569,364	52	36
N02	<i>I. fossor</i>	2,243,262	56	36
N1-10	<i>I. fossor</i>	46,055,489	50	36
N17-1	<i>I. fossor</i>	12,424,954	50	36
M01	<i>I. fossor</i>	63,962,035	55	36
NC3	<i>I. fossor</i>	146,230,589	51	37
N08	<i>I. fossor</i>	56,421,386	50	36
S11	<i>I. fossor</i>	51,350,095	51	36

Table 2.3. TopHat report on individuals of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* used in this project, individual information about the samples are included in Table 1.3. “Input reads” refers to the total raw reads entered for mapping. “Left hand reads” refers to the reads aligned to the reverse strand. “Right hand reads” refers to the reads aligned to the forward strand. “Aligned reads” refers to the position of a sequencing read corresponding to the reference genome. “Mapped reads” refers to the total count of reads aligned to a reference sequence to generate transcripts. “Unmapped reads” refers to the total count of reads that does not aligns to a reference sequence.

<b>Sample number</b>	<b>Species</b>	<b>Raw reads</b>	<b>Left hand reads</b>	<b>Right hand reads</b>	<b>Aligned reads</b>	<b>Mapped reads (%)</b>	<b>Unmapped reads (%)</b>
C13-03	<i>I. castaneus</i>	16,458,931	7,314,077	7,352,286	4,828,489	44.4	55.6
C20-03	<i>I. castaneus</i>	52,494,442	10,684,432	11,845,185	6,404,086	21.5	78.5
IC3	<i>I. castaneus</i>	96,717,768	11,629,494	11,707,925	5,136,857	12.1	87.9
IC1	<i>I. castaneus</i>	56,683,988	13,028,542	13,326,271	5,322,431	23.2	76.8
IC2	<i>I. castaneus</i>	53,569,364	17,853,945	18,057,938	8,630,585	33.5	66.5
N02	<i>I. fossor</i>	2,243,262	562,609	568,665	353,602	25.2	74.8
N1-10	<i>I. fossor</i>	46,055,489	7,101,631	7,193,979	3,107,177	15.5	84.5
N17-1	<i>I. fossor</i>	12,424,954	4,347,283	4,379,129	2,873,991	35.1	64.9
M01	<i>I. fossor</i>	63,962,035	22,347,410	21,448,641	11,928,619	34.2	65.8
NC3	<i>I. fossor</i>	146,230,589	27,404,754	24,134,108	11,337,147	17.6	82.4
N08	<i>I. fossor</i>	56,421,386	16,877,697	16,388,607	9,679,622	29.5	70.5
S11	<i>I. fossor</i>	51,350,095	11,374,084	11,568,940	4,799,618	22.3	77.7

Table 2.4. HTSeq report on individuals of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* used in this project, individual information about the samples are included in Table 1.3. “No feature” refers to the reads (or read pairs) which could not be assigned to any feature. “Ambiguous” refers to the reads (or read pairs) which contains more than one feature and not counted. “Alignment not unique” refers to the reads (or read pairs) with more than one alignment. “Total no of HTSeq counts” refers to the total number of reads reported which includes all ‘no feature’, ‘ambiguous’, and ‘alignment not unique’. “Total sum of reads counts by HTSeq” refers to the reads which contains feature. “% of genes count” refers to the total sum of genes reported by HTSeq/ Total number of HTSeq counts. “Total non-zero counts” refers to the read (or read pairs) whose numeric value is >0. “Gene count >5” refers to the reads (or read pairs) whose numeric value is >5.

<b>Samples</b>	<b>No feature</b>	<b>Ambiguous</b>	<b>Alignment not unique</b>	<b>Total no of HTSeq counts</b>	<b>Total sum of reads counts by HTSeq</b>	<b>% of genes counts</b>	<b>Total non-zero counts</b>	<b>Gene count &gt;5</b>
C13-03	5,519,743	132,270	5,901,286	11,553,299	2,757,196	23.86	9,058	7,385
C20-03	7,840,544	31,890	22,823,114	30,695,548	4,781,411	15.57	9,005	7,726
IC3	9,102,420	61,213	11,767,531	20,931,164	7,157,655	34.19	8,293	7,173
IC1	6,839,790	42,861	26,312,524	33,195,175	9,834,821	29.62	9,132	8,187
IC2	11,714,599	68,226	32,513,264	44,296,089	9,866,298	22.27	9,374	8,325
N02	470,584	934	1,020,321	1,491,839	106,212	7.11	5,855	2,895
N1-10	4,476,554	48,608	9,172,959	13,698,121	5,089,264	37.15	9,304	8,364
N17-1	2,835,641	19,299	4,421,807	7,276,747	2,219,182	30.49	8,930	7,692
M01	11,640,409	65,887	59,595,880	17,666,176	9,933,530	56.23	9,293	8,275
NC3	15,486,129	126,475	26,014,920	41,627,524	20,154,960	48.42	9,116	8,139
N08	11,580,489	59,995	15,273,634	26,914,118	9,201,891	34.2	8,836	7,739
S11	7,454,789	39,471	8,880,307	16,374,567	9,072,932	55.4	9,165	8,040

Table 2.5. Details on ovarian transcriptomes of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* used in DESeq2 as input. Samples from both species were pooled together according to gonadal stage to study the global pattern of gene expression during different stages of ovarian development (Figure 1.4, Table 1.1). Sample numbers C13-03, C20-03, IC1, IC2 and IC3 are chestnut lamprey and M01, NC3, N08 and S11 are northern brook lamprey.

<b>Early gonadal stages (stages 1-3)</b>	<b>Mid-gonadal stage (stage 4)</b>	<b>Late gonadal stages (early and late vitellogenesis)</b>
C13-03, C20-03, N1-10 & N17-1	IC1, IC2, IC3 & M01	NC3, N08 & S11

Table 2.6. Number of genes reported by DESeq2 that were regulated (up and down) in parasitic chestnut lamprey and non-parasitic northern brook lamprey at different stages of ovarian development based on *Padj* values of  $\leq 0.05$ . Samples from both species were pooled together based on three different stages of ovarian development: a) early gonadal stages (stages 1-3), b) mid-gonadal stage (stage 4) and, c) late gonadal stages (early and late vitellogenesis). “Total DESeq2” denotes genes with numeric value  $>0$ . All the genes reported in sea lamprey (*Petromyzon marinus*) are converted into zebrafish (*Danio rerio*) orthologs.

<b>Samples with Input factor 1</b>	<b>Samples with Input factor 2</b>	<b>Samples with Input factor 3</b>	<b>DESeq2 (up)</b>	<b>DESeq2 (up)</b>	<b>DESeq2 (down)</b>	<b>DESeq2 (down)</b>	<b>DESeq2 (nonzero)</b>
<b>(early gonadal stages)</b>	<b>(mid-gonadal stage)</b>	<b>(vitellogenesis)</b>	<i>Petromyzon marinus</i>	<i>Danio rerio</i>	<i>Petromyzon marinus</i>	<i>Danio rerio</i>	
C13-03, C20-3, N1-10 & N17-1	IC1, IC2, IC3 & M01	NC3, N08 & S11	2864	5401	1738	2730	9119

Table 2.7. Top 20 up-regulated genes reported by DESeq2 in parasitic chestnut lamprey and non-parasitic northern brook lamprey with respect to sea lamprey reference genome at different stages of ovarian development based on *Padj* values of  $\leq 0.05$ . Samples were pooled based on three different stages of ovarian development: a) early gonadal stage (stages 1-3), b) mid-gonadal stage (stage 4) and, c) late gonadal stage (early and late vitellogenesis).

<b>Input factor 1 (early gonadal stages)</b>	<b>Input factor 2 (mid-gonadal stage)</b>	<b>Input factor 3 (vitellogenesis)</b>	<b>Total genes Reported</b>
<b>C13-03, C20-3, N10 &amp; N17</b>	<b>IC1, IC2, IC3 &amp; M01</b>	<b>NC3, N08 &amp; S11</b>	<b>2864</b>
<b>Ensembl lamprey ID (up)</b>	<b>Gene name</b>	<b>Log2-fold change</b>	<b><i>Padj</i> value</b>
ENSPMAG00000000844	FYNRK (MANY)	9.038747235	2.76E-09
ENSPMAG00000003115	ST6GALNAC2 (MANY)	8.250720552	1.47E-09
ENSPMAG00000010431	ZGC:152698	8.210498108	1.60E-08
ENSPMAG00000006450	N/A	8.123657766	4.12E-14
ENSPMAG00000005953	N/A	8.056852228	3.28E-08
ENSPMAG00000006295	N/A	8.036945861	2.67E-07
ENSPMAG00000005601	N/A	8.003845434	3.28E-08
ENSPMAG00000009495	N/A	7.99855061	6.68E-07
ENSPMAG00000008606	N/A	7.964533751	1.27E-08
ENSPMAG00000006166	N/A	7.954650106	1.14E-05
ENSPMAG00000004336	N/A	7.89217638	1.01E-06
ENSPMAG00000000741	N/A	7.892161824	1.00E-07
ENSPMAG00000002471	N/A	7.8589445	2.67E-07
ENSPMAG00000001018	N/A	7.823718754	4.34E-10
ENSPMAG00000005208	N/A	7.81089371	1.01E-06
ENSPMAG00000006047	N/A	7.752047191	5.77E-08
ENSPMAG00000005593	N/A	7.751930139	5.77E-07
ENSPMAG00000005537	RAB27A	7.739848202	2.14E-06
ENSPMAG00000006650	N/A	7.727552027	1.01E-06

Table 2.8. Top 20 down-regulated genes reported by DESeq2 in parasitic chestnut lamprey and non-parasitic northern brook lamprey with respect to sea lamprey reference genome at different stages of ovarian development based on *Padj* values of  $\leq 0.05$ . Samples were pooled based on three different stages of ovarian development: a) early gonadal stage (stages 1-3), b) mid-gonadal stage (stage 4) and, c) late gonadal stage (early and late vitellogenesis).

<b>Input factor 1 (early gonadal stages)</b>	<b>Input factor 2 (mid-gonadal stage)</b>	<b>Input factor 3 (vitellogenesis)</b>	<b>Total genes reported</b>
<b>C13-03, C20-3, N10 &amp; N17 Ensembl lamprey ID (down)</b>	<b>IC1, IC2, IC3 &amp; M01 Gene name</b>	<b>NC3, N08 &amp; S11 Log2-fold change</b>	<b>1738 <i>Padj</i> value</b>
ENSPMAG00000003315	PHAX	-0.471457662	0.475302198
ENSPMAG00000001024	UBE3C	-0.513155891	0.457946119
ENSPMAG00000000497	N/A	-0.515444718	0.457946119
ENSPMAG00000002546	POFUT2	-0.517008323	0.460434485
ENSPMAG00000008768	HDAC3	-0.528852086	0.418234475
ENSPMAG00000005139	N/A	-0.539153919	0.419543514
ENSPMAG00000006746	N/A	-0.53950403	0.470766704
ENSPMAG00000000273	PSMA4	-0.541149803	0.445995562
ENSPMAG00000004360	TTC1	-0.541699975	0.476017216
ENSPMAG000000009320	PSMC3	-0.553565976	0.465364829
ENSPMAG00000003497	GPKOW	-0.565562814	0.340460052
ENSPMAG00000005951	N/A	-0.566713329	0.489514839
ENSPMAG00000003095	DGCR6	-0.576683188	0.385030835
ENSPMAG00000001389	THUMPD1	-0.577721227	0.475302198
ENSPMAG00000008143	MOSPD3	-0.578384164	0.45210274
ENSPMAG00000008807	N/A	-0.581116055	0.34328887
ENSPMAG00000001164	POLR2B	-0.583198113	0.402743333
ENSPMAG00000005802	AAK1A	-0.585666197	0.474337433
ENSPMAG00000006044	CCT5	-0.589927899	0.398585451
ENSPMAG00000003576	SSBP1	-0.595125101	0.414649333

Table 2.9 a. Statistics of up-regulated genes reported by GOrilla in zebrafish *Danio rerio*. “Genes identified” denotes the total number of input genes. “Duplicated genes” refers to genes which are reported more than once in the database. “Unresolved genes” refers to genes with no information in the database. “Total genes for annotation” refers to genes identified for annotation search. “Genes with GO terms” refers to genes identified with annotation terms whereas “Genes with no GO terms” refers to genes with no associated ontology terms in the database. All the genes are converted into sea lamprey orthologs.

<b>Input factor 1 early stage</b>	<b>Input factor 2 mid-stage</b>	<b>Input factor 3 late stage</b>	<b>Upregulated (<i>Danio rerio</i>)</b>	<b>Identified Genes</b>	<b>Duplicated genes</b>	<b>Unresolved genes</b>	<b>Total genes for annotation</b>	<b>Genes with GO terms</b>	<b>Genes with no GO terms</b>
C13-03 & C20-3	IC1 & IC3	-	1339	894	148	261	746	681	65
IC1, IC2 & IC3	NC3, N08 & S11	-	3523	2397	357	785	2040	1809	231
IC1, IC2, IC3 & M01	NC3, N08 & S11	-	4699	3143	488	1090	2655	2362	293
C13-03, C20-3, N1-10 & N1-17	IC1, IC2, IC3 & M01	NC3, N08 & S11	5401	3604	681	1254	2923	2609	314

Table 2.9 b. Statistics of down regulated genes reported by GOrilla in zebrafish *Danio rerio*. “Genes identified” denotes the total number of genes identified in the Gene Ontology Association (GOA) database. “Duplicated genes” refers to genes reported more than once. “Unresolved genes” refers to genes with no information in the database. “Total genes for annotation” refers to genes identified for annotation search. “Genes with GO terms” refers to genes identified with annotation terms whereas “Genes with no GO terms” refers to genes with no associated ontology terms in the database. All the genes are converted into sea lamprey orthologs.

<b>Input factor 1 (early stage)</b>	<b>Input factor 2 (mid-stage)</b>	<b>Input factor 3 (late stage)</b>	<b>Downregulated (<i>Danio rerio</i>)</b>	<b>Identified genes</b>	<b>Duplicated genes</b>	<b>Unresolved genes</b>	<b>Total genes for annotation</b>	<b>Genes with GO terms</b>	<b>Genes with no GO</b>
C13-03 & C20-3	IC1 & IC3	-	342	224	1	67	223	197	26
IC1, IC2 & IC3	NC3, N08 & S11	-	1725	1121	51	390	1070	957	113
IC1, IC2, IC3 & M01	NC3, N08 & S11	-	2179	1396	58	501	1338	1190	148
C13-03, C20-3, N1-10 & N1-17	IC1, IC2, IC3 & M01	NC3, N08 & S11	2730	1698	84	692	1614	1430	191

Table 2.10. The list of top 20 Gene Ontology terms for up-regulated and eight Gene Ontology terms for down-regulated genes reported by GOrilla within GO biological processes.

<b>Annotations by Gorilla</b>			
<b>Up-regulated genes</b>		<b>Down-regulated genes</b>	
<b>GO Term</b>	<b>Description</b>	<b>GO Term</b>	<b>Description</b>
GO:0007186	G-protein coupled receptor signaling pathway	GO:0050907	detection of chemical stimulus involved in sensory perception
GO:0007165	signal transduction	GO:0050906	detection of stimulus involved in sensory perception
GO:0043902	positive regulation of multi-organism process	GO:0009593	detection of chemical stimulus
GO:0043900	regulation of multi-organism process	GO:0051606	detection of stimulus
GO:0080154	regulation of fertilization	GO:0007186	G-protein coupled receptor signaling pathway
GO:2000243	positive regulation of reproductive process	GO:0050896	response to stimulus
GO:2000241	regulation of reproductive process	GO:0007165	signal transduction
GO:2000344	positive regulation of acrosome reaction	GO:0032012	regulation of ARF protein signal transduction
GO:1905516	positive regulation of fertilization		
GO:0009988	cell-cell recognition		
GO:0035036	sperm-egg recognition		
GO:0007339	binding of sperm to zona pellucida		
GO:0060046	regulation of acrosome reaction		
GO:0035803	egg coat formation		
GO:0050907	detection of chemical stimulus involved in sensory perception		
GO:0008037	cell recognition		
GO:0007608	sensory perception of smell		
GO:0007606	sensory perception of chemical stimulus		
GO:0009593	detection of chemical stimulus		
GO:0022412	cellular process involved in reproduction in multicellular organism		

Figure 2.1. Phylogenetic relationships between different parasitic and non-parasitic lamprey species derived from cytochrome b sequence data (from Potter *et al.*, 2015). Republished with permission of Springer Netherlands from the book “Lampreys: Biology, Conservation and Control” edited by Margaret Docker (2015), volume 37, 57; with permission conveyed through Copyright Clearance Center, Inc.

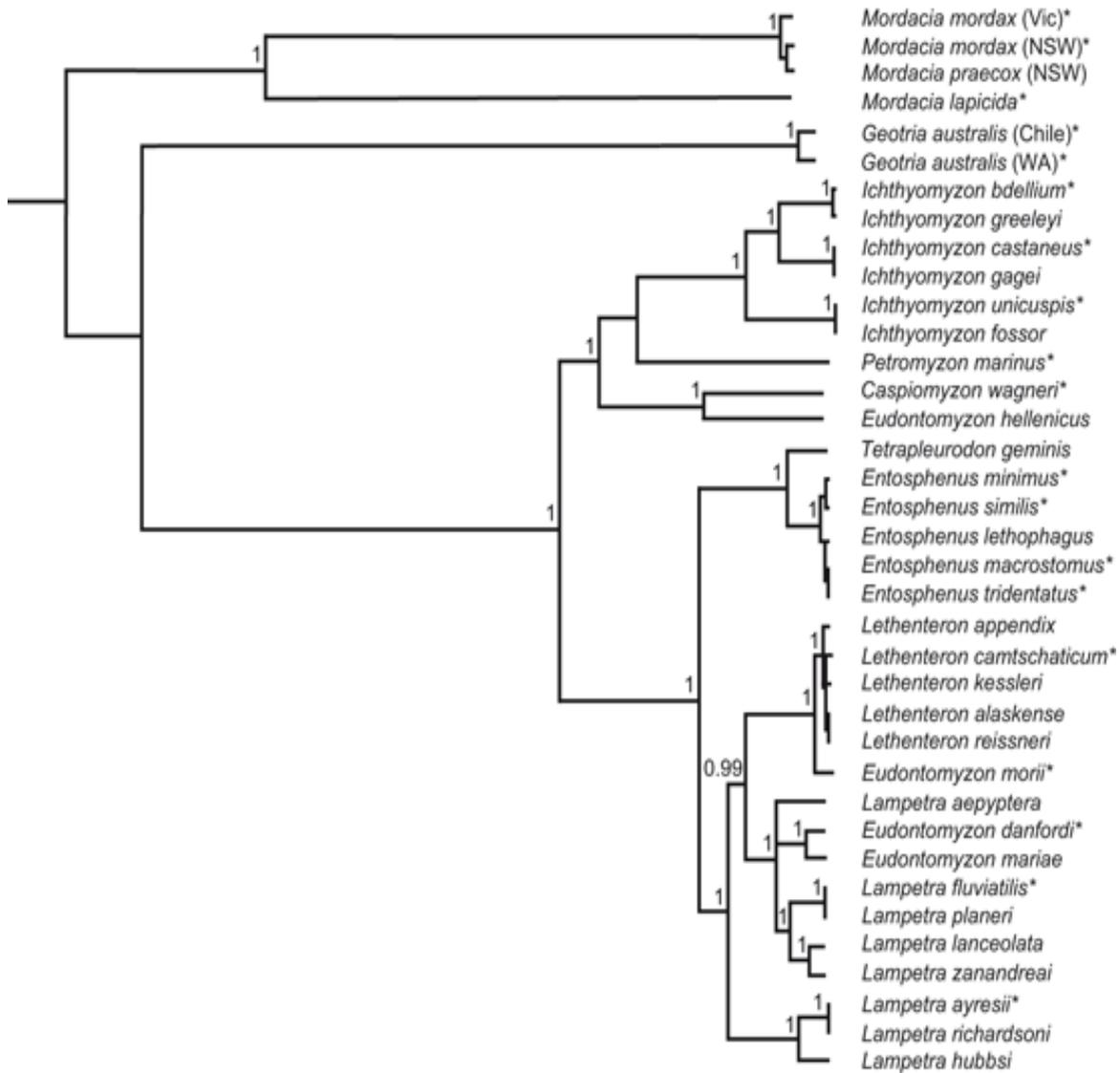


Figure 2.2. Transcriptomics data analysis pipeline designed for paired-end reads in the presence of a reference genome.

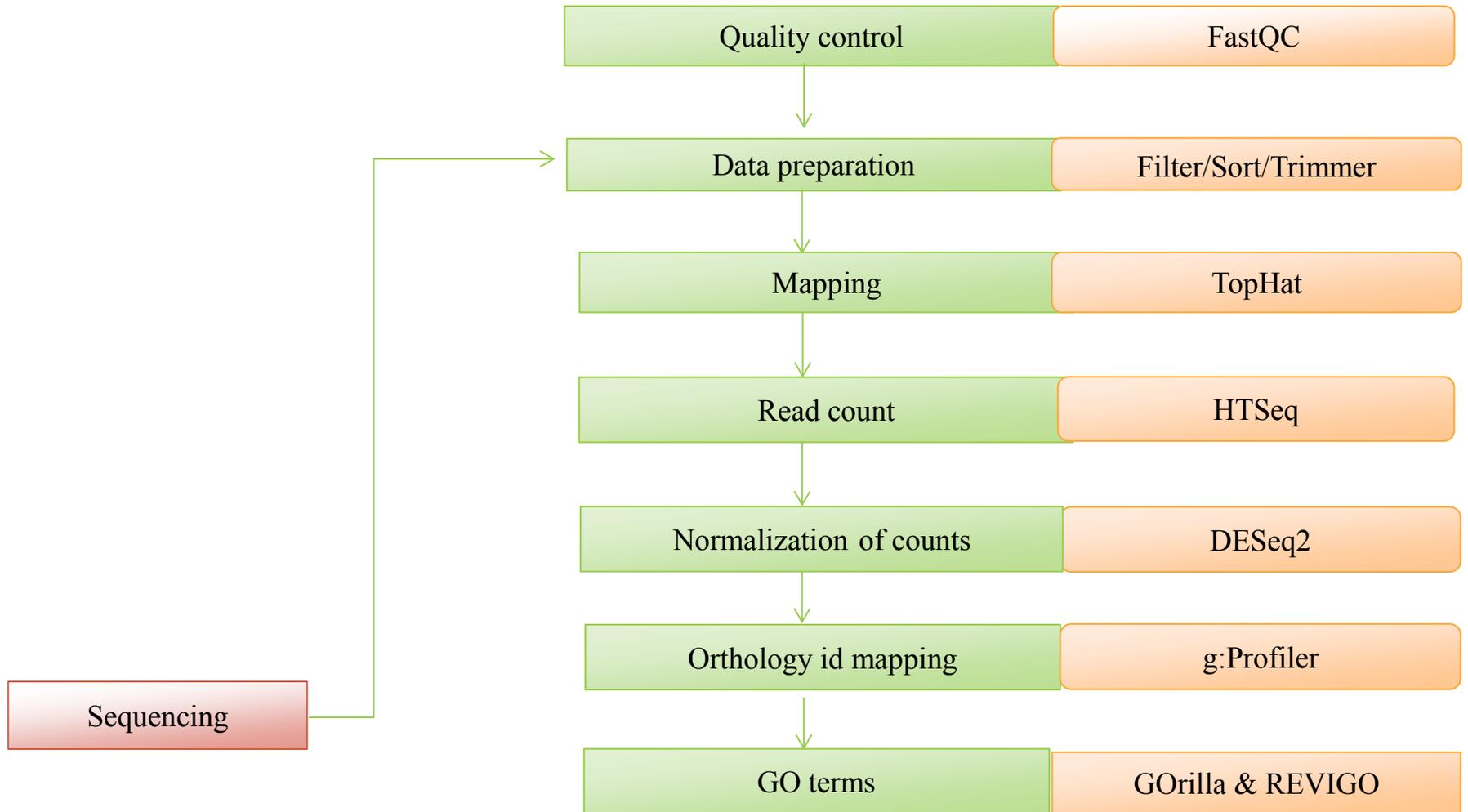
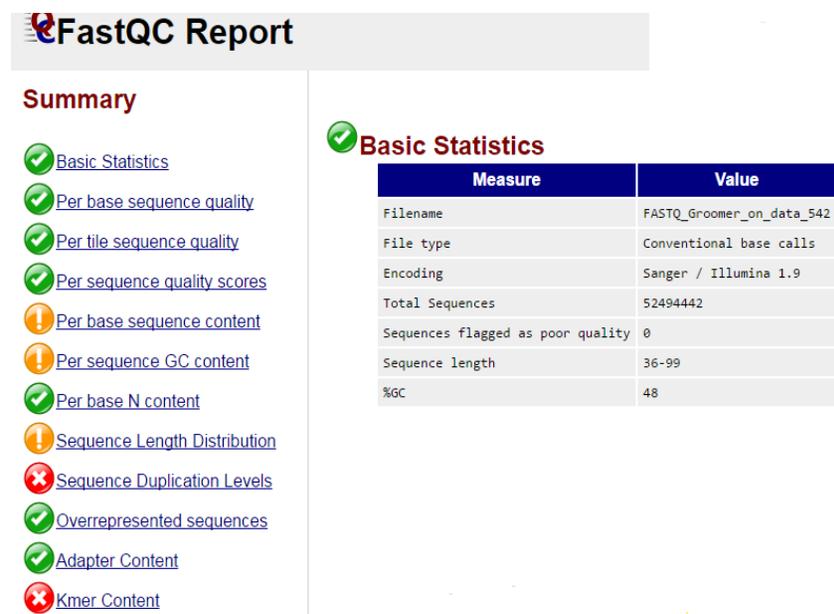
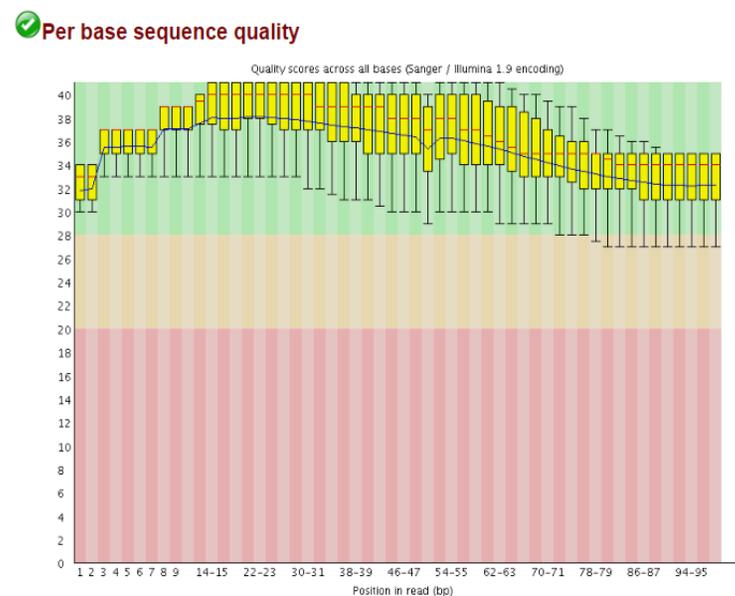


Figure 2.3 a. The summary report of FASTQC for chestnut lamprey C20-3 with detailed information about the total number of sequences processed and length of the shortest and longest sequence in the set. “%GC” denotes the overall GC content of all bases in all sequences. All other summary report figures are discussed below.

