TRANSCRIPTOME ASSEMBLY AND MOLECULAR EVOLUTIONARY ANALYSIS OF SEX-BIASED GENES IN THE GUPPY

(*Poecilia reticulata*)

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Eshita Sharma

aus Ambala, Haryana, Indien

Tübingen

2014

Tag der mündlichen Qualifikation:	16.03.2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Gerd Jürgens

Erklärung

Hiermit erkläre ich, dass ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind durch Angaben der Quellen kenntlich gemacht wurden. Die Doktorarbeit besteht teilweise aus Auszügen eigener Publikationen oder Publikationen, die derzeit in Vorbereitung sind.

Tübingen, November 2014
Eshita Sharma

Acknowledgements

"Guru Govind dono khade kake lagu paay
Balihari Guru aapne Govind diyo batay. "
- Kabir (1440 – 1518), India

The PhD journey has been a wonderful adventure and it's hard for me to believe that I finally wrote my thesis. Albert Einstein is quoted to have said, *"I am thankful to all those who said no, it is because of them that I did it myself",* but scientific study is rarely an individual effort and I will now attempt to thank all those who helped me in my research and without whom this document would not have been possible. Foremost, I am indebted to my supervisors, Christine Dreyer and Detlef Weigel, without whom I would never have got this opportunity to discover the fascinating world of guppies and evolutionary functional genomics. I consider myself privileged to have learned and developed under their instruction and support. For a biochemist with a prejudiced view of bioinformatics, the learning curve was a treacherously slow process and would not have been possible without their encouragement, support and patience. I would like to thank the members of my PhD Advisory Committee, Gerd Jürgens and Gunnar Rätsch, for their time and guidance that helped give a direction for the thesis. I would like to acknowledge Margarete, Verena, Bonnie and Axel for their invaluable collaboration over the course of my studies. Most of this work would not have been possible without Axel and Bonnie, who helped me with ideas and contributed in discussions. I am grateful to Eva, a former PhD student, and Gideon, a former Diploma student, members of the guppy group. They were my first instructors in bioinformatics. I thank Dino who often helped with discusions and analysis, many times at very early hours in the morning. I would like to thank Stefanja for always ensuring that our fish were healthy and happy and for supporting my attempts at conversations in German; Kristin, Philipp, Alexandra and Tim for helping with the fish photographs and for the wonderful chat sessions in the fish house; and Christa, Andrea and Jens in the genome lab for maintaining an exceptional lab and for helping with library preparation and sequencing. I would like to thank Maricris, Subhashini, Darya, Juan Diego, Xi, Marco, Dan, Beth and Jörg who helped me at various stages of the project. My research work would never be completed without the help of Andre Noll for cluster

computing facilities and data organization on the server. I am grateful to Huelya who helped organize a lot of my life in the institute and in Tuebingen. I am grateful to Rebecca for her help, suggestions and for all the chocolates and cocktails. I would like to thank Axel, Bonnie and Beth for reading and providing suggestions to improve my thesis. I am grateful to Verena for translating the abstract of my thesis into German. Unfortunately, my German is not as good as hers. I would also like to thank all members of the Weigel lab for providing a friendly and social atmosphere in the laboratory. I am also thankful to Cate and my colleagues at Oxford for their support and encouragement that was very helpful in the last stage of writing. I am deeply grateful for the friends I made during the course of my studies in Tuebingen - Cris Subu, Dino, JD, Johannes, Diep, Verena, Vini, Marco, Felipe, Eunyoung, David, Carmen, Jathish, Francelli, Edgardo, Anna-lena, Patricia, and many other members and visitors at Weigel lab. The good times together and their support that were of immense help all these years. Especially Cris, Subu, Vasuki, Janani and Dino have been there through the good and the not so good times. Thankfully, their presence made all times better. A special thanks also to Eunyoung and Sang-Tae, for the helpful discussions, the support and for the time they provided me with a roof. I would also like to thank my housemates Janna, Olga, Eva, Elisa, Fred, Can, Mohammad and Julie who made me experience the best of WG life and beyond. I owe you guys for the evening conversations, the steaming dinners and the desserts. I would like to thank Janani, Vasuki, Roopika, Ratna, Prajwal, Anurag, Vaishnavi and all the gangs of India. Thanks to them it was possible to have a home away from home. We would be nowhere without the ground we stand on and I am thankful to my friends and family in\India for being an integral part of my ground. I am extremely grateful to my parents and grandparents for keeping their faith in me and for pushing and motivating me when I needed it the most. I would like to thank my brother for showing me home experimentation that kindled my interest in science. I would like to thank my partner for making the effort to understand my work, for the many pep-up talks and for the celebratory champagnes. Last but not least, I would like to thank the German taxpayers for the financial support and my readers for their interest. My research would not have materialized without their contribution. I hope that this dissertation is helpful for others.

In accordance with the standard scientific protocol, I will use the personal pronouns “we” or “us” to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Table of figures

List of abbreviations and acronyms

BAC	Bacterial artificial chromosome
Bp	Base pair
CDS	Coding sequence
CPM	Counts per million
dN	number of non-synonymous substitutions per non-synonymous site
dS	number of synonymous substitutions per synonymous site
ECM	Extracellular matrix
EST	Expressed sequence tags
F_{adult}	Female adult
F_{brain}	Female brain
F_{embryo}	Female embryo
F_{gonad}	Female gonad
F_{tail}	Female tail
FC	Fold change
FDR	False discovery rate
FPKM	Fragments per kilo base per million fragments mapped
GO	Gene ontology
GRT	Guppy reference transcriptome
LG	Linkage group
M_{adult}	Male adult
M_{brain}	Male brain
M_{embryo}	Male embryo
M_{gonad}	Male gonad
M_{tail}	Male tail
NR	NCBI non-redundant protein database
ORF	Open reading frame
QTL	Quantitative trait loci
SBG	Sex-biased genes
SDL	Sex determining locus
SNP	Single nucleotide polymorphism
UTR	Untranslated region

Zusammenfassung

Es ist ein in der Natur weitverbreitetes Phänomen, dass Männchen und Weibchen einer Art aufgrund von geschlechtsspezifischem Selektionsdruck unterschiedliche Phänotypen besitzen. Die Evolution und Aufrechterhaltung von sexuellen Dimorphismen geht im Allgemeinen mit Unterschieden in der Genexpression in den zwei Geschlechtern einher. Gene, die stärker in einem Geschlecht exprimiert werden, sogenannte geschlechtsabhängige Gene, zeigen oft eine beschleunigte molekulare Evolution. Zudem sind geschlechtsabhängige Gene auf den X- oder Z-Chromosomen von etlichen Arten im Vergleich zu den anderen Chromosomen überrepräsentiert. Im letzten Jahrzehnt hat die Forschung bezüglich des Verständnisses von Geschlechtsdetermination in Drosophiliden, Säugetieren und Vögeln große Fortschritte gemacht, es ist jedoch relativ wenig über die Evolutionsrate und genomische Position von geschlechtsabhängigen Genen in Teleostei-Arten bekannt, die zum größten Teil undifferenzierte und evolutionär gesehen junge Geschlechtschromosomen besitzen.

Ein Paradebeispiel dafür ist der Guppy, auf den sich meine Arbeit konzentriert. Guppys sind Süßwasserfische mit einem XY-Geschlechtsdeterminationssystem und einer Y-gekoppelten Vererbung von Eigenschaften, die vorteilhaft für Guppymännchen sind. Guppys sind durch einen Geschlechtsdimorphismus in Größe, Pigmentierung und Verhalten gekennzeichnet, Eigenschaften, welche in der Natur sowohl durch natürliche als auch durch sexuelle Selektion beeinflusst werden. Um die geschlechtsabhängigen Gene des Guppys zu identifizieren, assemblierte ich zuerst ein Referenztranskriptom aus cDNA-Sequenzierdaten mit hoher Abdeckung.

Ich verglich unterschiedliche Transkriptomassemblierungsmethoden für RNA-Sequenzierungsdaten (RNA-seq) und erstellte und annotierte ein Referenztranskriptom bestehend aus einer Genom-unabhängigen und einer Genom-abhängigen Assemblierung. Danach untersuchte ich die Expression von geschlechtsabhängigen Genen im Gehirn, Schwanzbereich und den Gonaden, da diese Gewebe in adulten Guppys sexuell dimorph sind. Dabei fand ich gewebespezifische Expression, die mit dem sexuellen Dimorphismus des Phänotyps in Zusammenhang steht. Kurz zusammengefasst wurden Signaltransduktions-, Pigmentierungs- und Spermatogenesegene stärker in Männchen exprimiert, wohingegen Gene, welche in Weibchen stärker exprimiert wurden, mit Wachstum, Zellteilung,

Organisation der extrazellulären Matrix, Nährstofftransport und Follikulogenese in Zusammenhang stehen. Da die männliche Geschlechtsdetermination und -differenzierung im Guppy vermutlich mit den männlich-spezifischen Farbmustern assoziiert sind, habe ich die Genexpression und genomische Position von Guppy-Orthologen von Pigmentierungskandidatengenen analysiert, die in anderen Wirbeltieren eine Rolle in der Farbentwicklung spielen. Die Kandidatengene konnten genau auf dem weiblichen Genom positioniert werden und es konnte kein Kandidat, der nur in Männchen exprimiert wurde, identifiziert werden. Das Ausmaß und die Richtung der geschlechtspezifischen Expression war von etlichen Genen, die mit dem Geschlecht in Zusammenhang stehen und von einigen Pigmentierungsgenen, gewebespezifisch.

Als ich die Verteilung aller geschlechtsabhängigen Gene im Genom untersuchte, entdeckte ich, dass auf den Geschlechtschromosomen (Kopplungsgruppe 12) Gene, die mit den Ovarien assoziiert sind, über- und Gene, die mit den Hoden assoziiert sind, unterrepräsentiert sind. Ein genomweiter Vergleich der Evolutionsrate der geschlechtsabhängigen und -unabhängigen Gene, gemessen an dem Verhältnis der nicht-synonymen Austauschrate (d_N) zu der synonymen Austauschrate (d_S), deutete in allen drei Geweben darauf hin, dass Gene, die mit Hoden und Ovarien assoziiert sind, schneller evolvieren. Die Gene, die höher im weiblichen Gehirn exprimiert wurden, zeigten unabhängig vom Umfang und der Höhe der Expression ein erhöhtes Ausmaß an nicht-synonymen Austäuschen.

Nach ausführlicher Evaluation der vorhandenen Assemblierungsmethoden stellt diese Studie ein umfangreiches Referenztranskriptom des Guppys zur Verfügung. Das Referenztranskriptom stellt eine molekulare Ressource dar, von der ausgehend die komplexen adaptiven Merkmale des Guppys untersucht werden können. Der Vergleich der genomweiten differentiellen Expression zwischen männlichen und weiblichen Geweben führte zu der Identifizierung von Kandidatengenen, die vermutlich zu dem sexuellen Dimorphismus, der mit den Geweben assoziiert ist, beitragen. Die Liste an Kandidatengenen dient auch als eine Referenz für zukünftige Studien über reproduktive und somatische Geschlechtsunterschiede in Guppypopulationen und Poeciliiden sowie anderen Teleosten. Die unterschiedliche genomische Verteilung der Gene, die mit Ovarien und Hoden assoziiert sind, zeigt, dass es geschlechtsspezifischen Selektionsdruck gibt, der durch die ungleiche Verteilung der geringfügig differenzierten Geschlechtschromosomen des Guppys agiert. Die erhöhten Nukleotidsubstitutionsraten, die in Gonaden-abhängigen und weiblich-abhängigen Genen im

Gehirn beobachtet wurden, stimmen mit der Hypothese einer beschleunigten Proteinevolution ausgelöst durch sexuelle und entspannte purifizierende Selektion überein. Diese Ergebnisse bilden die Grundlage für zukünftige Experimente, die die Variation zwischen Guppys aus Populationen mit unterschiedlichem Ausmaß an sexuellem Dimorphismus und vermutlich variierenden sexuellem und natürlichem Selektionsdruck untersuchen werden. Zudem stellen die Ergebnisse eine Referenz zur Verfügung mittels der interspezifische Variationen in Genen, die möglicherweise vorteilhaft für ein Geschlecht sind, in eng verwandten Poeciliiden mit diversen Geschlechtsdeterminationsmechanismen erforscht werden kann.

Abstract

It is a phenomenon universally seen that males and females of a species show phenotypic differences as they evolve under often diverging sex-specific selection pressures. The evolution and maintenance of their sexual dimorphism is generally associated with gene expression divergence between the sexes. Genes that show enriched expression in one sex, also called sex-biased genes, often show rapid molecular evolution. Furthermore, sex-biased genes have also been found to be over-represented on X or Z chromosomes in several species with differentiated sex chromosomes or neo-sex chromosomes. While research on sex-biased genes in drosophilids, mammals and birds has developed in the last decade, there is relatively little known about sex-biased genes in teleost species with largely undifferentiated sex-chromosomes of recent origin.

A case in point is that of the Trinidadian guppy, *Poecilia reticulata*, which is the focal species of my thesis. In this dissertation, I investigate sex-biased gene expression in guppy, a fresh-water fish with XY sex-determination and Y-linked inheritance of male-advantageous traits. Guppies display sexual dimorphism in size, ornaments, and behavior, traits that are shaped by both natural and sexual selection in the wild. My first task was to assemble a transcriptome reference using deep sequencing of cDNA. I compared several methods of assembly with RNA sequencing (RNA-seq) data and assembled and annotated a reference transcriptome combining a genome-independent and a genome-guided assembly. Subsequently, I analyzed sex-biased gene expression in brain, tail and gonads, tissues with overt sexual dimorphism in adult guppies. I found tissue-specific expression generally related to the phenotypic sexual dimorphism. For example, genes related to signal transduction, pigmentation processes and spermatogenesis were expressed more in males; while female-biased genes related to growth, cell-division, extra-cellular matrix organization, nutrient transport, and folliculogenesis. As male sex-determination and differentiation in guppies is believed to be associated with the male-specific pigment patterns, I analyzed the gene-expression and genomic locations of guppy orthologs of candidate genes functional in these processes in other vertebrates. The list of candidate genes could be specifically aligned to the female genome and no male-limited candidate could be identified. I found tissue-specificity in the magnitude and direction of sex-bias in the expression of several sex-related and pigmentation genes.

I then studied the genomic distribution of all sex-biased genes. I observed the accumulation of ovary-biased genes on the putative sex linkage group, LG12. Genome-wide comparison of rates of evolution of sex-biased and unbiased genes, measured by the ratio of non-synonymous substitution rate (d_N) to the synonymous substitution rate (d_S), indicated faster evolution of testis-biased genes, and female-biased genes in all three studied tissues. Among these, the female-biased genes in brain showed elevated ratios of non-synonymous substitutions irrespective of the breadth and magnitude of expression.

In this study, I describe a comprehensive annotated guppy reference transcriptome that is compiled after extensive evaluation of different existing methods for assembly using *de novo* strategies as well as reference-guided strategies. The reference transcriptome of the guppy provides a resource for investigating the molecular genetics of the guppy's complex adaptive traits. The methods and pipelines are generally applicable for developing and utilizing transcriptomic resources in organisms with limited molecular resources.

Genome-wide differential expression between male and female tissues, allowed us to identify genes with strong characteristic differential expression in the differentiated gonads as well as genes with small but significant expression differences in the somatic tissues. These sets of sex-biased genes may be relevant for the tissue-associated sexual dimorphism. Differential genomic distributions of ovary- and testis-biased genes provide evidence for sex-specific selection pressures acting on the slightly differentiated sex chromosomes of the guppy. Elevated rates of molecular evolution observed in testis-biased and all categories of female-biased genes suggest evolution under distinct selection pressures on the reproductive versus non-reproductive tissues. Overall, these results are useful for guppy researchers and for further understanding the evolution of sex differences in diverse species.

Publications and Personal Contribution

Published papers:

Transcriptome assemblies for studying sex-biased gene expression in the guppy, Poecilia reticulata. *BMC Genomics* 2014 15:400.

This work has been published in BMC Genomics (Sharma, et al. 2014) with the following authors: Eshita Sharma (ES), Axel Künstner (AK), Bonnie A. Fraser (BAF), Gideon Zipprich (GZ), Verena A. Kottler (VAK), Stefan R. Henz (SRH), Detlef Weigel (DW) and Christine Dreyer (CD). ES, BAF and CD conceived the study and designed the experiments. CD, ES and BAF performed the dissections. ES performed the RNA extraction and library preparations. ES performed the data handling and assembly comparisons. ES and AK performed the transcriptome assemblies. ES performed the gene ontology (GO) annotations and ortholog identification. ES performed the differential expression analysis. BAF and ES performed the GO enrichment analysis. ES, VAK and CD performed the pigmentation ortholog analysis. ES and AK performed the molecular evolution analysis. ES, GZ and SRH explored assembly strategies and helped with data handling and scripts. DW contributed reagents, materials, helped with discussions and analysis. ES wrote the paper with contributions from AK, BAF and CD and comments and revisions from VAK, SRH, GZ, and DW. All authors read and approved the final manuscript.

Chapter 1: General Introduction

1.1 Sexual dimorphism

The evolution of differences between sexes was first described by Darwin when he proposed the theory of sexual selection to explain the extravagant display traits frequently seen in males of a species (Darwin 1871). The divergence between sexes in their morphology, behaviour, physiology as well as life-history traits is in fact ubiquitous in the eukaryotic domain, and its prevalence is attributed to differential selection pressures acting on the two sexes.

Sex-specific selection pressures

Sexually reproducing species differ in the reproductive investment by the male and female parents. This leads to conflict in their optimal reproductive strategies as one sex often becomes a limiting resource. Darwin described sexual selection as the competition between individuals of the non-limiting sex (typically males), for reproductive success with the limiting sex (typically females). This competition for reproductive advantage may manifest within members of the same sex (intra-sexual) or between the two sexes (inter-sexual) and often leads to sex-specific development of advantageous traits. Intra-sexual selection, such as male-male competition where males directly compete with each other for female access, often leads to the development of traits that help the competing sex, *e.g.* deer antlers or beetle horns. Conversely, inter-sexual selection, such as female choice, where reproductive access is determined by mate-choice often drives the evolution of traits in the non-limiting sex that are attractive to the limiting sex, *e.g.* the peacock's tail, reviewed in (Berglund, et al. 1996). These decorative or armour traits are advantageous for one of the sexes (often males), but unnecessary and usually absent in the other sex. In addition to competition for reproductive advantage, males and females of a species also experience sex-specific ecological pressures that affect their survival or fecundity differently (Darwin 1871; Fisher 1958; Selander 1972). Therefore males and females may optimize different fitness traits to suit their specific ecological niches.

Molecular divergence between the sexes

The overall competition between the sexes to enhance their specific fitness often results in sexually antagonistic selection of traits, *i.e.* selection of traits advantageous in one sex even at the cost of detrimental effect in the other sex. This sexual conflict is believed to be alleviated by sex-specific trait development ultimately resulting in sexual dimorphism (Bonduriansky and Chenoweth 2009; Lande 1980; Rowe and Day 2006; van Doorn 2009). But how does such spectacular dimorphism emerge in the presence of genetic constraints imposed by a shared genome? When male and female fitness differs for a shared trait, its equal expression in both sexes will be sub-optimal for the fitness of either sex. Therefore selection should favour its expression in the sex for which it is advantageous and suppress its expression where it has detrimental effect (illustrated in Figure 1.1). In species with genetic sex determination some traits can be separated between sexes by sex-linked or sex-limited genes; *e.g.* male-specific gene expression can be regulated by the Y chromosomes in XX-XY systems and female-specific gene-expression by W chromosomes in ZZ-ZW (Coyne, et al. 2008; Mank 2009; Rhen 2000; Rice 1984). But all differences between sexes are not sex-limited and a majority of sexually dimorphic traits are encoded on their shared genomes and expressed in both sexes. A common assumption made, is that genome-wide regulatory differences resolve ongoing sexual conflict by preferential gene-expression in the sex (and tissue) where it is advantageous (Ellegren and Parsch 2007; Parsch and Ellegren 2013). In most species a large number of autosomal genes show sex-differences in gene-expression, isoform-abundance, gene-splicing, imprinting, and sub-functionalization or neo-functionalization of duplicated genes (Connallon and Clark 2010; Gallach, et al. 2011). These mechanisms for sex-specific modulation of the genome can allow less-constrained evolution and expression of sexually dimorphic traits in species with or without genetic sex determination (Fisher 1931; Rhen 2000; Rice 1984).

1.1.1 Sex-biased gene expression

While genetic sex determination occurs at a critical period in early development, phenotypic inter-sexual differences manifest throughout embryonic to adult life. In fact males and females exhibit the strongest phenotypic differences as adults. Likewise, mature testes and ovaries are the most sexually dimorphic organs. This phenotypic divergence over time is complimented by an increase in gene expression divergence throughout development and is most explicit in the differentiated sexually mature gonads (Mank, et al. 2010; Vicoso, et al.

2013). Quantitative comparisons of cDNA from male and female tissues of diverse animal and plant species have shown that a large fraction of autosomal genes are differentially expressed between the sexes in their reproductive as well as non-reproductive tissues (Mank, et al. 2008a; Parisi, et al. 2004; Small, et al. 2009; Xia, et al. 2007; Yang, et al. 2006). Genes with differential expression between sexes are referred to as sex-biased genes (SBG), while genes that show equal expression in both sexes are called unbiased. SBG are subsequently divided, according to the sex which shows enriched expression, into male- and female-biased genes. Theoretically, if SBG contribute to the maintenance of intersexual phenotypic differences, their evolution should be subject to the same forces of natural and sexual selection that shape the evolution of sexual phenotypes. Accordingly, patterns of sex-biased expression have also been related to the extent of sexual dimorphism (Pointer, et al. 2013; Stuglik, et al. 2014), sexual antagonism (Innocenti and Morrow 2010), sex-linkage (Meisel, et al. 2012) and evolutionary turnover (Zhang, et al. 2007), partly reviewed in (Ingleby, et al. 2014).

