Evolution of the synapse transcriptome

by

Abril Izquierdo Barraza

January, 2019

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

School of Veterinary Medicine and Science

Advance Data Analysis Centre

Abstract

The vast cognitive repertoire seen within the animal kingdom from rudimentary forms of habituation and information-processing to highly complex cognitive processes that confers the ability to adapt to challenging environments is a topic of great interest. The presynaptic and postsynaptic terminals of the synapse form an immensely structured protein network, the origin of which has been proposed to precede the origin of multicellularity in elementary cell signalling pathways. Such molecules were central for the arrangement of macromolecular complexes through genome duplications and posterior diversification in the vertebrate evolution. Yet, mutations in the postsynaptic density (PSD) are associated to more than 130 neurological alterations. It is therefore fundamental to better understand brain gene expression and evolution of these genes. Proteomic analysis of the synapse have characterised more than 1,500 proteins, however strikingly, there is a lack of research using recent transcriptomics approaches.

This PhD thesis contributes to understanding of comparative synaptic biology by exploiting NGS technologies to generate a comprehensive analysis of gene expression of brain tissues. A *de novo* transcriptome assembly pipeline was developed and employed to that end. We sequenced and generated a *de novo* transcriptome from brain tissues of zebrafish, bat and lion to explore the presence of genes known to be essential in learning and memory (Emes and Grant, 2012). To adequately provide a richer understanding of neurological diseases in humans, it is essential to investigate the magnitude of which metazoan genes shared orthologs. Transcripts enriched and specific to each tissue were determined, along with the analysis of which mouse orthologous

genes were present in the brain, synaptosome (SYN) and PSD of zebrafish, bat and lion.

This research revealed a strong conservation of PSD and SYN components, where the genes with the highest expression in the three species, i.e., cell-adhesion and signalling enzymes represent the core adaptive machinery of the ancestral synapse. In addition, this work demonstrates a substantial connection of highly expressed genes with critical neurodegenerative diseases, highlighting the urgency to improve the understanding of synaptic dysfunction. Lastly, this study provides the first exploration of bat and lion transcripts encoded in the brain, SYN and PSD, in which species-specific adaptations were found, along with evidence of convergent evolution in the echolocating bat.

Publications

Joanna Moreton, Abril Izquierdo, Richard Emes (2016), "Assembly, Assessment, and Availability of De novo Generated Eukaryotic Transcriptomes". *Frontiers in Genetics*. **6**: 361. DOI: 10.3389/fgene.2015.00361

Alex Bàyés, Mark O. Collins, Rita Reig-Viader, Gemma Gou, David Goulding, Abril Izqui erdo, Jyoti S. Choudhary, Richard Emes, Seth G.N. Grant (2017), "Evolution of complexity in the zebrafish synapse proteome". *Nature Communications*. **8**: 14613. DOI: 10.1038/ncomms14613.

Abril Izquierdo, Martin Fahrenberger, Tania Persampieri, Mark Q. Benedict, Tom Giles, Flaminia Catteruccia, Richard D. Emes, Tania Dottorini (2018), "Evolution of gene expression levels in the male reproductive organs of Anopheles mosquitoes". *Life Science Alliance*. **2**: 1. DOI: 10.26508/lsa.201800191

Declaration

By this means, I declare that this PhD thesis was conducted in accordance with the requirements of the University of Nottingham Regulation and Code of Practice for Research Degree Programmes, and has not been, or will be submitted for any other academic award. The work presented hereby is my own and all work of other authors and material from other sources is properly recognized.

Abril Izquierdo

Acknowledgments

First of all, I like to express my deepest gratitude to Prof. Richard Emes for his patient guidance, motivation and excellent support throughout my life as a PhD student. Being under his supervision has been a wonderful experience. I could not have imagined a better mentor. I am also grateful to Dr. Lisa Chakrabarti for providing me with the brain samples used in my work, and her keen supervision in the laboratory part of my research.

A very special gratitude to Dr. Tania Dottorini for her tremendous enthusiasm and encouragement during my studies. Thank you for creating a delightful environment in the office. I appreciate the splendid support of the ADAC group; Dan, Tom, Jo, Niraj, and particularly Andrew Warry for the countless times of help to decipher R's error messages.

A special thank you to the Optics & Photonics Group, specially to Prof. Stephen Morgan, for taking me as an honorary member of his research group, his friendship and sharing special moments.

To all my friends at the Vet School; Necati, Purba, Grazi, Lamyaa, Ramon, Veronika. In a very special way; Lola Ruiz-Diaz, Aouatif Belkhiri and Ramzi Al-Agele for their genuinely friendship and valuable moments together that I would never forget. My best friends in Mexico; Gina, Evy, Ingrid and Tere.

I wish to extend my thanks to CONACyT (Consejo Nacional de Ciencia y Tecnología) for funding my project and permitting me to pursue my doctoral degree.

This journey would not have been the same without David Gomez, my partner and best friend during all my PhD. Thank you for your love, encouragement, support and all those memorable moments together.

Finally, I am profoundly thankful to my parents Sergio and Mayte for their vast love and support. My late big brother Sergio, your life has been an inspiration to me, and my dear brother Mauricio, I could not have had greater brothers. Thanks to my beloved grandmas. Without you all, I would never be here.

List of Figures

List of Tables

List of acronyms

AMPA ..... α-amino- 3-hydroxy-5-methyl-4-isoxazolepropionate

AD ..... Alzheimer's disease

CaMKII ..... Ca^{2+}/calmodulin-dependent protein kinase II

CNS ..... Central nervous system

CV ..... Coefficient of variation

Dlg ..... Disc large homologue

HMM ..... Hidden Markov Model

G2Cdb ..... Genes to cognition database

GKAP ..... Guanylate kinase-associated protein

GM ..... Generalized linear model

GRIP ..... Guanylate kinase-associated protein

GKAP ..... Guanylate kinase-associated protein

HD ..... Huntington's disease

iGluRs ..... Ionotropic glutamate receptors

LTD ..... Long-term depression

LTP ..... Long-term potentiation

MAGUK ..... Membrane-associated guanylate kinases

mGluRs ..... Metabotropic glutamate receptors

NGS ..... Next-generation sequencing

NMDA ..... N-methyl-D-aspartate

ORF ..... Open Reading Frame

PD ..... Parkinson's disease

PKC ..... Protein kinase C

PMCA ..... Plasma membrane calcium ATPase

PSD ..... Post synaptic density

PSD-95 ..... Post synaptic density 95

PSP ..... Post synaptic proteome

PrePSP ..... Pre-synaptic proteome

SD ..... Standard deviation

SNARE ..... Soluble N-ethylmaleimide-sensitive factor attachment receptor

SYN ..... Synaptosome

TPM ..... Transcripts Per Kilobase Million

TSGD ..... Teleost-specific genome duplication

VGSCs ..... Voltage-gated sodium channels

vPSD ..... vertebrate PSD

WGD ..... Whole genome duplication

Contents

Chapter 1

Introduction

“Let us understand what our own selfish genes are up to because we may then at least have the chance to upset their designs."

- Richard Dawkins -

The origin of synapses is a central event in the evolution of species, rebuilding the molecular pieces of the present puzzle, might shed light into the diversity of ingredients that make up a neuron different from other cell types, and at the same time, how these molecular kits have enabled animals to have remarkable cognitive skills. Thereby this chapter provides an introduction to the origin and evolution of the synapse, their role in the brain functioning, along with the main molecular machinery that makes all viable.

1.1 Origin and evolution of the synapse

Synapses represent the fundamental structures of the brain, and are key for the processing and transmission of information. Understanding how synapses originated is pivotal to comprehend how animals are able to make use of perception and forms of learning to adapt to their surroundings (Burkhardt, 2015). The synapse is a cellular apparatus embodied by the assembly of an interaction network of proteins to communicate between neuronal cells via electrical or chemical activity. During the latter, endogenous signals, i.e., neurotransmitters, are released from the presynaptic synapse to interact with receptors from the postsynaptic synapse transforming the chemical signal to an electrical impulse (Conaco et al., 2012). Thousands of regulatory and signaling proteins have been discovered in the mammalian pre- and postsynaptic synapses, known as the pre- and postsynaptic proteome (PrePSP and PSP), respectively (Emes et al., 2008).

While the process of learning and adaptation to different environments is a characteristic shared by all metazoans (Emes and Grant, 2012), invertebrates and even bacteria display elementary processes of cognition in the form of environmental stimuli and ontogenetic adaptation (Van Duijn et al., 2006). Proteins involved in environmental stimuli, cell-communication, cell-adhesion and cell-differentiation originated before the emergence of multicellularity (Ruiz-Trillo et al., 2007; King, 2004) (Figure 1.1).

Moreover, it is likely that modest forms of cognition observed in invertebrates, i.e., sensitization and habituation represent the foundation for more sophisticated processes present in vertebrates (Emes et al., 2008; Van Duijn et al., 2006).

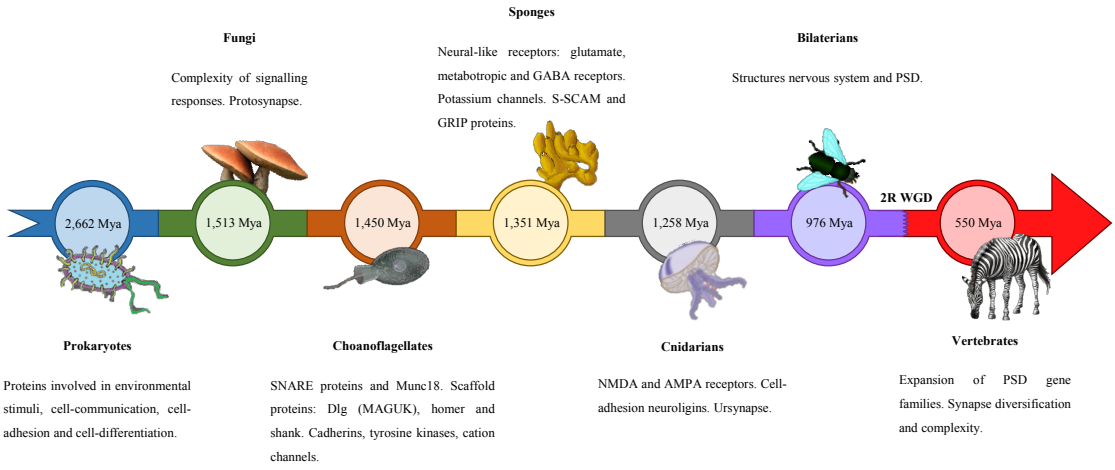


Figure 1.1: **Evolution of the synapse.** Timeline representing key taxonomic groups that gave rise to the synaptic origin (modified from Emes et al. (2008)). Dates denote the time (in million of years) of divergence.

Studies have shown high levels of protein conservation among unicellular and multicellular organisms. For example, the genome of the yeast *Saccharomyces cerevisiae* and amoeba *Dictyostelium discoideum* encode ancestral synaptic protein families, including PMCA (plasma membrane calcium ATPase) and protein kinase C (PKC) (Ryan and Grant, 2009). In metazoans, the former regulates excitatory synaptic transmission by controlling the neuronal calcium influx (Jensen et al., 2007), while the latter is fundamental for activity-dependent synaptic plasticity (Ramakers et al., 1997). The PSD (postsynaptic density) and MAGUK (membrane-associated guanylate kinases) are the main molecular machinery of excitatory synapses. Emes and Grant (2012) examined 570 mammalian PSD genes and 183 MAGUK scaffold proteins (MASC). 23% of these genes were found conserved in the yeast *S. cerevisiae*. Similarly, de Mendoza et al. (2010) identified scaffold MAGUK proteins in the protist *Capaspora owczarzaki* and choanoflagellate *Monosiga brevicollis*, but failed to find any of these proteins in fungi

and amoebozoans.

Comparing orthologs between divergent species and identifying a high degree of conservation is a potential indicator of shared function in equivalent pathways (Conaco et al., 2012). Using this approach, it has been possible to shed light on the proteins that were key at the origin of the synapse. Choanoflagellates, the closest living relatives of the animal kingdom represent splendid candidates (Carr et al., 2008). Several synaptic proteins formerly considered metazoan-specific have been found in this single-cell organism. For instance, *M. brevicollis* express cadherins and tyrosine kinases (Burkhardt, 2015). Both proteins are widely known to be involved in synaptogenesis and synaptic plasticity in metazoans (Abedin, 2010; Purcell and Carew, 2003). Yet, a role in environmental stimuli has also been suggested for tyrosine kinases. This is implied by the observation that the transfer of *M. brevicollis* from a poor nutrient environment to a rich one, produces a rapid phosphorylation (Ruiz-Trillo et al., 2007).

Rapid release of neurotransmitters from synaptic vesicles is achieved by the neurosecretory SNARE (soluble N- ethylmaleimide-sensitive factor attachment receptor) machinery protein family, which involves the expression of synaptobrevin 2, syntaxin 1 and SNAP-25, that are regulated by cystosolic proteins Sec1/Munc18 (Nouvian et al., 2011; Chen and Scheller, 2001). The choanoflagellates *M. brevicollis* and *Salpinogea rosetta* encode SNARE proteins and a Munc 18, serving as primitive neurosecretory machinery, in conjunction with PSD scaffolds; Homer, Dlg4 (PSD-95) and the Shank scaffolds (Burkhardt et al., 2014; Alié and Manuel, 2010; Emes et al., 2008). Further proteins identified in choanoflagellates include cation channels, which resemblance the voltage-gated sodium channels (VGSCs) responsible for action potentials, as well as a wide variety of metazoan plasma membrane calcium channels (Burkhardt et al., 2011).

The sponge *Amphimedon queenslandica* is considered the earliest branching surviving metazoan taxon, and therefore represents an additional interesting candidate to study

the evolution of the synapse. Sponges are capable for sensing and responding to their environment, however it is remarkable that although they lack a nervous system, their genome encodes the complete set of the mammalian PSP (Sakarya et al., 2007). Similar to choanoflagellates, sponges encode the PSD scaffolds Homer, Dlg and Shank (Kosik, 2009). However, these ancient animals encode a larger number of post synaptic scaffold proteins with an almost identical conservation of mammalian protein domains and ligands (Sakarya et al., 2007). Examples of genes in the sponge that are absent in yeast, fungi or choanoflagellates include orthologs of S-SCAM and GRIP (Ryan and Grant, 2009). Additionally, the expression of acetylcholine (ACh), a neurotransmitter in the cholinergic nervous system of metazoans has been identified in primitive forms of life, including; sponges, plants, bacteria and fungi, acting as a mediator (Horiuchi et al., 2003).

It is plausible that the fundamental features of synaptic transmission and adaptation that involves the backbone of the PSD scaffold, evolved before the origin of synapses and the appearance of the first metazoan. Possibly within the unicellular protozoan phylum Choanozoan during the transition to multicellularity (Emes and Grant, 2012). Nevertheless, genes that encode postsynaptic receptors, such as glutamate receptors (GABA and metabotropic), seemed to have evolved with the appearance of early metazoans, e.g., *A. queenslandica*, or preceding the evolution of cnidarians, such as the origin of N-methyl-D-aspartate (NMDA) and α-amino- 3-hydroxy-5-methyl-4-isoxazolepropionate (AMPA) receptors, together with cell-adhesion neuroligins (Sakarya et al., 2007). From above, it can be suggested that ionotropic glutamate receptors must have evolved in a cnidarian-bilaterian ancestor, prior to the origin of excitatory ionotropic glutamate receptors (Ryan and Grant, 2009).

It is likely that the first nervous system was first evolved in cnidarians or a closely related ancestor. This nervous system consist of a nerve net where collectively, glia cells and sensory, motor, and neurosecretory neurons are organized (Grimmelikhuijzen

and Westfall, 1995). This primitive nervous system permits complex active behaviours, particularly for feeding, which rely on coordinated movements. It is suggested that this nerve net correlates to simple forms of cephalization implicated in the coordination of behaviours (Miljkovic-Licina et al., 2004). What is more, it represents the last common ancestor of all synapses, the ursynapse (Emes et al., 2008).

On the other hand, the organization of proteins that characterizes the PSP protein-protein interaction networks (scaffold proteins, receptors, and enzymes) was already present in the genome of choanoflagellates, sponges and cnidarians. This arrangement of proteins represents the set of synaptic proteins which existed in ancient metazoans that lack a nervous system, the protosynapse (Emes and Grant, 2012), and it provides the basic signalling elements of synapses (Ryan and Grant, 2009). The identification of protosynaptic proteins lead to the conclusion that functional synapses evolved by the expansion of ancient genes and simple molecular machines with adaptation of regulatory pathways resulting in a coordinated neuronal expression (Conaco et al., 2012).

Diversification of the early synapse happened mainly by duplication or retention of duplicates of essential genes. It has been widely hypothesized that two rounds of whole genome duplication (2R WGD) took place around 550 Mya at the base of the chordate lineage (Dehal and Boore, 2005), during the evolution of vertebrates from early deuterostome ancestors. The first of which occurred in the Cambrian period (~510 Mya), while the second happened in the initial Devonian period (~400 Mya) (Catchen et al., 2009).

Drastic adaptive radiations and evolutionary innovations in the vertebrate genome have been associated to the 2R of WGD (Berthelot et al., 2014). Presumably, some gene families increased the number of ortholog genes, called onhologs, up to four times in vertebrates, compared to a single invertebrate protein (Dehal and Boore, 2005). Subse-

quently, in a process called fractionation, it is likely that most of these gene families lost one gene copy (Langham et al., 2004). For example, only 25% of the human genome are onhologues originated from the 2R WGD (MacKintosh and Ferrier, 2017).

Studies have demonstrated the involvement of most onhologs in significantly enriched processes, such as neuronal synapse development and function (Berthelot et al., 2014). To this extent, the complexity of a gene function is a determinant for retention (Guo, 2017), synaptic genes are retained at higher rates reflecting their molecular diversity and fitness (Bayés et al., 2017). Certainly, vertebrates have evolved novel synapse types and functions (Bayés et al., 2017). Examples of synaptic gene families with remaining onhologues are neurotransmitter receptors, such as GABA and glutamate, Dlg, cadherins, neuroligin, CaMKII (Ca^{2+}/calmodulin-dependent protein kinase II), PKC, guanylate kinase-associated protein (GKAP) and PMCA (Ryan and Grant, 2009).

Vertebrate independent duplications have also been observed in the MAGUK gene family and NMDA receptor. To illustrate, while MAGUKs originated before the evolution of metazoans; choanoflagellates and *Capsaspora owczarzaki* encode three classes of MAGUK proteins, i.e., DLG-like, MPP-like and MAGI-like. Later a MAGUK expansion occurred in the metazoan lineage, creating novel classes, i.e., CACNB, DLG5, DLG,CASK, ZO. However, vertebrates not only evolved two other MAGUKs, i.e., CARMA and MPP1 (de Mendoza et al., 2010), but also diverged various genes. For example, vertebrates expresses four *Dlg* genes (*Dlg1-4*), whereas, invertebrates encode a single *Dlg* gene. A similar case occurred with MAGI and ZO MAGUKs classes. From this, it has been shown that while *Dlg1* and *Dlg4* have likely retained its ancestral function, such as elemental forms of learning, *Dlg2* and *Dlg3* evolved to perform higher cognitive processes (Nithianantharajah et al., 2013).

The NMDA receptor, essential for regulating fast neurotransmission and synaptic plasticity (Sprengel et al., 1998), is a further example for vertebrate independent duplica-

tion. Chordate functional NMDA receptors consist of an NR1 subunit and at least one of four NR2 subunit types (NR2A to NR2D). These subunits are encoded by four genes, i.e., GRIN2A to GRIN2D, which have separate spatial and temporal expression patterns. In contrast, invertebrates have a single NR2 with an intracellular C-terminal domain five times smaller than the vertebrates (Teng et al., 2010). DLGs are the main scaffolding proteins that bind the C-terminal of the NR2 subunit. Since invertebrates have a single DLG gene and a shorter NR2 subunit, the number of interactions increased to 12-fold in vertebrates (Ryan and Grant, 2009).

Additional GRIN2 genes are expressed in teleost, i.e., GRIN2A-1 and 2, GRIN2B-1 and 2, GRIN2C-1 and 2, and GRIN2D-1 and 2 (Teng et al., 2010). In this respect, a further round of WGD took place ~300 Mya in the largest fish clade, called teleost-specific genome duplication (TSGD) (Bayés et al., 2017). As a result, teleosts have a larger number of protein-coding genes than any other vertebrate, eg., the zebrafish has 26,206 protein-coding genes (Howe et al., 2013a), whereas the human and mouse possess 19,042 and 20,210, respectively (Church et al., 2009). Enriched gene ontologies for zebrafish onhologs are neural activity and transcription factors (Howe et al., 2013a). Yet, even though zebrafish encode a larger number of synaptic proteins, its PSD complexity is lower when compared to that of the mouse (Bayés et al., 2017).

WGD events are key elements for the transition of a more sophisticated brain (Bayés et al., 2017). In this way, the synaptic proteome boosted its complexity through gene family duplication and diversification, rather than the formation of new protein types (Emes et al., 2008). Remarkably, it seems that the latest mam malian synaptic proteome integrations, are those that promoted divergence, implying a correlation between complexity and diversity (O'Rourke et al., 2012). In general terms, the evolution of complex molecular machineries entails the assemblage of numerous proteins, where each of them adds functionality to the integrated complex (Sakarya et al., 2007). The evolution of the synapse from eukaryotes to metazoans, and thereafter to chordates, overlies the

presence of multiple protein populations. This has not only permitted the diversification of the chordate brain functioning (Ryan and Grant, 2009), in the same way, varied forms of cognition have arisen.

1.2 Synapses and neurons

The nervous system is specialized to receive information from the environment and transform it into biological actions producing behaviour. It is arranged into two components, the central nervous system (CNS), i.e., brain and spinal cord that determines behaviour, and the peripheral nervous system (PNS), i.e., nerves and most sensory organs.

The neuron is the information-processing and information-transmitting basis of the nervous system. Whilst neurons have different forms depending on the specific role that they play, overall, they are comprised of the following structures: soma or cell body, dendrites, axon, and terminal buttons. The nucleus is contained in the soma, from which numerous dendrites extend. The latter, captures and propagates signals from other neurons. Axons, are long lean tubes that transmit information from the soma to the terminal buttons (Hughes, 2007). Neurons communicate between each other via synapses. The synapse is a subcellular arrangement constituted by a protein collection from the PreSP and PSP. The role of the synapse is to receive, process and transmit signals by the identification of distinct neural patterns from electrical activity, and transform it into intracellular biochemical cascades that alter the neurons properties (Collins et al., 2006). The presynaptic cell initiates the signal, while the postsynaptic neuron receives it within the extracellular space, the synaptic cleft (Kandel et al., 2000).

Signals are first transmitted via electrical events, known as action potentials that ini-

tiate at the edge of the presynaptic neuron (or axon, generally), close to the soma, heading to the terminal buttons. When an action potential reaches this point, depolarization of the membrane occurs and voltage-gated Ca^{2+} channels are activated. Differences in concentration between the inside and outside of the cell, drives the influx of Ca^{2+} (Kandel et al., 2000). Chemical neurotransmitters, which are retained in subcellular organelles called synaptic vesicles, are released from the active zone of the presynaptic terminal to the synaptic cleft, and eventually bind specific receptors at the postsynaptic membrane (Kandel et al., 2000).

1.3 The postsynaptic proteome (PSP)

Information from the outer environment is processed by the nervous system, which resolves learning and memory by molecular signalling networks in the postsynaptic terminal of synapses (Emes and Grant, 2011). Both the PreSP and PSP make up the complete synaptic protein network (Figure 1.2). The PreSP is primarily constituted by the vesicle exocytosis mechanism, by which neurotransmitters are released (Raiteri, 2001). Yet, the PSP has been most widely investigated, as it holds the signalling and neurotransmitter receptor machinery that underlies the overall synaptic functioning.

The PSP is remarkably complex with a large range of cellular autonomy. It is embodied by an assortment of protein classes. Whilst a minority of the PSP are neurotransmitter receptors, most of this protein repertoire is associated with a wide arrangement of signalling, adhesion, metabolic, structural, trafficking and regulatory activities (Roy et al., 2018; Emes et al., 2008). Mass spectrometry has recently been used to characterize the human, mouse, rat and zebrafish PSP, revealing a collection of ~1,000 highly conserved proteins (Roy et al., 2018), along with lineage-specific elements, which have promoted functional diversity (Bayés et al., 2017). Moreover, mutations affecting the PSP functioning have been linked to more than 130 brain diseases. It is therefore important

to recognize the extent by which the PSP functioning underlies cognition and disease (Bayés et al., 2011).

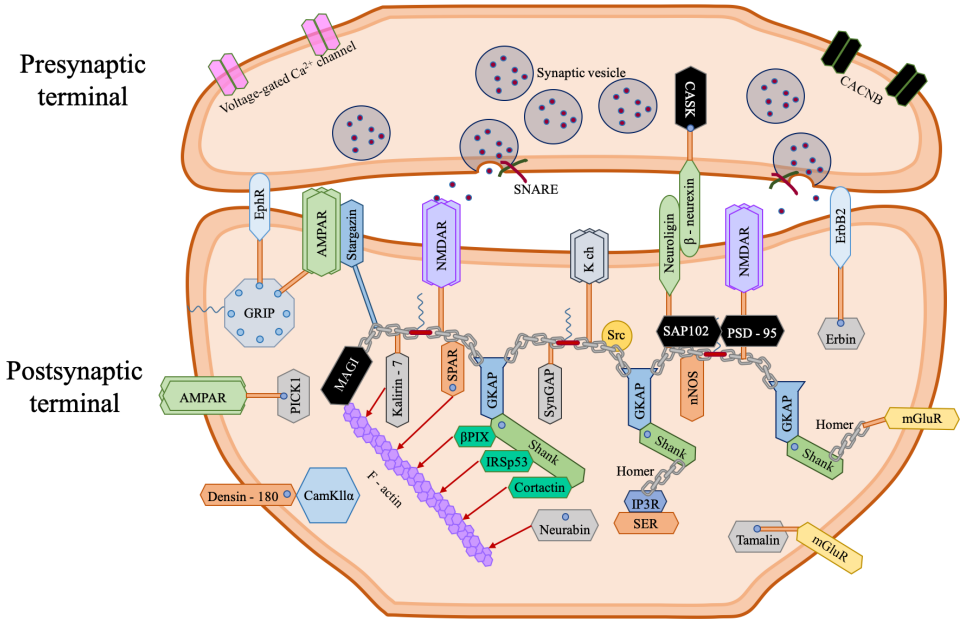


Figure 1.2: **Schematic illustration of the chemical synapse.** Chemical synapses envelop complex molecular machinery from the presynaptic terminal that mediate neurotransmitter release to the post-synaptic terminal upon depolarization, followed by a sophisticated cascade of multiprotein complexes that depending on the impulse (inhibitory or excitatory), fire an action potential. MAGUK proteins are depicted in black (modified from Kim and Sheng (2004)) .

One metazoan feature is the adaptive ability of neural tissue to change its organisation in response to different stimuli (Kolb and Whishaw, 2001). This feature is known as neuroplasticity, and it reflects activity-dependant changes in the intensity of neural activity, through strengthening, weakening, eliminating or creating novel synaptic connections (Pascual-Leone et al., 2011). Long-term potentiation (LTP) and long-term depression (LTD) are the most important molecular phenomena of synaptic plasticity, both of which represent the foundation for human memory (Ohno et al., 2011). Certainly, neuroplasticity constitutes the keystone of learning and memory, behaviour and mental illness (Pocklington et al., 2006). Yet, its molecular basis relies on PSP receptors, such as NMDA and mGluRs (metabotropic glutamate receptors) (Grant, 2006). Several

studies have suggested a significant role of PSPs in regulating the number of AMPA receptors at the synapse (Chater and Goda, 2014). However, the molecular machinery of cognition that underlies neuroplasticity, is yet to be fully elucidated (Grant, 2006).

PSD-95, a protein confined in the postsynaptic terminal at the postsynaptic density (PSD), binds to neurotransmitter receptors and ion channels to build signalling pathways that mediate neuroplasticity (Fernández et al., 2009; Dosemeci et al., 2007; Kornau et al., 1995). The PSD is an electron-dense specialized protein organization that includes the neurotransmitter receptor machinery of excitatory synapses (Emes and Grant, 2012; Collins et al., 2006). It is confined to the cytoplasmic part of the postsynaptic terminal membrane, opposite the active zone of the presynaptic terminal (Ziff, 1997). Functioning of the PSD widely depends upon the assembly of multiple protein classes, namely; cell-adhesion, cytoskeletal, scaffolding and adaptors, membrane-bound receptors, G-proteins, and signalling proteins (Böckers, 2006).

Neurotransmitter receptors in the PSD are physically linked by the PDZ domain of several scaffold proteins. For example, as mentioned previously, the PDZ domains of the Dlg protein family, which are PSD scaffolds (including PSD-95), interact with the C-terminl of the NR2 subunit of the NMDA glutamate receptor. Simultaneously, Dlgs bind to cytoplasmic signalling proteins. The assemblage of such complexes eases the binding of postsynaptic receptors to trafficking apparatus and downstream signalling mechanisms. This in turn, regulates synaptic strength, cytoskeletal rearrangements, and cellular responses (Cheng et al., 2006; Sheng and Kim, 2002). Studies have revealed 77 proteins involved in the NMDA receptor-PSD-95 complex (Husi et al., 2000). The PSD contributes to crucial roles of synaptic integration and regulation, together with neuroplasticity. Therefore, extensive efforts are centred to identify its protein constituents to further analyse its subcomponents and complexes (Grant, 2012; Bayés et al., 2017, 2011; Cheng et al., 2006; Walikonis et al., 2000).

The mammalian PSD roughly consist of 1,500 proteins (Grant, 2012). A large subset of these include MAGUKs, which coordinate the complex signalling machinery at the PSD (Reese et al., 2007). MAGUKs are scaffold proteins that possess significant protein-binding domains that cluster receptors, enzymes, NMDA and AMPA receptors at the excitatory synaptic core (Emes and Grant, 2012). MAGUKs regulate signal transduction by the coordination of multimolecular complexes at particular location in the membrane (Funke et al., 2005). The expression of MAGUKs is largely expanded in the brain and highly conserved throughout the metazoan evolution. Notwithstanding the large differences in size, each member of the MAGUK family (excluding MAGI) share a common structural core; an N-terminal cluster of three PDZ domains, a SRC homology 3 (SH3) domain and a guanylate kinase-like (GK) domain at the C terminus (Oliva et al., 2012; Zhu et al., 2011). Through multiple MAGUK-PDZ domain-containing proteins, a complex assemblage of several ion channels, such as K^+ and NMDA receptors that propagate glutamate responses is formed (Ziff, 1997). Knockout mouse studies in MAGUK proteins have unveiled numerous synaptic mutants that disrupt the PSD functioning and cause an altered cognition and mental illness (Emes and Grant, 2011; Nithianantharajah et al., 2013).

PSD-95, also known as Dlg4 and SAP-90 is the most studied protein from the MAGUK family and the most abundant PSD protein (Cheng et al., 2006; Kim and Sheng, 2004). It's main function is the recruitment of glutamate receptors at the postsynaptic terminal (Südhof, 2008). PSD-95 interacts with diverse proteins linking cytoplasmic signal transduction proteins and surface receptors (Verpelli et al., 2012). It is largely found in the forebrain, in the postsynaptic membrane, and in the presynaptic cerebellar basket cells (Elias and Nicoll, 2007). While DLG-MAGUK proteins are the prototype of the P-SP (Emes and Grant, 2012), a lack of expression at the presynaptic terminal halts the clustering of the calcium channel DMCa1A subunit, interrupting short-term plasticity (Astorga et al., 2016). PSD-95 binds directly to the NR2 subunit of the NMDA receptor via two PDZ domains and the cytoplasmic tail of Shaker-type voltage gated potassium

channels (Stathakis et al., 1997). Another class of transmembrane protein that binds directly to the PDZ domain of PSD-95 is neuroligin (NLGN). The latter, in conjunction with its ligand, neurexin (NRXN) are the best-characterized synaptic cell-adhesion molecule (SynCAMs) (Verpelli et al., 2012). This cell-adhesion proteins are pivotal to communicate the presynaptic and postsynaptic terminals, regulate signalling over the synapse, and frame the neural networks features. Disruption of the neurexin and neuroligin functioning is associated with cognitive diseases, such autism (Südhof, 2008).

The PSD-95-like subfamily includes PSD-93 (Chapysin-100 or Dlg2), SAP-102 (Dlg3) and SAP-97 (Dlg1). All of which anchor NMDA receptors to the submembrane cytoskeleton, contributing to signal transduction complexes at postsynaptic sites (Niethammer et al., 1996). At the PSD, PSD-95 regulates the amount of AMPARs interaction with other proteins, including Stargazin (Chen et al., 2015). Numerous studies have shown that an overexpression of PSD-95 increases the amplitude of AMPA receptors during excitatory postsynaptic currents (EPSC) without any changes in the rate of NMDA-EPSC. In the hippocampus, it raises the frequency and amplitude of miniature EPSCs. Overexpression of PSD-93 and SAP-102 also increases AMPA receptor EPSCs. Hence, a critical role of the Dlg protein family in synaptic trafficking of AMPA receptors is implied (Schnell et al., 2002). On the other hand, SAP-97 binds directly with the AMPA receptor subunit GluR1. During low levels of synaptic AMPAR, e.g., early development or PSD-95 disruption, SAP-97 can compensate for other MAGUKs (Howard et al., 2010).

In addition to the Dlg-like protein family, the scaffolds Shank and Homer also make up the PSD core (Hayashi et al., 2009; Sala et al., 2001). The Shank family contains various interaction domains, beginning with ankyrin-repeats adjacent to the N-terminal, an SH3 domain, a PDZ domain, a large proline-rich domain, and a SAM (sterile alpha motif) domain at the C-terminal (Sheng and Kim, 2000). Among the numerous protein-protein interactions of the Shank family, the most important is the one with a

PSD-95-associated protein, i.e., guanylate kinase-associated protein (GKAP) through it's PDZ domain. As a result, the complex NMDAR-PSD-95 is interplays with the Shank family (Naisbitt et al., 1999).

Unlike PSD-95, which is restricted to the neighbouring postsynaptic membrane, Shank is localized well immersed within the PSD, nearby its cytoplasmic face (Valtschanoff and Weinberg, 2001). In this manner, Shank might also function as an integrator of the PSD with the postsynaptic cytoplasm and cytoskeleton. A good example in this regard, is the interaction of Shank with cortactin, an F-actin binding protein abundant in dendritic spines that in response to extracellular stimuli, it is relocated in the cytoskeleton (Naisbitt et al., 1999). Three members make up the Shank family; Shank1, Shank2 and Shank3. The former two are broadly highly expressed in the brain, mainly in the cortex and hippocampus. Shank3 lacks the SH3 and ankyrin repeats domains, and it is more abundant in heart tissues (Alié and Manuel, 2010; Sheng and Kim, 2000). Studies have found that mutations in any of the three Shank members are related with neurodevelopmental and psychiatric disorders, including schizophrenia and autism (Bliss et al., 2014).

Shank is considered the "master scaffold" for bringing together NMDAR, mGluR and AMPAR complexes at the postsynaptic terminal (Kim and Sheng, 2004). Members of the Homer family bind Shanks proline-rich region via a single EVH1 Homer domain. They are largely sited at the mammalian PSD where they act as adapter proteins for numerous PSD proteins. Moreover, Homer is essential for an effective by mGluRs and IP3 receptors (Shiraishi-Yamaguchi and Furuichi, 2007). Knockout mice have linked the loss of Homer to behavioural anomalies, such as schizophrenia (Shiraishi-Yamaguchi and Furuichi, 2007). Recently, a study found decreased expression levels of Homer1 in the cingulate gyrus gray matter of patients diagnosed with schizophrenia, bipolar disorder and major depression. Thereby, sustaining the functioning of Homer proteins as role models in neuropsychiatric research (Leber et al., 2017).

Certainly, scaffold proteins at the PSD are central in the synaptic architecture and functioning, such as; trafficking, anchoring, clustering of glutamate receptors and adhesion. Even though that these proteins lack of enzymatic activity, are shaped by modular and specific domains capable of creating robust protein networks and molecular complexity (Verpelli et al., 2012). Alteration of these proteins gave rise to the phenotypes of at least 130 neurological diseases (Bayés et al., 2011). There is no other collection of proteins that configure the nervous system that can induce such a great number of diseases than the PSD (Grant, 2012).

Collectively the established literature suggests that the core synaptic proteins are evolutionary conserved in taxonomically diverse vertebrate species (Emes and Grant, 2012). This thesis tested the hypothesis that conservation is also seen in the transcriptome from different brain tissues of the zebrafish, bat and lion. To test this hypothesis, data was compared to mouse genome and proteome datasets. Mouse was chosen as it is the most extensively studied species within the comparative Genes to Cognition database of proteins relating to synaptic function and cognition (Croning et al., 2008) G2C.

In addition, this thesis tested a number of hypothesis; 1) Chapter 4 determined if *de novo* transcriptomics could be used to provide a comprehensive analysis of neural transcripts in novel sequenced species, 2) Chapter 4, 5 and 6 tested if lineage-specific gene duplications dominate the evolution of transcripts encoding the synapse and post synaptic density, 3) in Chapter 5 phylogenetics was used to test proposed convergent evolution of genes associated with echolocation.

Chapter 2

De novo transcriptome assembly

“The good thing about science is that it’s true whether or not you believe in it."

- Neil deGrasse Tyson -

This chapter presents a focused review scientific article that was published in the Frontiers in Genetics journal in 2016. This paper was an invitation from The Associate Editors to introduce a rather novel research topic, which in turn comprises a significant part of this PhD subject, which is "*De novo* transcriptome assembly". The review paper was written by Jo Moreton, a collaborator from the Advanced Data Analysis Centre (ADAC) at the University of Nottingham, Richard Emes, head of ADAC and Abril Izquierdo, author of the present thesis.

The present review paper, introduces the reader to the methodology of *de novo* transcriptome assembly (See Appendix A). The transcriptome represents the absolute estimated arrangement of transcripts in a particular cell type or tissue. The analysis has turned into a fundamental part not only for basic research, but also for major clinical studies. It promises to elucidate a comprehensible understanding of the complexity of gene expression in desired cell-types or tissues and in a desired organism, at an unsurpassed resolution. However, the correct assembly of a transcriptome might be challenging and entails an appropriate technique that can successfully piece together billions of sequencing reads as accurately as possible. This scientific article describes and compares the two main strategies for transcriptome assembly, along with the subsequent steps of this technique, i.e., quality assessment, annotation and its optional further availability.

Formerly, the study of a transcriptome depended only on techniques such as cloning cDNAs, EST (expressed sequence tag) and microarrays, which are still currently very useful, but not for large-scale analyses. The emergence of NGS (Next-Generation Sequencing) offered the exploration of the transcriptome in a much more cost-effective manner and in a remarkable degree of sensitivity and precision. Generally, reference-based transcriptome assembly is the preferred method used when a well annotated model organism is available. It relies on the correct alignment of the sequencing reads to a reference genome and then the overlapping alignments are assembled into tran-

scripts. This method has various advantages, such as; is highly sensitive, does not require a large computational load, sequencing artefacts do not represent a major issue, low-abundance transcripts are precisely assembled and it also allows the identification of novel transcripts that are not present in the genome model. However, in many cases a reference genome is not available, or lacks quality, therefore, *de novo* transcriptome assembly is a highly advantageous strategy that aims to rebuild overlaps among sequencing reads and assembles them into full-length transcripts. This approach usually makes use of a mathematical algorithm, called *De Bruijn* graph, which delimits a node by a specific number of nucleotides, termed *k-mers*, these are then linked by overlapping edges except for one nucleotide (*k*-1), which in turn overlaps another *k-mer*. These overlapping *k-mers* are condensed into a single linear string providing all the potential alternatives by which a string can be reconstructed. *De novo* assembly requires a higher sequencing depth (≥30x coverage), in conjuction with a great deal of computational work.

There are a number of software for *de novo* reconstructions that tackle different difficulties, yet a faultless algorithm does not yet exist. As a result, it is vital to assesses the quality of the transcriptome to eliminate in this way all possible errors in the data that might affect the interpretation of the transcriptome analysis. There are several approaches that are very useful and efficiently detect and removes aberrations in the assembly. Thereafter, the transcriptome is set for functional annotation, granting the comprehension of the desired study.

Chapter 3

Materials and methods

“Equipped with his five senses, man explores the universe around him and calls the adventure Science."

- Edwin P. Hubble -

Chapter 3 covers the general methodology workflow used for each of the species transcriptome assembly and analysis. A wet lab section is included, and the bioinformatics stepwise pipeline that represents the core of the thesis. Methodology was performed by myself, excluding the following; **a)** Brain tissue RNA extraction, except for *Panthera leo*, **b)** mRNA deep sequencing which was conducted at the University of Nottingham sequencing service (DeepSeq). Contribution by others is clearly stated in the corresponding chapters. Moreover, it is worth mentioning that although over the past decade efforts have been made to raise the availability of computational resources to assemble a transcriptome *de novo*, it is complex to determine the accuracy of these methods. In consequence, the described methodology reflects a process of comparing the performance of various metrics (many of them are not mentioned), to finally allow us to conclude which methods achieve the most robust results. See Table 3.3 for a summary of *in silico* tools applied in this research.

3.1 Ethics and source of tissues

This study has been approved by the School of Veterinary Medicine and Science Ethics and Clinical Review Panel, University of Nottingham (ERN# 2752 190520). All zebrafish tissues were obtained from Dr Martin Gering, School of Life Sciences, University of Nottingham, Queen's Medical Centre, Nottingham, NG7 2UH. Zebrafish were humanely killed by a trained individual using an approved schedule 1 method, in full accordance with UK Home Office guidelines. Bat and lion brain tissues were obtained from animals that had died from natural causes at an adult age. Bat carcasses were obtained from the West Yorkshire Bat Hospital with permission from Natural England. The lion was a female obtained from the Twycross Zoo, United Kingdom. Post mortem examination of the lioness was conducted at the Veterinary Pathology Service, School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, LE12 5RD. All brain samples were frozen immediately and transported on dry ice to the laboratory where tissues were dissected with the aim of dissection instruments.

Table 3.1: **Brain tissues prepared for RNA extraction.**

Specie	Tissue	Total
Zebrafish	whole brain	4
Zebrafish	telencephalon and olfactory lobe	1
Zebrafish	optic lobe	1
Bat	cortex (left and right cerebral hemispheres)	1
Bat	brainstem	1
Bat	cerebellum	1
Lion	forebrain (Figure 3.1)	2
Lion	brainstem (Figure 3.1)	2

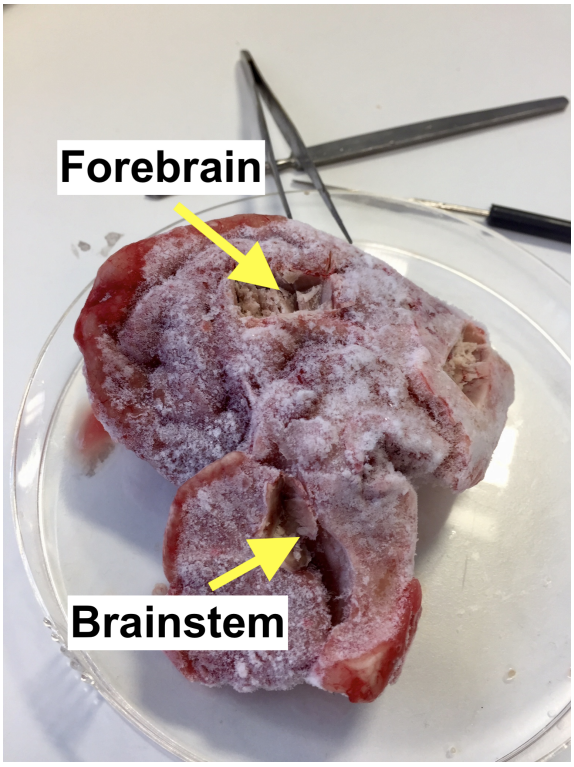


Figure 3.1: **Dissection of lion brain tissues**. Whole brain of a lioness was dissected into two samples of forebrain and two samples of brainstem utilizing dissection instruments .

3.2 RNA extraction

Tissues were collected from whole brains frozen at -80°. Brain tissues are rich in lipids, for this reason, RNA isolation can be challenging. For the wet lab in the present research, it was used a protocol developed by Dr. Lisa Chakrabarti at the University of Nottingham) to improve RNA isolation.

3.2.1 Sample homogenization

Total RNA for each tissue was extracted following the TRIzol Reagent by Life Technologies approach. All samples were weighted and deposited in gentleMACS^{TM} M tubes. 1 mL of TRIzol Reagent (Ambion^{TM} Life technologies, USA) per 50-100 mg of brain tissue was added. Homogenization was conducted using the gentleMACS Dissociator (Miltenyi Biotec, Germany).

After the homogenization, due to the rich content of fatty acids, all samples presented an evident fat monolayer. To minimize this contamination, an additional centrifuge step was included. Samples in the gentleMACSTM M tubes were centrifuged (Allegra X-22 Series-Beckman Coulter, Inc) at 12,000 g for 10 minutes at 4°C. In each sample, a supernatant containing the RNA, along with lipid layer on top of the supernatant, and a pellet that consisted of an extra cellular matrix, polysaccharides and DNA could be seen. The lipid layer was easily removed, and the supernatant was transferred to a 2 mL Eppendorf tube.

3.2.2 Phase separation and isolation

Samples were incubated at room temperature for 5 minutes. Next, 0.2 M of chloroform (BHD PROLABO, Belgium) per 1 mL of TRIzol Reagent (Ambion^{TM} Life technologies,

USA) was added. Samples were gently agitated by hand for 15 seconds, and incubated again for 3 minutes at room temperature. Subsequently, a centrifugation step at 12,000 g for 15 minutes at 4°C (Microcentrifuge 5417R, Eppendorf) was carried out. The aqueous phase was carefully transferred into a new tube.

0.5 mL of 100% isopropanol (Fisher Scientific, UK) per 1 mL of TRIzol Reagent (Ambion^{TM} Life technologies, USA) was added before centrifugation for 10 minutes at 12,000 g at 4°C (Microcentrifuge 5417R, Eppendorf). The supernatant was removed, and the pellet was washed using 75% of ethanol per 1 mL of TRIzol Reagent (Ambion^{TM} Life technologies, USA).

3.2.3 Resuspension

The samples were vortexed, and then centrifuged at 75,000 g for 5 minutes at 4°C (Microcentrifuge 5417R, Eppendorf), discarding any liquid. Pellets were left to air dry for 10 minutes at room temperature. At the end, each sample was eluted in 40 micro litres of RNAse-free water and pipetted thoroughly. For 15 minutes all samples were incubated in a heat block set at 60°C (QB Series Dry Block Heating Systems, Grant Instruments) and stored at -80 °C.

3.2.4 RNA quantitation

RNA quality and quantity was assessed by spectroscopy (NanoDrop 8000, Thermo Scientific) employing OD260 to estimate the concentration, along with the ratios A_{260}/A_{280} and A_{260}/A_{230}. Samples with ratios greater than 1.80 and 2.0 respectively, were considered to be of satisfactory quality. RNA integrity was also evaluated with the Agilent 2100 Bioanalyzer, provided by The University of Nottingham Immunology Division, A floor, west Block, Queen's Medical Centre.

Table 3.2: **RNA quality and quantity for all used tissues.**

Tissue	RIN	Concentration	μl sent for RNAseq
Zebrafish whole brain	8.7	274ng/μ	10
Zebrafish whole brain	9.0	266ng/μ	10
Zebrafish whole brain	9.0	206ng/μ	10
Zebrafish whole brain	8.9	217ng/μ	10
Zebrafish telencephalon and olfactory lobe	8.6	202ng/μl	40
Zebrafish optic lobe	8.7	35ng/μl	40
Zebrafish cerebellum	8.5	41ng/μl	20
Bat cortex	8.7	87ng/μl	10
Bat cerebellum	8.6	224ng/μl	20
Bat brainstem	8.4	87ng/μl	10
Lion forebrain	8.9	167ng/μl	20
Lion forebrain	8.5	160ng/μl	20
Lion brainstem	8.8	95ng/μl	20
Lion brainstem	8.6	94ng/μl	20

3.3 RNA sequencing

Construction of RNA-Seq libraries and sequencing were performed using Illumina NextSeq500 sequencing platform at the Deep Seq Next Generation Sequencing Facility of the University of Nottingham, Queen's Medical Centre, who provided the following information.

1 μgof Total RNA was used for enrichment of mRNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490). Illumina stranded whole transcriptome sequencing libraries were prepared using NEBNext Ultra Directional RNA library prep kit for Illumina (NEB, E7420S) and the NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1) (NEB, E7600S). Library quality control was performed using bioanalyser HS kit (Agilent biotechnologies, 5067-4626). Libraries were next quantified using qPCR

(Kapa Biosystems, KK4824) and pooled at desired concentrations. Finally, denaturing and loading for sequencing were perfomed according to manufacturer's instructions. Sequencing was achieved on the Illumina NextSeq500 sequencing platform to generate 2 x 150bp reads.

3.4 Transcriptome assembly

3.4.1 Quality Control of short reads

A common problem in this type of experimental data is the presence of sequencing adapter fragments in the short reads. These might mislead the alignment and complicate the assembly. Trimming of adapter fragments, along with poly-A tails, primers and other types of sequencing contaminants, was achieved with Cutadapt v1.9 (Martin, 2011). Both 5' and 3' sequence adapters were specified and removed (using -b and -B options). Additionally, before trimming, low-quality ends from reads were filtered out with a cutoff of 10 (-q option). Reads shorter than 50 bases were also removed after trimming (-m option). Below is an example of a bash code used to run Cutadapt showing the utilized parameters:

```
> cutadapt -q 10 -b AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -B AGATCGGAAGAG
CGTCGTCGTGTAGGGAAAGAGTGT -m 50 R1.fastq.gz R2.fastq.gz -o R1.trimmed.
fastq -p R2.trimmed.fastq
```

Since, both RNA ends were sequenced, there were two FASTQ files per sample, paired read 1 and 2.