Figure 2.3 b. The per base sequence quality view gives an overview about the range of quality values across all bases at each position in the FastQ file. The x-axis represents the position of each read and y-axis denotes the quality scores. The background of the graph is divided into three different colors based on the of quality calls. For very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red).



2.3 a.

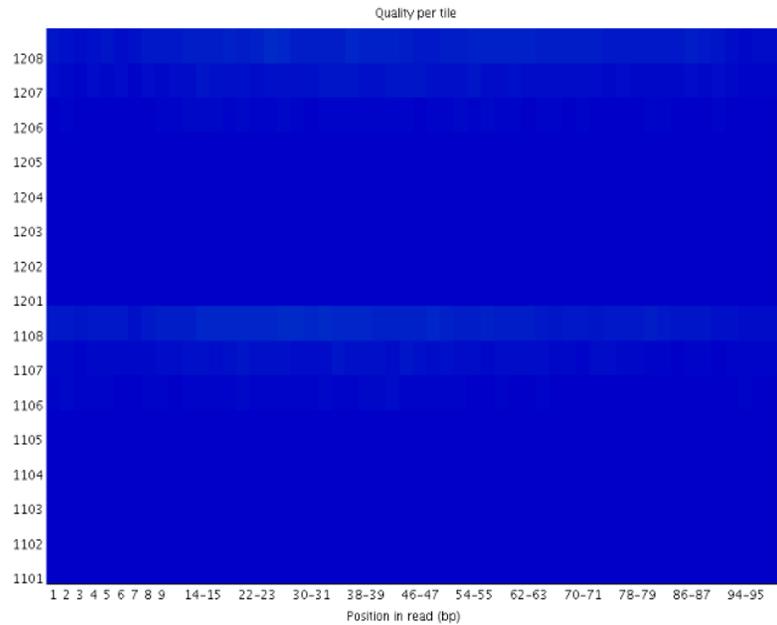


2.3 b.

Figure 2.3 c. The per tile sequence quality allows to look at deviation from the average quality for each tile. It generates the quality scores from each tile across all of bases and gives an idea about the quality loss (if any) associated with only one part of the flowcell. Cold colors denote that the quality is at or above the average for that base and hotter colors indicate that a tile had worse qualities than other tiles for that base. A good plot should be blue in color all over.

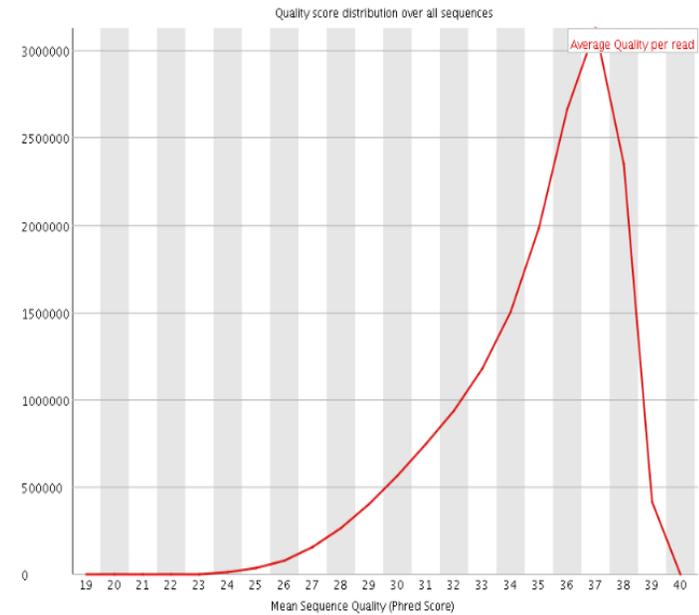
Figure 2.3 d. The per sequence quality score report gives an idea about the sequences that have universally low quality values.

✔ Per tile sequence quality



2.3 c.

✔ Per sequence quality scores

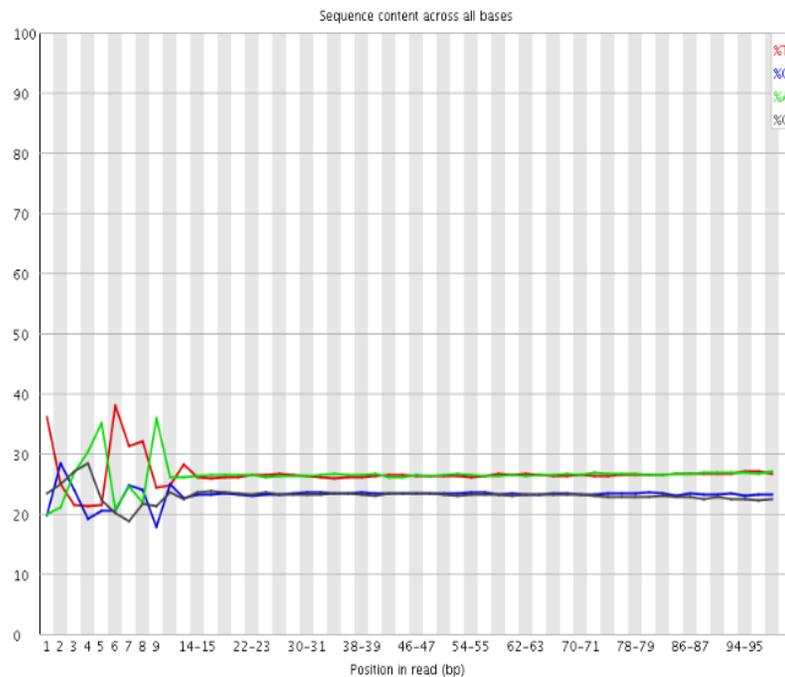


2.3 d.

Figure 2.3 e. The per base sequence content estimates the distribution of each base position for which each DNA bases (A, T, G and C) has been called and if the difference between A and T, or G and C is greater than 10% in any position, the module issues a warning.

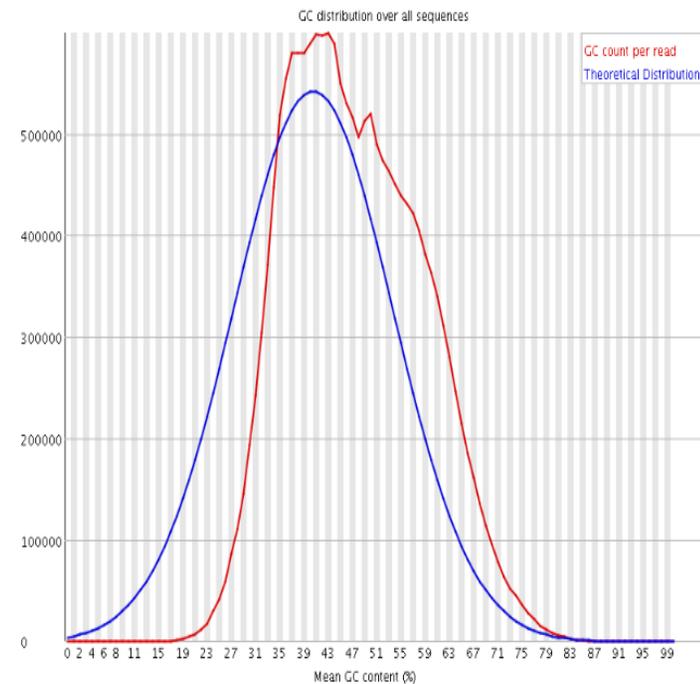
Figure 2.3 f. This per sequence GC content measures the GC distribution across the whole length of each sequence when compared to a normal distribution of GC content. If the sum of the deviations from the normal distribution is more than 15% of the reads, a warning is issued but if it is more than 30% it indicates a failure.

### ⚠ Per base sequence content



2.3 e.

### ⚠ Per sequence GC content

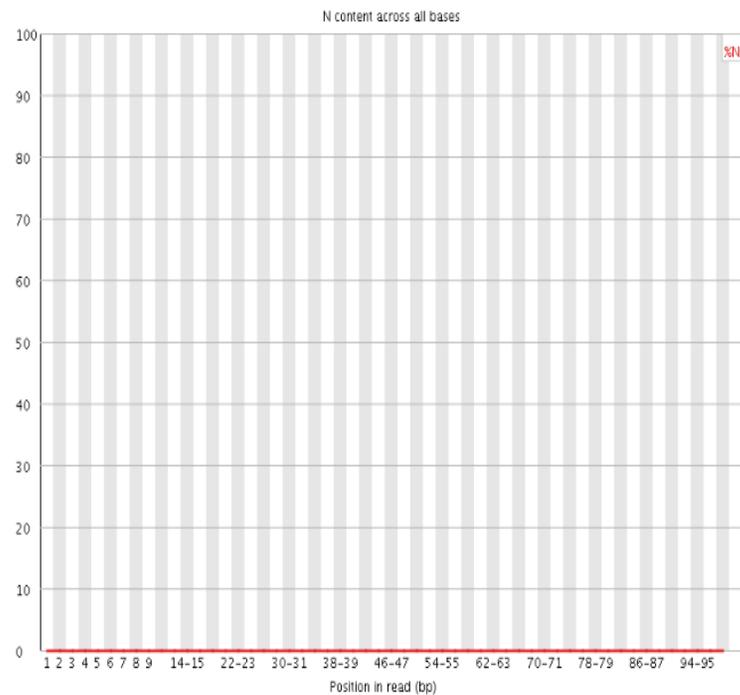


2.3 f.

Figure 2.3 g. The per base N content module estimates the percentage of base calls at each position for which an N was called. Generally, an N is substituted when a sequencer is unable to make a base call with sufficient confidence. A warning is issued when N content is >5%.

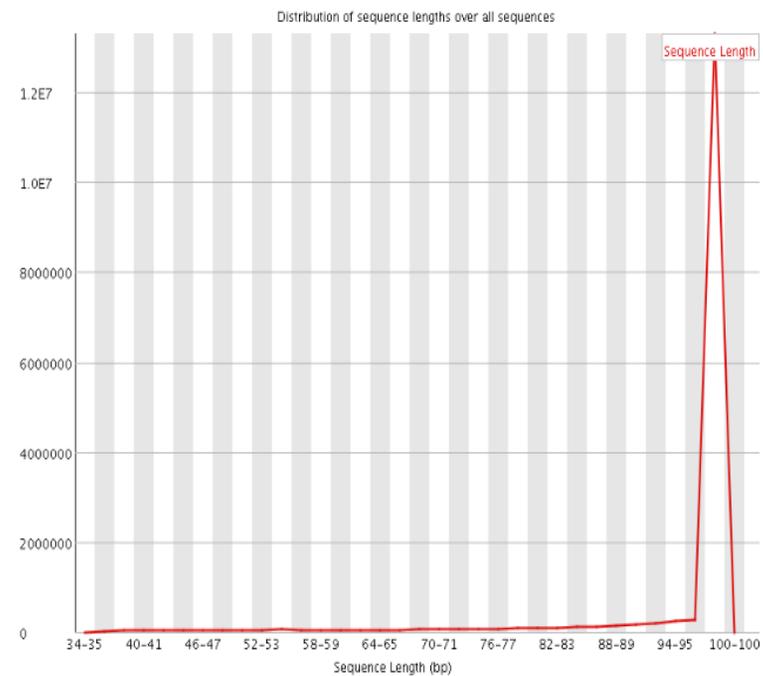
Figure 2.3 h. The sequence length distribution generates a simple graph showing the distribution of fragment sizes over all sequences and gives a single peak for same size of sequence fragment. A warning is issued if sequences are not of the same length.

 **Per base N content**



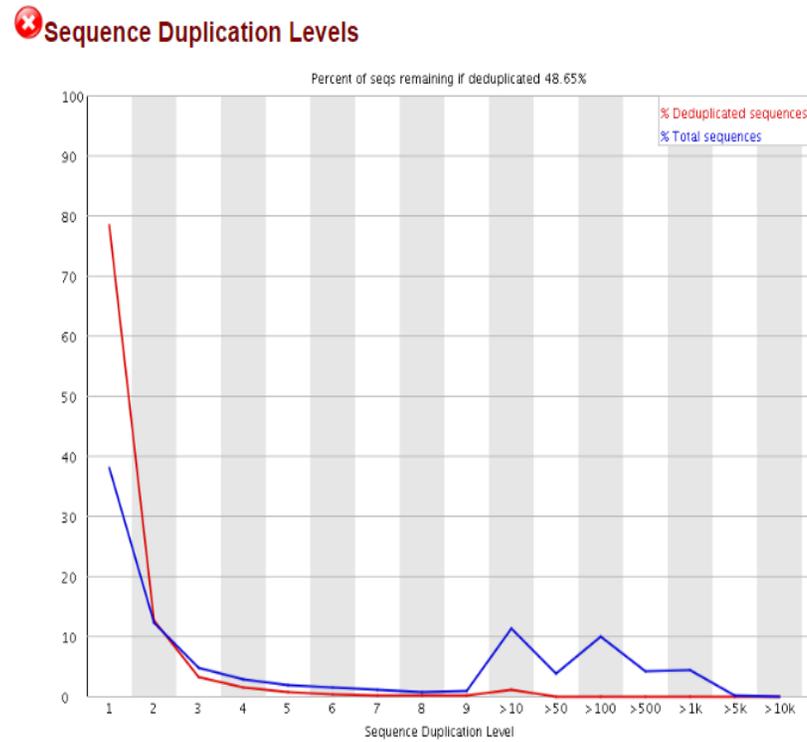
**2.3 g.**

 **Sequence Length Distribution**



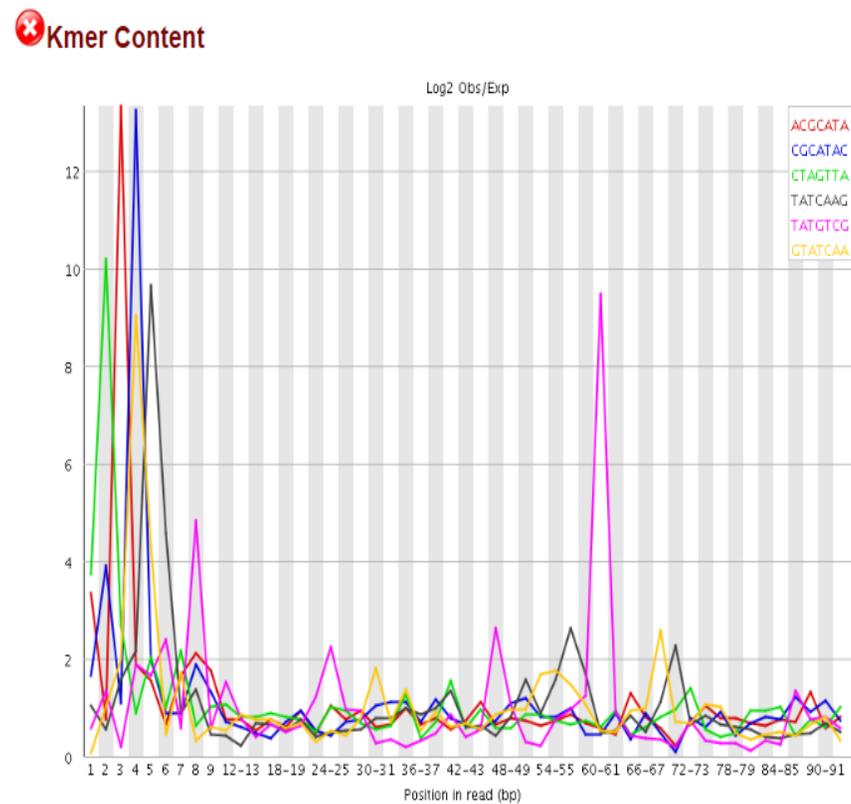
**2.3 h.**

Figure 2.3 i. The sequence duplication levels measure the degree of duplication for every sequence and plots sequences with different degrees of duplication. An error is issued if non-unique sequences are more than 50% of the total sequences.



**2.3 i.**

Figure 2.3 j. The k mer content module provides enrichment pattern about read length; generally it counts the enrichment of every 5-mer within the sequence library. It estimates an expected value based on the base content of the library at which this k mer should have been 'n' and calculates an observed/expected ratio for that k mer. A warning is raised if any k-mer is enriched more than 3-fold overall enrichment or a 5-fold enrichment at any given base.



2.3 j.

Figure 2.4 a. The coverage graph of mapped reads (contigs) for sample C13-03 chestnut lamprey generated from UCSC Genome Browser (Lamprey Assembly, WUGSC 7.0/petMar2).

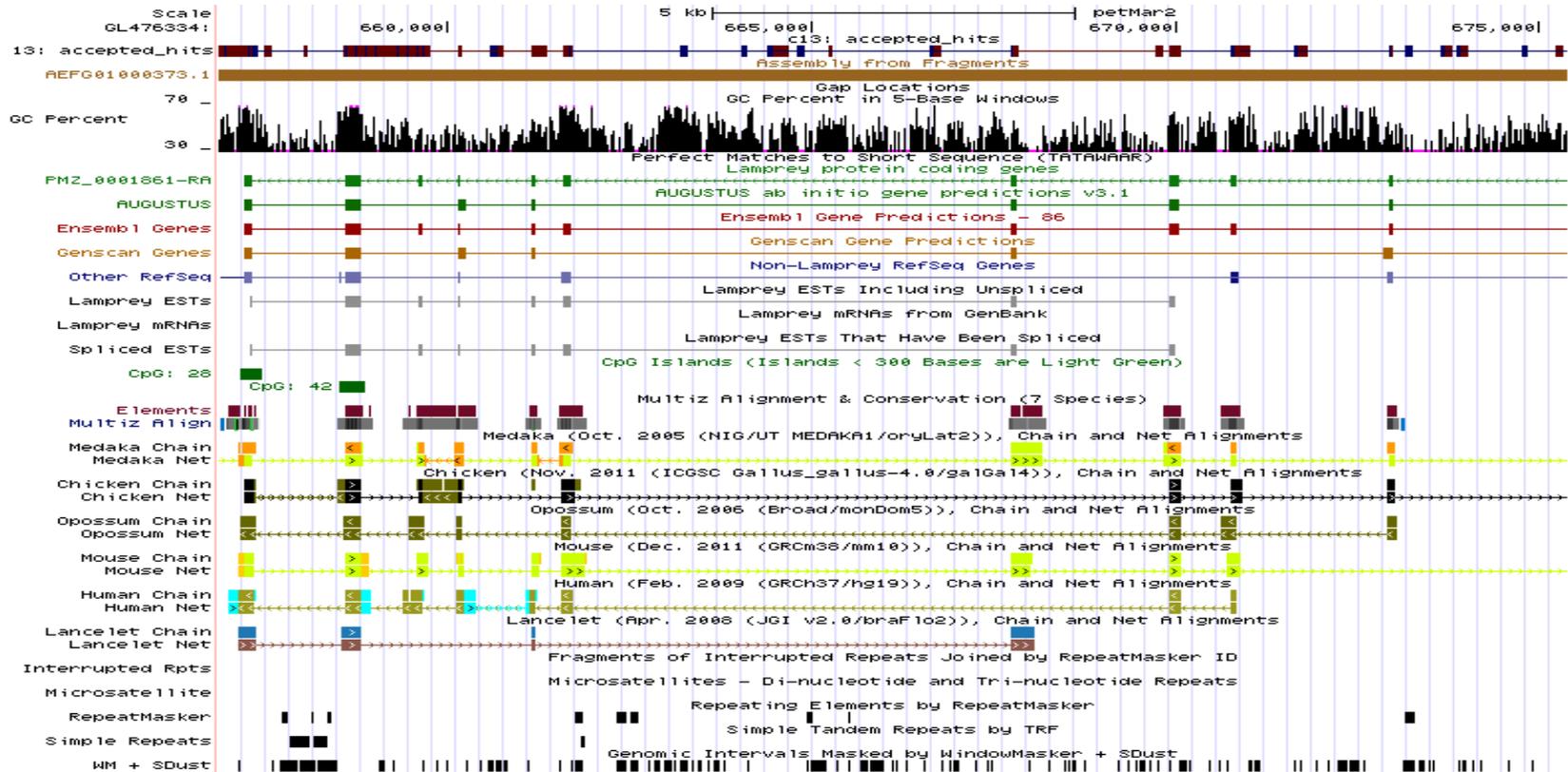


Figure 2.4 b. The coverage graph of mapped reads for sample C13-03 chestnut lamprey generated from Integrated Genome Browser. “The Accepted Hit File” from TopHat is used as an input which is represented by Y-axis and the length of reads is represented by X-axis.

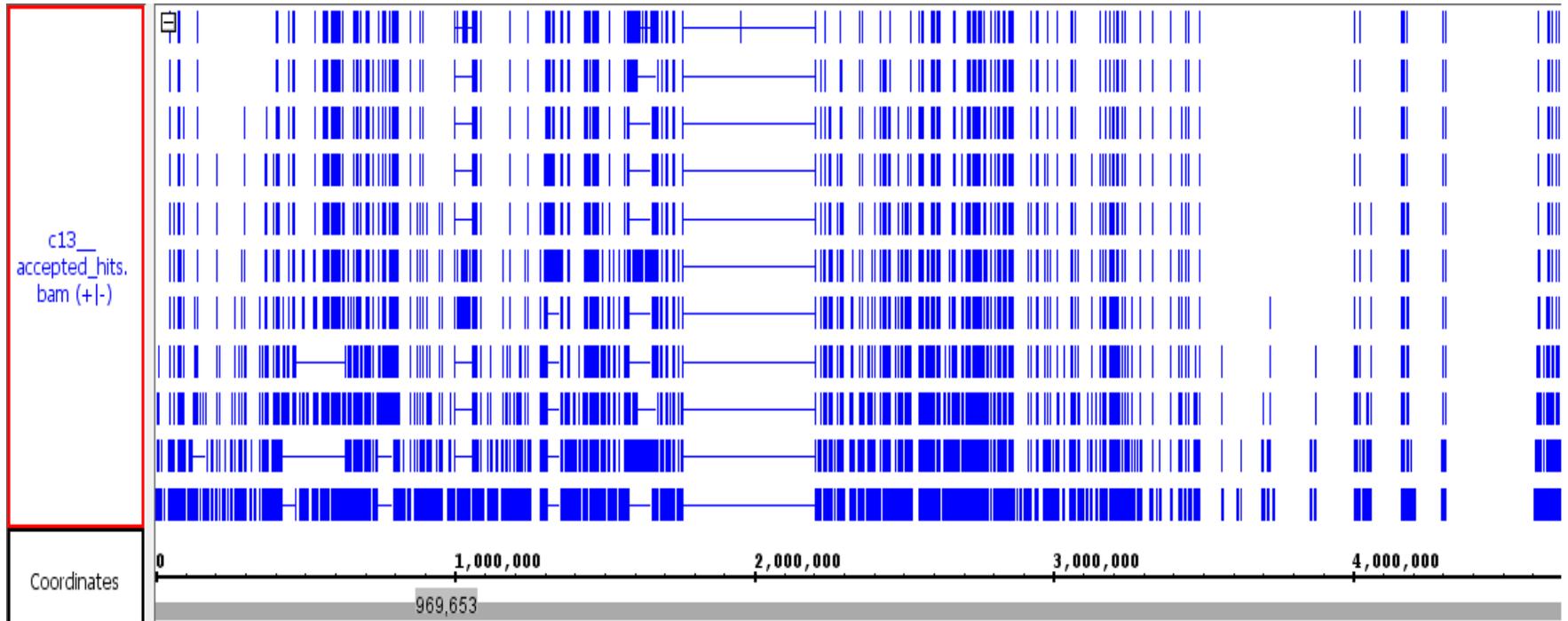


Figure 2.5 a. The coverage graph of mapped reads (contigs) for sample N17-1 northern brook lamprey generated from UCSC Genome Browser (Lamprey Assembly, WUGSC 7.0/petMar2).

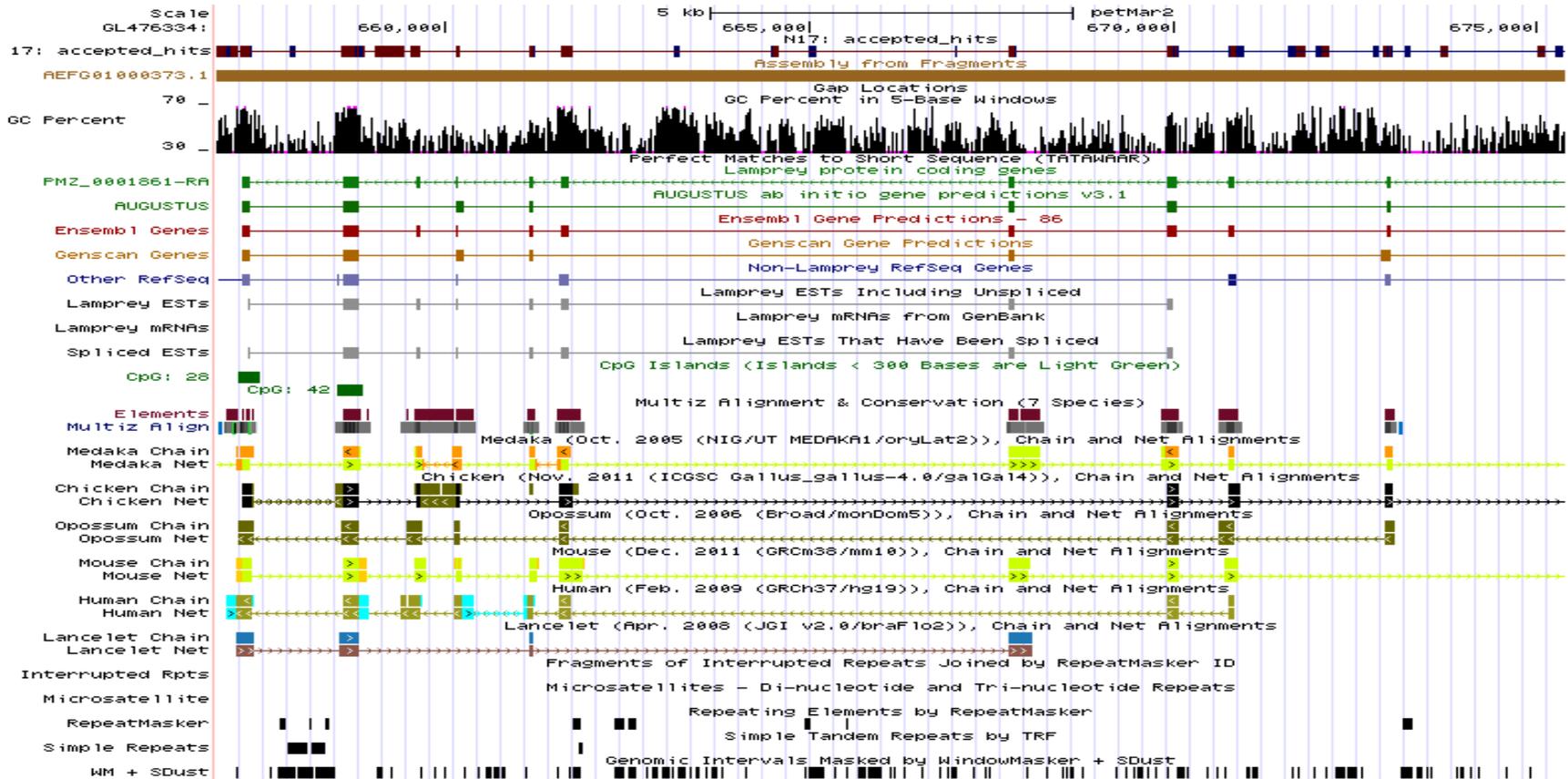


Figure 2.5 b. The coverage graph of mapped reads for sample N17 northern brook lamprey generated from Integrated Genome Browser. “The Accepted Hit File” from TopHat is used as an input which is represented by Y-axis and the length of reads is represented by X-axis.