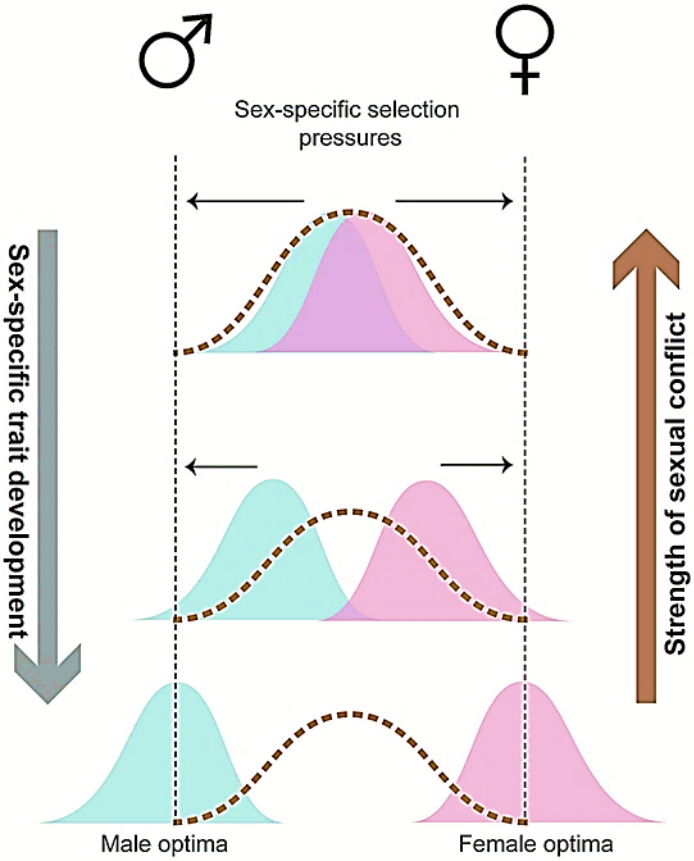

Figure 1.1: Differential fitness optima of males and females drives sex-specific trait development. Traits that are advantageous for males (blue curve) or females (pink curve) are at sub-optimal fitness when they are expressed at same levels in both sexes (brown dashed curve). Preferential expression of these traits at levels closer to their optimal fitness (black dashed lines) reduces sexual conflict and allows sex-specific trait

development.

1.1.2 Chromosomal location of sex-biased genes

In species with well-differentiated sex chromosomes, sex-linked genes are under differential sex-specific selection owing to the reduced population size and hemizygous state of the X or Z chromosome in males or females respectively. In addition to this, morphologically different (heteromorphic) sex chromosomes may also have unique properties such as dosage compensation (DC) and meiotic sex chromosome inactivation (MSCI). These biological features have been postulated to play a central role in the distribution of sex-biased genes as well as in sexual conflict resolution (Parsch and Ellegren 2013).

Previous research on genomic distribution of SBG in organisms with heteromorphic sex chromosomes has shown their non-random distributions. A positive association has been found between biased expression in the homogametic sex and X- or Z-linkage in mouse, *Mus musculus* (Khil, et al. 2004); silkworm, *Bombyx mori* (Arunkumar, et al. 2009); chicken, *Gallus gallus* (Ellegren 2011; Kaiser and Ellegren 2006); roundworm, *Caenorhabditis elegans* (Reinke, et al. 2004) and Drosophila species (Assis, et al. 2012; Khil, et al. 2004; Ranz, et al. 2003). This pattern of sex-linkage is also predicted by theoretical models of sex-specific selection where non-random distribution may possibly arise through any of the following three mechanisms that operate in species with non-recombining sex chromosomes (Gallach, et al. 2011; Mank 2009). Though these mechanisms may exist in both XY and ZW systems, here I describe them using the example of an XY system of sex determination.

i) **Sexual antagonism:** As the effective population size of the X-chromosome is greater in females than in males, selection can either favour X-linkage of female-beneficial alleles or disfavour X-linkage of male-beneficial alleles. The ultimate result is dependent on the degree of dominance of the sexually antagonistic mutation and the direction and magnitude of opposing selection coefficients in the sexes (Connallon and Clark 2010; Rice 1984).

ii) **Meiotic silencing of unsynapsed chromatin:** Meiosis specific pairing and exchange of genetic material between homologous chromosomes is either partial or missing in differentiated sex chromosomes. The lack of pairing between homologous chromosomes triggers meiotic silencing of unsynapsed chromatin/unpaired DNA (MSUC or MSUD). A special case of MSUC is the lack

of pairing in differentiated X and Y chromosomes called meiotic sex chromosome inactivation (MSCI) (Turner 2007). Demasculinization of the X-chromosome with respect to testis specific genes, can result from selection against this transcriptionally inactive state of the X-linked genes (Hense, et al. 2007).

iii) **Dosage compensation:** Sex-specific selection can further arise due to dosage constraints on stoichiometry sensitive gene products. In heterogametic XY males with non-uniform dosage compensation of X-linked genes (Ohno 1967), potential stoichiometric imbalance between expression of X-linked and autosomal male-beneficial genes makes the X chromosome an unfit location for male-beneficial genes (Meisel, et al. 2012; Parisi, et al. 2003; Vicoso and Charlesworth 2009).

However species with homomorphic and/or nascent sex chromosomes lack the need for global dosage compensation and meiotic sex chromosome inactivation as most X- or Z- linked genes are present in two copies. Therefore, the chromosomal distribution of sex-biased genes in species with less differentiated or homomorphic sex chromosomes can provide clues about the origin of these mechanisms and intermediate stages in sex chromosome evolution.

1.1.3 Molecular evolution of sex-biased genes

The relationship between sex-biased expression and coding sequence evolution is an area of ongoing research. In general, protein evolution through amino-acid substitutions can either be constrained for preservation of protein function (purifying selection), or accelerated by positive selection of favourable functional changes. Sequence changes can be quantified by comparison of homologous genomic sequences between diverged species to identify lineage-specific increases in nucleotide substitution. A basic assumption is that changes at non-synonymous sites that may cause amino-acid substitutions are rare unless accelerated by selection or relaxed constraints. By contrast changes at synonymous sites are silent and are a measure of neutral changes. An intuitively simple measure of excess substitutions relative to baseline mutation rate is the ratio d_N/d_S (also called ω or Ka/Ks). This is the ratio of number of non-synonymous substitutions per non-synonymous site (d_N) to number of synonymous substitutions per synonymous site (d_S) (Yang 2006). Higher than average rates of non-synonymous substitutions (d_N) or d_N/d_S values indicate accelerated evolution of the coding sequence.

Genomic comparisons of evolutionary dynamics of sex-biased genes with unbiased genes have shown faster evolution of sex-biased genes. Research on SBG in Drosophila species

suggest that male-biased genes evolve more rapidly at the protein level than female-biased or unbiased genes (Haerty, et al. 2007; Meisel 2011; Zhang, et al. 2004) reviewed by (Ellegren and Parsch 2007). However in some species with non-XY systems of sex determination, rapid evolution of female-biased genes has also been observed (Malone, et al. 2006; Mank, et al. 2007; Whittle and Johannesson 2013). Similar observations were made for female-biased genes in the somatic tissue of gonadectomized drosophila (Meisel 2011). Varying evolutionary dynamics of SBG suggests that different selection pressures may be dominant on the reproductive and somatic tissues. Initially the rapid divergence of SBG had been suggested to result from positive selection via male-male competition, female-choice or sex-specific natural selection on the sexually dimorphic phenotypes. Ergo several studies relate the rapid evolution of sex-related genes with the large inter- and intra-specific variation seen in secondary sexual characteristics (Andersson 1994; Parsch and Ellegren 2013; Svensson and Gosden 2007). However, recent analyses indicate that accelerated divergence of SBG may be largely apportioned to their narrow expression breadths in a tissue-specific or developmental stage-specific manner (Meisel 2011; Perry, et al. 2014). Gene dispensability or reduced functional pleiotropy of genes with narrow expression breadth may relax purifying selection and result in accelerated rates of non-synonymous substitutions (Ellegren and Parsch 2007; Mank and Ellegren 2009; Ellegren and Parsch 2013). Moreover, these mechanisms are not mutually exclusive and may all be involved in the observed rapid divergence of genes associated with sexual structures and reproductive roles (Parisi et al. 2004; Haerty et al. 2007; Ellegren and Parsch 2007). For example, sperm-surface proteins in the testis may be under sexual selection manifested by sperm competition or egg-coat proteins, but they also may have gonad-specific expression and functions and therefore evolve under relaxed purifying selection.

For a comprehensive understanding of evolution of inter-sexual divergence and the role of sex-biased genes, we need to gather more empirical data from studies on non-model organisms. A wider perspective in this research can be obtained by examination of sex-biased genes in species where the sex chromosomes are of recent origin and sex-determination mechanisms are variable and/or reversible. Such an opportunity is presented by different sexual phenotypes among species of teleost fish. Species within the *Poeciliidae* family exhibit varying mechanisms of genetic sex-determination accompanied by frequent sex-reversal as well as visible and often spectacular sexual dimorphism. Among poeciliids *Poecilia reticulata*, or the Trinidadian guppy, has an illustrious research history in regard to its sexual dimorphism yet the molecular mechanisms underlying these traits have not been explored

(introduction to Guppy biology in Chapter 1.3).

1.2 RNA sequencing and *de novo* assembly

Over the last two decades our understanding in molecular biology has grown in leaps and bounds with the advent of high-throughput genome profiling methods. Advances in DNA sequencing and the development of microarrays enabled genome-wide studies in comparative and functional genomics. System-level studies including transcriptome quantification, profiling of DNA-protein interactions, and characterization of genetic variation have contributed significantly towards disease classification, clinical diagnostics, therapeutics, agriculture, environment studies, evolutionary biology and many other disciplines (DeRisi, et al. 1996; Liao and Zhang 2006; Sotiriou and Pusztai 2009; Takata, et al. 2005; van 't Veer, et al. 2002; White, et al. 1999). However studies with DNA microarrays, although powerful, are dependent on prior knowledge of genome sequence information and gene anotations. This was changed by the development of low cost, massively parallel DNA sequencing methods hailed as the next generation of sequencing (NGS) (Metzker 2010; Shendure and Ji 2008; Wang, et al. 2009). These technologies have enabled the study of a diverse array of organisms enhancing our understanding of many facets of biology. In particular, RNA sequencing (RNA-seq) can be used to characterize trancriptomes of organisms by simultaneously sequencing genes, quantifying gene expression and identifying variants across multiple samples (Barbazuk, et al. 2007; Ozsolak and Milos 2011; Wang, et al. 2009).

1.2.1 RNA-seq for the study of ecological model organisms

Applications of reference based transcriptome assembly and hybridization-free count-based gene expression quantification have facilitated better transcriptome characterization, including analysis of strand-specificity, mapping of fusion transcripts, identification of novel non-coding RNAs and splice variants (Martin and Wang 2011; Mortazavi, et al. 2008; Trapnell, et al. 2010) as well as analysis of expression dynamics of single-cells (Hashimshony, et al. 2012; Islam, et al. 2011; Tang, et al. 2009). In addition to these a substantial advantage of RNA-seq lies in its use for large-scale molecular analysis of organisms where no previous genomic information exists or which have only draft genomes with non-validated and incomplete gene annotations. In fact, the use of RNA-seq for *de novo* transcriptome assembly in organisms important for environmental, agricultural, ecological and evolutionary research has led to the identification of novel candidate genes and pathways (Meyer, et al. 2011; Stewart, et al. 2013; Xu, et al. 2013) and expanded molecular resources for addressing a diverse array of species-

specific biological problems (Balakrishnan, et al. 2014; Moghadam, et al. 2013) reviewed in (Ekblom and Galindo 2011; Strickler, et al. 2012). A recurring problem inherent to these studies is the assembly and annotation of a comprehensive reference transcriptome from short-read data of variable coverage (Gongora-Castillo and Buell 2013; Martin and Wang 2011).

1.2.2 Assembly methods and assemblers

The process of reconstruction of full-length transcripts is a different bioinformatics challenge as compared to *de novo* assembly of genomes. Unlike genome sequences, the coverage of transcriptome sequences varies in accordance to gene-expression levels. Multiple splice-variants add an additional level of complexity over and above allelic variants, paralogs, homeologs, and pseudogenes. Moreover, transcripts encoded by adjacent loci can be erroneously fused to form a chimeric transcript. Transcriptome assembly algorithms also face some challenges similar to genome-assemblers such as – i) accurate reconstruction using short-reads with sequencing errors; ii) uneven coverage across sequence length due to sequencing biases; and iii) ambiguities introduced due to conserved domains in closely related and duplicated genes. The existing assemblers use varying strategies to address these challenges and assembler-specific parameters can be modified to optimize assemblies. Broadly speaking, there are two types of transcriptome assembly strategies, reference-independent or *de novo* assembly and reference-guided or *ab initio* assembly. Transcripts can be reconstructed using either of these or a combined strategy that merges the two depending on the availability of a reference genome (Martin and Wang 2011).

De novo transcriptome assembly

Reference-independent assemblers assemble short reads using the *de Bruijn* graph approach in which reads are broken down into sequences of length *k* (*k*-mers) that form nodes and are connected by edges based on *k-1* bp overlap to build the sequence of the contig (Compeau, et al. 2011; Flicek and Birney 2009; Zerbino and Birney 2008a). *De novo* transcriptome assemblers using this approach include OASES (Schulz, et al. 2012), TRANS-ABYSS (Birol, et al. 2009; Robertson, et al. 2010), TRINITY (Grabherr, et al. 2011a) and SOAPDENOVO-TRANS (Xie, et al. 2014). Assembly optimization studies using different parameter values suggest that transcriptome assemblies are sensitive to *k*-mer length and coverage and assembly with a single *k*-mer often does not recover the complete gene-expression repertoire (Gruenheit, et al. 2012; Haznedaroglu, et al. 2012; Surget-Groba and Montoya-Burgos 2010). Theoretical

expectations and experimental evidence suggest that longer *k*-mer length results in fewer but long and high-coverage contigs while low-expressed genes are better assembled at shorter *k*-mers (Gibbons, et al. 2009; Zerbino and Birney 2008a).

Ab initio transcriptome assembly

Reference-guided assemblers first align reads to a reference genome followed by construction of a connectivity graph representing splice variants. Full-length transcripts are then assembled by traversing the graphs (Martin and Wang 2011). Splice-aware aligners such as BLAT (Kent 2002), TOPHAT (Trapnell, et al. 2009), TOPHAT2 (Kim, et al. 2013), GSNAP (Wu and Nacu 2010) and STAR (Dobin, et al. 2013) use either exon-first or seed-and-extend strategies for read alignment and splice-site predictions. Exon-first approaches are faster as they use the computationally intensive step of read splicing only for a subset of unaligned reads that potentially lie on an exon-junction. On the other hand, seed-and-extend strategies are more sensitive and can align a large number of reads and usually predict more splice variants (Garber, et al. 2011). Genome-guided assemblers like SCRIPTURE (Guttman, et al. 2010) and CUFFLINKS (Garber, et al. 2011) use an exon-first splice-aware aligner, TOPHAT/TOPHAT2, and have been successfully used for *ab initio* reconstruction of transcripts from coding and non-coding genes.

Another approach for transcriptome assembly is utilized in Genome-guided TRINITY (reference link in Chapter 2.2.4), which uses a combined strategy. RNA-seq reads are first aligned (using a seed-and-extend aligner GSNAP) to the genome and partitioned in read clusters according to genomic locus, followed by *de novo* transcriptome assembly at each partitioned locus. The TRINITY assembled transcripts may then be aligned back to the genome with GMAP (Wu and Watanabe 2005) and assembled into complete transcript structures with PASA (Haas, et al. 2008), thereby using information about genomic proximity, splice-sites and read support.

Nevertheless, for all these assembly approaches accurate splice-site identification and transcript reconstruction is challenging and the biological meaning of the numerous transcripts and splice-variants has not been demonstrated. Both *de novo* and genome-guided approaches have their own advantages and biases (Martin and Wang 2011; Steijger, et al. 2013). *De novo* assembly tools are unbiased and independent of errors and gaps in genome sequence. On the other hand, genome-guided assemblers are computationally less resource-intensive, can utilize genomic information to assemble full-length transcripts including low-abundance transcripts and are less affected by sequencing biases and errors. It has also been

found advantageous to use genomes of closely-related organisms, in the absence of a cognate reference genome, for *ab initio* transcriptome assembly or for improving *de novo* transcriptome assembly (Toth, et al. 2007; Ward, et al. 2012). In recent years, pipelines and algorithms have also been developed to augment transcript reconstruction by combining the output from different assemblers (Bao, et al. 2013; Jain, et al. 2013; Melicher, et al. 2014; Zhao, et al. 2011).

1.2.3 Metrics for comparing transcriptome assemblies

In order to appropriately utilize the assembled sequences, it is necessary to assess the quality and the biological content of the transcriptome assembly. Inspite of several studies comparing assembler performances, there is still no consensus on the criteria and metrics to gauge assembly quality. Most metrics require a set of well-established expressed transcripts as a reference (Martin, et al. 2010; Martin and Wang 2011). As a fallout of initial comparative studies of genome assemblies, transcriptome assemblies of previously non-sequenced organisms have largely been compared using count-based, assembly-size based and length-based metrics (Lu, et al. 2013). Maximising metrics such as singleton and contig count, average coverage, N50 , and overall assembly size is generally considered to be indicative of assembly completeness and complexity. However, a recent study assessing the accuracy, consistency and employability of these metrics shows that these metrics do not necessarily improve with assembly quality and alternate annotation-based metrics provide a more informative and biologically meaningful comparison(O'Neil and Emrich 2013). Associating assembled sequences with homologous proteins, however, depends on the evolutionary distance between species and is complicated due to lineage specific gene duplications, gene-losses, gene-expansions. Furthermore, it is dependent on the assembly quality of the distant relative. Annotation-based metrics may be a useful indicator of assembly completeness and redundancy, but still have limited-utility in assessment of the large number of splice-variants predicted by both *de novo* and *ab-initio* assembly methods. Therefore, accurate splice-variant annotation and transcript-level quantification is another challenging aspect of RNA-seq workflows. Recent large-scale studies for assessment of transcriptome assemblers (Steijger, et al. 2013) and RNA-seq read alignment tools (Engström, et al. 2013) show that no one method excels in all metrics. The authors of most comparison studies conclude that choice of appropriate assembly tool often depends on the type of data, organism under study and the ultimate research goal.

1.3 Guppy as a model system

The focal species of my dissertation is the Trinidadian guppy, *Poecilia reticulata*. The guppy is a live-bearing freshwater fish native to Trinidad and north-eastern South America. It has XY genetic sex-determination and shows an elevated degree of sexual dimorphism. Guppy populations have been intensely studied in evolution, ecology and behaviour for their vivid sexually dimorphic traits that evolve under natural and sexual selection in wild (Magurran 2005). Research on evolution of guppies in different ecological habitats separated by barrier waterfalls and mountain ranges has shown that variation in predator regimes results in marked differences among populations in colour patterns, behaviour and life history traits (Endler 1983; Reznick and Endler 1982; Reznick 1989). Guppies from high predation sites exhibit less intense colouration, rapid maturation, and have larger and more frequent broods with smaller offspring than their counterparts from low predation localities. Artificial introduction of guppies from high-predation to low-predation sites show rapid evolution of traits adaptive to the low-predation sites (Endler 1980; Gordon, et al. 2012a; Kemp, et al. 2009; Reznick, et al. 1997). Therefore, the guppy's population structure and easily identifiable phenotypes make it a particularly advantageous model to study evolution and ecological adaptation in wild.

1.3.1 Sexual dimorphism and sexual conflict

Of particular interest is its sexual dimorphism and Y-linked inheritance of male-advantageous sexually antagonistic loci. Male guppies display highly polymorphic colour patterns, while females show grey body colour resembling gravel present on the river-bottom in its habitat. Some male colour patterns are inherited by strictly male-specific Y-linked loci (Endler 1980; Winge 1927). The colour patterns are formed by combinations of pigment cells, mainly xanthophores, melanophores and iridophores together forming spots and stripe patterns of three predominant colours - "orange" formed by carotenoids and pterins; "black" formed by melanin; and "iridescent" structural colour including blue, green, violet and white (Kottler, et al. 2014). The conspicuous colour patterns are positively associated with mating success but also make the fish more visible to predators, therefore, the evolution of divergent colour patterns in natural guppy populations arises due to interplay of predator-intensity and female-preferences. In addition to the colour traits, male body-size and shape traits also show autosomal and Y-linked inheritance (Tripathi, et al. 2008; Tripathi, et al. 2009b). Study of the guppy's ornamental traits on and off its nascent Y-chromosome is important for

understanding the evolution of sex chromosomes for resolution of sexual conflict (Fisher 1931; Mank, et al. 2006; Postma, et al. 2011).

Sexual conflict is expected to be important in the guppy's mating system. Guppies have highly promiscuous resource-free mating where both sexes mate multiply. Fertilization is internal and the male uses his modified anal fin (the gonopodium) to transfer sperm to the female. The female can store sperm from multiple males thus providing an opportunity for post-copulatory selection by sperm-competition and cryptic female choice (Magurran 2001). Before copulation females display mate-choice with varying preferences for a number of ornamental and behavioural traits (Brooks and Endler 2001; Houde and Endler 1990). Male guppies exhibit both courtship display ("Sigmoid display") and sneaky mating ("Gonopodial thrusts") in an attempt to secure access to females dependent on female receptivity and ecological factors (Houde 1997; Liley 1966; Magurran 2005). Although females potentially gain substantial benefits from polyandrous mating (Evans and Magurran 2000; Ojanguren, et al. 2005), the females are receptive only as virgins or 3-4 days after parturition (Liley 1966). Female preferences for nuptial ornaments are not driven by fecundity benefits (Pilastro, et al. 2007). Overall the males benefit more from multiple mating and female guppies face high sexual harassment due to large number of mating attempts by males. In wild mature females receive on average one coercive mating attempt per minute with more harassment faced by females from high-predation environments (Magurran and Seghers 1994a).