3.4.2 *De novo* Transcriptome assembly

Trimmed reads were *de novo* assembled, therefore no reference genome was used to map the reads. This was carried out by Trinity v2.1.1 (Haas et al., 2013). Trinity pieces together short reads of length k (k-mer) that overlap. The software incorporates three

different software: Inchworm, Chrysalis, and Butterfly that together attempt to rebuild full-length transcripts including alternative spliced isoforms.

Trinity was installed on a Linux workstation (32 CPUs and 189 GB RAM) with the prerequisite Bowtie v1.1.2. Further plug-ins for transcriptome downstream analysis, such as RSEM, were built by the command "make plugins". Trinity was run with flags indicating the paired-end nature of the reads. In addition, since the assemblies involved a considerable number of short reads, the minimum *k*-mer coverage was set to 2, i.e., singleton *k*-mers are not integrated in the initial contigs, therefore reducing the chances of attaining transcriptional "noise". The rest of parameters were set as default. Below is an example of a code used to run Trinity.

```
> Trinity --seqType fq --max_memory 30G --left R1.trimmed.fastq --rig
ht R2.trimmed.fastq --min_kmer_cov 2 --CPU 6
```

3.5 Post-assembly assessment

Once the assembly was accomplished, a measure of its accuracy and completeness was required. Note that each assembly represented a single sampled tissue.

TransRate v1.0.1 (Smith-Unna et al., 2016) was used to map all short reads to the *de novo* assemblies. Quality statistics were generated per contig to evaluate the entire assemblies. Only well-supported transcripts that accurately mapped reads were integrated to generate an improved assembly.

TransRate took as input the Trinity-assemblies in FASTA format, along with left and right paired-end reads already trimmed by Cutadapt. The program was run as the following command line code for each sample:

```
> transrate --assembly trinity_out --left R1.trimmed.fastq --right R1
.trimmed.fastq --output transrate_out
```

Utilizing the optimized assembly generated by TransRate, transcripts whose abundance was estimated to be less than 0.5 TPM were removed. For this, a perl script was utilized (See Appendix B and available here through a github repository), which in turn makes use of Salmon v0.4.2 (Patro et al., 2015).

To improve the quality of the final assembly and reduce redundancy, each sample was used to generate an independent assembly. These were then combined and filtered to identify transcripts generated in multiple independent assemblies. Transcripts from each assembly were named with a unique identifier and pooled. CD-HIT-EST was used to identify transcripts from multiple assemblies. The longest transcript in each cluster was then used for final mapping for isoform identification and transcript quantification.

Assemblies for each sample were combined together and then grouped based on sequence identity by CD-HIT-EST v4.6 (Li and Godzik, 2006). The assembly was sorted by sequence length, and clustered by similarity with a threshold of 90% of sequence identity (-c option). Cluster similarity was estimated based on identical "short word" algorithm that grouped together similar substrings of bases, setting this number to 10 (word length, -n option). Short sequences in each cluster were removed from the assembly:

```
> cd-hit-est -i combined.assembly.fasta -o combined.assembly.cdhit.fa
sta -c 0.9 -n 10  -T 6
```

Following clustering, a greedy re-filtering was completed considering cluster size. Using a perl script (See Appendix B and accessible through a github repository available here) to generate an assembly of the longest transcript sequences per cluster. This

master non-redundant assembly served as the transcriptome model to measure the proportion of reads which mapped back to the transcriptome.

3.5.1 Reads Mapped Back to Transcript

Hisat v2.0.4 (Kim et al., 2015) was initially used to create a set of systematic indexes from the generated non-redundant master assembly (hisat2-build command). This in turn, provided accurate genomic regions to aim the mapping of the entire set of trimmed paired-end short reads. Hisat2 was run using default parameters:

```
> hisat2-build -p8 master.assembly.fasta master.assembly.index
> hisat2 -p8 -x master.assembly.index -1 R1.trimmed.fastq -2 R2.trimm
ed.fq -S sample.SAM
```

SamTools v1.3.1 (Li et al., 2009) was used to parse and manipulate the alignments files. The SAM files were initially converted to BAM files (samtools view) and then coordinate sorted (samtools sort), an example of the used bash code is:

```
> samtools view -bS sample.BAM sample.SAM
> samtools sort sample.BAM > sample.sorted.BAM
```

The computed sorted read alignments were next assembled into potential, full and partial-length transcripts, including a variety of isoforms by StringTie v1.2.3 (Pertea et al., 2015). The computed assemblies were merged together (-merge option) to generate a single GTF (General Feature Format) file. This way, uniformity was achieved across every sample for subsequent analyses. A prefix was added to the transcripts (-l option). Finally, this file was used to feed a final assembly step to output accurate transcripts, in conjunction with their expression levels.

```
> stringtie sample.sorted.BAM -o sample.GTF
> stringtie --merge path_gtf.txt -o Merged.GTF -l UoN.deNovo -T 0.5
> stringtie sample.sorted.BAM -o sample.GTF -G Merged.GTF -A sample.a
bundances.txt -b sample.ctables
```

Notice that StringTie was run a total of three times. The first run, was to assemble the aligned reads and output a GTF file per sample. The purpose of the next StringTie run, is to create a consistent general GTF that incorporates every single GTF file.

As input, a text file containing the directory paths of every single GTFs was used. In this step, transcripts were labelled as the University of Nottingham short name "UoN" using a perl script (See Appendix B and available here), and a filter of minimum expression was added (-T option), i.e., all transcripts with expression levels less than 0.5 TPM were filtered out.

The third StringTie run re-assembled all transcripts utilizing the merged GTF as the reference annotation. Tab-delimited transcript tables containing coverage data were computed (-A and -b options) for further differential expression tests. Lastly, transcript abundances in TPM (Transcripts Per Kilobase Million) units were also estimated using two methods; StringTie and Kallisto v0.43.0 (Bray et al., 2016). The later computed transcript levels using a 'pseudo alignment" method that has been reported to decrease noise in quantification.

Transcript abundance in TPM units is computed as:

$$TPM = \frac{r_g * rl * 10^6}{fl_g * T} \tag{3.1}$$

Where r_g is the number of reads that mapped to a gene region, multiplied by *rl* (read length); the average number of nucleotides mapped per read, multiplied by 10^6 (kilo-

base scaling factor). Divided by fl_g (feature length), which is the number of nucleotide in a mapable gene region, multiplied by T, the total number of transcripts sampled in a RNA sequencing run (Wagner et al., 2012).

Kallisto was run with default parameters in two steps, which involved construction of an index using the non-redundant master assembly as the reference genome, and the quantification algorithm.

```
> kallisto -i kallisto.index master.assembly.fasta
> kallisto quant -i kallisto.index -o sample.abundance.txt R1.trimmed.
fastq R2.trimmed.fastq
```

3.6 Additional assessment of the transcriptome

Additional quality and completeness assessment of the transcriptome was conducted with not filtering of the data. This procedure was carried out after the initial Trinity-generated assembly, and before transcriptome annotation.

Examination of the quality was performed by exploring the number of *de novo* that display full-length or close to full-length. Transcripts were aligned to a known database (Uniprot or a reference genome) using BLAST (Camacho et al., 2009) and the portion that each transcript aligned to a known protein was determined.

```
> blastx -query assembly.fasta -db uniprot.fasta -out blastx.out
 -evalue 1e-20 -num_threads 15 -max_target_seqs 1 -outfmt 6
```

The percentage of a matched known protein (from a known databse) that aligned to the assembly was determined using a perl script (See Appendix B) provided by Trinitiy

developers (Grabherr et al., 2011).

The completeness of the trasncriptome was assessed using BUSCO v3 (Simão et al., 2015) that searched against a database of highly conserved single-copy genes.

```
> python run_BUSCO.py -i assembly.fasta -o busco.out -m transcriptome
-l vertebrata_odb9 -c 16
```

3.7 Transcriptome Functional Annotation

Before annotation, a perl script (See Appendix B and available through github here) was utilized to extract transcripts from two files; the merged GTF file (product of StringTie -merge), and the master non-redundant assembly in FASTA format (file which served as reference genome for mapping back the trimmed reads). A FASTA file of the *de novo* assembly was output, and then annotated using Dammit v0.2.5 (Scott, 2016).

Dammit dependencies were installed independently such as LAST, BLAST+ (Camacho et al., 2009), CRB-BLAST (Conditional Reciprocal Best BLAST) https://github.com/cboursnell/crb-blast, HMMER v3.1b2, which allowed the in-depth search of databases to compare sequence similarity and compute high-confidence orthologs. Other dependencies, including TransDecoder, which identified ORF's (open reading frames) and predicted protein coding regions, and BUSCO v3 (Simão et al., 2015), which aside from assessing completeness of the assembly, also comprised a Metazoan database to aid annotation of the assembly. Together with BUSCO, Dammit took advantage of a variety of other databases, including Pfam-A v29 (Finn et al., 2016) to search protein families, and non-coding transcripts with Rfam v12 (Nawrocki et al., 2015), OrthoDB v9.1 (Zdobnov et al., 2016) for retrieving orthologs, lastly, UniRef (Suzek et al., 2007), which implemented a collection of clustered sequences from UniProtKB. The Dammit software was run in a straightforward command line:

```
> dammit annotate assembly.fasta --n_threads 8
```

3.7.1 Orthology inference

For comparative genomics purposes, the mouse genome was used as the key model to understand mammalian biology and disease. Aside from Chapter 4, orthologs were obtained using Inaparanoid v.4.1. The annotated assembly by Dammit was used as input for retrieving orthologs.

```
> perl inparanoid.pl annotated_assembly.pep mouse.pep
```

In chapter 5, ortholog genes between species were obtained using the biomaRt bioconductor package (Durinck et al., 2009).

3.7.2 Classification of Synaptic proteins

To examine the conservation of the brain and synaptic proteome within mouse and the species in question, this thesis used two *Mus musculus* datasets; 1) the Genes to Cognition Database (G2Cdb) (Croning et al., 2008) to obtain sets of genes and proteins that have been isolated and experimentally validated to be constituents of the mouse synapse and postsynaptic density, 2) to retrieve a whole brain set of genes and proteins it was aimed to use the genome-wide Allen Brain Atlas (Jones et al., 2009), however it was found the supplementary proteome data of the high-resolution mass spectrometry-based analysis from (Sharma et al., 2015) a more suitable option as the Allen Brain Atlas lacks a downloadable list of mouse brain proteins.

3.8 Transcriptome analysis

Comparative transcriptome examination was achieved using RStudio 1.1.423 (Racine, 2012). Datasets integrating transcript expression levels in TPM (Transcriptome per

Million), orthology information and functional annotation per tissue, together with an assortment of R scripts were created to carry out an exhaustive analysis of the assembly, concluding the following;

a) Enriched protein domains :

HMMER v3.1b2 (Eddy, 1998) - hmmscan. Top 35 enriched Pfam (Finn et al., 2016) domains were identified in the annotated transcriptomes. HMMER uses a hidden Markov model (HMM) algorithm to search sequence homologs against a database. The top enriched protein domains represent the hmmscan hits with the highest full-sequence bit-score (i.e., log-odds ratio score testing the likelihood of the profile HMM to the probability of a null hypothesis).

b) Synaptic conservation :

Per tissue, examination of orthologous transcripts expressed in the mouse brain, synaptosome (SYN) and postsynaptic density (PSD). Brain data was taken from (Sharma et al., 2015), while the Genes to Cognition database (Croning et al., 2008) was used to determine SYN and PSD mouse genes.

c) Gene expression :

Gene clustering, distribution and expression per tissue.

d) Tissue enrichment analysis :

Examination of transcripts enriched and specific to each tissue.

e) Synaptic proteome :

HMMMER v3.1b2 (Eddy, 1998) in conjunction with Pfam v31.0 (Finn et al., 2016), Blastp v2.2.28 (Camacho et al., 2009), and UniProt were used to generate a G2C (Genes2Cognition) library. The later included particular synaptic proteins relevant to alterations of cognition. The whole assembly was then searched against the created library, transcripts per tissue were identified, and their expression levels were compared.

f) Gene ontology analysis :

Using an R script (NIPA, See Appendix B and available in github here) enriched GO terms were obtained from Biomart (Durinck et al., 2009) following an hypergeometric distribution. The script was modified if using the zebrafish transcriptome, bat transcriptome or lion transcriptome.

g) Phylogenetic analysis :

Homolog sequences were aligned using SeaView and phylogenetic trees were visualised using FigTree.

h) Evolutionary rates :

Homolog sequences were aligned using ParaAT v1.0. Evolutionary rates (dN/dS) were estimated by KaKs calculator v2.0

Table 3.3: Summary of tools used in transcriptome assembly and assessment.

Tool	Aim	Version	Author
Cutadapt	QC short reads	1.9	(Martin, 2011)
Trinity	Transcriptome assembly	2.1.1	(Haas et al., 2013)
TransRate	Transcriptome assessment	1.0.1	(Smith-Unna et al., 2016)
Salmon	Transcriptome quantification	0.4.2	(Patro et al., 2015)
CD-HIT-EST	Clustering and redundancy	4.6	(Li and Godzik, 2006)
HISAT2	Read alignment	2.0.4	(Kim et al., 2015)
SamTools	Parsing of alignment files	1.3.1	(Li et al., 2009)
StringTie	Assembly and quantification	1.2.3	(Pertea et al., 2015)
Kallisto	Transcriptome quantification	0.43.0	(Bray et al., 2016)
Dammit	Transcriptome annotation	0.3.2	(Scott, 2016)
Inparanoid	Orthology	4.1	(O'Brien et al., 2005)
RStudio	Analysis and Statistics	1.0.143	(Racine, 2012)
HMMER	Sequence identification	3.1b2	(Eddy, 1998)
SeaView	Sequence alignment	4.6.3	(Gouy et al., 2009)
FigTree	Phylogenetic tree	1.4.3	(Rambaut and Drummond, 2008)
ParaAT	Sequence alignment	1.0	(Zhang et al., 2012b)
KaKs Calculator	Evolutionary rates	2.0	(Zhang et al., 2006)

Figure 3.2: **De novo assembly protocol.** Diagram depicting the overall bioinformatics pipeline developed for the present study.

Chapter 4

De novo Assembly of the Zebrafish brain transcriptome

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

- Marie Curie -

Synapses are central in the functioning of the brain, mutations of genes encoding synaptic proteins are associated to more than 130 neurological alterations (Bayés et al., 2011). It is therefore of fundamental interest to understand evolution and gene expression of the mammalian brain. As a contribution to this, the zebrafish (Danio rerio) has emerged as a key model for researching vertebrate gene function in neuroscience (Davey et al., 2010).

Zebrafish are members of the teleost infraclass, which is considered to have originated ~340 million years ago from the common vertebrate ancestor. Significant, was that teleosts underwent an extra round of whole-genome duplication (WGD), known as teleost-specific genome duplication (TSGD). As a consequence, genes were duplicated in the teleost genome (called ohnologues). Zebrafish express more genes (26,206) than any previously sequenced vertebrate, together with a greater number of species-specific genes than any mammal (Howe et al., 2013a).

Recently, the proteome and ultrastructure of zebrafish synapses were reported (Bayés et al., 2017). Noteworthy in this study was the reduction of complexity of the post synaptic density proteome compared to mammals, even though the zebrafish underwent a teleost-specific genome duplication. To adequately provide a richer understanding of neurological diseases in humans, it is essential to investigate the magnitude of which zebrafish genes have mammalian orthologs. To compliment the study of proteins known to be essential in learning and memory (Emes and Grant, 2012), *de novo* transcriptome assemblies were presented from three major regions of the zebrafish brain; the optic lobe, olfactory lobe and hindbrain, together with four zebrafish whole brains (Figure 4.1). Transcripts enriched and specific to each tissue were determined, together with orthologs of mouse genes present in the brain, synaptosome and post synaptic density were identified.

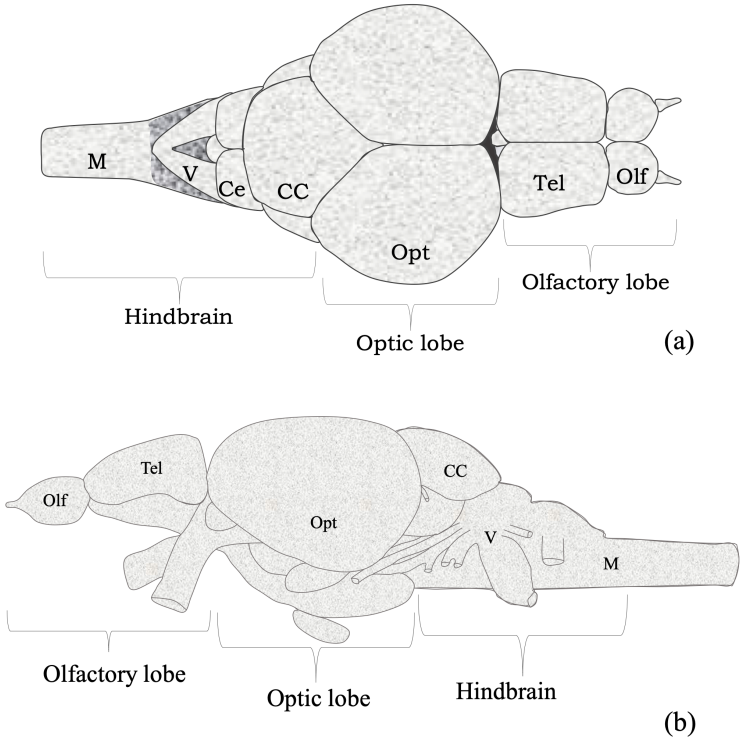


Figure 4.1: **Dorsal (a) and lateral (b) representation of the adult zebrafish brain.** showing the three studied regions; Olf (olfactory lobe) and Tel (telencephalon) both anatomical structures together were assigned the same name; Opt (optic lobe); and hindbrain, which comprised the CC (cerebellum), V (vagal lobe) and M (medula spinalis).

4.1 RNA sequencing

Total RNA was isolated and sequenced from dissected brain tissues of female and male adult *Danio rerio* (See chapter 3). These tissues include single samples (no replicates) of the olfactory lobe, the optic lobe, hindbrain, along with four whole brain (WB) replicates. To enhance transcript coverage, cDNA libraries were sequenced on four distinct lanes of an Illumina NextSeq500 sequencing platform. A total number of 302,222,856 strand-specific paired end reads 150 bp were obtained from the seven tissues with an average of 49.4 million per sample. cDNA library information is provided in Table 4.1. To ease transcriptome assembly, trimming of adaptors used in the cDNA library construction was performed with Cutadapt v1.8.3 (Martin, 2011). After removal out 0.2%

of the RNA-seq reads, the rest (119.8Gb) were used as input for *de novo* assembly to achieve a set of non-redundant transcriptomes.

Table 4.1: **cDNA library summary of the RNA sequencing yield.**

Tissue	Number of RNA-seq reads	Number of RNA-seq trimmed reads	RNA-seq trimmed reads Mbp
Hindbrain	49,325,929	49,225,244	3,754.54
Olfactory lobe	39,885,917	39,803,227	3,035.86
Optic lobe	48,429,737	48,288,479	3,682.22
WB 1	35,037,156	34,733,008	2,641.88
WB 2	35,124,670	34,833,682	2,656.59
WB 3	40,938,594	40,534,803	3,090.30
WB 4	53,480,853	53,227,383	4,061.51

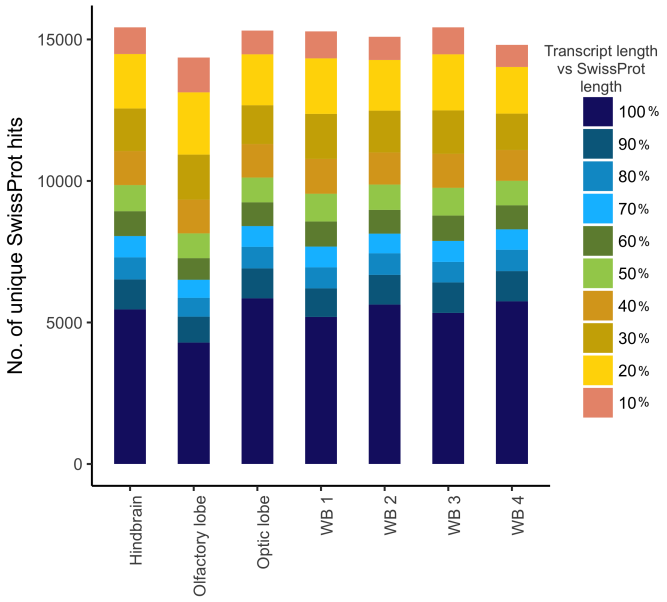
4.2 De novo transcriptome assembly and assessment

Trimmed reads were assembled into contigs (transcripts) without a reference genome using the 3-module *de novo* assembler Trinity (Haas et al., 2013). First, Inchworm assembled unique contigs from adapter-trimmed RNA-seq reads using a *k*-mer size of 25, next, Chrysalis grouped Inchworm-generated contigs into components and output a De Bruijn graph for each. Lastly Butterfly evaluated each graph and enumerated each possible transcript, including alternative splicing forms. Seven transcriptomes (one for each biological sample) were successfully assembled into 89,595 ~ 109,847 contigs. Each assembly represented around 83-109 Mbp (698,207 contigs; 746 Mbp in total). Transcripts less that 200 bp were removed resulting in a mean contig length of 699-869 bp. Detailed metrics for all assembled transcriptomes are represented in Table 4.2.

Table 4.2: **Trinity *de novo* assembly statistics summary.**

Tissue	No. of contigs	Base pairs (Mbp)	Mean contig length (Bp)	Median contig length	SD contig	%GC	% read content
Hindbrain	102,900	82.864	805	404	978	43.59	86.13
Olfactory lobe	89,595	62.624	699	376	795	43.45	87.66
Optic lobe	100,740	85.597	850	425	1,013	43.50	87.42
WB 1	98,285	78.338	797	405	964	43.95	85.81
WB 2	95,957	82.443	859	426	1,049	43.63	85.57
WB 3	100,883	81.878	812	402	1,001	43.91	86.40
WB 4	109,847	95.488	869	421	1,070	43.69	86.95

Using Bowtie2 (Langmead and Salzberg, 2012), reads were mapped to the *de novo* assemblies in order to evaluate their content. Overall alignment rates were ~91% with a total percentage of read content ranging from 85.57% to 87.42%. A good quality assembly is expected to be represented by a minimum of 80% of its RNA-Seq reads (Haas et al., 2013). The fraction of the transcriptome that represented either full-length or close to full-length was next examined. Figure 4.2 illustrates the number of unique "top hits" based on BLASTX sequence similarity (transcript alignments that represent the highest score per database entry with an E-value cutoff of e-20) compared to the SwissProt database and *Danio rerio* RefSeq transcripts (GRCz10).



(a)

Figure 4.2: **Percentage of the transcript length that align to a known protein.** A metric for quality assessment of the assembly is to explore the portion of the transcripts that align a known protein. If ≥90% of the total length of a transcript aligns to a known protein, this transcript is considered "full-length" . BLASTX (e-value of e-20) was carried out to align the Trinity-generated assembly to (a) the SwissProt database; and (b) the *Danio rerio* reference transcriptome (GRCz10). This used to assess the Trinity-generated transcriptome, with no filtering of the data.

Transcripts are determined as "full-length" if they represent over 90% of the reference transcript length (Mamrot et al., 2017). In summary, an average of 21,200 and 25,726 unique BLAST hits were obtained against the SwissProt protein database and the *Danio rerio* RefSeq respectively for all assemblies. From the average of 21,200 unique matches with the SwissProt database, around 19% (3,963) corresponded to *Danio rerio*, which in turn represented the highest percentage of identical matches (greater than 90%). Detailed results of unique BLASTX alignments are displayed in Table 4.3.

Table 4.3: **Trinity transcriptomes BLASTX (E-value of e-20) summaries and percentage that align from the Transrate-optimised > 0.5 TPM** for a completeness assessment of the transcriptome and not filtering practice was carried out.

Tissue	Unique SwissProt BLASTX hits	Unique RefSeq BLASTX hits
Hindbrain	31,448 - 31.2%	41,176 - 40.8%
Olfactory lobe	28,269 - 31.6%	37,556 - 42.0%
Optic lobe	30,851 - 30.6%	39,821 - 39.6%
WB 1	31,325 - 31.9%	40,637 - 41.4%
WB 2	30,029 - 31.3%	38,778 - 40.4%
WB 3	31,722 - 31.5%	41,297 - 41.0%
WB 4	33,050 - 30.1%	42,829 - 39.0%

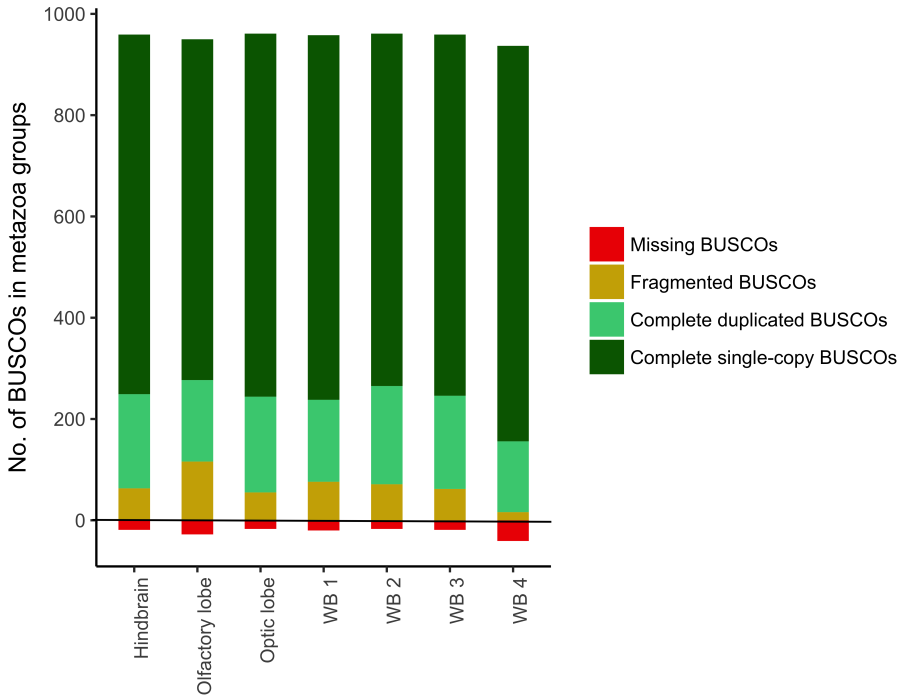
Optimized transcriptome BLASTX (E-value of e-20) summaries and percentage that align from the Transrate-optimised > 0.5 TPM for a completeness assessment of the transcriptome and not filtering practice was carried out.
Nonetheless, there are a few points to be considered, such as separate portions from one transcript can have different best top BLAST hit, this may indicate misassembly of these transcripts. Moreover, the assemblies might consist of sequences that are artificially fragmented, thus these are designated as “distinct genes". Consequently, unique BLAST hits might not reflect the true number of unique genes. Additionally, the proportion of transcripts that did not match any SwissProt and RefSeq proteins, likely constitute either non-coding, or transcript fragments insufficient to reach the E-value cutoff of e-20.

The Trinity assemblies were further assessed by BUSCO (Benchmarking Universal Single-Copy Orthologs) using metazoan and vertebrata (Simão et al., 2015) gene datasets (Figure 4.3 and Table 4.4). BUSCOs completeness is categorized as; Complete and single-copy (recovered transcripts are within 95% expectation of the BUSCO group mean

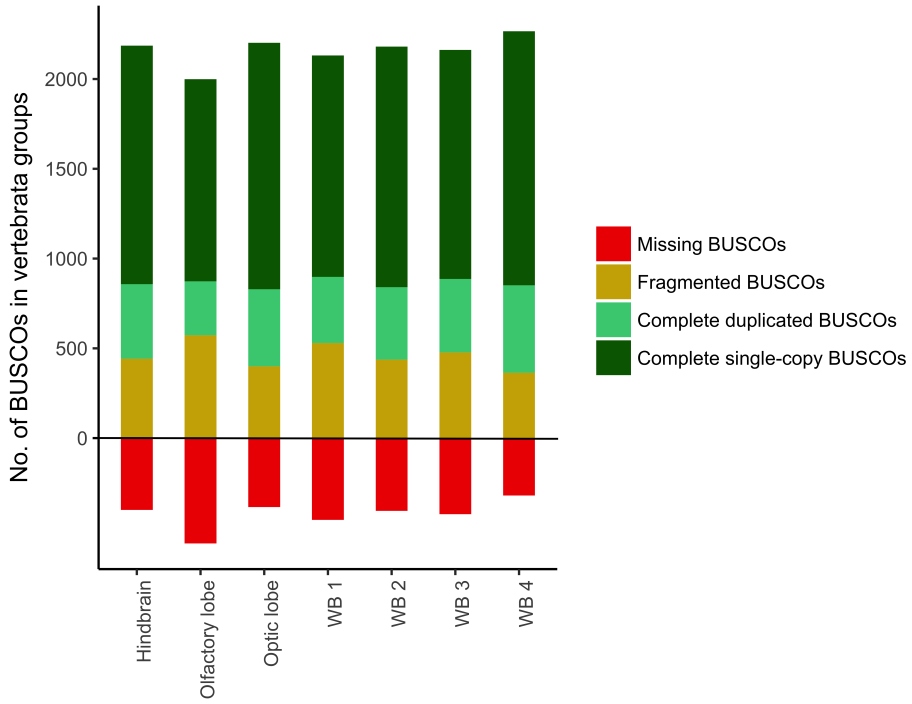
length); Complete (transcripts with more than one single-copy); Fragmented (incomplete recovered transcripts), and; Missing (not recovered transcripts). Such metric was only used for assessment of the transcriptome but not filtering procedure was carried out.

From 978 single-copy orthologs in the metazoan dataset used to evaluate the relative completeness, the assemblies were 85.3-93.8% complete, 5.2-11.9% were fragmented, and only a small number of transcripts were missing; 1-2.3%. Yet, while assessing completeness with 2,586 single-copy orthologs for the vertebrata dataset, the proportion of detected complete BUSCO single-copy and duplicated decreased to 43.5-54% and 11.6 -8.8%, respectively, moreover, it also increased the proportion of fragmented and missing BUSCO to 22-12.4%.

It should be noted that the metazoa dataset consists of several phyla that provides a broader coverage of the tree of life, whereas the larger vertebrata clade is more precisely defined, therefore offering higher resolution. Regardless, both BUSCO datasets presumes that all genes are expressed and detectable, yet this study is focused only in brain tissues. For that reason, genes that are expressed in a given cell or tissue but not in the brain, would be assessed as "missing" despite the fact that the transcriptome is "complete".



(a)



(b)

Figure 4.3: **BUSCO completeness assessment of Trinity assemblies.** BUSCO analysis is categorized as; Complete and single-copy (recovered transcripts are within 95% expectation of the BUSCO group mean length); Complete and duplicated (transcripts with more than one single-copy); Fragmented (incomplete recovered transcripts), and; Missing (not recovered transcripts). a) Assessment using metazoan BUSCO set, representing 843 genes. (b) Assessment using vertebrata BUSCO set, representing 3,023 genes.Such metric was only used for assessment of the transcriptome but not filtering procedure was carried out.

Table 4.4: **% of BUSCO Metazoa (M) and Vertebrata (V) completeness assessment of Trinity assemblies with no filtering of the data.**

Tissue	Complete single-copy		Complete duplicated		Fragmented		Missing	
	M	V	M	V	M	V	M	V
Hindbrain	72.6	51.4	19.0	16.1	6.5	17.1	2.0	15.5
Olfactory lobe	68.8	43.5	16.5	11.6	11.9	22.2	2.9	22.7
Optic lobe	73.3	53.0	19.3	16.6	5.6	15.5	1.7	14.9
WB 1	73.7	47.7	16.6	14.2	7.8	20.5	2.1	17.6
WB 2	71.2	51.9	19.9	15.6	7.3	16.9	1.7	15.7
WB 3	73.0	49.4	18.8	15.8	6.3	18.5	2.0	16.4
WB 4	75.4	54.7	18.4	18.8	5.2	14.0	1.0	12.4

Results from BUSCO were found to be consistent with high-quality reference transcriptomes from various BUSCO taxa (Simão et al., 2015; Mamrot et al., 2017), therefore, suggesting a favourable transcriptome completeness (full-length transcripts). The number of duplicated sequences, likely reflects evidence of teleost-specific genome duplication (TSGD). Small differences were found across the seven assemblies. The olfactory lobe assembly resulted in the lowest BUSCO completeness score, this is likely due to the lower quality of the transcriptome as seen by the reduced number of BLASTX hits reported in Table 4.2 and Figure 4.2.

4.3 *De novo* assembly Quality Control

While Trinity has proven to effectively reconstruct a high quality *de novo* assembly (Zhao et al., 2011; Grabherr et al., 2011; Mamrot et al., 2017), additional transcriptome-quality measures can be implemented to correctly identify poorly assembled transcripts, so unreliable data do not affect downstream analyses. Transrate (Smith-Unna et al., 2016) was used for quality and completeness evaluation of the seven assemblies. An optimized score varying from 0.392 to 0.412 was estimated in each assembly. This

score represents a quality transcriptome suitable for downstream studies, i.e. assemblies with an optimized scores of 0.35 or greater are considered of sufficient quality (Smith-Unna et al., 2016).

Transrate correction condensed the Trinity contig set down to 40,997-59,238 transcripts, on average, 51.5% of data was filtered out for each transcriptome (n=359,633 contigs). Detailed Transrate assessment data is presented in Table 4.5. Transrate improved assemblies were quantified using Salmon (Patro et al., 2015). Transcripts whose abundance was estimated of less than 0.5 TPM were excluded, removing around 2.3% of the total sequences. The set of contigs was compressed to 43,930-57,646 (n= 331,624 contigs; 389Mb) with an average contig length of 1,213.5 bp. Filtering out low-expression generated transcripts is a frequent practice as it removes RNA sequencing errors that most likely are mirrored in the generation of low-expression transcripts (Sha et al., 2015). However, it is challenging to appropriately select a rationale of TPM filtering, as there is a lack of a consensus TPM threshold filtering. In this research, a 0.5 TPM filering was selected for avoiding RNAseq noise and also be able to identify low-expressed transcripts. Nevertheless, in Chapter 7 it was preferred a more stringent TPM threshold to increase confidence in the identification of PSD homologs.

Table 4.5: **Transrate quality control assembly summary.**

Tissue	Transrate optimized score	No. transrate transcripts	% retained transcripts from Trinity-generated transcriptomes	No. transcripts > 0.5 TPM
Hindbrain	0.3917	59,238	57.57	57,646
Olfactory lobe	0.4024	40,997	45.76	40,393
Optic lobe	0.4088	47,871	47.52	46,728
WB 1	0.4053	46,954	47.77	46,223
WB 2	0.3997	49,388	51.47	48,456
WB 3	0.4391	43,973	43.59	43,248
WB 4	0.4117	50,183	45.69	48,930

To allow a systematic comparative analysis across different tissues, it is sensible to integrate all transcriptomes into a consolidated set of transcripts, followed by mapping back RNA-seq data for expression estimation. CD-HIT-EST (Li and Godzik, 2006) was used to merge redundant transcripts (n=331,624). Highly similar transcripts (> 90% similarity) were clustered, which resulted in a drastic reduction in the number of transcripts to 96,378 (102 Mb). A custom PERL script was used to further filtered out the assembly, retaining only the longest transcript per CD-HIT-EST cluster. This way, the assembly comprised 45,775 redundant-free unique transcripts, and served as a reference transcriptome to map back all reads in the subsequent step of the pipeline. Among these transcripts, 26,628 (58%) were longer than 1,000 bp, with a range of 204 - 2,1804 bp. The mean contig length was 1,663 bp.

As described in Chapter 3, following the "new Tuxedo" pipeline (Pertea et al., 2016), RNA-seq reads from each sample were mapped against the non-redundant assembly (n=45,775 transcripts). Mapping results are illustrated in Figure 4.4.

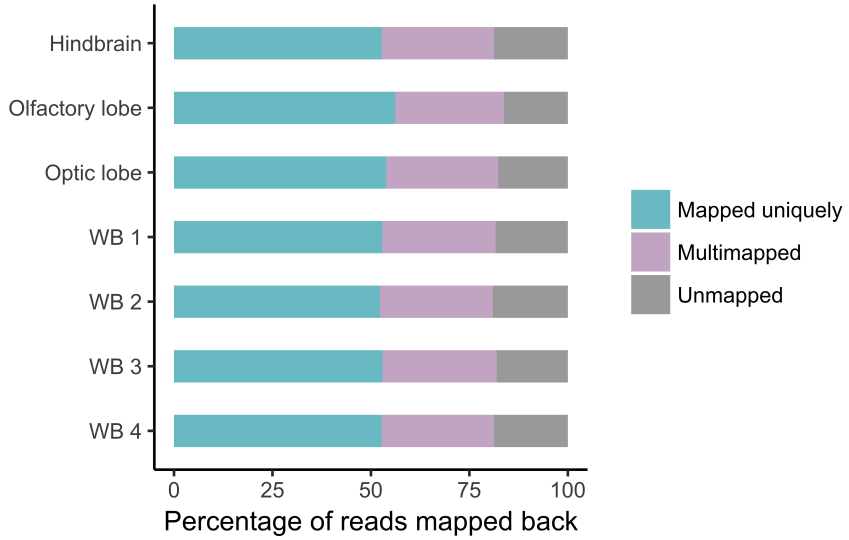


Figure 4.4: **Reads mapped back to transcriptome.** The non-redundant assembly (n=45,775 transcripts) was used as reference transcriptome to map back trimmed reads from all tissues. Proportion of the mapping summaries are represented in different colours.

Overall alignments obtained ranged from 85.40% to 87.77% among samples, which can be interpreted as satisfactory mapping rates, as has been reported in previous studies (Trapnell et al., 2009; Conesa et al., 2016). Alignments were then assembled resulting into a uniform contig size distribution through all transcriptomes, as shown in Table 4.6.

Table 4.6: **Assembly statistics after mapping reads back to the optimized Trinity non-redundant transcriptome.**

Tissue	% Overall alignment	No. of contigs that aligned	Mean contig length	SD contig length	% GC content
Hindbrain	85.40	46,430	1,644	1,398	42.75
Olfactory lobe	87.77	45,294	1,636	1,376	42.70
Optic lobe	87.07	46,550	1,648	1,398	42.76
WB 1	86.26	46,092	1,644	1,393	42.75
WB 2	86.10	46,199	1,643	1,395	42.75
WB 3	86.44	46,227	1,643	1,393	42.77
WB 4	86.10	46,707	1,653	1,403	42.77

In an improvement to the initial Trinity-generetad transcriptomes, in which the average contig length was of 699 ~ 869 bp, this final set of transcriptomes comprised longer transcripts, with an average contig length of 1,636 bp (for the smallest assembly) to 1,653 bp (for the largest assembly). In addition, the number of transcripts was reduced, ranking from 45,294 to 46,707 contigs; initially 89,595 ~ 109,847 bp. The large number of Trinity contigs suggest that individual genes were formed by multiple contigs, possibly because of assembly of incomplete reads.

A final set of transcript sequences was created for annotation and subsequent analyses. This assembly consolidated the Trinity/CD-HIT-EST non-redundant assembly (n=45,775) along with a merged GTF file computed by Stringtie -*merge* function, which consolidated transcript structures from all samples. This final assembly included 47,979 transcripts (79 Mbp) with an average contig length of 1,683 bp (SD of 1,426). The longest and shortest transcript was 21,804 and 201 bp respectively, and 28,482 transcripts, or 59% of the assembly was longer than 1,000 bp.

In order to assess whether the *de novo* optimized transcriptome (n=47,979) was consis-

tent with the *Danio rerio* reference genome (GRCz10), the contig size distribution was compared (Figure 4.5a). Whilst, there was a larger number of smaller contigs (less than 1,000 bp) in the *de novo* transcriptome, there is clear consistency. Additionally, both assemblies were compared and assessed by BUSCO vertebrata dataset (Figure 4.5b and Table 4.7).

Table 4.7: **BUSCO vertebrata assessment of Trinity optimized assembly and zebrafish reference genome for a completeness assessment of the transcriptome and not filtering practice was carried out.**

Assembly	**Complete single-copy**		**Complete duplicated**		**Fragmented**		**Missing**	
	No.	%	No.	%	No.	%	No.	%
Optimized *de novo*	1,609	62.2	428	16.6	209	8.1	340	13.2
***Danio rerio*-GRCz10**	1,498	58.0	1,011	39.1	48	1.9	29	1.1

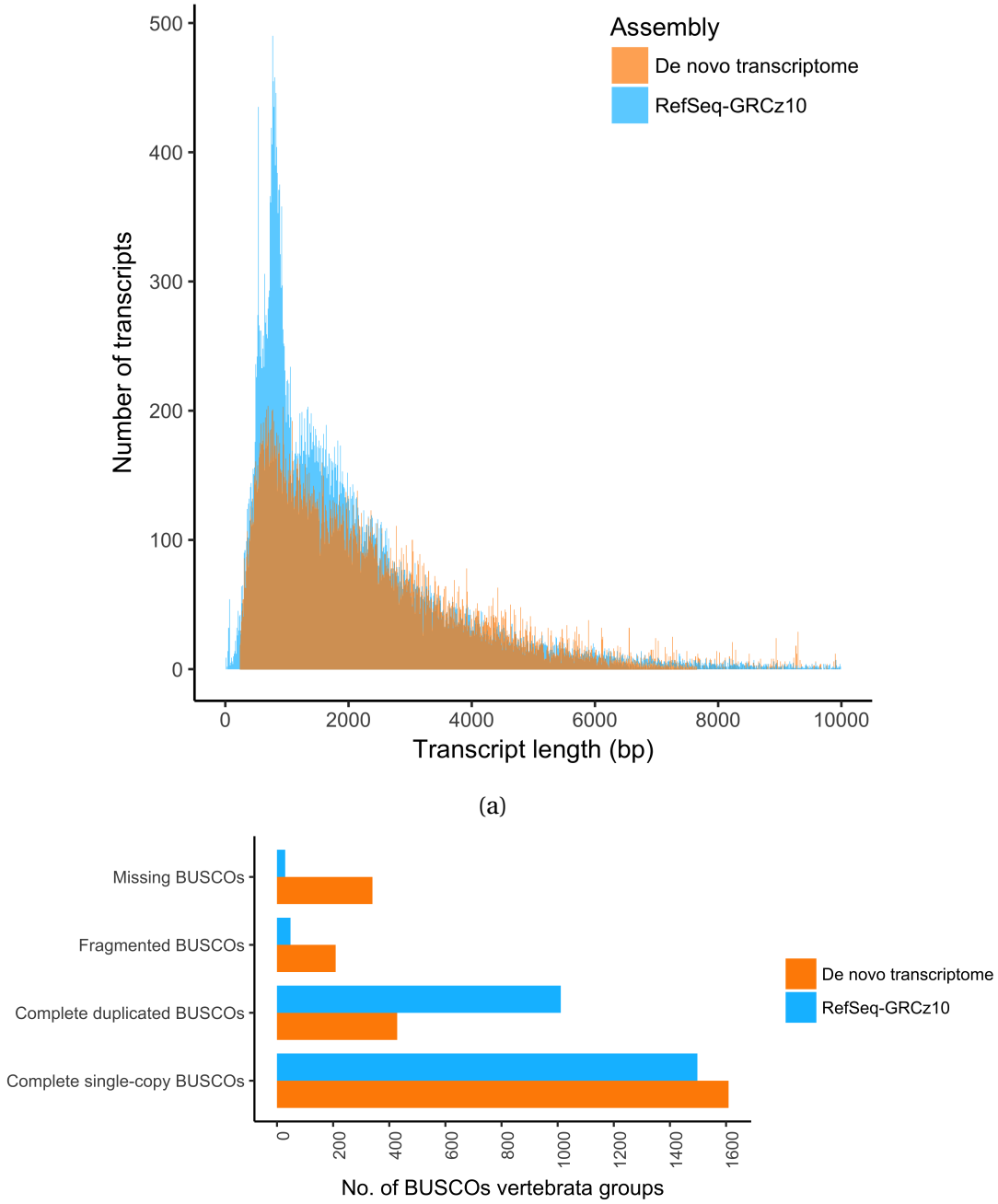
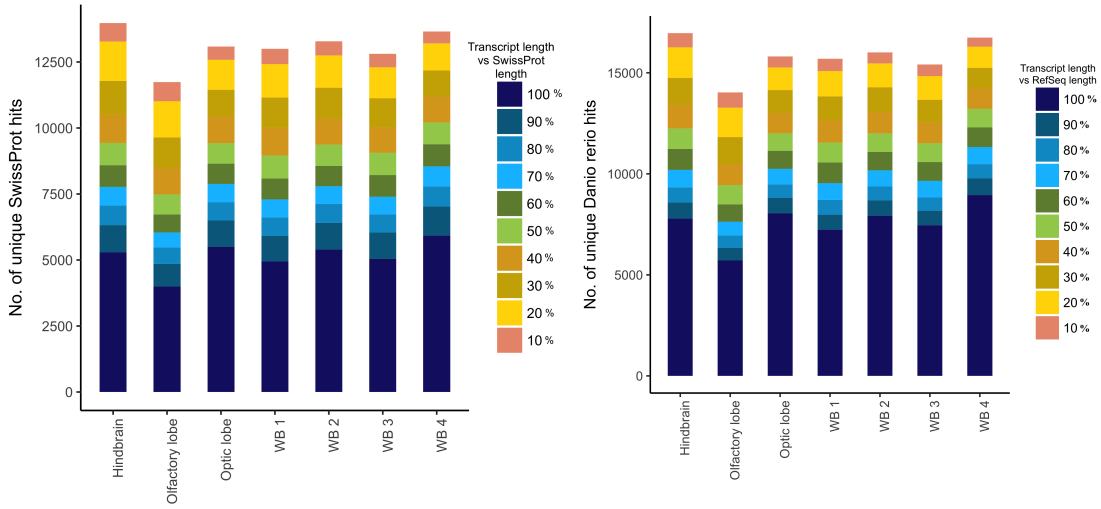


Figure 4.5: **Assessment of the optimized *de novo* assembly and *D. rerio* (GRCz10).** a) Contig length distribution overlapping the number of transcripts (x-axis) with their length in base pairs (y-axis). (b) BUSCO-vertebrata assessment showing a comparison of completeness between the optimized assembly and the *D. rerio* reference genome. Such metric was only used for assessment of the transcriptome but not filtering procedure was carried out.

Danio rerio has been widely studied for vertebrate gene function, therefore, its genome

reference (GRCz10) is well-annotated, which yielded 97% BUSCO completeness, with only a few fragmented and missing portions (1.1% and 1.9%, respectively). The above might be due to a technical BUSCO limitation, rather than a slightly incomplete assembly. There was a high number of complete-duplicated BUSCOs in the reference genome, which is consistent with the teleost-specific genome duplication (TSGD). The optimized assembly showed a substantial improvement in comparison with the non-optimized assembly, yielding 78% BUSCO completeness (62% single-copy and 17% duplicated). The proportion of fragmented and missing BUSCO also decreased to 8% and 13% respectively (as illustrated in Figure 4.3). Nonetheless, 13% was not found by the BUSCO analysis, and 8% was determined as fragmented.

The *de novo* optimized assembly was compared to the SwissProt database and Danio rerio-GRCz10 using BLASTX (Figure 4.6 and Table 4.8). Using an E-value cutoff of 1e-20, 25,891 unique “single best" BLAST hits resulted from the SwissProt databse, which 24,682 contained an ORF and 1,755 were alternatively spliced. Likewise, 27,900 transcripts (or 58% of the non-redundant transcriptome) had a unique BLASTX hit against *Danio rerio*, of which 1,857 were isoforms and 26,225 contained an ORF. In contrast with the non-optimized *de novo* assemblies (Figure 4.2), the optimized assembly showed less redundancy and higher number of full-length transcripts than the original Trinity-generated assembly (See Table 4.3 for comparison).



(a) (b)

Figure 4.6: **Percentage of the transcript length that align to a known protein.** A metric for quality assessment of the assembly is to explore the portion of the transcripts that align a known protein. If ≥90% of the total length of a transcript aligns to a known protein, this transcript is considered "full-length" . BLASTX (e-value of e-20) was carried out to align the optimized assembly to (a) the SwissProt database; and (b) the *Danio rerio* reference transcriptome (GRCz10). This used to assess the optimized transcriptome, with no filtering of the data.

Table 4.8: **Optimized transcriptome BLASTX (E-value of e-20) summaries and percentage that align from the Transrate-optimised > 0.5 TPM for a completeness assessment of the transcriptome and not filtering practice was carried out.**

Tissue	Unique SwissProt BLASTX hits	Unique RefSeq BLASTX hits
Hindbrain	23,777 - 41.3%	41,176 - 40.8%
Olfactory lobe	18,530 - 45.9%	37,556 - 42.0%
Optic lobe	21,573 - 46.2%	39,821 - 39.6%
WB 1	21,788 - 47.1%	40,637 - 41.4%
WB 2	21,968 - 45.3%	38,778 - 40.4%
WB 3	21,280 - 49.2%	41,297 - 41.0%
WB 4	23,226 - 47.5%	42,829 - 39.0%

4.4 Transcriptome Annotation

Dammit! (Scott, 2016) was used to annotate the optimized non-redundant assembly (n=47,979), retrieving annotations for 34,182 transcripts or 71.2% of the total transcriptome, of which 2,173 were spliced forms. TransDecoder (Haas and Papanicolaou, 2012) estimated that 28,754 of the transcripts contained an ORF, which likely represent candidate protein sequences, 6.7% of these transcripts (1,932) were identified as alternatively spliced. Pfam in conjunction with HMMER v3.2.1 (Eddy, 1998) identified that 25,384 transcripts (53%) matched to 11,203 protein domain families, 7% (1,786) of these transcripts were spliced forms. LAST search retrieved 33,371 (69.6%) matches for known protein sequences in the UniRef90 database, of which 6.3% (2,103) corresponded to alternative spliced forms.