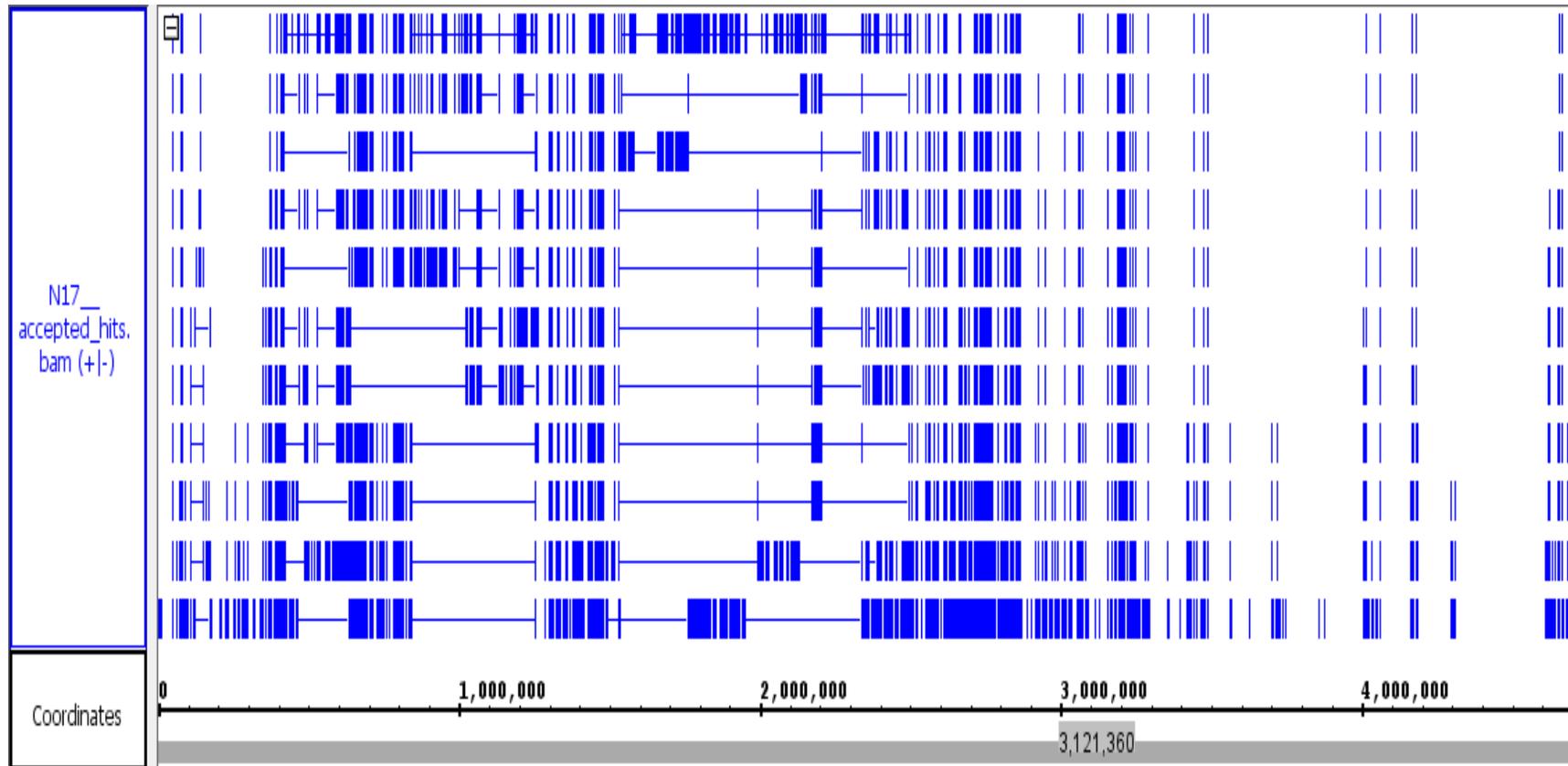


Figure 2.6. Principal Component Analysis (PCA) plot of pooled samples of chestnut lamprey (parasitic) samples (C13-O3, C20-03, IC1, IC2 & IC3) and northern brook lamprey (non-parasitic) samples (N1-10, N17-1, M01, N08 & S11) at three different time points of ovarian development (early stage (1-3), mid-stage (4) and, late stage (early and late vitellogenesis) from DESeq2 wherein **triangle** symbolizes parasitic lamprey and **rectangle** represents non-parasitic lamprey. The samples are plotted in the two-dimensional plane spanned by their first two principal components and is useful for visualizing the overall effect of experimental covariates and batch effects.

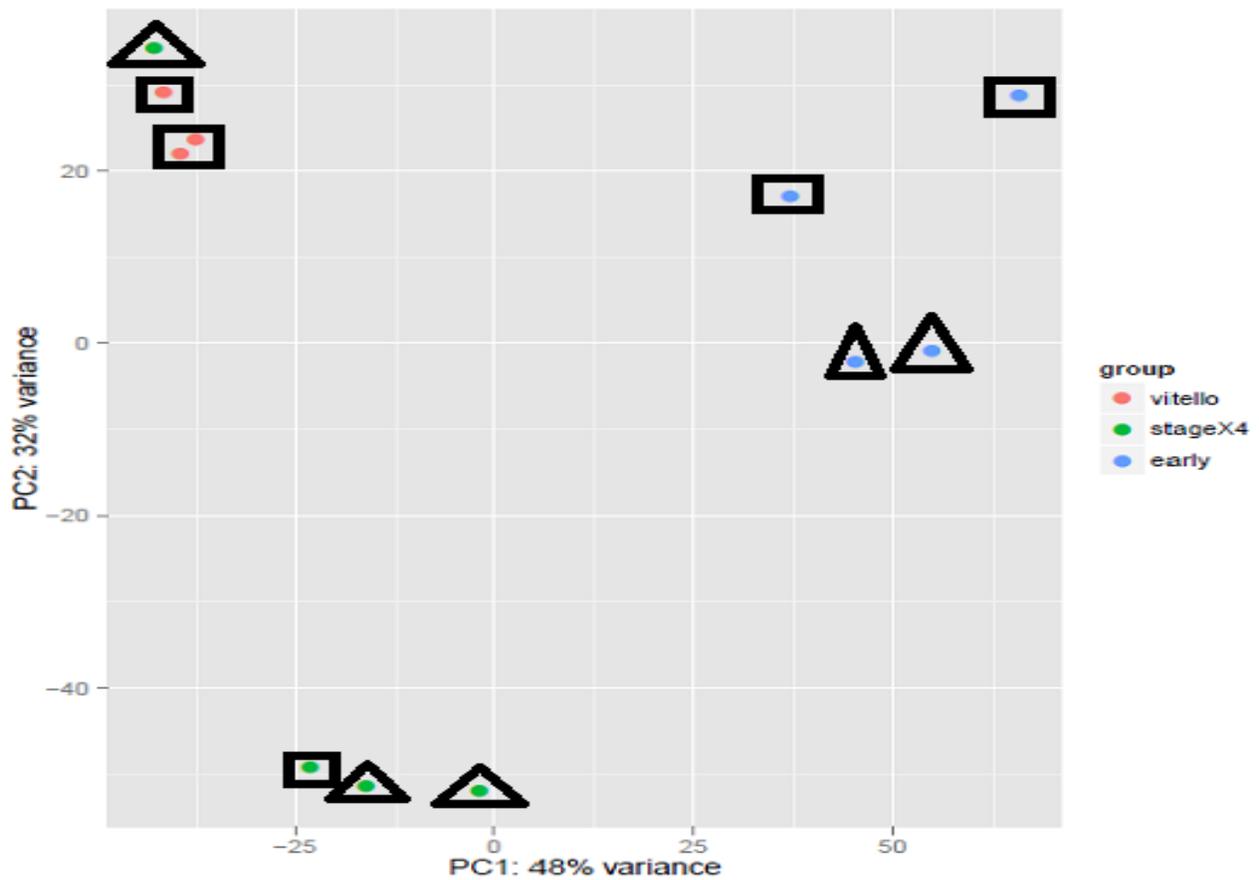


Figure 2.7. Heatmap of the sample-to sample distances of chestnut lamprey (parasitic) samples (C13-O3, C20-03, IC1, IC2 & IC3) and northern brook lamprey (non-parasitic) samples (N1-10, N17-1, M01, N08 & S11) at three different time points of ovarian development (early, mid, and late stages; Figure 1.4, Table 2.5) from DESeq2. The heatmap shows the Euclidean distance (value) between samples as calculated from the log transformed standardized counts. Darker colors refer to lower Euclidean distances. The five chestnut lamprey sample labels are outlined in green and the six northern brook lamprey are outlined in yellow. Based on the similarities and dissimilarities of genes expressed, the samples clustered in three blocks with darker blue squares (the first four samples, the next three, and the last four). The samples in the blocks are more similar to each other as compared to the samples in other blocks. In the extreme right top, three samples of northern brook lamprey of late gonadal stages clustered with one chestnut lamprey of mid-gonadal stage which indicates that these samples had similar gene expression profiles.

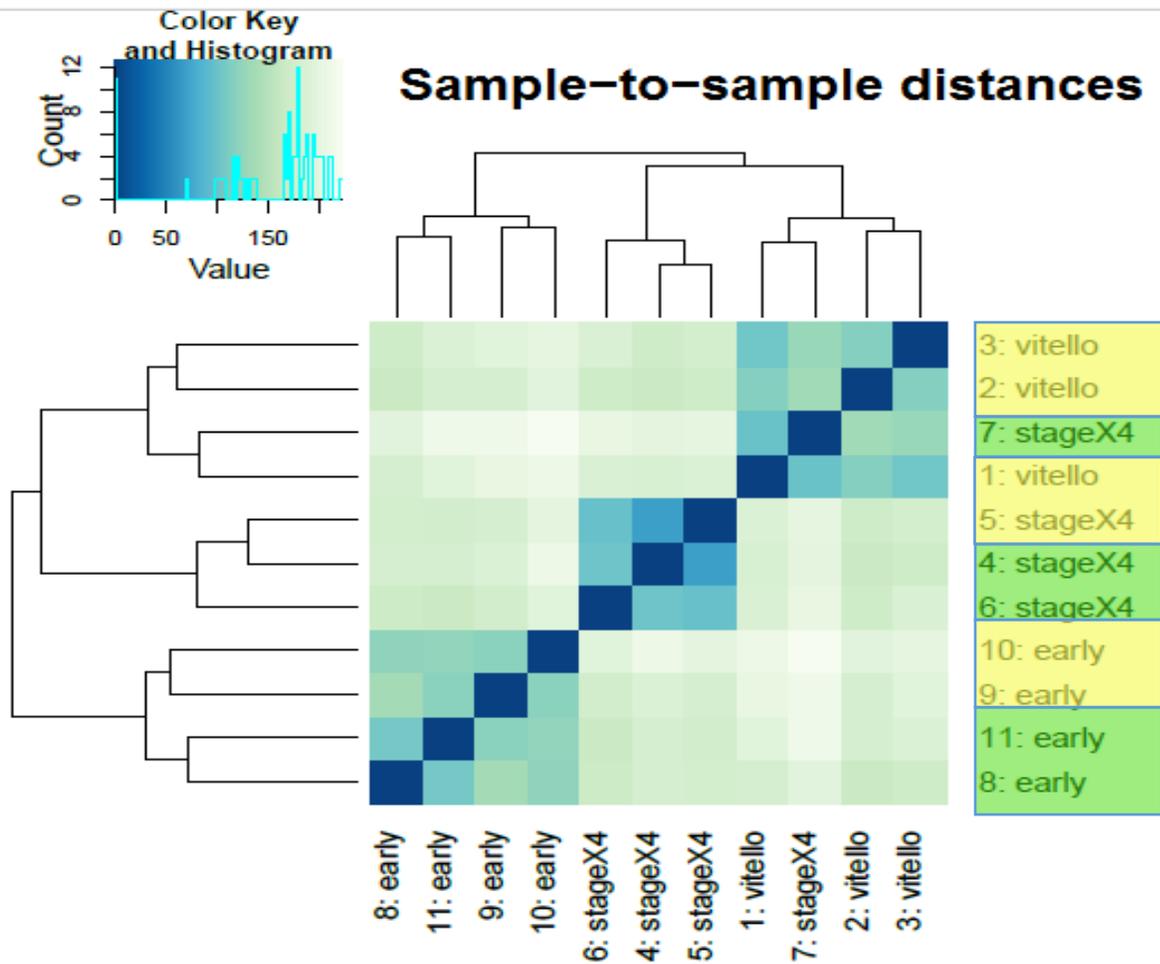


Figure 2.8. Dispersion estimate of chestnut lamprey (parasitic) samples (C13-O3, C20-03, IC1, IC2 & IC3) and northern brook lamprey (non-parasitic) samples (N1-10, N17-1, M01, N08 & S11) at three different time points of ovarian development (early, mid, and, late stages; Figure 1.4, Table 2.5). The graph shows the gene-wise estimates (black), the overall fitted model (red) based on sample size, variation among samples in gene-wise estimates and the number of coefficients (factors) in the model, and the final estimates (blue) after model correction. It estimates the final shrunk from the gene-wise estimates towards the fitted estimates. Depending on different factors such as sample size, the number of coefficients, the row mean and the variability of the gene-wise estimates the amount of shrinkage varies from more or less.

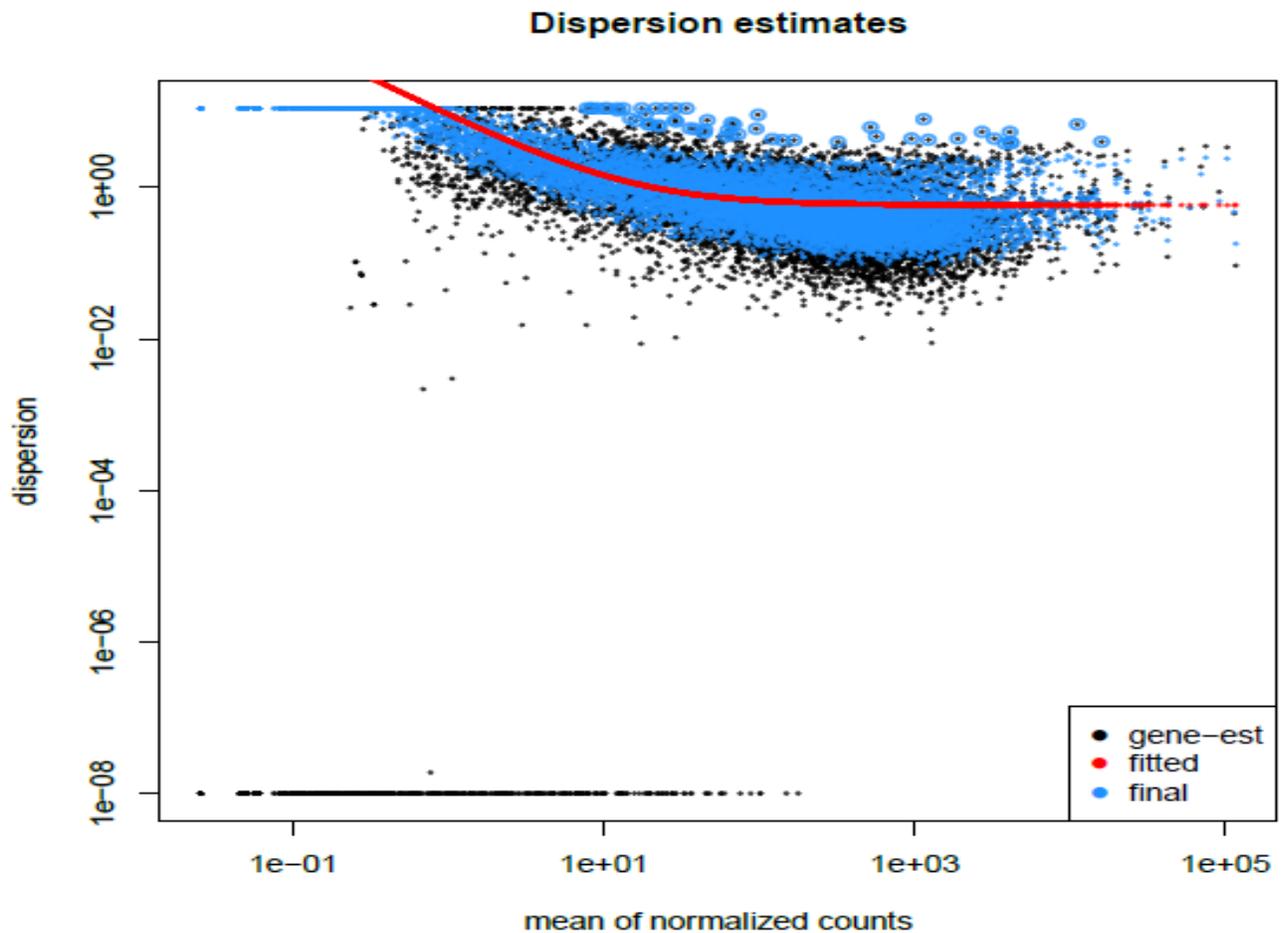


Figure 2.9. Histogram of chestnut lamprey (parasitic) samples (IC1, IC2 & IC3) and northern brook lamprey (non-parasitic) samples (M01, N08 & S11) during mid and, late stages based on  $p$ -values that are distributed more or less uniformly. The subsets that pass the filtering are shaded in blue while the area in khaki indicates those subsets that do not pass filtering.

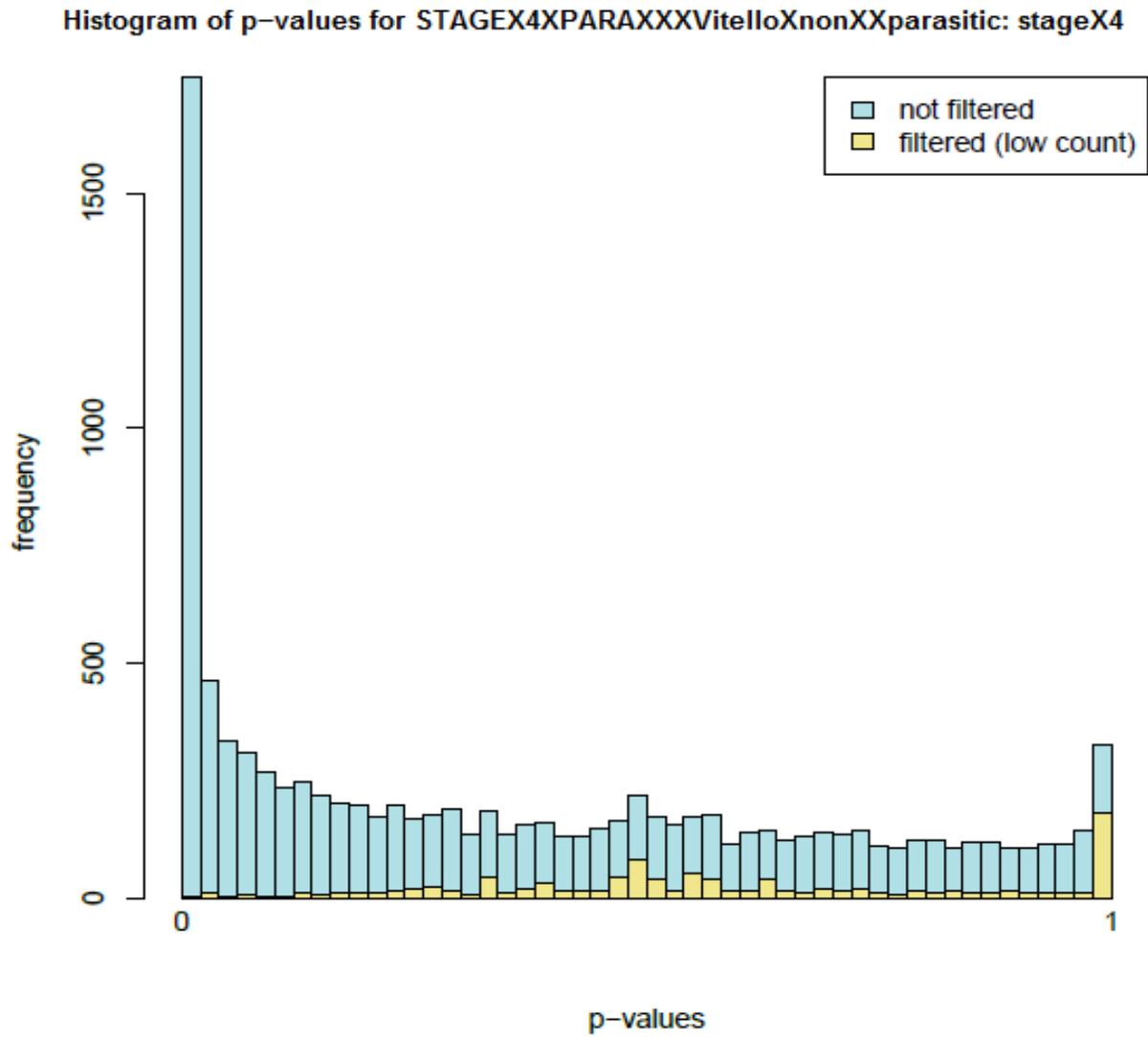


Figure 2.10. The MA (Mean difference) plot of chestnut lamprey (parasitic) samples (IC1, IC2 & IC3) and northern brook lamprey (non-parasitic) samples (M01, N08 & S11) during mid and, late stages from DESeq2. It estimates log<sub>2</sub>-fold change over the log-average mean of normalized counts for all the samples in the DESeq2 data set. If the adjusted *p*-value is less than 0.1, points are colored red. Points which lie out of the window area are plotted as open triangles pointing either up-regulated or down-regulated.

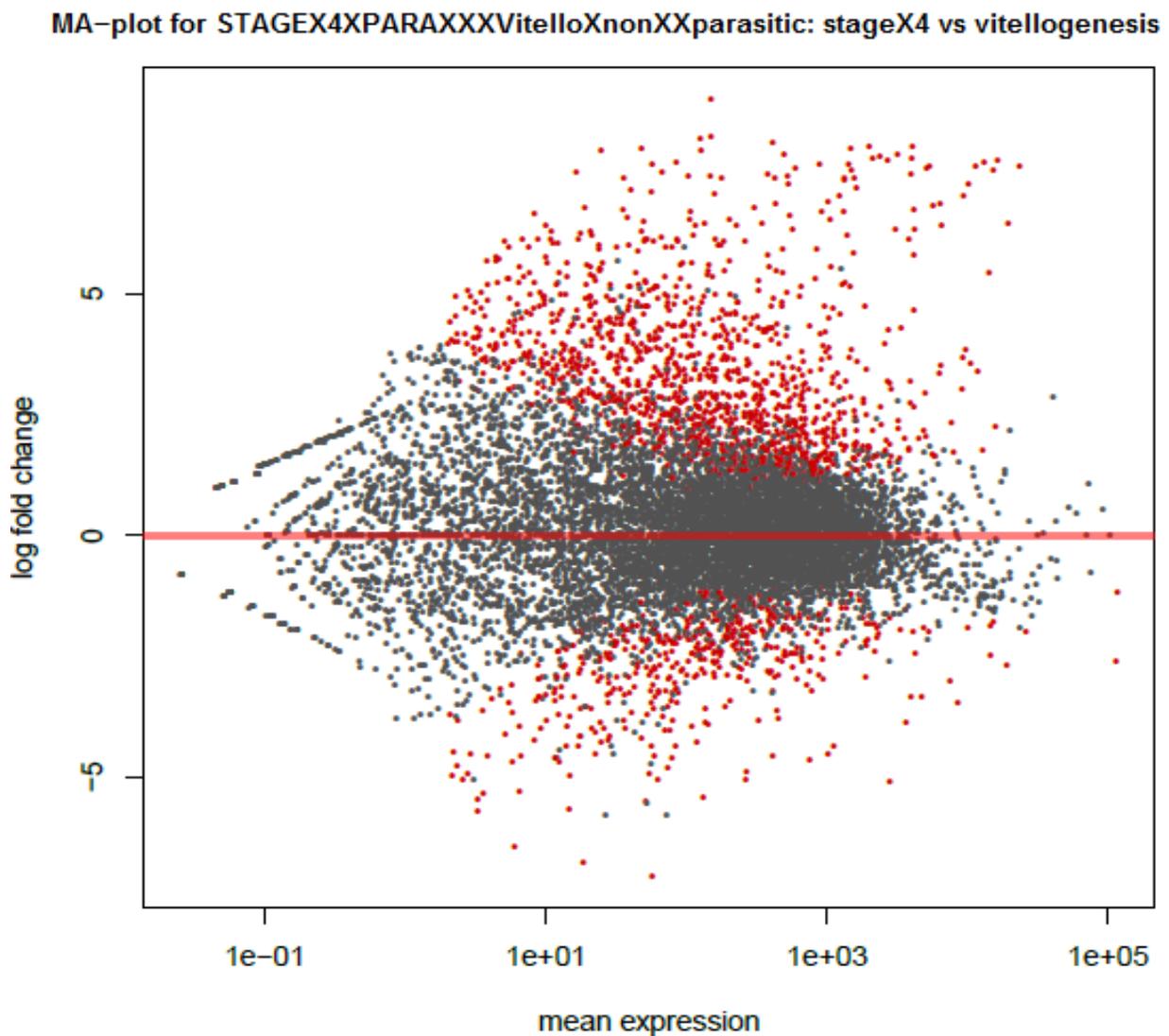


Figure 2.11. The “Scatterplot & Table” view output of up-regulated genes of zebrafish *Danio rerio* from the GOrilla output was given as input to REVIGO. The “Scatterplot” view gives an idea about the terms which are reduced due to redundancy and are showed as a cluster representative in a two-dimensional space by using semantic similarities. In the lower part, the table view “Lists” different processes related to GO terms. “Black denotes” cluster representatives and other cluster members are represented by gray letters. The parameters are easily customizable, “Bubble color” refers to the user provided  $p$ -value (legend in upper right-hand corner); “Size of the bubble” indicates the frequency of the GO term in the Gene Ontology Association (GOA) database.