Guppies show sex differences in body size, growth pattern, foraging behaviour, predator avoidance and other non-reproductive life-history traits (Magurran and Garcia 2000). Female guppies grow throughout their lives while male growth slows down after puberty. Female fecundity is a product of longevity and foraging efficiency and the female invests in maximizing its energy intake and having larger broods (Magurran and Seghers 1994b). Therefore, females devote more time to foraging to satisfy their energy needs. They are also more cautious in the presence of predators and have lower mortality than males in nature (Rodd and Reznick 1997). Female guppies give birth to broods of live young but are considered lecithotrophic species where the maturing oocyte stores all nutrients for maternal provisioning before fertilization (Thibault and Schultz 1978; Turner 1940). The developing embryo receives nourishment from the fully provisioned yolk and there is no placenta-like exchange of nutrients (Constantz 1989; Reznick and Yang 1993).

Guppy males and females exhibit this tremendous sexual dimorphism potentially as a direct or indirect consequence of gender-inequalities in their mating system and ecological habitats. Among teleosts, guppies are possibly the only species with such well-characterized

dimorphism associated with rapid evolution of potentially sexually antagonistic traits under varying ecological pressures. Considerable research interest has revolved around the linked loci for sex-determination and male-ornaments on the short male-specific Y region of the guppy's largely pseudo-autosomal sex chromosome (Breden and Lindhom 2011) . While the evolutionary ecology of the guppy and its sexual dimorphism has been studied with respect to heredity and adaptation, the molecular mechanisms governing this dimorphism have not yet been identified.

1.3.2 Available molecular resources

At the beginning of my research, the transcriptomic resources available for the guppy were a library of Sanger-sequenced expressed sequence tags (ESTs) roughly corresponding to 9,000 unique genes from mixed populations in Trinidad and Venezuela (Dreyer, et al. 2007). The available genomic resources were randomly sequenced Bacterial Artificial Chromosome (BAC) end sequences from a population from Cumana River in Venezuela. Using single nucleotide polymorphisms (SNPs) developed from these resources a detailed linkage map of the guppy was generated integrating mapping crosses between the Quare and Cumana guppies (Tripathi, et al. 2009c). Several quantitative trait loci (QTL) influencing male size, shape and colour traits were mapped to sex-linked and autosomal linkage groups (LGs) and a linkage map comprising 23 LGs had been described. Sex as a trait was mapped to the distal end of LG12, but no male-limited markers could be identified (Tripathi, et al. 2009a).

During the course of my research work a 454 sequenced transcriptome was assembled using cDNA from several guppy populations (Fraser, et al. 2011). Currently our lab has also assembled a draft assembly for the guppy genome using genomic DNA from an inbred female and male from the Guanapo population in Trinidad (Künstner *et al.* submitted, GenBank ID GCA_000633615.2). For my research, I used the draft genome of the female for genome-guided assembly. The EST and 454-equenced transcriptomes were mainly used for quality assessment. Genomic resources and protein annotations for *Danio rerio* (zebrafish), *Gasterosteus aculeatus* (stickleback), *Oryza latipes* (medaka), *Oreochromis niloticus* (tilapia), *Takifugu rubripes* (fugu), *Tetraodon nigriviridus* (tetraodon), *Gadus morhua* (cod), have been available since the start of my research. However, the recent major update in the zebrafish genome and annotations (Howe, et al. 2013) had considerably more gene models and annotations and the latest assembly was therefore incorporated. The annotated genome of the closely-related platyfish, *Xiphophorus maculatus*, was officially released in early 2013 (Schartl, et al. 2013). The availability of annotated sequences from the platyfish was useful

for homology-based validation and annotations.

1.4 Outline of the thesis

In Chapter 2, I describe the comparison of various transcriptome assembly tools. The transcript output from different assemblers is compared using length-based, mapping-based, and annotation-based metrics. The ultimate objective was to select the appropriate assembly that is a comprehensive representation of full-length transcripts from the complex cDNA pool from guppy tissues.

In Chapter 3, I present the compilation and annotation of a reference transcriptome. The combined reference transcriptome was used to study the differential expression between sexes in three sexually dimorphic tissues from adults. I further relate the tissue-associated sex-biased gene-expression to the phenotypic dimorphism of the guppy. I also present the identification and expression pattern of candidate genes orthologous to pigmentation and sex-determination related genes in other species.

In Chapter 4, I address the evolutionary aspect of sex-biased gene expression in the guppy. I first explored the genomic distribution of sex-biased genes to identify non-random distributions that suggest sex-specific selection pressures at certain genomic regions. In the second part of this chapter I compared the rates of coding sequence change in sex-biased genes versus unbiased genes to look for signatures of accelerated sequence evolution, as hypothesized for genes under sexual selection.

In Chapter 5, I summarize all the results and discuss them in light of our current knowledge and touch upon future aspects of the research.

Chapter 2: Comparison of reference-based and reference-independent transcriptome assemblers

2.1 Introduction

The choice of a suitable strategy and quality metrics for transcriptome assembly is largely dependent on the study organism and research objective. Till now there is no gold standard or standardized protocol for transcriptome assembly from organisms with limited molecular resources. The primary goal of my work was to obtain a set of predominantly full-length transcripts that represent a majority of the guppy's protein coding genome at any given time. With this purpose, I chose to extract RNA from multiple adult tissues and an embryonic developmental. I used cDNA prepared from a single population of the guppy to reduce population associated polymorphisms in the data for short-read assembly. Assembly of vertebrate transcriptomes is complicated due to a large number of splice-variants. In addition, teleost genomes have duplicated chromosomes and gene expansions and losses that arose due to divergence after a teleost specific whole genome duplication (Brunet 2006). Therefore, genes with high sequence similarity may be difficult to assemble and annotate uniquely. Comparisons with genomes of closely related species are expected to be the most informative in this regard.

The landscape of non-model genomics and transcriptomics has rapidly evolved in the last 5 years. Although the evolution in sequencing methods since 2010 does not affect this analysis, the development of new software and improvements in existing tools for assembly and downstream analysis has reduced computational times and enhanced read utilization. I have tried to keep my research updated with the latest assembly tools and with availability of published and personally communicated molecular data. For background information for the reader I will briefly explain the various assemblies that I performed, in the context of time line of assembler development. The first set of RNA-seq reads was assembled *de novo* only with OASES, which was the first transcriptome assembly tool for Illumina reads. TRINITY *de novo* assembler was released in the beginning of 2011 and the assembly was performed several times since then in order to optimize the strategy. Genome-guided TRINITY was developed by the end of 2012. All genome-guided assemblies were performed after the assembly of the female draft genome was considered ready for release towards the end of 2012. The *de novo*

assemblies were subsequently repeated in early 2013, using all the data and latest versions of assemblers to ensure uniform input and up-to-date comparisons.

2.2 Materials and methods

2.2.1 Fish strains, husbandry and dissection

All tissues were prepared from laboratory-reared guppies that were descendants of wild female fish caught in 2003 from Upper Quare river, East Trinidad (Tripathi, et al. 2009c; Willing, et al. 2010). The fish were reared at 25°C in a 12-hour light and dark cycle in uniform conditions of food and water for a population size of 5-6 individuals per 1.5 l tank. Female organs were prepared from virgin adult fish that were separated from males at the age of 3-4 weeks to avoid premature insemination and sperm storage in ovaries. Whole embryos were isolated from gravid females that had been reared with males and had given birth to 1-2 broods. Mature adult guppies aged 5-6 months, were isolated and kept in clean freshwater tanks in fungicide treated water for 44-48 hours prior to dissections. The fish were not fed during fungicide treatment to allow for clearing of the gut in order to minimize bacterial contamination of the sample. Fish were anaesthetized in 0.1% neutralized MESAB and washed in ice-cold PBS before dissection. Brain, eyes, liver, spleen, skin, tail and gonads were isolated from adult males and females grouping brain with eyes and liver with spleen during isolation. For the embryonic tissue, whole embryos at late-eyed to very late-eyed stages stage of development were isolated from gravid females (Martyn, et al. 2006). A small fin-clip was taken from each embryo and stored in 95% ethanol for genotyping sex. All samples were washed with ice-cold PBS, frozen in liquid nitrogen and stored at -80°C till RNA isolation.

2.2.2 Library preparation and Illumina sequencing

Non-barcoded libraries: Four Illumina cDNA libraries were independently prepared from guppy females and males using i) late-eyed stage embryos and; ii) adult tissue pool comprising total RNA from brain, eyes, liver, spleen, skin, tail, and gonad. Embryos were first genotyped using genomic DNA isolated from fin-clips with markers 229 and 230 with sex-specific single nucleotide polymorphisms (SNPs) in Quare population (Tripathi, et al. 2009a). All tissue samples were homogenized in TRIzol® reagent (Invitrogen) using a Polytron® homogenizer (PT 1200, Kinematica AG, Switzerland). Total RNA was extracted from the Trizol homogenate according to manufacturer's instructions. After removal of contaminant

DNA, using DNaseI (Invitrogen), purified RNA was quality-checked and quantified (Nanodrop ND-2000, ThermoScientific peqlab®). For libraries from male and female adults, 75µg RNA starting material was prepared by pooling 15µg total RNA isolated from each tissue. For libraries from male and female embryos, 75µg total RNA was isolated from 15 individual embryos of each sex. Subsequently, purified polyA+ mRNA (Dynabeads® Oligo(dT), Invitrogen) was used for preparation of paired-end RNA libraries with insert-size of 200-300bp, using the mRNA-seq Sample Preparation Kit (Illumina, San Diego, CA) or the NEBNext® mRNA Library Prep Reagent Set for Illumina (NEB), according to manufacturer's instructions. Library quality and concentration were assessed using the Agilent DNA 1000 Bioanalyser assay (Agilent Technologies, Germany). Each library was sequenced on a separate GAIIx lane (Illumina, San Diego, CA, read length 101bp). Hereafter, I will refer to these four datasets as female and male adult (F_{adult}, M_{adult}) and female and male embryo (F_{embryo}, M_{embryo}).

Barcoded libraries: I prepared barcoded cDNA libraries for quantitative analysis of gene-expression differences. The following tissues were isolated from adult male and female guppies: brain and eyes, tail (containing skin, muscle, bone and cartilage), and gonads (ovaries from virgin females or testes from males). All tissues were individually homogenized in TRIzol Reagent (Invitrogen, Carlsbad, CA, USA). Homogenization was done using steel beads in plastic tubes for tissue disruption by high-speed shaking(Qiagen TissueLyserII with Qiagen TissueLyser Adapter Set 2 x 96). Total RNA was extracted from the TRIzol homogenate using DirectZol RNA extraction kits with in-column DNaseI treatment. Purified total RNA was quality-checked on agarose gels and quantified using the Qubit RNA Assay Kit (Invitrogen, Carlsbad, CA, USA). I prepared libraries from six biological replicates for each tissue and sex type, except the female brain. For brain sample from females, I prepared libraries from 7 biological replicates and included two technical replicates. All samples were randomized and individually barcoded during library preparation using TruSeq mRNA-seq Sample Prep Kit (Illumina, San Diego, CA, mRNA-seq Sample Prep Manual v2 protocol). In total 39 paired-end libraries were prepared. Libraries were pooled with 13 libraries per pool and sequenced on 3 lanes of the HiSeq™ 2000 (Illumina, San Diego, CA, read length 101bp). I will refer to these barcoded cDNA libraries from adult tissues as: Female brain (F_{brain}), Male brain (M_{brain}), Female tail (F_{tail}), Male tail (M_{tail}), Female gonad (F_{gonad}), and Male gonad (M_{gonad}). Table 2.1 summarizes the samples used to prepare the libraries and the number of reads obtained in each set.

2.2.3 Quality filtration and read trimming

I filtered the resulting reads in the non-barcoded datasets using the following tools; i) low complexity reads were removed with SHORE v0.6 (Schneeberger, et al. 2009); ii) PCR duplicates were removed with an in-house script for matching 60 bp of both reads of a pair, keeping unique pairs and 3 potential duplicates with highest quality scores; iii) Homopolymer sequences (polyA/T/G/C) over 22 bp length were trimmed using CUTADAPT V1.2.1(MARTIN 2011); iv) low-quality nucleotides were trimmed using CONDETRI v2.2 (Smeds and Kunstner 2011) with cut-offs of phred20 quality, 35 bp length and all other default parameters. In the barcoded datasets, I only removed PCR duplicates and quality filtered the reads to reduce interference in count-normalization and expression quantifications (Table 2.1).

Table 2.1: Description of RNA samples, Illumina cDNA libraries and sequenced datasets. I describe the number of individuals, the organ composition, total RNA, library preparation and sequencing protocol, and number of filtered sequenced reads in each dataset. Each barcoded library (*) represents a single tissue from an individual guppy. In female brains (**), data from all 9 samples (7 biological replicates and 2 technical replicates) were used for assembly but only 6 biological replicates were used for differential expression analysis.

Library (No. of individuals)	**Organ(s)**	**Amount (μg)**	**Library preparation protocol**	**No. of read pairs after phred20 filtering**	**Dataset**
Female adult: F_{adult} (9)	Brain, Eyes, Liver, Skin, Tail, Ovaries	15 μg of total RNA from each organ pooled ~75μg used for polyA+ purification	Paired End RNA library prepared with NEB RNA kit for Illumina (Each library sequenced separately on a single lane of Illumina GAII)	26,393,787	Non-barcoded
Male adult: M_{adult} (9)	Brain, Eyes, Liver, Skin, Tail, Testes			29,481,947	
Female embryo: F_{embryo}(15)	Fin-clipped embryos	75μg of total RNA used for polyA+ purification		24,138,679	
Male embryo: M_{embryo}(15)	Fin-clipped embryos			18,775,577	
Total read pairs :				98,789,990	
Female brain: F_{Brain} (9,6**)	Brain and eyes	3 μg of total RNA each	Paired End RNA library prepared with Illumina TruSeq RNA kit : Each organ individually barcoded (3 x 13 libraries multiplexed and sequenced on 3 lanes of Illumina HiSeq)	111,941,790, 79,016,273**	Barcoded*
Male brain: M_{Brain} (6)	Brain and eyes			70,950,871	
Female tail: F_{Tail} (6)	Tail (Muscle, Skin)	2 μg of total RNA each		75,180,682	
Male tail: M_{Tail} (6)	Tail (Muscle, Skin)			58,020,495	

Female gonad: F_{Gonad} (6)	Ovaries	1 µg of total RNA each		53,602,790	
Male gonad: M_{Gonad} (6)	Testes			53,065,339	
Total read pairs :				422,761,967	

2.2.4 Transcriptome Assembly

Genome independent assemblies

I assembled the reads using two *de novo* assemblers, TRINITY (trinityrnaseq_r2012-06-08) (Grabherr, et al. 2011b) and VELVET: V1.2.03–OASES: V0.2.06 (Schulz, et al. 2012; Zerbino and Birney 2008b). Both these assemblers are designed for genome independent assembly of short read RNA-seq data using *de Bruijn* graphs constructed from a series of overlapping *k*-mers. First, two strategies of data pooling were compared for assembly of maximum unique transcripts from available *k*-mer coverage. This analysis was done using only the non-barcoded datasets (F_{adult}, $M_{adult,}$ F_{embryo}, M_{embryo}) and TRINITY (trinityrnaseq_r2011-11-26). In the first pooling strategy, I performed individual assemblies and subsequently clustered the resulting transcripts using CD-HIT-EST v4.6 with default parameters (Fu, et al. 2012; Li and Godzik 2006). In the second strategy, I began with pooled reads and performed a single assembly followed by clustering of resulting transcripts. As I obtained more unique transcripts with better coverage of the EST dataset using the second strategy (Appendix: Table A2.1), I chose it for subsequent *de novo* assembly.

A) TRINITY: The high coverage datasets were first individually normalized for *k*-mer coverage using the TRINITY package associated script for *in silico* read normalization, *normalize_by_kmer_coverage.pl,* with default parameters. The normalized reads (*k*-mer: 25, Coverage: 30) from all datasets were pooled (F_{adult} + M_{adult} + F_{embryo} + M_{embryo} + F_{brain} + M_{brain} + F_{tail}+ M_{tail}+ F_{gonad}+ M_{gonad}). This dataset was half the initial size amounting to a total of nearly 258,000,000 read pairs. These reads were *de-novo* assembled with TRINITY using *k*-mer coverage 2, minimum length 200bp and other default parameters. I refer to this transcriptome assembly as Trinity.

B) VELVET-OASES: *De novo* assembly with VELVET-OASES failed due to memory limitations when the above-mentioned pooled TRINITY normalized read dataset was used. Therefore, to reduce the read data further I performed an additional single pass *in silico* read normalization to *k*-mer coverage 20, size 19bp, using DIGINORM (Brown, et al. 2012). Digital normalization with DIGINORM substantially reduced the dataset to

30,000,000 read pairs. Transcripts were assembled from normalized read data using single *k*-mers (21, 25, 27, 31, 35). The OASES single-*k* assemblies are referred to as Oases_k21, Oases_k25, Oases_k27, Oases_k31, Oases_k35 or in some figures as k21,k25,k27,k31,k35 based on space constraints. All *k*-mer assemblies were merged using *k*-mer 27, coverage cut-off 5 and transcript length 200bp to have a multiple-*k* assembly (Oases_merge). Multiple *k*-mer assemblies were also merged by pooling all assemblies and clustering transcripts with 90% identity using CD-HIT-EST. The clustered assembly is referred to as Oases_clust.

Genome-guided assemblies

We performed genome-guided assemblies using the draft version of the female guppy genome (Künstner *et al.* submitted, GenBank ID GCA_000633615.2) with TOPHAT – CUFFLINKS – CUFFMERGE v2.0.4 transcriptome assembly pipeline (Trapnell, et al. 2012; Trapnell, et al. 2010); and genome-guided TRINITY r2012-10-05 transcriptome assembly pipeline (http://www.vcru.wisc.edu/simonlab/bioinformatics/programs/trinity/docs/genome_guided_trinity.html, last accessed 17.09.2014). Genome guided assemblers infer exon-intron junctions through alignment of spliced reads on the genome; therefore, to infer appropriate tissue-specific or development-stage specific splicing (Merkin, et al. 2012) we performed dataset specific assemblies and merged the final assemblies.

A) **TOPHAT-CUFFLINKS-CUFFMERGE:** Reads from each RNA-seq sample were first individually mapped to the reference genome using TOPHAT2 (Kim, et al. 2013), a BOWTIE2 v2.0.4 (Langmead and Salzberg 2012) based aligner that allows spliced alignments of reads. CUFFLINKS used the resulting alignments to generate a transcriptome assembly for each dataset (F_{adult}, M_{adult}, F_{embryo}, M_{embryo}, F_{brain}, M_{brain}, F_{tail}, M_{tail}, F_{gonad} and M_{gonad}). These assemblies were then merged together to give a combined assembly with CUFFMERGE. I refer to this assembly as Cufflinks_GG.

B) **Genome-guided TRINITY:** Reads from each RNA-seq sample were first normalized using *in silico* read normalization with TRINITY (identical to genome-independent assembly with TRINITY). Normalized read datasets were individually assembled using the genome-guided TRINITY assembly pipeline. Reads were first mapped to the reference genome using GSNAP v2012-07-20 (Wu and Nacu 2010). The mapped reads were partitioned into read-covered regions of the genome and reads in each partition were *de novo* assembled with TRINITY. The TRINITY assembled transcripts from each dataset were combined and clustered with CD-HIT-EST (default parameters) to remove

redundant transcripts. I refer to this assembly as Trinity_GG.

C) **Genome-guided PASA[1]:** The Trinity_GG transcripts were used for gene-model prediction using the PASA (PASA_r2012-06-25) assembly pipeline (Haas, et al. 2003; Haas, et al. 2008). To reduce redundant input sequences, transcripts with 80% sequence identity were further clustered using USEARCH v6.0.307 implemented tool *cluster_fast* (Edgar 2010) followed by CD-HIT-EST (sequence identity threshold 0.8). The set of clustered transcripts were first cleaned with SEQCLEAN (PASA_r2012-06-25), then aligned back to the genome using GMAP v2012-07-20 (Wu and Watanabe 2005), and finally assembled into full-length transcript structures with PASA (maximum intron length of 100,000 bp). I refer to this assembly as Pasa_GG.

D) **EVM[1]:** A fourth approach for reference-guided gene-set prediction combined *ab initio* gene prediction, *de novo* assembled RNA-seq transcripts, and an orthology-based approach. For *ab initio* prediction, the AUGUSTUS web-server (Stanke, et al. 2008) was first trained with 4,434 guppy EST sequences (randomly chosen from NCBI (Pruitt, et al. 2014)). Then, *ab initio* prediction was performed and a total of 33,527 gene models were predicted. For the RNA-seq approach gene models assembled with the Trinity-PASA pipeline (Pasa_GG) were used. For the orthology-based approach protein sequences from *Danio rerio* (zebrafish), *Gasterosteus aculeatus*(stickleback), *Gadus morhua* (Cod), and *Oryzias latipes* (medaka) were downloaded from ENSEMBL (Release 70). For each gene, we extracted the longest amino acid sequence and combined all sequences (86,176) to cluster the sequences using CD-HIT version 4.6.1 with sequence identity threshold set to 0.7. This resulted in 48,803 protein sequence clusters. Next, the clusters were blasted (TBLASTP version 2.2.27+, maximum intron length 100,000 bp, e-value < 1 $x10^{-5}$) against the draft genome sequence. All protein clusters with confident hits (45,526 sequences) were aligned against the guppy draft assembly using EXONERATE version 2.2.0 (Slater and Birney 2005) with the *protein2genome* model, at least 60% match to the genome and maximum intron length of 200,000 bp. This approach resulted in 10,627 orthologous gene models. Gene models from all three approaches were combined to build the guppy reference gene set. We used EVIDENCEMODELER version r2012-06-25 (Haas, et al. 2008) and put different weights on the prediction methods (*ab initio* 4, protein 5, transcript 10). EVIDENCEMODELER was run on 1 Mb genome segments with 100 kb overlap to

[1] These assemblies were done by a post-doctoral researcher, Axel Künstner (Künstner et. al. in preparation).

reduce computational burden. The resulting reference gene set contained 31,902 protein-coding sequences.