Transcripts containing Pfam domains related to protein kinases PDZ, SH3, C2, ribosom e-binding and ankyrin repeats (Ank) were among the most significantly enriched group (Figure 4.7). These domains have been well-conserved from the earliest branching animals with synapses. For example, the conservation of PDZ ligand sequences are entirely conserved between sponges and humans. PDZ and ankyrin repeat-containing scaffolds bind proteins, such as PSD-95 and Shank into molecular complexes that mediate the synaptic size and potency (Kim and Sheng, 2004; Böckers et al., 2001). Moreover, MAGUK (membrane-associated guanylate kinase)-associated signalling complex among the PSD, are comprised of an array of multiprotein complexes, e.g., kinases, phosphatases, cell adhesion, receptors and ion channels, which comprises the main post-synaptic machinery implicated in synaptic transmission and plasticity and its origin precedes the evolution of the nervous system (Ryan and Grant, 2009; Emes et al., 2008; Sakarya et al., 2007).

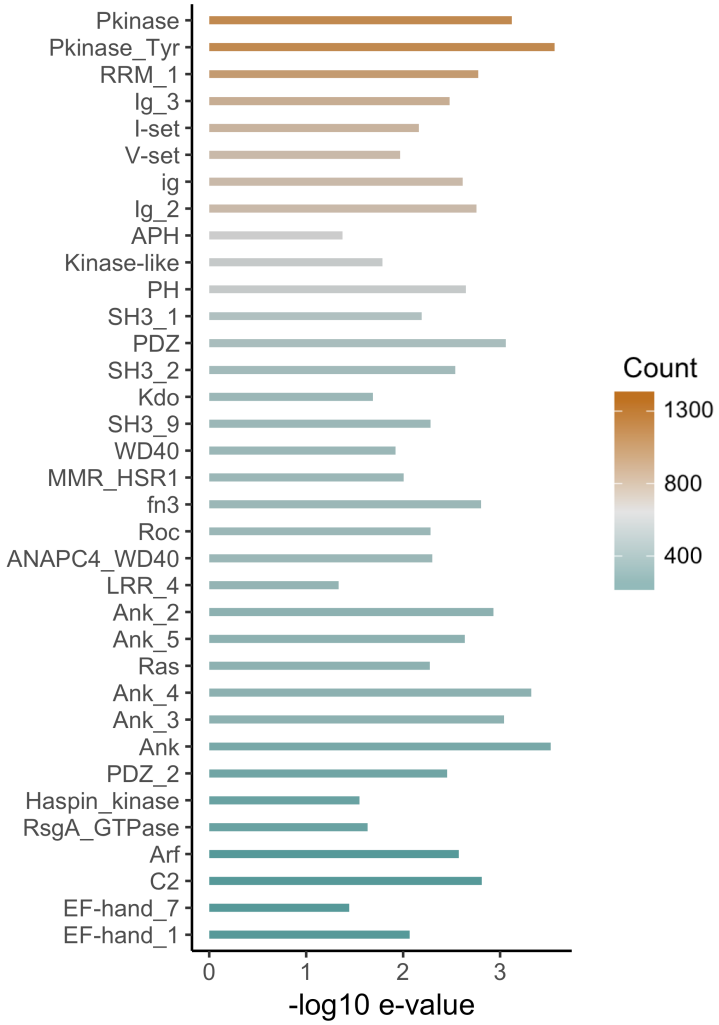


Figure 4.7: **Top 35 enriched Pfam domains.** Annotated transcripts were investigated for protein domains using Pfam and HMMER hmmscan (See chapter 3). The y-axis shows the top enriched domains found with an e-value cut-off of 0.05. Colours indicate the number of transcripts found in each domain; orange comprises the highest, while blue the lowest.

Additionally, the predicted protein-coding transcripts (n=28,754) from Dammit! were compared to the zebrafish reference genome (GRCz10) using BLASTp and an e-value cut-off of e-10. 95% of these transcripts (n=27,273) showed unique blast hits, which mat-

ched to 16,240 proteins and 14,402 genes of the *D. rerio* reference genome. Given that zebrafish have become a widespread model organism for vertebrate studies, its reference genome is of very high quality (Howe et al., 2013a). Thereby, *de novo* transcripts

that aligned (n=27,273) to the reference genome were selected for investigating orthology and gene expression. Summary of the *de novo* assembly and annotation is illustrated in Figure 4.8.

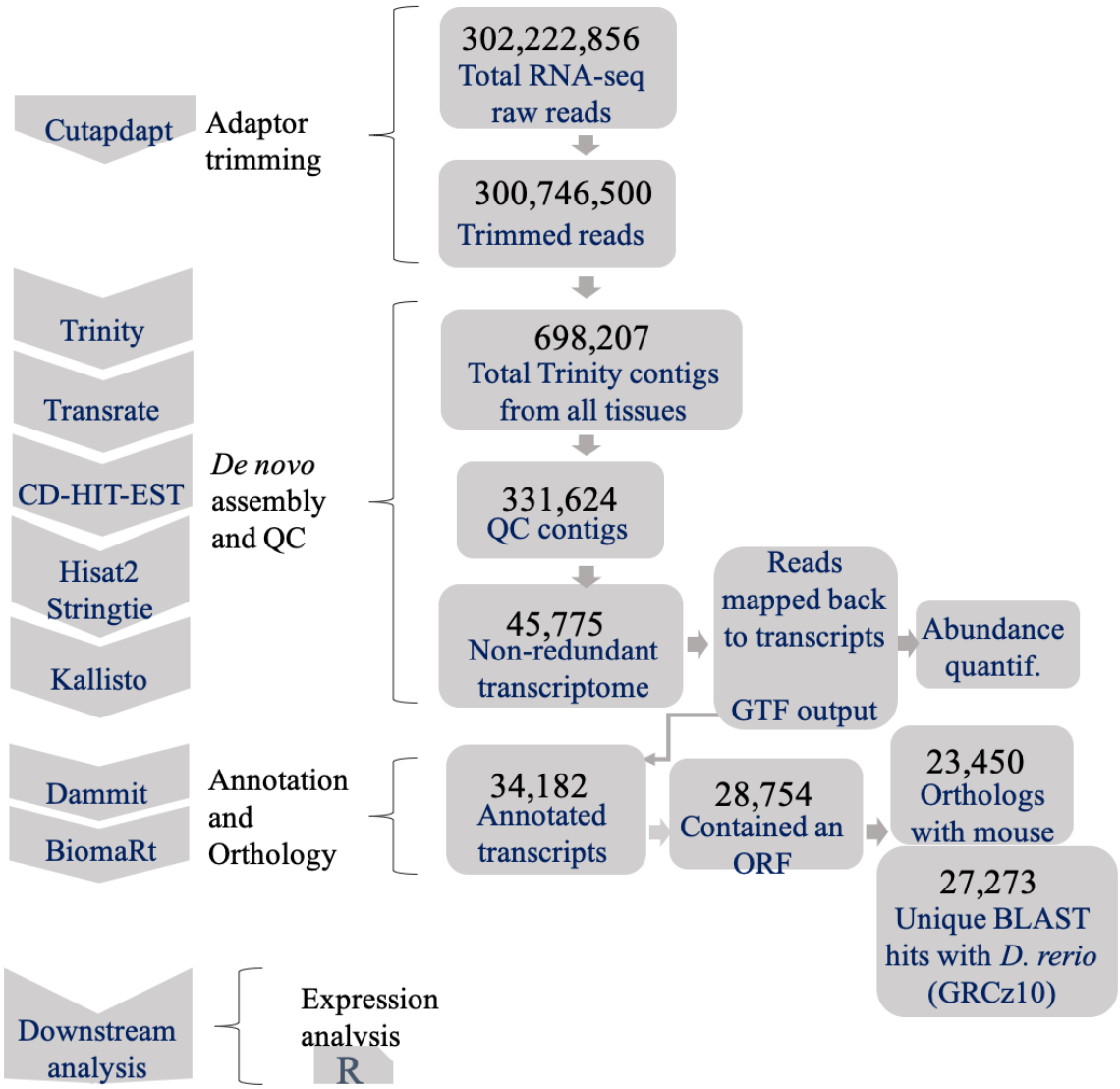


Figure 4.8: **Summary of *de novo* transcriptome assembly and annotation pipeline.** RNA-seq reads were pre-processed using Cutadapt and *de novo* assembled using Trinity. Quality control of Trinity contigs was performed by Transrate. CD-HIT-EST was used to obtain a non-redundant transcript set retaining only the longest contig per CD-HIT cluster. This generated a high quality transcriptome, which was used as reference to map back the RNA-seq reads from all tissues to its further quantification. Using the Dammit platform, the non-redundant transcriptome was annotated and protein-coding genes were predicted by Transdecoder. This set of protein-coding genes was BLASTp obtaining a final set of transcripts that was used to infer orthology and expression analysis.

4.5 Transcriptome Orthology

Efforts to characterize the postsynaptic proteome in mice have allowed a better comprehension of the structure, function and disease of the synapse (Bayés et al., 2012; Collins et al., 2006). Therefore, as a strategy to elucidate vertebrate synaptic evolution along with neurological alterations, direct comparison was conducted of the *de novo* zebrafish transcripts with *Mus musculus* (Ensembl version GRCm38.p5) genome. For each transcript that obtained a unique BLASTp hit (27,273 transcripts) with *Danio rerio* (version GRCz10), the equivalent gene was determined and orthologous genes among both species were identified using the biomaRt bioconductor package (Durinck et al., 2009).

In total, 23,450 transcripts were identified to have a mouse ortholog (6.8%, n=1,587 were spliced forms) with 11,816 *Mus musculus* genes. Given the above, 86% of the transcriptome have a mouse orthologous gene, yet, studies have identified that approximately 71.4% of the human genome have a zebrafish orthologue. And mouse and human have a median amino acid sequence identity of 78.5% (Gharib and Robinson-Rechavi, 2011; Howe et al., 2013b). The present study found a greater number of orthologs between zebrafish and mouse than was expected. However, studies have reported that the evolution of proteins expressed in the brain, is considerable slower in comparison with other tissues, which may explain the larger number of orthologs found (Bayés et al., 2017, 2011).

Among the set of orthologs found, relationships were estimated, i.e., (zebrafish:mouse) 1:1, many:1, 1:many, many:many and unique to zebrafish (Figure 4.9(a)). 50.7% of all orthologous transcripts (n=13,819 in 5,666 mouse genes) had a 1:1 or single-copy orthology relationship, which likely reflects the significant conservation in the vertebrate synaptosome. These genes diverged through a speciation event and hence are likely to have kept equivalent functions. Whereas the 1:many and many:many type of or-

thologs are usually referred as “paralogs” and not considered “true orthologs", since these genes have diverged through duplication events, and might unveil functional differences (Creevey et al., 2011; Gabaldón and Koonin, 2013). Moreover, the term “onholog" is commonly used to define paralogs that arose from a whole genome duplication event, such as the teleost-specific genome duplication (TSGD) ((Howe et al., 2013a)).

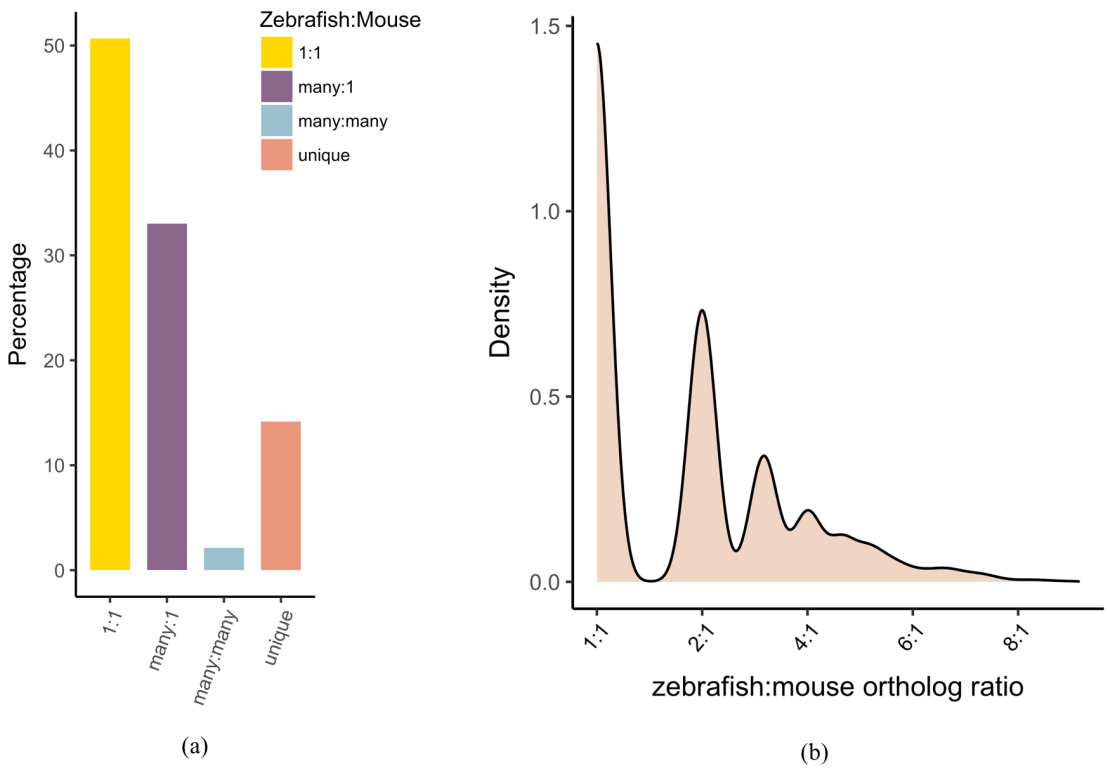


Figure 4.9: **Orthology distribution among *de novo* transcripts and mouse genes.** a) Each transcript that retrieved a unique blastp hit with *D. rerio*-GRCz10 (e-value cut-off of e-10) (27,273 transcripts) was used to determine orthology with *M. musculus*-GRCm38.p5 using the biomaRt bioconductor package (Durinck et al., 2009) Zebrafish:Mouse ratio of orthologs were represented as 1:1, many:1, many:many, and unique to zebrafish. b) For each pair of orthologs a density distribution was produced.

The second largest class of orthologs found after the 1:1, was many:1 (zebrafish:mouse). 33% (9,007 transcripts) of the total set of orthologs were identified to be expanded in the zebrafish genome. From this set of transcripts, 46% of the ortholog pairs were determined to have a 2:1 ratio; i.e., two zebrafish transcripts with

one orthologous gene in the mouse (Figure 4.9(b)), these duplicated pairs in the zebrafish genome supports the TSGD.

A *de novo* assembler frequently computes numerous transcripts compared to the number of genes that these express. In this respect, studies have shown that the number of expressed transcripts to genes is determined to be on an average ratio of 1.12, i.e., there are more than one transcript per expressed gene. Alternative splicing events also increase the number of transcripts per express gene (Gonzàlez-Porta et al., 2013). Coupled with above, algorithms of *de novo* transcriptome asssembly might result into fragmented assemblies of a substantial number of transcripts (contigs), which in reality are sub-sequences of the underlying true transcript. Figure 4.9 illustrates this phenomenon with the many:1 (zebrafish:mouse) ratios; 3:1 (21.5%), 4:1 (11.8%), 5:1 (6.8%), 6:1 (5%), 7:1 (2.7%), and so on up to 24:1 (0.016%) orthology ratios.

To investigate whether the generated orthology ratios generated (Figure 4.9) for the *de novo* assembly differ in the *Danio rerio* reference genome, rations were also estimated for the later (Figure 4.10). A greater density of 1:1 in contrast to the many:1 ratio of orthologs was observed, yet the larger majority of many:1 were determined as 2:1 with negligible amounts of >2:1.

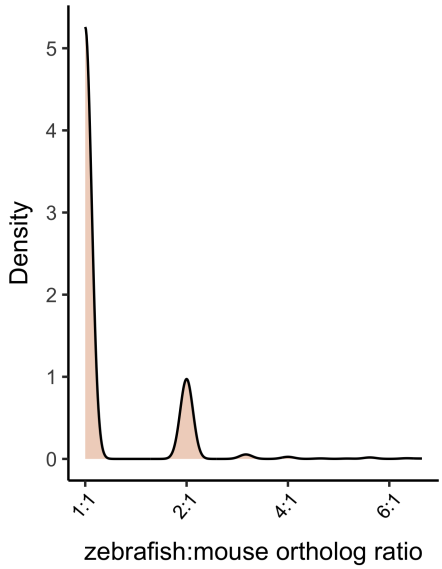


Figure 4.10: **Density of genes by ortholog ratio.** For each pair of orthologs (zebrafish:mouse) the density of its distribution was calculated. *Danio rerio* genes obtained by Blastp were used. It is shown only two ratio peaks; 1:1 that encompass the majority of orthologs and a 2:1 smaller peaks.

As described, these zebrafish genes (many:1), are not “true orthologs", since they likely originated as a result of the teleost-specific genome duplication (TSDG) that occurred ~300 million years ago, i.e., ~200 million years after the two rounds of whole-genome duplication (2R-WGD) at the based of the vertebrate clade. As a consequence, a larger number of species-specific genes have been found in the zebrafish genome, than in the chicken, mouse and human genome. Indeed, zebrafish has a larger number of protein-coding genes than any formerly sequenced non-fish chordate (Howe et al., 2013b). In this study 14.3% of transcripts (n = 3,911; 2,048 *D. rerio* genes) were determined to be zebrafish species-specific. To ascertain the overall functionality of the zebrafish species-specific genes, enrichment of GO terms (Figure 4.11), along with pathways and protein domains was investigated using R tools and HMMER.

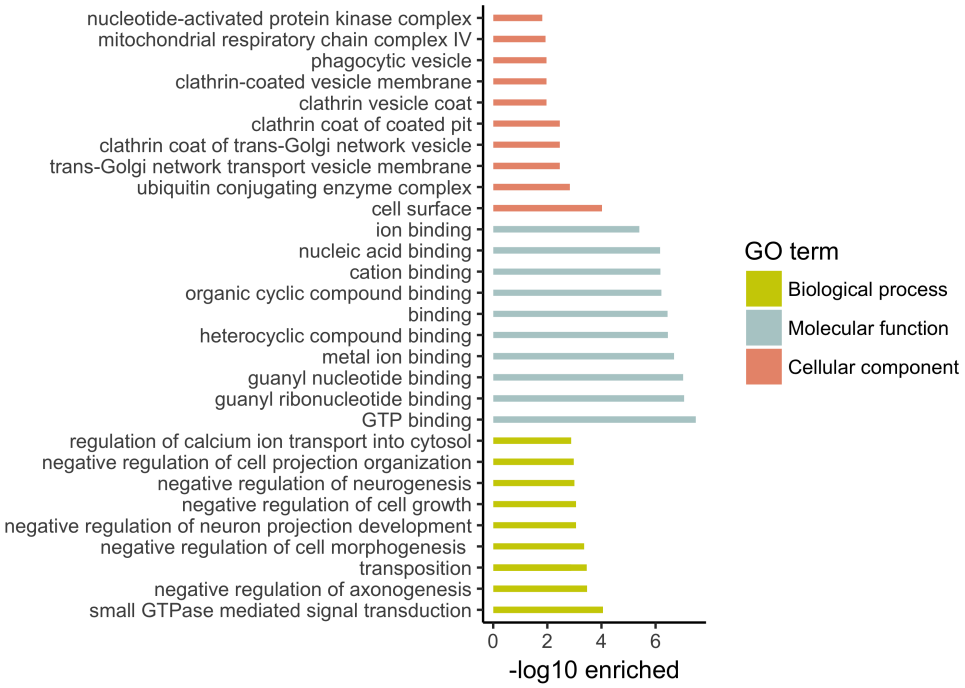


Figure 4.11: **GO analysis of species-specific genes in the zebrafish.** Zebrafish-specific genes (n=3,911) were annotated for Gene Ontology (Biological Process, Molecular Function and Cellular Component) using R analysis tools. The top most significantly enriched GO terms were plotted against the negative log10 P-value and coloured based to their functional category. P values were estimated based on hypergeometric distribution and adjusted by false discovery rate (FDR) control procedure applying a cut-off of 0.05.

Several of these species-specific genes were found to be enriched in clathrin-vesicles components (ap1g2, clta and slc18a3a). Clathrin-mediated endocytosis recycles synaptic vesicles from in and around the synaptic cleft to the presynptic terminal within a fraction of a second by an aggregation of membrane in the synaptic region, thereby allowing an effective neurotransmission (Pelassa et al., 2014). Protein binding molecular function was found enriched, genes involved in this ontology included agouti-related protein family members (asip1, asip2, agrp1, and agrp2). This protein family are neuropeptides that play a role in energy balance, and are believed to have evolved early in vertebrate evolution. Moreover, agouti-related protein family members are an example of teleost-specific genome duplication (TSGD) as tetrapods express two, while teleost have four (Braasch and Postlethwait, 2011). A high density of zinc finger pro-

teins were found involved in cation, metal and ion binding (e.g., zgc:92594, zgc:112998, zgc:171971, zgc:153704, zgc:110239). Negative regulation of neurogenesis and development were enriched in semaphorin proteins. These are group of phylogenetically conserved proteins expressed in various organs, including the brain, in which play a role in several aspects of the nervous system development and stability among the inhibitory and excitatory synaptic transmission (Koropouli and Kolodkin, 2014). In addition, 7,024 protein domains were observed (e-value = 0.05), among the most enriched, were domains involved in protein-protein interactions, including PDZ and a considerable number of duplicated zinc-fingers. Nonetheless, proteomic studies have demonstrated that even though teleost species underwent a further WGD than mammals, which caused the duplication of genes in their genome, the zebrafish showed fewer PSD protein families, hence lower genome complexity (Bayés et al., 2017).

Only 2% (584 transcripts) of the data were classified as many:many, which might illustrate to a greater extent the high conservation in vertebrates after the TSGD. Further, studies have proposed that the expansion of the synapse proteome has occurred by increasing the number of already existent protein domain types through gene family duplication and diversification, instead of giving rise to new proteins (Emes et al., 2008).

The proportion of transcripts that had an ortholog with genes expressed in the mouse brain, synaptosome (SYN) and postsynaptic density (PSD) was next examined using the supplementary proteome data from (Sharma et al., 2015) (brain orthologs) and the G2C database (Croning et al., 2008) (SYN and PSD orthologs). 57% (15,535 transcripts) of the *de novo* assembly had an ortholog with a gene expressed in the mouse brain, 25.4% (6,915 transcripts) had an identifiable ortholog with a mouse SYN gene and 17.6% (4,782 transcripts) in the PSD.

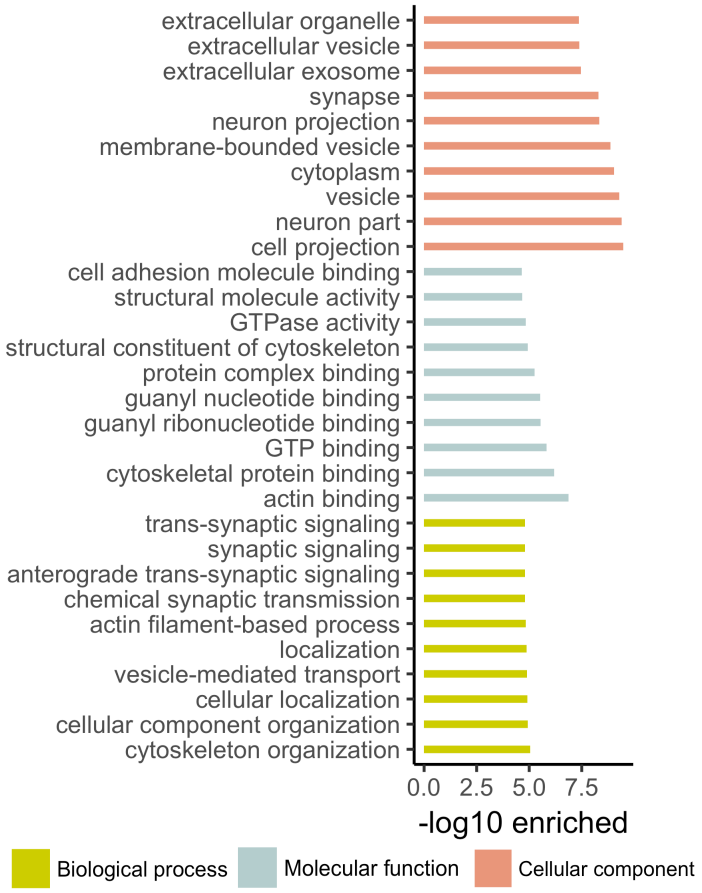
153 PSD genes were not found in the assembly, and also did not have any ortholog

with the *Danio rerio* reference transcriptome (GRCz10) using the biomaRt bioconductor package. Therefore, this gene set were determined as mouse specific, which might imply gene gain in mammals or gene loss in teleost fishes, and in turn could have led to mammalian PSD functional divergence (See Appendix C.1 for a full list of the 153 genes and their function). In order to investigate a potential significance in the synaptic functioning, GO terms and pathways were investigated, as seen in Figure 4.12. Enriched terms related to synapse, synpatic signalling and transmission, and neuron projections were found to be enriched, along with vesicle-mediated transport and receptor binding. This latter were not unlikely to be observed, since vesicle-mediated trafficking is paramount for secretion of neurotransmitters during synaptic transmission.

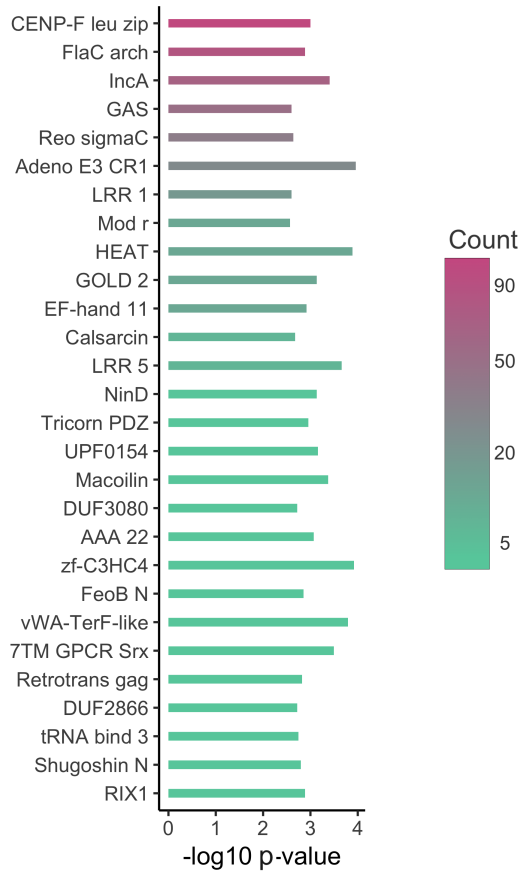
Additionally, within this set of mouse-specific PSD genes, 737 (unique) Pfam domains were observed (e-value cut-off of 0.05). Among the most enriched domains found were leucine-rich repeats (LRR), which were identified in the mammalian homer scaffolding protein 2. LRR domains are known to be localized at the postsynaptic side of excitatory synapses, where they interact with a vast number of PSD proteins (such as PSD-95), including NMDA receptors. Regulation of presynaptic and postynaptic elements is achieved by LRR proteins during development of axons, dendrites and synapses. Thereby, the role of these elements is critical for the control of synaptic connections into functional neural circuits. Alterations of LRR proteins have been linked to impairment of learning and memory, abnormal startle response in transgenic mice, schizophrenia, bipolar disorders and Rett syndrome in humans (de Wit and Ghosh, 2014; Woo et al., 2009; Shiraishi-Yamaguchi and Furuichi, 2007).

SNARE proteins were also observed among the most enriched domains, such as IncA. Mammals possess over thirty SNARE family members, which play an essential role over intracellular membrane trafficking and membrane fusion. In the presynaptic terminal, vesicles containing neurotransmitters are fused in a calcium-dependant manner with the membrane to cause the release of their content into the synaptic cleft. SNARE fam-

ily proteins are key in long-term modulation of synaptic strength, which is associated with learning and memory. Among mouse-specific PSD genes containing SNARE domains, some did not have an identifiable ortholog with the zebrafish, such as: *Tpm1, Hook3, Tmem109, Homer2, Hap1, Zwint, Crocc*. This is surprising as proteins involved in the process of neurotransmitter influx are conserved in metazoans and choanoflagellates such as *Monosiga brevicollis* and *Salpinogea rosetta*, which possess homologs of the three SNARE proteins (synaptobrevin 2, syntaxin 1 and SNAP-25) (Burkhardt, 2015; Paumet et al., 2009; Chen and Scheller, 2001).



(a)

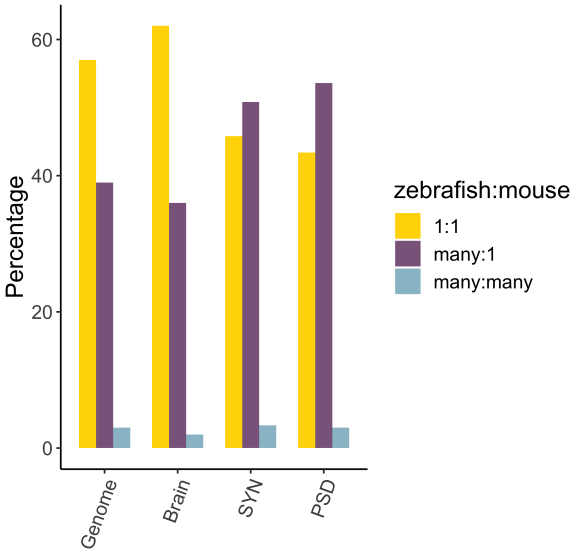


(b)

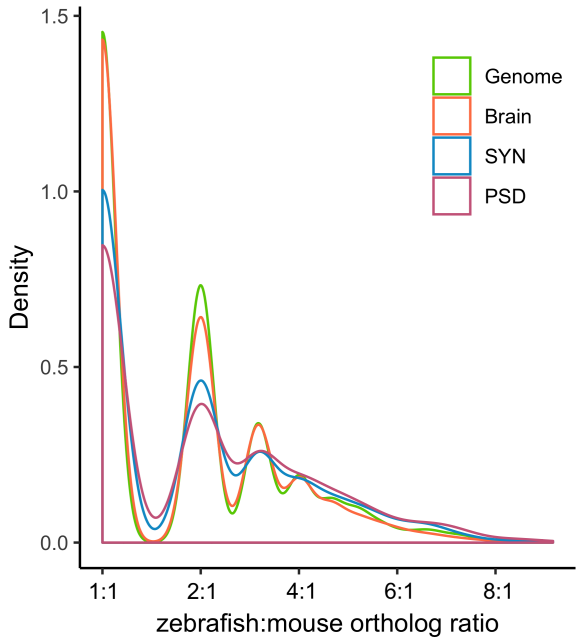
Figure 4.12: **PSD mouse-specific genes.** a) 153 mouse-specific genes were annotated for GO terms using R analysis tools. P-values were estimated based on hypergeometric distribution and adjusted FDR, applying a cut-off of 0.05. Highly significant enriched GO terms with a minimum genes cut-off of 2 were selected and plotted against their negative log10 P-value. (b) Pfam domains were determined using HMMER and R tools. Enriched domains were estimated and plotted using an P-value of cut-off of 0.05 and a FDR cut-off of 0.01. Domains are coloured according to the number of appearance and plotted against their -log10 P-value.

Moreover, as an effort to compare orthology classes (zebrafish:mouse; 1:1, many:1 and many:many) in the PSD, SYN and brain transcript ratios, percentages were determined. A larger proportion of many:many orthologs were detected in the brain, SYN and PSD compared to the whole genome. A higher percentage of “true orthologs" was identified within the brain orthologs, which was also coupled with a lower number of many:1 (Figure 4.13). Yet, this last orthology class was increased in the SYN, and interestingly, in the PSD as well, even though that the zebrafish PSD proteome is 17% smaller than the mouse PSD. Bayés et al., 2017 obtained similar results using proteomic data, suggesting that the synaptic proteome is more strongly conserved than in other tissues.

Genome duplication is a critical evolutionary mechanism which leads to the origin of novel functions. Studies have compared a substantial number of teleost gene families with retained duplicates along with singleton gene families. Revealing that genes with retained duplicates were considerably longer (27.9-38.2%) and contained a higher number of functional domains (20.5-26-5%) than singletons. Hence, genes which encode longer proteins, and which also have a greater number of functional domains were selected to be retained (Guo, 2017). These duplicated genes have a greater biological importance and its sequence is subjected to a stronger functional constrain (Jordan et al., 2004).



(a)



(b)

Figure 4.13: **Orthology types distribution and ratios within brain, SYN and PSD.** a) For each *de novo* transcript that obtained a unique BLASTp hit (e-value cut-off of e-10) with *D. rerio* orthology classes were estimated with mouse brain, SYN and PSD genes, as (zebrafish:mouse) 1:1, many:1and many:many. (b) Distribution of pair of orthologs (zebrafish:mouse) expressed in the brain, SYN and PSD.

4.6 Transcript expression in brain tissues

Transcript abundance was quantified for each contig in the three brain tissues (hindbrain, olfactory lobe and optic lobe), and all four WB replicates, used in this study. To achieve a better quantification accuracy, StringTie (alignment-dependant) (Pertea et al., 2015) and Kallisto (alignment-independant) (Bray et al., 2016) quantification tools were used and outputs were compared (Figure 4.14). Both tools displayed comparable results with minor discrepancies, yet Kallisto computed marginally higher expression levels and fewer lowly expressed transcripts in all samples.

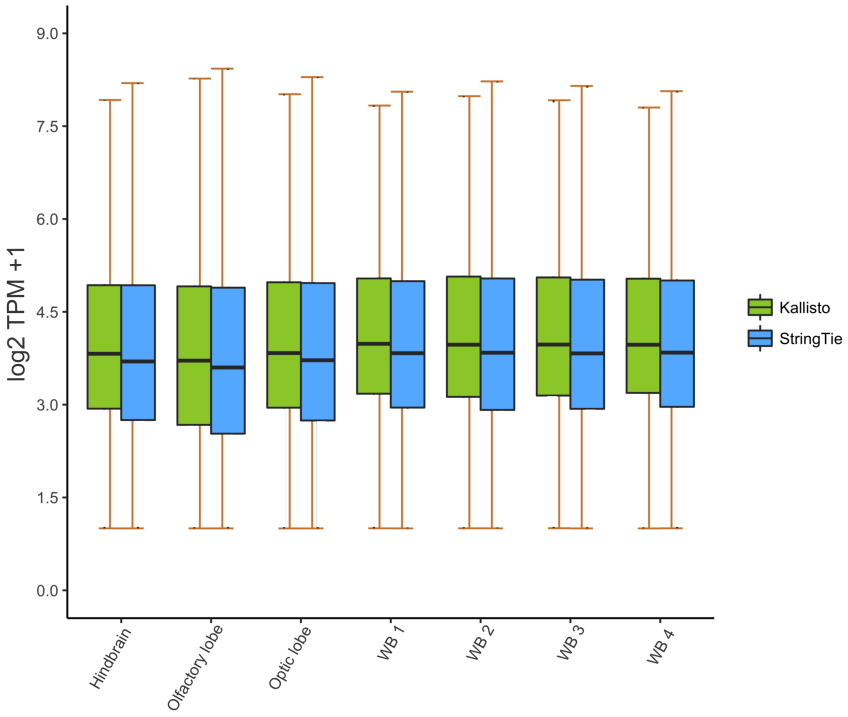


Figure 4.14: **Representation of transcript expression using Stringtie and Kallisto.** Transcript abundance were log2(+1) transformed, and transcripts whose expression was >0.5 were removed from the dataset. StringTie transcripts expression levels are computed in blue, while Kallisto's are in green.

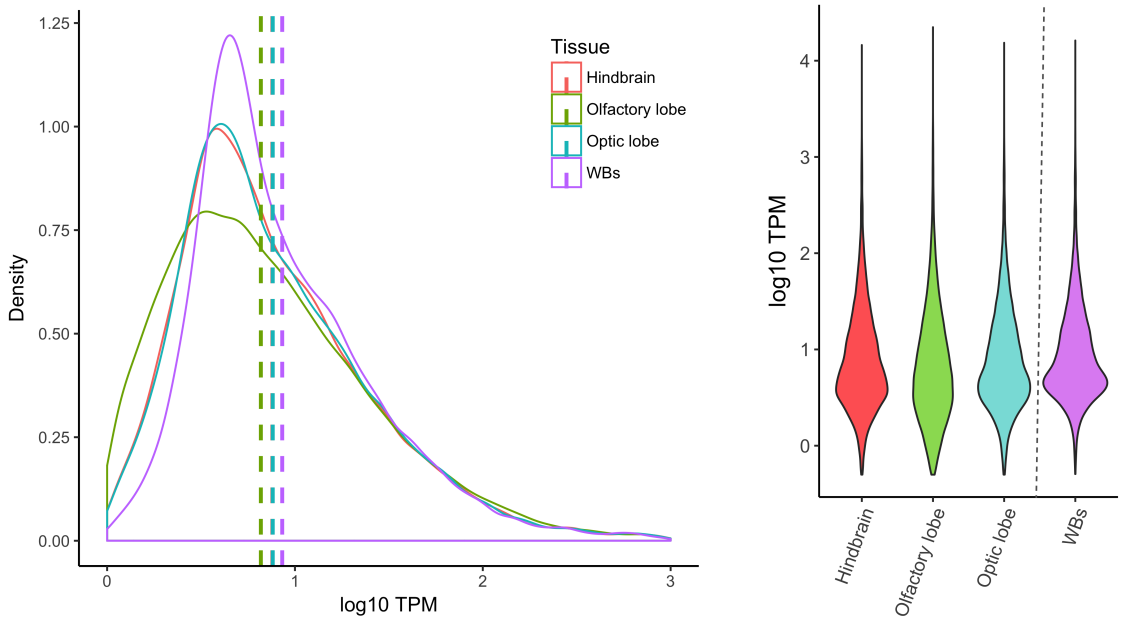
Considering the comparison of the models that have been developed for abundance estimation are out of scope in the present research. Kallisto was the tool of choice, since its alignment-independent quality, made its run substantially faster using min-

imal memory requirements. Hence, for further analysis in this research, Kallisto usage eliminates some computational bottlenecks. Moreover, as an attempt to obtain a more stringent expression analysis, transcripts computed by Kallisto that did not meet the pipeline annotation criteria were removed from further analysis, together with the ones which exhibited low expressed quantification (<0.5 TPM).

To determine the variation in the expression of the three different zebrafish brain tissues and four WB replicates, the annotated transcriptome (27,273 transcripts) from Dammit and BLASTp was used. Summaries of transcript expression levels in all tissues are seen in Table 4.9 and Figure 4.15.

Table 4.9: **Number of annotated transcripts at different TPM threshold.**

Tissue	>0.5 TPM	>1.0 TPM	>5.0 TPM	>10 TPM	>25 TPM
Hindbrain	26,352	25,830	15,117	9,190	4,055
Olfactory lobe	25,364	24,013	13,432	8,423	3,938
Optic lobe	26,367	25,840	15,111	9,190	4,116
WB's	26,409	26,002	16,985	9,905	4,272



(a) (b)

Figure 4.15: **Distribution of transcript expression levels,** estimated as TPM (transcripts per million) in three brain tissues (hindbrain, olfactory lobe and optic lobe) along with the mean expression of four whole brain replicates (indicated as WBs). Lowly expressed transcripts (<0.5 TPM) were filtered out, in addition to unnanotated transcripts output by Kallisto. a) Density distribution of log10 transformed transcripts illustrating differences in expression in all samples, along with the mean expression of each shown as dash lines. b) Violin plot showing distribution of log10 transformed TPMs and its probability density.

At a TPM threshold of 0.5, all three tissues, and the mean expression of the four WB replicates showed similar expressions levels. The olfactory lobe showed fewer annotated transcripts (n=25,364) than the optic lobe and hindbrain (n=26,367 and 26,352, respectively). The WB replicates displayed the following number of expressed transcripts; 26,350, 26,417, 26,391 and 26,477 (mean=26,408.8), as expected the WB replicates obtained a larger number of transcripts.

The mean expression for the hindbrain, olfactory lobe and optic lobe transcripts was; 24.7 (SD=162.4; max=14,556.7), 26.0 (SD=188.9; max=22,305.8) and 24.4 (SD=159.7; max=15,351.8), respectively. With the olfactory lobe exhibiting the highest mean expression even though it also showed the fewest transcripts. All WB replicates showed

a mean expression of; 24.7, 23.8, 24.4 and 24.3 (mean=24.3; SD(mean)=155.8, max (mean)=16,905).

4.6.1 Highly expressed genes in whole brain replicates

The top 20 most enriched genes (mean TPM of 2,860) in the four whole brain replicates were identified. Among the top *D. rerio* genes that represented these *de novo* transcripts (Table 4.10), the majority were found expressed in the mouse brain, SYN or PSD. For example, ependymin the highest expressed gene identified, is implicated in neuroplasticity and neuronal regeneration. This gene is expressed in the cerebrospinal fluid of various vertebrates species, however, its expression has also been identified in invertebrate deuterostomes (Suárez-Castillo and García-Arrarás, 2007).

Likewise, a considerable number of the highly expressed genes are nervous system-specific, such as *Mbpa* (myelin basic protein). Mbpa is the second most abundant protein in CNS (after the proteolipid protein). The role this protein is to bond the cystosolic surface of the multilayered compact myelin, and interacts with a number of the highly expressed WB genes, such as actin and calmodulin proteins (Boggs, 2006). *Snap25a* (synaptosomal-associated protein 25), a member of the SNARE complex is crucial for the accelerated release of neurotransmitters from their synaptic vesicles. Its domain is highly conserved in metazoans and choanoflagellates (Burkhardt, 2015; Emes and Grant, 2012).

Table 4.10: **Top highly expressed genes in the whole brain replicates.**

Gene ID	Mean TPM	Protein	Function	Mouse orthology
epd	16904	Ependymin	Neuroplasticity and regeneration	-
mt-nd1	10587	NADH dehydrogenase	Mitochondrial respiratory chain	SYN
mt-co3	5160	Cytochrome c oxidase	Mitochondrial respiratory chain	SYN
mt-nd5	4796	NADH dehydrogenase 5	Mitochondrial respiratory chain	SYN
mt-nd6	3315	NADH dehydrogenase 6	Mitochondrial respiratory chain	Genome
mbpa	2819	Myelin basic protein a	Myelin adhesion	-
snap25a	2817	Synaptosomal-associated	Intracellular membrane fusion	PSD
actb2	2654	Actin, beta 2	Mediator of internal cell motility	PSD
cd59	2303	Complement defense 59	Inhibition of MAC assembly	-
ckbb	2054	Creatine kinase, brain b	Cellular energy	SYN
ba1	2050	Hemoglobin subunit beta-1	Oxygen transport	Brain
aldocb	1785	Fructose-bisphosphate aldolase C-B	Glycolysis	-
stmn1b	1710	Stathmin 1b	Tubulin binding	SYN
slc25a5	1634	Solute carrier family 25	Transmembrane transport	-
calm1a	1607	Calmodulin 1a	Calcium ion binding	-
gapdhs	1569	Glyceraldehyde-3-phosphate dehydrogenase	Glycolysis and gluconeogenesis	PSD
rpl36a	1563	Ribosomal protein L36A	RNA binding	Brain
marcksl1b	1514	MARCKS-like 1b	Neural development	SYN
cox5b	1357	Cytochrome c oxidase subunit 5B	Mitochondrial respiratory chain	SYN
rplp1	1318	Ribosomal protein, large, P1	Elongation step of protein synthesis	PSD

Several highly expressed WB genes are encoded by the mitochondrial genome and involved in the mitochondrial respiratory chain, e.g.,*mt-nd1, mt-co3, mt-nd5, mt-nd6 and cox5b*. This might be associated to either biases in the sample quality or its biology. Mitochondria are abundant in the cytoplasm of mammalian cells, such as neurons, which survival depends on the production of mitochondrial energy. Alterations in mitochondria results is associated to a multi-systemic disease, however, the brain is most susceptible to these defects, implying that mitochondria mediates elementary phases of brain function (Picard and McEwen, 2014).

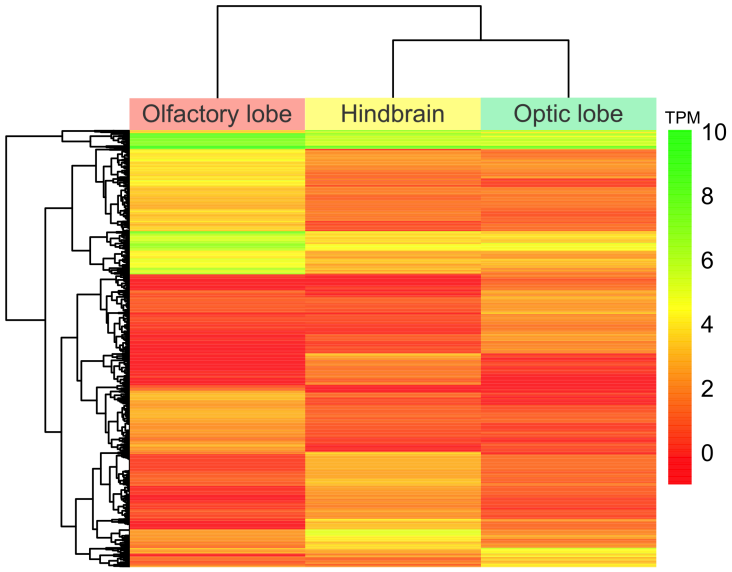
Mitochondrial dysfunction causes ATP depletion and the accumulation of super-oxide radicals prompting an abnormal cycle of oxidative stress. Neurodegenerative diseases, particularly Alzheimer's disease (AD) has been associated with deficiency of the mentioned genes, since these trigger cerebral hypometabolism and defective homeostasis in the redox status causing neuronal cell death (Grimm et al., 2016; Kim et al., 2001; Nicholls and Budd, 2000; Maurer et al., 2000; Chandrasekaran et al., 1994). Genes such as *cd59* (complement defense 59) (Yang et al., 2000), *ckbb* (creatine kinase, brain b) (Aksenov et al., 2000), and the calmodulin gene family (O'Day and Myre, 2004) have also been linked to neurodegenerative diseases.

4.6.2 Tissue-enriched and specific gene expression

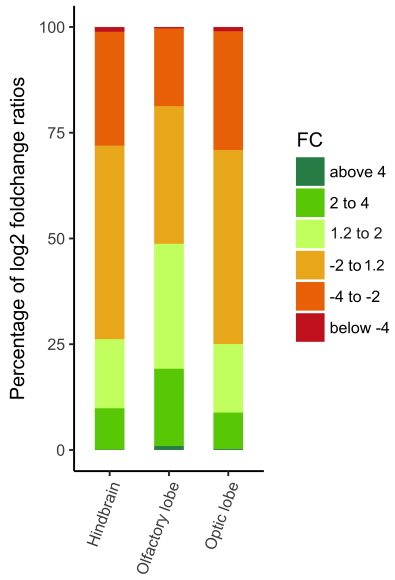
To identify transcripts differing in expression between brain regions, transcripts were classified into three categories; tissue-enriched, tissue-specific and non-specific. A threshold of 1.2 fold-change was selected to determine tissue-enriched transcripts, as a more stringent threshold (e.g., 1.5) generated insufficient transcripts for downstream analyses. Therefore, transcripts that exhibited at least 1.2 fold-change higher in a particular tissue, and concurrently exhibited at least 1.2 fold-change lower for the same transcript in the other two tissues, were classified as "tissue-enriched".

By the same token, transcripts expressed in one of the three tissues with no expression in the other two, were determined as “tissue-specific". Lastly, transcripts that were broadly expressed in all tissues, were categorized as tissue "non-specific".

Initially, tissue-enriched transcripts were determined and further examined. A similar expression pattern was observed between the hindbrain and optic lobe-enriched transcripts (Figure 4.16), which might be consequence of their close localization within the



(a)



(b)

Figure 4.16: **Foldchange tissue enrichment**. a) Heatmap showing expression levels of tissue-enriched transcripts; FC > 1.2 in one tissue and FC < 1.2 for the same transcript in the other two tissues. Expression values are transformed TPM. Dendogram clustering on the X-axis indicates sample similarity, while Y-axis dendogram clustering groups transcripts with similar expression (b) Distribution of tissue-enriched transcripts showing in the Y-axis percentages of foldchange ratios. Different colours indicate relative expression levels.

brain. Higher expression levels were seen in the olfactory lobe-enriched transcripts, along with a considerable higher amount of enriched transcripts in comparison with

the other two brain tissues.

When comparing tissue-specific transcripts, a greater co-expression of transcripts was identified in the optic lobe and hindbrain (n=1,152; 4.3% of the tissue-specific transcripts) than both tissues with the olfactory lobe (n=228; 0.8% and 233; 0.9%). The number of tissue-specific transcripts in the optic lobe and hindbrain were higher than the olfactory lobe (Figure 4.17). Lastly, 24,784 transcripts were determined to be non-specific for any tissue.