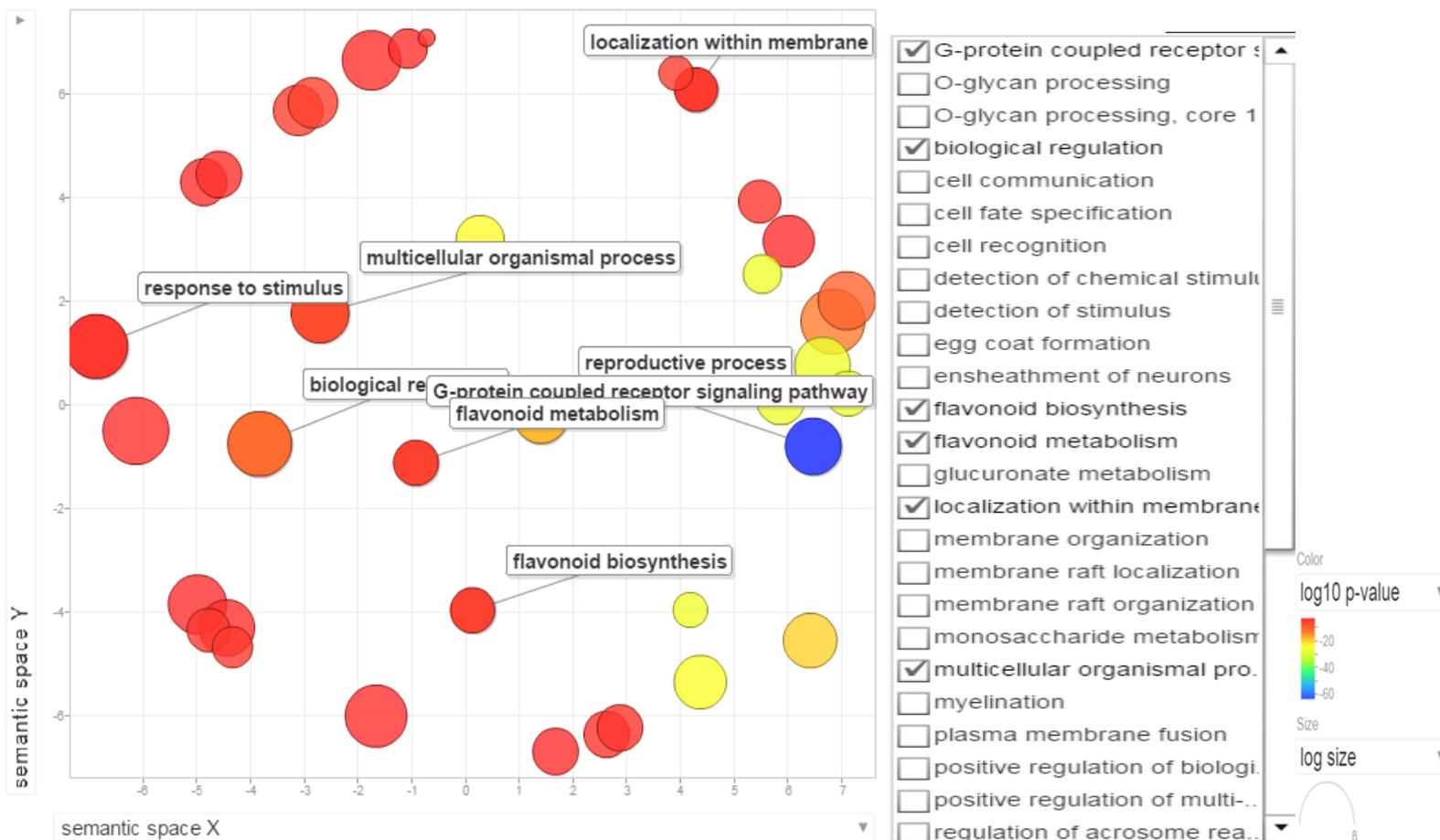


Figure 2.12. The “Interactive graph” view of up-regulated genes from REVIGO. “Bubble color” refers to the user provided  $p$ -value (legend in upper right-hand corner); “Size of the bubble” indicates the frequency of the GO term in the database. Edges are used for linking the highly relevant similar GO terms and line width indicates the degree of similarity.

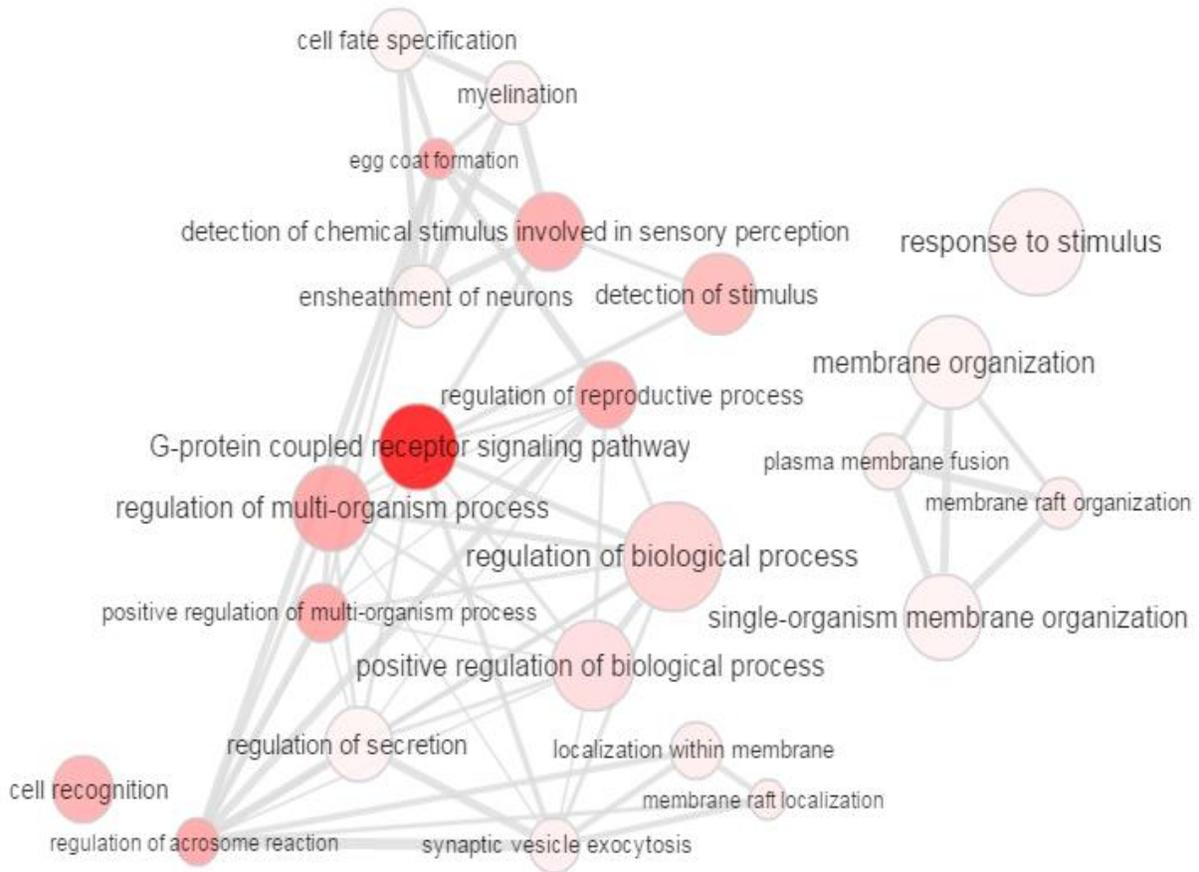


Figure 2.13. The “TreeMap” view of up-regulated genes from REVIGO where rectangle is used for representing different clusters. Clusters are joined together into ‘superclusters’ of related terms and are color coded. The size of the rectangle can be adjusted based on p-value or the size (frequency) of the GO terms in the Gene Ontology Association (GOA) database.

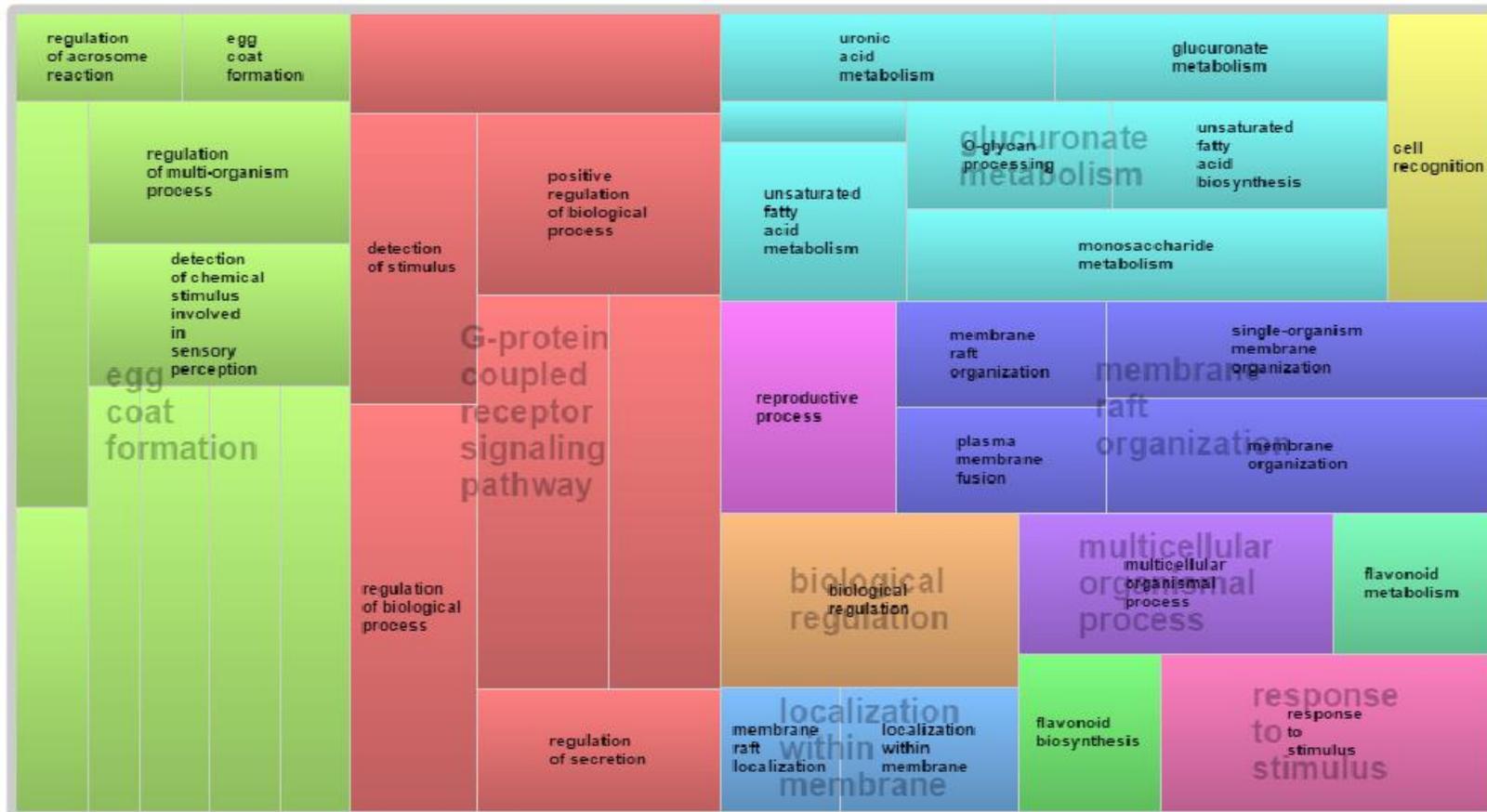
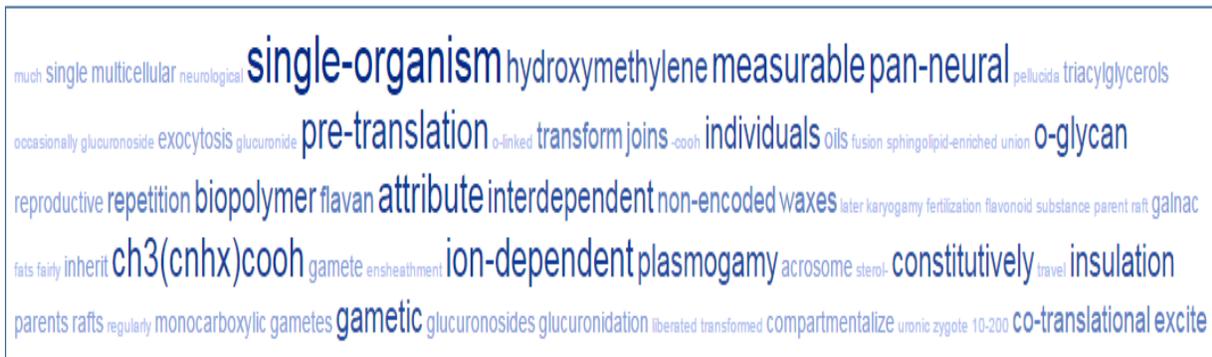


Figure 2.14. The “Tag Cloud” view of up-regulated genes from REVIGO. It displays words which are based on the relatively similar GO terms in the user-supplied list. Large and dark letters represent stronger overrepresentation. Underrepresented keywords or terms are not displayed in the Tag Cloud.



Frequent keywords within your set of GO terms:



Keywords that correlate with the value you provided alongside GO terms:



Figure 2.15. GOrilla analysis output where 2,730 down regulated genes of *Danio rerio* from the DESeq2 were ranked according to their differential expression based on *Padj* value and was given as input to GOrilla. The resulting GO terms were enriched and visualized using a DAG graphical representation and were assigned different colors based on their degree of enrichment. The nodes in the graph are clickable and provides additional information about the genes and their enrichment terms (GO).

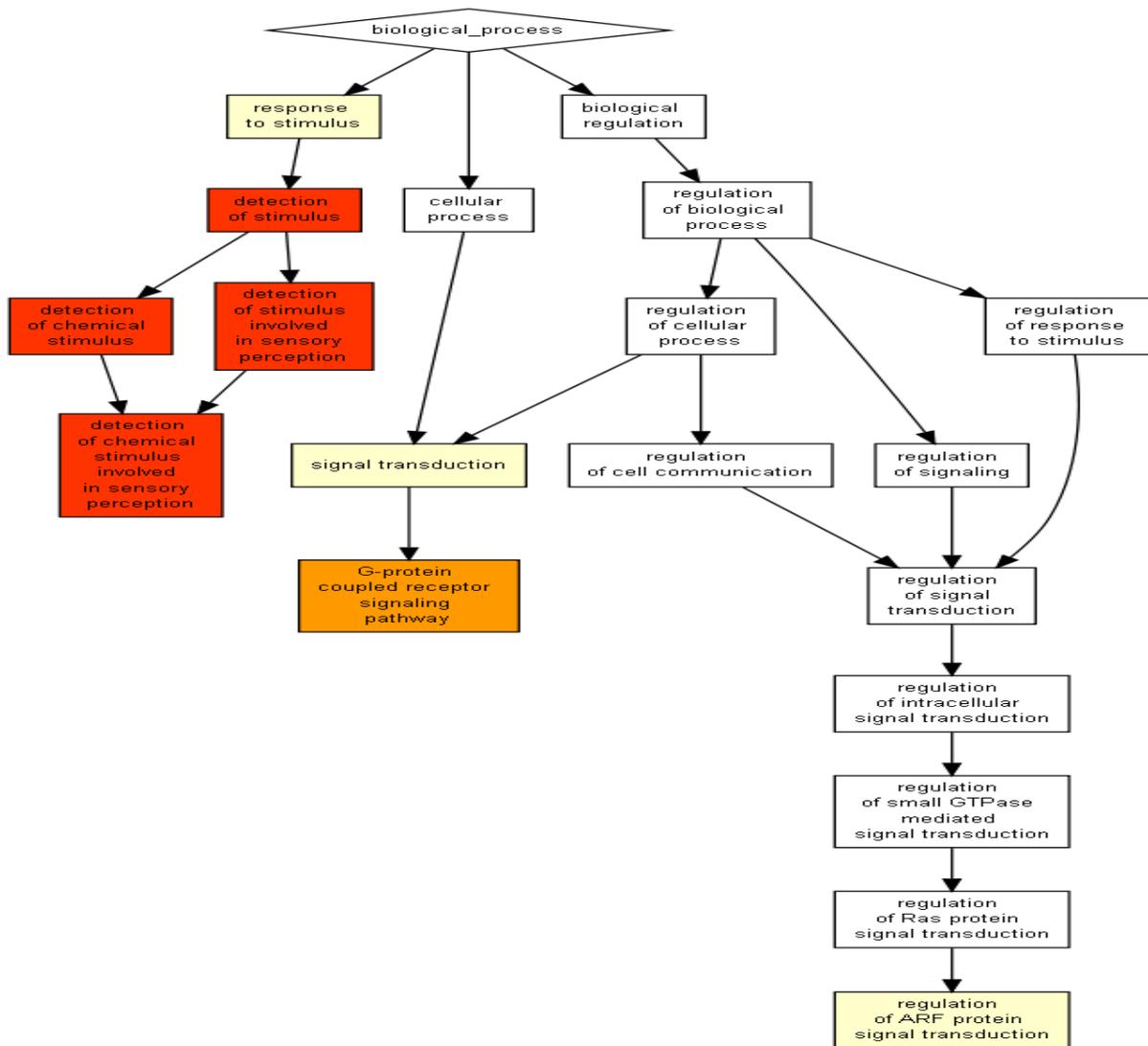


Figure 2.16. The “Scatterplot & Table” view output of down regulated genes of *Danio rerio* from the GOrilla output was given as input to REVIGO. The “Scatterplot” view gives an idea about the terms which are reduced due to redundancy and are showed as a cluster representative in a two-dimensional space by using semantic similarities. In the lower part, the table view “Lists” different processes related to GO terms. “Black denotes” cluster representatives and other cluster members are represented by gray letters. The parameters are easily customizable, “Bubble color” refers to the user provided  $p$ -value (legend in upper right-hand corner); “Size of the bubble” indicates the frequency of the GO term in the Gene Ontology Association (GOA) database

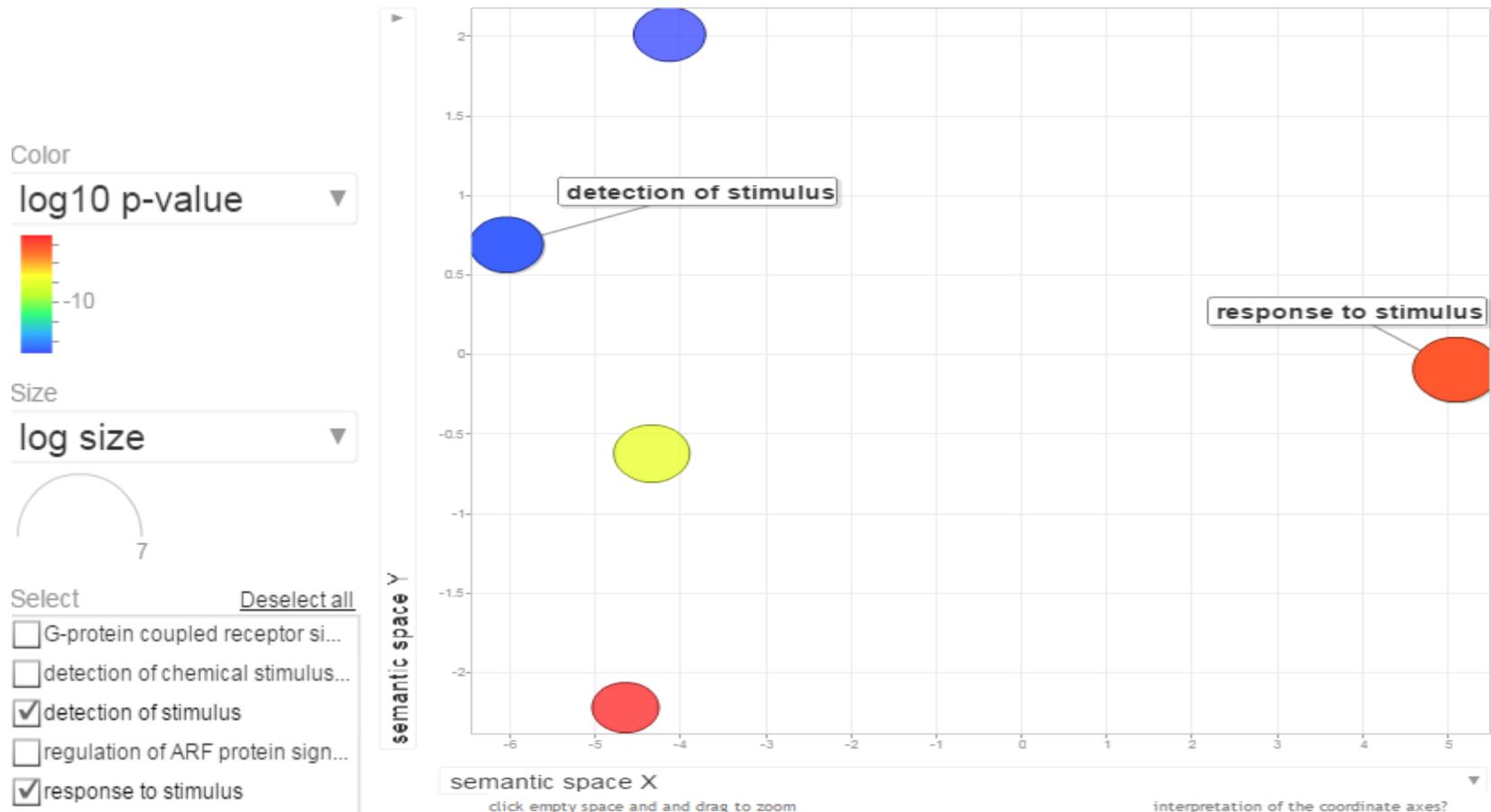


Figure 2.17. The “Interactive graph” view of down regulated genes from REVIGO. “Bubble color” refers to the user provided  $p$ -value (legend in upper right-hand corner); “Size of the bubble” indicates the frequency of the GO term in the database. Edges are used for linking the highly relevant similar GO terms and line width indicates the degree of similarity.

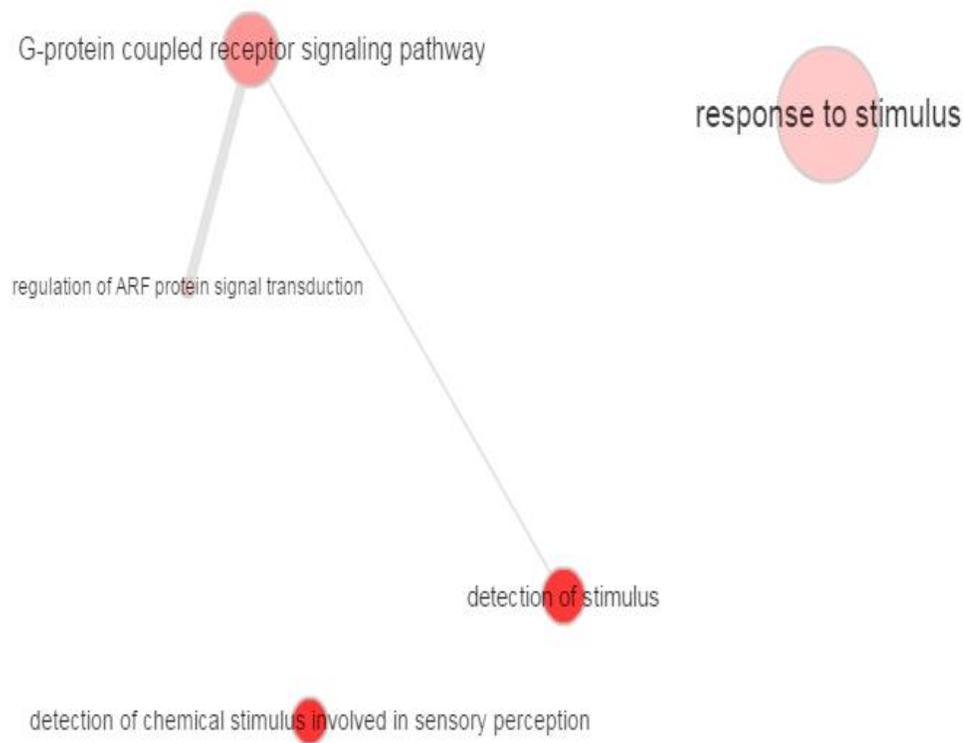


Figure 2.18. The “TreeMap” view of down regulated genes from REVIGO where rectangle is used for representing different clusters. Clusters are joined together into ‘superclusters’ of related terms and are color coded. The size of the rectangle can be adjusted based on *p*-value or the size (frequency) of the GO terms in the Gene Ontology Association (GOA) database.



Figure 2.19. The “Tag Cloud” view of down regulated genes from REVIGO. It displays words which are based on the relatively similar GO terms in the user-supplied list. Large and dark letters represent stronger overrepresentation. Underrepresented keywords or terms are not displayed in the Tag Cloud.



Frequent keywords within your set of GO terms:

single multicellular communication **neurological** restricted downstream single-organism events matrix state receive expression trigger light transmission signal cascade ends stimulus than organ carried switching signalling convert signaling modulates required **arf** substance interactions tissues member occurring mediates mediated sensory physiological organs transduction surroundings series ras organism signals molecular level environment receptors **constitutively** organismal activated **gtp-bound** begins further **gtpase-mediated** active characterize extracellular received absence gtpase pathway process converted cellular recognize detection **perception** attachment



Keywords that correlate with the value you provided alongside GO terms:

proceeds expression **received** gtp production taste dissociates enzyme gpcr results promoting **chemical** movement converts exchange transcription cellular extracellular secretion ends activating receptor-ligand electric begins gamma-subunits cell interaction **molecular** agonist organism signals further **detection** associated touch sound alpha-g-protein then heterotrimeric smell g-protein result **series** complex gtp-bound **events** gdp **perception** stimulus chemoperception sensing response **substance** receptor activated signalling beta transmit pathway absence **converted** g-protein-coupled alpha-subunit signaling physiological **sensory** downstream coupled basal **part**

Figure 2.20. Graphs of genes like *oxytocin receptor*, *zona pellucida glycoprotein 3a*, *neuropeptide Y receptor Y8b* and *Wilms tumor 1 associated protein* reported by genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts for both the species. All these genes (a, b, c and d) showed an increased expression in gonadal stages 2, 3 & 4 in non-parasitic northern brook lamprey.

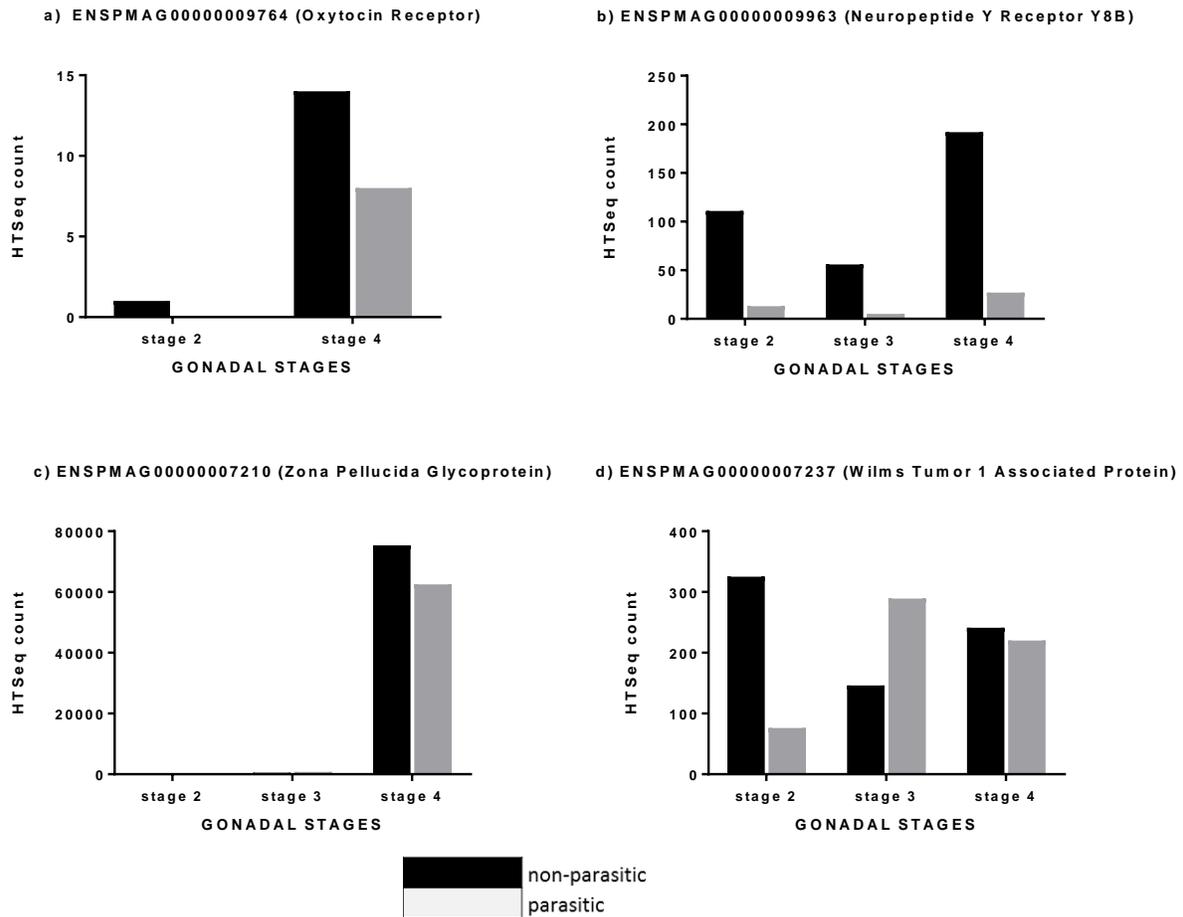


Figure 2.21. Graphs of genes like *thyroid stimulating hormone receptor*, *adrenoceptor alpha* and *estrogen receptor 2*, reported by genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts for both the species. All these genes (a, b and c) showed an increased expression in gonadal stage 4 in parasitic chestnut lamprey.