Longest transcribed isoform (LTI) Assembly: Since exact splice variant prediction requires more elaborate algorithms and was not the focus of our study we used only the longest isoform for each locus (OASES), component (TRINITY) or gene group (CUFFLINKS) for further analysis. We describe and compare the reduced assemblies comprising only the transcripts that remain on keeping the longest transcribed isoforms in the Trinity, Oases single-*k*, Oases_merge and Cufflinks assemblies.

Clustered assemblies: For Oases_clust and Trinity_GG assemblies, transcripts from several independent assemblies were combined and clustered with CD-HIT-EST. I evaluate only the reduced assemblies that remain on keeping the longest transcribed sequence in each cluster.

ORF prediction with TRANSDECODER: Open reading frames (ORF) were predicted for the Oases, Trinity, Trinity_GG, Cufflinks and Pasa_GG transcripts from the reduced assemblies using the program TRANSDECODER implemented in *transcripts_to_best_scoring_ORFs.pl*, script in Trinity pipeline. TRANSDECODER annotates coding sequence boundaries using hexa-nucleotide frequencies learnt from a first pass on the data to calculate likelihood scores for predicted sequences, similar to GENEID (Blanco and Abril 2009; Blanco, et al. 2007). Predicted coding sequences (CDSs) were further clustered to remove sequences with 90% redundancy using CD-HIT-EST.

2.2.5 Database of guppy EST and 454 sequences

Sequences from the Sanger-sequenced EST database were combined with the 454 sequenced transcriptome assembly. The sequences were clustered using CD-HIT-EST with default parameters to remove sequences with greater than 90% redundancy. This set of 58,418 sequences is referred to as Guppy_454EST.

2.2.6 Identification of orthologous proteins in other teleosts

Orthologous genes in other vertebrate species were identified using translated CDS for the genome-guided and genome-independent assemblies. Peptide sequences for *Danio rerio* (zebrafish), *Gasterosteus aculeatus* (stickleback), *Oryza latipes* (medaka), *Xiphophorus maculatus* (platyfish), *Oreochromis niloticus* (tilapia), *Takifugu rubripes* (fugu), *Tetraodon nigriviridus* (tetraodon), *Gadus morhua* (cod), *Homo sapiens* (human), and *Mus musculus* (mouse) were downloaded from ENSEMBL (Release 71). Single-copy (1:1) orthologs were

identified using PROTEINORTHO v4.26 (Lechner, et al. 2011) (Parameters: BLASTP v2.2.21, e-value < $1x10^{-10}$, alignment connectivity: 0.8, coverage: 40%, identity: 30%, adaptive similarity: 0.95, including pairs: 1). PROTEINORTHO evaluates the pairwise reciprocal blast results for best scoring hits that match user-defined filtration criteria and transforms the results into a graph. Highly connected graph components represent closely related proteins and are reported as orthologs.

2.2.7 Alignment against Platyfish genome

The quality of each assembly was further assessed by comparing the transcript coverage and exon recovery using the genome of platyfish, a closely related poeciliid fish. Comparisons were performed using the last stable release of the *X.maculatus* genome, Xipmac4.4.2. Cross-species gene structures were predicted by aligning the transcripts using GMAP v2012-07-20 (Wu and Watanabe 2005) against the repeat-masked toplevel genome downloaded from ENSEMBL, Release 74 (Flicek, et al. 2013). Predicted mRNA features in each gene feature file (GTF format) were parsed for coverage of the aligned sequence in the platyfish genome after removing sequences with alignment identity less than 75%. Annotated exon features from each GTF file were compared against exon annotations in the GTF reference for platyfish (ENSEMBL Release 74). Exon coverage was calculated using *coverageBed*, BEDTOOLS version 2.16.2 (Quinlan and Hall 2010). Exon recovery was calculated from the fractions of total protein coding exons that were completely assembled and annotated (100% covered with a single feature) and completely missing from the annotation (0% covered).

2.2.8 Read alignment and calculation of FPKM

Paired-end reads from digitally normalized read dataset were aligned to each transcriptome assembly using BOWTIE2 v2.0.4 (default parameters for sensitive local alignment). Mapped reads were counted using EXPRESS v1.3.1 (Roberts and Pachter 2013) and Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) were calculated for transcripts in each assembly.

2.2.9 Transcript reconstruction of *X. maculatus titin* homolog

The full-length reconstruction of a single transcript from the longest vertebrate gene, *titin*, was studied in more detail. Using the nucleotide sequence of *X.maculatus titin*-like mRNA (RefSeq Accession: XM_005798187.1, GI: 551493350), I identified reciprocal best-blast hit

orthologs in each assembly. The orthologous sequences were also confirmed by matching the identity of translated nucleotides with protein orthologs obtained in the analysis with PROTEINORTHO.

2.3 Results and discussion

The genome-guided and genome-independent assemblies were evaluated using length-based, mapping-based, and annotation-based metrics: i) the total length of assembly and mean length of assembled transcripts; ii) number of full-length predicted open reading frames (ORFs); iii) number of reads used (completeness); iv) number of correctly oriented read pairs (accuracy); v) extent of redundancy; vi) number of orthologs identified using reciprocal blast against other validated sequence databases as reference; vii) exon coverage using gene models from a closely related species; viii) transcript coverage using genome of a closely related species; and ix) full-length reconstruction of transcript from the longest vertebrate gene.

2.3.1 Length based metrics for transcriptome comparisons

I compared the numbers of assembled transcripts, average length of assembled transcripts, and N50 statistics for the six assemblers. The total number of assembled fragments varied across assemblers but was always greater than number of predicted gene models in EVM assembly (Figure 2.1). The number of transcripts and predicted CDS were highest for the clustered transcriptomes, Trinity_GG and Oases_clust.

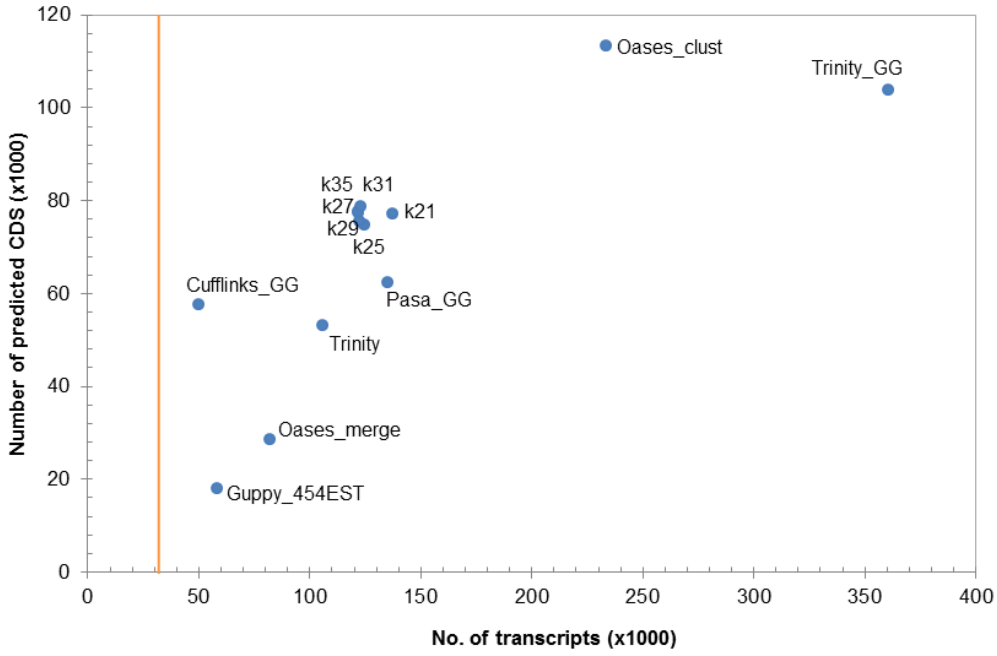

Figure 2.1: Count-based comparison of transcriptome assemblies. Total number of assembled transcripts against number of predicted CDS in each assembly and the Guppy_454EST database. Orange vertical line shows the number of gene models predicted by EVM assembly. All assembler names are shown next to the respective data-points and the Oases single *k*-mer assemblies are represented as k21-k35.

Further comparisons of total length of assembly against average length (Figures 2.2: A) or half-length of assembly against N50 length (Figures 2.2: C) show that Cufflinks_GG assembly contains the longest transcripts. The overall transcriptome assembly sizes of Cufflinks_GG, Pasa_GG and Oases single-*k* assemblies were comparable to each other being smaller than the assembly size calculated from the EVM gene models. Eukaryotic transcriptome assemblies should be smaller than the assembled size of associated gene models as introns may be spliced out by co-transcriptional or post-transcriptional splicing in the poly-adenylated mRNA (Bentley 2014). Both Oases_clust and Trinity_GG had an unusually large total assembly size suggesting that redundant sequences may still be retained post clustering. The small assembly size and average length, N50 length of the Oases_merge assembly suggests that there is a significant loss of sequence information in the merge step.

Comparison of transcript lengths (Figure 2.2) shows more about the quality of assembled data in comparison to assembly size. As transcript length may be different from length of putative coding sequences, I also compare the average and N50 length of predicted CDS against total length of CDS assemblies including the EVM predicted CDS as a benchmark (Figures 2.2: B, D). The average length and N50 length of Trinity_GG and Cufflinks_GG were highest, with genome-independent Trinity a close second. Overall the predicted CDS from Trinity and Cufflinks_GG were of comparable total size and N50 length as the EVM assembly. Though, Oases single-*k* and Oases_clust assemblies had longer transcripts than Pasa_GG and Trinity, the predicted CDS from Oases assemblies were shorter. Similar to transcriptome sizes, the total length of Trinity_GG and Oases_clust were much greater than EVM.

The length-based comparisons show that Cufflinks_GG assembles most contiguous transcripts and is comparable to the EVM assembly. A comparison of the other genome-guided assemblies, Trinity_GG and its subsequent assembly Pasa_GG, shows that Pasa_GG is able to reduce the redundant information in Trinity_GG. However, assembly with PASA also reduces the contiguity of transcripts and CDS. Among *de novo* assemblers, OASES assembles the longest transcripts with a slight increase in contiguity with increasing *k*-mer length. Clustering multiple *k*-mer assemblies resulted in retention of longer transcripts but did not remove all redundant transcripts, while combining multiple-*k* assemblies with OASES merge resulted in a considerably short assembly. *De novo* assembly using TRINITY resulted in shorter transcripts than the other assemblers but the lengths of predicted CDS were comparable to the

Cufflinks_GG assembly. This suggests that the other assemblies contain a lot of non-coding sequence, while Trinity transcripts may have shorter putative UTR's but retain coding sequence information. The Guppy_454EST dataset was much smaller and shorter than all other assemblies showing the markedly greater complexity in the assemblies from Illumina data.

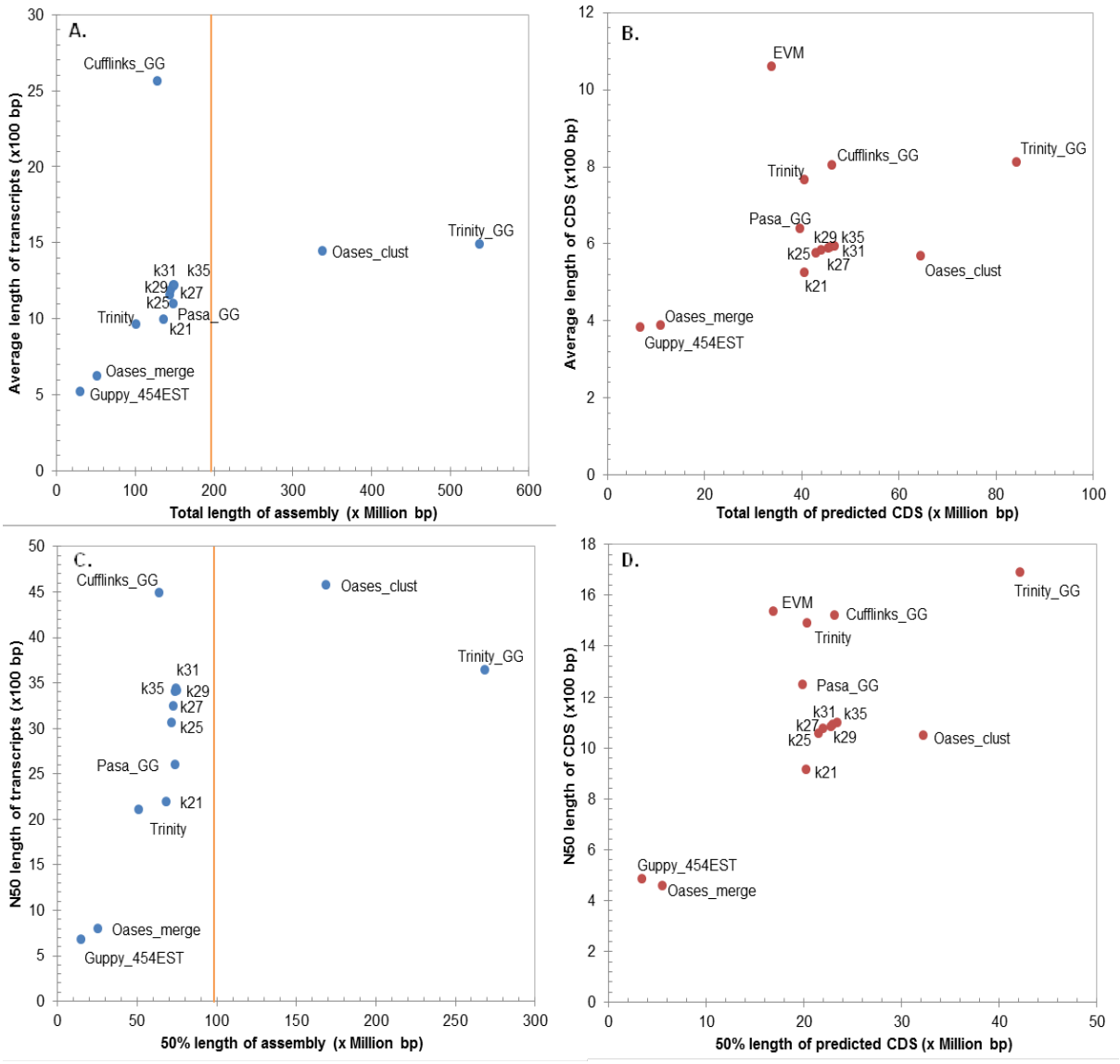

Figure 2.2: Length-based comparison of transcriptome assemblies. Scatter plots showing length based comparisons of assembled transcripts (Blue dots) or predicted CDS (Red dots). The average lengths (A, B) or N50 lengths (C, D) are plotted on Y-axis against the total sum of lengths (A, B) or half of the total sum of lengths (C, D) on X-axis. Vertical orange lines in figures A and C shows the sum total of lengths of predicted gene models in EVM assembly. All assembler names are shown next to the respective data-points and the Oases single *k*-mer assemblies are represented as k21-k35.

2.3.2 Mapping based metrics for transcriptome comparisons

The total number of reads incorporated in each assembly and the orientation of read pairs were compared as measures of completeness and accuracy. Read pairs aligning in correct orientation (concordantly) and to a single assembled region (unique) indicate assembly of long non-redundant transcripts. On the other hand multiple concordant alignments indicate the presence of transcripts with similar sequence (possible redundancy), while discordant and unpaired alignments reflect the fragmented or mis-assembled transcripts.

Expectedly, the maximum number of reads aligned to the longest assemblies, Trinity_GG and Oases_clust (Figure 2.3). However, these assemblies also had the largest number of multiply aligned reads. All Oases single-*k* assemblies had comparable number of overall and unique alignments but there was a gradual increase in multiple-alignments with increase in *k*-mer length. The Oases_merge had very low number of aligned reads. This shows that there is a clear loss of sequence information in the merge step. Re-assembly of Trinity_GG transcripts using PASA (Pasa_GG) reduced multiple mappings indicating the retention of unique transcripts. However, this step also reduces the overall alignment rates. The overall alignments of Trinity and Cufflinks_GG were comparable but Trinity had a much lower percentage of multiple alignments. The total percentage of reads that aligned to EVM cDNA were less than all other assemblies (except Oases_merge), perhaps reflecting the difference in sequence content of protein coding sequences versus full-length transcripts reconstructed from RNA-seq.

Comparison of density distribution of normalized expression (FPKM) by transcript length further shows the differential results obtained by each of the reconstruction methods (Figure 2.4). While every assembly method produced a number of long transcripts with non-zero expression, we observed that only EVM and Cufflinks_GG assemblies had major proportion of total transcripts in this area (dark-spot in middle-right). However, Cufflinks_GG also assembles a large number of transcripts that are longer than 1000 bp and have low expression ($\log_2$FPKM ~ 0).The Oases single-*k* assemblies (except *k*-mer 25) and Trinity_GG had a large proportion of non-expressed transcripts over varying lengths ($\log_2$FPKM ~ 10). These may correspond to assembly artifacts and redundant transcripts where composing reads show alternate alignments with better scores. Comparison of Pasa_GG against Trinity_GG indicates that this step removes the non-expressed or redundant transcripts shifting the density distribution towards transcripts with non-zero expressions.

In summary, the analysis with mapping-based metrics indicates that TRINITY, among *de novo*

assemblers, and CUFFLINKS, among genome-guided assemblers, incorporate the read data with high accuracy. Both these assemblers effectively reconstruct a large proportion of total transcripts with long lengths and considerable expression. Amongst the remaining assemblies, Oases_k25 and Pasa_GG assemblies also comprise a high proportion of long well-expressed transcripts and very few non-expressed transcripts.

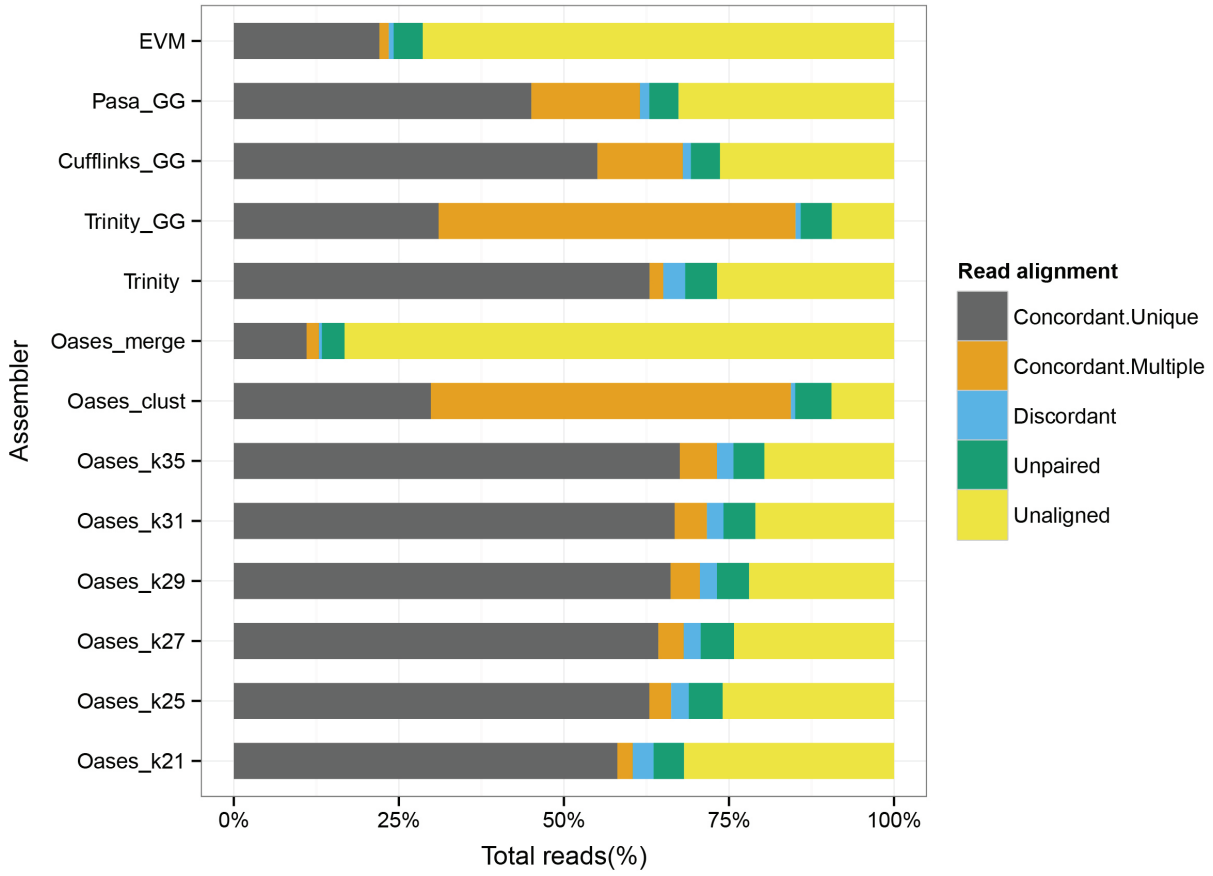

Figure 2.3: Read alignment statistics. Barplots show percentage of total reads mapped to each assembly. Paired reads with concordant and unique alignment (grey), paired reads with concordant and multiple alignment (orange), paired reads with discordant alignment of mates (blue), unpaired singletons with single or multiple alignments (green), unaligned reads (yellow).