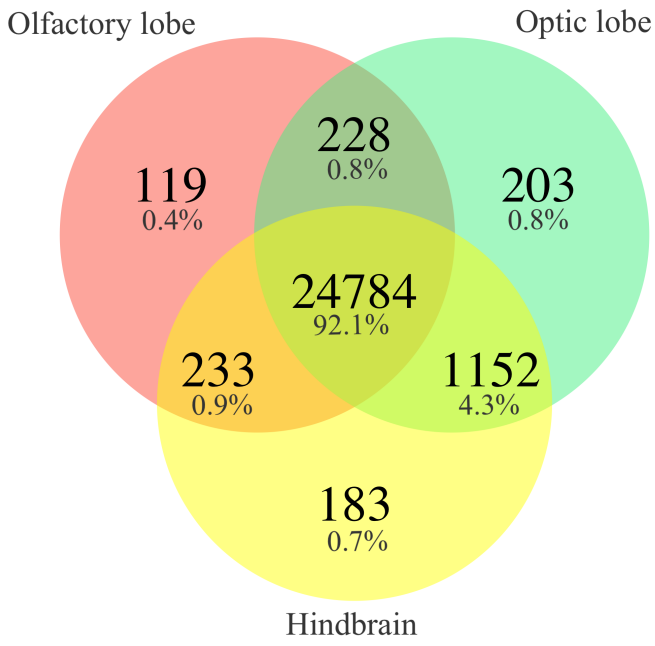


Figure 4.17: **Tissue-specific Venn diagram**. Tissue-specific transcripts; 203 optic lobe; 183 hindbrain; 119 olfactory lobe. 24,784 annotated transcripts were broadly expressed in the three tissues.

Gene orthology enrichment was determined for the tissue-enriched transcripts and proportions of orthologs expressed in the mouse brain, SYN and PSD. (Table 4.11). This showed a more similar distribution between the optic lobe and hindbrain. A high-

er percentage of PSD mouse orthologs was revealed in the optic lobe, with fewer in the olfactory lobe.

Table 4.11: **Tissue-enriched and tissue-specific orthology statistics.**

Tissue	Trans-cripts	Genes	%Brain	%SYN	%PSD
		Tissue-enriched			
Hindbrain	228	219	43.8	22.8	16.9
Olfactory lobe	424	361	41.2	15.2	11.6
Optic lobe	218	207	52.0	23.1	17.9
		Tissue-specific			
Hindbrain	183	159	50.0	28.3	23.3
Olfactory lobe	119	115	57.5	25.2	18.3
Optic lobe	203	182	53.2	29.1	20.3

Orthology relationships (zebrafish:mouse; 1:1, 1:many, many:many and unique) were determined for the tissue-enriched and tissue-specific transcripts in each tissue 4.18). The optic lobe displayed the highest percent of 1:1 (“true orthologs") orthologs in both tissue-enriched and tissue-specific. In the other hand, the hindbrain displayed the highest proportion of zebrafish-unique transcripts.

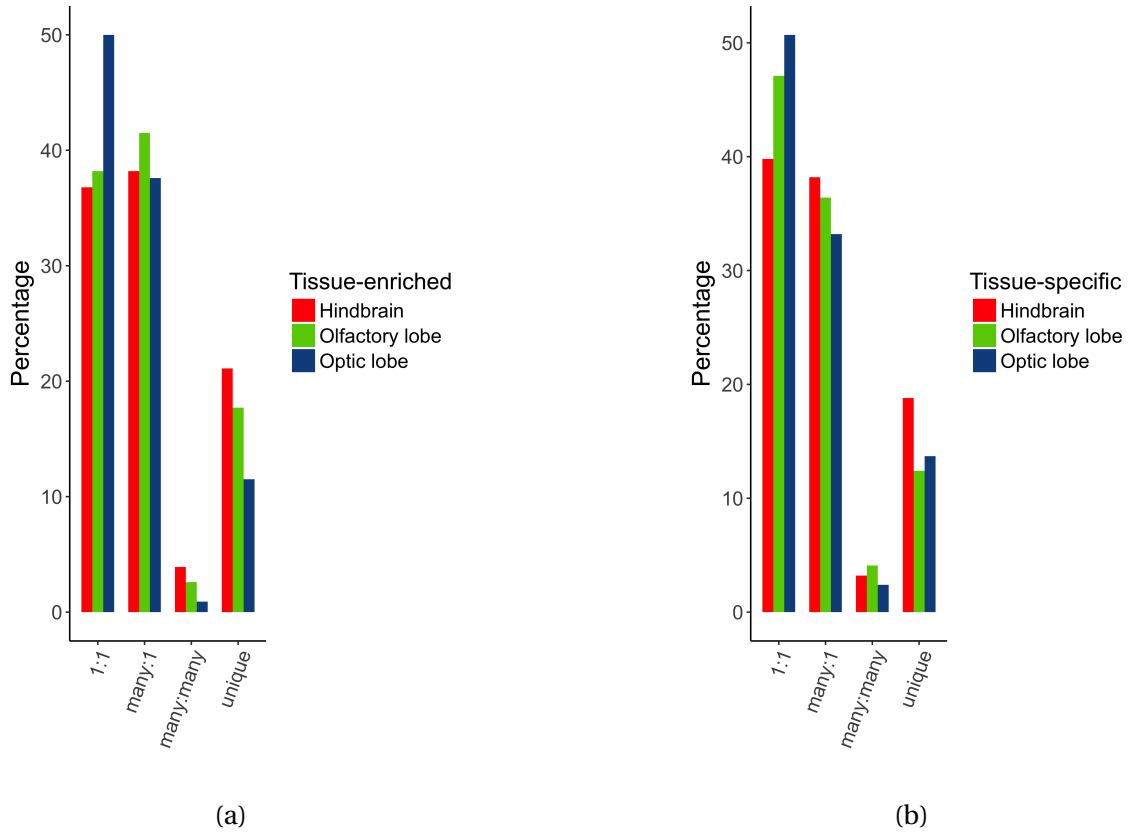


Figure 4.18: **Tissue-enriched and tissue-specific orthology types distribution,** between zebrafish and mouse. For each set of transcripts, the zebrafish:mouse portion was obtained a) Tissue-enriched transcripts showed >1.2 FC expression levels b) Tissue-specific transcripts were only expressed in one tissue.

Enriched GO terms were examined initially for the olfactory lobe-enriched and restricted sets (Figure 4.19). Ontologies such as “dentrite", “neuron", “neuropeptide receptor activity", or “SNARE binding" were observed only in the olfactory-enriched set. Yet, a particular enrichment of genes expressing G-protein-coupled receptor (GPCRs), celullar signalling was observed in both olfactory-enriched and specific classes. GPCRs play a significant role in a vast quantity of extracellular signalling pathways including, sensory perception (smell), neurotransmission and cell communication (Tuteja, 2009).

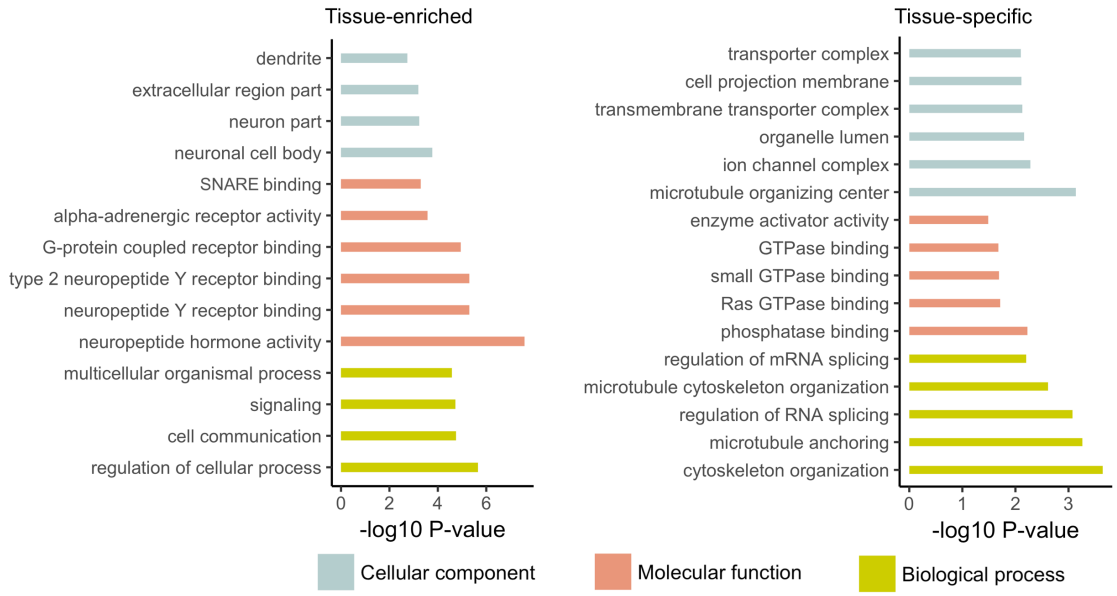


Figure 4.19: **Olfactory lobe-enriched and specific GO analysis.** Figures showing the distribution of GO terms with statistical significant differences (hypergeometric distribution applying a p-value and FDR cut-off of 0.5). The two panels correspond to 361 olfactory lobe-enriched (left) and 115 olfactory lobe-specific (right). Different colours are used for each GO term .

Odorant receptor (OR) genes constitute the largest known family of GPCRs in vertebrates, which have been highly conserved during evolution (Miyasaka et al., 2014; Kratz et al., 2002). OR genes are widely expressed in neurons of the olfactory epithelium, allowing the zebrafish to detect and discriminate an extensive range of water-soluble molecules. This underlies complex neuronal circuits, which are transferred from the olfactory bulb to different forebrain regions and finally translated into correct output responses (Miyasaka et al., 2013; Friedrich et al., 2004).

The optic lobe enriched and restricted sets revealed enriched terms more relevant to synaptic biology, such as “postsynaptic specialization", “excitatory synapse", “postsynaptic density", “synapse", among others (Figure 4.20). The entire biological process ontologies for the optic lobe, were related to neurogenesis or nervous system development. It was additionally observed several GABA receptor activity in both cellular component and molecular function terms. As neuronal activity is regulated by the re-

lease of GABA and glutamate this may explain an enrichment of these terms (Spitzer, 2006).

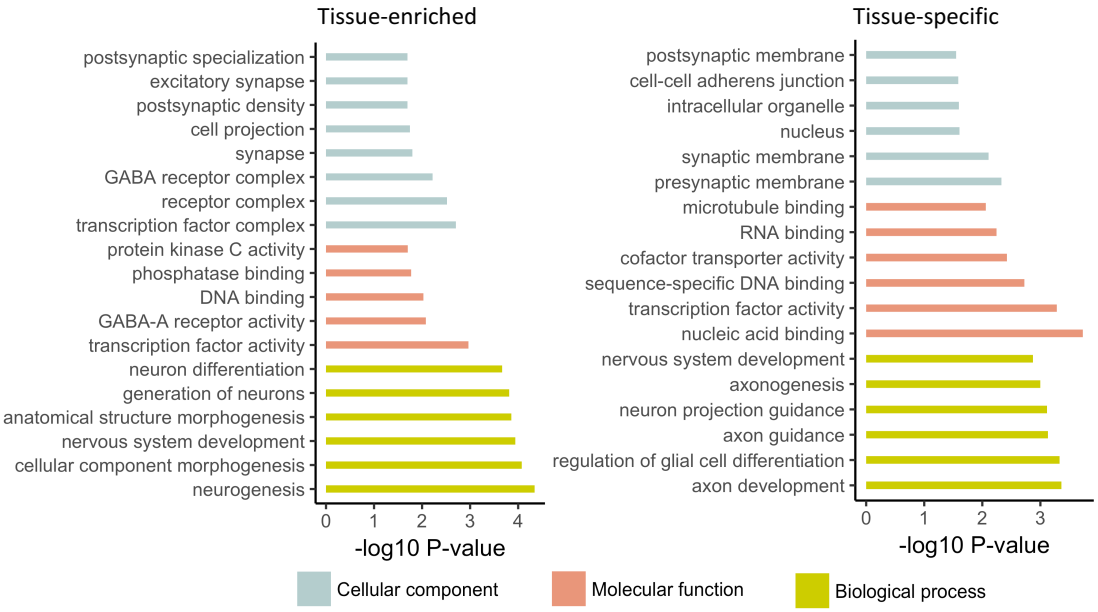


Figure 4.20: **Optic lobe-enriched and specific GO analysis.** Figures showing the distribution of GO terms reflecting statistical significant differences (hypergeometric distribution applying a p-value and FDR cut-off of 0.5). The two panels correspond to 207 optic lobe-enriched (left) and 182 optic lobe-specific (right). Different colours are used for each GO term.

Studies have demonstrated a persistent neurogensis ability throughout the nervous system during the adulthood of teleost fishes compared to any other vertebrate. Zebrafish have a neuronal proliferation rate with a potential 10 to 100 times greater than mammals (Ganz and Brand, 2016; Zupanc et al., 2005). These new neurons originate from cells with stem cell-like characteristics, that in contrast with the mammalian brain where adult neurogenesis only occurs in the hypothalamus (Sorrells et al., 2018; Yoo and Blackshaw, 2018), teleosts have an exceptional number of proliferation zones all over the brain (Nieuwenhuys et al., 2014; Zupanc, 2011; Alunni et al., 2010). The optic lobe in anamniotes (amphibians and fishes) is of special interest in vertebrate neurogenesis, since they possess the capability to constantly generate and replace adult retinal cells with additional regrowth of optic axons with complex brain connections

(Ito et al., 2010; Becker et al., 2004; Becker and Becker, 2000).

Lastly, enriched GO terms in the hindbrain sets revealed ontology enrichment particularly for calcium ion binding (Figure 4.21). Genes associated with this ontology were expressed in both enriched and restricted sets. This is of special interest, since voltage-gated calcium is a pivotal part in neuronal development, synaptic transmission and plasticity. Imbalances in genes expressing this ontology (such as *cacna2d3, cacng1b*) have been demonstrated to cause dysregulation in homoeostasis of voltage-gated calcium, which in turn, causes several pathological mechanisms, mainly neurodegenerative disorders, including Parkinson's disease, bipolar disorders, schizophrenia and Alzheimer's disease (AD) (Sulzer and Surmeier, 2013; Green et al., 2010; Ferreira et al., 2008).

Studies in subjects with AD have shown increased levels of intracellular calcium. An increase of amyloid metabolism is the main cause of AD, yet evidence has suggested that amyloid proteins induce calcium influx into neurons that alter neuronal excitability (Villela et al., 2016; Berridge, 2010; Small, 2009). Further investigations in the mechanisms of these genes could be a key step in the pathogenesis of neurodegenerative diseases.

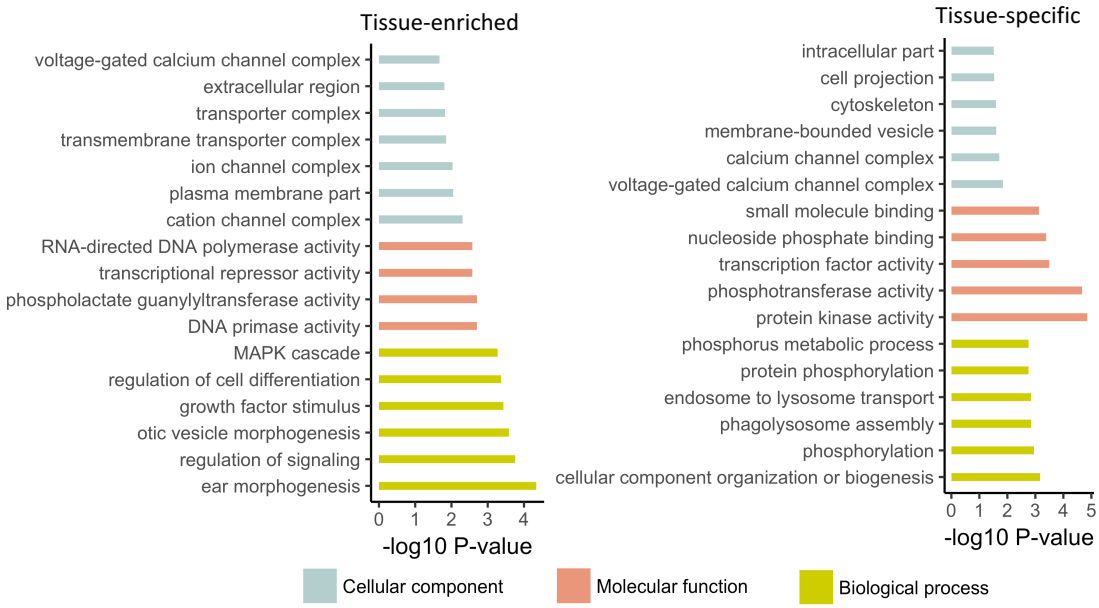


Figure 4.21: **Hindbrain-enriched and specific GO analysis.** Figures showing the distribution of GO terms reflecting statistical significant differences (hypergeometric distribution applying a p-value and FDR cut-off of 0.5). The two panels correspond to 219 hindbrain-enriched (left) and 159 hindbrain-specific (right). Different colours are used for each GO term.

Additionally, enriched terms in the hindbrain, such as “ear morphogenesis”, “otic vesicle morphogenesis” and “Notch signaling pathway”, were particularly predominant with the expression of genes (i.e., *jag1b, tmie, msx3, fgf8a, stm, msx3, fgf8a*), which are crucial to the inner ear development, but also linked to inner ear defects (Ma and Zhang, 2015; Gleason et al., 2009; Wang et al., 1996). Most of these genes belong to the ancient gene family Homeobox (Hox genes) that are pivotal in the regulation of embryonic development of the CNS (Wang and Lufkin, 2005). Phylogenetic analysis revealed four Hox clusters in mammals, while teleosts encode seven Hox clusters. Therefore, the exploration of zebrafish Hox gene duplicates can improve the understanding of multiple gene loss and retention throughout evolution (Moens and Prince, 2002; Amores et al., 1998).

Investigation of PSD genes that were enriched and specific for each tissue showed a similar number of combined tissue-specific and enriched PSD genes expressed in the

olfactory lobe (n=55) and optic lobe (n=56), the hindbrain expressed a considerably fewer amount (n=25).

The number of each GO categories were considered for each tissue and showed a substantially higher number of biological processes and cellular components in the optic lobe (n=60 and 21, respectively) than the olfactory lobe (n= 44 and 6) and hindbrain (n=6 and 11). Yet, the hindbrain and optic lobe displayed the same number of molecular function (n=17), whereas the olfactory lobe showed the fewest (n=6).

As expected, most of the PSD enriched and specific-optic lobe genes showed and enrichment in ontologies related to synaptic functioning, namely; “synapse", “PSD specialization", “neuron", and functions encompassing actin filament binding, protein kinases and receptor signalling activities. The optic lobe PSD GO enriched ontologies displayed several developmental processes related mainly to the nervous system, therefore reflecting its ability for neurogenesis. The olfactory lobe PSD exhibited enriched GO terms particularly in various signalling activities and membrane trafficking, but also in kinases and phosphatases pathways. Conversely, the hindbrain PSD genes were mostly involved in terms associated with the cytoskeleton organization and RNA binding proteins. The latter probably reflects localized translational control, since effective synaptic function relies on the continuous regulation of local proteins. This is achieved by the synaptic capacity to undergo local translation (Rangaraju et al., 2017).

4.6.3 Transcripts with highly variable expression

In addition to gene specificity, gene expression variability contributes to significant understanding of how genes function in biological processes. It is considered as an essential element of population fitness and adaptability. Moreover, researchers have associated genes with the highest expression variability to the onset of numerous dis-

eases. In fact, analyses of enriched pathways for gene expression variability within the brain are recognized to be associated with neurodegenerative disorders, including AD, Parkinson's disease (PD), dementia and schizophrenia (Ran and Daye, 2017; Zhang et al., 2015; Mar et al., 2011; Li et al., 2010a). For that reason, studying expression variability could shed light to the evolution and differentiation of vertebrate gene expression.

For each transcript the expression variability was calculated by computing its coefficient of variation (CV). This technique is a standarized measure of variability that is estimated by taking the ratio of the standard deviation (SD) and the average mean expression per each transcript across the three tissues, based on a generalized linear model (GLM) (Figure 4.22). To minimize bias driven by low expression levels, as an excess of zeros might distort model estimation, transcripts of which TPM was lesser than 0.5 were excluded.

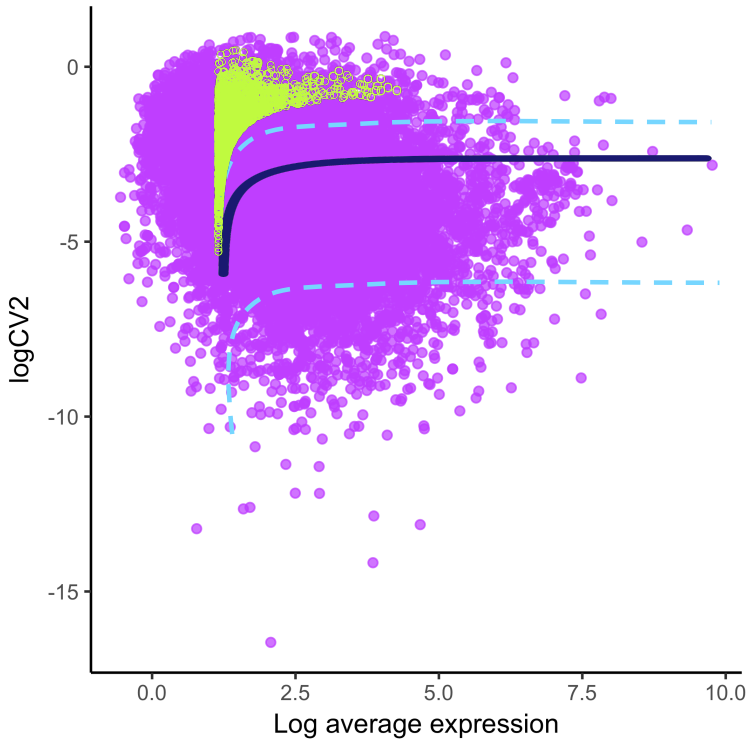


Figure 4.22: **Squared coefficient of variation (CV2)**. For each transcript the CV2 (ratio of the standard deviation of its expression to the mean value) was calculated. The x-axis shows the log mean expression of the 3 tissues; the y-axis represents the log transformed CV2. The solid blue curve represent the fitted variance-mean dependence; the dashed lines reflects a 95% confidence interval; green dots correspond to transcripts which CV2 is significantly higher than 50% (CV2 > 0.25).

1,189 transcripts were determined as significantly highly variable (p-val < 0.001), corresponding to 1,081 genes and 1,004 mouse orthologs. GO terms were analysed (Figure 4.23), and showed an enrichment of biological processes, on which the majority were involved in development. A substantial number of coordinated changes in gene expression are implicated in the developing mechanisms of multicellular organism, over various cell and tissue types (Francesconi and Lehner, 2014). Several genes involved belong to the aforementioned homeobox gene family, which encode transcription factors (TFs) that have a pivotal role in regulating axon guidance and synaptic formation (Polleux et al., 2007; Meyer, 1998).

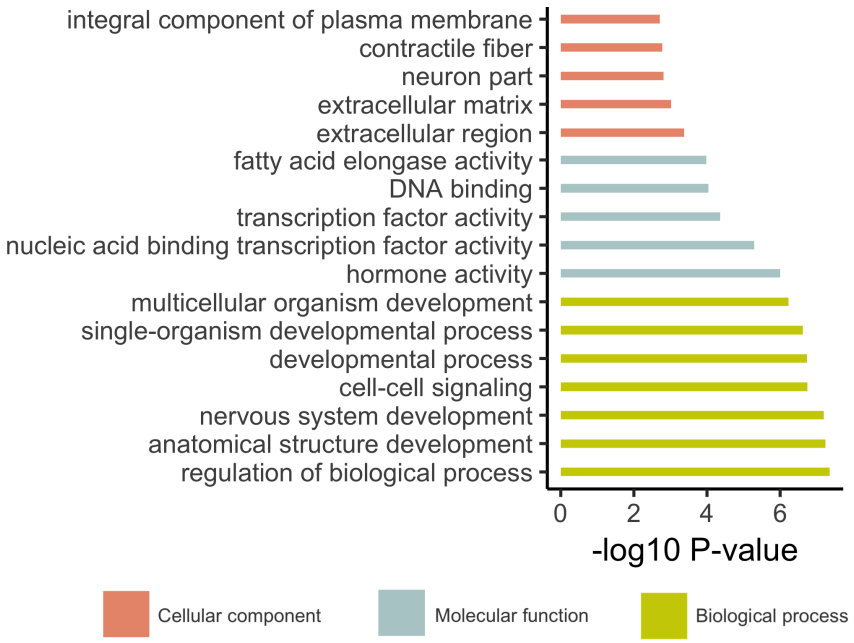


Figure 4.23: **GO analysis for highly variable genes**. 1,081 were determined as highly variable across the three tissues. Enriched GO terms were analysed, reflecting statistical significant differences (hypergeometric distribution applying a p-value and FDR cut-off of 0.5)

TFs were among the most enriched molecular function terms found. In fact, studies have observed that genes involved in regulatory functions, such as TFs, commonly display higher levels of variability. Evidence have exhibited a rapid evolution of transcription factors as an outcome of selection, contributing to novel phenotypes (Lin et al., 2017). For example, pelvic loss of the three-spined stickleback fish occurred by regulatory mutations during vertebrate evolution, deleting a tissue-specific transcription factor (Shapiro et al., 2004; Chan et al., 2010).

Fatty acid enzymatic activity was also noted enriched in the highly variable genes. Studies have associated dysregulation of unsaturated fatty acid metabolism in the brain of patients with different degrees of AD, as well as neuronal ceroid lipofuscinoses (Snowden et al., 2017; Vesa et al., 1995). Fatty acid elongases (elovl), are essential for the biosynthesis of the retina and brain, keeping the structural and functional integrity of synapses (Hopiavuori et al., 2016; Astarita et al., 2011).

Within cellular component ontologies, genes were found to be expressed in the cell-membrane or the extracellular region. A study found that genes with high expression variability were involved in signal transduction pathways at the periphery of the cell (Mar et al., 2011). Members of the CLC family of chloride channel/transporters were amongst the most enriched genes, which mutations have been associated with epilepsy and blindness (D'agostino et al., 2004). To further investigate whereby the TSGD influenced gene variability in the PSD, the distribution of orthology types was next examined.

There was no difference among the highly variable genes expressed in the genome, brain, SYN and PSD. However, the majority of the SYN and PSD highly variable genes were many:1 (zebrafish:mouse), in comparison to all orthologs, in which the majority had a 1:1 relationship (Figure 4.24). Therefore, it might be suggested a role of whole genome duplication and increased gene variability in the synaptosome. In this regard, orthologs are homologs expressed in different species (1:1 relationship) that evolved from a common ancestral gene, and which function is generally retained. On the contrary, paralogs are homolog genes that evolved by gene duplication (many:1) and commonly endow functional innovations that are conserved for environmental adaptation (Peters et al., 2012). It is thereby possible that gene duplication is a major force that contributes to the relaxation of biological constrains leading to genes with highly variable expression, or high variation contributes to gene retention.

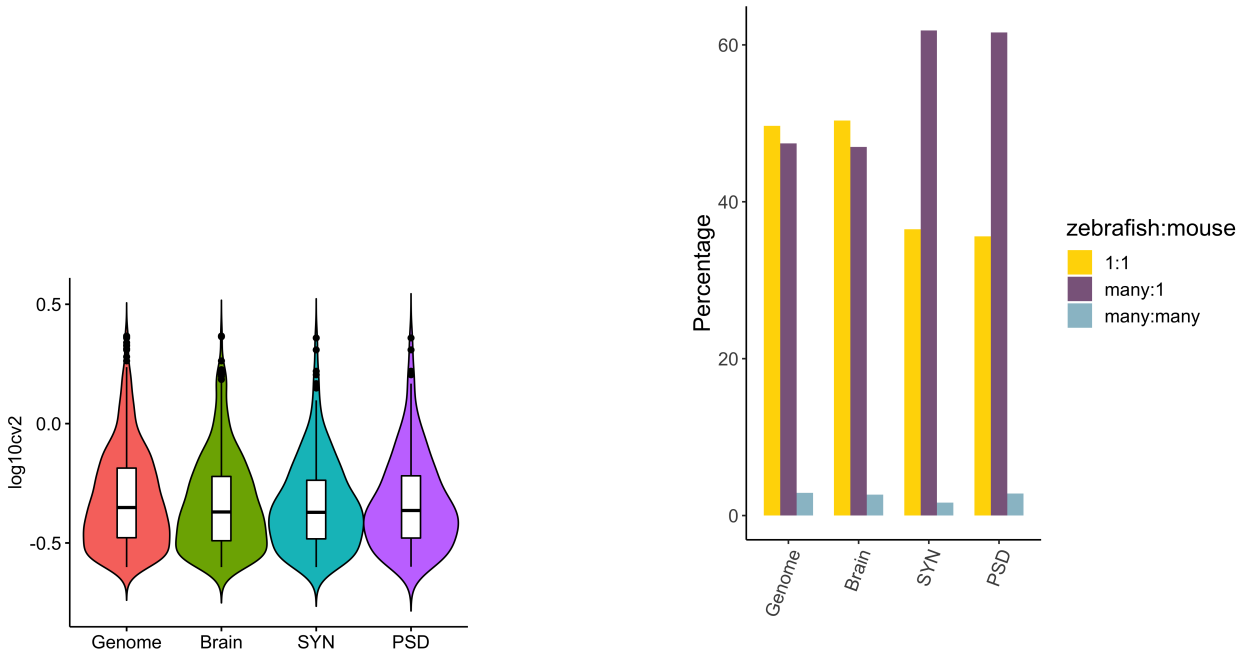


Figure 4.24: **Orthology distribution across highly variable genes**. a) log10 squared CV distribution across mouse genome, brain, SYN and PSD orthologs. b) For each highly variable gene (n=1,081), its orthology ratio (zebrafish:mouse) was determined (including unique zebrafish genes), brain, SYN and PSD.

The significance of gene duplication in the evolution of genetic novelty is a known concept, distinct copies of duplicated genes can be retained if both are advantageous. Through substantial sub-functionalization, duplicated genes might become specialized, conferring different expression patterns in terms of tissue specificity, hence a higher expression variability might be expected to allow a better adaptation to environmental change. Therefore, variability in gene expression may contribute to significant phenotypic evolution (Kliebenstein, 2008; Gu et al., 2004). A great proportion of highly variable genes were among the tissue-enriched sets (Figure 4.25). Particularly, 72% of the enriched-olfactory, were determined as highly variable.

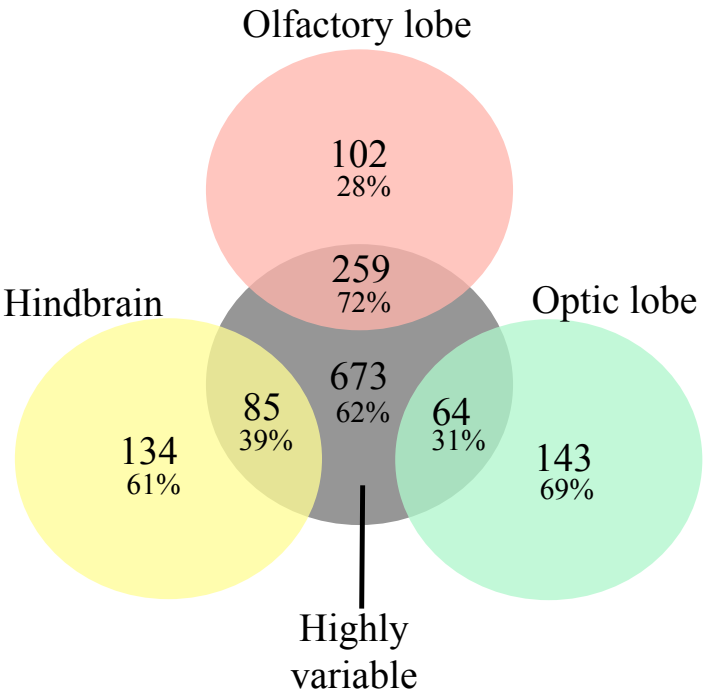


Figure 4.25: **Highly variable tissue-enriched genes**. Venn diagram reflecting the proportion of tissue-enriched genes that were additionally determined as highly variable.

4.6.4 Gene expression of key synaptic genes

The synaptic functioning is dependant of a cascade of protein activity, moreover, some proteins play a major role and is alteration or absence inhibits the correct synaptic activity. Such proteins are defined as key for the synapse functioning. The extent of key synaptic protein coding genes were further compared between the three tissues. Using literature exploration a set of key synaptic genes was compiled (Bayés et al., 2017, 2012) (Table 4.12) and HMMER was used to identify matching genes. Additionally, gene expression patterns were subjected to hierarchical clustering in order to identify co-expression of key synaptic genes in the different anatomical brain regions of the zebrafish. Despite the fact that the olfactory lobe clustered separately, in general similar

expression patterns were observed among the tissues (Figure 4.26). This conservation implies that the general synaptic machinery is evolutionary conserved. However, the handful of genes that do display inconsistent regional expression patterns, might reflect species-specific adaptations, particularly in the olfactory lobe.

Table 4.12: **Key elements of the postsynaptic density.** Synaptic fuctioning depends on the expression of such key elements.

Gene	**Protein description**
Ablim 1	Actin binding
Baiap2	Brain-specific angiogenesis inhibitor
Bdnf family	Brain-derived neurotrophic factor
Cacng2	Calcium channel
Camk family	Ca^{2+}/calmodulin-dependent protein kinase
Cntnap2	Contactin Associated (neurexin)
Dlg family	Disks large homolog (MAGUK member)
Dlgap family	Disks large-associated
Gabra family	Gamma-aminobutyric acid receptor subunit
Gria family	Glutamate ionotropic receptor AMPA type subunits
Grin family	Glutamate ionotropic receptor NMDA type subunits
Homer family	Homer scaffold
Iqsec family	IQ motif and Sec7
Magi family	Membrane-associated guanylate kinase
Mapk family	Mitogen-activated kinase
Ncam1	Neural Cell Adhesion
Nlgn family	Neuroligin
Nsf family	Vesicle-fusing ATPase
Nrxn family	Neurexin
Shank family	Proline-rich synapse-associated
Snap-25 family	Synaptosomal-associated
Stx family	Syntaxin (SNARE)
Syngap family	Synaptic Ras GTPase-activating
Vamp family	Vesicle-associated membrane protein (SNARE)

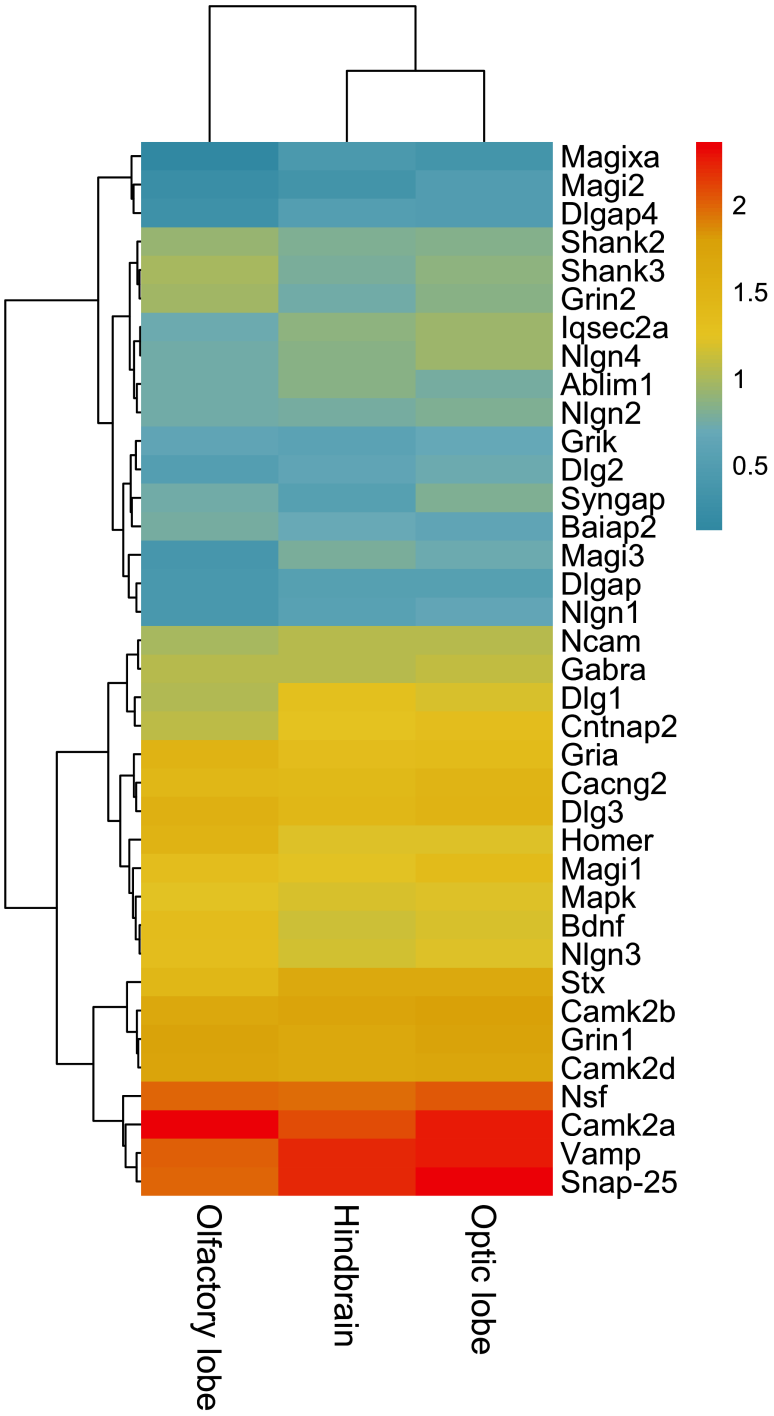


Figure 4.26: **Expression of key synaptic genes**. Synaptic fuctioning depends on the expression of such key genes. Tissues and genes were hierarchically clustered at the top (tissues) and left (genes) of the heatmap. Considering the presence of splice-forms, the mean expression value was obtained per gene, and its expression in TPM was log10 transformed. Different colours distinguish highly and lowly expressed genes in red and blue, respectively.

A large proportion of duplicates (many:1; zebrafish:mouse) was evident in our *de novo* assembly. As seen in Figure 4.13, and similarly to Bayés et al. (2017), compared to the whole genome, the many:1 orthology type is greater in the SYN and PSD. Duplicated genes that play essential roles in numerous cell processes, are comprised of more functional domains, together with a larger protein sequence. Such genes are more likely to be retained after a genome duplication (Bayés et al., 2017; Guo, 2017).

Between the most highly expressed synaptic genes found, included Neurexin/N-ethyl maleimide-sensitive factor (*nsf*), Synaptobrevin/Vesicle associated membrane proteins (*vamp*), Synaptosomal-associated protein 25 (*snap-25*) and calcium/calmodulin-dependent protein kinase family (*camk2*). Neurexins (*nsf*) forms a family of proteins that acts as neuronal cell-surface receptor. The key synaptic role of neurexins, is the binding of neuroligins (*nlgn*) by its β-sheet. Together both proteins form a critical synaptic cell-adhesion complex, where they activate the presynaptic (neuroligins) and postsynaptic (neurexin) differentiation, and bind together both terminals during a synaptic transmission (Südhof, 2008; Craig and Kang, 2007; Dean et al., 2003; Scheiffele et al., 2000). Moreover, whilst these gene families have been extensively linked to cognitive functions, have also been involved in neurodevelopmental diseases, such as schizophrenia and autism spectrum disorders (ASD) (Reichelt et al., 2012; Südhof, 2008). However, given the advantages that the developing zebrafish provides for the detection of gene expression at level of singular neurons, the characterization of neurexins and neuroligins in *D. rerio* is of particular interest. The expression of such proteins have been detected during the firsts stages of the embryonic development, suggesting new roles in neural specification and migration. Likewise, a potential delivery of paternal RNA to the embryo has been suggested, in which a neurexin isoform expressed in the adult testis and in the earliest stages of development has been identified (Wright and Washbourne, 2011; Davey et al., 2010; Rissone et al., 2006).

3 neurexin and 4 neuroligin genes have been determined in mammals, except for hu-

mans and higher primates, which expresses 5 of the latter. In the zebrafish 6 neurexin mammalian homologs and 7 neuroligin genes are known, 3 of which constitute duplicates of mammalian genes. A higher gene conservation between human and zebrafish neuroligin 4 has been observed in comparison with mouse (82% protein identity zebrafish-human and 60% for zebrafish-mouse) (Wright and Washbourne, 2011; Davey et al., 2010; Rissone et al., 2006). It is intriguing to further analyse the similarities among zebrafish and human neuroligin 4, and whether these genes possess equivalent synaptogenic function.

In this study, 2 neurexin isoforms were found, all of which were classified as many:1 (zebrafish:mouse), along with 7 neuroligin isoforms; one classified as 1:1, one as unique to zebrafish, and the rest many:1 (See Appendix C.2 for a list of all transcripts corresponding to key synaptic genes). Neuroligin 3 showed higher expression in the 3 tissues in comparison with neuroligin1,2 and 4, however, the latter was the lowest expressed. The optic lobe displayed a slightly higher expression for both neurexin and neuroligin genes than the hindbrain and olfactory lobe. Analysis of expression patterns in the embryonic zebrafish have recognized a higher detectable signal of neuroligin in the *tectum optic* than other brain regions (Rissone et al., 2006).

Most highly expressed key synaptic genes enriched in the optic lobe, belong to SNARE protein complex, namely; *vamp, snap-25* and syntaxin (*stx*). These proteins make up the core machinery that fuses the membrane of neurotransmitter-containing vesicles for their release to the synaptic cleft in a calcium-dependent manner (Byrne et al., 2014; Chen and Scheller, 2001). In this study, it was identified a high conservation of the SNARE complex between the *de novo* zebrafish assembly and mouse.

However, despite the fact that SNARE proteins showed an enrichment in the optic lobe and hindbrain (~1.5x fold change), the olfactory lobe-enriched set displayed a GO analysis enrichment associated with "SNARE binding" (Figure 4.19). Genes related

to this ontology were members of the complexin subfamily of proteins (*cplx2, cplx3b, cplx3a*), synaptotagmin (*syt5b, syt5a*). Members of the SNARE family have comparable expression patterns over cortex, cerebellum and hippocampus (Prescott and Chamberlain, 2011). Yet, phosphorylated syntaxin by *camk2* is enriched in the presynaptic membrane of retinal ribbon synapses, which regulates the docking of SNARE complexes (Liu et al., 2014).

Camk2 one of the most abundant protein kinases, comprised one of the highest key synaptic genes expressed in this study. Camk2 (or CamkII) is a Ca^{2+} activated enzyme abundant in the brain, making 1-2% of the total protein. This kinase is fundamental in various neuronal functions, namely; neurotransmitter synthesis and release, modulation of ion channel activity, long term potentiation, synaptic plasticity, as well as learning and memory (Jahn and Fasshauer, 2012; Yamauchi, 2005; Strack et al., 1997). Gene conservation around 92-95% of identity has been observed among *D. rerio* and human. This *de novo* assembly showed all 7 camk2 expressed in the zebrafish. It was identified one isoform unique to the zebrafish (*camk2d2*), two 1:1 (zebrafish-mouse) isoforms, and the rest many:1, with no particular enrichment within a specific tissue. Studies of expression patterns in this gene family during the zebrafish early development, have demonstrated a complex pattern of gene expression consistent with pleiotropic functions during development (Hsu and Tseng, 2010; Rothschild et al., 2007).

4.7 Summary and Comments

Using high throughput Illumina sequenced reads obtained from 3 main regions of a zebrafish brain, together with 4 whole brain replicates, is presented an effective pipeline for *de novo* transcriptome assembly and annotation. The quality of the pipeline was assessed comparing the *de novo* assembly to the *D. rerio* Ensembl genome (version GRCz10). It was therefore demonstrated that the pipeline outputs robust and good

quality transcripts.

As part of the recent characterization of the zebrafish synapses by Bayés et al. (2017), the PSD between mouse and zebrafish was compared. In addition, it was attempted to examined variations in gene expression associated with different regions of the zebrafish brain and their similarity with the mammalian PSD. The data showed a high proportion of orthology (86%) between *M. musculus* and the *de novo* assembly. A large fraction of orthologs (33%) were subject to gene duplication as a result of the TSGD, and therefore supporting Bayés et al. (2017) findings. Nonetheless, only a small portion of the resulted ortholog genes (25.5%) corresponded to mouse synaptic genes, suggesting that this proportion is likely conserved among all vertebrates. Moreover, 153 mouse-specific PSD genes were identified.

Inasmuch as tissue-enriched and specific expression, it was observed a higher similarity between the hindbrain and optic lobe, compared to the olfactory lobe. Yet, the optic lobe displayed the highest proportion of mouse PSD ortholog genes, together with a higher number of 1:1 orthology types. Hence, it is proposed that the zebrafish's optic lobe is the brain region with the highest similarity to the mammalian brain, which may be a useful for prospects comparative studies. It was also identified a higher teleostean conservation within the hindbrain genes.

Furthermore, when considering SYN and PSD genes with highly variable expression levels, it was observed a greater number of many:1 (zebrafish:mouse) orthologs, suggesting that genome duplication is a factor that boost expression variability, or that expression variability is a cause of gene retention.

Chapter 5

De novo Assembly of the Bat brain transcriptome

“Success is a science; if you have the conditions, you get the result."

- Oscar Wilde -

Bats evolved around 50 million years ago (Wilson and Reeder, 2005), and are the only mammal with the ability to fly, which has permitted their broad distribution over all continents (except Antarctica). In fact, these animals encompass ~20% of all known mammalian species (Lee et al., 2015; Shaw et al., 2012). This wide diversification has resulted in inconsistent phylogenetic conclusions. To date, molecular data have established that the bat order Chiroptera is comprised by Yangoquiroptera (most microbat families) and Yinpterochiroptera (megabats and a few microbats), and is placed in the Laurasiatheria mammalian clade (e.g., horses, carnivores, shrews and whales) (Teeling et al., 2016; Madsen et al., 2001; Murphy et al., 2001).

The availability of multiple bat genomes has generated valuable material that facilitates future scientific research on the molecular biology of bat's remarkable capabilities (Fang et al., 2015). For example, evidence of positive selection in mitochondrial and nuclear *OXPHOS* genes involved in energy metabolism, is proposed as an adaptation of bats to satisfy the large energy consuming flight activity (Shen et al., 2010). Moreover, metabolic adaptations, particularly in the growth hormone (*GHR*) and insulin-like growth factor 1 (*IGF1*) receptor, have been attributed to the notably long lifespan in bats, compared to their small body size (Seim et al., 2013).

Toothed whales (e.g. dolphin and killer whale), microbats and a few megabats species are capable of echolocation using ultrahigh frequency sounds (Shen et al., 2012; Jones and Teeling, 2006). Evidence of convergent evolution has been reported between these unrelated species by means of natural selection in the "hearing-gene" *Prestin*, which is involved in the cochlear amplification of the mammalian ear (Li et al., 2010b). Echolocation functions via adaptive sensorimotor systems, which together permits location and tracing of sonar objects while flying. For this reason, efforts to identify genes acting on echolocation have concentrated mostly on the auditory and vocalization systems (Teeling, 2009).

Echolocation is a complex phenotypic trait and the functional genomics relating to this process remains unresolved. Motivated by this, here it is employed NGS technologies to generate a comprehensive analysis of gene expression from three brain regions (cortex, brainstem and cerebellum) of the microbat, common pipistrelle (Pipistrellus pipistrellus). The synaptic complexity involved in learning and memory is also investigated for the first time in this species.

5.1 De novo transcriptome assembly

A *de novo* assembly approach was exercised to produce a transcritptome without depending on a reference genome (See chapter 3). In the previous chapter, the assembly pipeline was assessed against a reference genome, wherein its ouput was considered robust, and therefore suitable to be exploited for the common pipistrelle assembly, which lacks a reference genome. Tissues included in this study were; cortex, brainstem and cerebellum (Figure 5.1) that yielded 100,537,958 paired-end reads. Reads were quality processed, decreasing their number to 97,817,253 (14,880.31 Mpb) (Table 5.1).

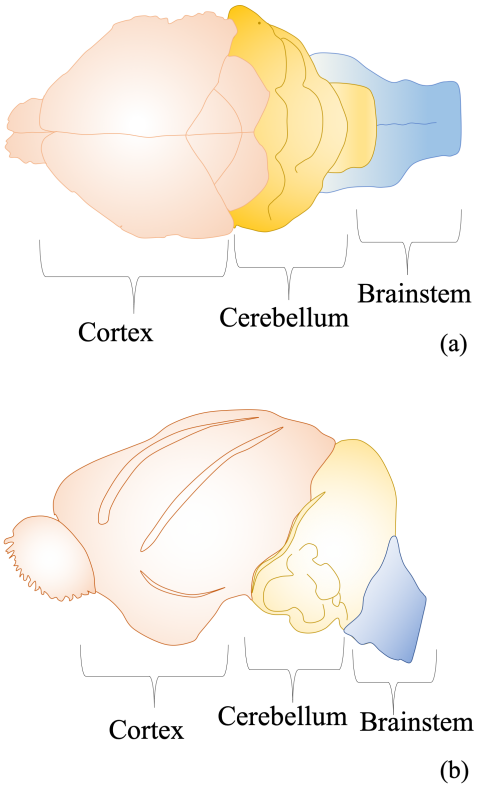


Figure 5.1: **Dorsal (a) and lateral (b) representation of the adult common pipistrelle brain** depicting tissues used for the present study; cortex (left and right cerebrum hemispheres), brainstem and cerebellum.

Table 5.1: **cDNA library summary of the RNA sequencing yield**

Tissue	No. of raw reads	No. of trimmed reads	Base pairs (Mbp) of trimmed read
Cortex	37,631,874	36,720,241	5,588.52
Brainstem	31,196,640	30,429,698	4,629.45
Cerebellum	31,709,444	30,667,314	4,662.34

High quality reads were assembled using Trinity. This process generated 215,827 contigs together in all three assemblies (cortex, brainstem and cerebellum). The assembly obtained a mean contig length from 940.52 to 1,044.23, with the cerebellum assembly

showing the lowest number of contigs (see Table 5.2).

Table 5.2: **Trinity *de novo* assembly and statistics summary.**

Tissue	No. contigs	Base pairs (Mbp)	Mean contig length	Median contig length	SD contig	%GC
Cortex	73,521	76.8	1,044.23	525	1,176.38	52.96
Brainstem	74,868	79.8	1,065.24	540	1,198.48	53.04
Cerebellum	67,438	63.4	940.51	487	1,023.6	52.92

Quality control

As before, quality assessment was achieved using Transrate by mapping RNA-seq reads to Trinity contigs, and evaluating its alignments. Adequately optimized scores for downstream analyses were computed (see Table 5.3). Poorly supported transcripts, together with those that had expression levels lower than 0.5 TPM were removed. Resulting in a reduction of approximately 36.8% of the original Trinity assembled contigs.