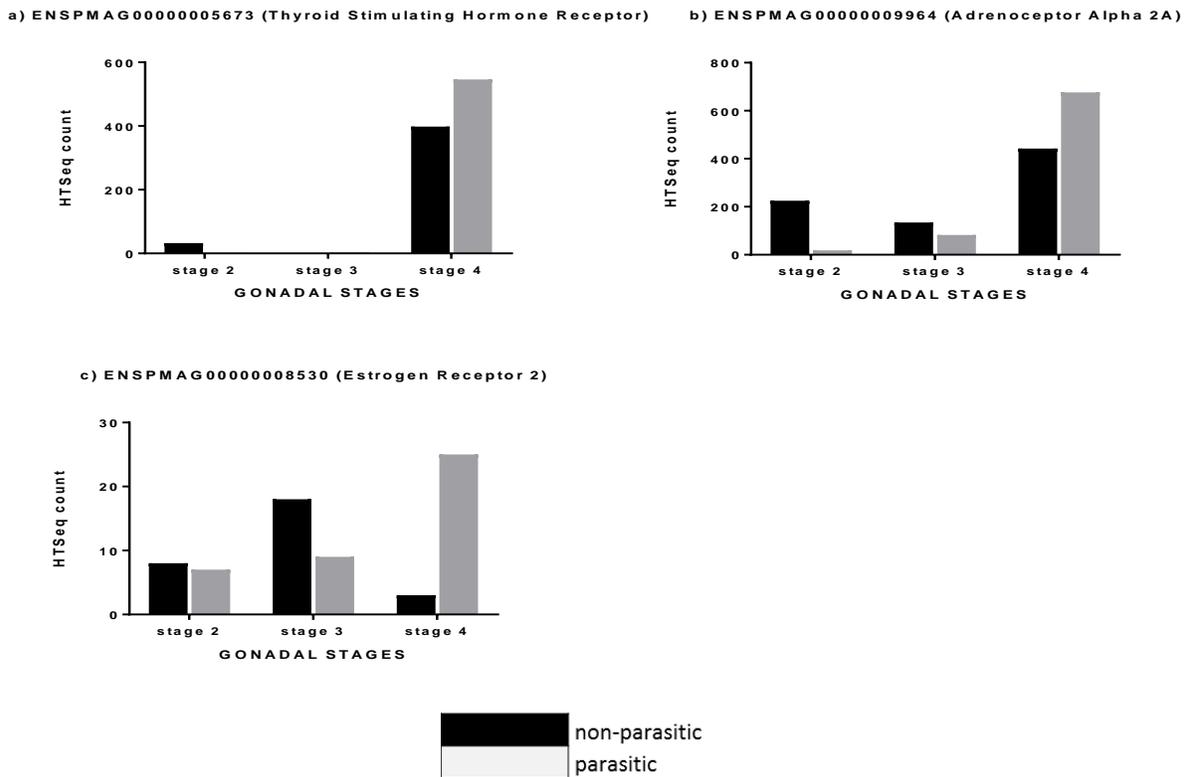


Figure 2.22. Graphs of genes like *calcitonin receptor* and *forkhead box 12*, *SRY* (*sex determining region Y*)-*box 2*, reported by genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts for both the species. Both are expressed in gonadal stages 2, 3 and 4 in parasitic chestnut lamprey.

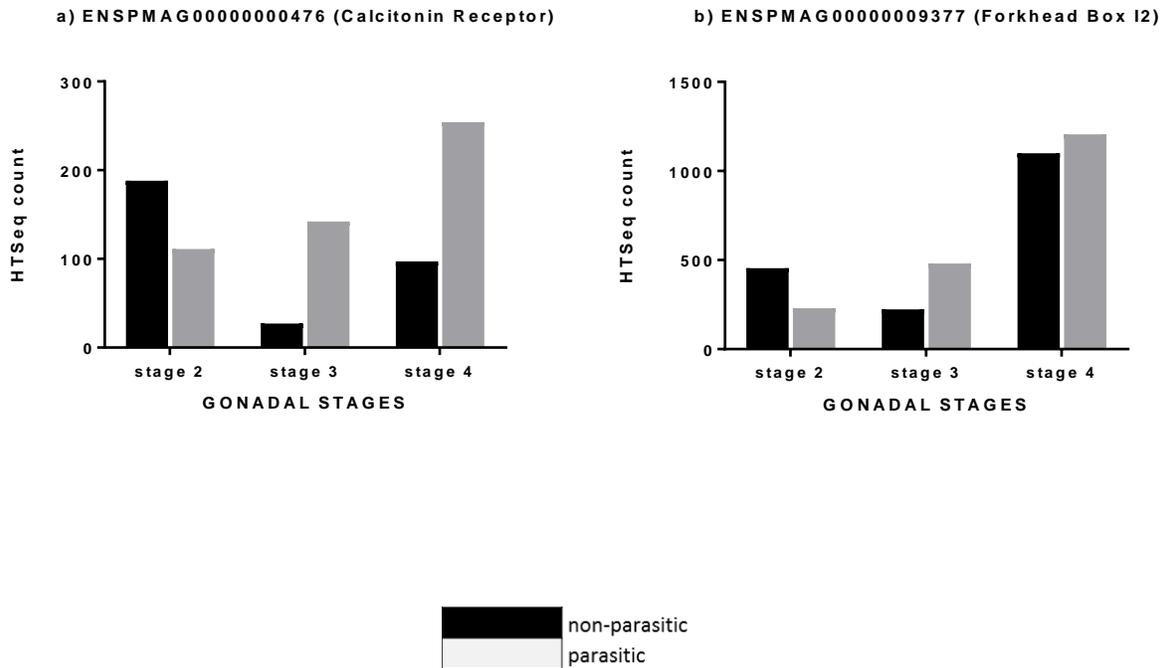


Figure 2.23. Graphs of insulin family genes- rxfp1 (novel), igf1r or insrr and rxfp3-1 reported by the genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts for both the species. All these genes (a, b, c, d and, e) showed an increased expression in gonadal stages 2, 3 & 4 in northern brook lamprey; the putative name of these genes are assigned based on genomic locations and data mining

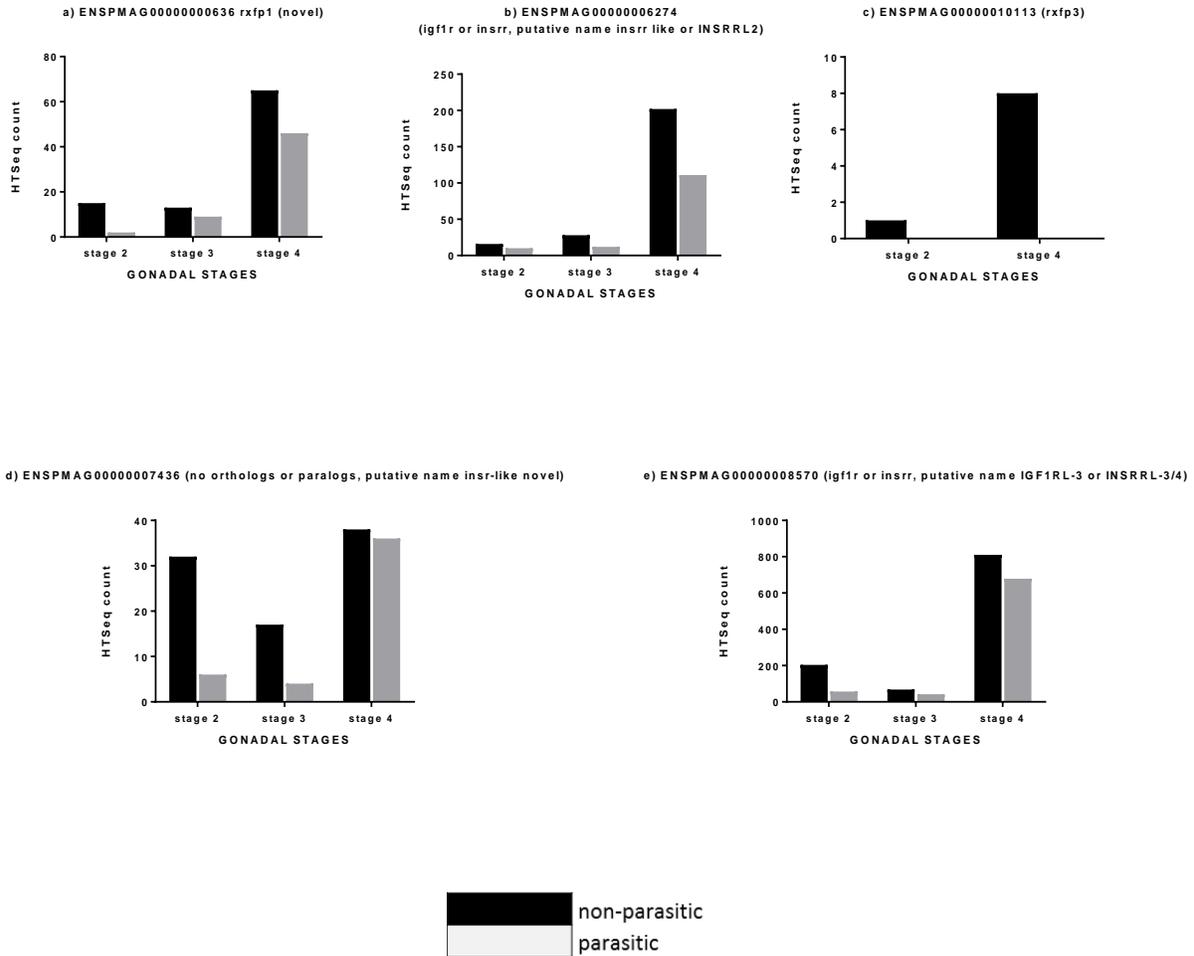
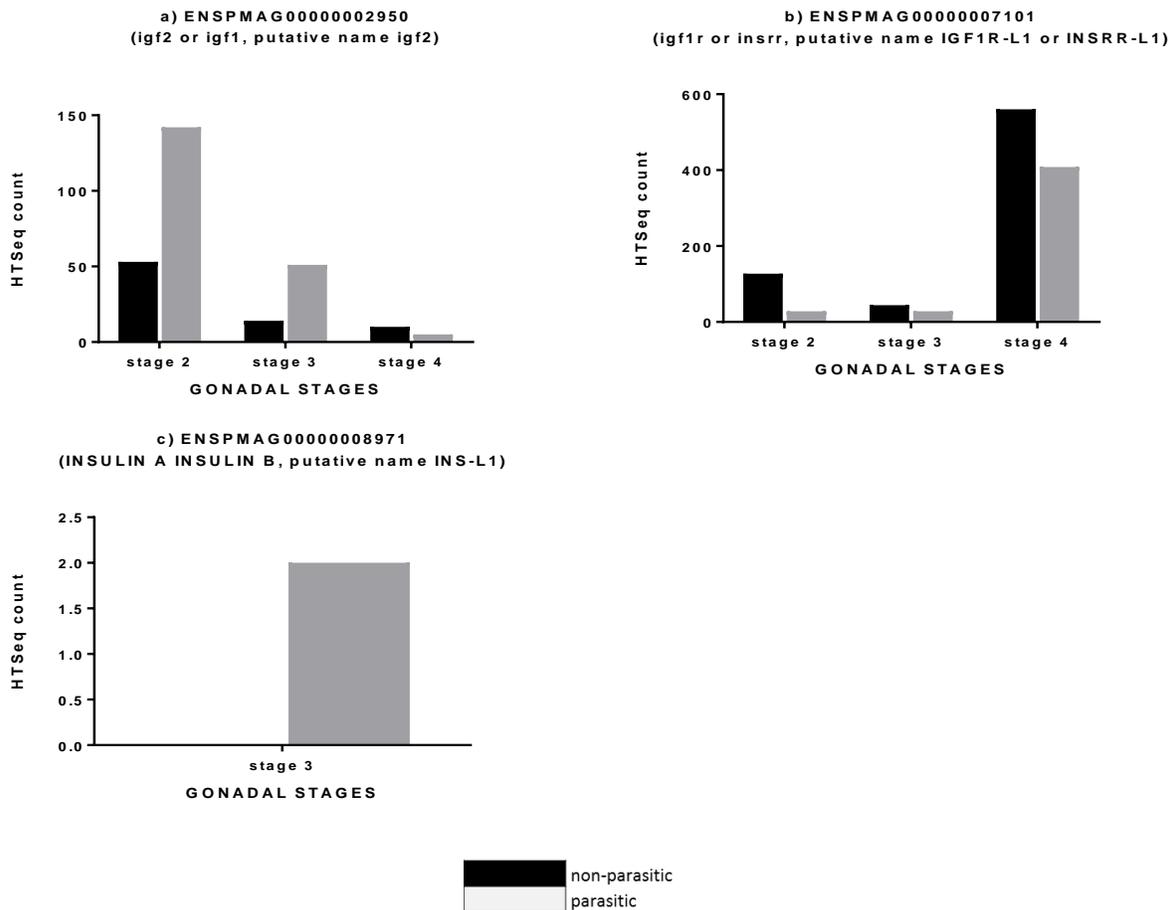


Figure 2.24. Graphs of insulin family genes- igf2 or igf1, igf1r or insrr and INSULIN A INSULIN B reported by genome-guided assembly, where the x-axis represented the gonadal stages and y-axis represented the HTSeq gene counts. In (a), igf2-igf1 gene was expressed in parasitic chestnut lamprey during gonadal stages 2 and 3 but in gonadal stage 4 a sharp decrease in gene expression was observed in chestnut lamprey; in (b), igf1r or insrr was highly expressed in non-parasitic northern brook lamprey during gonadal stages 2, 3 and 4 as compared to the chestnut lamprey; and in (c) INSULIN A INSULIN B was only expressed in chestnut lamprey during gonadal stage 3. The putative name was assigned to these ID's based on bioinformatic data mining.



## Chapter 3

### ***De novo* approach for generating reference transcriptome of lampreys for identifying novel and specific genes across different developmental stages**

#### **3.1. Abstract**

In this chapter, RNA reads of ovarian samples from two species of lamprey were used to reconstruct full length transcripts in a process known as ***de novo*** (note that terms that are bolded on first use within each chapter are defined in the glossary in Appendix 2) sequence assembly. The pipeline was designed to assemble reads without a **reference genome**, which is a strategy used for diverse purposes including genome annotation, identification of **novel** transcripts and alternatively spliced transcripts including tissue specific variants. To this end, the complete set of RNA-Seq data from the “best” ovarian samples obtained from the two focal species of this thesis, *Ichthyomyzon castaneus* and *I. fossor*, were assembled *de novo* using the program Trinity. The *de novo* assembled transcriptome for each species was then used as a reference to map and to count the number of times each transcript was identified in all the samples for each species using the program RSEM (RNA-Seq by Expectation Maximization). Finally, to obtain tentative gene identities, the *de novo* assembled transcripts were queried against a custom-made BLAST database containing all the RefSeq entries from 10 chordates with well-annotated genomes, and the names of the top three best hits were retained.

To test the validity of the pipeline and to assess its ability to identify novel transcripts and to provide **putative** ID's, we specifically looked for the number and identity of the insulin superfamily of genes in the output based on ***a priori*** expectations and then compared the insulin

superfamily genes and their counts as identified by the reference-guided (Chapter 2) and *de novo* (Chapter 3) approaches. The strengths and weaknesses of the genome-guided approach versus *de novo* assembly are briefly discussed.

### 3.2. Introduction

With recent advances in high-throughput sequencing, RNA-Sequencing has become a popular tool to study gene expression in non-model organisms (e.g., Bernatchez *et al.*, 2010; Wolf *et al.*, 2010), including understanding changes in gene expression over developmental stages and between life history types. For example, RNA sequencing (RNA-Seq) in lake whitefish *Coregonus clupeaformis* determined that adults of the dwarf form overexpressed more genes related to energy metabolism, whereas the adults of the normal form overexpressed more genes related to protein synthesis; these differences correspond to metabolic and growth differences between the life history types (Jeukens *et al.*, 2010). RNA sequence is widely used to assemble transcriptomes for understanding changes in gene expression in different stages or treatments, but it is also useful for single nucleotide polymorphism (SNP) discovery even in the absence of a reference genome or genome sequences (Quinn *et al.*, 2008; Vera *et al.*, 2008; Kristiansson *et al.*, 2009; Renaut *et al.*, 2010).

Primarily, raw RNA-Seq data can be converted into transcripts by using two methods: through genome-guided assembly or via *de novo* assembly (Hass & Zody, 2010; Martin & Wang, 2011). The genome-guided assembly method is widely used for model organisms with a well assembled reference genome and both public and commercial assembly packages are available to perform robust transcriptomic analyses using a reference genome to test for differences in gene expression, SNP discovery, alternative splicing and other read outs (Trapnell *et al.*, 2010; Guttman *et al.*, 2012). However, the kinds of analyses pursued typically depend on the study questions; in fact, variation in the quality of the reference genome, the relatedness of the species for which RNA-Seq data is available (and of interest) compared to the one for which

the reference genome is available, and the quality of the annotation of the reference genome all affect the kinds of analyses that can be performed using reference-based transcriptomic analyses.

Thus, *de novo* transcriptome assembly is often required (Hass *et al.*, 2013) for studies performed on taxa not closely related to a reference genome, or when the annotation of the reference genome is incomplete. *De novo* analysis is also used by genome consortia groups as part of the essential pipeline for annotating genomes and typically genome consortia perform deep RNA-Sequencing on a range of tissues to cover as much of the transcriptome as possible to perform gene annotation. Several software packages and tools are now available for *de novo* assembly of RNA-Seq data. Trans-ABYSS (Robertson *et al.*, 2010), Velvet-Oases (Schulz *et al.*, 2012), SOAPdenovo-trans (Luo *et al.*, 2012), and Trinity (Grabherr *et al.*, 2011) are some of them.

### 3.3. Goals of *de novo* assembly

RNA-Seq data is typically used for three different types of analyses a) read alignment; b) transcript assembly and/or genome annotation; and c) transcript and gene quantification. *De novo* assembly is often used to **annotate** genomes which have been fully sequenced, and for this purpose transcriptomic data from multiple tissues is usually obtained to ensure that most of the genes, and alternative transcripts, in the genome are contained in the RNA library. However, it is often challenging to assemble and identify reads obtained from *de novo* assembly because the reads contain 5' and 3' untranslated regions (UTRs) which are often longer than the genes themselves and which align to non-coding segments of the genome.

In this chapter, a *de novo* assembly pipeline was used in which raw reads of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* of different gonadal

stages were merged to generate a “reference ovarian transcriptome” of each species by using the program Trinity. These *de novo* assembled transcriptomes were then used as a reference to align and obtain normalized counts of transcripts from each developmental stage using the program RSEM (RNA-Seq by Expectation Maximization). Each *de novo* assembled transcript has a unique **Trinity ID**, but since no annotation was involved in the assembly, we aligned the transcripts using the BLASTX algorithm (Altschul *et al.*, 1990) against those contained in a customized BLAST database consisting of sequences from 10 chordate genomes. This allowed us to obtain the name of the most closely related transcript in the customized database and thereby assign a putative or tentative ID to the *de novo* assembled transcripts (Figure 3.1).

To test the functionality of this *de novo* assembly pipeline, I used insulin family genes because of *a priori* knowledge of their involvement in ovarian development (Table 3.3). Insulin family genes, relaxin family peptide receptors (Rxfps) and their ligands, relaxin (RLn) and insulin-like (Insl) peptides play important roles in reproductive and neuroendocrine processes in mammals, and have evolved a range of diverse functions in vertebrate neuroendocrine regulation and reproduction (Good *et al.*, 2014). Relaxin family peptides are very different from insulin and IGFs, and are involved in reproduction and neuroendocrine regulation whereas insulin and IGFs play important roles in carbohydrate and fat metabolism, growth and development (Duan *et al.*, 1997, Nakamura *et al.*, 1998; Lu *et al.*, 2005; Wood *et al.*, 2005; Good *et al.*, 2012). Research indicates that some early deuterostomes harbored relaxin-like peptides, Rxfp1/2, which played an important role in regulating reproduction approximately 840 million years ago (Good *et al.*, 2014), but in lampreys, this gene family is not well studied and characterized. Based on the evolutionary history of RLN/INSL peptides and their two classes of GPCR receptors throughout vertebrate history, Yegorov & Good (2012) demonstrated that there was a single RLN/INSL

molecule, a single RXFP3/4 ancestral molecule and one RXFP1/2 receptor in the vertebrate ancestor. Following 2R, there were a total of four RLN/INSL peptides (namely, RLN3, RLN, INSL3 and INSL5), four RXFP 3/4 type receptors (RXFP3-1, 3-2, 3-3 and RXFP3-4) and three RXFP1/2 type receptors (RXFP1, RXFP2, and RXFP2-like) (Yegorov & Good, 2012). Although this evolutionary model has been confirmed in later diverging vertebrates, such as spotted gar *Lepisosteus oculatus* and coelacanth *Latimeria chalumnae*, in lampreys it could not be verified, in large part due to the incomplete annotation and small size of genomic **contigs** available in the first drafts of the lamprey genome (Yegorov *et al.*, 2014). Nevertheless, based on the evolutionary model presented in Yegorov *et al.* (2014), the lamprey genome should contain four RLN/INSL peptides (RLN, RLN3, INSL3, INSL5) and seven RXFP (RXFP1, RXFP2, RXFP2-like, RXFP3-1, RXFP3-2, RXFP3-3 and RXFP3-4) receptors barring no gene loss.

Unpublished work in Dr. Good's laboratory at the University of Winnipeg also provides insight into the early evolution of the INS/IGF peptides and their RTKs. Bioinformatic data mining and analysis of early insulin like peptide (ILPs) and the insulin-related receptors suggests that the vertebrate ancestor harbored single INS and IGF genes and a single INS, and at least two but probably three IGF genes were retained, giving rise to INS/IGF2 (which are linked), IGF1, and IGF3 in the post 2R vertebrate genome. Additionally, analysis suggests that there was a single insulin related RTK receptor in the vertebrate ancestor, but this receptor diversified during 2R to give rise to three receptors in the post 2R genome (INSR, INSRR, and IGF1R). This suggests that the lamprey genome could contain up to four INS/IGF ligands, and three receptors for these genes (INSR, INSRR and IGF1R). Previous difficulties in determining the full repertoire of insulin superfamily genes in lampreys appear to be associated with the relatively poor assembly of the sea lamprey genome, in addition to the difficulty of data mining and

confirming the sequence of small fast evolving peptides. Of the nine possible insulin superfamily ligands (ins, igf1, igf2, igf3, rln, rln3, insl3, insl5 and insl5a), only two (ins and igf2) are annotated on the current assembly of the lamprey genome. Of the ten possible receptors (insr, insrr, igf1r, rxfp1, rxfp2, rxfp2-like, rxfp3-1, rxfp3-2, rxfp3-3 and rxfp3-4), only five of them are annotated in the current assembly of the genome (igf1r, rxfp1, rxfp3-1, rxfp3-2 and rxfp3-3).

Thus, with this *de novo* RNA-Seq pipeline, I hope to identify currently annotated genes that are expressed in the ovary (proof of principle, Table 3.3) but also aim to identify any previously unannotated genes, several of which are also expected to be expressed in ovary. The pipeline has successfully identified some novel genes of the insulin family and can be potentially used for identifying any specific genes of interest. Each of the designed pipelines – genome-guided and *de novo* assembly – has its own strengths and weaknesses; hence, depending on the research objectives, both pipelines can be used for studying transcriptomes in model and non-model species with and without a reference genome.

### **3.4. Material and methods**

#### **3.4.1. Sample collection**

Chestnut lamprey and northern brook lamprey were collected using a backpack electroshocker (Smith-Root LR-24) as per collection permit protocol (for 2011, SCP 05-11 and SECT 73 SARA C&A 11-012; for 2012, SCP 15-12 and SECT 73 SARA C&A 12-009, Spice *et al.*, 2014). All samples of chestnut lamprey were collected from the Rat River in St. Malo, Manitoba, whereas northern brook lamprey were collected from two different locations - Birch River near Prawda, Manitoba and McKinnon Creek near Sault Ste. Marie, Ontario. The individuals were sacrificed as per the Animal Use Protocol guidelines, F11-019. The details

about the samples used in the project are discussed in Table 1.3. The gonad was removed from each individual and put in liquid nitrogen, and then subsequently stored at  $-80^{\circ}\text{C}$  until use. Total RNA was extracted from each gonad by using Qiagen RNeasy Mini Kit by following the manufacturer's instructions. A DNase digestion step was incorporated along with the RNase-free DNase Set (Qiagen) to remove the genomic DNA. Total concentration of RNA was measured using a NanoVue Spectrophotometer (GE) and RNA was stored at  $-80^{\circ}\text{C}$  until use. Complementary DNA (cDNA) was synthesized using the QuantiTect Reverse Transcription Kit (Qiagen) from 100 ng RNA, following the manufacturer's instructions. cDNA was ten-fold diluted and stored at  $-30^{\circ}\text{C}$  until use (Spice *et al.*, 2014).

### **3.4.2. RNA Sequencing**

Samples from both lamprey species, *I. castaneus* and *I. fossor*, were sent for RNA sequencing to two different sequencing facilities, Oklahoma Medical Research Foundation (Oklahoma City, Oklahoma) and Hussman Institute for Human Genomics (Miami, Florida). In short, RNA-Seq is a process in which messenger RNA (mRNA) is reverse-transcribed to complementary DNA (cDNA) fragments and adaptors are attached at the end of one or both the fragments. cDNA is then sequenced from one or both ends (Hudson 2008; Wang *et al.*, 2009; Tucker *et al.*, 2009; Ekblom & Galindo 2011). Sample numbers IC3, M01, and, N1-10 (Spice, 2013), IC1, IC2, NC3, and S11 (Table 1.3) were sequenced at the Oklahoma Medical Research Foundation wherein messenger RNA (mRNA) was isolated from ribosomal RNA and small modulating RNA using a poly-T bead step. For sample numbers C13-3, C20-3, and N17-1 (Spice, 2013) that were sent to the Hussman Institute for Human Genomics, mRNA was isolated using a poly-A step. Illumina TruSeq DNA Kit and Epicentre ScriptSeq Kit was used for

creating non-normalized libraries. Sequencing was performed in both forward and reverse directions using 75-100 base pair (bp) by using TrueSeq sequence adaptors on both ends on an Illumina Hi-Seq 2000 for paired-end reads. The adaptors were 60 bp long and the insert size was 300 bp for both the paired-end reads. The inner mean distance was calculated for all the samples by subtracting the fragment size from the insert size. Thus, the inner mean distance for all the paired-end reads in both forward and reverse directions was 220 bp.