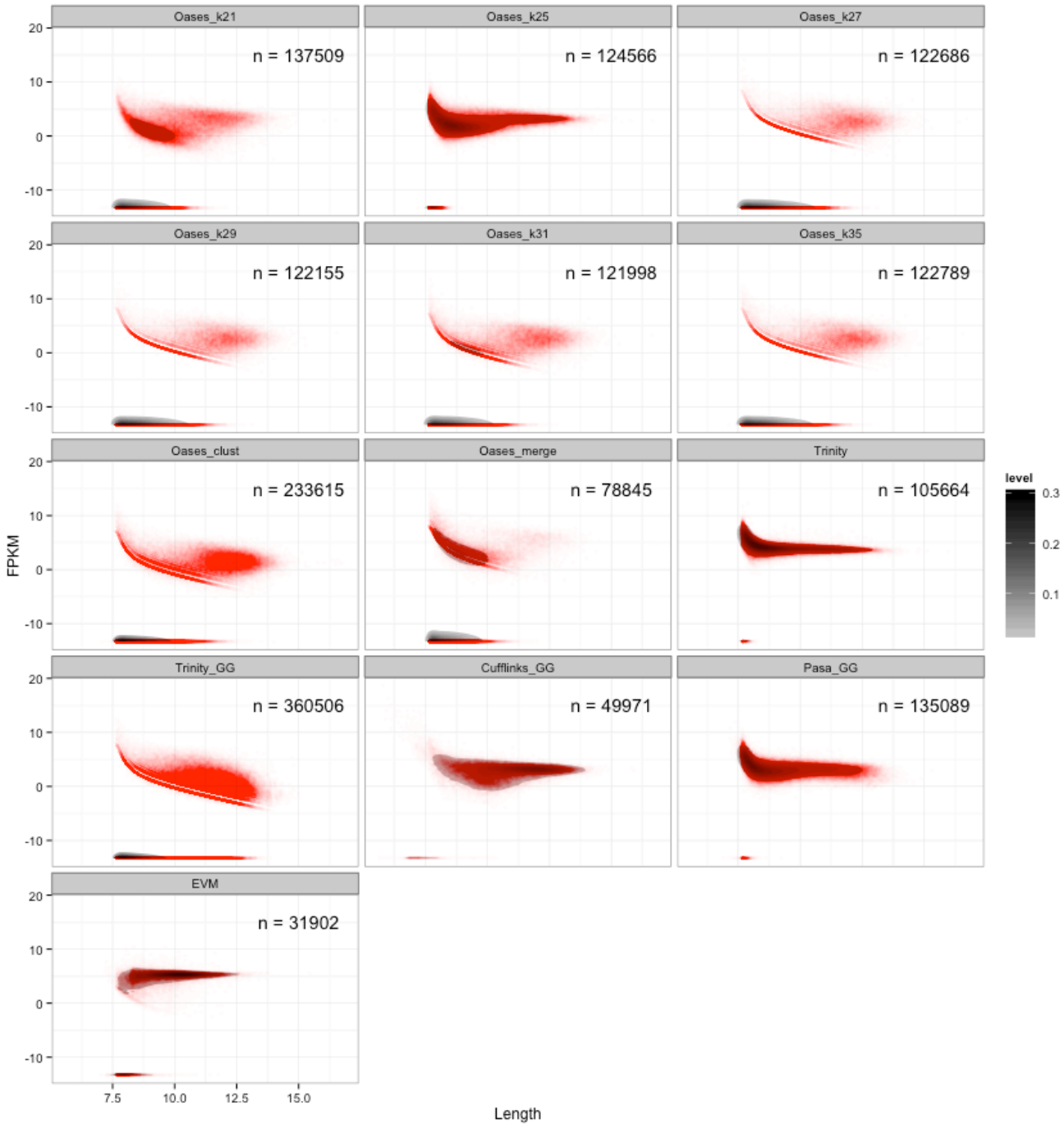

Figure 2.4: Density comparison of transcript length and expression. Scatterplots show the $\log_2$-transformed base pair lengths *vs* FPKM expression of transcripts. Each translucent red dot represents a transcript. Plot region with high transcript density has high intensity of red. Density distribution of points are also plotted using shaded contours in a greyscale (light grey to dark grey) showing the region where most of the data points lie. For each assembly method the total number of transcripts are summarized on top right.

2.3.3 Annotation based metrics for transcriptome comparisons

To enumerate reconstruction of unique protein-coding genes I evaluated the putative protein sequences using validated sequences of other vertebrate species. I checked the total numbers of reciprocal BLASTP orthologs to measure transcriptome completeness. Numbers of unique query sequences with alignments relative to ratio of unique hits per query sequence were evaluated, to measure redundancy as deviation from a 1:1 homology. The longest assemblies, Trinity_GG and Oases_clust, reported the maximum numbers of reciprocal best-BLASTP orthologs (Figure 2.5: A), but they also had the highest number of reported hits per query sequence (Figure 2.5: B). The Oases single-*k* assemblies show an increase in number of putative orthologs with increase in *k*-mer length (Figure 2.5: A). Cufflinks_GG, Trinity and EVM assemblies showed a high number of putative orthologs for the least number of redundant protein sequences (Figure 2.5: A, B). Among these three assemblies, Cufflinks_GG had the greatest number of cross-species and single species orthologs and the ratio of unique hits per query sequence was similar to EVM. Overall EVM was the most unique assembly. The greater number of unique orthologs identified in Oases_clust over any Oases single-*k* assemblies suggests that each *k*-mer assembly has subsets of unique sequences that have valid protein orthologs. This highlights the advantage of a multiple *k*-mer strategy for the assembly of maximum number of orthologs.

2.3.4 Assembly of *X. maculatus* protein coding exons

I assessed the assembly of protein-coding exons in each method, using annotations from the closely related genome of platyfish. I aligned transcripts against the *X. maculatus* genome to predict gene features. Comparing the reconstructed features with the reference annotations, I evaluated the percentage of total reference protein coding exons that were fully assembled or completely missing in each transcript assembly (Figure 2.6). Genome-guided methods, Cufflinks_GG and Trinity_GG, showed the recovery of the greatest percentage of exons. While Cufflinks_GG assembled the highest number of complete exons, Trinity_GG had the lowest number of completely missing exons. Among the *de novo* methods, TRINITY had the least percentage of missing exons than any of the Oases single-*k* assemblies as well as the genome-guided Pasa_GG and EVM assemblies. Once again the clustered multiple *k*-mer assembly, Oases_clust, recovered more exons than other *de-novo* assemblies, showing that certain transcripts/coding regions are better assembled with specific *k*-mer lengths. I did not plot Oases_merge and Guppy_454EST as these methods showed very poor exon recovery.

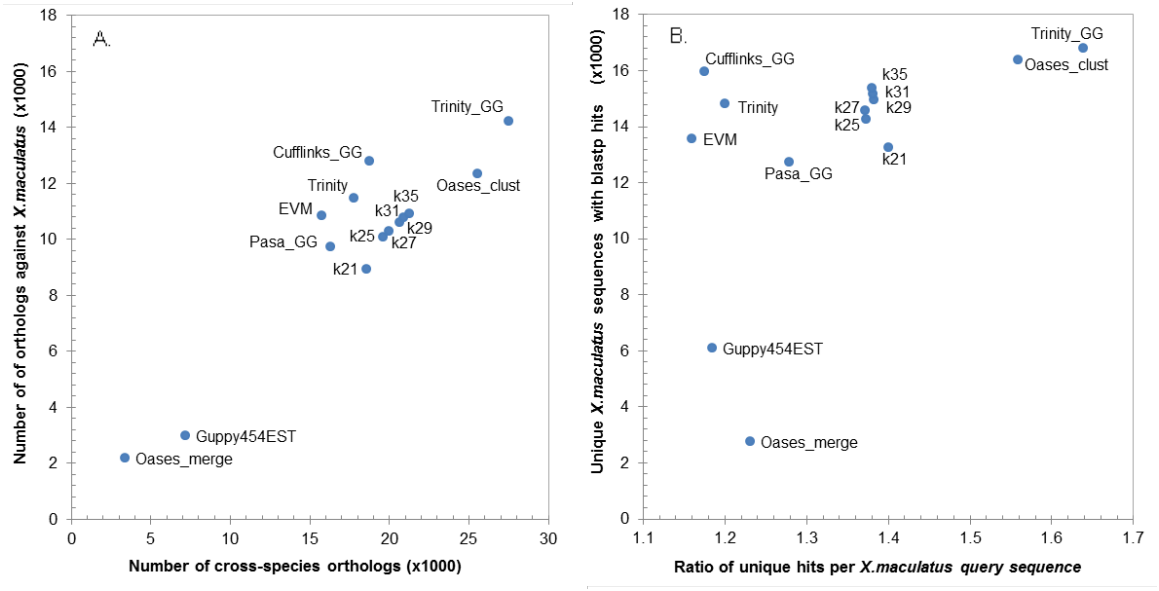

Figure 2.5: Number of orthologs identified to protein coding sequences in other teleosts. The plots show number of orthologs (A) or best-BLASTP hits (B) obtained in each transcriptome assembly. (A) Number of reciprocal blast-hit (RBH) orthologs identified against protein sequences of a single closely related species, platyfish (Y-axis) and total sequences orthologous to protein sequences from teleosts (including platyfish), human or mouse databases (X-axis). (B) The total number of unique platyfish protein sequences that were identified with best BLASTP alignment (e-value < $1x10^{-20}$) plotted with respect to the ratio of unique guppy transcript per platyfish query.

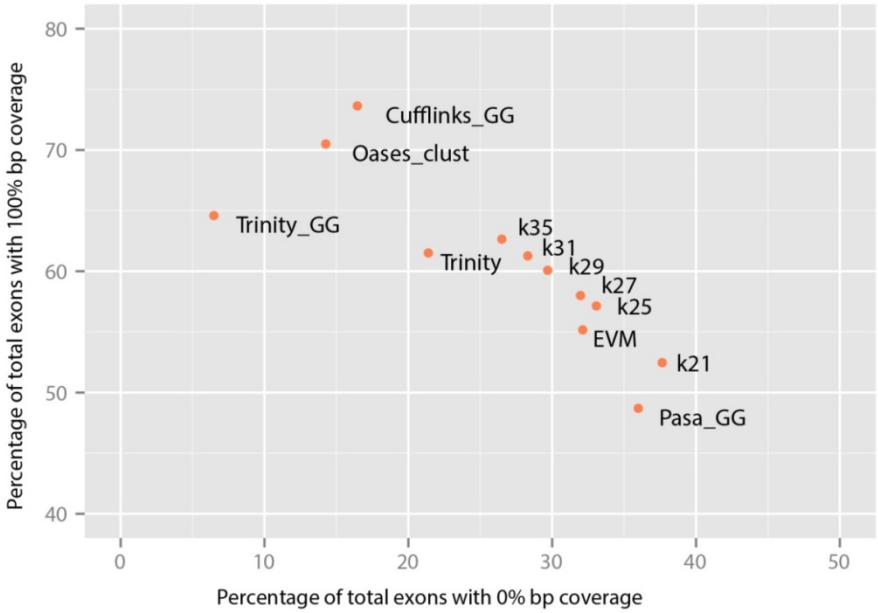

Figure 2.6: Exon recovery comparison across transcriptome assemblies. The plot shows assembly of homologous protein coding exons as the percentage of all coding exons of *X.maculatus* genome. Percentage of reference exons that match predicted exons from the transcriptome assembly with 100% feature coverage (Y-axis) are plotted against percentage of reference exons that are completely missing, 0% coverage in the assembled transcripts.

2.3.5 Trancriptome coverage in *X. maculatus* genome

I further assessed the nucleotide-level coverage of assembled transcripts in the genome of the platyfish. Transcripts in each assembly were aligned to the platyfish genome assembly and analyzed by the proportion of transcript length covered (Figure 2.7). The density distributions, for all assemblers except EVM and CUFFLINKS, showed that a majority of assembled transcripts have either 0-10% coverage, or 75-100% coverage. While the EVM cDNA sequences included very few unaligned transcripts and the greatest proportion of long and full-length alignments (95-100% coverage), Cufflinks_GG assembly differs as it includes long transcripts with intermediate coverage in the platyfish genome. A substantial proportion of Cufflinks_GG transcripts show short (10-25%) to medium (25-75%) coverage; therefore, indicating that these transcripts have unique regions that are not contiguous in the platyfish genome. Moreover, Cufflinks_GG and to a lesser extent Trinity_GG assemblies also had the greatest proportion of putative chimeric transcripts with secondary alignments in the genome (Appendix: Figure A2.1).

2.3.6 Transcript reconstruction of *X. maculatus titin* homolog

Since the longest vertebrate protein, the structural protein Titin (TTN), is usually the longest transcribed gene with the maximum number of exons; therefore, I evaluated the contiguity of the *titin* homolog reconstructed by each assembler. The homologous transcript was identified by the top ten BLASTN alignments identified using the *X.maculatus titin* mRNA as query against subject databases constructed from each assembled transcriptome. Comparing the total assembled length and aligned length of each reconstructed transcript, I found that the most complete and contiguous transcript is assembled by Oases_k27 and Oases_k31. Both these transcripts along with the other *titin* homologs from the single-*k* assemblies are not clustered in the Oases_clust assembly, while, all the transcripts are missing in the Oases_merge assembly. Among the other assembly methods TRINITY assembled a longer, contiguous transcript with exons from the 5' and 3'- end. All *titin* homologs assembled with genome-guided methods have a long contiguous 3'- end and several shorter contigs that align to the remaining coding exons. On closer inspection, I found that the transcripts in genome-guided assemblies were assembled using reads aligning to two different genomic scaffolds. Therefore, the transcript contiguity is broken due to incomplete assembly of the genomic scaffolds.

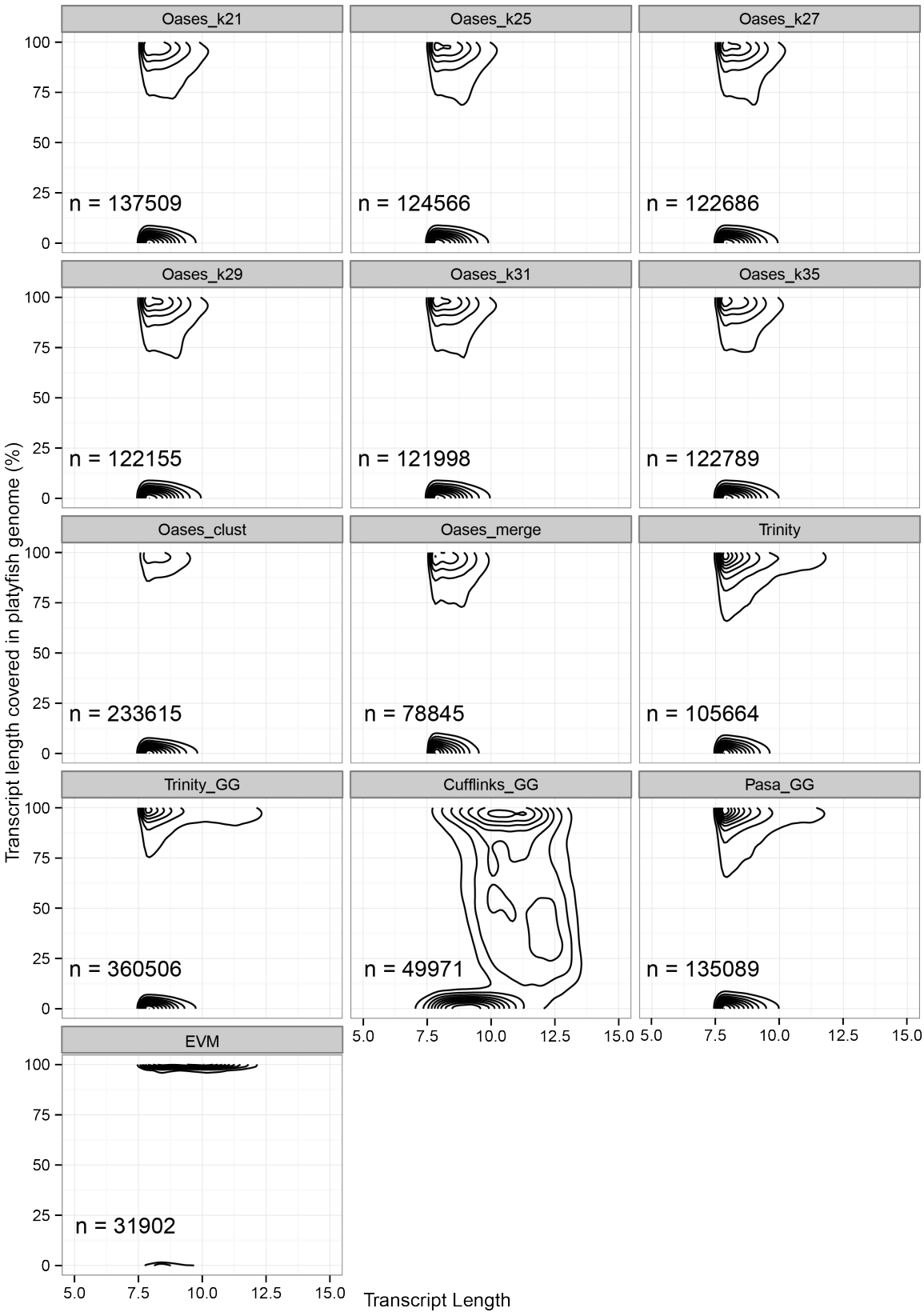

Figure 2.7: Nucleotide-level coverage of guppy assembly transcripts in the platyfish genome. Contour lines show the 2-dimensional density of data points in each assembly. Density distributions are calculated for points in a scatterplot showing percentages of total transcript length covered in the platyfish genome (Y-axis) versus $\log_2$-transformed transcript length (X-axis).

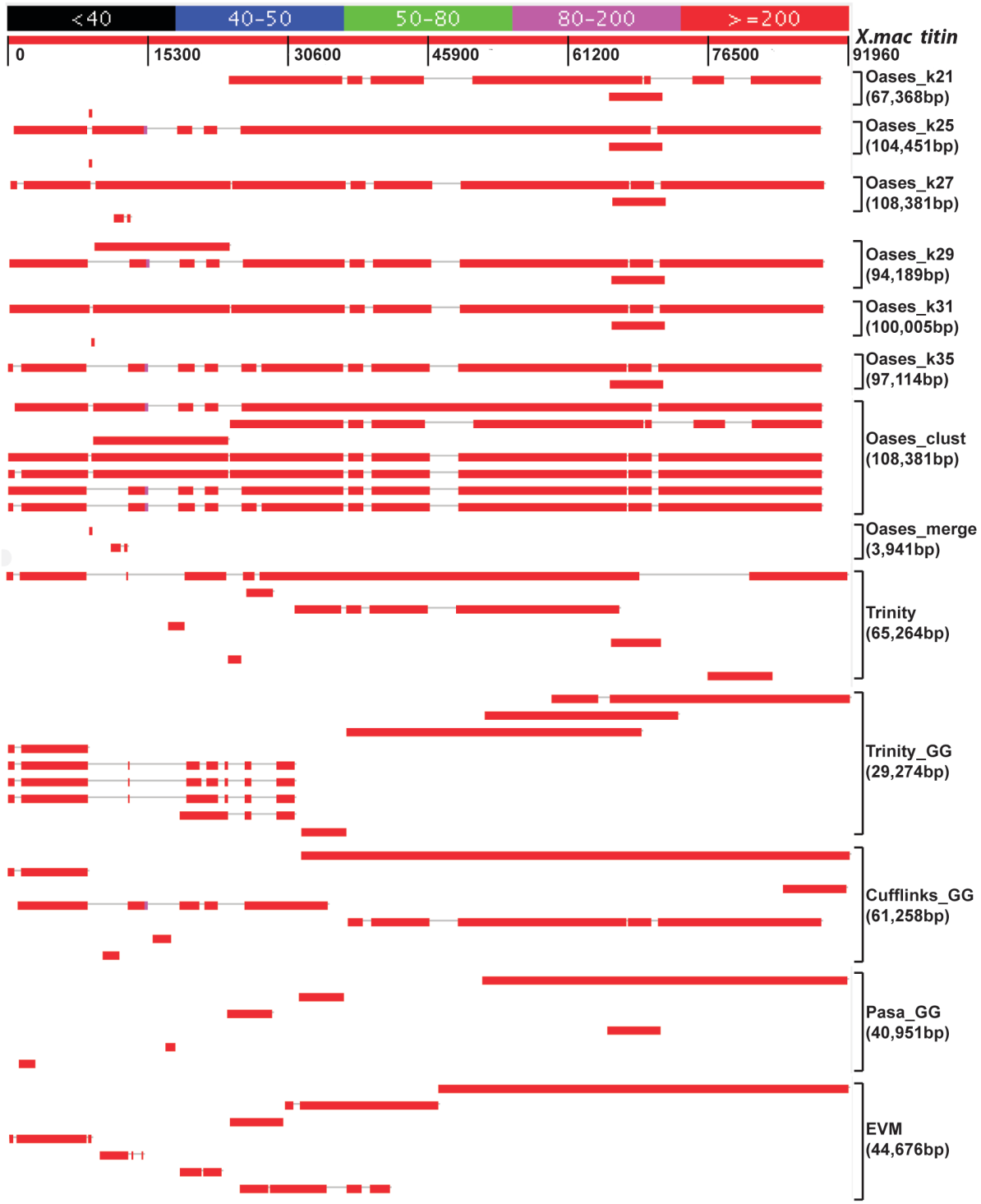

Figure 2.8: Transcript reconstruction of *titin* mRNA homolog. Transcripts with high coverage of *X.maculatus titin* are aligned using BLASTN for alignment of two or more sequences. The red colour corresponds to alignment score greater than 200 as shown in the colour legend on top. Square brackets indicate the best-score alignments (maximum 10) per assembly. The assembler name and length of the longest contiguous transcript is shown on the right of the bracket.

2.4 Conclusions

Reconstructing a comprehensive transcriptome from short reads has different computational challenges than a genome assembly. Firstly, unlike genomic sequences that have mostly uniform coverage, sequencing depth of transcripts can vary several orders of magnitude depending on their expression level and length. Secondly, transcriptome assemblers need strand-specific information to resolve overlapping sense and antisense transcripts (Makalowska, et al. 2005). Thirdly, splice variants from the same gene can share exons and hence are difficult to resolve unambiguously. Finally, the complexity of eukaryotic transcriptomes is increased in the presence of gene and genome duplications that give rise to paralogous and homeologous transcripts, many of which acquire tissue-specific expression and splice patterns or become pseudogenes.