Table 5.3: **Transrate quality control assembly summary**

Tissue	Transrate optimized score	No. transrate contigs	No. contigs >0.5 TPM	% remained contigs	Base pairs (Mbp)	Mean contig length
Cortex	0.37	45,308	44,865	61	56.1	1,251.53
Brainstem	0.38	46,350	45,989	61.4	58.0	1,260.91
Cerebellum	0.35	45,946	45,542	67.6	48.2	1,058.31

Full-length transcript analysis

As a measure to evaluate the quality of the Transrate assessed assemblies, the number of full-length transcripts was determined. Commonly, in a reference-guided transcriptome, contigs are aligned against its reference. Although that the present transcriptome was assembled without any reference genome, the Swiss Prot database and a related bat species *Myotis lucifugus*-Myoluc2.0 peptide sequence available from Ensembl, were used as a proxy for a reference genome. Length coverage was examined by the amount of unique top matching proteins which aligned over more than 80% of a known protein length using BLASTX with and E-value of e-20 (Figure 5.2). 5,998 ~ 7,032 and 5440 ~ 6324 proteins were encoded by near or full-length transcripts, comprising more than the 80% of alignment coverage for the SwissProt database and Myoluc2.0, respectively. A total of 26,562 ~ 28,002 and 25,647 ~ 27,088 transcripts were found to had unique hits to SwissProt and Myoluc2.0, respectively.

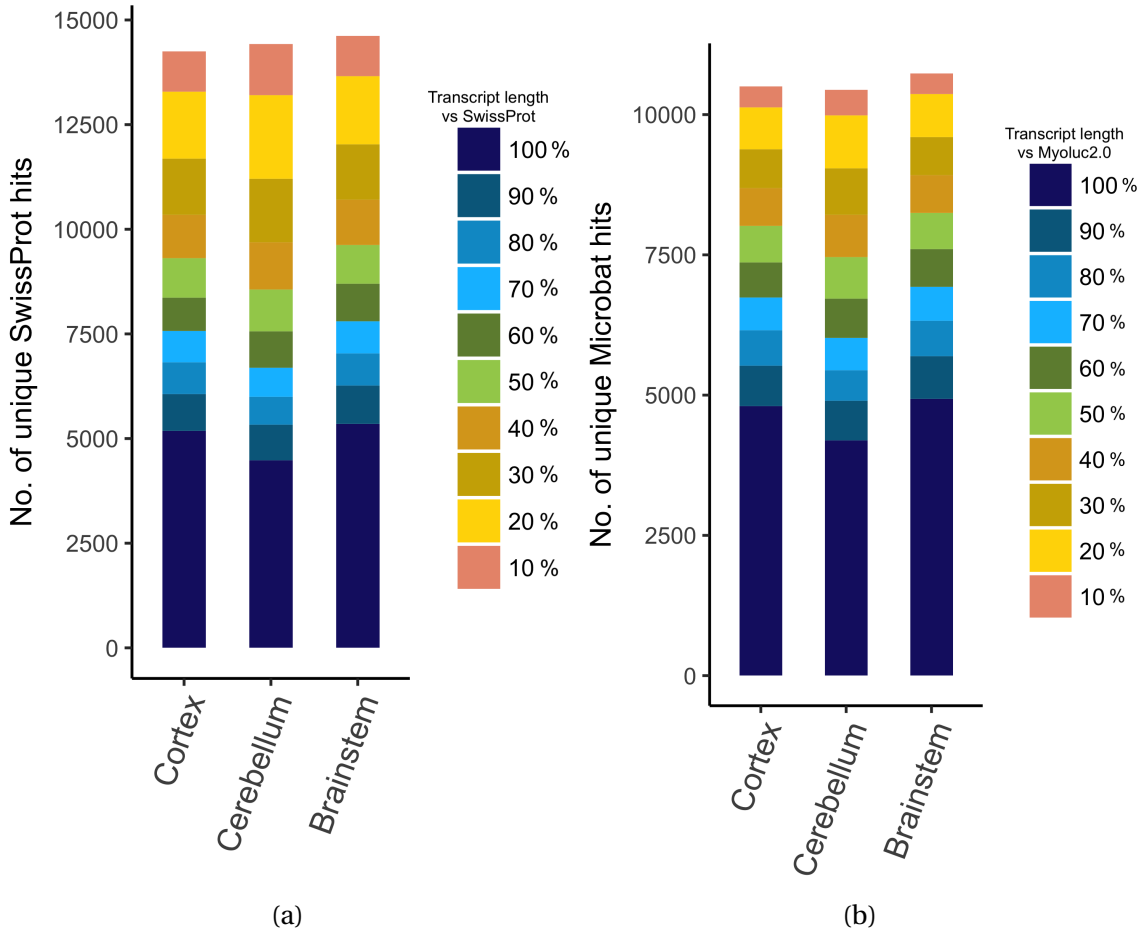


Figure 5.2: **Percentage of the transcript length that align to a known protein.** A metric for quality assessment of the assembly is to explore the portion of the transcripts that align a known protein. If ≥90% of the total length of a transcript aligns to a known protein, this transcript is considered "full-length" . BLASTX (e-value of e-20) was carried out to align the optimized Trinity-generated assembly to (a) the SwissProt database; and (b) *Myotis lucifugus*-Myoluc2.0 reference transcriptome. This used to assess the optimized Trinity-generated transcriptome, with no filtering of the data.

Completeness transcript analysis

As a further assessment of the Transrate assemblies, their completeness was evaluated by BUSCO. Although BUSCO is intended for complete genome assemblies, it provides an estimated relative assessment guide for these brain transcriptomes. Single-copy orthologs from vertebrata and metazoan BUSCOs gene datasets were used. Results were consistent across the assemblies, and suggested a completeness proportion of 78.7 to

82.1% and 51 to 60.2% for the metazoan and vertebrata datasets, respectively (Figure 5.3).

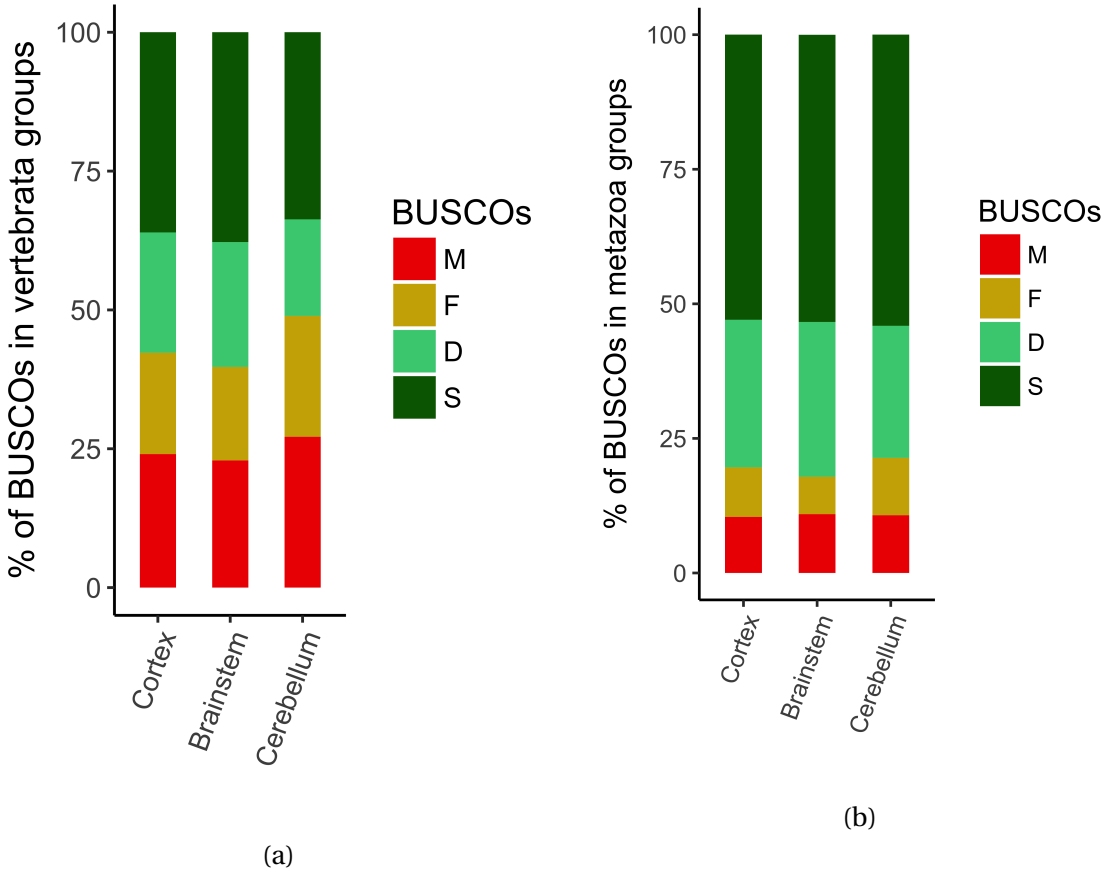


Figure 5.3: **Distribution of BUSCOs categories in the optimized *P. pipistrellus de novo* assemblies**. BUSCO completeness is categorized as; (S) Complete and single-copy (recovered transcripts are within 95% expectation of the BUSCO group mean length); (D) Complete and duplicated (transcripts with more than one single-copy); (F) Fragmented (incomplete recovered transcripts), and; (M) Missing (not recovered transcripts). (a) Assessment using vertebrata BUSCO set, representing 3,023 genes, (b) Assessment using metazoa BUSCO set, representing 843 genes. Such metric was only used for assessment of the transcriptome but not filtering procedure was carried out.

Generation of a non-redundant assembly

The number of contigs varied across the three assemblies owing to their dissimilar expression levels and sequencing depth. In order to perform downstream comparative analyses, contigs from the three tissues were combined together into a single assem-

bly (136,396 contigs), and the redundancy content was evaluated by CD-HIT-EST. In this way, a non-redundant unified assembly was computed, consisting of 26,463 transcripts, of which 2,277 represented spliced forms. This assembly served as reference for mapping back preprocessed reads using Hisat2. Overall alignments rates among the assemblies displayed satisfactory results; 84.14%, 83.71% and 83.49%, for cortex, brainstem and cerebellum, respectively (Figure 5.4). Expression levels in TPM for the three assemblies were next obtained by Kallisto.

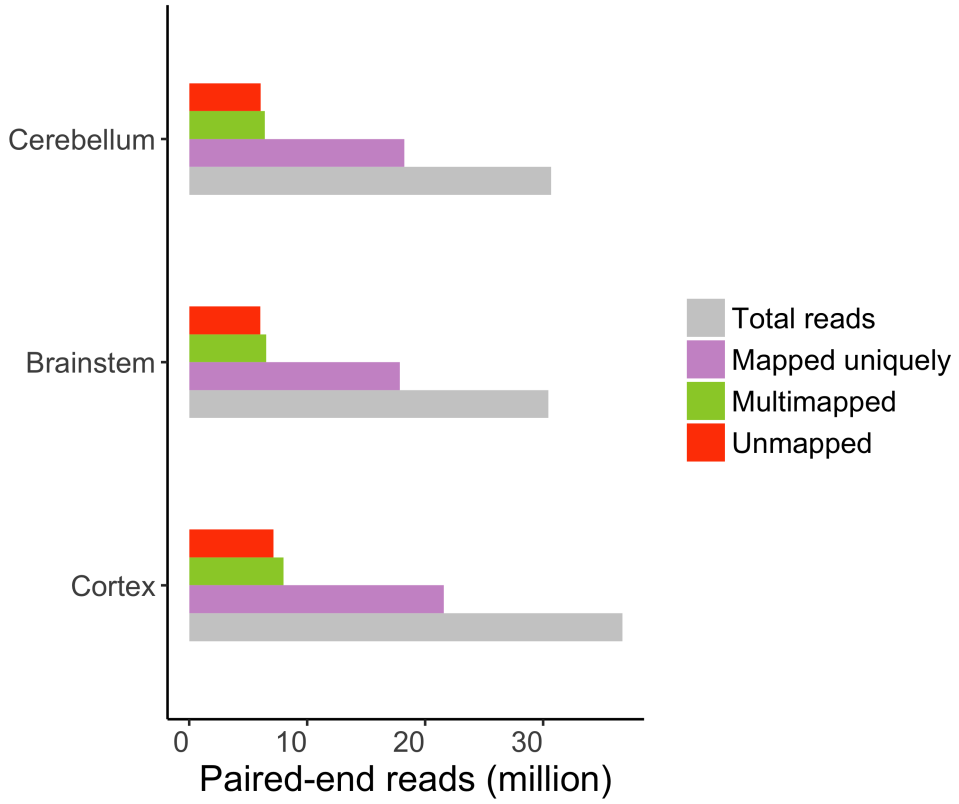


Figure 5.4: **Number of reads mapped back to transcriptome**. Preprocessed reads from the three *P. pipistrellus* brain tissues (cortex, brainstem and cerebellum) were mapped back to a unified non-redundant assembly (136,396 transcripts). Different colours indicate the total amount of preprocessed paired-end reads (grey); reads that aligned concordantly 1 time (uniquely) (purple); reads that aligned more than 1 time (multimapped) (green); reads that did not aligned at all (unmapped) (red).

To investigate whether there was any loss of valid transcripts in the obtained unified non-redundant assembly, the assembly was re-evaluated by BUSCO (Figure 5.5). The

proportion of missing BUSCO's decreased slightly in comparison with Transrate assemblies for both BUSCO's metazoa and vertebrata ortholog datasets. Additionally, the proportion of BUSCO's complete single-copy increased ~15%, whereas the complete-duplicated category, showed a reduction. Hence, it was validated an improvement in the quality of the assembly.

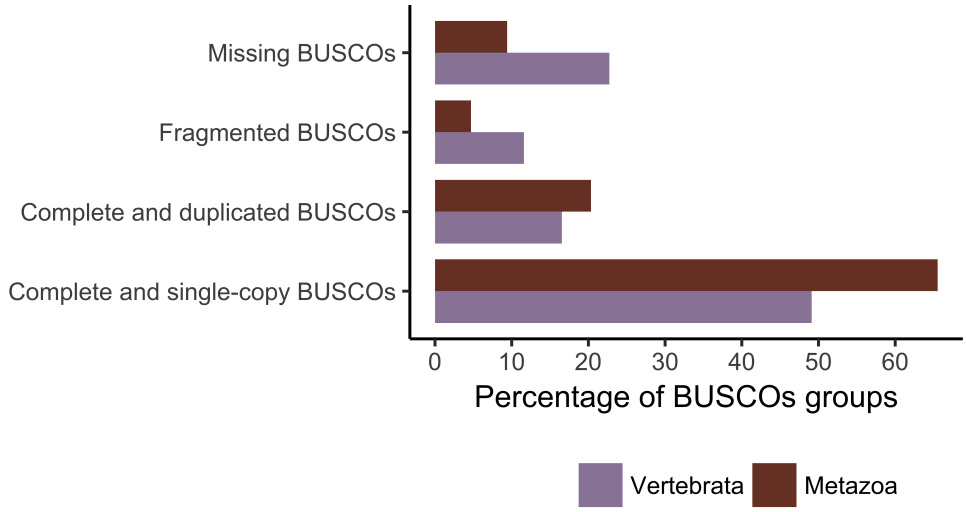


Figure 5.5: **Distribution of BUSCO assessment in the non-redundant *P. pipistrellus* assembly**. The unified non-redundant assembly (26,463 transcripts) was evaluated for loss of valid transcripts by BUSCO vertebrata and metazoa orthologs shown in different colours. Such metric was only used for assessment of the transcriptome but not filtering procedure was carried out.

5.2 Transcriptome Annotation

Dammit! (Scott, 2016) was employed to annotate the non-redundant *de novo* assembly (26,463 contigs). Transdecoder estimated that 18,747 of the contigs contained an ORF, and hence likely corresponded to protein-coding transcripts. Among these, 8,990 contained a complete ORF (with a start and stop codon). Whereas 6,042, 1,338 and 2,377 of the contigs included a partial 5', partial 3' and an internal ORF, respectively. Summary of the *de novo* pipeline assembly and annotation results are shown in Figure 5.6.

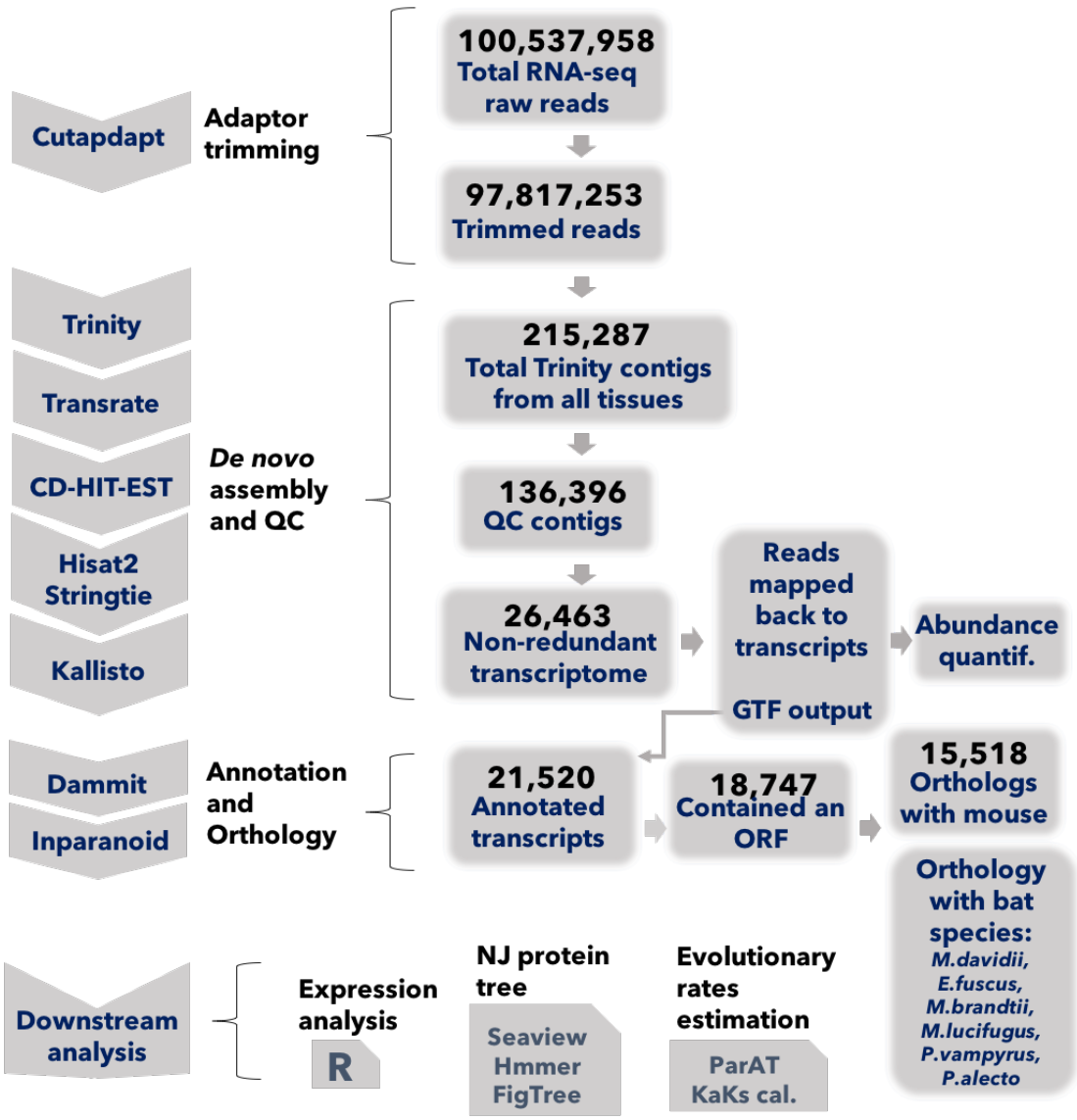


Figure 5.6: **Summary of *de novo* transcriptome assembly and annotation pipeline.** RNA-seq reads were trimmed (Cutadapt) and assembled (Trinity). Low-quality contigs were removed (Transrate) and a non-redundat transcriptome was obtained (CD-HIT), which served as reference to map back the RNA-seq reads from all tissues (Hisat and Stringtie) and its expression abundance was quantified in TPM (Kallisto). The transcriptome was annotated (Dammit) and ORF estimated (Transdecoder). Orthology was inferred (Inparanoid) for further downstream analyses.

Protein domain analysis

Analysis of protein domain annotation was used to improve the characterisation of the transcriptome. With the use of HMMER, a total of 9,889 unique Pfam domains were

identified, 7,299 of these had an E-value below the threshold of <0.05. Results of top enriched domains in the common pipistrelle brain (see Figure 5.7) agree with the assumption that synapses first evolved in simple eukaryotes with subsequent stepwise development of more sophisticated synaptic molecular complexes, such as the vertebrate PSD (vPSD).

Emes et al. (2008) and Bayés et al. (2017) identified a conserved group of synaptic proteins (1,101) common in all vertebrates. The most conserved core elements of this vPSD set were kinases and phosphatases, which were also found as the most enriched domains in the bat *de novo* brain assembly. Certainly, with hundreds of kinases encoded within the mammalian genome, practically every signal transduction process in the nervous system is induced by interlinked phosphorylation events (Chico et al., 2009; Yamauchi, 2005).

Among the most enriched protein domains in the assembly and also components of the vPSD were key elements in synaptic transmission and plasticity, namely, ribosomal proteins, for example RRM (Ribonucleoside-diphosphate reductase), GTPases and core domains of the MAGUK protein family, such as SH3, PDZ and Ank domains. Other critical elements for the proper function and maintenance of neural circuits, such as calcium-binding EF-hand domain, cell-adhesion Leucine-rich repeat (LRR) and the Immunoglobulin domain superfamily (Ig) were widely enriched in the bat brain.

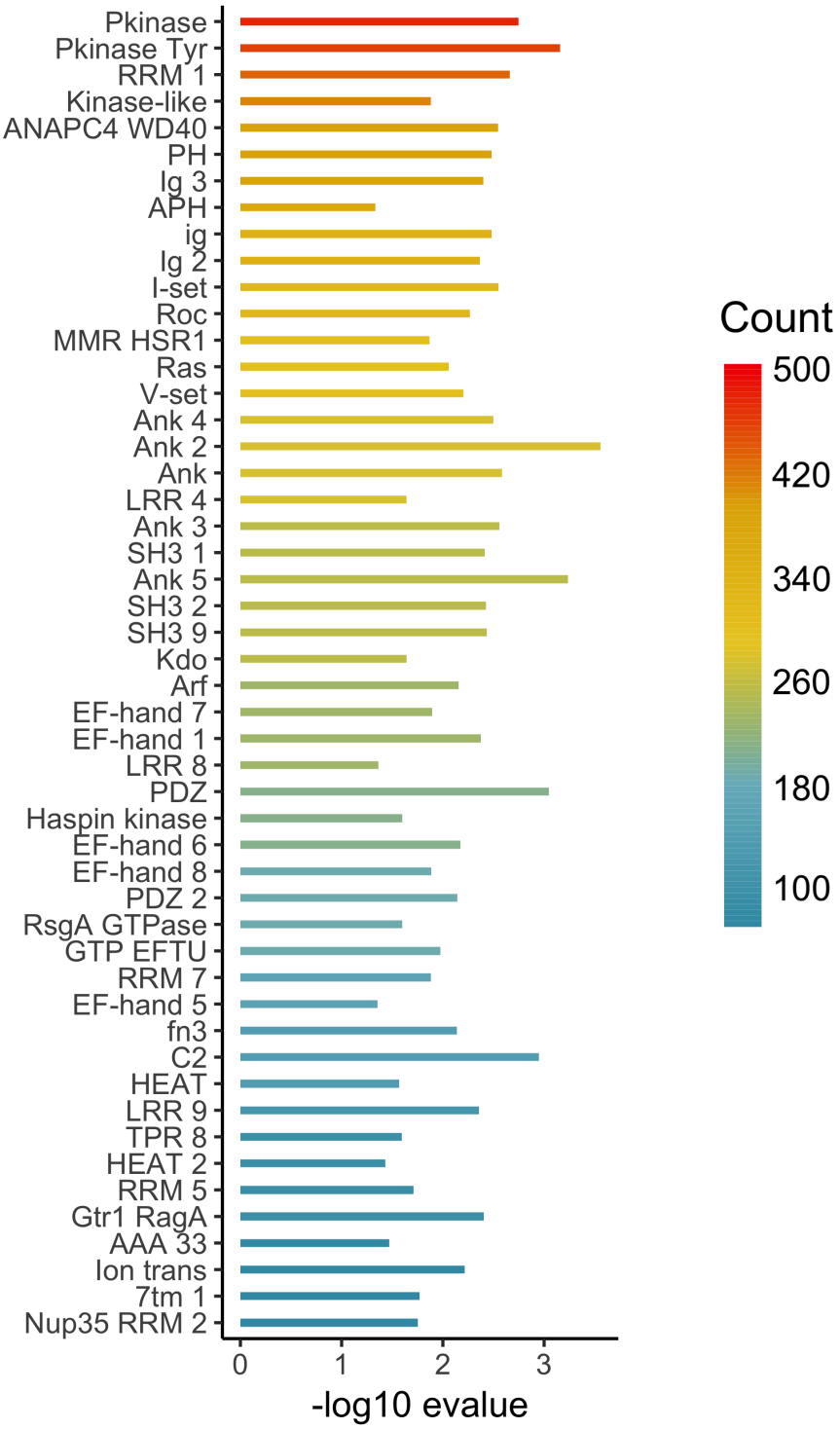


Figure 5.7: **Enriched Pfam domains.** Annotated transcripts were investigated for protein domains using HMMER hmmscan that searched the *de novo* transcripts against the pfam (Finn et al., 2016) database using probabilistic models called profile hidden Markov models. The y-axis shows the top enriched domains found with an e-value cut-off of 0.05. Colours indicate the number of transcripts found in each domain; red comprises the highest, while blue the lowest.

The assembly captures most of bat trancriptomes

The annotated assembly, was compared to the transcriptome of other bat species from the NCBI Eukaryotic genomes database, namely; *Myotis davidii* (David's myotis), *Eptesicus fuscus* (big brown bat), *Myotis brandtii* (Brandt's bat) and *Pteropus alecto* (black flying fox), along with 2 bat species available from Ensembl; *Myotis lucifugus* (little brown bat) and *Pteropus vampyrus* (large flying fox). The mouse transcriptome was used as an out-group.

Using Blast (with an alignment cut-off of 70% identity and an E-value of 1e-4) results were consistent with the published evolutionary history of bat species (see Figure 5.8). Considering that the most notable phylogenetic arrangement in bats is the division into two subordinal taxa, such as; Yinpterochiroptera (most megabats) and Yangochiroptera (most microbats). Megabats (*P. alecto* and *P. vampyrus*) are herbivores and have developed an acute sense of sight and smell, whereas microbats (*M. lucifugus*, *M. davidii*, *E. fuscus*and *M. brandtii*, including the common pipistrelle) have evolved the ability of laryngeal echolocation to orient in complete darkness. As expected, the common pipistrelle assembly demonstrates higher similarity with echolocating bats than non-echolocating.

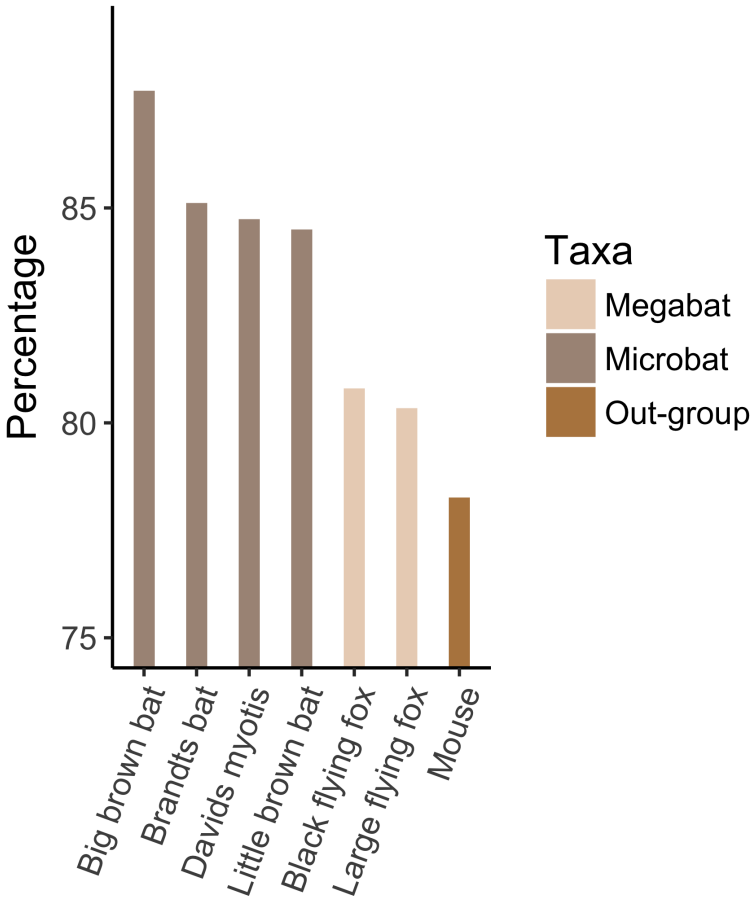


Figure 5.8: **Assembly bat coverage**. The annotated assembly was compared with available bat genomes (Blast; ≥70% identity; E-value 1e-4). Unique transcriptome hits are shown as percentage. Colours depicts taxa of bat species; microbat (echolocating bats; including the common pipistrelle) and megabat (non-echolocating). The mouse is represented as an outgroup.

5.3 Transcriptome Orthology

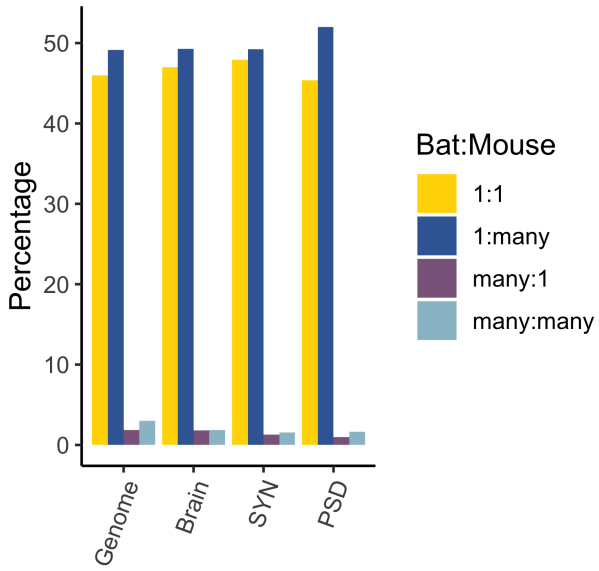
For the purpose of exploring synaptic genes expressed in the common pipistrelle *de novo* assembly, orthologs in the mouse were determined. From the supplementary data of (Sharma et al., 2015) and G2C database (Croning et al., 2008), genes expressed in the mouse brain, SYN and PSD were outlined and their equivalent transcript in the bat assembly was identified. Inparanoid estimated that from 18,747 protein-coding tran-

scripts, 83% had an ortholog pair with 9,474 mouse genes. Additionally, 70.2% had an ortholog pair with mouse genes encoded in the brain; 18% in the SYN and 12.18% in the PSD. With reference of these pair of orthologs, its corresponding orthology type was computed i.e., bat:mouse; 1:1, 1:many, many:1, many:many and unique to bat. Very few many:1 homologs were detected (301 transcripts, 1.93%). This was mirrored with a large proportion of 1:many orthologs, which in part may reflect the incompleteness of the bat transcriptome but also suggests that lineage specific duplications dominate in the mouse geneome (Figure 5.9a and 5.9b).

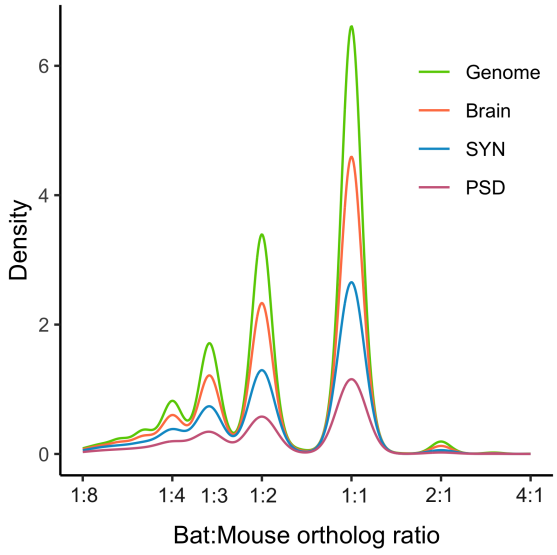
Certainly, researchers have recognized the smaller size of bat and bird genomes, compared to the size of other vertebrates (Gregory, 2002). For example, the latest assembly of the microbat *Myotis lucifugus* (Myoluc2.0) contains 19,728 coding genes, whereas the latest assembly of the house mouse (GRCm38.p6) encompass 22,604 coding genes. It has been associated high metabolic demands to a smaller cell size and hence, a constrained genome in vertebrates that evolved powered flight (Hughes and Hughes, 1995). A correlation between wing loading (measurement of the total mass over the wing area) has also been seen. For example, Smith and Gregory (2009) compared the genome size between megabats and microbats, noticing even stronger levels of genome constraint in megabats, which in turn have higher wing loading than microbats (Norberg and Rayner, 1987). Similarly, Andrews et al. (2009) conducted a comparison of cell and genome sizes across 74 bird species (*Passeriformes*). The same positive relation among genome size and wing loading was found, therefore suggesting gene constraint as an adaptation for a more efficient flight.

Moreover, the 1:many orthology type was generally higher in the PSD. It is reasonable to consider gene constraint in the bat as a potential factor in this matter. As it has been previously reported, the PSD consists of hundreds of interacting proteins (Emes and Grant, 2012). The extent of evolutionary rate for an individual protein is negatively correlated by the number of its interactions, which might be caused by biological con-

straints that are essential to sustain its multiple interactions (Fraser et al., 2003).



(a)



(b)

Figure 5.9: **Orthology distribution among bat transcripts and mouse genes.** Inparanoid estimated that 15,518 transcripts have an ortholog mouse gene. a) Bat:Mouse ratio of orthologs were represented as 1:1, many:1, many:many, and unique to bat in genes expressed in the genome, brain, SYN and PSD. b) For each pair of ortholog it was estimated the density of its distribution.

Considering that the most common cause of new functional genetic material is gene

duplication (Proulx, 2011), the many:1 *de novo* transcripts may represent unique bat specializations. From this group of transcripts (n=301), homologous genes from *M. lucifugus* (little brown bat) (n=216) were determined and gene ontology was conducted using Biomart (Figure 5.10). Bats are well-known to host a wide range of zoonotic viruses (Calisher et al., 2006). Noteworthy, among the most enriched GO terms, were those associated with virus susceptibility. For example, the major histocompatibility complex (MHC) class II, which plays a significant role on the activation of the adaptive immune response, has been characterised in bats (Ng et al., 2017, 2016). In such studies, unique bat features were found, including; the existence of a class II locus away from the MHC-II region, which confirms an ancient MHC-II duplication block; additionally, unique insertions within the antigenic-peptide binding groove were detected. Insertions are fundamental mechanisms that confer phenotypic differences between species (Volfovsky et al., 2009). It is likely that these unique bat features have play a significant role in their evolutionary history.

Regulation of the apoptotic process which may also be related to intracellular infection, was the most enriched GO term identified, moreover, it has been long acknowledged that the mitochondria has a role in apoptosis. Notably, proteins bound to the mitochondrial intermembrane space (found as enriched GO term) are required for viral immunity (Brook and Dobson, 2015). These unique bat characteristics, are likely to have shaped the ability of bats to manage infection without manifesting disease.

Atg/ULK1 kinase complex was among the most statistically significant observed GO terms. ULK1 (unc-51 like autophagy activating kinase 1; Atg1 complex in yeast) is a kinase involved in the selective initiation of autophagy, either dependently or independently of a nutrient and energy status (Lin and Hurley, 2016). Autophagy is an intracellular mechanism that degrades detrimental lysosomal contents (Mizushima, 2010). Dysregulation of autophagy has been implicated in major neurodegenerative diseases, such as AD (Nixon, 2013; Barnett and Brewer, 2011; Lipinski et al., 2010). Protein lipoy-

lation, also a statistically significant term, is an uncommon and highly conserved lysine post-translational modification in mammals. Dysregulation of this process has been linked to metabolic alterations (Rowland et al., 2018), and significantly, to the pathological accumulation of the microtubule-associated protein tau, primary trigger of AD (Thomas and Yang, 2017).

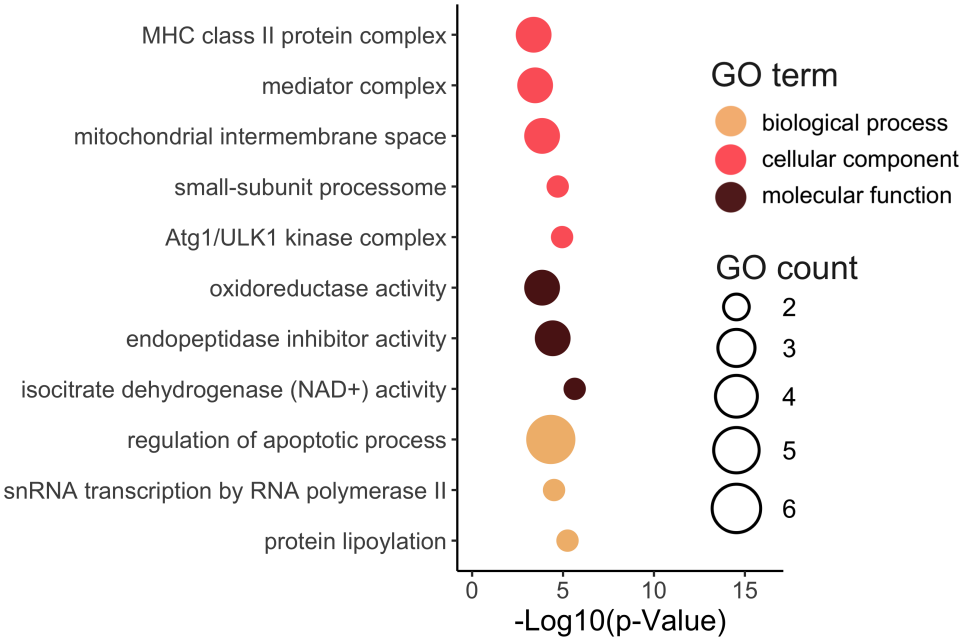


Figure 5.10: **GO analysis of bat gene duplication.** The many:1 *de novo* transcripts expressed in the brain were annotated for GO terms using R analysis tools. P-values were estimated based on hypergeometric distribution and adjusted FDR, applying a cut-off of 0.05. Highly significant enriched terms were selected and plotted against their negative log10 P-value. Different size in circles indicates the enrichment.

Inparanoid estimated that 604 of the protein-coding transcripts did not have an identifiable ortholog in the mouse gene set. To validate this finding, the transcripts were compared to the mouse proteins using BLAST. This reduced the number of bat specific transcripts to 89. Of these, 72 had homologs in *M. lucifugus*. GO analysis of these 72 proteins was performed, in which 4 basic enriched terms were identified, including intracellular, nucleic acid binding, endonuclease activity and regulation of transcription (Figure 5.11). Most of the bat-unique genes are novel or not yet characterised, however, several zinc finger proteins.

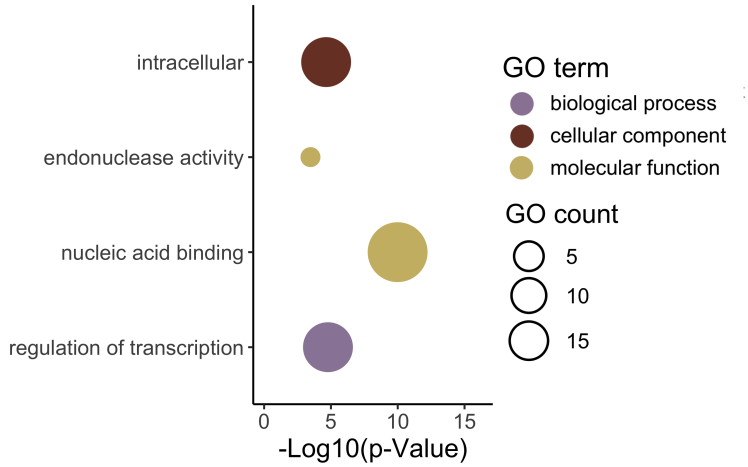


Figure 5.11: **GO analysis of bat-specific genes.** Bat transcripts that did not have an identifiable mouse homolog were annotated for GO terms using R analysis tools. P-values were estimated based on hypergeometric distribution and adjusted FDR, applying a cut-off of 0.05. Highly significant enriched terms were selected and plotted against their negative log10 P-value.

5.4 Transcript expression in bat brain tissues

Transcript abundance was quantified for the 18,747 protein-coding transcripts and the expression in the different tissues was compared. Summaries of transcript expression levels at multiple thresholds are exhibited in Table 5.4. All three tissues displayed comparable expression levels. Nonetheless, the brainstem showed a higher number of expressed transcripts at all thresholds, followed by the cortex.

Table 5.4: **Number of annotated transcripts at different TPM threshold.**

Tissue	>0.5 TPM	>1.0 TPM	>5.0 TPM	>10 TPM	>25 TPM
Cortex	18,157	18,075	16,596	12,403	5,984
Brainstem	18,190	18,116	16,893	12,835	6.053
Cerebellum	18,033	17,918	15,460	11,420	5,654

For downstream analyses, transcripts with low expression (TPM ≤0.5) were removed. Mean expression for the cortex, brainstem and cerebellum were as following; 42.05 (SD=121.92; max=5,801.72), 41.88 (SD=115.97; max=6,086.15) and 41.75 (SD=132.95;

max=7,588.67). While the three tissues showed uniform expression levels (p-value = 0.25 multiple pairwise Kruskal-Wallis test), the cerebellum was the most distinct (Figure 5.12). Additionally, based on gene expression data, evolutionary relationships were reconstructed. Again the cerebellum depicted lower correlation with the cortex and brainstem. This in turn might underlie different mechanisms regulating gene expression and hence function in the bat.

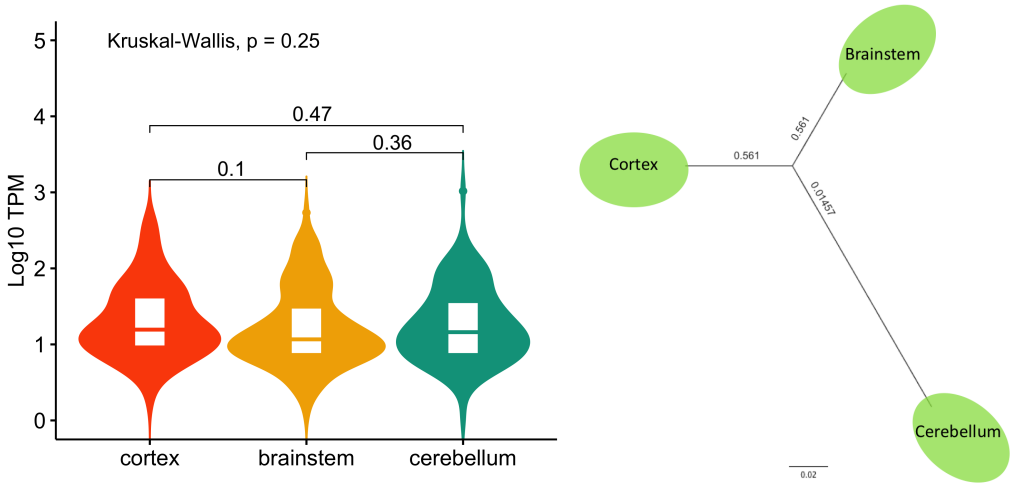


Figure 5.12: **Distribution of transcript expression levels,** estimated as TPM (transcriptome per million) in brain tissues (cortex, brainstem and cerebellum). Lowly expressed transcripts (<0.5 TPM) were filtered out. a) Combined violin and boxplot showing the distribution of log10 transformed TPMs and its density. P-value=0.25 (multiple pairwise Kruskal-Wallis test) c) Tree based on expression distance matrices between the tissues.

5.4.1 Ubiquitously highly expressed genes in the bat brain

The mean TPM expression was determined in the 3 brain tissues (cortex, cerebellum and brainstem), and the 21 highest expressed transcripts were defined as the most ubiquitously high expressed (mean TPM; 2,584). These top highly expressed transcripts shared a mouse ortholog (Table 5.5) with various genes expressed in the PSD.

The highest ubiquitously expressed transcript in the common pipistrelle brain was transthyretin (TTR) (expression in TPM; cortex 1,409; brainstem 1,374; cerebellum 1,215). In the choroid plexus of mammals, reptiles and birds, TTR is the primary synthesized protein, where it forms around 20% of the total protein (Sousa et al., 2007). TTR is distributed through the cerebrospinal fluid (CSF) and serum carrying thyroid hormones (THs) (Blay et al., 1993). mRNA expression profiling has shown the presence of TTR in the hipocampus, cortex and cerebellum (Sousa et al., 2007).

Studies based on TTR have suggested paraphyly in microbat species (Khwanmunee et al., 2016). More effective transport of THs hormones has been observed in long-lived mammals, compared to short-lived, thereby linking TTR to bats longevity Buffenstein and Pinto (2009). In this matter, reduced levels of TTR in the CSF lead to the accumulation of amyloid plaques and the onset of Alzheimer's disease (AD) (Merched et al., 1998; Serot et al., 1997). TTR has been shown to bind amyloid-beta peptide *in vitro* and inhibit the formation of amyloid fibers (Sousa et al., 2007). Certainly, many of the most expressed genes in the three brain tissues have been linked to AD. For example, apolipoprotein E (APOE) is the greatest genetic risk factor for the late-onset of AD. Via immunomodulatory mechanisms, APOE activates dysfunctional microglia that disrupts the clearance machinery of the brain, hastening amyloid plaque formation (Shi and Holtzman, 2018; Corder et al., 1994).

Table 5.5: **Top 21 highly ubiquitously expressed genes in the 3 bat brain tissues.**

Gene ID	Mean TPM	Protein	Function	Mouse ortholo-gy
TTR	6616	Transthyretin	Transport protein	SYN
APOE	6492	Apolipoprotein E	Fat metabolism	Brain
PNMA1	3716	PNMA family member 1	Paraneoplastic antigen	SYN
SYT1	3563	Synaptotagmin 1	Calcium-binding synapse	PSD
OAZ1	3283	Ornithine decarboxylase antizyme 1	Metabolism	Brain
SNAP25	3230	Synaptosome associated protein 25	Intracellular membrane fusion	PSD
GNAS	2666	GNAS complex locus	Signal transduction pathways	PSD
COX6A1	2106	Cytochrome c oxidase subunit 6A1	Mitochondrial respiratory chain	PSD
CKM	2061	Creatine kinase M-type	Energy homeostasis	Brain
COX4I1	1972	Cytochrome c oxidase subunit 4 isoform 1	Mitochondrial respiratory chain	Brain
PCP4	1915	Purkinje cell protein 4	Synaptic plasticity	Brain
GAPDH	1900	Glyceraldehyde-3-phosphate dehydrogenase	Energy metabolism	PSD
PSAP	1860	Prosaposin	Myelinotrophic and neurotrophic factor	SYN
CALM1	1810	Calmodulin 1	Calcium-binding	Brain
GLUL	1794	Glutamate-ammonia ligase	Synthesis of glutamine	Brain
SLC25A3	1741	Solute carrier family 25 member 3	Transmembrane transport	SYN
HSPA8	1570	Heat shock protein family A8	Molecular chaperone	PSD
SYT1	1502	synaptotagmin 1	Calcium binding	PSD
YWHAQ	1501	Tyrosine 3-monooxygenase	Signal transduction	PSD
RPS24	1494	Ribosomal protein S24	Component of the large ribosomal subunit	SYN
SYP	1477	Synaptophysin	Major synaptic vesicle protein	SYN

Amongst other highly expressed genes in the bat brain, were genes that regulate hibernation. Torpor is an exceptional adaptation, in which the brain plays a central role to effectively decrease metabolic activity with reduction of heartbeat, respiration, fuel consumption and expensive cell processes with no evidence of brain damage (Gautier et al., 2018). Moreover, Lei et al. (2014) identified a set of up-regulated genes while studying bats in hibernation, among this gene set, *SYT1, SNAP25, CALM1, HSPA8* and *RPS24* were identified highly expressed in the *de novo* assembly.

5.4.2 Tissue-enriched gene expression

Genes with regionalized expression patterns provide insights into the delicate functional and structural organization of each brain region, hence in this study, transcripts significantly enriched per tissue were determined (cortex, brainstem and cerebellum- "tissue-enriched"). The approach used to classify the contigs was as follows; those that showed at least 1.2x fold-change (FC) greater in a specific tissue and simultaneously showed at least 1.2x FC lower in the other two tissues. Likewise, transcripts expressed in one tissue only ('tissue-specific") were looked for, however, from the 18,747 protein-coding transcripts, none were identified.

The cerebellum displays a higher number of tissue-enriched transcripts (n=213), than the cortex (n=94) and brainstem (n=89). Parallel expression patterns were also observed in the cortex and brainstem, compared to the cerebellum, reflecting a distinct function (Figure 5.13a). Moreover, to assess the degree of enrichment in each tissue, different ratios of log FC were calculated (Figure 5.13b). While the brainstem and cerebellum displayed similar number of transcripts in the different ratios, the cortex reflected higher levels of tissue-enrichment.