#### **3.4.2.1. Evaluation of raw data**

The raw sequences generated for both the chestnut and northern brook lampreys were returned in FASTQ format and a quality check was performed to remove low-quality reads and identify over-represented sequences using the program, FASTQC (Andrews, 2010). This identified high-representation of the adaptor sequences which were removed in the following step.

#### **3.4.2.2. Removal of adaptor sequences**

After selecting high-quality reads, the tool Trimmomatic (Bolger *et al.*, 2014) was used to remove adaptor sequences from the paired-end reads of both species of lampreys. During sample library preparation, the RNA is fragmented and size-selected. Because the size selection captures a range of fragment sizes (200-1000 bp) and up to 75-100 bp of sequence was additionally obtained from the right and left hand adaptor, fragments smaller than ~100 bp may have the entire fragment sequenced (including an overlapping portion in the middle) in addition to part of the adaptor sequence. The adaptor sequences were removed and paired-end read files (both right

and left) were generated and subject to a second round of quality control selection using FASTQC as described above.

### **3.4.2.3. Format conversion**

The files obtained from Trimmomatic were “groomed” using the tool, FASTQ Groomer (Blankenberg *et al.*, 2010), which converts the FASTQ data files into other FASTQ variant file formats for downstream analyses. The FASTQ data files were converted into fastqsanger format.

### **3.4.2.4. Concatenation of forward and reverse reads**

Based on the mapping percentage of all the samples (Table 2.3), the forward and reverse reads of sample numbers C13-3, C20-3, IC1, and IC2 from chestnut lamprey and N1-10, N17-1, M01, and N08 from northern brook lamprey were merged using the program “Concatenate datasets” (Table 3.1). Thus, all the FASTQ/ FASTA files of ‘left’ and ‘right’ reads of paired-end data of each species were merged into single ‘left.fq’ and ‘right.fq’ and was used as an input for generating an ovarian-wide transcriptome.

### **3.4.2.5. Reference-free transcriptome assembly**

The TopHat alignment summary results obtained in Chapter 2 (Table 2.3) indicated that 12-45% reads of chestnut lamprey and 16-36% reads of northern brook lamprey reads mapped to the genome, while 55-88% of the genes do not map and were discarded by the reference-guided pipeline (Chapter 2). To investigate what lies in the unmapped portion, a *de novo* assembly (Chaisson *et al.*, 2009) was performed on each of the species. All the overlapping forward and reverse reads from both the chestnut and northern brook lamprey were assembled separately into

putative transcripts (contigs) in FASTA format by using the package Trinity version v2.2.0 (Grabherr *et al.*, 2011). The Trinity software is supported by Linux, and can be run through command-line interface. It is capable of assembling reads generated from RNA-Seq Illumina platform. Normally, at least 1 GB of RAM per 1 million paired-end reads configuration is recommended; however, smaller datasets can be executed with less memory allocation. The Galaxy platform also has a GUI for performing Trinity. Using this platform, all the forward and reverse reads were assembled and used to create two reference-free “Transcriptome assemblies”, one for each species. These FASTA formatted files, “Assembled Transcripts”, are composed of all the *de novo* assembled transcripts including potentially alternatively spliced variants, and each transcript is given a unique **Trinity ID** including sub-scripts regarding details about the read cluster, gene and isoform. The program Trinity assembles all the transcripts into clusters based on shared sequence content and calls them a ‘gene’. The output of Trinity is FASTA-formatted and each Trinity cluster has a unique accession ID (Trinity ID). The Trinity accession contains important information about Trinity ‘gene’ and ‘isoform’. The example is discussed below where the Trinity accession “TRINITY\_DN142776\_c0\_g1\_i1” is the read cluster, ‘g1’ indicates the gene and ‘i1’ isoform. Thus, Trinity assembles many clusters of reads separately and assigns unique ‘gene id’ (TRINITY\_DN142776 g1’) and ‘isoform id’ (TRINITY\_DN142776 g1 i1) explained below.

```
>TRINITY_DN142776_c0_g1_i1 len=346 path=[324:0-345] [-1, 324, -2]
GTTTCATAGTTGCTATCAAAGCATTGAGGTGTCAGGATTGTTGCTTTAAATTCAGTG
TATTACCTGCCTTAACATATATGGCAGGTCATACACGTATCTTACTTCAGACCTCCTT
CCTGTCCACTCATGTCTCACCTCTTCATGCATCTCACCGCAGACTGCACATATATTCC
TCTCAAACCATGCAGCTCCCATCTCGACTTCACACATATCCTGCCACGCATCTCCTTC
CTCCGATTCTGTGCATCATCATTCTCATGCAGTATATCTCCACCTCGTATGCCAAGT
GCATTGTGAAATGGTTGGTATATGTGCCCAGGATATGATATAACTGTAGATTTCCG
```

#### **3.4.2.6. Statistics of Trinity assembled files**

Based on the length of the generated assembled transcriptome contigs, various statistics were calculated on both the Trinity assembled files. A total of 128,361,110 bases were assembled in chestnut lamprey with a total of 136,772 transcripts, whereas in northern brook lamprey, 237,081,771 total bases were assembled and 333,747 total transcripts were computed (Table 3.2).

#### **3.4.2.7. Transcript estimation by RSEM**

The next step in the pipeline aimed to determine which transcripts are the isoforms of the same gene. In Trinity, alternative transcripts of the same gene are called ‘**isoforms**’ that have similarity with the gene sequence but are structurally different; sometimes reads map to multiple genes or isoforms and it is often difficult to determine which transcripts are isoforms of the same gene. For this, a transcript quantification package RSEM (RNA-Seq by Expectation Maximization; Li & Dewey, 2011) is used for quantifying the gene and isoform abundance from paired-end reads of both chestnut and northern brook lampreys. The “Trinity-generated transcript” file is used as an input to RSEM package to count the occurrence of transcripts across different stages of ovarian development in both lamprey species. By considering positional biases, it calculates the probabilities of the reads being derived from each transcript and hence, estimates the gene and isoform levels from RNA-Seq data. It generates two output files which contain information about the abundance estimates; a) RSEM.isoform.results; and b) RSEM.gene.results. The RSEM isoform result file contains information about the fragments that are derived from a given isoform or gene, whereas RSEM gene results computes what fraction of transcripts are made up by a given isoform or gene; hence ‘gene level’ is estimated by using the

Trinity component as a proxy for the gene (Hass *et al.*, 2013). This package is particularly useful for *de novo* transcriptome assemblies and enables accurate transcript quantification for species without sequenced genomes. When Trinity is installed, RSEM is automatically installed; hence it is an in-built package inside Trinity. I used Galaxy (usegalaxy.org) for running Trinity and RSEM. The RSEM.gene.results files of both species were used further for downstream analysis.

#### **3.4.2.8. Identification by BLAST (Basic Local Alignment Search Tool)**

In the next step, I aimed to assign tentative names to the assembled transcript ID's to determine the gene showing the highest homology to the transcript in a customized database. This homology search was done using the algorithm BLASTX (Altschul *et al.*, 1990). The entire set of assembled transcriptomes of chestnut lamprey and northern brook lamprey were queried against a custom-made BLAST database of 10 fully sequenced chordate genomes containing all known protein sequences (FASTA) of selected species of interest (i.e., elephant shark *Callorhinchus milli*, Florida lancelet *Branchiostoma floridae*, western clawed frog *Xenopus tropicalis*, spotted gar *Lepisosteus oculatus*, coelacanth *Latimeria chalumnae*, human *Homo sapiens*, chicken *Gallus gallus*, mouse *Mus musculus*, zebrafish *Danio rerio*, and sea lamprey *Petromyzon marinus*, Table 1.4). The advantage of using this strategy is that it assigns Gene Ontology terms to the genes and transcripts identified in sea lamprey using the customized database rather than the full non-redundant BLAST database.

BLASTX function was used with customized parameters (E-value of 1e-07, max hsps 5, num align 5 and outfmt 5) and the top 5 best hits were retrieved. “**E-value**,” also known as expectation value, is a statistical parameter for estimating the probability of obtaining hits or match by chance; lower E-values indicate significant match. “Alignment” is a process in which

two or more biological sequences (nucleotide or amino acid) are matched to identify the similarity level between sequences. “**High-segment scoring pairs**” (hsps) is used for generating an alignment without gaps between sequences that share a high level of similarity or homology. In the BLAST, stand-alone program, the user can obtain results in several output formats (outfmt); for this thesis, I used output format 5 for generating results in xml format with gene descriptors.

The BLAST xml output file was then converted into a 12-column tabular format where each column represented query seq-id, subject seq-id, percentage of identical matches, E-value, bit score, number of mismatches, alignment length, number of gap openings, start of alignment in query and end of alignment in query. This allowed us to assign putative gene names to all assembled **Trinity ID**'s (contigs). The list of putative genes identified by this *de novo* assembly pipeline as belonging to the insulin superfamily was compared to the list of **annotated** insulin superfamily genes identified in the genome-guided pipeline (Chapter 2).

#### **3.4.2.9. Specific genes of interest**

One of the main goals of the *de novo* assembly pipeline was to look for expression of any focal genes of interest that we have *a priori* reasons to think may be involved in ovarian development in chestnut and northern brook lampreys and/or are of interest to us (e.g., insulin family genes). For example, insulin superfamily genes are known to play an important role in ovarian steroidogenesis in Nile tilapia *Oreochromis niloticus* (Li *et al.*, 2012) and oocyte maturation in zebrafish (Li *et al.*, 2011), but in lampreys, the role of these genes is largely unknown (Spice *et al.*, 2014). Of the nine possible insulin superfamily ligands (ins, igf1, igf2, igf3, rln, rln3, insl3, insl5 and insl5a), only two (ins and igf2) are annotated on the current

assembly of the lamprey genome. Of the ten possible receptors (insr, insrr, igf1r, rxfp1, rxfp2, rxfp2-like, rxfp3-1, rxfp3-2, rxfp3-3 and rxfp3-4), only five of them are annotated in the genome (igf1r, rxfp1, rxfp3-1, rxfp3-2 and rxfp3-3). Thus, to identify both (known and novel) genes of the insulin family expressed in ovarian development in lampreys, the BLAST 12-column tabular output of both chestnut and northern brook lampreys was searched to identify the Trinity ID's associated with insulin superfamily genes, resulting in identification of several potential family members in each species. The FASTA sequences of selected Trinity ID's (i.e., insulin family genes) was queried again using the non-redundant (nr) BLAST database available at NCBI (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to further assess the confidence of the ID's based on the custom-made BLAST database. This resulted in positive identification (re-confirmation) of the **putative** ID of some transcripts, but clear identification of false positives for other transcripts. There can be several reasons for this discrepancy, including size and content of the custom-made database, selection of search parameters (E-value), and the transcriptome used for querying. One of the most important criteria for adjusting the number of true-positive (i.e., close homologs) is to adjust the E-value. I originally used a value of  $E < 1e-20$  which resulted in identification of very few transcripts (~21,358) and then reduced this to  $E < 1e-07$ , which resulted in the identification of 35,793 and 69,757 transcripts of chestnut and northern brook lampreys, respectively. However, 20-30% of these transcripts exhibited homologies based on a small section of the gene and thus were false positives (potentially distant homologs). In a subsequent iteration, the customized database should be expanded to include a few additional taxa, and the Expectation threshold could be reduced to  $E < 1e-08$ , or adjusted appropriately for the size of the database. For the purposes of this study, only the insulin family genes that were shown to be

probable homologs when queried against the full NCBI database were retained for further analysis.

#### **3.4.2.10. Putative gene name assignment to Trinity ID's of insulin family genes**

The next and most difficult step in this pipeline was to assign putative gene names to the insulin superfamily genes identified using the *de novo* transcript assembly pipeline. To determine the gene ID, FASTA sequences of the Trinity ID's sub-selected as belonging to the insulin superfamily were queried using BLAT (for searching against the sea lamprey genome) or BLASTN (for searching against non-lamprey genomes) using the Ensembl Genome browser. Gene names were assigned two ways: if the sequences matched an annotated (but **novel**) gene on the lamprey genome, I searched the **orthologs** and **paralogs** of this gene in other taxa and examined the gene tree to assess its putative orthologs. In some cases, there is more than one gene name reported in the orthologs and paralogs; in that case, the gene name with the highest percentage identity was selected. In the second case, when the query sequence did not match an annotated region of the genome in lamprey, I used the transcript to search the genome of other taxa (e.g., human and zebrafish) to assess if the transcripts aligned to an annotated gene with high homology, and additionally used the sequence to re-query NCBI. Lastly, in some cases, unpublished data from Dr. Good's laboratory provided the genomic location and evidence for the existence of some insulin superfamily genes in lampreys. Using these various approaches, putative gene names were assigned to all the Trinity ID's of insulin family genes. One of the main advantages of this approach is that it is an efficient method of finding novel genes of interest from a *de novo* transcriptome assembly.

#### **3.4.2.11. RSEM count and BLAST**

The next step in the pipeline was to retrieve the RSEM counts of the selected Trinity ID's pertaining to the insulin family genes for measuring the change in transcript expression at different stages of ovarian development in chestnut and northern brook lampreys. The two outputs – RSEM.gene.output file (of each species) and BLAST 12-column tabular file of each species – were merged together to get the overall transcript count. The generated table contained a list of Trinity ID's with their putative gene name and RSEM count across different gonadal stages (table available upon request). The Trinity ID's of the insulin superfamily genes were plotted using GraphPad (Figures 3.2 – 3.6).

### **3.5. Results**

#### **3.5.1. Trinity statistics**

The “Assembled Transcripts” file of chestnut and northern brook lampreys were evaluated to determine the quality of the assembled reads by using the script “Trinity.Stats.pl”. The transcriptome of chestnut lamprey C13-03, C20-3, IC1 and IC2 had 128,361,110 total assembled contigs. The total number of genes (RSEM count) and transcripts identified were 121,601 and 136,772, respectively. Based on the assembled contigs length, chestnut lamprey had an N50 length value of 1,845 bases. “N50” is the maximum length of the contig above which the assembly contains at least half the total number of bases. On the other hand, the northern brook lamprey transcriptome had 237,081,771 total assembled contigs, 244,098 total genes and 333,747 transcripts were reported by RSEM count. The N50 length value was 737 bases (Table 3.2).

### **3.5.2. BLASTX for Trinity ID's**

For chestnut lamprey, the BLASTX function identified a total of 37,793 transcript ID's with putative gene names whereas for northern brook lamprey it was 69,757 (table available upon request).

### **3.5.3 Combined output of RSEM and BLASTX.**

The BLASTX results of each species were merged with their RSEM expected counts across different gonadal stages which indicated that about 15,041 Trinity ID's had RSEM count with putative gene names in chestnut lamprey. However, northern brook lamprey had 19,038 Trinity ID's with RSEM count and putative gene name (table available upon request). This indicates that about 20,000 transcripts from chestnut lamprey and 50,000 transcripts from northern brook lamprey had no clear homolog in the custom database.

### **3.5.4. Insulin family genes identified**

From this *de novo* assembly, 12 Trinity ID's (TRINITY\_DN50463, TRINITY\_DN43997, TRINITY\_DN41783, TRINITY\_DN56949, TRINITY\_DN57356, TRINITY\_DN56120, TRINITY\_DN69516, TRINITY\_DN55983, TRINITY\_DN20203, TRINITY\_DN65114, TRINITY\_DN51388 and TRINITY\_DN58444, Table 3.3) belonging to insulin family genes, some of which have not been previously identified in the sea lamprey genome, were identified. In chestnut lamprey, TRINITY\_DN43997 and TRINITY\_DN50463 and, in northern brook lamprey, TRINITY\_DN56949 and TRINITY\_DN56120 were identified as ligands of insulin superfamily. The putative/tentative name of these Trinity ID's was assigned as igf1 and igf2 (Figures 3.2 & 3.3). Five Trinity ID's, TRINITY\_DN41783, TRINITY\_DN69516,

TRINITY\_DN55983, TRINITY\_DN20203 and TRINITY\_DN65114 (one from chestnut lamprey and four from northern brook lamprey, Tables 3.2 & 3.3), were identified as putative receptors of insulin family (insr, insrr and igf1r). In northern brook lamprey, one putative ligand (TRINITY\_DN57356, relaxin-like) and two putative receptors of relaxin family (TRINITY\_DN51388, rxfp1 and TRINITY\_DN58444, RXFP2-like) were also identified (Table 3.3, Figure 3.5). Thus, in chestnut lamprey, only three Trinity ID's associated with insulin family genes were reported; however, in northern brook lamprey, nine Trinity ID's belong to insulin family genes (table available upon request).

### **3.5.5. Genes reported by genome-guided and *de novo* assembly**

One of the tests of this pipeline was to determine whether the *de novo* assembled transcripts can identify all of the insulin superfamily genes that were found in the genome-guided pipeline and identify additional novel members of the family. Almost all of the genes – igf2, insr-like, igf1r-like, insrr-like – reported by the HTSeq pipeline were also identified in the *de novo* assembly pipeline. This shows that the results obtained from both these pipelines complement each other and were able to identify the insulin family genes that are both expressed in ovary and annotated in the current version of the sea lamprey genome. However, as expected and hoped, the *de novo* assembly identified some novel genes not reported by the genome-guided pipeline (Table 3.3).

### **3.5.6. Novel genes of insulin family identified**

One of the main contributions of this pipeline is that it identified four novel genes of the insulin family (two ligands - relaxin-like & igf1 and two receptors – insrr & RXFP2-like). Both

igf1 and insrr are ligands and receptors of the insulin family, and relaxin-like and RXFP2-like are ligands and receptors of the relaxin family (Figures 3.2-3.4, Table 3.3). The putative name was assigned to these genes based on bioinformatic data mining and previous information. Igf1 was identified in both chestnut and northern brook lampreys. During gonadal stage 3 and 4, chestnut lamprey shows high expression of igf1 as compared to the northern brook lamprey. Relaxin-like, RXFP2-like and insrr is only expressed in the non-parasitic northern brook lamprey during gonadal stage 4. All these genes have not yet been identified in the sea lamprey genome; this information will contribute to updating the known genes of lamprey.

### **3.5.7. Expression of insulin family genes across different gonadal stages**

#### **a) igf1 and igf2**

These genes were identified in both chestnut and northern brook lampreys; the transcript expression plot of RSEM (Figures 3.2 & 3.3) showed that during gonadal stage 2, igf1 was expressed more in northern brook lamprey as compared to chestnut lamprey. However, at gonadal stages 3 and 4, a sharp decline of igf1 expression was noticed in northern brook lamprey, suggesting that it may be involved in accelerating ovarian differentiation in northern brook lamprey, whereas igf2 was only expressed in chestnut lamprey during gonadal stages 3 and 4.

#### **b) igf1r and insrr**

igf1r was reported in both chestnut and northern brook lampreys, whereas insrr was only identified in northern brook lamprey. In gonadal stage 4, northern brook lamprey showed high expression of igf1r compared to chestnut lamprey (Figure 3.5), whereas insrr was highly

expressed during vitellogenesis, suggesting that it may be involved in undergoing sexual maturation in northern brook lamprey (Figure 3.4).

c) RXFP-1 and RXFP-2 like

These genes were expressed only in northern brook lamprey, during gonadal stages 2 and 3; RXFP-2 like was more highly expressed than RXFP-1 (Figure 3.6).

### 3.6. Discussion

This study was designed to identify both novel and known genes involved in lamprey ovarian development by using a *de novo* assembly. The genome-guided pipeline used in Chapter 2 reported low mapping percentages and later it was learned that samples had a substantial amount of contamination, mostly bacterial in origin (unpublished results). Hence, the mapped percentage was in the range of 12-45% (Table 2.3), so ~55-88% of the reads were unmapped and could not be used for downstream analysis. Additionally, the HTSeq counting method (Anders *et al.*, 2015) identified 6,913 **Ensembl ID**'s with an associated gene name but ~6,000 of these Ensembl ID's had no gene name and were identified as "novel protein coding gene". Thus, a major proportion of the reads (~13,000) were not counted and were treated as having "No feature" by HTSeq since they could not be mapped to an annotated region of the genome. This is not surprising given that the sea lamprey has an estimated 26,046 genes (Smith *et al.*, 2013) but only ~13,000 genes are annotated on the current version of the Ensembl genome annotation.

Thus, to identify novel genes expressed in the lamprey ovary but not included in the current annotation of the genome, this *de novo* assembly pipeline was used. This pipeline assembled 136,772 transcripts of chestnut lamprey and 333,747 transcripts of northern brook

lamprey. To screen these transcripts, I employed a custom-made database of ten fully sequenced chordate genomes. The BLAST database identified 35,793 transcripts in chestnut lamprey and 69,757 transcripts in northern brook lamprey; approximately 20,000 and 50,000 transcripts, respectively, did not have a homolog in the custom-made blast database. Although this could be due to a low significant threshold, I modified the E-value threshold to be  $1e-07$  to allow for high sequence divergence between the lamprey and other chordate genomes and to take into consideration the small size (450 MB) of the custom database. This suggests that between 65 and 80% of the transcripts had no chordate homolog – a percentage close to that contained in the unmapped reads in the reference guided assembly.

To test the functionality of the *de novo* assembly pipeline and to check its ability to identify specific genes of interest, I looked for genes belonging to the insulin superfamily, because of *a priori* reasons to suggest that several members of the superfamily play roles in ovarian development. This *de novo* assembly pipeline identified twelve Trinity ID's of insulin family genes and reported four novel genes (igf1, insrr, relaxin-like and RXFP2-like) not included in the current annotation of sea lamprey and not identified previously (see Table 3.3, Figures 3.2, 3.3, 3.5 & 3.6), but confirm prior hypotheses about the origin of the family during 2R (Yegorov & Good, 2012).

Although this pipeline is designed to identify novel genes, it remains challenging to assemble and definitively identify the gene ID from the *de novo* assembly for many reasons. Firstly, the *de novo* transcripts contain 5' and 3'untranslated regions (UTRs) which are often longer than the genes themselves and which align to non-coding segments of the genome. This clearly suggests that the *de novo* assembly pipeline can be used successfully for identifying novel and alternatively spliced variants of a gene, but it is difficult to assign the coding regions.

Secondly, even once the coding regions are identified, it is notoriously difficult to determine gene orthology, particularly for divergent species such as lampreys. Thus, in future work, the final assessment of the novel insulin superfamily genes in lampreys should be assessed through more detailed phylogenetic and sequence analyses.

Despite these issues, the *de novo* pipeline performed better for the northern brook lamprey than the chestnut lamprey (Section 3.5.4) and identified most of the genes reported by the genome-guided (HTSeq) pipeline (Table 3.3). The HTSeq gene counting method reported several Ensembl ID's that had no gene name. Through data available from Dr. Good's laboratory, it was found that eight of the Ensembl ID's belong to insulin family genes and had no gene name; the **putative** name was assigned to these Ensembl ID's based on genes reported in orthologs and paralogs and their position in the gene tree (Section 2.7.6). Hence, the *de novo* assembly pipeline identified insulin family genes and the putative name of these genes are igf2, insr-like, igf1r, insrr, and rxfp1. Both the pipelines successfully identified insulin family genes, but the *de novo* assembly pipeline reported novel genes igf1, insrr, relaxin-like and RXFP-2 like, which are not included in the current annotation of sea lamprey (Table 3.3).

There were some technical difficulties reported. Firstly, one of the main goals of this pipeline was to identify the percentage of overlapping genes reported by the HTSeq gene counting method versus the RSEM transcript counting method. The HTSeq method used in the genome-guided pipeline (Chapter 2) provided estimated counts of genes (including multiple splice variants) annotated in the lamprey genome by using the sea lamprey reference Gene Feature File (GFF). A major proportion of the genes (~13,000) were not counted by the HTSeq gene counting method, potentially because of the incomplete annotation of the sea lamprey genome or because they were not reported by the algorithm of HTSeq. All possible efforts were

made to retrieve the genes which were assigned as “No feature” by the program. There is a chance that some important genes were missed and thus were not available for any further downstream analysis. Therefore, *de novo* assembly was performed which generated assembled transcripts of each lamprey species and calculated the normalized counts using the program RSEM (RNA-Seq by Expectation Maximization) across developmental stages. Thus, to identify the percentages of genes reported by both genome-guided and *de novo* assembly pipelines, Trinity assembled transcripts (FASTA sequences) were queried by BLASTX to get the putative gene name. The results of BLASTX for each of the species was downloaded and their GI number and RefSeq ID was converted to Ensembl ID’s of species used in the customized BLAST database. It was not possible to convert all the GI number’s or RefSeq ID’s into Ensembl ID’s because some of the species which were used for making the customized BLAST database (Table 1.4) did not have gene names published in the Ensembl Genome browser and had no Ensembl ID’s. Thus, it was statistically not possible to identify the proportion of overlapping genes reported by both pipelines. However, the *de novo* approach identified many novel genes of the insulin family (Table 3.3, Figures 3.2, 3.3 & 3.5) which is a great contribution towards the already known lamprey genes and this approach can be used for any non-model organism. Future studies should focus on mapping all the transcripts obtained from the *de novo* assembly onto the sea lamprey genome to identify the proportion of genes reported by both pipelines. One simple approach would be to use all of the transcripts that had homologs in the custom-made database in a reference genome pipeline to assess whether the *de novo* assembled transcripts map to the same ~6,900 annotated genes in the sea lamprey genome.

Lastly, the *de novo* pipeline performed better for the northern brook lamprey than the chestnut lamprey. In chestnut lamprey, only three Trinity ID’s associated with insulin family

genes were reported; however, in northern brook lamprey, nine Trinity ID's belonged to insulin family genes. There can be several reasons for this, including selection of the samples for generating the reference transcriptome for each species. It was observed that samples of chestnut lamprey that were used for generating transcriptomes had low mapping percentages reported in genome-guided assembly (Chapter 2) and a cursory examination of the samples indicated that a substantial fraction of the unmapped reads was bacterial in origin (especially from the one company that didn't select poly-A tails, Section 2.3.2). So, a major fraction of the assembled reads (contigs) were not recognized by the customized BLAST database. All possible efforts were made to get true hits; the E-value was customized (lowered from 1e-20 to 1e-07 because the former gave too few hits), but unfortunately the BLAST results for chestnut lamprey recovered many fewer transcripts with homologs in the custom database. Future studies should be designed by including more species for making a custom database and the parameters of the BLASTX query should be modified to find the optimal values given the nature and size of the database.

The present study identified novel genes which are unannotated in the sea lamprey genome so this *de novo* approach can be used for any other non-model organism for identifying specific genes of interest. Additionally, the list of putative genes generated in chestnut and northern brook lampreys through BLASTX can be used to design PCR based studies such as quantitative real-time reverse transcriptase PCR (qRT-PCR) and for designing primers. All the sequence data generated through this *de novo* assembly can be used for building phylogenetic trees and may help to understand questions related to the evolution of particular genes in vertebrates. In addition, this project identified some of the key problems of RNA-Seq studies in non-model organisms, so future studies should be designed by taking at least six replicates

belonging to different gonadal stages. A recent study by Schurch *et al.* (2016) suggested that for identifying significantly differentially expressed genes in RNA-Seq data at least six replicates should be used for measuring all fold changes. The transcriptomes generated in this project may help to update the present lamprey genomic and transcriptomic resources (e.g., Smith *et al.*, 2013; Chung-Davidson *et al.*, 2013; Chung-Davidson *et al.*, 2015).