Here, I evaluated the guppy transcriptomes reconstructed with genome-guided and genome-independent assemblers using length-based, mapping-based, and annotation-based metrics similar to those proposed for comparisons of genome and transcriptome assemblies (Martin and Wang 2011; O'Neil and Emrich 2013). In agreement with other such comparisons (Jain, et al. 2013; Lu, et al. 2013; Steijger, et al. 2013), we found that no single method excelled in all metrics. For length-based metrics, the average lengths of transcriptomes did not necessarily compare with average lengths of coding sequences, N50 lengths, and total lengths. Genome-guided assembly with CUFFLINKS, *de novo* assembly with TRINITY and combined assembly with EVM showed comparable results for predicted CDS but not for full-length transcripts. While the length-based metrics reflect the contiguity and completeness of genome assemblies, these metrics are not consistent with the quality of transcriptome assemblies (O'Neil and Emrich 2013). Comparison of percentage of reads used and proportion of uniquely and concordantly mapped read pairs suggested that the *de novo* assemblies with OASES and TRINITY incorporate more reads with greater specificity, while the genome-guided assemblies may contain a greater number of redundant and chimeric transcripts. Examination of transcript lengths and expression together highlights Oases_k25, Trinity, Cufflinks_GG, Pasa_GG, and EVM assemblies to have the largest proportions of long, expressed transcripts. An outcome of evaluating the expression along lengths of reconstructed transcripts was the distinction between Oases_k25 and the other Oases single-*k* assemblies. This superiority of the k25 assembly is not apparent in most other comparisons. A second interesting observation was the high proportions of transcripts with zero expression in the very huge Oases_clust and Trinity_GG assemblies.

Overall the length- and mapping-based metrics provide no information on whether the

assembled transcripts represent plausible mRNA sequences that are similar to an organism's genes (Misner, et al. 2013). Therefore, we further assessed the number of reciprocal best-blast hit protein sequence orthologs, exon-level coverage and nucleotide-level coverage using the annotated genome of the closely related teleost, platyfish. All comparisons clearly indicate that no one assembly can provide a complete set of the minimum contigs that fully reconstruct the guppy's protein-coding transcriptome. On one hand, assembly with assistance of genomic information enables reconstruction of more full-length sequences with CUFFLINKS and more coding exons with genome-guided TRINITY. On the other hand, these assemblies contain artifacts such as chimeric transcripts and non-expressed transcripts. Possibly, non-expressed genomic regions or overlapping genes may be linked together as indicated by the long UTRs in Cufflinks_GG. Additionally, due to gaps and mis-assemblies present in incomplete draft genomes, several transcripts may be missing, have fragmented assembly, or incorporate errors present in the genome assembly. Combining the output of several individual assemblies, such as clustering multiple *k*-mer assemblies of OASES, or tissue-specific assembly with genome-guided TRINITY results in the recovery of more plausible coding sequences but suffers from the problem of overwhelming numbers of redundant transcripts.

An advantageous approach would be to include size-based and FPKM-based filtering in addition to clustering, to remove potential mis-assemblies and redundant transcripts. Re-assembly of Trinity_GG transcripts using PASA and subsequently EVM, considerably reduces the redundancy, as transcripts with high read support are extended with genomic information (PASA) and gene structure prediction algorithms (EVM). However, the Pasa_GG transcripts, and to a lesser extent the EVM transcripts, recover fewer homologous exons and protein sequence orthologs. This suggests that a sizable proportion of valid sequence information is also lost due to stringent filtration. It may be possible to optimize these parameters in an assembly specific manner so as to ensure the maximum recovery of valid sequences.

Our results indicate that *de novo* assembly with TRINITY outperforms all single-*k* assemblies with OASES in terms of valid protein sequences and exon recovery. TRINITY also seems to reconstruct long, accurate and unique transcripts with the aid of a large number of the paired-end short reads. However, certain sequences are assembled only with alternate *k*-mers or/and genomic information. Considering genome-guided assemblers, although Cufflinks_GG was not the assembly with maximum number of orthologs, CUFFLINKS outperforms the other strategies for the recovery of the minimum set of transcripts with the maximum coverage of biologically relevant coding sequences.

Chapter 3: Annotation and analysis of the combined reference transcriptome

3.1 Introduction

Gene-expression studies provide a snapshot of the functional status of a genome at an experimental point with respect to its environment. At the fundamental level, gene-expression is the process by which a genotype is connected to its phenotype. This may be studied at the level of RNA transcription (for both coding and non-coding genes) or protein translation (for coding genes). In this chapter, I describe the use of RNA-seq to identify the tissue-specific sex-bias in gene transcription in somatic and reproductive tissues of the guppy in order to understand the molecular mechanisms underlying the phenotypic differences between the sexes.

Identification and quantification of relative transcript abundance using RNA-seq requires a representative reference of gene-models whose transcription status must be studied. The first step in the analysis involves alignment of paired- or single-end reads against the reference transcriptome. Transcripts missing from the reference will be lost at the step of read alignment. The second step is the quantification of aligned reads. In an approach that differs from hybridization-based quantification of microarrays, RNA-seq enables direct-sequencing and read-count based quantification. This requires the use of appropriate mathematical models that not only account for biological variations but also the biases introduced due to sample preparation, sequencing methods and other technical effects. Appropriate quantification of multiple alignments and transcript isoforms makes this process more challenging and is often ignored in more simplistic analyses. Relative expression of transcripts is then calculated using statistical tests to compare the difference between mean-counts in two or more sample distributions. The values for statistical significance must then be corrected to account for the large number of comparisons. Finally, using an *a priori* cut-off for significance a set of differentially-expressed transcripts or genes can be identified.

Further information about gene-models is obtained from their genomic locations and functional annotations. For non-model species this is often a rate-limiting step as only a few genes may have been characterized functionally. Therefore, annotations are usually obtained using functional information available for putative orthologs identified using sequence-homology based approaches. I have used a similar approach for the annotation of the guppy reference transcriptome. A large number of assembled transcripts could not be annotated

using this approach; therefore, for my analysis I focus on the annotated transcripts alone. Furthermore, I used this annotated reference transcriptome to examine whether sex-biased gene expression in the adult guppy's brain, tail, and gonad tissues reflects the morphological and physiological differences in these tissues.

3.2 Materials and methods

3.2.1 Combined reference transcriptome and functional annotation

I merged the genome-independent and genome-guided assemblies by pooling the predicted CDS from both assemblies followed by clustering sequences with 90% identity using CD-HIT-EST to create a guppy reference transcriptome (GRT). Annotations were found using BLASTX with query GRT sequences against the NCBI non-redundant protein database (Pruitt, et al. 2014). The contigs from the guppy reference were then annotated using BLAST DESCRIPTION ANNOTATER (BDA) implemented in BLAST2GO® v2.7.0 (Gotz, et al. 2008). Gene ontology categories were assigned by mapping GO terms and the ANNEX-based annotation augmentation using BLAST2GO. Translated peptide sequences of transcripts assembled by CUFFLINKS and TRINITY were separately annotated using probable orthologous sequences in other teleost databases. Orthologs were identified using reciprocal best-blast hit comparisons with PROTEINORTHO as described in chapter 2.2.6.

3.2.2 Alignment against female genome

Genomic coordinates of predicted CDS of the reference transcriptome were obtained by aligning them against the repeat-masked draft female genome using GMAP v2012-07-20 (Wu and Watanabe 2005). In cases of ambiguous alignments (Total: 607), the alignment with the highest total coverage and identity was kept.

3.2.3 Differential expression analysis

Each barcoded sequenced library from the organ datasets (F_{brain}, M_{brain}, F_{tail}, M_{tail}, F_{gonad}, M_{gonad}) were individually aligned to the guppy reference transcriptome using BOWTIE2 v2.0.04. Mapped reads were counted using EXPRESS v1.3.1 (Roberts and Pachter 2013). Read counts from six individually barcoded biological replicates per tissue were used for differential expression analysis between male and female tissues using the BIOCONDUCTOR (Gentleman, et al. 2004) package EDGER v3.0.8 (Robinson, et al. 2010). First low abundance CDS with less than two counts per million mapped reads (< 2 CPM/sample) across six samples were removed. Read counts were normalized for sequencing depth using TMM normalization (Robinson and Oshlack 2010). Differential expression between the sexes was tested with a modified exact test implemented in EDGER,(Robinson and Smyth 2007). P-values were corrected for multiple comparisons by estimating False Discovery Rates (FDR)

(Benjamini and Hochberg 1995). CDSs with significant expression difference between the sexes (FDR < 0.1 or if mentioned FDR < 0.05) were classified as sex-biased and CDSs with no significant difference between the sexes (FDR > 0.1) were called unbiased. All sex-biased sequences identified showed at least a 1.2 fold difference ($log_2FC > 0.3$ or < -0.3) in expression between the sexes. I refer to CDSs with preferential expression in males or females as male-biased or female-biased respectively. Genes with sex-specific functions may have varying levels of expression divergence in different tissues (Assis, et al. 2012; Meisel 2011; Pointer, et al. 2013). Therefore, assuming that a greater sex-bias in expression suggests increased sex-specificity, I further categorized the sex-biased CDS by fold-change, keeping CDSs with greater than median-fold difference in expression between sexes, log_2 (Male/Female) in each study tissue. These median-fold cutoffs were, Brain: 1.5 fold ($log_2FC > 0.6$ or < -0.6); Tail: 1.7 fold, ($log_2FC > 0.8$ or < -0.8); and Gonad: 3.2 fold ($log_2FC > 1.8$ or < -1.8).

3.2.4 Gene-set enrichment analysis

Gene ontology categories that were over-represented ($p<0.01$, N >=3) among median-fold sex-biased sequences were compared to all annotated sequences in the guppy reference using a Fisher's exact test with the ELIM algorithm implemented in the R package: TOPGO v2.10.0 (Alexa and Rahnenfuhrer 2010). Enriched GO terms were reduced to unique and informative terms that were visualized in a force-directed graph using REViGO (Supek, et al. 2011) and CYTOSCAPE v3.0.2 (Cline, et al. 2007).

3.2.5 Sex-biased expression of pigmentation and sex-related candidate genes

To identify the guppy homologs of candidate genes, I parsed the reciprocal best BLASTP-hit orthologs obtained using PROTEINORTHO for corresponding candidate gene annotations in teleost, human and mouse databases. List of candidate genes expected to be involved in pigmentation, patterning; and sexual development were prepared from available literature. The compiled lists included 54 candidate genes for sex determination and differentiation (Berbejillo, et al. 2012; Forconi, et al. 2013; Mank and Avise 2009) and 132 candidate genes for pigmentation and patterning (Braasch, et al. 2009; Schartl, et al. 2013). I extracted the transcripts encoding putative candidate proteins as well as their paralogs and the checked the identity of each retrieved transcript through NCBI BLASTN by homology. Finally, I parsed the list of differentially expressed sex-biased genes in each of the study tissues to identify sex-bias in expression (FDR < 0.1) in each of the candidate genes.

3.3 Results and discussion

3.3.1 Comparing the outputs of TRINITY and CUFFLINKS

As transcriptome assembly with each assembler can produce a set of unique transcripts or fragments, a comprehensive reference may be generated by combining different assembly strategies. But, even so, a combined assembly may include more erroneous transcripts as errors of each assembly algorithm are compounded. Therefore, I further compared the output of TRINITY, among *de novo* assemblers, and CUFFLINKS among *ab initio* assemblers to evaluate the benefits of combining their resulting transcripts. These assemblies were chosen as they contained the least redundant set of transcripts with long CDS, maximum exon coverage and maximum valid protein coding sequences. Since my objective was to assemble and annotate a comprehensive reference of plausible coding sequences, I compared the number of single-copy orthologs identified from translated coding sequences (CDS) of the guppy assemblies against other teleost, human, and mouse protein sequence databases.

The total number of orthologs found between guppy and other species reflects the phylogenetic distance between the guppy and the other species (with the exception of medaka, *Oryzias latipes*, possibly due to the smaller size of the medaka protein database) (Figure 3.1).

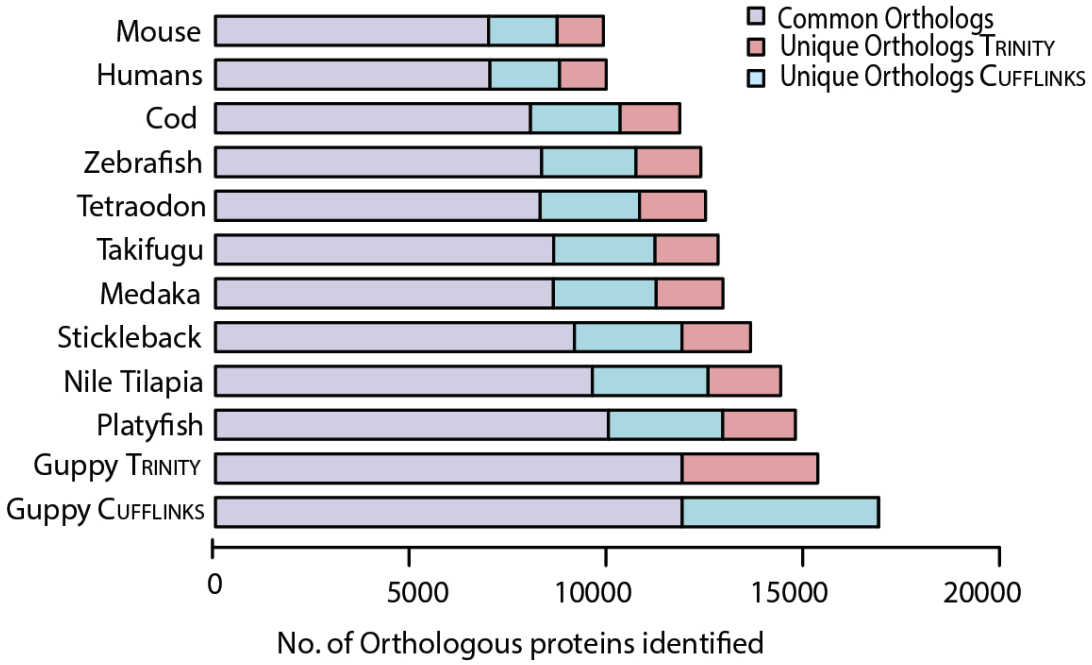

Figure 3.1: Barplots showing the number of protein sequence orthologs identified in other teleosts. Orthologs were identified in two-way reciprocal best-BLASTP hit comparison between platyfish, tilapia, medaka, stickleback, takifugu, tetraodon, zebrafish, cod, human, and mouse proteins. The stacked bars show the number of orthologs common between Cufflinks_GG and Trinity (purple), unique to Cufflinks_GG (blue) and unique to Trinity (red).

I identified 24,020 reciprocal best-blast hits shared between the genome-guided and genome-independent assemblies (Table 3.1). For approximately half of these overlapping set of peptide sequences (12,006), orthologous protein sequences were identified in other vertebrates. An additional 11,721 vertebrate protein orthologs were identified from only one of the two assemblies (Table 3.1). In addition to the identified reciprocal best-blast hit orthologs, 30-40% of the remaining translated CDS predicted from both genome-guided and genome-independent assemblies had significant sequence similarity (E-value < $1x10^{-20}$, alignment length > 50 amino acids) with protein coding sequences of the other vertebrates (Table 3.1). These may represent partially assembled sequences, incomplete CDS predictions and perhaps alternative splice-variants.

Table 3.1: Comparison of guppy transcriptomes assembled with genome-guided and genome-independent assemblers.

	TRINITY: Genome-independent assembly (GIA)	CUFFLINKS: Genome-guided assembly (GGA)
Total length (bp)	416,036,223	301,476,740
Length with longest isoforms per locus (bp)	101,831,430	128,048,246
No. of transfrags	213,088	91,126
No. of transcripts (Unique components/gene groups)	105,664	49,971
Mean length (bp)	1,952	3,308
Longest contig (bp)	65,264	61,058
Overall mapping (%)	73.21	73.64
Concordant and unique mapping (%)	62.98	55.10
Total no. of ORFs*	53,537	63,520
No. of complete ORFs*	29,309	49,535
Mean length ORF*(bp)	766	803
Longest ORF*(bp)	63,897	54,732
Total length of assembly with CDSs only (bp)	40,889,623	48,745,723
Number of best BLASTP alignments** (Orthologs***)		
Against guppy (GGA against GIA or vice-versa)	40,973 (24,020)	35,147 (24,020)
Xiphophorus maculatus	19,680 (13,399)	19,941 (14,934)
Oryzias latipes	17,925 (11,102)	18,197 (12,455)
Gasterosteus aculeatus	19,139 (11,758)	19,429 (13,096)
Orthologs in only one assembly	4,767	6,954

NOTE- * Predicted ORFs with minimum length greater than 50 amino acids
** Best BLASTP hits against other protein sequence databases (E-value < $1x10^{-20}$)
*** Reciprocal best BLASTP hits identified using PROTEINORTHO

3.3.2 Generation of a guppy reference transcriptome

As there was a clear advantage in merging the output of both assemblers, I pooled the predicted coding sequences from both Trinity and Cufflinks_GG followed by clustering to generate a non-redundant set of putative coding sequences. This strategy for merging plausible CDS from each assembly was chosen to minimize potential errors due to mis-joins and artifacts in UTR regions, non-coding transcripts as well as erroneous annotations of possible fusion transcripts (Grabherr, et al. 2011a). The final dataset consisted of 74,567 (46,798 from Cufflinks_GG and 27,769 from Trinity) sequences. These predicted CDS were used for subsequent analysis and are referred to as guppy reference transcriptome (GRT) (Figure 3.2).

3.3.3 Functional annotation of the guppy reference transcriptome

I performed *de novo* annotation of the GRT using two amino acid sequence similarity based procedures. The first set of annotations were performed using the gene names obtained from annotated gene models encoding orthologous protein sequences (reciprocal best-BLASTP similarity) in other teleosts, humans and mouse. Gene names were assigned in the order of availability of annotated orthologs in zebrafish, platyfish, tilapia, medaka, stickleback, takifugu, tetraodon, human or mouse. The first preference was given to annotations obtained from orthologs in zebrafish, the teleost with most experimentally obtained functional annotations, followed by annotations from closely related species in order of divergence between species. Using this approach I annotated 18,357 contigs of the guppy reference transcriptome with 18,290 unique gene annotations. For the second set of annotations, I performed BLASTX against the NCBI non-redundant protein sequence database (NR). In total, 30,643 (41.1% of the GRT) sequences showed significant alignment (BLASTX E-value$< 1x10^{-15}$) to 22,780 different known or predicted proteins. The taxonomic classification of hits from the NR database is presented in Figure 3.3. The BLASTX E-value distribution and similarity distribution is presented in the appendix (Appendix: Figure A3.1).

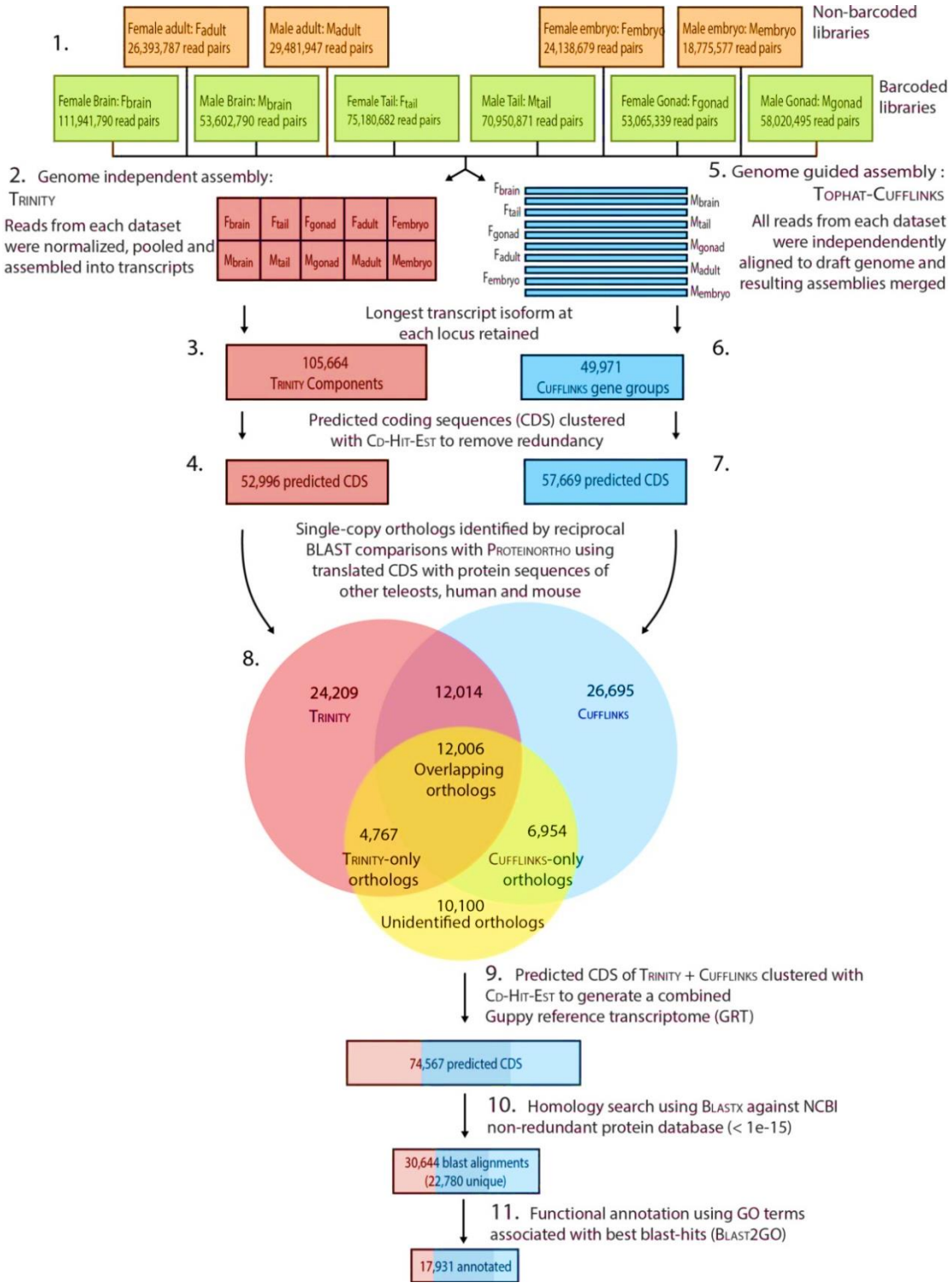

Figure 3.2: Flowchart describing sequencing data, assembly strategy, comparison, merging and annotation of the guppy reference transcriptome. (1) The high quality paired-end reads from each sequenced dataset (non-barcoded: orange and barcoded: green) were assembled into transcripts. Putative coding sequences were predicted for the genome-independent assembly, TRINITY (2-4, red), and genome-guided assembly, CUFFLINKS (5-7, blue). (8) Venn diagram showing number of protein sequence orthologs identified between at least two species. Orthologs were identified using translated sequences from the two guppy assemblies (red, blue), and protein sequence databases from eight teleosts, mouse, and human (yellow); (9-11) Merging of predicted CDS from both assemblies and functional annotation of the guppy reference transcriptome (GRT).