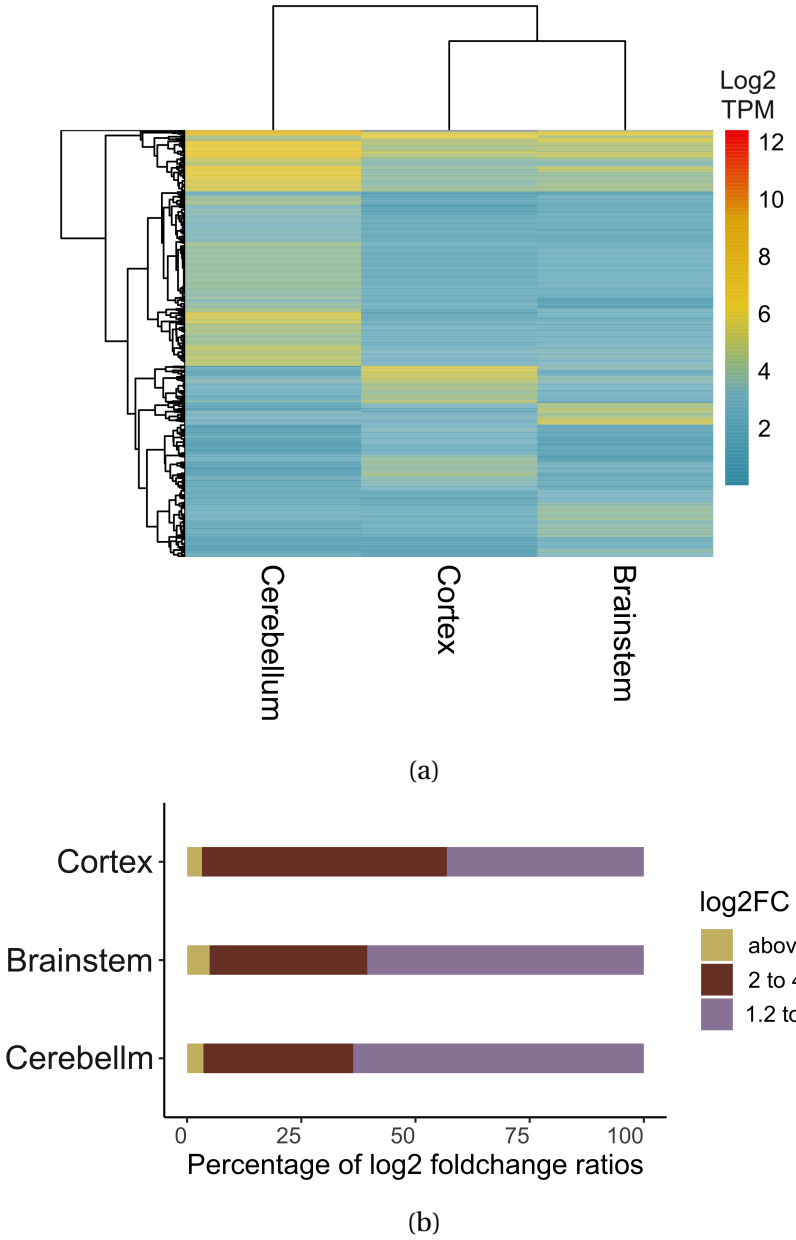


Figure 5.13: **Log2 foldchange of tissue enrichment.** a) Heatmap showing expression levels of tissue-enriched transcripts; FC > 1.2 in one tissue and FC< 1.2 for the same transcript in the other two tissues. Expression values are transformed TPM. Dendogram clustering on the X-axis indicates sample similarity, while Y-axis dendogram clustering groups transcripts with similar expression (b) Distribution of tissue-enriched transcripts showing in the Y-axis percentages of foldchange ratios. Different colours indicate relative expression levels.

A greater number of tissue-enriched transcripts was observed in the cerebellum, in conjunction with a lower proportion of mouse orthologs for the genome, brain, SYN and PSD (Table 5.6). This might reflect a lower similarity of the bat cerebellum with

the mouse, compared to the other bat brain regions. Next, for each pair of orthologs (Bat:Mouse) its orthology type was determined (Figure 5.14). A higher number of 1:1 orthologs were observed in the brainstem, along with a larger amount of many:1, which might suggest that various duplication events took place in enriched-brainstem genes.

Table 5.6: **Tissue-enriched transcripts and percentage of mouse orthology.**

Tissue	**#Contigs**	**%Genome**	**%Brain**	**%SYN**	**%PSD**
		Mouse orthology			
Cortex	94	98.94	86.10	24.47	17.02
Cerebellum	213	87.32	61.94	16.43	11.74
Brainstem	89	95.5	86.92	24.72	16.86

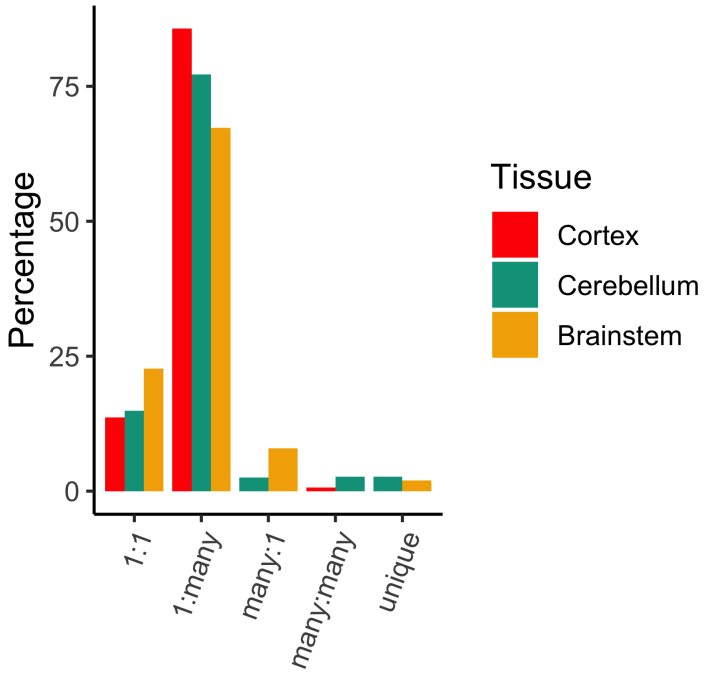


Figure 5.14: **Tissue-enriched orthology types distribution,** between bat and mouse. For each pair of orthologs its orthology type was obtained (Bat:Mouse; 1:1, 1:many, many:1, many:many or unique to bat.)

Gene ontology enrichment was determined for each bat brain region, in conjunction

with PSD mouse ortholog genes. In general, a larger number of GO terms were observed for the cortex (Figure 5.15). Moreover, ontologies relevant to the basic machinery of the PSD were also enriched in the cortex, namely; "synapse", "postsynaptic density", "postsynaptic membrane", "dendrite", "calmodulin binding", etc. In addition, ontologies that might be related to bats and echolocation, such as; "social behavior", "vocal learning" and "vocalization behavior" were also enriched in the cortex. Genes involved in social behavior, included; MAPK8IP2 (mitogen-activated protein kinase 8 interacting protein 2), PTCHD1 (patched domain containing 1) and neural-specific BRINP1 (BMP/retinoic acid inducible). The latest is expressed in abundance since early development stages of the central nervous system and play an important role in neural development (Motomiya et al., 2007; Nakatani et al., 2005). Knockout mouse has shown that the absence of BRINP1 causes abnormal behaviours analogous to human schizophrenia and attention-deficit disorder (Kobayashi et al., 2014).

Bats have stunning forms of sophisticated communication, including; vocal dialects, calls for trouble, courtship and territorial songs (Rodenas-Cuadrado et al., 2015). The ability to adjust vocalizations in response to auditory inputs, termed vocal learning, has been observed only in a few non-human animals, such as birds and a handful of mammalian species, including bats (Knörnschild et al., 2010). Bat genes found involved in vocal learning (GO:0042297) included; neurexin (Nrxn1 and 2), Forkhead box P2 (Foxp2), contactin (Cntnap2), Shank3, stimulated by retinoic acid gene 6 (Stra6) and huntingtin (Htt). Studies have found that initially young pups manifest an assorted repertoire of calls (such as the babbling of babies), which is not present in adults, yet after exposure to adult vocalizations, pups are able to learn and repeat these acoustics (Boughman, 1998; Knörnschild et al., 2010; Prat et al., 2015). While the evolution of human language remains unclear, primitive vocal learning process in the bat cortex might give us insights to the evolutionary base in the complexity of language formation.

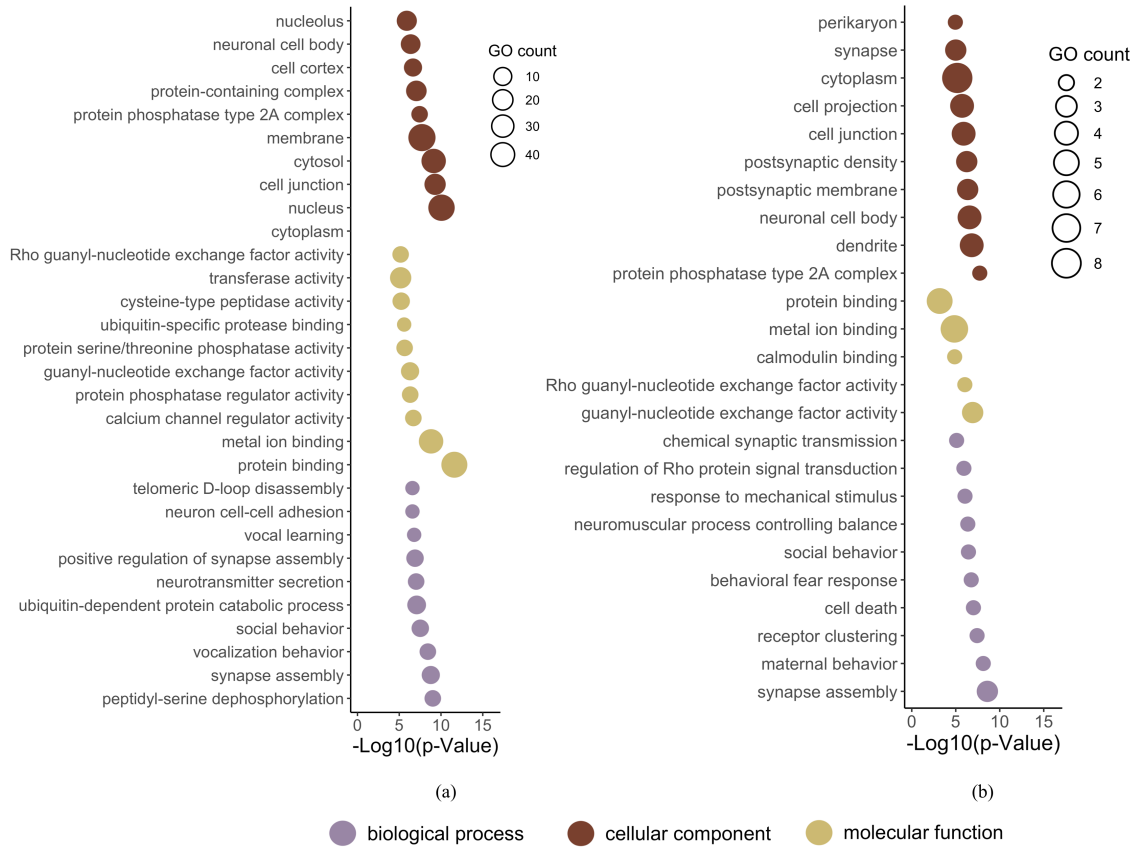


Figure 5.15: **Cortex-enriched GO analysis.** Distribution of GO terms reflecting statistical significant differences (hypergeometric distribution applying a p-value and FDR cut-off of 0.5). Different sizes in circles reflects the number of genes involved in each term. a) Cortex-enriched (n=94). b) PSD ortholog genes (n=16).

The bat brainstem has been a focal point when researching vocalization. Echolocation calls are generated through an intricate coordination of motor actions, concerning the governing of laryngeal, respiratory and articulary muscles of throat, mouth and nose (Schuller and Radtke-Schuller, 1990). The mesencephalic part of the brainstem is recognized to play a functional role in motor coordination (Schuller and Radtke-Schuller, 1990). The bat brainstem encompass several neurons adapted to biological crucial parameters of sound (signal interval, frequency-modulated sweeps, etc) through their interaction of time-delayed excitatory and inhibitory functioning (Rodenas-Cuadrado et al., 2015; Covey and Casseday, 1999).

In addition, glucocorticoid receptor activity (GO:0004883) was observed enriched in the brainstem. Glucocorticoids mediate homeostasis, but also their release has been linked to the circadian peak of the activity phase of the animal. Interestingly, studies have reported high levels of this opioid receptor in nocturnally active animals in the course of their active period (Dickmeis, 2009; Yoshida et al., 2005).
Enriched GO terms in the brainstem as well as in the cortex, also displayed several ontologies associated to synaptic functioning (Figure 5.16).

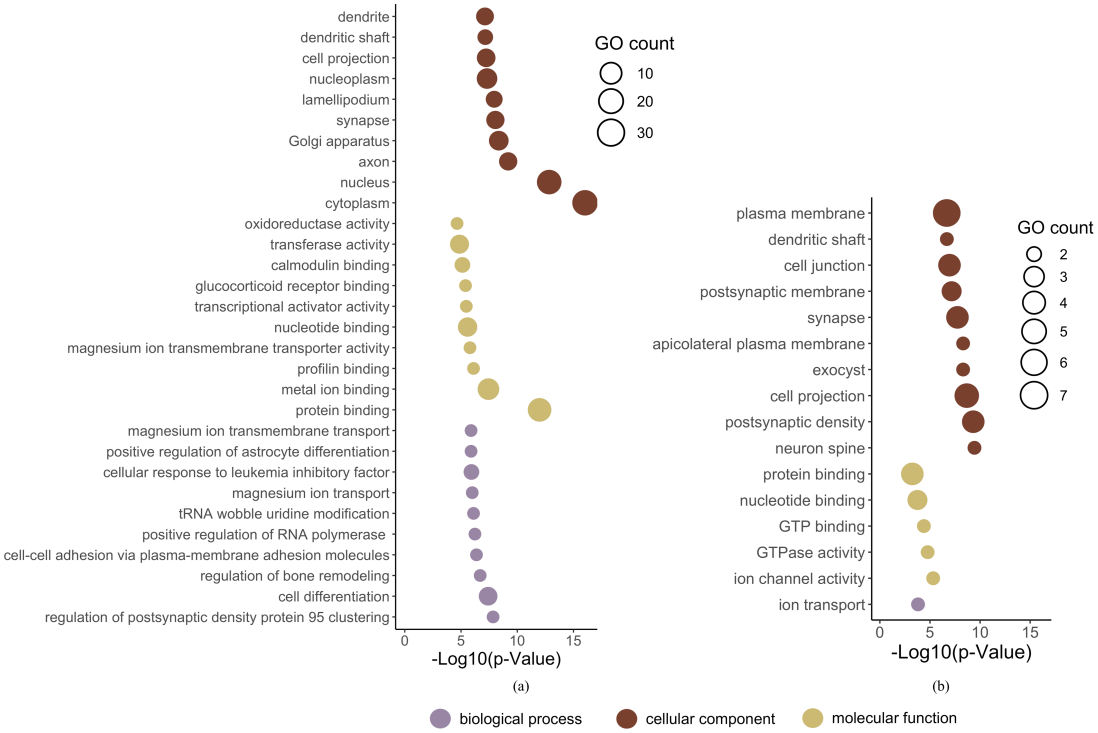


Figure 5.16: **Brainstem-enriched GO analysis.** Distribution of GO terms reflecting statistical significant differences (hypergeometric distribution applying a p-value and FDR cut-off of 0.5). Different sizes in circles reflects the number of genes involved in each term; a) Brainstem-enriched (n=89), b) PSD ortholog genes (n=15).

Ontologies involving motor neuron and spinal cord were observed within the cerebellum-enriched GO analysis (Figure 5.17). Bats need to be able to accurately coordinate their flight pattern upon an echo source. The cerebellum of bats is not only capable of coordination of motor movements, but it has been reported that neurons in the cere-

bellum can directly respond to acoustic stimuli (Horikawa and Suga, 1986; Jen and Schlegel, 1980). Noteworthy was the enrichment of the vitamin D binding GO term (GO:0005499). Studies have reported the expression of the vitamin D receptor in specific brain regions, including; the cerebellum, termporal lobe, amygdala, thalamus and hippocampus. Vitamin D has been acknowledged for its important role in the regulation of bone metabolism. However, the presence of vitamin D receptor is susceptible to ageing, and has been linked to dysfunction of cognition and dementia (Buell and Dawson-Hughes, 2008; Anjum et al., 2018).

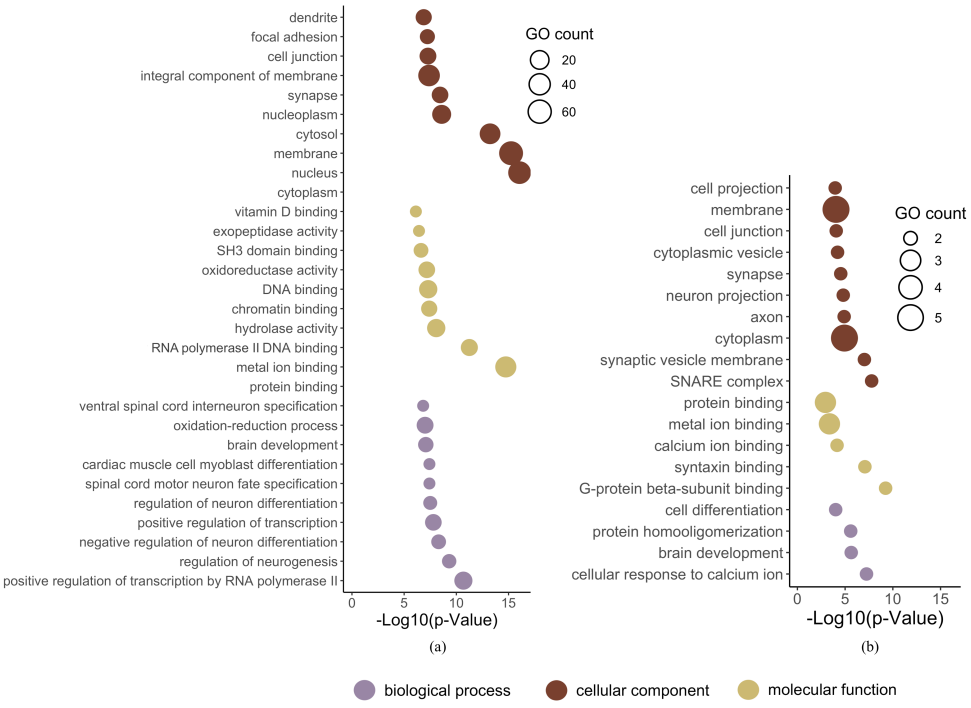
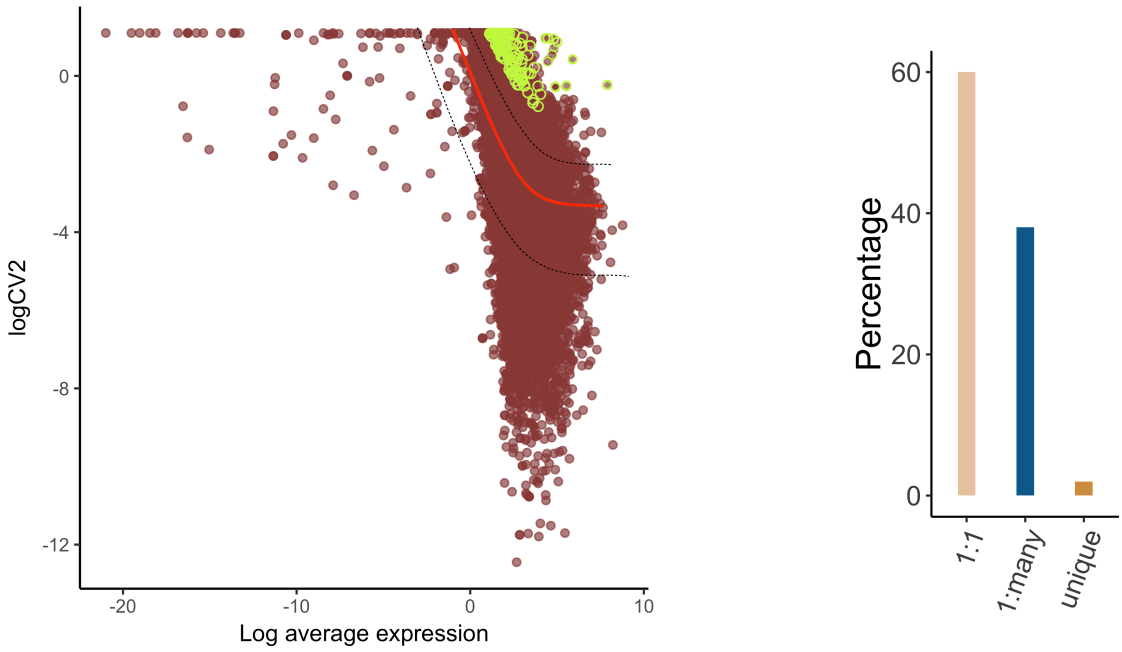


Figure 5.17: **Cerebellum-enriched GO analysis.** Distribution of GO terms reflecting statistical significant differences (hypergeometric distribution applying a p-value and FDR cut-off of 0.5). Different sizes in circles reflects the number of genes involved in each term; a) Cerebellum-enriched (n=213), b) PSD ortholog genes (n=25).

5.4.3 Transcripts with highly variable expression between tissues

Transcripts with highly variable expresssion were determined based on the per gene squared coefficient of variation (CV2) in the three tissues (Figure 5.18). A total of 42 transcripts were determined as highly variable (p-Value ≤1e-3, chi-squared distribution), of which 12 and 6 corresponded to mouse ortholog genes expressed in the SYN and PSD. In contrast to the constraint level seen in the bat genome (Figure 5.9), the 25 of the highly variable transcripts were 1:1 orthology type (true-orthologs). Variability of gene expression is a mechanism that produces diversity and provides insight to phenotypic variation (Raj et al., 2010). Hence, potentially this variability in expression is a major type of population variation that permits evolutionary change to become fixed in a population.

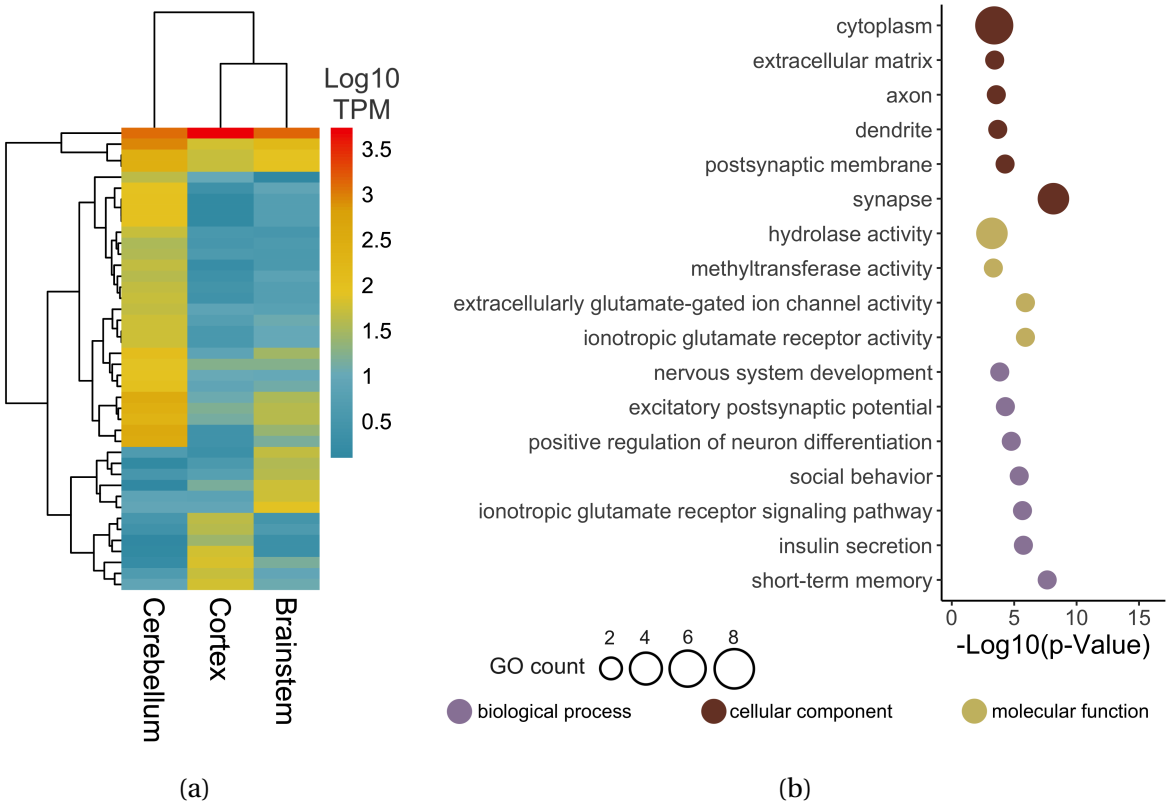


(a) (b)

Figure 5.18: **Squared coefficient of variation (CV2).** a) For each transcript it was calculated the CV2 (ratio of the standard deviation of its expression to the mean value). The x-axis shows the log mean expression of the 3 tissues; the y-axis represents the log transformed CV2. The solid red curve represents the fitted variance-mean dependence; the dashed lines reflects a 95% confidence interval; green dots correspond to transcripts which CV2 is significantly higher than 50% (CV2 > 0.25), b) Orthology types of highly variable transcripts.

Several of the variable transcripts were highly expressed in the cerebellum (22 transcripts > 50 TPM), and less expressed in the cortex (8 contig > 50 TPM) (Figure 5.19a). 22 of the highly variable transcripts corresponded to the cerebellum-enriched set of transcripts, while only 7 and 5 corresponded to cortex and brainstem-enriched, respectively. Therefore, this suggest the cerebellum as the most diverse region of the bat brain. This is also consistent with obtained results, i.e., the cerebellum differs in expression in comparison with the cortex and brainstem, clusters separately and shows a larger number of tissue-enriched contigs, thereby it is placed as an out-group. The capability of echolocation might be an explanation in this matter, since the cerebellum is the brain region responsible for motor functions, and echolocation requires high-

ly coordinated motor actions. Specialized neurons that respond to auditory stimuli have been observed in the echolocator bat (Jen and Schlegel, 1980). Moreover, toothed whales and microbats have a proportionally larger cerebellum in comparison with other species (Marino et al., 2000). Several enriched GO terms involving the synaptic functioning were observed in the highly variable transcript set, such as; the neuron, axon and synapse, but also mechanisms regulating synaptic vesicle fusion (Figure 5.19b).



(a) (b)

Figure 5.19: **Highly variable transcripts.** a) Heatmap representation of highly variable transcript levels. Tissues and transcripts were hierarchically clustered. Colours distinguish highly expressed genes (red) from lowly expressed (blue). b) GO analysis of the 44 highly variable transcripts. Enriched GO terms are represented in different colours, while the size in circles illustrates transcripts count.

5.4.4 Gene expression of key synaptic genes

Expression of key synaptic protein coding genes was explored in the bat brain (Figure 5.20). Notwithstanding that within the cerebellum, lower expression profiles were

observed in some genes (i.e., *CAMK, DLG, SYNGAP, HOMER, GABRA4*), comparable expression patterns were obtained in the three brain tissues. Considering that most synaptic proteins contribute to the rich interaction networks (Emes and Grant, 2012; Bayés et al., 2012; Bayés and Grant, 2009), it is therefore comprehensible that an overall conserved organization would be found within different regions of the vertebrate brain.

Several of the highly expressed synaptic protein coding genes have been linked to the molecular mechanism of hearing. For example, in a study conducted to assess whether listening to classical music triggers any influence in the human transcriptome, *NRGN (neurogranin)* was an up-regulated gene after exposure to music (Kanduri et al., 2015). *NRGN* encodes a protein kinase substrate that is the main postsynaptic scaffold controlling the availability of calmodulin in the absence of calcium (de Arrieta et al., 1999). While a study reported lack of expression of this gene in the brainstem and cerebellum of rat brain (Represa et al., 1990), in this study it was ubiquitously enriched in the three bat brain regions.

VDAC (voltage-depended anion channel), a mitochondrial outer membrane protein, is pivotal for the influx regulation of ions and molecules. This protein is also essential in mitochondrial-mediated apoptosis (Shoshan-Barmatz et al., 2010). Similarly, VDAC has been linked with impairment of the inner ear hair cells function, leading to hearing loss (White et al., 2018; Seo et al., 2017).

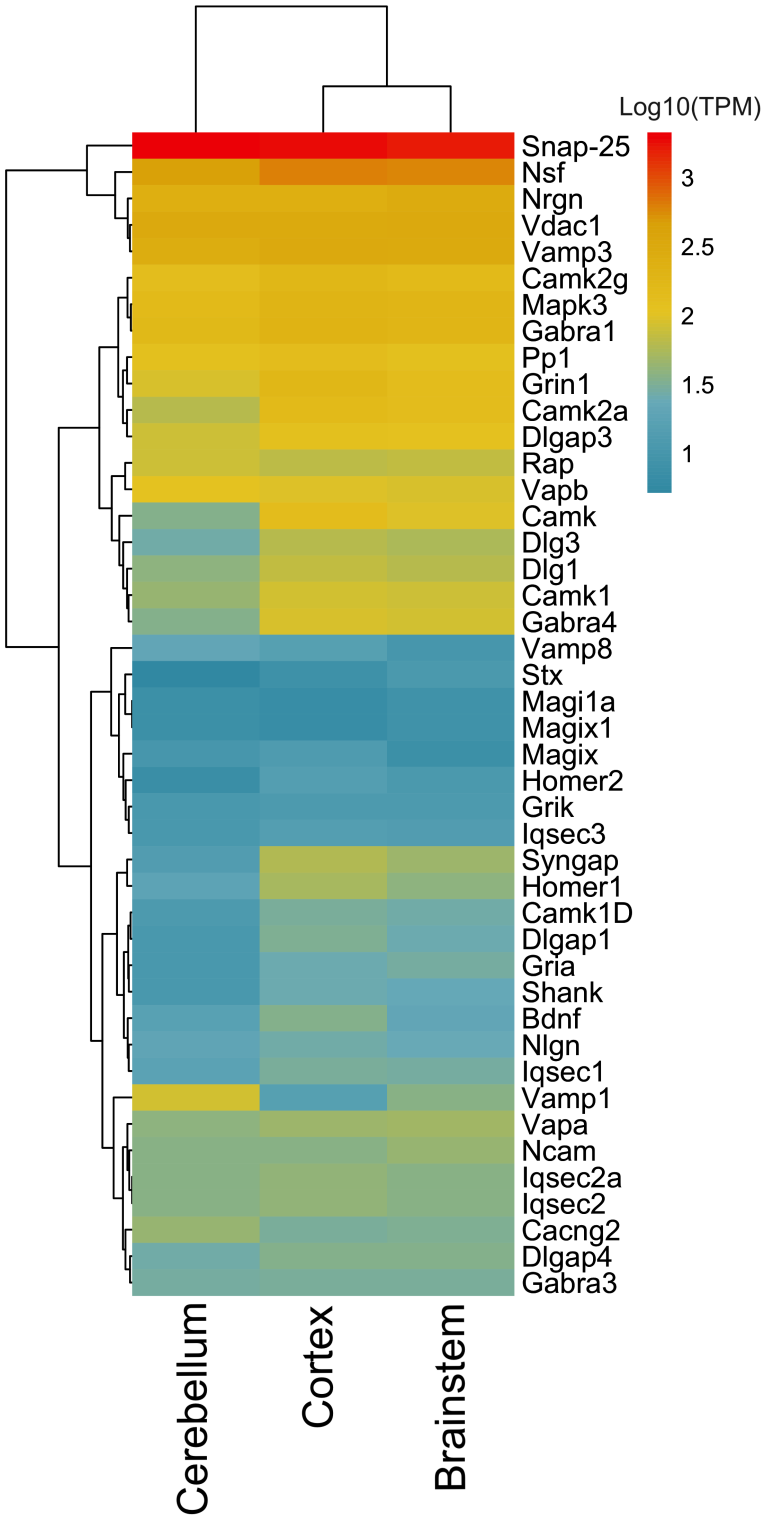


Figure 5.20: **Expression of key synaptic genes.** Tissues and genes were hierarchical clustered. Considering the presence of splice-forms, the mean expression value was obtained per gene, and its expression in TPM was log10 transformed. Colours distinguish highly expressed genes (red) from lowly expressed (blue)

Members of the SNARE complex (e.g., *SNAP-25*), which are involved in vesicle docking and synaptic transmission, play a significant role in molecular mechanisms of synaptic exocytosis in the inner ear hair cells. Mutations in the *OTOF (otoferlin)* gene inhibit its binding with SNARE proteins causing deafness (Ramakrishnan et al., 2009). An over-expression of *SNAP-25* has also been linked to neural plasticity during hibernation in bats (Lei et al., 2014). Enriched expression in ear hair cells and therefore important for the molecular basis of hearing, is also seen in *MAPK* (mitogen-activated protein kinases) (Jamesdaniel et al., 2011) and *CAMK2* (Ca2+/calmodulin-dependent protein kinase). Suppression in the expression of *CAMK2*, causes malformation of the inner ear (Rothschild et al., 2013).

Brain-specific genes in mammals generally exhibit lower rates of evolution compared to genes expressed elsewhere (Wang et al., 2006). Therefore, it was next investigated the existence of synaptic genes under positive selection between mouse and bat. Using ParaAT and KaKs calculator, non-synonymous to synonymous substitutions dN/dS were estimated for 1:1 orthologs of 3,217 genes known to be expressed in the synapse. A total of 11 synaptic genes were identified to have undergone positive selection (ω>1) (Table 5.7).

Table 5.7: **Synaptic genes with evidence of positive selection.**

Gene	TPM	Omega	Function
AHCYL2	57.10	1.33	NAD binding and adenosylhomocysteinase activity
AKAP7	15.73	1.05	Nucleotide binding and protein kinase A binding
CHPT1	22.39	1.06	Diacylglycerol binding and cholinephosphotransferase activity
DGKZ	61.78	1.09	Protein C-terminus binding and NAD^+ kinase activity
HUWE1	28.98	5.46	Ligase activity
MCF2L	10.37	1.10	Rho guanyl-nucleotide exchange factor activity and 1-phosphatidylinositol binding
SCN8A	47.50	1.10	Ion channel activity and voltage-gated sodium channel activity
SLC12A5	189.04	1.22	Protein kinase binding and potassium:chloride symporter activity
SLC25A10	59.81	1.04	Antiporter activity and phosphate ion transmembrane transporter activity
TANC2	13.42	1.28	Ankyrin-Repeat Containing Protein
UQCC1	15.12	1.52	Mitochondrial respiratory chain complex II-I assembly

Among these genes, *HUWE1* displayed the highest omega value (ω=5). HUWE1 induces mitochondrial autophagy (mitophagy) by the interaction with AMBRA1 (autophagy/beclin-1 regulator-1) (Di Rita et al., 2018). Interestingly, mitophagy prevents ageing, and has widely investigated in this respect (Diot et al., 2016). Bats have an exceptional longevity (Brunet-Rossinni and Austad, 2004), thereby it is likely that HUWE1 plays a critical role in the evolution of ageing retardation in bats. In addtion, mutations on *HUWE1* have been reported to contribute to mental retardation (Froyen et al., 2008), as well as several types of cancer when an over-expression occurs (Wang et al., 2014).

On the other hand, *SLC12A5* and mainly *SLC25A10* solute carrier revealed a notably high expression in the 3 brain tissues. Such genes are responsible for cochlear amplification in echolocating bats (Alvarez-Leefmans and Delpire, 2009). More importantly, the solute carrier *SLC26A5* protein (*Prestin*), not only has been reported to have undergone convergence evolution between echolocating Yinpterochiroptera and Yangochiropterans bats (Li et al., 2008), a more striking convergence between toothed whales and echolocating bats was reported (Li et al., 2010b). *Prestin* represents a key element of the cochlear amplifier that supports the high sensitivity of the inner ear of mammals (Liberman et al., 2002).

5.5 Phylogeny of echolocating genes

Echolocation is the generation of sonar signals to the environment and interpretation of the returning echoes for navigation, evasion of obstacles and prey capture. Animals capable of echolocation use a complex interaction of systems involved in vocalization, auditory and neural that have been subjected to evolutionary changes (Teeling et al., 2016). Echolocating bats posses physiological and morphological adaptations. For example, echolocating bats and toothed whales have the most sophisticated auditory systems for the detection of ultrasonic sounds (≥200 kHz) (Davies et al., 2013).

Fossil evidence suggests that the bat's common ancestor (~64 mya) possessed very small eyes and an auditory brain structured to enable laryngeal echolocation. It is presumed that this trait might have evolved once in bats and afterwards been lost in Old World fruit bats (pteropodids). Alternatively, echolocation independently evolved at least two times in echolocating bats (Thiagavel et al., 2018). In contrast, Old World fruit bats, which eat only plants and fruits, evolved a larger body size, excellent olfaction and large eyes which enable them to have an exceptional dim light vision (Marshal, 1983).

Besides, considering the sophisticated mechanism of echolocation, bats are not the

only mammal to have evolved this ability. In the same manner as bats, toothed whales (Odontoceti) independently evolved this trait. However, the way that bats and whales echolocate differs slightly. In both species the sonar is generated in the larynx, but in bats it is propagated through the nostrils or mouth (Teeling, 2009). Whereas whales have specialized air-space nasal structures in their forehead, called the monkey-lips-dorsal-bursae (MLDB) that allows the pressurization of air. This sonar is then transmitted to a large acoustic fat body, called the melon (also in the forehead) that focus the sound beam just before its emission to the water (Huggenberger et al., 2016; Madsen et al., 2005). The ability of laryngeal echolocation represents an example of parallel or convergent evolution by natural selection (Shen et al., 2012).

In this research, it was set to test these assumptions using the generated data. Available homologous (1:1) protein sequences of mammalian species were compiled, including echolocating and non-echolocating bats and whales. Finally, using neighbor joining (NJ) methods (100x bootstrapping), protein trees were reconstructed from the previously aligned sequences.

Based on their association with hearing and echolocation five genes were analysed, namely; *FOS* (Fos proto-oncogene, AP-1 transcription factor subunit) (Figure 5.21(a)), *SLC45A2* (solute carrier family 45 member 2, also called *MATP*) (Figure 5.21(b)), *RGS7BP* (regulator of G protein signaling 7 binding protein) (Figure 5.22(a)), *USH1G* (Usher syndrome type-1G) (Figure 5.22(b)), *TMC1* (transmembrane channel like 1) (Figure 5.23).

Implicated in the vocalization of echolocating bats, *FOS* has been suggested as a candidate gene involved in this trait (Zhang et al., 2012a; Schwartz and Smotherman, 2011). Despite that Figure 5.21(a), did not display a clustering among echolocating bats and whales. Surprisingly, the hyrax did cluster together with echolocating bats. Hyraxes are small mammals closely related to elephants, moreover, the male generates complex vocalizations (or songs) emitted from the larynx that depict their own identity, age, so-

cial position and body condition (Demartsev et al., 2017; Koren and Geffen, 2011).

Phylogenetic reconstructions of *SLC45A2*, *RGS7BP*, *USH1G* and *TMC1* protein sequences, showed distinctions between the yinpterochropteran (echolocating) and yangochiropterans (most non-echolocating) bat clades. Along with the separated branches of toothed whales (echolocating) and baleen whales (non-echolocating). Similarly as Li et al. (2010b), evidence of molecular adaptation and strong sequence convergence between unrelated echolocating bats and toothed whales was found in the transmembrane protein *TMC1*. This gene serves as a sensory transduction in the inner and outer hair cells, mutations yield dominant and recessive deafness in humans and mice (Pan et al., 2013; Marcotti et al., 2006). Of note, the Egyptian fruit bat is the only Old World Bat that uses echolocation, which differs to the laryngeal echolocation used by yinpterochropteran bats and toothed whales (Jones and Teeling, 2006).

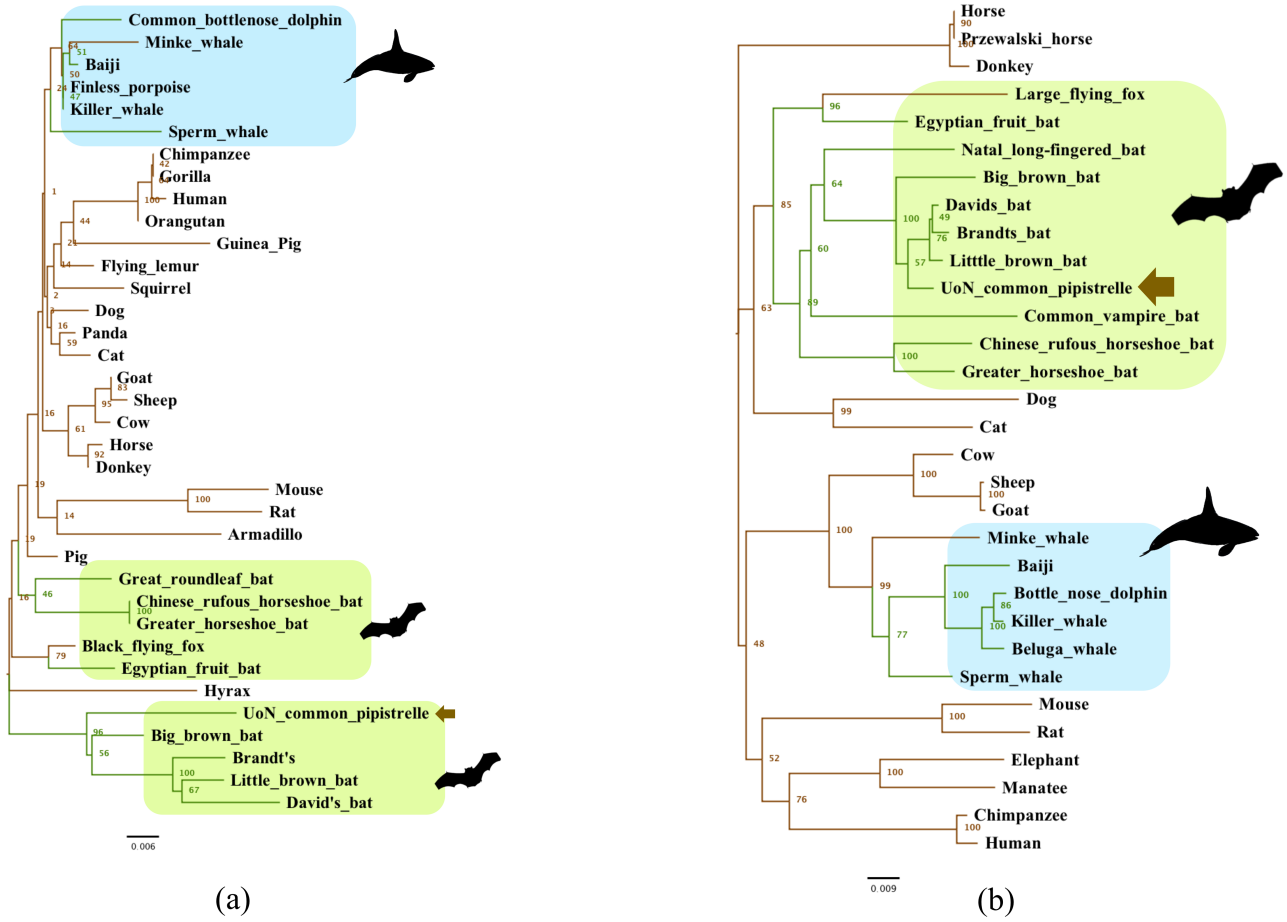


Figure 5.21: **Neighbor joining phylogenetic tree,** based on the protein sequence of; a)***FOS*** and b) ***SLC45A2***. Trees were constructed using homologous sequences of 31mammalian species with bootstrap values indicated on each branch (100x). Green branches depict the phylogentic position of echolocating bats and echolocating toothed-whales. Bats and cetaceans are shown in a green and blue shadow, respectively. An arrow indicates the position of the common pipistrelle.

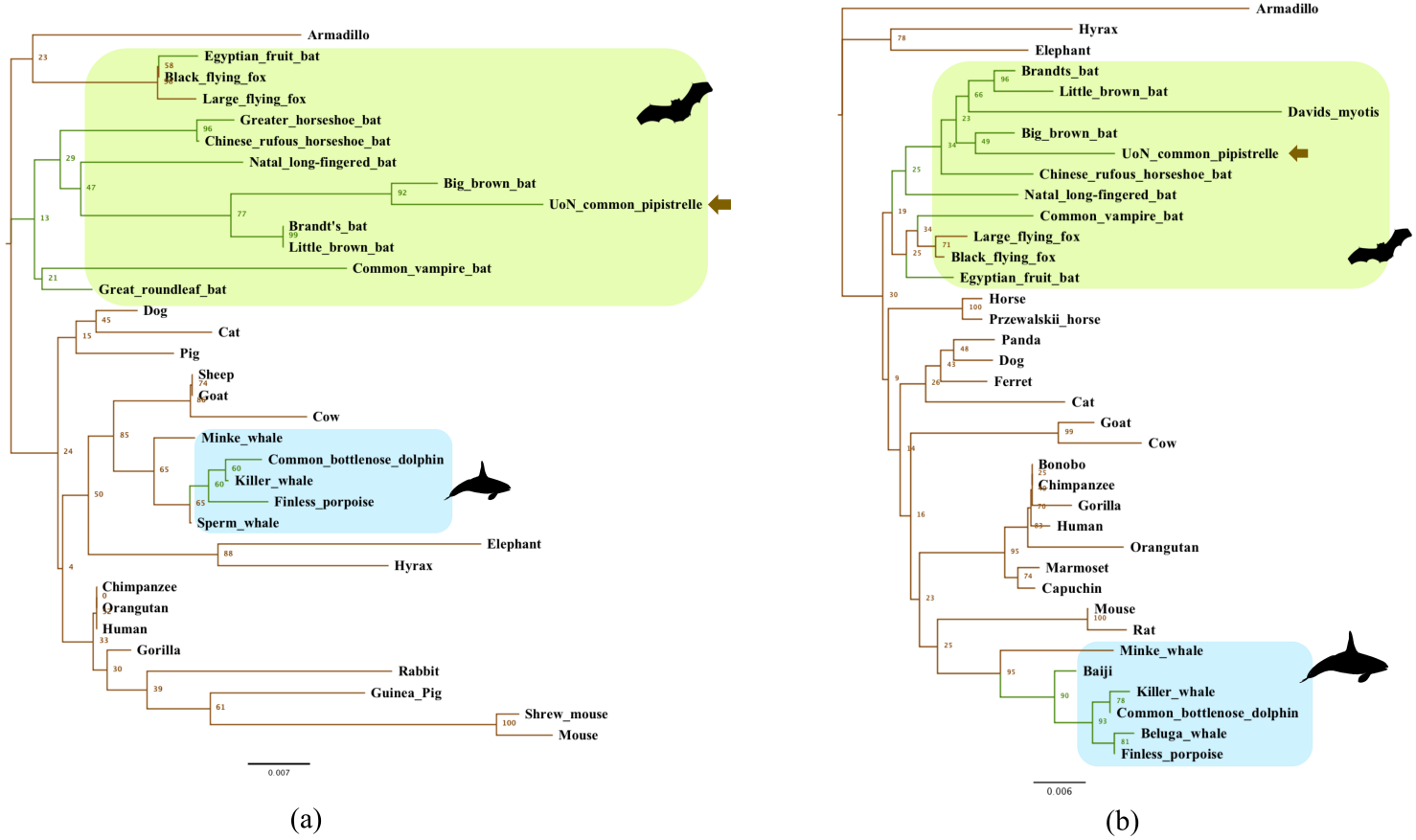


Figure 5.22: **Neighbor joining phylogenetic tree,** based on the protein sequence of; a) ***RGS7BP*** and b) ***USH1G***. Trees were constructed using homologous sequences of 34 and 37 mammalian species with bootstrap values indicated on each branch (100x). Green branches illustrate the phylogentic position of echolocating bats and echolocating toothed-whales. Bats and cetaceans are shown in a green and blue shadow, respectively. An arrow indicates the position of the common pipistrelle.