### 3.7. References

- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Altschul S, Gish W, Miller W *et al.* (1990) A basic local alignment search tool. *Journal of Molecular Biology*. **215**, 403–410.
- Anders S, Pyl PT, Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
- Blankenberg D, Gordon A, Von Kuster G *et al.* (2010) Manipulation of FASTQ data with Galaxy. *Bioinformatics*, **26**, 1783-1785.
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
- Chaisson MJ, Brinza D, Pevzner PA (2009) *De novo* fragment assembly with short mate-paired reads: does the read length matter? *Genome Research*, **19**, 336-346.
- Chung-Davidson Y-W, Priess MC, Yeh C-Y *et al.* (2013) A thermogenic secondary sexual character in male sea lamprey. *The Journal of Experimental Biology*, **216**, 2702–2712.
- Chung-Davidson Y-W, Yeh C-Y, Bussy Ugo *et al.* (2015) Hsp90 and hepatobiliary transformation during sea lamprey metamorphosis. *BMC Developmental Biology*, **15**, 47.

- Duan C (1997) The insulin-like growth factor system and its biological actions in fish. *American Zoologist*, **37**, 491–503.
- Ekblom R & Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.
- Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.
- Good S, Yegorov S, Martijn *et al.* (2012) New insights into ligand-receptor pairing and coevolution of relaxin family peptides and their receptors in teleosts. *International Journal of Evolutionary Biology*, **2012**, 310278.
- Guttman M, Garber M, Levin JZ *et al.* (2010) *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, **28**, 503–510.
- Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.
- Haas BJ & Zody MC (2010) Advancing RNA-Seq analysis (2010). *Nature Biotechnology*, **28**, 421–423.
- Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.
- Jeukens J, Renaut S, St-Cyr J *et al.* (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology*, **19**, 5389–5403.

- Kristiansson E, Asker N, Forlin L *et al.* (2009) Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. *BMC Genomics*, **10**, 345.
- Li J, Liu Z, Wang D *et al.* (2011) Insulin-like growth factor 3 is involved in the oocyte maturation in zebrafish. *Biology of Reproduction*, **84**, 473–486.
- Li B & Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
- Li M, Wu F, Gu Y *et al.* (2012) Insulin-like growth factor 3 regulates expression of genes encoding steroidogenic enzymes and key transcription factors in the Nile tilapia gonad. *Biology of Reproduction*, **86**, 1–10.
- Lu C, Lam HN, Menon RK (2005) New members of the insulin family: regulators of metabolism, growth and now reproduction. *Pediatric Research*, **57**, 70R–73R.
- Luo R, Liu B, Xie Y *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.
- Martin JA & Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*, **12**, 671–682.
- Nakamura M, Kobayashi T, Chang XT *et al.* (1998) Gonadal sex differentiation in fish. *Experimental Zoology*, **281**, 362–372.
- Quinn NL, Levenkova N, Chow W *et al.* (2008) Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics*, **9**, 404.
- Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp Salmonidae). *Molecular Ecology*, **19**, 115–131.

- Robertson G, Schein J, Chiu R *et al.* (2010) *De novo* assembly and analysis of RNA-Seq data. *Nature Methods*, **7**, 909–912.
- Schulz MH, Zerbino DR, Vingron M *et al.* (2012) Oases: robust *de novo* RNA-Seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.
- Schurch NJ, Schofield P, Gierlinski M *et al.* (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**, 839–851.
- Smith JJ, Kuraku S, Holt C *et al.* (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genetics*, **45**, 415–421.
- Spice EK (2013) Ovarian Differentiation in an Ancient Vertebrate: Timing, Candidate Gene Expression, and Global Gene Expression in Parasitic and Non-parasitic Lampreys (*M.Sc. thesis*). University of Manitoba, Winnipeg, Canada.
- Spice EK, Whyard S, Docker MF (2014) Gene expression during ovarian differentiation in parasitic and non-parasitic lampreys: Implications for fecundity and life history types. *General and Comparative Endocrinology*, **208**, 116–125.
- Trapnell C, Williams BA, Pertea G *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–515.
- Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, **10**, 57–63.

- Wolf JBW, Bayer T, Haubold B *et al.* (2010) Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology*, **19**, 162–175.
- Wood AW, Duan C, Bern HA (2005) Insulin-like growth factor signaling in fish. *International Review of Cytology*, **243**, 215–285.
- Yegorov S, Bogerd J, Good SV (2014) The relaxin family peptide receptors and their ligands: new developments and paradigms in the evolution from jawless fish to mammals. *General Comparative Endocrinology*, **209**, 93–105.
- Yegorov S, Good S (2012) Using paleogenomics to study the evolution of gene families: origin and duplication history of the relaxin family hormones and their receptors. *PLoS ONE*, **7(3)**, e32923.

### 3.8. Tables and Figures

Table 3.1. Details on ovarian transcriptomes of chestnut lamprey *Ichthyomyzon castaneus* and northern brook lamprey *I. fossor* pooled together for generating an “Assembled transcript” of both the species by using Trinity. Sample numbers C13-03, C20-03, IC-1, and, IC-2 are chestnut lamprey and MO-1, NC-3, NO-8 and S11 are northern brook lampreys. “Gonadal stage” refers to the stage of ovarian development in lampreys.

<i>I. castaneus</i>	Gonadal stages
C13-03, C20-03, N1-10 & N17-01	IC-1, IC-2, IC-3 & M0-1
<i>I. fossor</i>	
N-10, N-17, MO-1 and NO-8	2, 3, 4 and vitellogenesis

Table 3.2. Trinity statistics of the “Assembled Transcripts” of *I. castaneus* and *I. fossor* used in this project. Total Trinity ‘genes’ is the genes reported by Trinity after assembling all the contigs. Total trinity ‘transcripts’ is the total assembled transcripts. Percent GC refers to the total GC content. ‘N50’ is the length of the contig above which the assembly contains at least half the total number of bases. Total ‘Assembled bases’ is the total sequences assembled from forward and reverse reads.

	<i>I. castaneus</i> (transcriptome)	<i>I. fossor</i> (transcriptome)
Total trinity 'genes' (RSEM count)	121,601	244,098
Total trinity ‘transcripts’	136,772	333,747
Percent ‘GC’	54.7	52.42
‘N50’	1,845	1,144
Total ‘Assembled bases’	128,361,110	237,081,771
Transcripts identified by BLASTX	35,793	69,757
RSEM count assigned to identified transcripts obtained from BLASTX	15,041	19,038

Table 3.3. Comparison of insulin family genes identified by both genome-guided versus *de novo* assembly pipeline in *I. castaneus* and *I. fossor*.

Insulin family genes	Gene name on Ensembl	Genome-guided pipeline (Ensembl ID)	<i>De novo</i> assembly (Putative gene name **)
<b>Relaxin family</b>			
<b>Ligands</b>			
Rln	n/a	n/a	n/a
rln3	n/a	n/a	relaxin like*** (TRINITY_DN57356#)
insl3	n/a	n/a	n/a
insl5	n/a	n/a	n/a
<b>Receptors</b>			
rxfp1	no name on Ensembl	ENSPMAG0000000636 *	relaxin receptor 1 isoform (TRINITY_DN51388#)
rxfp2	n/a	n/a	relaxin receptor 2-like*** (TRINITY_DN58444#)
rxfp2-like	n/a	n/a	n/a
rxfp3	rxfp3	ENSPMAG0000001344*, ENSPMAG00000010113*	n/a
rxfp3-1	n/a	n/a	n/a
rxfp3-2	n/a	n/a	n/a
rxfp3-3	RXFP3-3A3	ENSPMAG00000006421 *	n/a
rxfp3-4	n/a	n/a	n/a
<b>Insulin family</b>			
<b>Ligands</b>			
INS-L1	Insulin	ENSPMAG00000008971 *	n/a
IGF1	n/a	n/a	insulin-like growth factor I isoform X1 or igf1*** (TRINITY_DN56949# & TRINITY_DN43997##)
IGF2	igf2/igf1 ortholog	ENSPMAG00000002950*	ortholog of igf2 (TRINITY_DN56120# & TRINITY_DN50463##)
IGF3	n/a	n/a	n/a
<b>Receptors</b>			
Insr	insr/insra/insrb orthologs	ENSPMAG00000005896*, ENSPMAG00000006274*, ENSPMAG00000007436*	insulin receptor (TRINITY_DN55983# & TRINITY_DN41783##)
Insr	igf1r or insrr orthologs	ENSPMAG00000007101*, ENSPMAG00000008570*	insulin receptor or insulin related receptor*** (TRINITY_DN20203, TRINITY_DN69516# & TRINITY_DN41783##)
igf1R	igf1r or insrr orthologs	ENSPMAG00000008570*, ENSPMAG00000007101*	insulin like growth factor 1 receptor (TRINITY_DN65114 # & TRINITY_DN41783##)

\* genes that were having Ensembl ID but no gene name, through data mining gene name was found.

\*\* putative gene name assigned based on phylogeny and BLAST

\*\*\* novel genes identified through de novo pipeline

# reported in northern brook lamprey

## reported in chestnut lamprey

Figure 3.1. Transcriptomics data analysis pipeline also known as *de novo* assembly for paired-end reads in the absence of a reference genome by using customized BLAST database.

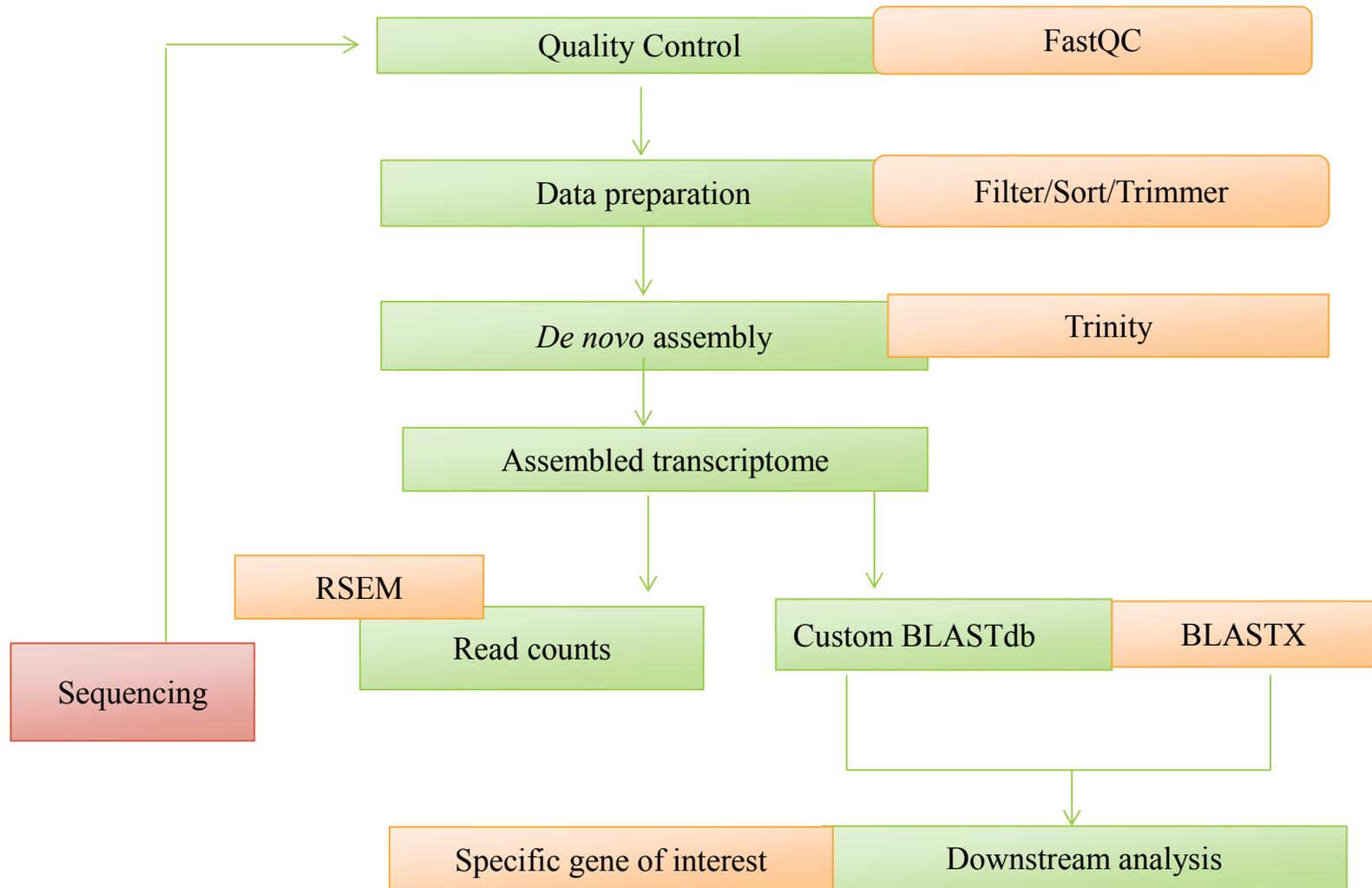


Figure 3.2. Plots of insulin family genes (igf1, igf2 and igf1r or insrr) reported in chestnut lamprey (parasitic) samples, where the x-axis represented the gonadal stages and y-axis represented the RSEM count. Genes like igf1 and igf1r or insrr are reported as novel genes and are not annotated in the sea lamprey genome.

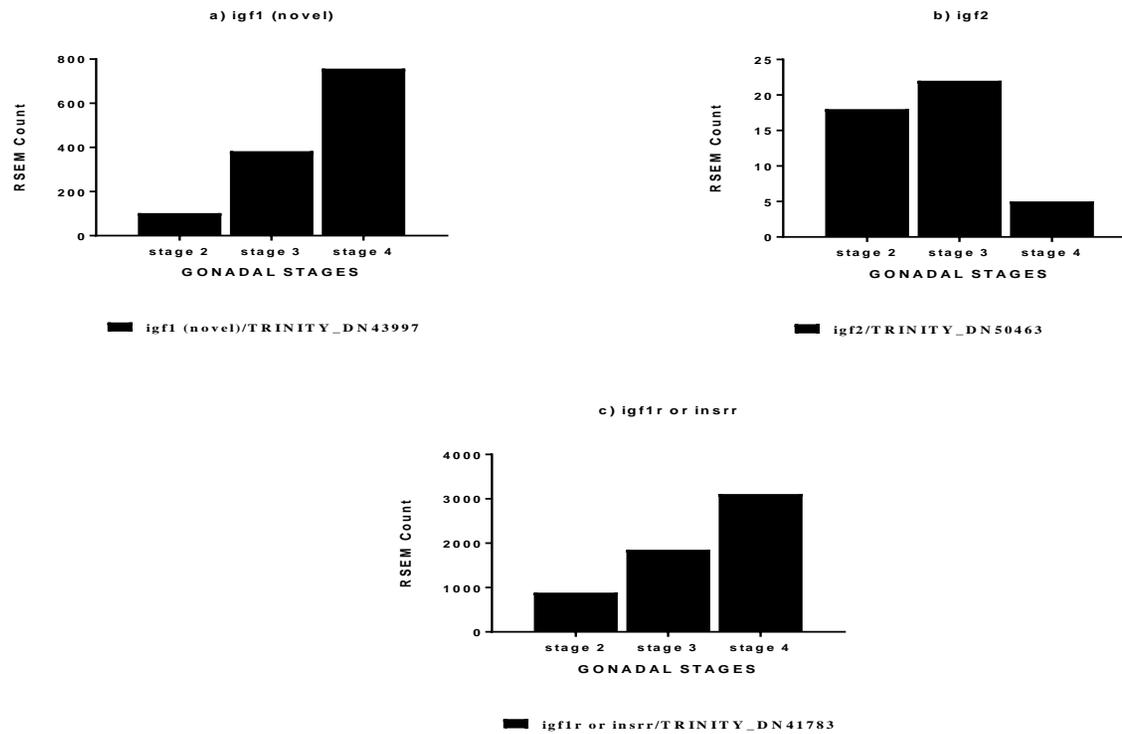


Figure 3.3. Plots of insulin family genes (*igf1*, *igf2*, *insrr*, relaxin-like, RXFP1 and RXFP-2 like) expressed in northern brook lamprey (non-parasitic) samples, where the x-axis represented the gonadal stages and y-axis represented the RSEM count. Genes like *igf1*, *insrr*, relaxin-like and RXFP2-like are reported as novel genes and are not annotated in sea lamprey.

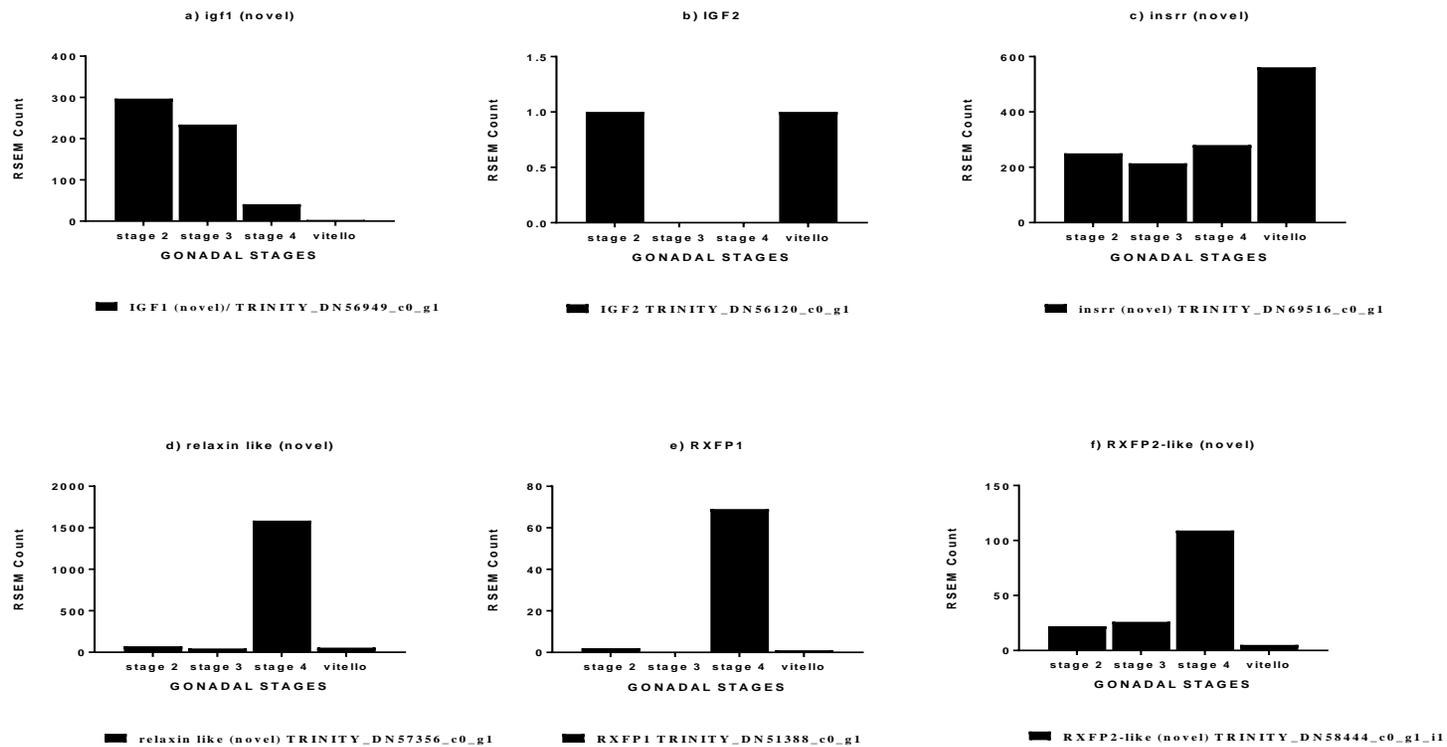


Figure 3.4. Plots of insulin family genes (igf1r or insrr) reported in northern brook lamprey (non-parasitic) samples, where the x-axis represented the gonadal stage and y-axis represented the RSEM count.

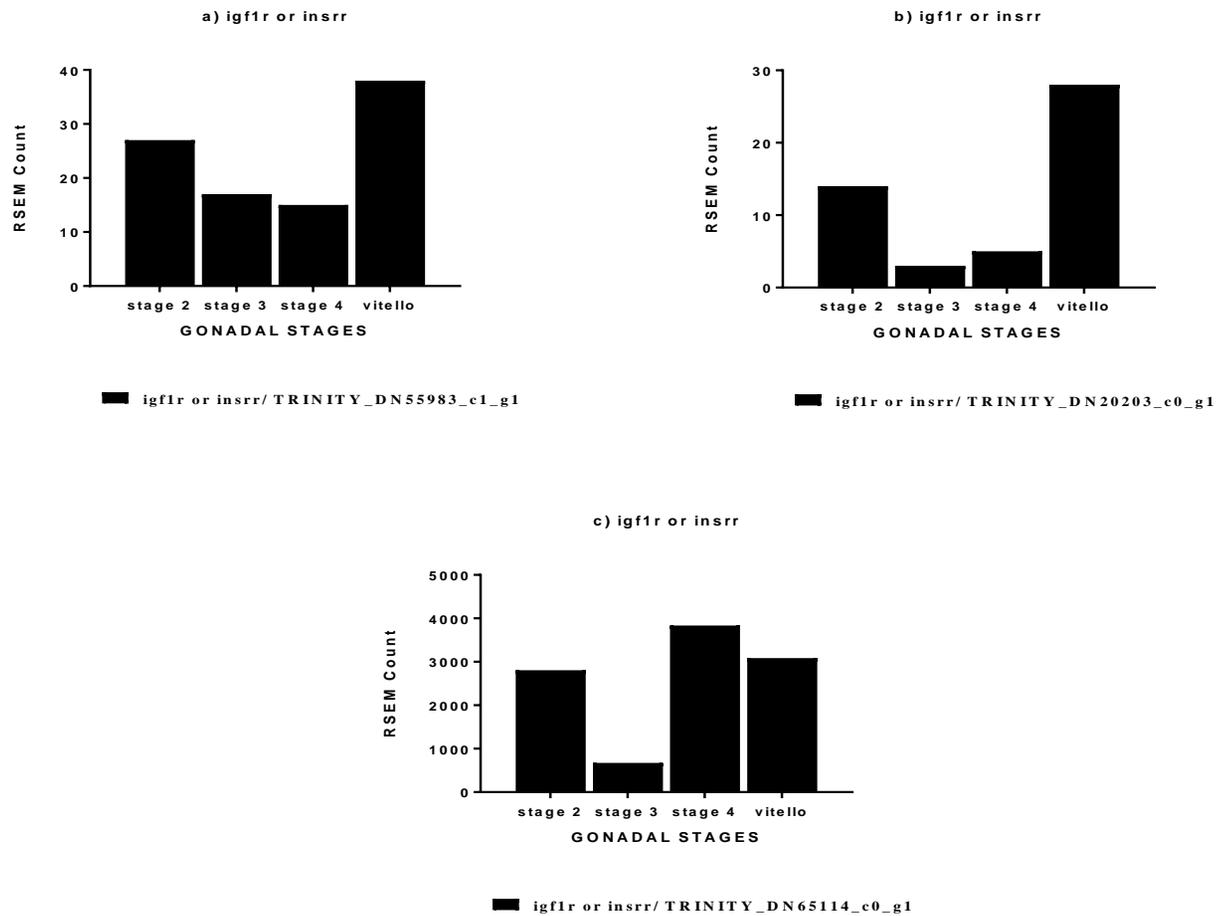


Figure 3.5. Plots of insulin family genes (igf1, igf2 and igf1r) reported in chestnut lamprey and northern brook lamprey, where the x-axis represented the gonadal stage and y-axis represented the RSEM count. Gene igf1 is reported as novel and is not annotated in sea lamprey.

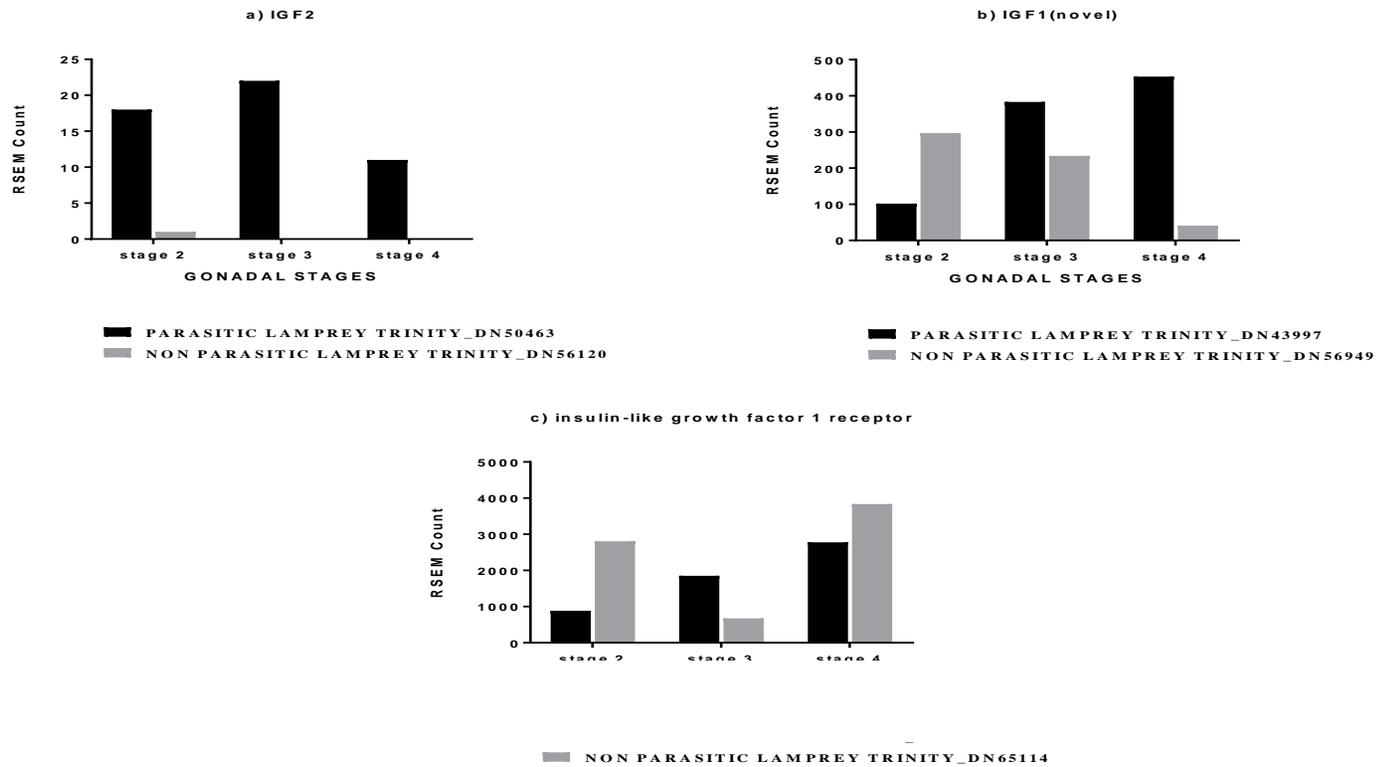
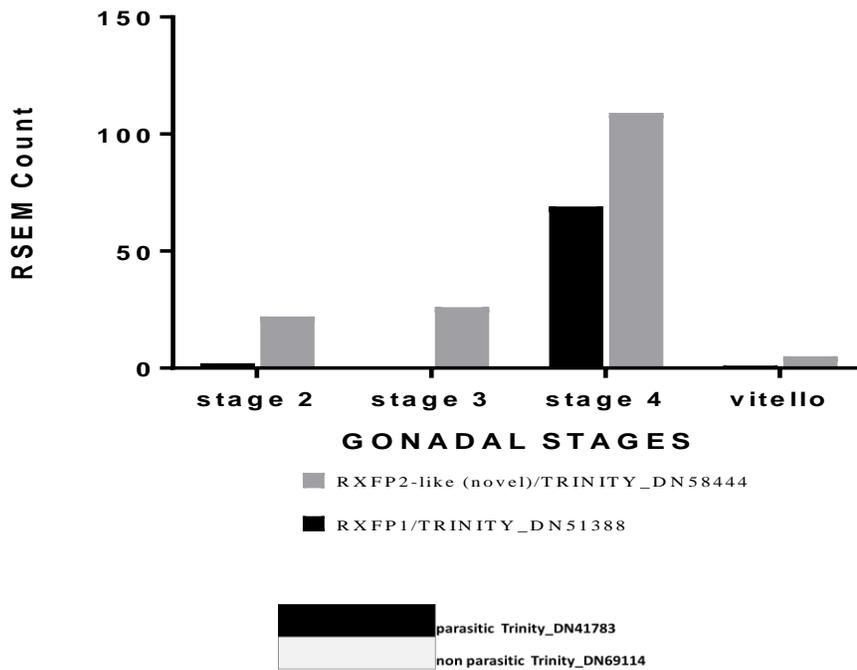


Figure 3.6. Plots of relaxin family peptides receptors – RXFP1 and RXFP-2 like novel reported only in northern brook lamprey (non-parasitic) samples, where the x-axis represented the gonadal stage and y-axis represented the RSEM count. RXFP-2 like is reported as a novel gene (i.e., is not annotated in the sea lamprey genome) and is expressed in all stages of ovarian development.