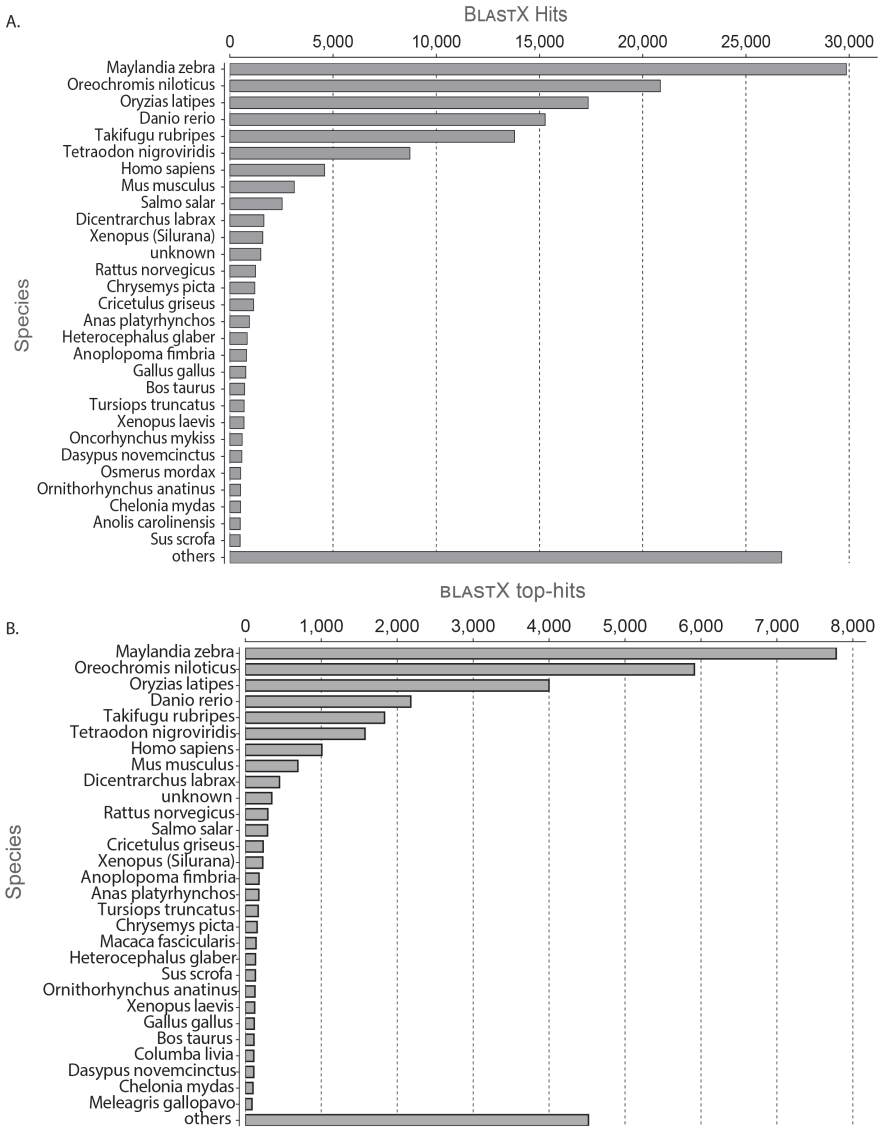

Figure 3.3: Taxonomic classification of annotations from BLASTX hits against NR database. The bars show the species distribution of top five (A) and best (B) **BLASTX** alignments (E-value< 1x10-15) of contigs from the guppy reference transcriptome against the non-redundant protein database.

3.3.4 Gene Ontology (GO) annotations

By mapping gene ontology terms associated with **BLASTX** homologs, 17,931 guppy reference contigs were annotated with 76,875 GO terms. The level of assigned GO category is calculated using the hierarchical vocabulary structure of GO's directed acyclic graph (DAG). The numeric level is indicative of a general (low-level) or specific (high-level) functional classification. The mean level of assigned categories was 6.2. The distribution of GO levels and evidence codes is shown in the appendix (Appendix: Figure A3.2). Figure 3.4 shows the number of putative coding sequences annotated with high-level (greater than level 3) terms for each of the GO domains, biological process, molecular function and cellular component.

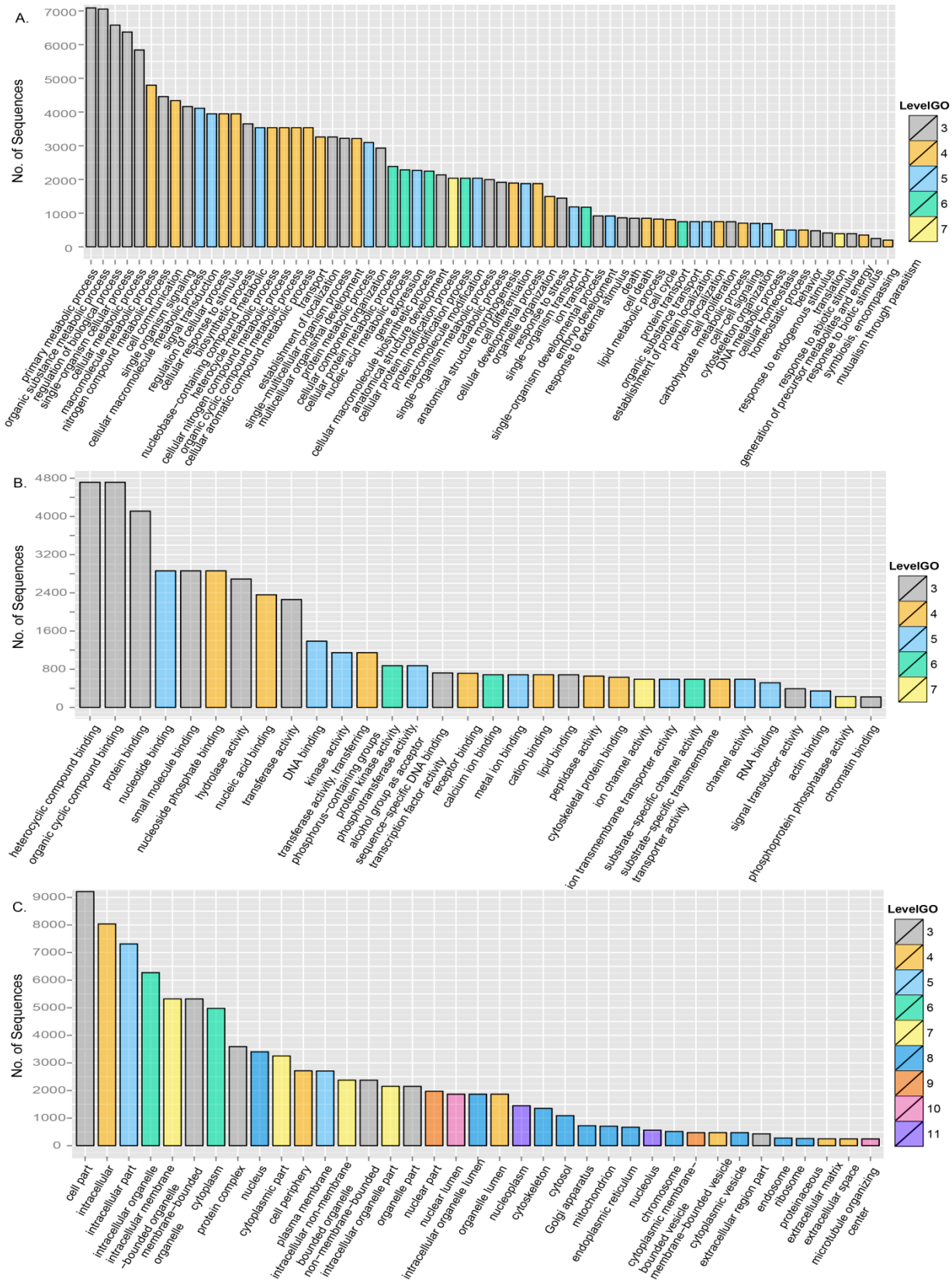

Figure 3.4: Distribution of Gene Ontology categories for the guppy reference transcriptome. The bar-plot represents the GO term names corresponding to annotations, greater than level 3 and Blast2GO confluence score 200, for biological process (A), molecular function (B), and cellular component (C) domains. Bars are coloured by level of GO annotation as specified in the plot legends (right).

3.3.5 Alignment to the draft genome of the female guppy

A total of 73,518 contigs of the guppy reference transcriptome could be aligned to the draft genome of the female guppy (Figure 3.5). Of these, 67,882 aligned to genomic scaffolds that were assigned to guppy linkage groups and 5,636 sequences aligned to scaffolds that could not be assigned to linkage groups (Unplaced: Un) (Künstner *et al.* submitted, GenBank ID GCA_000633615.2). All sequences that did not align to the female genome (1,044) were from the genome-independent assembly. I further attempted to align these contigs to the draft genome of the male (Künstner *et al.,* submitted; GenBank ID GCA_000633615.2) and to the genome of the platyfish (Schartl, et al. 2013). Among these, 179 sequences aligned to the genome of the male guppy (PrM), 435 aligned to both genomes (PrMXm), 186 aligned only to the platyfish genome (Xm), and 270 remained unaligned (NA). I then compared the genomic distributions of annotated and unannotated contigs with respect to total number of contigs in the alignment group to see if any group showed deviations from expected distribution. The number of annotated contigs was less than expected in the contig groups that aligned to Unplaced scaffolds (Un), and all groups that did not align to the female genome; PrM , Xm, PrMXm, and NA. In accordance, the distribution of unannotated contigs was greater than expected in a few of these contig groups, namely Un, PrM and NA (Figure 3.5). Overall, from the 1,044 contigs that could not be aligned to the female genome I could annotate only 223 contigs and could not identify any obvious male-specific sex or pigmentation related candidate among these (see Chapter 3.3.12).

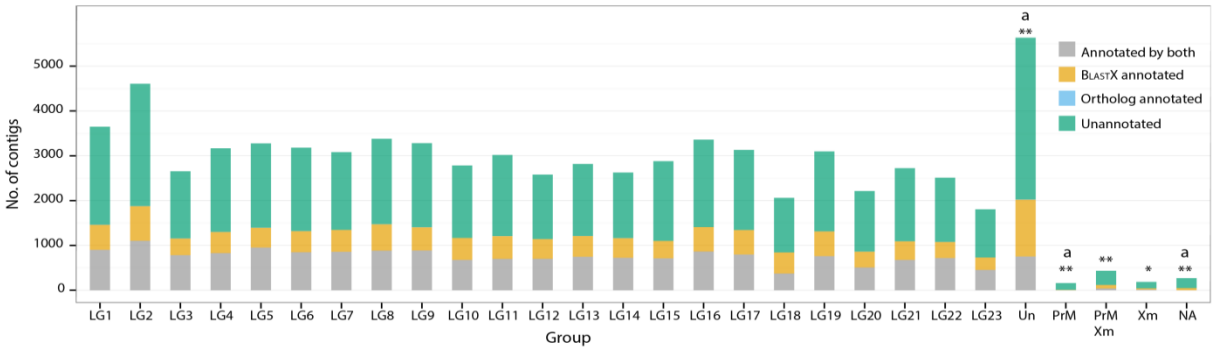

Figure 3.5.: Genomic distribution of contigs of the guppy reference transcriptome. The bar-plot shows the distribution of contigs that align to scaffolds from draft female genome (assigned to linkage groups: **LG1-LG23**; and unassigned to any linkage group: **Un**), draft male genome (**PrM**), platyfish genome (**Xm**), both male genome and platyfish genome (**PrMXm**) and contigs that remain un-aligned (**NA**). The colours show annotation by both methods (grey), **BLASTX** alone (yellow), orthologs alone (blue) and unannotated (green). The asterisk and alphabets above the bars mark the contig groups where I found a significant under-representation of annotated contigs (*) or over-representation of unannotated contigs (a). ** $p < 0.005$, * $p < 0.05$; **a** $p < 0.05$

3.3.6 Differential gene-expression between the sexes

To assess the extent of sex-biased expression in the guppy, I compared gene expression between males and females in three tissues with phenotypic sexual dimorphism in adult guppies (Figure 3.6).

1) **Brain (isolated with the eyes):** The guppy, like several other poeciliids, displays dimorphism between the brains of males and females (Alexander, et al. 2012). The brain tissue is also presumed to reflect some of the sex-associated hormonal and behavioural dimorphism.
2) **Tail:** The post-anal tissue included skin, muscle, bones, cartilage and end of the spinal cord, is presumed to reflect the pigmentation pattern and growth associated dimorphism.
3) **Gonads:** The guppy, being a gonochoristic fish species, has terminally differentiated gonads. Therefore, ovary and testis are the most sexually divergent organs and are expected to show the greatest degree of gene-expression divergence.

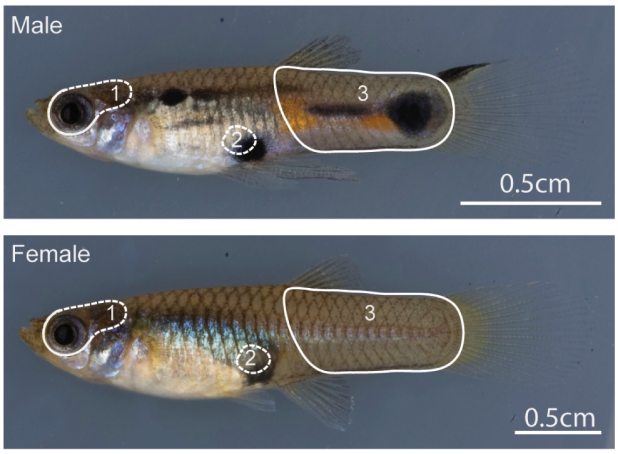

Figure 3.6: Phenotypic sexual dimorphism in the guppy. Males (top) are smaller than females (bottom) and have complex colour patterns on the body. The encircled region (white outline) indicates the tissues that were used for preparing the barcoded libraries, 1) brain and eyes; 2) Male testis and female ovary; and 3) tail.

A separate comparison of the somatic and reproductive tissue allowed us to isolate the degree of sex-biased gene expression in each of the study tissues. By mapping reads to predicted CDS of the guppy reference transcriptome, instead of transcripts, I tried to increase the accuracy of read assignment to putative genes but lost the information from reads that represent untranslated regions (UTRs). Therefore, I also performed differential expression

analysis after mapping reads to both the genome-guided and genome-independent assemblies and to the full-length transcripts in the merged GRT. Since the four analyses produced similar results (data not shown), I will present only the results obtained by mapping against the predicted CDS (hereafter referred as genes).

3.3.7 Sex-biased gene expression is tissue-specific

Following the pipeline for differential expression analysis (Figure 3.7 A), I first looked at the pattern of gene expression across all individual samples. Samples from the same somatic tissue show a strong correlation in gene-expression (Spearman's correlation $\rho > 0.85$, $p < 1x10^{-10}$), suggesting only a few differences between the sexes (Figure 3.7 B). As expected, the greatest sex related difference was observed between the ovary and testis where overall expression clustered by sex. By sub-setting the data to analyze only the expressed genes ($\log_2$CPM > 2), I found that the brain tissue had the highest number of expressed genes followed by the gonads and the tail (Figure 3.7 C).

There was a considerable overlap of expressed genes across any two tissues or all three tissues. Using normalized read counts, I then identified the genes with a significant difference in expression (FDR < 0.1) between the sexes, also referred as sex-biased genes. Expectedly, the gonads were the most sexually dimorphic tissues as seen by comparing the magnitude of differential expression (Figure 3.7 D) and total number of sex-biased genes in each tissue (Figure 3.8A). I further compared sets of all expressed genes and sex-biased genes across tissues in order to assess tissue-specificity. I found a significantly smaller overlap between sex-biased expression tissues as compared to the expression in tissues genes ($P < 1 \times 10^{-16}$, χ^2 -test for equality of proportions) suggesting tissue-specificity in sex-bias. By comparing the overlap between tissues, of all sex-biased genes with that of male-biased genes (Figure 3.8B) and female-biased genes (Figure 3.8C), I saw that male-biased genes showed significantly less overlap ($P < 1 \times 10^{-16}$, χ^2 -test for equality of proportions). This indicates that male tissues have more specific sex-biased genes than the female tissues.

As the degree of differential expression between sexes varied in the reproductive and non-reproductive tissues, I further categorized the dataset to study genes with high sex-bias within a tissue. I chose tissue-specific medians as the threshold fold-change required for sex-biased genes. Figure 3.8 shows the number of all sex-biased genes identified in a tissue and across multiple tissues before and after median-fold cutoff. The number of expressed genes and sex-biased genes in each category are summarized (Appendix Table A3.1). In chapters 3.2.8 - 3.2.11, I shall present further analysis of only the median-fold sex-biased genes.

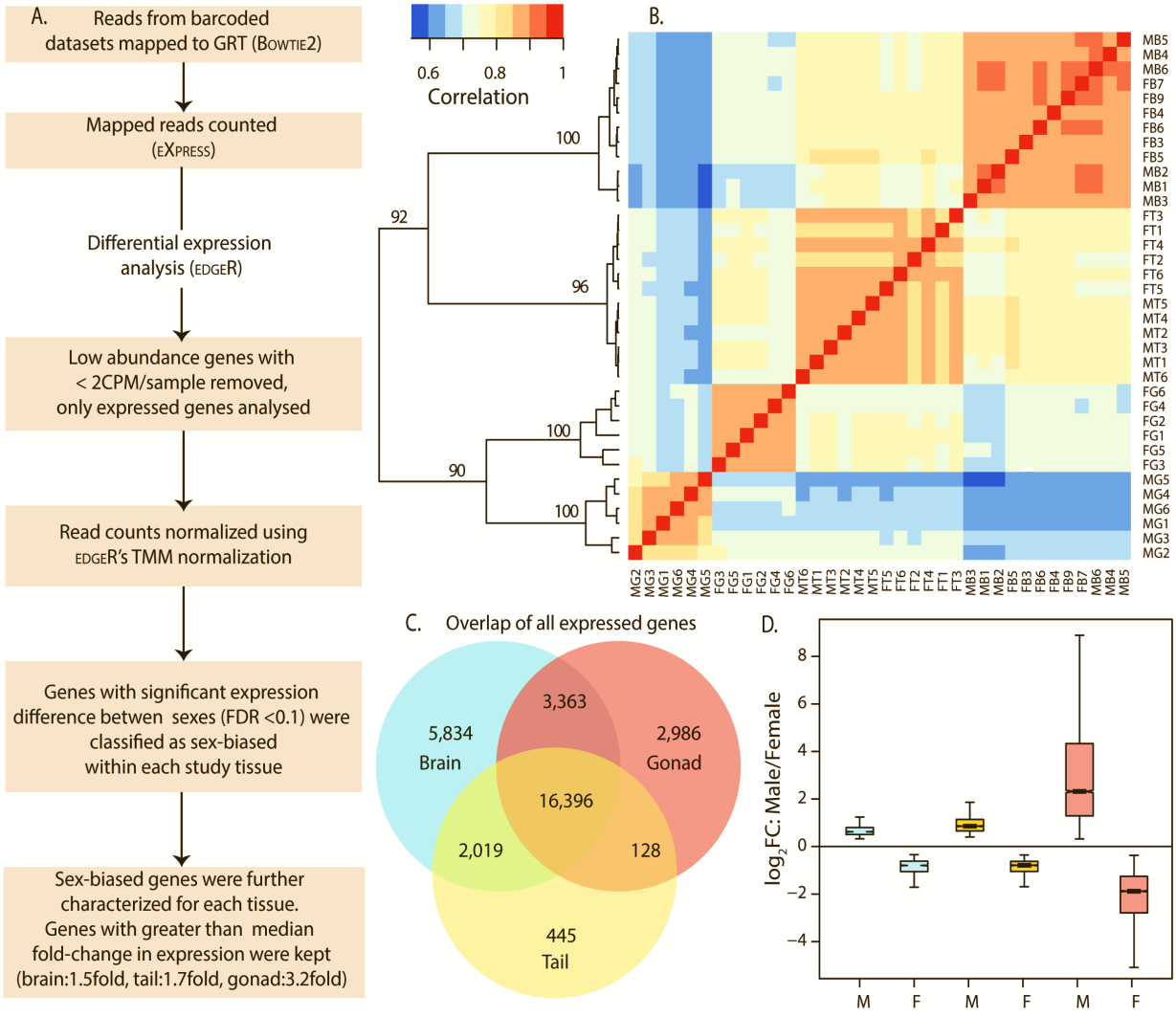

Figure 3.7: Tissue-specific analysis of sex-biased gene expression. (A) Flowchart shows steps used for abundance estimation and tissue-specific differential gene-expression analysis. (B) Spearman's correlation of normalized counts for adult tissue datasets. Heatmap displays spearman's correlation between samples. The dendrogram shows the bootstrapped agglomerative clustering (Ward's) by correlation in gene expression. Samples cluster by tissue-type, except for the gonads. The gonads show distinct expression from the somatic tissues and cluster by sex (Female Brain FB; Male Brain MB; Female tail FT; Male tail MT; Female gonad FG). (C) Venn diagram shows the overlap of expressed genes (log_2CPM > 2) in each tissue. (D) Boxplots (coloured by tissue in the same scheme as the Venn diagram) show the distribution of log_2FC (Fold change: Male/Female). The lower median of each pair was used as cutoff for significant fold change for that comparison (brain = 0.6; tail = 0.8; gonad = 1.8)