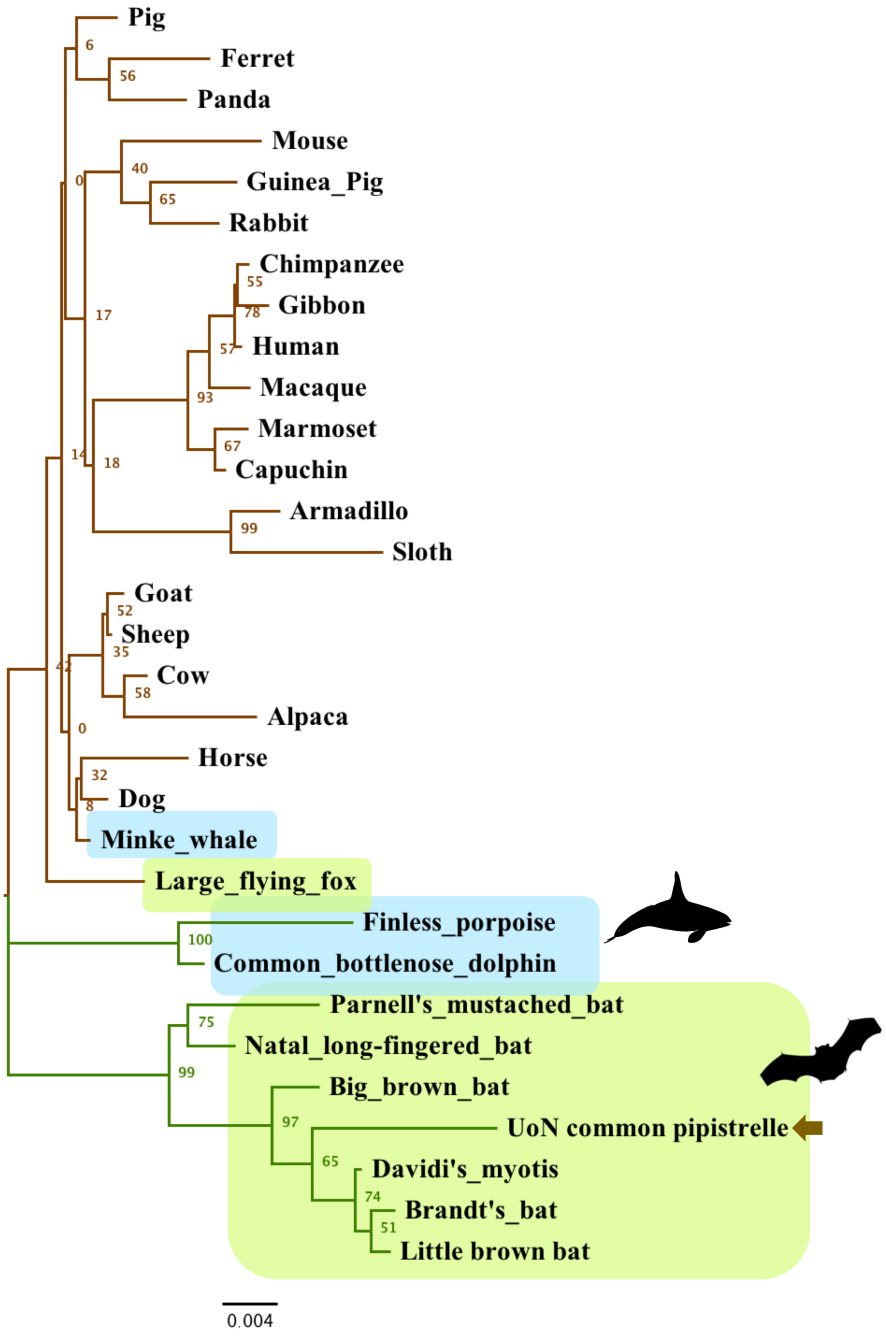
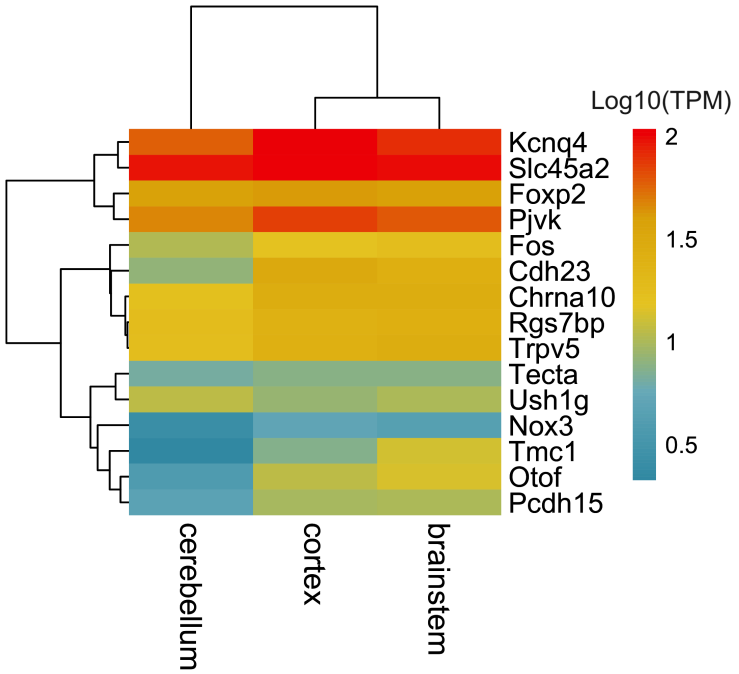


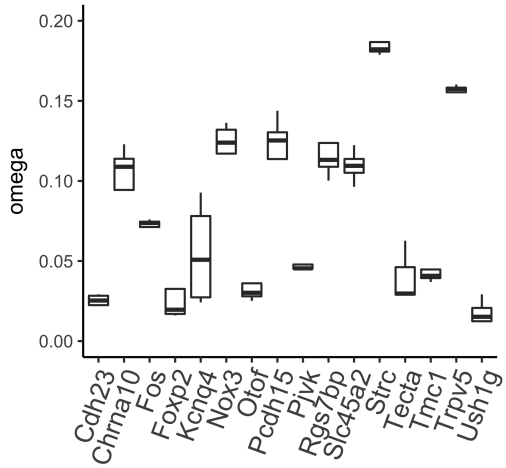
Figure 5.23: **Convergent evolution of echolocating bats and toothed whales,** based on the protein sequence of ***TMC1***. The NJ tree was constructed using homologous sequences of 31 mammalian species with bootstrap values indicated on each branch (100x). Green branches reflect the phylogentic position of echolocating bats and echolocating toothed-whales. Bats and cetaceans are shown in a green and blue shadow, respectively. An arrow indicates the position of the common pipistrelle.

Expression levels of key genes associated with echolocation were analysed in the common pipistrelle assemblies (Figure 5.24a). In general, a comparable expression pattern was noted in the three tissue, yet, the cerebellum showed lower expression, particularly for *OTOF*, *CDH23*, *PCDH15* and *TMC1*. Shen et al. (2012) investigated the expression of *OTOF* (which encodes the protein Otoferlin) in bat brains, and reported a 70-fold expression higher in the cortex compared to the cerebellum. This protein is a Ca^{2+} sensor, which regulates neurotransmitter release at the ribbon synapse of cochlear hair cells. Mutations trigger an autosomal recessive nonsyndromic form of prelingual and sensorineural deafness (Varga et al., 2003; Yasunaga et al., 1999). *CDH23* and *PCDH15* that encode cadhedrin 23 and protocadhedrin 15, are necessary to hair bundle motility of the inner ear (Ahmed et al., 2006; Siemens et al., 2004).

KCNQ4 (Potassium voltage-gated channel, KQT-like subfamily, member 4) displayed the greatest expression level. It is expressed in the outer hair cells, and in humans, mutation of this gene causes progressive hearing loss at a young age (Nie, 2008). In bats, a monophyletic group was reported in bats that use laryngeal echolocation. Also, the evolution of *KCNQ4* has shown various parallel patterns seen in *SLC26A5 or Prestin*, which underwent convergent evolution between echolocating taxa (Liu et al., 2011). It has been proposed, that sequence convergence of *Prestin* is a result of positive selection. Subsequently, omega values were estimated from 1:1 orthologs between mouse and the common pipistrelle assembly (Figure 5.24b). Little evidence of positive selection was found in the hearing related genes.



(a)



(b)

Figure 5.24: **Key genes associated with echolocation.** a) Expression of putative echolocation genes in the cortex, brainstem and cerebellum of the common pipistrelle. Tissues and genes were hierarchical clustered. Colours distinguish highly expressed genes (red) from the ones with a lower expression (blue); b) Evolutionary rates (omega) of key hearing related genes associated with echolocation. Estimated based on pairwise alignments between mouse and echolocating bats.

5.6 Summary and comments

By means of NGS, reads were obtained from 3 regions of a common pipistrelle brain (cortex, brainstem and cerebellum). In this study the effective bioinformatics pipeline for *de novo* transcriptome assembly and annotation developed, was used. The quality of this bioinformatics pipeline was evaluated comparing the common pipistrelle assembly to 6 bat species and mouse, which showed consistent results with the evolutionary adaptation of these bats.

This work provides the first exploration of bat transcripts (and genes) encoded in the brain (n=13,160) including those encoding the synaptosome (SYN) (n=3,374) and post-synaptic density (PSD) (n=2,283). Forty four key genes known to be crucial for learning and memory were found expressed in the common pipistrelle. Genes with the highest expression levels are important for the molecular mechanism of hearing, i.e., *NRGN, V-DAC, CAMPK, MAPK*, in addition to members of the SNARE complex; molecular mechanisms of synaptic exocytosis in the inner ear hair cells and neural plasticity while hibernation (Figure 5.20). Twelve synaptic genes were also identified with evidence of positive selection.

A substantial number of 1:many (Bat:Mouse) orthologs were identified, suggesting the presence of strong gene constraint in the bat genome. This is likely a consequence of high metabolic demands. Yet, when estimated highly variable transcripts, the majority were 1:1 orthology type (true-orthologs). Therefore, it was implied that variability in expression is a major factor generating evolutionary change.

Different analyses generated in this study are consistent and demonstrate that the bat cerebellum is the most distinct region when compared to the brainstem and cortex. As well as less involved in cognition and echolocation. This is also invariable with comparative neuroanatomy, which has indicated the cerebellum as an outlier from the rest

of brain regions (Strand et al., 2007).

Finally, evidence of convergent evolution acting on unrelated echolocating mammals (bats and cetaceans) in the hearing gene *TMC1* was found. In conclusion, these comparative data generates a robust basis for future comparative studies of bat in the framework of evolution, disease and ageing, along with the understanding of the molecular mechanisms underpinning echolocation.s

Chapter 6

De novo Assembly of the Lion brain transcriptome

“El que lee mucho y anda mucho, ve mucho y sabe mucho."

- Miguel de Cervantes -

The lion (*Panthera leo*), the second largest Felidae species and Africa's key predator, is a formidable, and for many people a charismatic carnivore. Lions have captivated human populations since pre-historical ages, being inspirational animals in several cultures (Antunes et al., 2008). Although its population is nowadays restricted to sub-Saharan Africa and a confined area in India, fossil evidence suggests that lions originated in the late Pliocene in the grasslands of east Africa (~2-1.5), but they then spread over the majority of Africa, Europe, Asia, North America and regions of South America (Turner, 2000). As a result, lions are considered to have influenced the evolution of other sympatric carnivores through direct and indirect competition. Hence evolution of the lion has had crucial implications for paleoecology studies and evolutionary exploration of more cat species (Yamaguchi et al., 2004).

Despite the genetic diversity that has been reported within several subpopulations in Africa, two subspecies are officially accepted by the International Union for the Conservation of Nature (IUCN), the African lion (*Panthera leo leo*) and the Asiatic lion (*Panthera leo persica*) (Bauer et al., 2012). Estimates of wild lions in Africa in a 2004 inventory was about 16,500 to 30,000 individulas (Bauer and Van Der Merwe, 2004). However, the African lion is categorized as "vulnerable" on the Red list of Threatened species (Bauer et al., 2012). These species currently faces habitat loss, a broad prey base depletion, improper regulated sport hunting, urge for traditional Chinese and African medicines, and the defensive killing for human and livestock safety (Bauer et al., 2015). Additionally, some lion populations have been affected by viral diseases, such as; canine distemper virus, and several feline specific viral diseases, including; retrovirus, parvovirus, calicivirus and herpesvirus (Martella et al., 2007; Packer et al., 1996).

Studies of nuclear and mitochondrial loci have elucidated the phylogeography and population genetics of lions (Barnett et al., 2006; Antunes et al., 2008). Low coverage genome sequence (1.9-fold coverage) of the domestic cat, have also helped to resolve the evolution of the Felidae family (Pontius et al., 2007). To date, no whole-genome

reference sequence has been published for the lion. In this study, it is aimed to provide the first description of gene expression in tissues of the brain and explore the expression of key synaptic genes, pivotal for learning and memory. In addition, it is provided a comparison with mouse, and four other feline species, including; the domestic cat (*Felis catus*), cheetah (*cinonyx jubatus*), leopard (*Panthera pardus*) and tiger (*Panthera tigris*).

6.1 *De novo* transcriptome assembly

A paired-end RNA-seq library was generated consisting of two brain tissues, i.e., forebrain and brainstem (including the cerebellum) with two replicates each (Figure 6.1), collected from one adult female lion (*Panthera leo*), donated from Twycross zoo in England. A total of 174,914,518 raw reads were produced and processed as described in Chapter 3, which reduced the raw read set to 174,636,130 (Table 6.1).

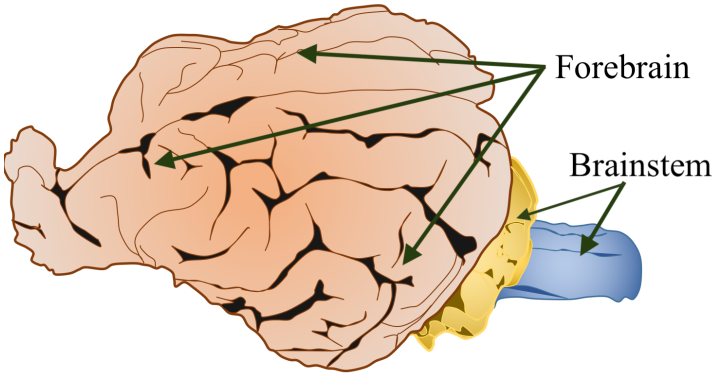


Figure 6.1: **Lion brain representation,** depicting the investigated brain tissues.

Table 6.1: **Summary of the RNA sequencing yield.**

Tissue	Raw reads	Clean reads	Base pairs (Mbp)
Forebrain 1	42,278,992	42,212,456	6,424.39
Forebrain 2	38,894,872	38,833,114	5,910.082
Brainstem 1	52,308,406	52,222,953	7,947.906
Brainstem 2	41,432,248	41,367,607	6,295.811

Cleaned reads were used to build a *de novo* assembly for each sample using Trinity. A total of 870,928 contigs (417 Mbp) were generated with a minimum sequence length of 201 base pairs and a maximum of 18,520 nucleotides. The assemblies were processed with Transrate to improve quality. These optimised assemblies (490,470 contigs, 234 Mbp) were additionally enhanced by discarding contigs which abundance was beneath 0.5 TPM. The assemblies were compressed to 456,398 transcripts with an average contig length of 979.408 bp (Table 6.2).

Table 6.2: **Quality improved Trinity *de novo* assemblies.**

Tissue	Contigs	Base pairs (Mbp)	Mean contig length	Median contig length	SD contig	%GC
Forebrain 1	98,524	97.88	993.43	508	1,162.10	49.32
Forebrain 2	107,834	95.28	883.56	453	1,049.10	48.52
Brainstem 1	128,398	126.15	982.49	479	1,212.28	49.71
Brainstem 2	121,642	127.70	1049.80	535	1,256.16	49.25

The proportion of transcripts that emerged as full-length or nearly full-length was examined to evaluate the quality of the assemblies. The BLASTX algorithm (cutoff E-value e-20) was used to compare each transcript to the SwissProt database (Figure 6.2). As a result of a larger size of both brainstem assemblies, these contain a larger number

of full-length transcripts (9,849 and 9,596 transcripts ≥ 80% length coverage) than both forebrain assemblies (9,202 and 8,262 transcripts ≥ 80% length coverage). Numbers of unique transcript hits to SwissProt were also higher for the brainstem assemblies (38,101 and 36,977) than the forebrain ones (34,163 and 32,504).

Additionally, transcriptome completeness of each assembly was evaluated by BUSCO (Figure 6.4). This software searches against a database of highly conserved 1:1 ortholog genes in vertebrates. The BUSCO results denoted that the assemblies were 72% to 81% complete. These BUSCO criteria values are agreeable with reported transcriptome assessments (Kordonowy and MacManes, 2016).

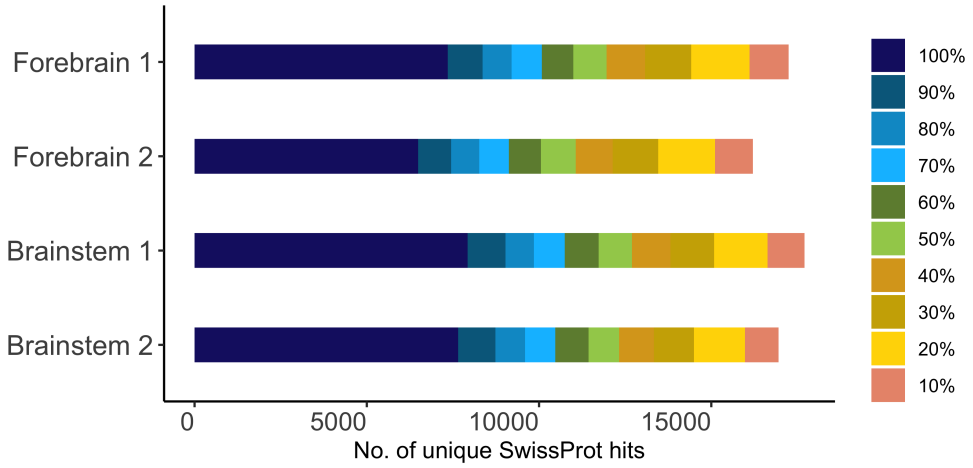


Figure 6.2: **Percentage of the transcript length that align to a known protein.** A metric for quality assessment of the assembly is to explore the portion of the transcripts that align a known protein. If ≥90% of the total length of a transcript aligns to a known protein, this transcript is considered "full-length" . BLASTX (e-value of e-20) was carried out to align the optimized Trinity-generated assembly to the SwissProt database. This used to assess the optimized Trinity-generated transcriptome, with no filtering of the data.

Generation of a non-redundant assembly

During the *de novo* assembly approach, multiple transcripts were built for single genes, likely consequence of the assemblage of incomplete RNAseq reads. These duplicated

transcripts depict redundancy in the assemblies, which can be problematic for expression analyses. All four assemblies were combined into a single non-redundant transcriptome (456,398 transcripts) using CD-HIT-EST. The non-redundant assembly constituted 44,941 non-redundant transcripts (7.2% were alternative splicing forms). This assembly was used as a single reference transcriptome for mapping back cleaned reads from all samples, and hereby compute unified assemblies for each sample. Alignment rates reflected the high quality of the assembly; 83.03%, 83.24%, 82.99% and 82.64% for forebrain 1 and 2 and brainstem 1 and 2, respectively with very few numbers of unmapped reads (Figure 6.3).

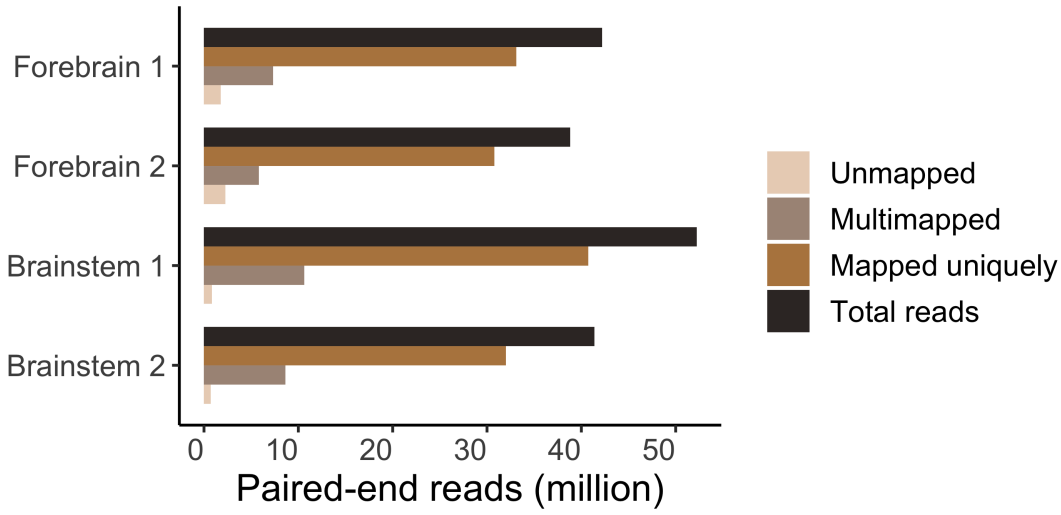


Figure 6.3: **Number of reads mapped back to transcriptome.** Trimmed reads from *Panthera leo* brain tissues (2x forebrain and brainstem) were mapped back to the generated non-redundant transcriptome (49,093 transcripts).

Transcriptome quality and completeness was evaluated additionally for the unified non-redundant assembly to identify any possible loss of valid transcripts (Figure 6.4). Results for this assembly improved the percentage of complete transcripts (83%) than the Trinity improved assemblies. Thereby, substantiating the use of this assembly for transcriptome annotation and downstream analyses.

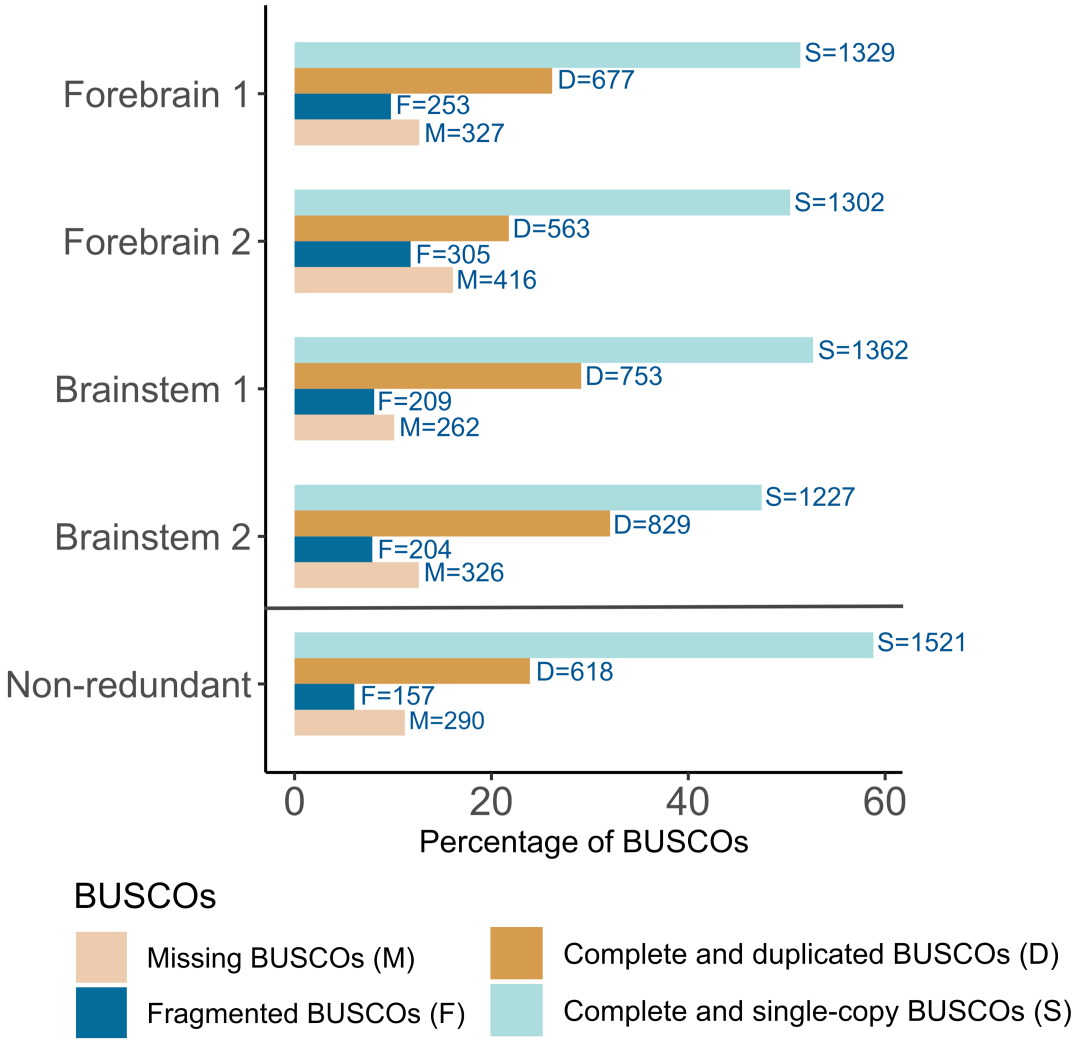


Figure 6.4: **Distribution of BUSCOs categories in the lion *de novo* assemblies.** Transcriptome completeness of the Trinity improved assemblies, in conjunction with the non-redundant assembly (44,941 transcripts), were assessed by BUSCO. Percentages of each BUSCO category is shown in different colours. The number of each category is depicted beside each bar. Such metric was only used for assessment of the transcriptome but not filtering procedure was carried out.

6.2 Transcriptome annotation

Annotation was performed by Dammit! (Scott, 2016), which successfully annotated 28,504 transcripts (63% of the non-redundant transcriptome). Transdecoder determined that 21,236 of the annotated trancripts potentially represented protein-coding

transcripts, as these comprised an ORF. 18,081 of these transcripts contain a complete ORF (containing a start and stop codon). Among other annotations, 5' and 3'-UTRs were detected (22,345 and 23,805, respectively). LAST search identified 90% of the transcripts (25,654) matched to the UniRef90 database, but only 2.7% matched to the Rfam database for ncRNAs. An outline of *de novo* pipeline assembly and annotation results are shown in Figure 6.5.

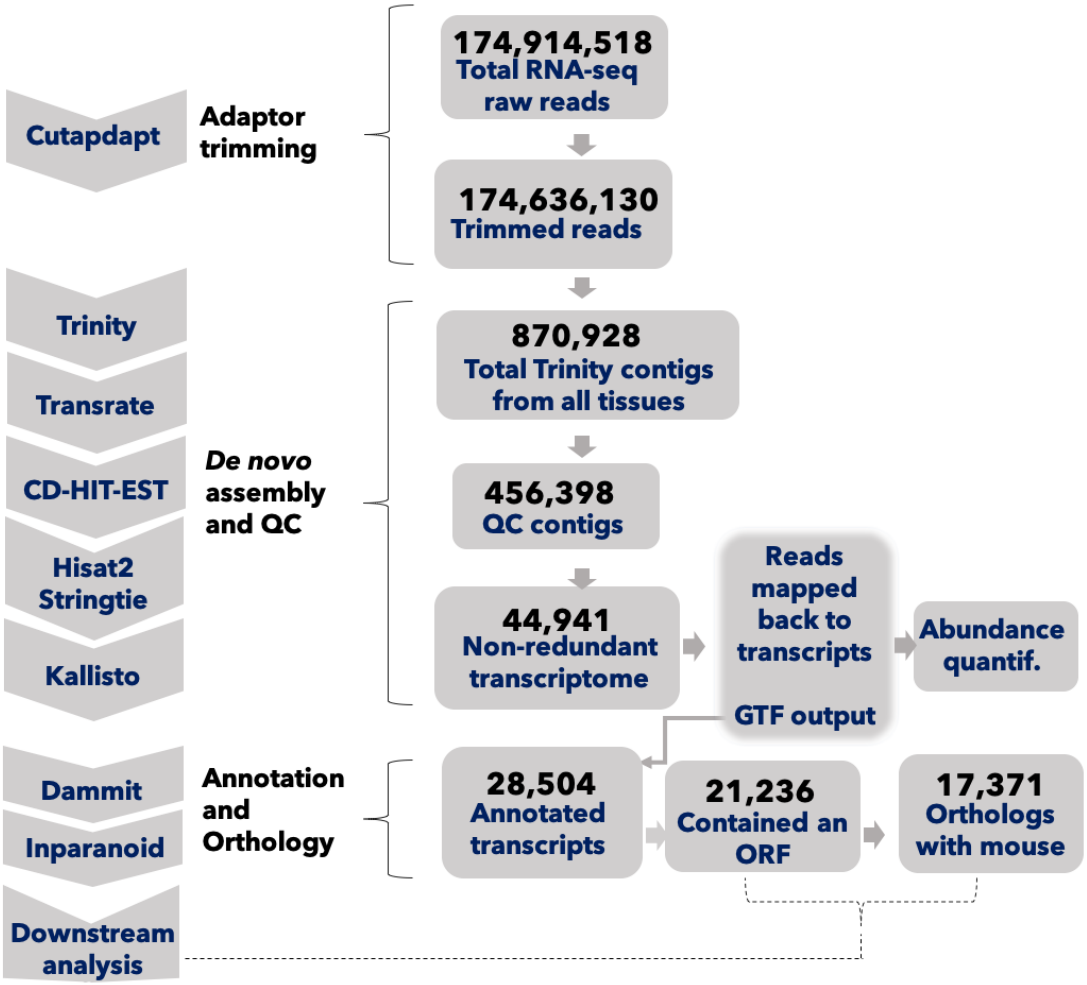


Figure 6.5: **Summary of de novo transcriptome assembly and annotation pipeline.** RNA- seq reads were trimmed (Cutadapt) and assembled (Trinity). Low-quality contigs were removed (Transrate) and a non-redundant transcriptome was obtained (CD-HIT), which served as reference to map back the RNA-seq reads from all tissues (Hisat and Stringtie) and its expression abundance was quantified in TPM (Kallisto). The transcriptome was annotated (Dammit) and ORFs were estimated (Transdecoder). Orthology was inferred (Inparanoid) for further downstream analyses.

Protein domain analysis

Protein domains (Pfam-A) were predicted for translations of all putative protein-coding transcripts (21,236 transcripts) using HMMER. 79.4% of the assembly contained 10,565 unique entries of the protein family database. From these transcripts, 43.12% had an E-value threshold of ≤ 0.05 to 5,325 pfam families. The top 50 enriched domains are shown in Figure 6.6. A diverse range of zinc finger domains were present in the assembly, i.e., 143 different types in 2,299 transcripts. The zinc finger family is a vast and diverse set of proteins that interact with DNA, RNA, proteins and small molecules. These are implicated in several cellular processes, such as transcription, translation, folding, DNA replication and repair, cell proliferation, signal transduction. Therefore, it is not surprising that their disruption has been linked with neurological problems, including schizophrenia, bipolar diseases and intellectual disability (Sun et al., 2015; Chasapis et al., 2012).

The zinc-finger double domain (zf-C2H2 2, PF13465.5) was the most enriched pfam family observed in the lion brain, found in 791 protein-coding transcripts. C2H2-zinc fingers protein domains represent one of the largest highly conserved gene families of higher eukaryotes (Fedotova et al., 2017). Independent expansions most likely occurred in the ancestral gene family in various lineages, which contributed to adaptive evolution modifying DNA-binding specificity, and thereby providing a mechanism for fast transcriptional evolution (Najafabadi et al., 2015; Emerson and Thomas, 2009). Moreover, mutations of domain-containing C2H2-ZF protein ZNF81 are implicated in non-specific X-linked mental retardation (Kleefstra et al., 2004).

Diverse classes of protein kinase domains (31) were present in the lion assembly, of which four types were among the 50 most enriched, namely; Pkinase, Pkinase Tyr, Kinase-like and Haspin kinase. In the PSP, proteins-containing kinases represent one of the most common and highly conserved protein classes in all species that possess

synaptic structures (Emes et al., 2008). Protein phosphorylation is essential for a wide number of neuronal functions including learning and memory. For example, this family is critical to initiate the influx mechanism of Ca^{2+} through the NMDA receptor and regulate downstream signalling events in both vertebrates and invertebrates (Purcell and Carew, 2003). Likewise, phosphorylation by protein kinases and protein phosphatases (also enriched in the lion brain) is suggested to mediate LTP and LTD expression. Identified protein kinases and phosphatases implicated in synaptic plasticity are; CaMKII, PKA, PKC, MAPK, tyrosine kinases, PP1, PP2A, calcineurin (Lüscher et al., 2000).

RNA and ribosomal binding protein domains were also enriched, including; RRM 1, RRM 7, RRM 5, MMR-HSR1. In the cell, RNAs are interlinked with RNA-binding proteins to make-up ribonucleoprotein complexes. Moreover, RNA-binding proteins mediate the RNAs structure and interactions, playing pivotal roles in their biogenesis, processing (such as splicing, editing and polyadenylation), and cellular localization (Glisovic et al., 2008; Birney et al., 1993). At the brain, the relatively long distances between the cell body and the synapse makes it challenging to achieve a rapid altering of environment of synaptic inputs. RNA binding proteins effectively overcome this by locally mediating protein synthesis and translation in the dendrites, and thereby modifying the synapse directly. This is essential for synaptic plasticity and learning and memory where aberration of these proteins has been implicated in neurodegenerative diseases (Sephton and Yu, 2015).

One of the most common motifs present in nature, the ankyrin repeat-containing domain were found enriched in the brain. They occur in a wide range of functionally diverse proteins involved in numerous cellular functions (Mosavi et al., 2004). For example, at the PSD, the ankyrin repeat-rich membrane spanning protein (ARMS) by forming macromolecular complexes that associate trafficking machinery, glutamate receptors and cytoskeletal mediators, are essential in the branching of dendrites dur-

ing development, and the turnover of spines in the cortex and hippocampus (Wu et al., 2009).

The PSD scaffold proteins such as Shank, are significant in the regulation of substantial receptor and effector complexes. The prototypical Shank proteins are equipped with a large set of protein-protein interaction domains, namely; six ankyirin repeat domains, along with, an SH3, PDZ, proline rich and SAM domains (McWilliams et al., 2004). All of these protein domains were contained in the assembly, particularly PDZ and SH3 were among the most abundant. Through the PDZ domain, Shank binds the C-terminal of PSD-95, conjoining NMDAR/PSD-95 complexes and bonding them to mediators of the actin cytoskeleton (Naisbitt et al., 1999). Significant, the prototype postsynaptic complex MASC, is constituted by several protein-binding domains, mainly PDZ and SH3, potentially allowing the assembly of complexes that regulate synaptic transmission and plasticity (Oliva et al., 2012; Emes et al., 2008).

Forty four different classes of EF-hand domains were also observed. These are found in a large family of calcium-binding proteins, consequently EF-hand domains are critical in the synaptic functioning. Calmodulin (CaM), for example, is capable to bind four calcium ions via its EF-hand domains. CaM is a regulatory protein that controls the activity of, and grants Ca^{2+} sensitivity on, numerous pivotal signalling molecules which are essential for plasticity. The significance of this protein in the brain is mirrored by its high concentrations (10 to 100 μM) (Xia and Storm, 2005).

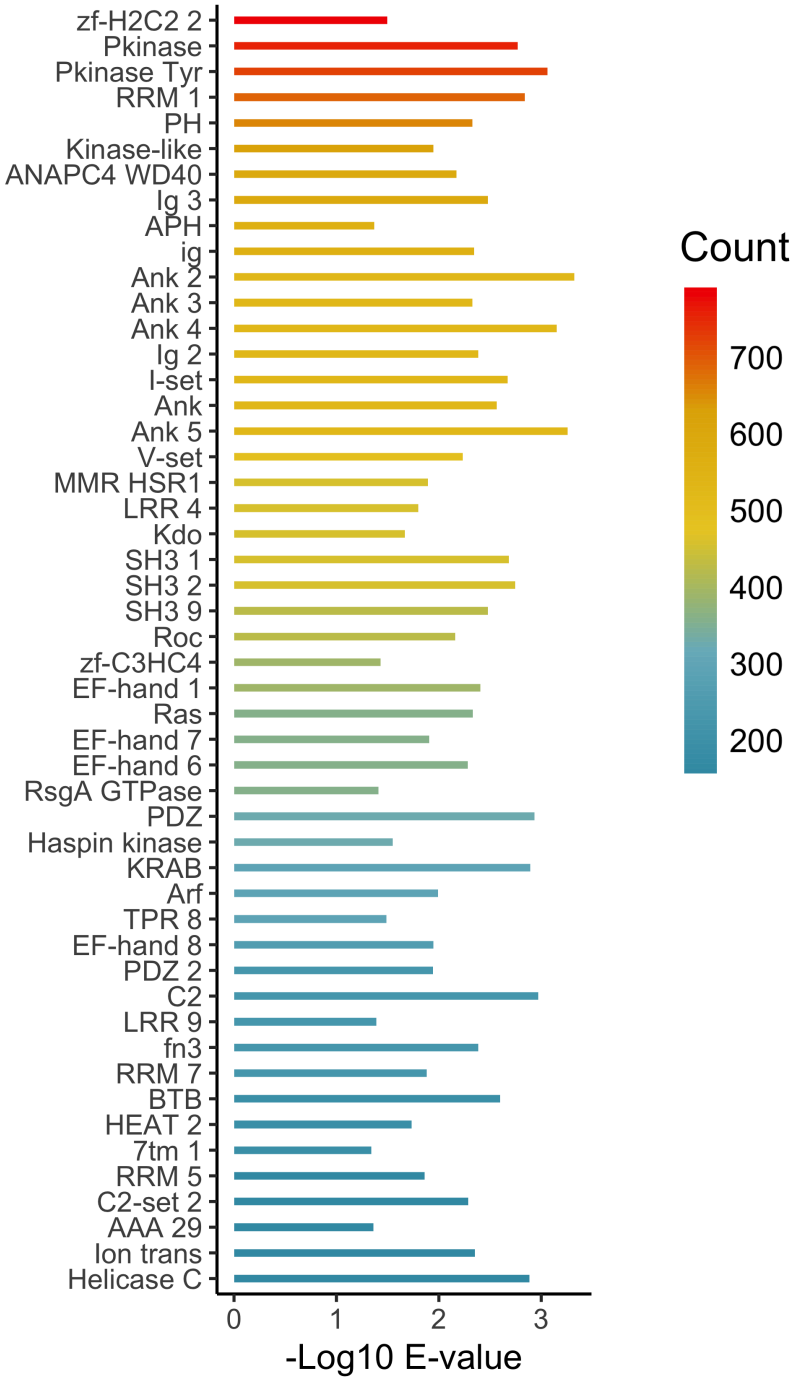


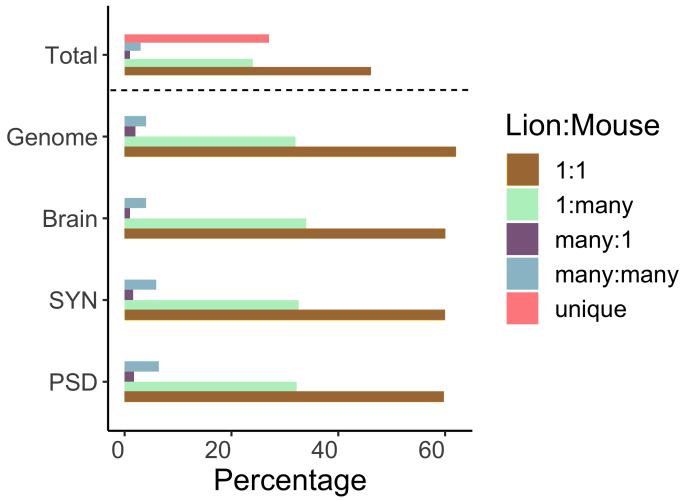
Figure 6.6: **The most enriched protein domains in the lion brain.** Annotated transcripts containing an ORF (21,236) were investigated for protein domains using Pfam and HMMER. The y-axis shows the top 50 enriched domains found (e-value 0.005). Colours indicate the number of transcripts found in each domain; red comprises the highest and blue the lowest.

6.3 Transcriptome orthology

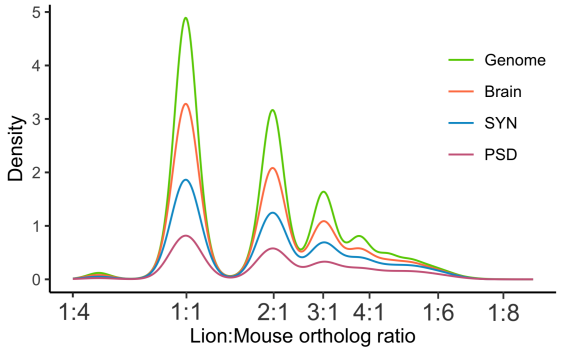
As a further attempt to evaluate the *de novo* lion assembly, it was compared to available *Felidae* assemblies (cat, leopard, tiger and cheetah). Resources used to obtain this data were ensembl (cat and leopard) and the NCBI genome databases (tiger and cheetah). Inparanoid was employed to obtain orthologs in the feline species. A comparable proportion of orthologs were obtained for the leopard, tiger and cheetah (73%, 72.3% and 72.4%, respectively). A higher number of orthologs was obtained with the cat assembly (92.7%), which is most likely due to the higher quality of the cat transcriptome. 13,817 *de novo* contigs or 9,238 ortholog gene clusters were identified to be shared in all 5 cat species.

Orthology was likewise determined with the mouse proteome. Orthologs expressed in the mouse brain, SYN and PSD from the supplementary data taken from (Sharma et al., 2015) for the former and the G2Cdb for the two later, were identified. From the 21,236 protein-coding transcripts, 17,371 or 81.8% of the protein-coding transcripts had an ortholog with the mouse genome. From this set of ortholog transcripts 69.0% had an ortholog pair expressed in the mouse brain, 17.8% (n=3,078) and 12.5% (n=2,164) matched to genes expressed in the mouse brain, SYN and PSD. 380, 344 and 221 genes were uniquely expressed in the mouse brain, SYN and PSD respectively.

Orthology types were determined, i.e., lion:mouse; 1:1, 1:many, many:1, many:many and unique to lion. In the light of all group of orthologs that were determined and illustrated in Figure 6.7 ("total" refers to lion's unique transcripts, and mouse orthologs in the genome, brain, SYN and PSD), their orthology types were generally constant.



(a)



(b)

Figure 6.7: **Orthology distribution between lion transcripts and mouse genes.** Inparanoid estimated 17,371 lion transcripts matching to a mouse ortholog (81.8% protein-coding transcripts). Lion orthologs with a mouse ortholog gene expressed in the brain, SYN and PSD determined. a) Lion:Mouse ratio of orthologs are represented as 1:1, many:1, many:many, and unique to lion in genes expressed in the genome, brain, SYN and PSD. b) For each pair of ortholog it was estimated the density of its distribution.

The largest orthology type found in the assembly was the 1:1 (Lion:Mouse). Despite the fact that the complete genome of *Panthera leo* remains unsequenced, it is plausible that its size is slightly smaller than that of the mouse. This is considering that the latest mouse assembly GRCm38.p6 (ensembl.org) includes 22,619 and 15,795 coding and non-coding genes, whereas the lion closest sequenced species, the leopard assembly PanPar1.0 (ensembl.org) (Bagatharia et al., 2013) comprises 19,688 genes and 3,900 non-coding genes. Likewise, the cat latest assembly Felis catus 9.0, INSDC (ensembl.org) is slightly smaller that the mouse, it contains 19,446 genes and 6,557 non-

coding genes.

A smaller number of many:1 transcripts were identified in the brain, this increased in the SYN and also in the PSD. This was also mirrored by a substantial proportion of 1:1 orthologs reduced in the PSD. Possibly reflecting lineage-specific gene duplication within the mammal PSD evolution. Additionally, 3,865 transcripts were identified as unique to lion (or lion-specific). From this set of transcripts, 326 were not present in any of the feline assemblies. Further Blastp (E-value 1-10) using the SwissProt database was performed and only 8 transcripts out of the 326 identified a homolog in human, rat or zebrafish. To substantiate if these 318 transcripts are novel, experimental validation is required but is out of the scope of the present study.

GO terms of the unique to lion set were determined using homologous cat genes (n=2,575) and the Biomart tool (Durinck et al., 2009) (Figure 6.8). Particularly enriched terms were collagen type V trimer (GO:0005588) and cytoskeletal anchoring (GO:0090286); functions linked to motor activity. Since lions eat large quantities of protein and very small amounts of carbohydrate, it has been proposed that carnivores do not have the requirement to deplete surplus glucose from their blood stream. Response to carbohydrate (GO:0009743) was found significantly enriched (Wang et al., 2013), which in turn might be also correlated with lipid kinase activity (GO:0001727). Terms involved in negative regulation of acute inflammatory response (GO:0002674) and tolerance induction (GO:0002507) were observed enriched, which relates to a study among eight carnivore species, which common GO terms were linked to carbohydrates and immune responses (Kim et al., 2016). All-trans retinal binding (GO:0005503) has been found as an adaptation for night vision (Nagata et al., 2018). Expressed genes associated with this annotation included; cytochrome P450 27C1(CYP27C1) and alcohol dehydrogenase 4 (ADH4).

Lastly, terms associated with neuronal activities were identified, such as compact myelin

(GO:0043218), sodium:potassium and sodium:potassium:chloride symporter activity (GO:0006814 and GO:0008511) and axon midline choice point recognition (GO:0016199).

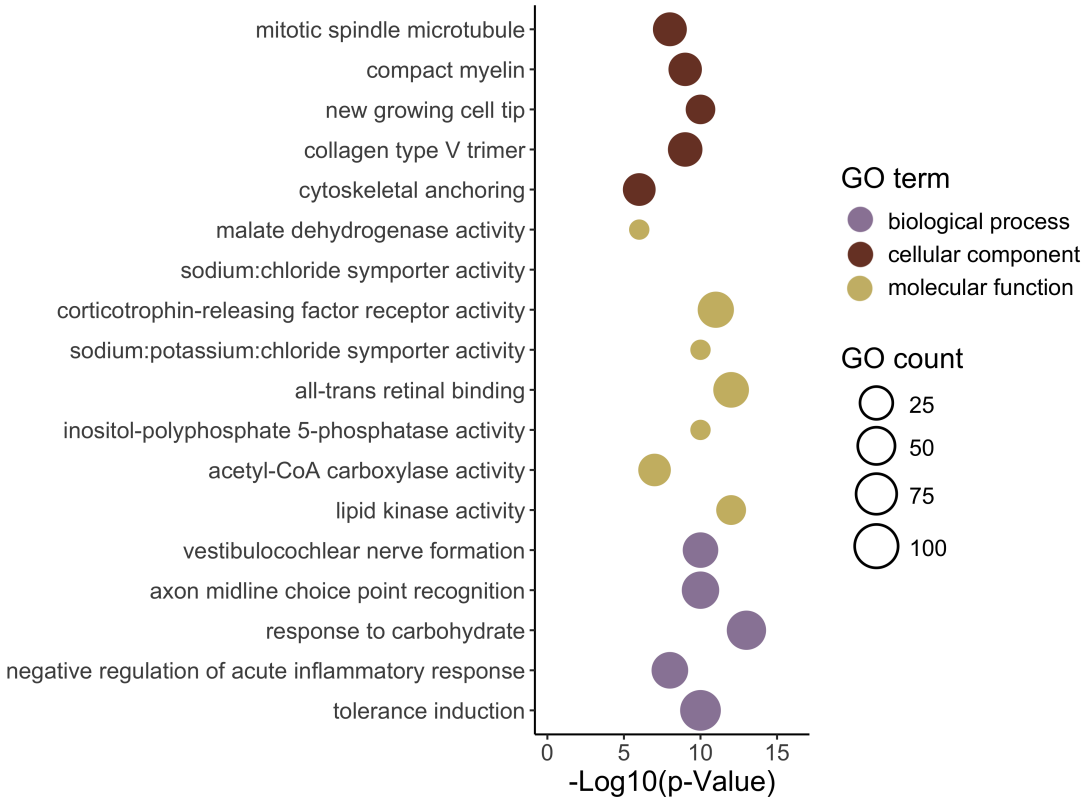


Figure 6.8: **GO analysis of transcripts unique to lion.** Lion transcripts that did not have an identifiable mouse homolog were annotated for GO terms using R analysis tools. P-values were estimated based on hypergeometric distribution and adjusted FDR, applying a 0.05 cut-off. Highly significant enriched terms were selected and plotted against their negative log10 P-value.

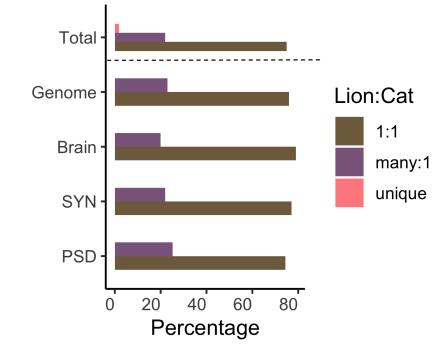
Nonsynonymous and synonymous substitution rates (dN/dS) of the unique to lion transcripts were compared to homologous *Felidae* proteins. Evidence of positive selection (dN/dS>1) were revealed in 6 protein coding genes (Table 6.3). A total of 35 GO terms were associated to these genes, yet the cellular component GO term, mitochondrion (GO:0005739) was the only significantly enriched (p-value of 0.18^{-3}). Genes annotated to this term are; *MRPL52, TRMT61B* and *MAVS*. Suggesting an adaptive metabolic functioning in the *Felidae* family. In this matter, carnivores have higher basal metabolic rates in comparison with herbivore species (Muñoz-Garcia and Williams, 2005). The mitochondria not only modulates metabolism but also plays a role in an-

tiviral innate immunity in vertebrates. This process relies on the triggering of receptors, signal transduction pathways and the involvement of MAVS (mitochondrial antiviral signaling protein) (Koshiba et al., 2011). A study of the innate immune response conducted in various carnivore species, revealed a largely higher immunocompetence in felines (including cheetah, leopard and lion) than other species formerly reported (Heinrich et al., 2016).

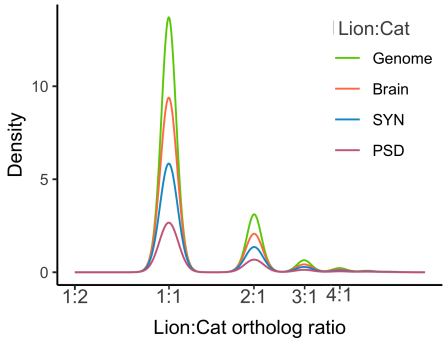
Table 6.3: **Positive selection genes among unique Felidae.**

Gene	Description	dN/dS (omega)
MRPL52	Mitochondrial ribosomal protein L52	3.00
C19orf44	Chromosome 19 open reading frame 44	2.85
UPF3A	Regulator of nonsense mediated mRNA decay	1.38
TRMT61B	TRNA methyltransferase 61B	1.38
MAVS	Mitochondrial antiviral signaling protein	1.03
AP5S1	Adaptor related protein complex 5 subunit sigma 1	1.02
MOGAT3	Monoacylglycerol O-acyltransferase 3	1.01

Additionally, the lion assembly was compared to the cat genome (Genome assembly: Felis catus 9.0) in the same manner that was compared to the mouse (Figure 6.9). 326 lion transcripts were not found in the cat genome (1.8% of the assembly). The vast majority of orthologs (75% total) belong to the class 1:1 (lion:cat), and only a small amount were determined as many:1 (22% total). The later is possibly associated to the process of *de novo* transcriptome assembly, which frequently multiple generated transcripts are sub-sequences of an underlying true transcript.