## Chapter 4: General Discussion

### 4.1. Background

This thesis has shed light on different aspects of RNA-Seq data analysis and discussed the benefits and drawbacks of using two different approaches; a genome-guided and *de novo* (note that terms that are bolded on first use within each chapter are defined in the glossary in Appendix 2) assembly to assemble and analyze RNA-Seq reads. In the genome-guided approach, the raw sequencing reads are aligned to a reference genome and then overlapping reads are assembled into transcripts/genes based on the annotation of the reference genome, whereas in the *de novo* assembly method, overlapping reads are assembled into contigs and then mapped onto a *de novo* assembled reference transcriptome that was pre-assembled by combining multiple samples. Given the absence of an annotation for the *de novo* assembled transcripts, a tentative ID was assigned to each transcript by querying transcripts against a custom-made database containing sequencing data from ten fully sequenced and annotated chordate genomes. Both approaches can be used potentially for organisms with a **reference genome** but for non-model organisms without a sequenced genome, *de novo* assembly is generally preferred to identify novel transcripts and known genes of interest. It is often observed that parameters should be tuned per the user's research objectives to get optimal results. Both the pipelines discussed in this project can be used to improve the annotation of reference genomes, such as the sea lamprey genome referenced here, and may be useful for a variety of other downstream analyses. This project took advantage of whole-genome RNA-Seq transcriptomic data from ovarian samples of parasitic chestnut lamprey *Ichthyomyzon castaneus* and non-parasitic northern brook lamprey *I. fossor* to identify suites of genes expressed during ovarian development. Chestnut lamprey are

parasitic and continue to feed after metamorphosis whereas northern brook lamprey are non-parasitic, stop feeding and undergo sexual maturation following metamorphosis and then reproduce and die (Docker, 2009). In non-parasitic lampreys, **sexual differentiation** and oocyte development generally occur earlier, fecundity is reduced, and **sexual maturation** is accelerated following metamorphosis compared to parasitic lampreys.

A primary goal of the thesis was to identify genes expressed at successive gonadal stages and to determine which molecular pathways are turned on at different stages of ovarian development. To this end, two different pipelines were designed, a genome-guided pipeline and *de novo* assembly (Figures 2.2 & 3.1). The genome-guided pipeline was used explicitly to identify the genes and pathways which were turned on or off during different gonadal stages. Samples from both chestnut and northern brook lampreys were pooled together based on their gonadal stages and then the reference-based transcriptome assembly pipeline was used that allowed us to a) assemble the reads, b) **map** these reads to the sea lamprey genome, c) extract the estimated number of reads that mapped to each gene annotated in the sea lamprey genome assembly in Ensembl, d) test for differences in gene expression between early stages (1, 2 and 3), mid-stage (4) and late stages (early and late vitellogenesis) stages of ovarian development, e) tabulate the genes found to be up- or down-regulated over temporal stage, and f) estimate the functions and molecular pathways of networks of genes found to be up- or down-regulated.

In the Principal Component Analysis (PCA), it was observed that samples from both chestnut and northern brook lamprey differed in their overall pattern of gene expression (based on where dots cluster together) as shown in Figures 2.6 & 2.7. In the early gonadal stages, all four samples (two from each species) had similar overall gene expression (blue dots close together on the PCA graph), while in the mid-gonadal stage, samples from two parasitic and one

non-parasitic lamprey also had similar gene expression (green dots). The same pattern was observed in the vitellogenesis stage; the three samples of non-parasitic lamprey clustered together (red dots), but a different pattern was observed in that one sample from the parasitic lamprey in gonadal stage 4 (green dot) clustered with the other non-parasitic samples during vitellogenesis. This sample was IC3 and, interestingly, this sample was from a lamprey with ovaries in gonadal stage 4, but morphologically was in late metamorphosis (stages 5-6), while the other two parasitic samples with ovaries in gonadal stage 4 were in a early metamorphosis (stages 2-3) (Manzon *et al.*, 2015). This result suggests that the later stage metamorphosing sample from the parasitic species had a similar gene expression profile to the non-parasitic species at a similar stage of metamorphosis (stages 5-8), even though the non-parasitic species has ovaries undergoing vitellogenesis whereas the parasitic species have not yet started to mature sexually. This suggests that up-regulation of many of the genes involved in vitellogenesis may occur well in advance of histological evidence of vitellogenesis. These preliminary but interesting observations should be followed up with further experimental work (with more individuals and with histological analysis to confirm stage of ovarian development) to confirm the relationship between stage of metamorphosis and ovarian development.

In the genome-guided pipeline used in Chapter 2, a reference genome of sea lamprey (Smith *et al.*, 2013) was used for mapping the raw reads to annotated genes and then obtaining counts of gene expression across gonadal stage. The selection of different parameters such as choice of a reference genome, calculation of inner mean distance between samples, quality check of reads, percentage of mapped and unmapped reads, and software for data analysis influence the outcome (Sections 2.3.2 & 2.4.1-2.5.9). For example, the low mapping percentages reported by the genome-guided pipeline was a matter of concern, but later it was learned that some of the

samples had substantial amounts of contamination that resulted in many unmapped reads. Only 55-88% of the reads were unmapped and thus could not be used for downstream analysis. In addition, a major proportion of the genes (~13,000) were not counted by the HTSeq gene counting method (Anders *et al.*, 2015) and therefore were not used further for differential analysis by DESeq2 (Love *et al.*, 2014) or ontology assignment by GOrilla (Eden *et al.*, 2009). Despite these pitfalls, the genome-guided pipeline successfully identified 6,913 Ensembl ID's with the associated gene name and reported three insulin family genes (rxfp3, rxfp3-1 and rxfp 3-3) that are annotated in the sea lamprey genome. The remaining Ensembl ID's (~6,000) had no gene name and were identified as "novel protein coding gene" (tables available upon request). Hence, through data available from Dr. Good's laboratory, it was found that eight of these **Ensembl ID's** belong to insulin family genes, and the putative name was assigned to these Ensembl ID's based on genes reported in **orthologs** and **paralogs** and their position in the gene tree (Section 2.7.6).

Although the results of Chapter 2 are quite comprehensive and provided interesting insights into overall patterns of change in gene expression across gonadal and metamorphic stages, a problem of this approach for lamprey species is that the genome annotation and the number of genes putatively identified with orthologs in other chordate taxa is low. Thus, to discover new genes expressed in the lamprey ovary which are not included in the current annotation of the genome, I also employed a second pipeline known as a *de novo* assembly pipeline (Figure 3.2) to identify novel lamprey genes without a reference genome. This pipeline assembled 136,772 transcripts of chestnut lamprey and 333,747 of northern brook lamprey. To assess which of these assembled transcripts pertain to the same genes already annotated in the sea lamprey genome identified in Chapter 2 versus those which are novel in Chapter 3, I used *a*

*priori* predictions and sequence information about the insulin superfamily of ligands and receptors in lamprey to a) assess the robustness of this pipeline, and b) identify **novel** members of the insulin superfamily. I chose the insulin superfamily of genes to compare these two pipelines because; a) it is a relatively large family estimated to contain approximately seven ligands and ten receptors in lampreys; b) we have *a priori* reasons to suggest that several members of this insulin superfamily play roles in ovarian development and thus would be expressed in these samples; and c) prior and on-going phylogenomic research on the insulin superfamily in Dr. Sara Good's laboratory means that we had *a priori* predictions about gene number and sequence variation of the genes. Moreover, of the nine possible ligands, only two (ins and igf2) are in the current lamprey annotation while of the ten possible receptors, only two are annotated (rxfp3, rxfp3-3), another one (rxfp3-1) is reported as novel but is almost completely **annotated** and another six are present as fragments of insulin receptors (insr, insrr and igf1r) and are partially but not fully annotated (Tables 1.2 & 3.3). Ultimately, this pipeline identified 12 **Trinity ID's** of insulin family genes and reported four novel genes (igf1, insrr, relaxin-like and RXFP2-like); these genes are not included in the current annotation of sea lamprey (Table 3.3). Thus, by focusing on this gene family, we could help resolve problems in gene annotation while also contributing information about the expression of these genes, including potentially novel members of the family in lampreys. This clearly suggests that the *de novo* assembly pipeline can be used successfully for identifying novel and specific genes of interest. Thus, both the genome-guided and *de novo* assembly pipelines identified novel (i.e., those not yet identified in the sea lamprey genome) and unannotated genes expressed during ovarian development in lampreys (Table 3.3). These pipelines are user friendly and easily reproducible; hence, they can be used for any non-model organism.

## **4.2. Relevance of results obtained from both the pipelines**

One of the main objectives of this project was to determine whether the *de novo* assembly can identify the genes of the insulin family reported by the genome-guided pipeline. Almost all the genes – *igf2*, *insr-like*, *igf1r-like*, *insrr-like* reported by the HTSeq pipeline were also identified in the *de novo* assembly pipeline. This indicates that the results obtained from both these pipelines complement each other. However, the *de novo* assembly identified some novel genes not reported by the genome-guided pipeline (Table 3.3).

## **4.3. Findings of the study**

One of the main contributions of this study is that both the pipelines worked well and can be used to address specific biological questions. The genome-guided pipeline identified 1,552 up-regulated and 750 down-regulated Ensembl ID's in lampreys during ovarian development (table available upon request). This can be used to study the expression of these genes in more detail using gene-specific approaches such as qRT-PCR which can provide clues about genes that are turned on or off during different processes related to ovarian differentiation and sexual maturation, and may therefore help in determining differences in lamprey life history types. However, the sample size used in this study was quite low and so differential analysis was performed by pooling samples from different gonadal stages. This analysis uncovered that the gene expression of samples in the same gonadal stage were generally the same, irrespective of whether samples were obtained from parasitic or non-parasitic taxa, with the important exception of one ovarian sample from the parasitic chestnut lamprey which was in gonadal stage 4 but whose gene expression clustered with samples from the non-parasitic northern brook lamprey undergoing vitellogenesis (Chapter 2, Figures 2.6, 2.7 & 2.8). These results suggest that gene

expression across gonadal stages is overall similar between species, regardless of adult life history type, but as the adults mature, gene expression patterns diverge. It also suggests that the ovaries of parasitic lampreys that are in late metamorphosis (even as early as July, even though they will not spawn for another approximately 1.8 years; Docker, 2009) are already preparing for vitellogenesis and sexual maturation by up-regulating a suite of genes involved in these processes. This list of normalized count can be used for targeting specific genes of interest and for designing primers for qPCR.

One of the novel approaches taken in this study was the method by which I assigned **Gene Ontology** terms to the genes and transcripts identified in lampreys. Generally, researchers use a program known as BLAST2GO (Conesa *et al.*, 2005) for this purpose, but in this project, I employed a different approach primarily because the BLAST2GO algorithm is very slow (up to 203 months per sample). In lampreys, since few annotated genes have been assigned gene ID's, most genes do not have assigned GO terms and it is not possible to categorize genes based on processes they are involved in. To overcome this problem, all the up- and down-regulated lamprey Ensembl ID's reported by DESeq2 were converted to their **putative** orthologs in zebrafish (5,401 up-regulated and 2,730 down-regulated). This gene list was then ranked by level of expression or of differential expression by GOrilla and REVIGO. Using this approach, all the up-regulated and down-regulated genes were assigned GO terms based on the *p*-value, FDR *q*-value and enrichment were visualized in two-dimensional plots, interactive graphs, tree maps and tag clouds. It was observed that up-regulated genes had Gene Ontology terms related to cellular and reproductive processes and down-regulated genes were mostly involved in signaling pathways. This suggests that up-regulated genes were involved in regulating different processes related to ovarian differentiation and sexual maturation in lampreys. This approach led to the

identification of genes such as *zona pellucida glycoprotein 3a* as a putatively up-regulated gene. This gene is known to be expressed in growing oocytes in many different fish species (Bobe *et al.*, 2008; Lubzen *et al.*, 2010). A second gene, *neuropeptide Y receptor Y8b*, was also found to be up-regulated in the northern brook lamprey during gonadal stage 4 (Figures 2.20 a, b, c & d); this gene is known to play diverse roles in regulating satiety, neuroendocrine axes, vasoconstriction, and cardiovascular remodeling (Pedrazzini *et al.*, 2003). On the other hand, *adrenoceptor alpha* was found to be up-regulated in chestnut lamprey during gonadal stage 4; this gene plays an important role in regulating embryogenesis and adulthood (Héctor *et al.*, 2017). Genes like *estrogen receptor 2* known to play a vital role in hormone signaling (Baker, 1997, 2003; Escriva *et al.*, 1997; Escriva, 2000; Thornton *et al.*, 2003), oogenesis and vitellogenesis (Gustafsson, 2003; Heldring *et al.*, 2007; Hess, 2003; Nilsson *et al.*, 2001) was found to be down-regulated in northern brook lamprey during gonadal stage 4. Previous studies reported that during ovarian maturation the expression of thyroid stimulating hormone receptor (TSH-R) increased in European sea bass *Dicentrarchus labrax* and channel catfish *Ictalurus punctatus* (Rocha *et al.*, 2007; Goto-Kajeto *et al.*, 2009); the same pattern was observed in chestnut lamprey during gonadal stage 4 (Figures 2.21 a, b & c). This suggests that these genes may be involved in processes related to ovarian differentiation in lampreys as well. However, future investigation is required (Section 2.7.6).

#### **4.4. Technical difficulties and limitations of the study**

The original intention of this study was to compare ovarian gene expression in two species of lamprey, *I. castaneus* and *I. fossor*, to identify genes that are differentially expressed during ovarian differentiation and sexual maturation. Because of small sample sizes, species-

wise comparison was not possible (Chapter 2, Section 2.5.6), thus limiting the ability to compare gene expression in species that show different trajectories related to ovarian development (Section 1.1.3.1). Samples from both species were pooled to identify genes that were overall up-regulated and down-regulated in lampreys during three different time points of gonadal development.

In the reference-guided pipeline as discussed above when the sea lamprey genome was used, the percentage of mapped reads was 12-45% and preliminary analysis suggested that some samples had contamination during sequencing; a major portion of the reads therefore did not align to the reference genome, which can be one of the reasons for low mapping percentage; in addition the parameters were customized in the pipelines during library preparation might have adversely affected the results. Thus, the genome-guided pipeline discarded all the unmapped reads, which is one of the limiting factors. Secondly, HTSeq calculated gene count only on mapped reads and the program did not allow recovery of information pertaining to reads which were counted as “No feature” (Table 2.4). It was also not possible to get the Gene Ontology terms directly related to lampreys, because most of the genes in sea lamprey do not have associated GO terms; thus, all the up- (2,864) and down-regulated (1,738) genes of DESeq2 were converted into orthologs of zebrafish to get the associated ontology in GOrilla. Thus, annotations were provided to zebrafish genes which were later re-converted back into lamprey orthologs. The results from GOrilla provided an overview of co-occurring or related GO terms found to be up- or down-regulated over gonadal stage.

In the second, *de novo* assembly, pipeline, a customized database was used for querying the transcriptomes of each species and putative gene names were assigned to ~35,793 transcript ID's of chestnut and ~69,797 to northern brook lampreys. For chestnut lamprey, the pipeline did

not work well; it recovered fewer transcript ID's showing homology to the chordate custom-made database, and even then, the tentative ID's contained many more false positives. Also, the *de novo* assembly of the chestnut lamprey genome identified only three Trinity ID's associated with the insulin gene family in chestnut lamprey while in northern brook lamprey, nine Trinity ID's were identified which belong to insulin family genes (Table 3.3).

One of the major difficulties was to assess and compare which genes in the *de novo* assembly pertain to the genes contained in the annotation of the sea lamprey reference. One approach would be to take the GI numbers or RefSeq ID's reported by BLAST and convert them into Ensembl ID's and then cross-reference them back to the lamprey orthologs. However, this proved impossible since some of the species contained in the database are not available on Ensembl and furthermore the orthologous Ensembl ID of many of the NCBI GI and RefSeq ID's could not be obtained. Hence, this approach was not useful for comparing the overall results. An alternate approach that will be attempted in future studies would be to map the transcripts obtained from Trinity to the sea lamprey genome to determine if *de novo* assembled transcripts identifies approximately the same 6,900 genes as reported by HTSeq. Ultimately, however, both pipelines were useful: the genome-guided pipeline (Figure 2.2) provided a more detailed analysis of already known genes with respect to the annotated reference genome, but the *de novo* assembly (Figure 3.1) was more useful for discovering new genes and splice variants in an efficient and relatively fast approach. Using the *de novo* analysis pipeline, several novel genes of the insulin superfamily (Table 3.3, Figures 3.2, 3.3 & 3.5) were confirmed, which, on its own, is a useful contribution to lamprey research.

#### 4.5. Directions for future research

- a) The results obtained in this thesis through the genome-guided and *de novo* assembly pipelines are based on different software which might have affected the accuracy of the results. Future studies could investigate the role of different software and their capability to generate true results.
- b) The list of up- and down-regulated genes and normalized count of genes generated in this thesis can be used as a starting point of evidence for many genes that may play an important role during ovarian differentiation and sexual maturation in lampreys. For examples, genes like *zona pellucida*, *oxytocin receptor*, *calcitonin receptor*, *thyroid stimulating hormone receptor*, *estrogen receptor*, *adrenoceptor alpha* are expressed in lampreys during different gonadal stages; the expression of these genes can be subsequently verified by using qRT-PCR.
- c) Another route that is worth pursuing is to use different tissues for locating insulin family genes or other specific genes of interest reported in the result and compare their gene expression across different gonadal tissues to understand the evolution of the insulin family of genes in lampreys.
- d) For identifying up- and down-regulated genes in lampreys, different gonadal stages of chestnut and northern brook lampreys were pooled together; given the small sample sizes, the present study could not determine differentially expressed genes between species. Future studies should be designed by taking multiple samples (at least six replicates) for each gonadal stage for differential analysis of genes.
- e) The lampreys used for this thesis were sampled from a limited number of rivers and locations; northern brook lamprey was sampled from three rivers and chestnut lamprey from only one. Thus, results may be affected by environmental variation. Future studies can be designed to study differential gene expression in parasitic and non-parasitic lampreys kept under the same

conditions to ensure that observed species-specific differences are not due to environmental factors.

f) All the sequence data generated in this thesis can be used for annotating new genes and will be useful for building phylogenetic trees of specific genes of interest.

g) For *de novo* assembly, a custom-made BLAST database of 10 chordates genomes was used for identifying the putative name of the all the transcripts. Future studies should focus on expanding the database by adding more species to the custom-made BLAST database and the parameters E-value should be modified to obtain more hits.

#### **4.6. References**

Anders S, Pyl PT, Huber W (2015) HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Baker ME (1997) Steroid receptor phylogeny and vertebrate origins. *Molecular and Cellular Endocrinology*, **135**, 101-107.

Baker ME (2003) Evolution of adrenal and sex steroid action in vertebrates: A ligand-based mechanism for complexity. *BioEssays*, **25**, 396-400.

Bobe J, Nguyen T, Mahé S *et al.* (2008) *In silico* identification and molecular characterization of genes predominantly expressed in the fish oocyte. *BMC Genomics*, **9**, 499.

Conesa A, Götz S, García-Gómez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Docker MF (2009) A review of the evolution of nonparasitism in lampreys and an update of the paired species concept. *American Fisheries Society Symposium*, **72**, 71–114.

- Eden E, Navon R, Steinfeld I *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.
- Escriva HF, Delauna F, Laudet V (2000) Ligand binding and nuclear receptor evolution. *BioEssays*, **22**, 717-727.
- Escriva HR., Hanni M-C, Langlois P *et al.* (1997) Ligand binding was acquired during evolution of nuclear receptors. *Proceeding in National Academy of Sciences U.S.A*, **94**, 6803-6808.
- Goto-Kazeto R, Kazeto Y, Trant JM (2009) Molecular cloning, characterization and expression of thyroid-stimulating hormone receptor in channel catfish. *General Comparative Endocrinology*, **161**, 313-319.
- Gustafsson JA (2003) What pharmacologists can learn from recent advances in estrogen signaling. *Trends in Pharmacological Sciences*, **24**, 479– 485.
- Hector A Cespedes, Kattina Zavala, Juan Opazo (2017) Evolution of the  $\alpha$ 2-adrenoreceptors in vertebrates: ADRA2D is absent in mammals and crocodiles, *bioRxiv*, **106526**, doi: <https://doi.org/10.1101/106526>.
- Heldring N, Pike A, Andersson S *et al.* (2007) Estrogen receptors: how do they signal and what are their targets. *Physiology Review*, **87**, 905–931.
- Hess RA (2003) Estrogen in the adult male reproductive tract: a review. *Reproduction Biology and Endocrinology*, **1**, 52.
- Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, **15**, 550.
- Lubzens E, Young G, Bobe J *et al.* (2010) Oogenesis in teleosts: how fish eggs are formed. *General Comparative Endocrinology*, **165**, 367–389.

- Manzon RG, Youson JH, Holmes JA (2015) Lamprey metamorphosis. In: Docker MF, editor. *Lampreys: Biology, Conservation, and Control*, volume **1**. Springer, 139-214.
- Nilsson S, Mäkelä S, Treuter E *et al.* (2001) Mechanisms of estrogen action. *Physiology Review*, **81**, 1535–1565.
- Pedrazzini T, Pralong F, Grouzmann E (2003) Neuropeptide Y: the universal soldier. *Cell and Molecular Life Sciences*, **60**, 350–377.
- Rocha A, Gómez A, Zanuy S *et al.* (2007) Molecular characterization of two sea bass gonadotropin receptors: cDNA cloning, expression analysis, and functional activity. *Molecular and Cellular Endocrinology*, **272**, 63-76.
- Smith JJ, Kuraku S, Holt C *et al.* (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genetics*, **45**, 415–421.
- Thornton JW, Need E, Crews D (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science*, **301**, 1714–1717.

## **Appendix 1: Acronyms**

1R and 2R refers to WGD events

BLAST Basic Local Alignment Search Tool

bp Basepair

cDNA Complementary DNA

CNS Conserved non-coding sequence

DNA Deoxyribonucleic acid

FSH Follicle-stimulating hormone

GH Growth hormone

GnRH Gonadotropin-releasing hormone

GO Gene Ontology

GOA Gene Ontology Association

GPCR G-protein coupled receptor

HP Hypothalamic-pituitary

HPG Hypothalamic-pituitary-gonadal

ID Identity document

IGF Insulin growth factor

INS Insulin

INSL Insulin-like

LH Luteinizing hormone

mRNA Messenger RNA

Mya Million years ago

NGS Next Generation Sequencing

PCR Polymerase chain reaction

QC Quality control

qRT-PCR Quantitative real-time PCR

RNA-Seq RNA sequencing

RLN Relaxin

RXFP Relaxin family peptide receptor

RNA Ribonucleic Acid

RSEM

RTK Receptor tyrosine kinase

SNP Single Nucleotide Polymorphism

UTR Untranslated region

WGD Whole Genome Duplication

WTSS Whole transcriptome shotgun sequencing

WGS Whole Genome Shotgun

## Appendix 2: Glossary

**A priori:** A rationale based on previous experimental or non-experimental observations.

**Ab initio method:** It is a method used for making predictions about genes based on different statistical properties of the input DNA sequence.

**Alignment:** It is a process in which two or more biological sequences (nucleotide or amino acid) are matched to identify the similarity level between sequences.

**Annotated:** It refers to the genes which have gene name and genomic location available on Ensembl Genome browser.

**Contigs:** In *de novo* assembly, the RNA-Seq reads are assembled into larger fragments to generate mRNA transcript. These assembled nucleotide sequences are known as contigs.

**De novo:** When a reference genome is not available, RNA-Seq reads are assembled to reconstruct full length transcripts by using different software package.

**E-value:** It is also known as expectation value; a statistical parameter used in BLAST for estimating the probability of obtaining hits or match by chance; lower E values indicates significant match.

**Ensembl ID:** It is the unique accession identifier assigned to gene or transcripts by Ensembl Genome browser.

**Gene Ontology:** It is used to classify genes and their products into different categories: Biological Process, Cellular Component and Molecular Function.

**High-segment scoring pairs (hsps):** It is used for generating an alignment without gaps between biological sequences (nucleotide or amino acid) that share a high level of similarity or homology.

**Isoforms:** In Trinity, alternative transcripts of the same gene are called as isoforms that have similarity with the gene sequence but are structurally different.

**Mapping:** It is the process of aligning short reads to a reference genome to identify genes and transcripts.

**Novel:** It refers to the genes which are not identified even in the reference genome.

**Orthologs:** Genes that evolved from a common ancestor through speciation.

**Paralogs:** Genes which have evolved from duplication within a genome.

**Putative gene:** It is a nucleotide sequence that has open reading frame but has not been assigned any gene name.

**Reference genome:** The annotated genome which contains information about the assembled chromosomes, coding and non-coding genes of a species genome assembly.

**Sex determination:** The potential of gonads to develop into an ovary or a testis.

**Sex differentiation:** The process by which the phenotypic sex of an individual is produced.

**Sexual maturation:** It is the final stage where changes occur in the gonad (ovary or testes) prior to reproduction.

**Transcriptome:** The transcriptome is the collection of all transcripts present in a cell, a population of cells (e.g., tissue or organ), or an organism at a developmental stage or functional form.

**Trinity ID:** It is the unique accession identifier assigned to assembled contigs (FASTA) in *de novo* assembly by the program Trinity.