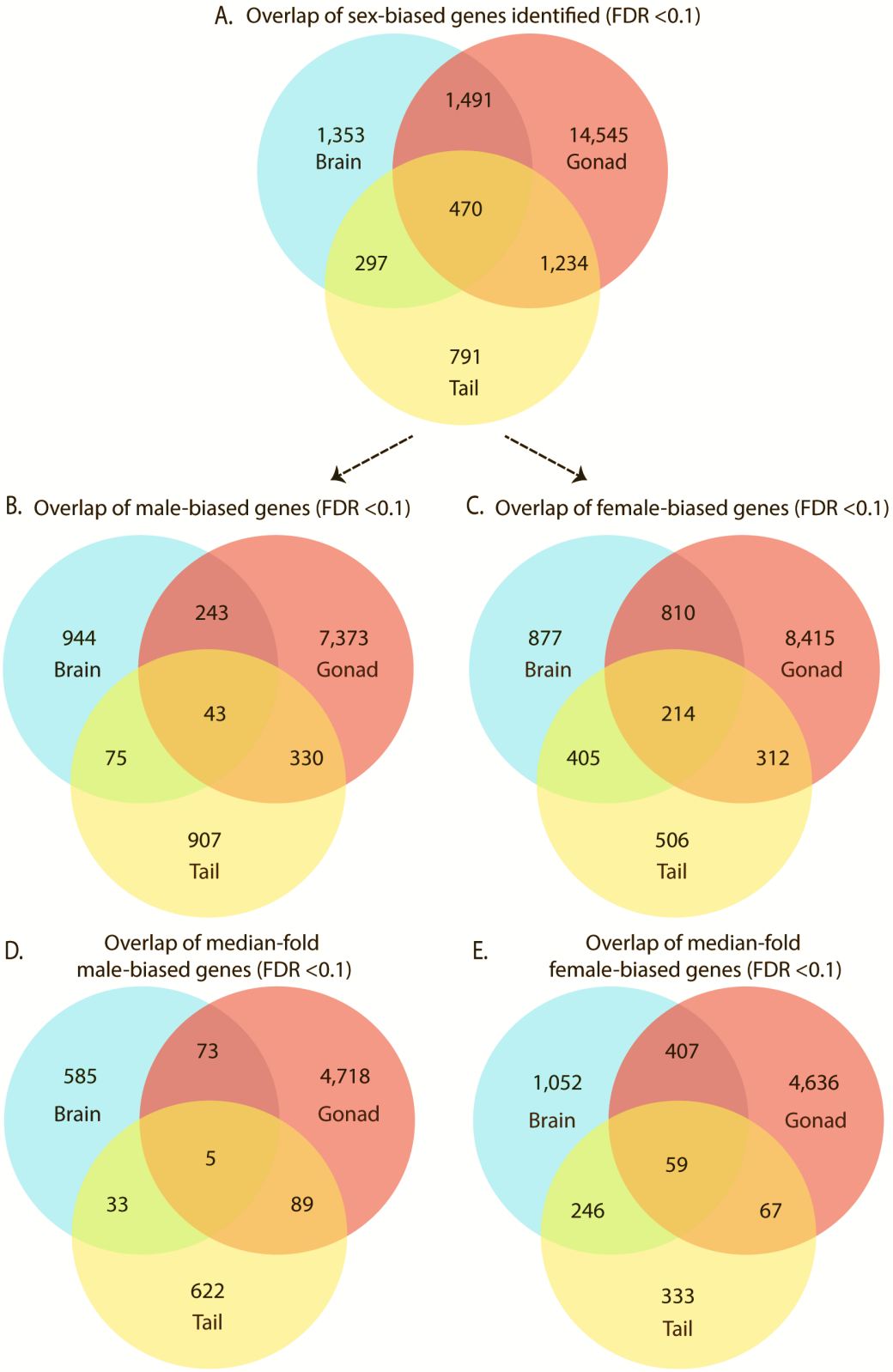

Figure 3.8: Tissue-specificity of sex-biased genes. Venn diagrams show the overlap between tissues in the three study tissues: brain (blue), tail (yellow) and gonad (red). All sex-biased genes (FDR < 0.1) (**A**); all male-biased genes (FDR < 0.1, $\log_2$FC > 0.3) (**B**); all female-biased genes (FDR < 0.1, $\log_2$FC < -0.3) (**C**); median-fold male-biased genes (FDR < 0.1) (D); and median-fold female-biased genes (FDR < 0.1) (**E**).

3.3.8 Female-brain has greater number of sex-biased genes

In the brain tissue, genes with female-biased expression greatly outnumbered those with male-biased expression (Figure 3.9A). Most genes identified as female-biased were expressed in both sexes but had significantly higher expression in females. Top female-biased transcripts encoded peptide hormones, e.g. growth hormone-1, chorionic gonadotrophin beta-1, prolactin, and the calcium binding proteins parvalbumin-2 and calsequestrin-1. The gene encoding teleost brain-specific aromatase, cytochrome P450 19A1b, was 5-fold higher expressed in the female than male brain (Figure 3.9B, Table 3.1A). Enriched gene ontology process terms included several related to DNA replication, growth, development, cell adhesion and migration, glycolysis, and immune response (Figure 3.10A, Table 3.1A). Notably, the most enriched cellular component term was proteinaceous extracellular matrix. Female-biased transcripts associated with the proteinaceous extracellular matrix, encoded basement membrane components nidogens, laminins, fibronectins, collagens, as well as specific matrix remodeling proteases metalloproteinases (Mmp-2-14) and members of a disintegrin and metalloproteinase with thrombospondin motifs (Adamts) family.

Annotated genes with the strongest male bias in expression encoded hypocretin/orexin transmembrane receptors, GABA receptors, Na^{+}- K^{+}- and Ca^{2+}- cation transport channels, and lens crystallins Crygm2d11 and Crygmxl2 (Figure 3.9B, Table 3.2A). Significantly high male-biased expression was also found in genes encoding some neuropeptide precursors: galanin prepropeptide, urotensin related peptide1, and CART prepropeptide (Figure 3.9B, Table 3.2A). Enriched gene ontology process terms among the male-biased genes were related to signal transduction, regulation of transmembrane ion transport, transmembrane receptors and cellular response and the most enriched cellular component term was integral to membrane (Figure 3.10B, Table 3.2A).

Female-biased expression of genes encoding cell-cycle and growth related hormones perhaps relates to the life-long growth observed in female guppies. Moreover, transcripts of the neurogenic zone associated aromatase, *cyp19a1b*, were higher expressed in the female brain, suggesting sexual dimorphism in adult neurogenesis in the guppy (Kaslin, et al. 2008; Le Page, et al. 2010). I found a female-bias in expression of many ECM components. Extracellular matrix (ECM) proteins and matrix proteinases have previously been associated with neurogenesis and synaptic plasticity (Fujioka, et al. 2012; Wlodarczyk, et al. 2011). The suggested greater plasticity in female brain in comparison to male guppies also relates to their behavioural dimorphism (Lucon-Xiccato and Bisazza 2014), based on predator avoidance, kin-recognition, and mate choice preferences in wild (Griffiths and Magurran 1998; Houde

1997; Magurran and Garcia 2000; Reader and Laland 2000). On the other hand, male-biased transcription of mRNA encoding neuropeptides and transmembrane receptors suggests sex-differences in signal transduction. These may relate to the male-associated responses to stimuli such as predator risk and mating opportunities that largely determine their courtship behaviour (Godin 1995; Magurran and Seghers 1990). Among male-biased neuropeptides, galanin is known to be involved in the neuroendocrine regulation of growth and reproduction in fish (Mensah, et al. 2010). Galanin neuropeptide and its receptor have also been shown to be highly expressed in parts of the brain of male sailfin mollies (*Poecilia latipinna*) (Cornbrooks and Parsons 1991a; Cornbrooks and Parsons 1991b).

3.3.9 Sex-biased genes in tail relate to growth and pigmentation dimorphism

We found similar numbers of male- and female-biased genes in the tail although the number of tail-specific genes was higher in the male-bias set than female-bias set (Figure 3.8, Figure 3.9C). Among female-biased genes, gene ontology biological process categories for cell-division, mitosis, DNA replication, DNA repair, recombination and glycolysis were over-represented (Figure 3.10C, Table 3.2B). The top enriched cellular component terms were collagen, myosin filament, and proteinaceous extracellular matrix. Differentially expressed genes with growth-related functions included mitotic cell-cycle factors cyclin B1, cyclin A2, cyclin dependent kinase-1, and mini-chromosome maintenance (MCM) replication initiation factors (Figure 3.9D, Table 3.2B). GO terms related to signaling pathways, vesicle transport, transmembrane transport and pigment biosynthesis were over-represented among the male-biased sequences (Figure 3.10D, Table 3.2B). Particularly, several top male-biased genes encoded proteins with functions in pigmentation processes (Figure 3.9D, and Chapter 3.2.12).
Sex-biased gene expression in tail relates to the phenotypic dimorphism in this multi-tissue sample. In adult guppies the tail tissue, comprising skin, muscle, bone, cartilage and end of the spinal cord, is characterized by complex pigment-patterns on male skin and life-long body growth in females. The male-biased expression of transcripts encoding vesicle transport and pigment biosynthesis proteins presumably reflects the greater concentration of pigment granule containing cells in the male skin (Kottler, et al. 2013; Kottler, et al. 2014). Similarly, the female-bias in a number of transcripts encoding cell-cycle, DNA replication, metabolism and growth-related proteins are indicative of the observed indeterminate growth and high-energy intake of female guppies.

3.3.10 Testis-biased genes show higher fold-change in expression than ovaries

Nearly 77% of all expressed genes in the gonads showed sex-biased expression (Figure 3.8, Appendix Table A3.1). I also found a number of genes with probable sex-limited expression in ovary or testis (black line in Figure 3.9E). Female-limited and female-biased genes include those encoding aromatase A (Cyp19a1a), the zona pellucida glycoproteins Zp1 and Zp2, oocyte specific proteins Zar1, Zar1l and growth differentiation factor Gdf9 (Figure 3.9F).

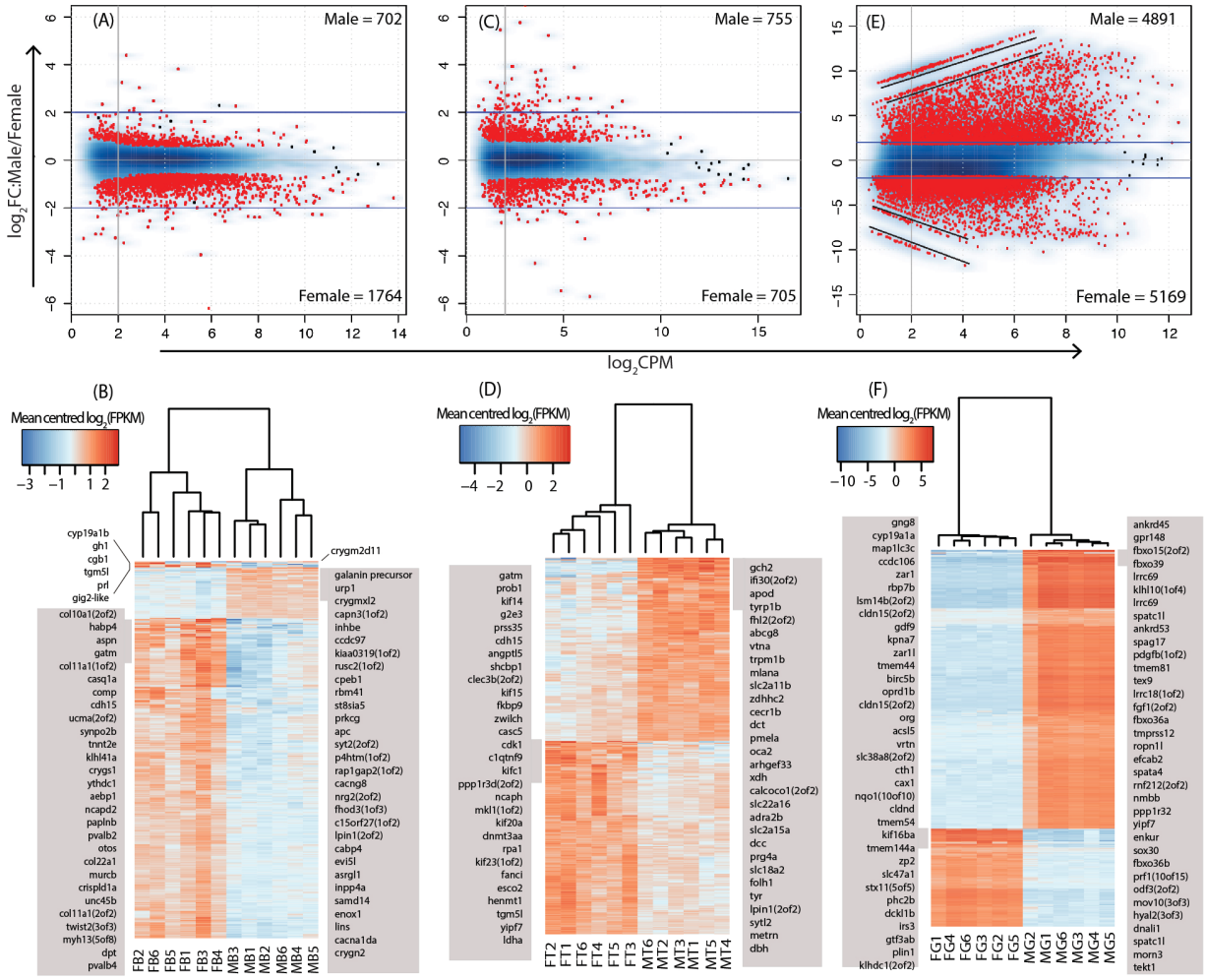

Figure 3.9: Quantitative differences in gene expression between sexes. Male/Female expression ratios (log_2FC, Fold-change) plotted against the average expression intensity (log_2CPM, Counts per million) in (**A**) brain, (**C**) tail, and (**E**) gonads. Genes with greater than median-fold bias (FDR <0.1) are in red while others are shown by black dots or smoothened (blue). Genes with higher expression in males have positive log_2FC, while those with higher expression in females have negative log_2FC. Blue lines mark 4-fold difference in expression between sexes. Genes with sex-limited expression are underlined in black in (E). Heatmaps (**B**), (**D**), and (**F**) show mean centered log_2FPKM (Fragments Per Kilo base per Million) for differentially expressed genes (FDR < 0.001). Expression levels of genes with greater than 1.5 fold-change (B, brain), 1.7 fold-change (D, tail), and 32 fold-change (F, gonad) in expression between the sexes are indicated by colour (red: high, blue: low). Grey boxes show the top 30 sex-biased genes in each tissue : left box (female-biased), right box (male-biased).

Over-represented GO terms associated with female-biased genes were blood vessel development, regulation of BMP signaling pathway, amino acid transport, focal adhesion, cell migration involved in gastrulation, FGF receptor signaling, apical protein localization, regulation of body-fluid levels, and gas transport (Figure 3.10E, Table 3.2E). Male-limited and male-biased transcripts also showed greater magnitude of fold-changes than the female-biased transcripts (Figure 3.9E, F). Several top testis-biased genes could not be annotated or encoded proteins with repeat rich domains, e.g. leucine rich repeats (Lrr) and ankyrin repeat domain (Ankrd). Others encoded sperm associated antigens, ciliary and flagellar proteins (e.g. Spag17, Spag6, Tekt-1), spermatogenesis related Spatc1l and Spata4, and testis expressed Tex9 (Figure 3.9F). Enriched GO-terms associated with male-biased genes included cilium assembly, spermatogenesis, microtubule-based movement, meiosis I (Figure 3.10F, Table 3.2E). I also examined the expression bias of sex differentiation and development associated genes in more detail (see Chapter 3.3.12). Expectedly, the differentiated adult gonads showed extremely divergent gene-expression profiles. The enrichment of terms related with spermatogenesis and testis-maintenance for male-biased genes and oogenesis and ovary-maintenance for female-biased genes relates to the sex-specific specializations of the gonads. The enrichment for follicular vascularization factors in female-biased genes is in accordance with the lecithotrophic developmental strategy of guppies (Thibault and Schultz 1978). In lecithotrophic species, oocyte maturation is accompanied by the transport of yolk precursors, amino acids and other metabolites from the blood to the maturing oocyte through a specialized highly vascularized follicle (Jollie and Jollie 1964; Turner 1940).

3.3.11 Genes with common sex-biased expression in brain and tail

Considering the overlap of sex-biased gene expression in two or all three tissues, we observe that a greater number of female-biased genes than the male-biased genes show a common direction of expression bias (Figure 3.8D, E). Despite considerable overlap between sex-biased genes in reproductive and somatic tissues, I evaluated GO enrichment in sex-biased genes with common expression bias in brain and tail alone. This was done to avoid ambiguity in comparison of genes with widely different fold-changes in expression. Over-represented GO terms among genes with female-biased expression in both brain and tail included glycolysis, DNA replication and recombination as biological process terms, and extracellular matrix, collagen and myosin as the cellular component terms (Figure 3.11A, C). For genes with male-biased expression in both brain and tail, all enriched biological process terms related to cation transmembrane transport and response to stimulus (Figure 3.11B).

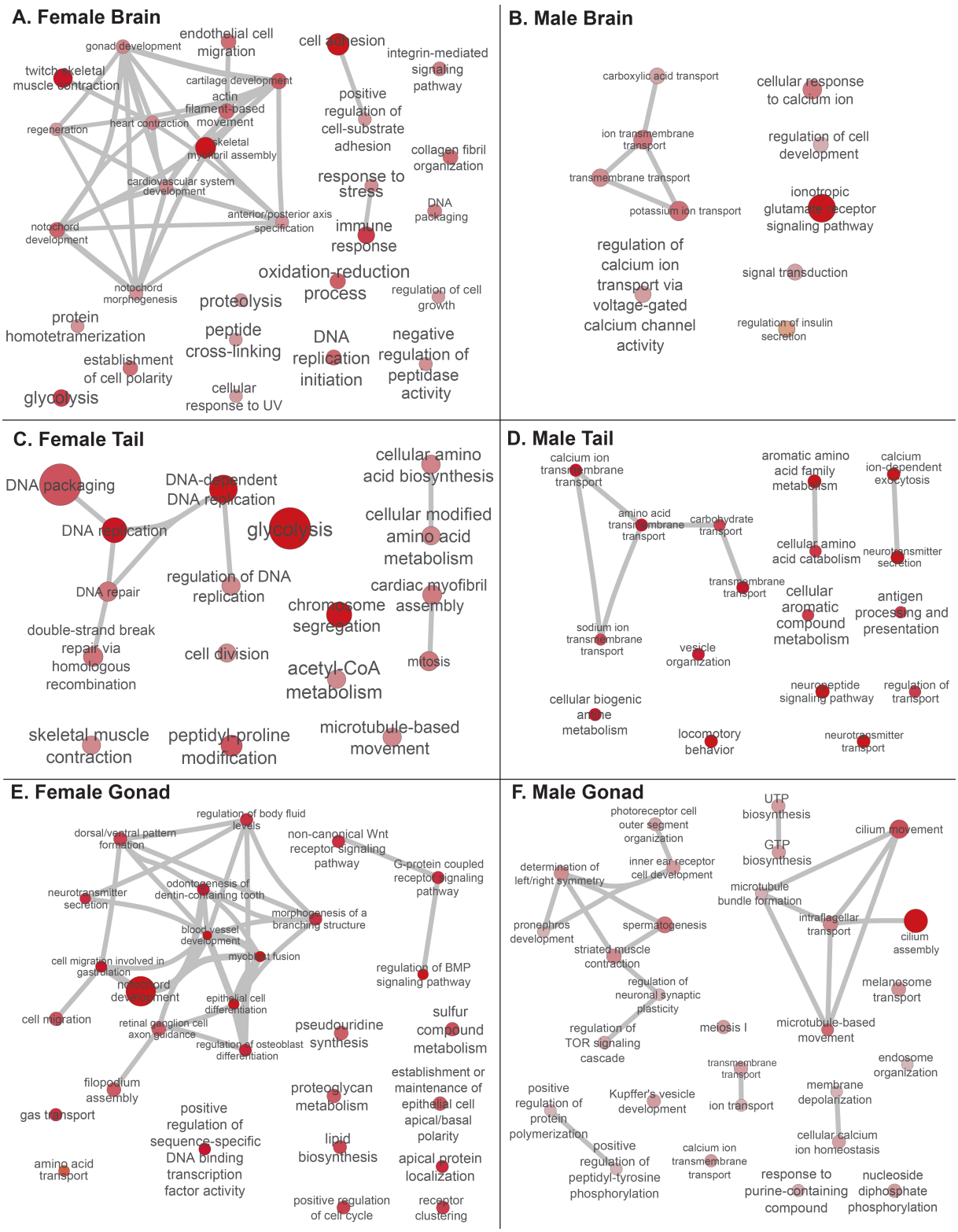

Figure 3.10: Gene ontology biological process terms enriched among sex-biased genes. The force-directed graphs show the enriched biological process terms after slimming. The node size correlates to the -log10 p-value of enrichment. Edges between nodes show the connected biological process terms. Label size represents the uniqueness of the term as determined by ReViGO in comparison to Uniprot database. Figures A,B; C,D; and E,F show enriched GO terms among female- and male- biased genes in brain, tail and gonad respectively.