(a)



(b)

Figure 6.9: **Orthology distribution between lion and cat.** Inparanoid estimated 16,103 lion transcripts matching to a mouse ortholog (92.7% protein-coding transcripts). Lion orthologs with a mouse ortholog gene expressed in the brain, SYN and PSD determined. a) Lion:Cat ratio of orthologs are represented as 1:1, many:1 and unique to lion in genes expressed in the genome, brain, SYN and PSD. b) For each pair of ortholog it was estimated the density of its distribution.

6.4 Differential expression of forebrain and brainstem lion brain tissues

To investigate signatures that determine differences in complexity and functioning between the forebrain and brainstem, a differential expression analysis was carried out using the Bioconductor software package EdgeR (Robinson et al., 2010), which uses an empirical model to identify differentially expressed transcripts. Utilizing a stringent adjusted p-value cutoff of <0.01 and log2 fold change of ≥2, 1,415 differentially expressed transcripts were determined for the forbrain vs. brainstem comparison. 834 of these transcripts were up-regulated in the forebrain, whereas 581 were up-regulated

in the brainstem (Figure 6.10). The top 20 up-regulated transcripts for each group are listed in Table (6.4).

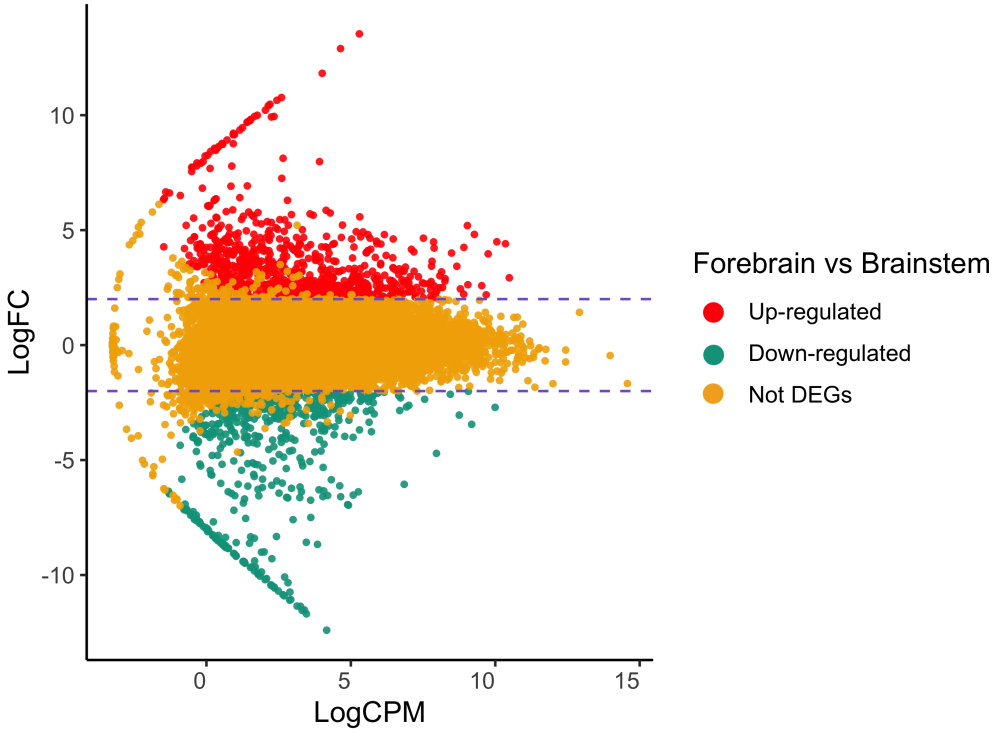


Figure 6.10: **MA plot of differentially expressed transcripts for forebrain and brainstem lion tissues.** From the 21,236 annotated lion transcripts, 1,415 were identified as significantly differentially expressed for forebrain vs. brainstem tissues (q-value <0.01 and ≥2 log2FC). Up-regulated forebrain transcripts have positive MA values and are depicted as red (n=834), whereas brainstem up-regulated transcripts (n=580) have negative MA values and are shown in green.

Table 6.4: **Top 20 differentially expressed genes in the lion forebrain and brainstem.**

Gene	Description	logFC	adj P-value	Ortho-logy	PSD
	Forebrain up-regulated				
GRHPR	Glyoxylate and hydroxypyruvate reductase	13.54	2.38e^{-67}	many:1	no
LPCAT4	Lysophosphatidylcholine acyltransferase 4	12.90	1.93e^{-41}	many:1	no
ABLIM3	Actin binding LIM protein family 3	11.82	4.36e^{-38}	many:1	yes
ADD2	Adducin	10.77	1.48e^{-17}	many:1	yes
SLC25A28	Solute carrier family 25, member 28	10.65	4.41e^{-05}	many:1	no
TPCN2	Two pore segment channel 2	10.47	5.12e^{-05}	many:1	no
DOPEY1	Dopey family member 1	10.40	2.52e^{-12}	many:1	no
GRIA1	Glutamate receptor, ionotropic, AMPA1	10.22	5.50e^{-05}	1:1	yes
MIEF1	Mitochondrial elongation factor 1	10.00	8.74e^{-13}	1:1	no
	Brainstem up-regulated				
ABCA8B	ATP-binding cassette, sub-family A, member 8b	12.40	6.70e^{-37}	many:1	no
SLC38A2	Solute carrier family 38, member 2	11.54	3.26e^{-27}	1:1	no
AP3M1	Adaptor-related protein complex 3, mu 1 subunit	11.52	9.35e^{-19}	1:1	yes
ZDHHC1	Zinc finger, DHHC domain containing 1	11.36	6.00e^{-19}	many:1	no
ADAM22	Disintegrin and metallopeptidase domain 22	11.35	4.94e^{-05}	many:1	no
UBE3C	Ubiquitin protein ligase E3C	11.07	1.74e^{-18}	1:1	no
EFHD1	EF hand domain containing 1	10.90	9.73e^{-18}	1:1	no
PGAP2	Post-GPI attachment to proteins 2	10.85	7.52e^{-17}	many:1	no
SPG7	PG7, paraplegin matrix AAA peptidase subunit	10.70	1.13e^{-05}	many:1	no

GO terms were examined in all differentially expressed forebrain vs. brainstem genes. 32 significantly enriched terms were observed in the forebrain up-regulated (and brainstem down-regulated), whilst 26 were seen in the up-regulated brainstem (forebrain down-regulated). In overall, a larger number of terms, which are key for the synaptic functioning and plasticity were observed in the up-regulated forebrain (Figure 6.11), namely; postsynaptic density (GO:0014069), regulation of synaptic plasticity (GO:004 8167) and long term memory (GO:0007616).

Learning and memory are highly sophisticated mechanisms, which permit the information that is frequently captured by the brain to be processed and stored by interconnected neuronal networks. These in turn rely on a highly accurate systematization between signalling cascades of axons and dendrites. Essentially, protein kinases and phosphatases modulate all phases of learning and memory (Mansuy, 2003). Amongst them, calcineurin (GO:0005955), a calcium-dependent protein phosphatase, is the most Ca^{2+}-sentitive, and the only Ca^{2+} active protein phosphatase localized in the brain (Klee et al., 1979). A study based on knockout mice have targeted forebrain-specific calcineurin in Schaffer collateral-CA1 synapses of the hippocampus. This showed a largely reduced LTD with a considerable variation in the LTD/LTP modification threshold, accompanied with a defective episodic-like memory and hippocampus-dependent learning (Zeng et al., 2001).

Calcium- and calmodulin-dependent protein kinase complex (GO:0005954), voltage-gated potassium channels (GO:0005267), and most of all, various voltage-gated calcium channels (GO:0005245, GO:0005891, GO:0005244, GO:0008331, GO:0086056) were identified in the up-regulated forebrain. In this regard, Ca^{2+} influx plays a major part in the modulation of synaptic transmittion, such as the induction and recovery from every kind of short and long term synaptic plasticity (Zucker, 1999).

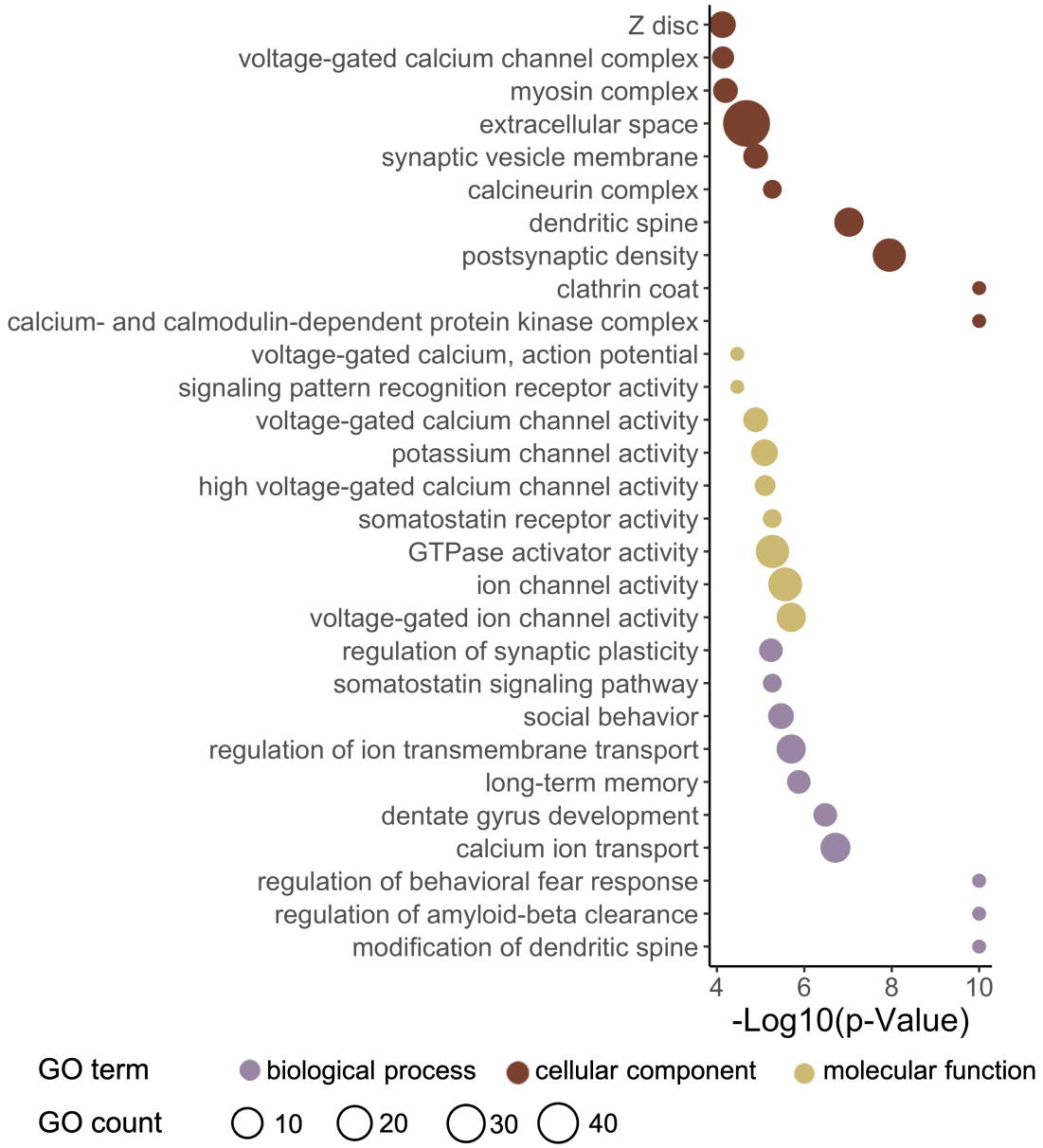


Figure 6.11: **Forebrain up-regulated GO analysis.** Distribution of GO terms reflecting statistical significant differences (hypergeometric distribution applying a p-value and FDR threshold of 0.5) based on 833 up-regulated transcripts in the lion forebrain vs. brainstem (q-value <0.01 and ≥2 log2FC). Different colours represent each GO domain, while sizes reflects the number of genes involved in the term.

Additionally, terms associated with the forebrain structures such as; regulation of fear response (GO:2000822) (Fanselow, 1994) and social behaviour (GO:0035176) (Goodson, 2005) were significantly enriched. The former is likely linked to somatostain receptor activity (GO:0004994) and somatostatin signaling pathway (GO:0038170), as the

stress response is mediated by five somatostatin receptors that are widely distributed in the brain (Stengel et al., 2015). Myosin complex (GO:0016459) GO term, which is typically associated with muscle function, was also significantly enriched in the forebrain up-regulated transcripts.

The forebrain is the largest region of the brain and where cognition is processed. Therefore, it is not surprising to observe several enriched terms essential to the synaptic functioning. In contrast, the brainstem is the portion of the brain that links the spinal cord and where 10 of the 12 cranial nerves originate. It embodies the cerebellum, midbrain, pons and medulla oblongata of the hindbrain. It is largely responsible for the autonomic nervous system, which controls breathing, heart rate, digestion, urination, pupillary response, along with motor and sensory innervation (McCorry, 2007; Gabella, 2001).

Significantly enriched GO terms for the up-regulated genes in the lion brainstem (Figure 6.12) included various involved in sensory responses, such as photoreceptor connecting cilium (GO:0032391), glossopharyngeal nerve morphogenesis (GO:0021615). Structure and motor proteins, including; axon (GO:0030424), microtubule (GO:0005874) and microtubule motor activity (GO:0003777), kinesin complex (GO:0005871). Kinesins are molecular motors that transport cargoes (vesicles, organelles and chromosomes) through microtubules. In the developing brainstem, mutations in protein coding genes of the kinesin family (KIF21A) have been shown to cause disruptions in the connectivity of ocular motoneurons, along with defects in axon growth. This in turn leads to various complex oculomotility syndromes or strabismus, known as congenital cranial dysinnervation disorders (CCDDs) (Engle, 2007, 2006).

GO terms linking cerebellar Purkinje cells were also found enriched, such as; cerebellar Purkinje cell differentiation (GO:0021702) and cerebellar Purkinje cell-granule cell precursor cell signalling (GO:0021937).

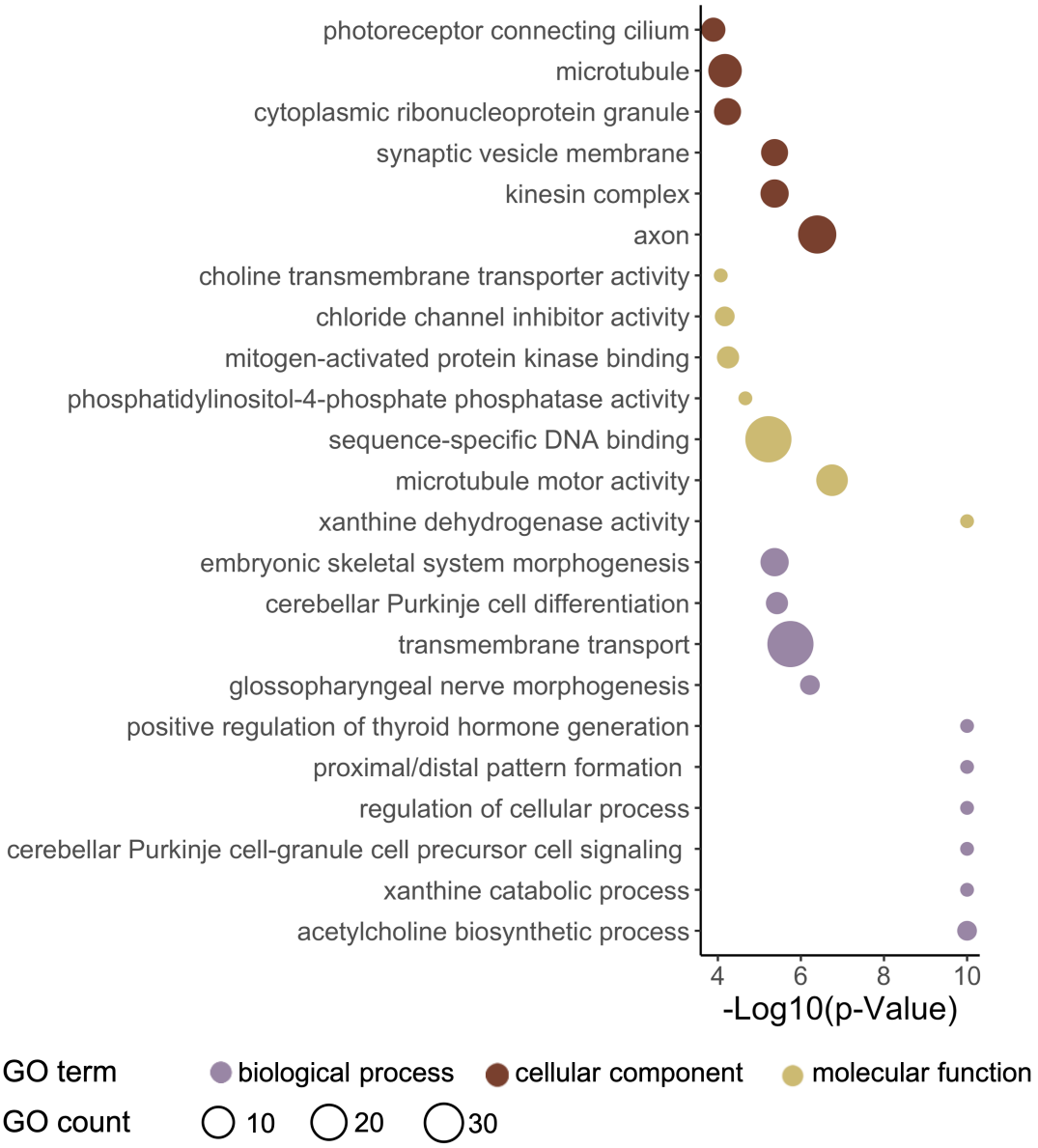


Figure 6.12: **Brainstem up-regulated GO analysis.** Distribution of GO terms reflecting statistical significant differences (hypergeometric distribution applying a p-value and FDR threshold of 0.5) based on 580 up-regulated transcripts in the lion brainstem vs. forebrain (q-value <0.01 and ≥2 log2FC). Different colours represent each GO domain, while sizes reflect the number of genes involved in the term.

These large neurons constitute the only output of the cerebellar cortex that extend to the deep nuclei (DCN), where these arrange GABAergic synapses. Granule cells promote differentiation and migration of Purkinje cells, which occur in the early embryonic stages, before the existence of basket or stellate cell synaptic inputs. Thus, Purkin-

je cells trigger the development of synapses in the cerebellar cortex (Watt et al., 2009). In this respect, positive regulation of thyroid hormone generation (GO:2000609) was up-regulated. Thyroid hormones, particularly T3, are foremost for the CNS development, these modulate neurogenesis, neural migration and differentiation, myelination and synaptogenesis through specific time windows (Bernal, 2007). Lack of thyroid hormone, not only induces severe hypoplasia of Purkinje cells, but also breaks granule cell differentiation. These results in congenital hypothyroidism with neurological evidences (ataxia and altered motor movement) in humans and mice (Heuer and Mason, 2003). Terms associated with developmental and differentiation processes were also observed, such as embryonic skeletal system morphogenesis (GO:0048704), glossopharyngeal nerve morphogenesis (GO:0021615), proximal/distal pattern formation (GO:0009954).

Among other enriched terms found in the up-regulated brainstem included; transmembrane transport (GO:0055085), chloride transmembrane transporter activity (GO:0015108), choline transmembrane transporter (GO:0015220) and acetylcholine bi osynthetic process (GO:0008292). The latter two suggest an enrichment of cholinergic synapses in the brainstem structures. Xanthine dehydrogenase activity (GO:0004854) are the final steps in the purine catabolic pathway, and plays a role in the neuroprotection against hyperammonemia (by reduced ATP levels), which in turn is mediated by the NMDA receptor (Kaminsky and Kosenko, 2009).

Gene ontology analysis comparing the top 20 most differentially expressed transcripts using GO-Slim tools, revealed highly similar proportion of terms (Figure 6.13). Particularly, biological process, cellular component and molecular function were commonly enriched. Yet, cellular component assembly (GO:0022607) and cell-cell signalling (GO:0007267) were uniquely showed in the forebrain up-regulated, while cytoskeleton (GO:0005856) and DNA binding were only observed in the brainstem (GO:0003677).

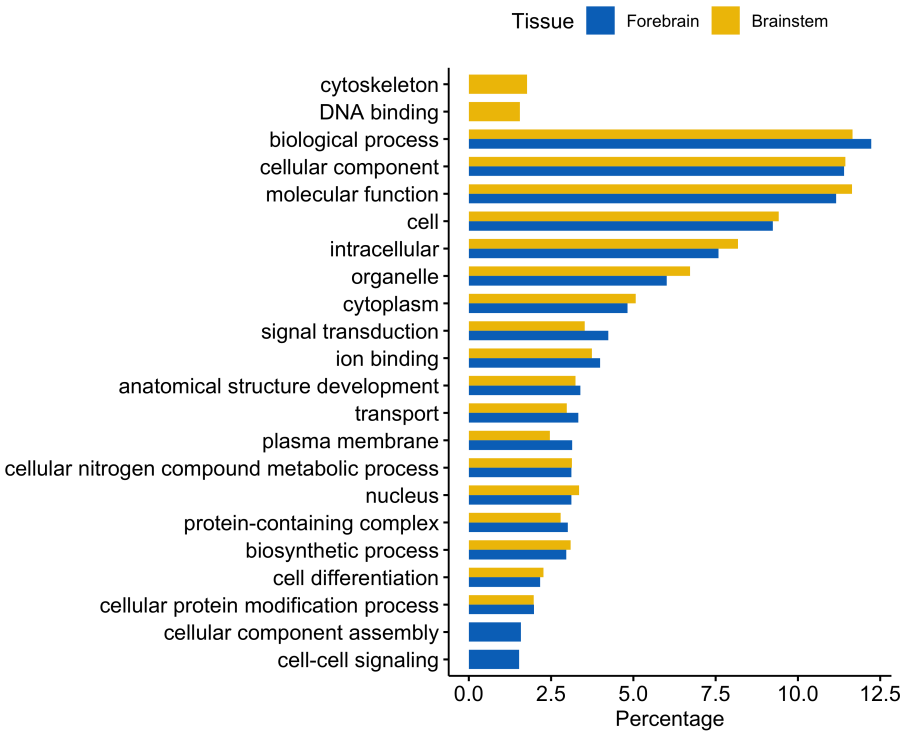


Figure 6.13: **GO-Slim analysis of the top 20 up-regulated transcripts in the lion forebrain and brainstem**. Differentially expressed transcripts were sorted according to their significance, and the top 20 for each brain tissue were analysed for GO-Slim terms.

Expression of the transcript set unique to lion (3,865) (Section:6.3) was examined. Whilst a larger proportion of these are found in the forbrain up-regulated transcripts (n=174) compared to brainstem up-regulated (n=114), this was not significantly different (Wilcoxon test, p-value 0.52) (Figure 6.14). The most enriched GO terms in the lion-specific forebrain up-regulated transcripts included; membrane-bound transcription factors involved in SREBP signaling pathway (GO:0032933), regulation of ventricular cardiac muscle cell action potential (GO:0098911) and embryonic forelimb morphogenesis (GO:0035115). Brainstem up-regulated transcripts were involved in mitogen-activated protein kinase binding (GO:0051019), cardiac septum development (GO:0003279) and eye photoreceptor cell development (GO:0042462).

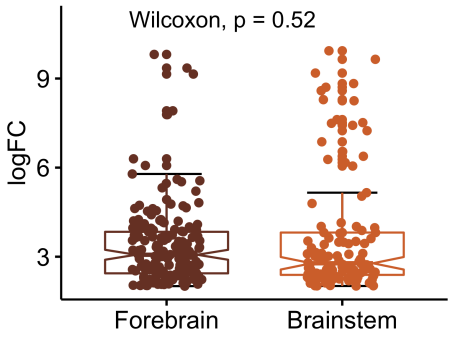


Figure 6.14: **Up-regulated and lion-unique transcripts.** From the 3,865 lion-unique transcripts, 174 were up-regulated in the forebrain and 114 were up-regulated in the brainstem. Transcripts are plotted against their log fold change.

6.4.1 Gene expression of key synaptic genes

Expression of key synaptic genes was examined using the lion annotated transcripts for both forebrain and brainstem tissues (Figure 6.15). On the whole, comparable patterns of expression were seen for both brain tissues, yet the forebrain displayed higher levels of expression of various key genes. Members of the SNARE protein family were found highly enriched in the lion brain tissues. SNAP-25 (Synaptosomal-associated protein of 25kDa), VAMP (vesicle-associated membrane protein, also called synapobrevin), STX (syntaxin) and NSF (N-ethylmaleimide-sensitive fusion protein) are involved in intracellular membrane vesicle docking, fusion and synchronization of neurotransmitter release. While the latter 3 proteins bind the membrane via C-terminal transmembrane domain, SNAP-25 is anchored by palmitoylation (Chen and Scheller, 2001).

SNAP-25 was the foremost enriched gene. This protein-coding gene mediates synaptic vesicle exocytosis over the arrangement of a SNARE complex and the interplay with various classes of voltage-gated calcium channels, impeding their function and thereby, decreasing calcium responsiveness to neuronal depolarization (Braida et al., 2016). Expression of *SNAP-25* has been found extensively dispersed in synapse enriched areas throughout the brain (Yamamori et al., 2011). Low expression levels of SNAP-25 have been identified in patients with schizophrenia or attention-deficit/hyperactivity

disorder.

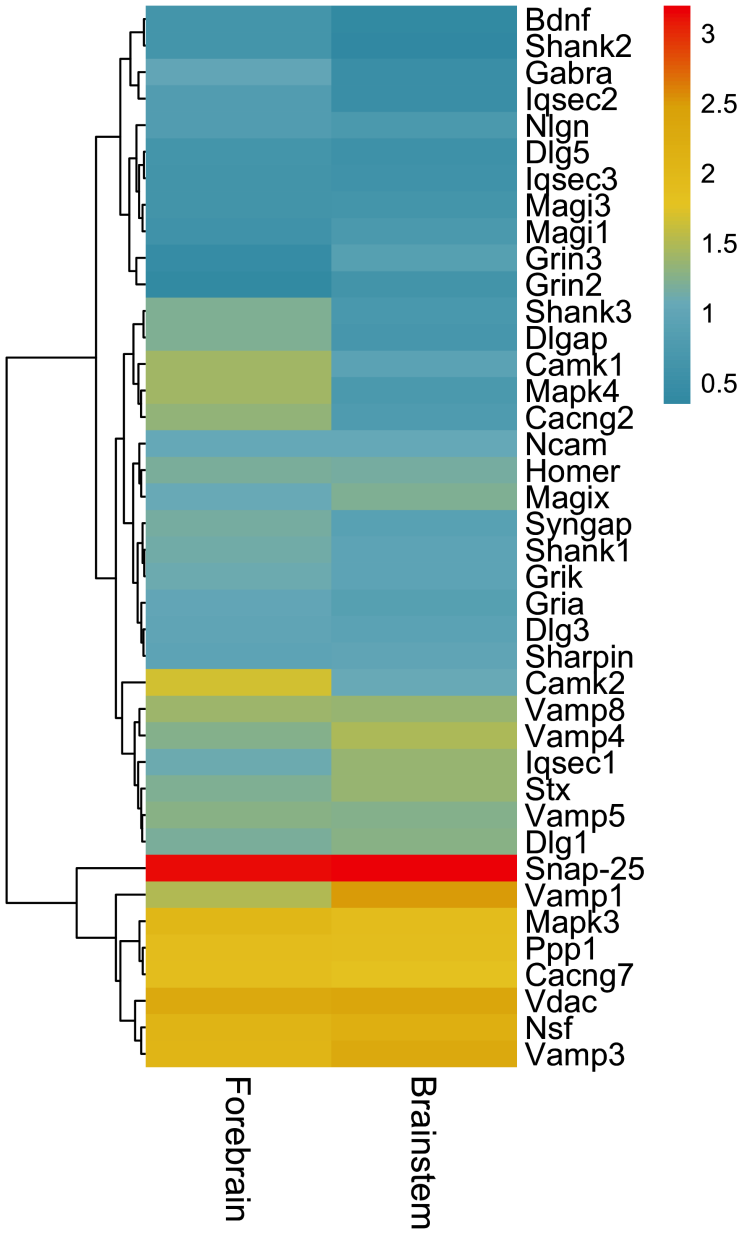


Figure 6.15: **Expression of key synaptic genes.** Genes that are essential for the synaptic functioning were investigated in the lion brain. Tissues and genes were hierarchical clustered. Considering the presence of splice-forms, the mean expression value was obtained per gene and its expression in TPM was log10 transformed. Colours distinguish highly expressed genes (red) from lowly expressed (blue).

ADs patients have shown an expression decrease of *VAMP, SNAP-25* and *STX*, mainly in neocortical regions (Berchtold et al., 2013). The expression of the two former genes,

has been found in the plasma membrane of oesophageal circular smooth muscle cells in felines, which function is suggested to regulate muscle excitability and contractility (Ji et al., 2002).

Other highly enriched synaptic genes that were found in the lion brain, included; *V-DAC, MAPK3, PPP1* and *CACNG7*. Mammalian central neurons possess numerous voltage-dependent anion channels (VDAC), along with calcium voltage-gated channels, such as calcium voltage-gated channel auxiliary subunit gamma 7 (CACNG7). The neuronal electrical mechanism depends on a diverse number of voltage and ligand-gated ion channels that are porous to inorganic ions, including; calcium, sodium, potassium, chloride. Whilst the latter three ions sustain an electrogenic role, calcium ions differ in that they do not just change the potential of the membrane, but further function as a vital signalling unit (Clapham, 2007). The opening of voltage-gated calcium channels causes the influx of calcium and the electrochemical gradient, which activates a number of calcium-dependent processes, namely; neurotransmitter release, neural outgrowth, and the triggering of calcium-dependent enzymes (CaMKII, PKC) (Simms and Zamponi, 2014). Four subunits constitute the channels; the pore that frames the α_1 subunit and the auxiliary $\alpha_2\delta$, β and γ subunits. These are encoded by different genes with alternative splicing variants and are expressed in a tissue specific manner. The activity of *CACNG7* is to decrease the current amplitude, and its expression is distributed in the brain, heart, lung and testis (Arikkath and Campbell, 2003; Moss et al., 2002). *CACNG2* was found with a higher expression in the forebrain (3.3x) than the brainstem. Among the activities of *CACNG2* include; inhibitory effect, kinetics activation/inactivation and AMPAR trafficking (Moss et al., 2002).

Similarly, VDAC proteins shape selective pores in the membrane of neurons yielding various characteristics for intrinsic electrical excitability. In this way, an abundant repertoire of firing behaviours are exhibited in mammalian neurons through an extensive scope of stimuli and firing frequencies (Vacher et al., 2008). A number of neu-

rological diseases, i.e., bipolar disorders, schizophrenia and AD have been linked to a decreased expression of VDAC proteins, while an elevated expression has been observed in patients with Down's syndrome (César Rosa and de Cerqueira César, 2016).

MAPK (Mitogen-activated protein kinase) are a conserved family of Ser/Thr protein kinases that have a role in synapse assembly, shape, function and plasticity. Expression of MAPK is enriched in the adult brain, which drives to the activation of extracellular signal-regulated kinases-1 and -2 (ERK1 and ERK2) via excitatory glutamatergic signalling, thus inferring a role in synaptic plasticity. Inhibition of ERK1 and ERK2 have demonstrated a role learning and memory, i.e., spatial learning and fear conditioning (Thomas and Huganir, 2004). In this study, *MAPK3* was highly enriched in the forebrain and brainstem, however, *MAPK4* was found enriched in the forebrain but not in the brainstem (5x).

While protein kinases, such as MAPK are critical for various cellular processes. Together with their counterpart, phosphatases, they provide equilibrium to the brain activity through a continuous push/pull of regulatory elements, e.g., increase/decrease of synaptic strength, on/off of neuronal firing rates, and excitation/inhibition of neural circuits, where the functioning of proteins is regulated by phosphorylation and dephosphorylation (Woolfrey and Dell'Acqua, 2015). PPP1 (protein phosphatase 1) is greatly enriched in the brain where, it has a significant role for the firing and initial sustainment of NMDAR-dependent LTD (Munton et al., 2004; Morishita et al., 2001). Activity of PPP1 is conjugated with ERK1 and ERK2, which firing is increased by the inhibition of PPP1, thus synaptic plasticity and memory is mediated through this mechanism. Moreover, brain restoration from oxygen/glucose restriction or ischemia when PPP1 is inhibited, has been shown altered, highlighting its role in brain neuroprotective pathways (Hédou et al., 2008).

Ultimately, while contrasting the expression between the two brain regions, a number

of genes that play a role in synaptic plasticity and are vital for the synaptic functioning, such as; *CAMK1, CAMK2, SHANK3, DLGAP* and *MAPK4* displayed higher expression levels in the forebrain than the brainstem. For example, *CAMK2*, was 5X more enriched in the forebrain. This protein is predominantly expressed in the brain, where it makes up 1 to 2% of the entire protein. When Ca^{2+} enters via the NMDAR, CAMK2 is able to detect this increment and triggers a biochemical cascade that strengthens synaptic transmission, i.e., LTP, thus it underlies various forms of leaning and memory. Additionally, one of its key functional features is its capability to autophosphorylate and dephosphorylate, suggesting that CAMK2 may also function as a molecular switch that permits long-term memory storage (Lisman et al., 2002).

6.5 Positive selection on the lion brain

Positive natural selection leads to the fixation of advantageous traits, and it has an important role in the evolution of a species (Sabeti et al., 2006). We sought to detect signatures of positive selection in genes expressed in the lion brain. Using the complete homologous (1:1 orthologs) dataset between the *de novo* lion transcripts and the mouse proteome (n=7,360) to calculate (dN/dS) for each ortholog pair.

Sixty-four genes showed positive selection ($\omega > 1$) between lion and mouse. Evolutionary rates among felines; including the lion, cat, cheetah, leopard and tiger (n=9,238) were also compared. 194 genes were identified as undergoing positive selection. Generally, as seen in Figure 6.16, proteins under positive selection were enriched in enzymatic activity. For instance, palmitoyl hydrolase activity (GO:0008474) which has a pivotal role in controlling protein traffic across synaptic membranes, and might also mediate synaptic plasticity. A considerable number of synaptic proteins are palmitoylated and depalmitoylated, where protein depalmitoylation (GO:0002084) was also within the most enriched GO terms. Example of these synaptic proteins, include; members of the SNARE and synaptotagmins (membrane-trafficking proteins) in the

pre-synaptic terminal, and PSD95, AMPAR and NMDAR in the post-synaptic terminal (Conibear and Davis, 2010).

Other brain adaptations found, included; CCR chemokine receptor binding (GO:0048020), which has been observed to have an important neuroprotective activity (De Haas et al., 2007). Mitochondrion (GO:0005739) was the most enriched term, likey due to the higher mutation rate in mitochondria. This is seen in many studies, for example, mtDNA in indigenous populations around the world have significant variations. These have permitted them to respond distinctly to varying environmental and pathological conditions (Panov et al., 2014). Mitochondria play a principal role in several essential physiological functions, i.e., ATP production, mediation of Ca_{2+}, ROS (reactive oxygen species) signalling, lipid synthesis, and the capacity to activate apoptosis. Thereby, these organelles are associated with the pathogenesis of an extensive number of neurodegenerative diseases, such as AD, PD, HD and amyotrophic lateral sclerosis (Lee et al., 2018; Ly and Verstreken, 2006).

Coupled with above, is fatty acid β-oxidation (FAO) (GO:0006635). FAO is a fundamental metabolic pathway in the mitochondria, whereby fatty acids are broken down by different tissues (liver, heart and skeletal muscle) to generate energy. During a restriction of glucose, FAO is of special importance (Houten and Wanders, 2010). Fatty acids comprised a major source of energy in carnivores, such as the lion, therefore, these species possess various adaptations that suit such diets (Schermerhorn, 2013). Additionally, although debated (e.g., (Schönfeld and Reiser, 2013)), it has been proposed that in conjunction with oxygen-dependent metabolism of glucose, up to 20% of all the brain energy is generated by mitochondrial oxidation of fatty acids (Panov et al., 2014). To this end, $ACAD_9$ a type of Acyl-CoA dehydrogenase that catalyzes the first steps of the mitochondrial FAO, has been found highly expressed in the brain, particularly in the cerebellum (Wanders et al., 2010). Hence, it is conceivable that this process is of particular importance in the lion brain as an adaptation for the lack of glucose in

their diet.

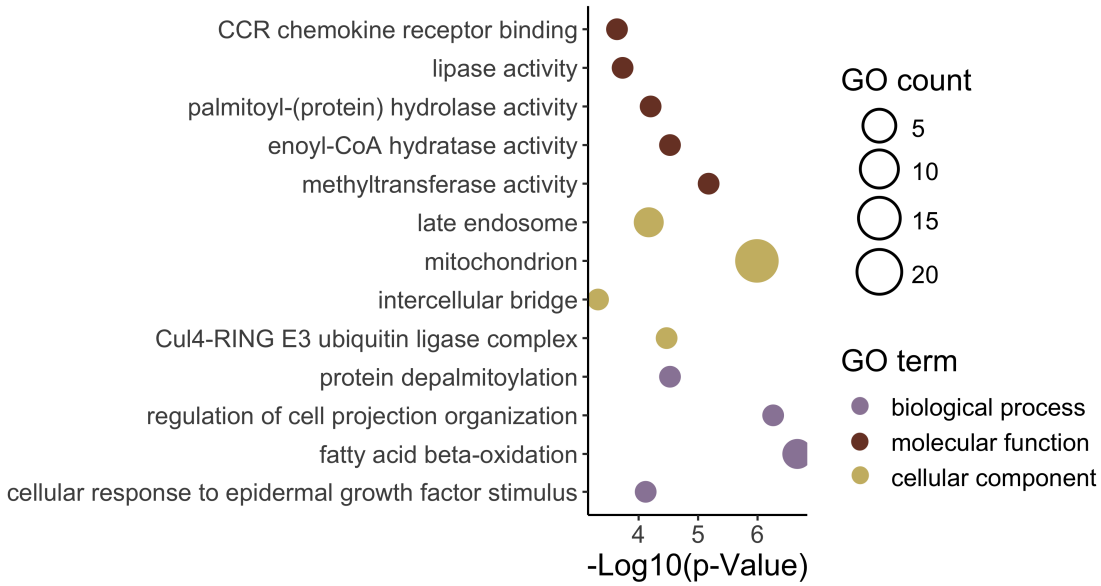


Figure 6.16: **GO analysis for genes under positive selection in the lion brain.** Positive selection in the lion transcripts was tested and GO terms were examined using R tools. P-values were estimated based on hypergeometric distribution and adjusted FDR, applying a threshold of 0.05. Highly significant enriched terms were selected and plotted against their negative log10 P-value.

Evolutionary rates across the up-regulated forebrain and brainstem genes were compared, and no significant difference was found (Wilcoxon test, p-value = 0.053) (Figure 6.17).

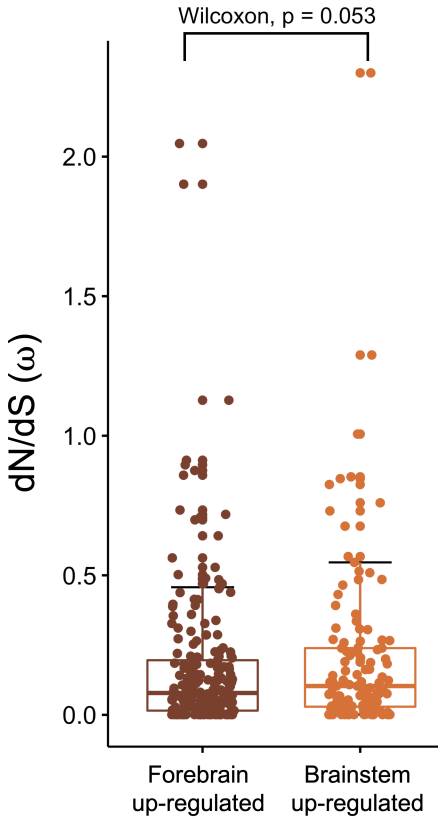


Figure 6.17: **dN/dS comparison between the up-regulated forebrain and brainstem.** dN/dS was examined among the up-regulated transcripts of the two tissues.

A total of 194 protein-coding genes were identified as undergoing positive selection, of which 28 are part of the synaptosome (dN/dS analysis using lion and mouse sequences; 11 SYN proteins, and dN/dS analysis using all felines sequences; 17 SYN proteins) (Figure 6.18). Overall, the expression of these proteins did not differ between the two brain regions, yet a negligible higher expression was observed in the forebrain. A wide number of these proteins were involved in the mitochondria, including; mitochondrial antiviral signaling protein (MAVS), demethyl-Q 7 (COQ7), NADH:ubiquinone oxidoreductase subunit B9 (Ndufb9), succinate dehydrogenase complex (Sdhd), mitochondrial ribosomal protein (Mrpl48), 3-hydroxyisobutyryl-Coenzyme A hydrolase (Hibch), thioesterase superfamily member 4 (Them4), erb-b2 receptor tyrosine kinase 4 (Erbb4) and coiled-coil-helix-coiled-coil-helix domain containing 3

(Chchd3). Thereby, validating the significance of mitochondrial adaptation in the lion brain.

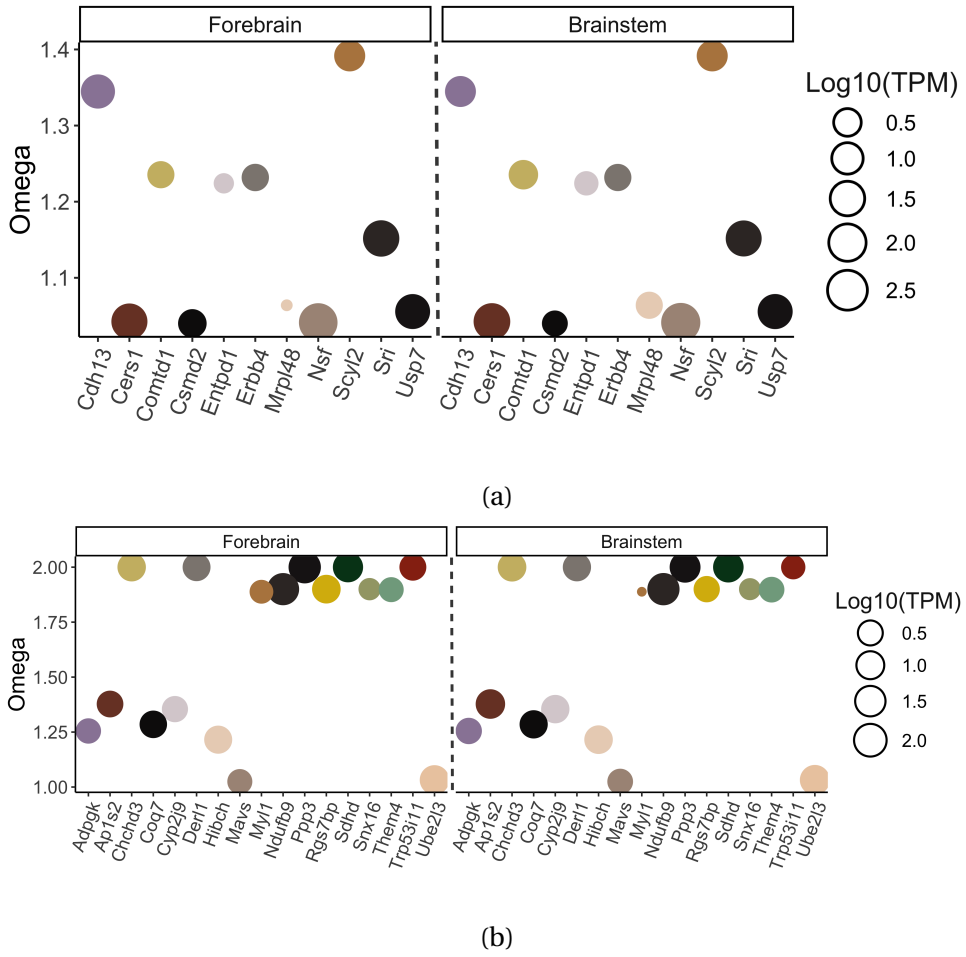


Figure 6.18: **Evolutionary rates of genes expressed in the synapse.** dN/dS (omega) between 1:1 ortholog pairs mouse-lion with a ratio greater than 1 (positive selection) against the lion expression values (TPM) in the 3 tissues. a) Lion and mouse sequences; b) all felines (lion, cat, leopard, cheetah and tiger) sequences

Several synaptic proteins under positive selection were also involved in the mechanism of protein ubiquitination, including; *COQ7, SDHD, CHCHD3, DER1*-like domain family, member 1 (*DERL1*), ubiquitin-conjugating enzyme E2L 3 (*UBE2L3*). In the synapse, protein ubiquitination pathways are crucial, acting upon various phases of the cell differentiation, such as synaptogenesis to synapse elimination, together with

activity-dependent plasticity and synaptic remodeling. Mutation of these genes have been linked to neurodegenerative diseases, such as; AD and PD, as well as a number of neurodevelopmental diseases (Haas and Broadie, 2008; Jason and Ehlers, 2005). A member of the SNARE complex and key synaptic protein, *NSF* was found enriched in both tissues and with signatures of positive selection. Suggesting the accumulation of fitness benefits for the rapid neurotransmitter release mechanism in the whole brain.

6.6 Summary and comments

Utilizing NGS, this research provides the first exploration of expression of lion synapse and PSD proteins. Here, brain tissues were dissected from an adult female *Panthera leo* (See chapter 3), accomplished an extraction of RNA and its subsequent sequencing. Tissues included in this study were two replicates of forebrain and brainstem. A *de novo* sequencing assembly pipeline was conducted, and the *de novo* transcriptome quality was evaluated using different bioinformatics tools, displaying satisfactory results.

21,236 likely protein-coding genes were identified of which 17,371 had a mouse ortholog. The major orthology type was the 1:1, followed by the 1:m. Additionally, orthologs with feline species (cat, cheetah, leopard and tiger) were determined, and 9,238 common orthologs in the five cat species were found. From these ortholgs, 2,575 were identified as Felidae-specific. Annotations from these genes reflected traits that benefit the consumption of large quantities of protein and lipids with a minor carbohydrate intake. Furthermore, we identified 318 potential novel transcripts.

1,415 differentially expressed transcripts were identified for the two replicates of forebrain and brainstem; 834 up-regulated transcripts for the former, and 580 for the latter. Forebrain up-regulated transcripts were associated with key synaptic function (PSD, synaptic plasticity, neurotrasmission). Brainstem up-regulated transcripts were char-

acterized by sensory response attributes, Purkinje cells qualities and morophogenesis.

Several key synaptic genes were expressed in the lion brain, members of the SNARE protein family complex and voltage-gated were particularly enriched. A number of transcripts that are pivotal for the syanaptic functioning and plasticity were enriched in the forebrain, compared to the brainstem. Therefore, it is emphasized differences in cognition and neural basis among forebrain and brainstem.

Signatures of positive selection (dN/dS >1) were tested in two groups; the first comprised lion and mouse, while the second group contained five feline species (i.e., lion, cat, leopard, tiger and cheetah). 64 genes presented evidence of positive selection in the first group, whereas in the second group 194 genes were identified. Several proteins encoded by genes under positive selection detected, were involved in mitochondrial pathways.

Chapter 7

Comparative study of the assembled species

“Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world"

- Albert Einstein -

This research focused on the exploration of neural transcriptomes in taxonomic diverse vertebrate species and comparison with the *Mus musculus* genome using the genes to cognition (Croning et al., 2008) and (Sharma et al., 2015) datasets. A comparative analysis using human data was out of the scope in this thesis, mice has turned into the choice mammalian proxy for genetic research, due to its wide feasibility to carry out genetic manipulation techniques that cannot be possible using human samples (Ellenbroek and Youn, 2016). Moreover, next-generation sequencing efforts have shown low sequence variation and major conservation across a set of human and mice essential genes, further supporting the use of mice as a valuable resource that enables the interpretation of sequence data in human disease research (Georgi et al., 2013). Comparison of synaptic proteins has been carried out between human and mouse with the objective ascertain the appropriateness of mice as prototype of human brain study and disease (Bayés et al., 2012). Such study identified a widely similar expression profile with a large proportion of orthologues (≥ 70%) in the PSD, along with an enrichment of genes associated to Huntington's and Parkinson's disease.

The availability of the novel transcriptomes described in this thesis, will allow a better comparison of PSD proteins from different species and will provide an improved understanding of the evolution of the synapse, together with the molecular networks by which genes function. With the aim to provide a groundwork for subsequent synaptic comparative studies, as a summary of previous work, in this chapter a comparative transcriptomic analysis of mouse PSD orthologs found in the *de novo* assembled species (from the previous chapters), i.e., zebrafish, bat and lion, along with an assessment of PSD gene expression across different brain tissues is conducted.

7.1 Shared PSD proteins

Genes encoding the mouse proteomes (genome, brain, SYN and PSD) were compared to those identified in the former chapters (Figure 7.1). A larger number of mouse orthologs were observed in the zebrafish for all proteomes. This may reflect the greater number of protein-coding genes expressed in this species, which is a result of the TSGD that occurred about 300 Mya. Hence, a larger number of potential protein-coding genes were obtained from the *de novo* assembly and annotation of the zebrafish (28,754), in comparison with the bat and lion assembly (18,747 and 21,236, respectively). Considering that the bat genome is generally smaller than other mammals (Gregory, 2002), it is therefore plausible the reason that the *de novo* assembly of the bat was also the smallest. The number of mouse orthologs found in each specie agreed with the observed proportion of assemblies (23,450, 15,518 and 17,371, for zebrafish, bat and lion, respectively).

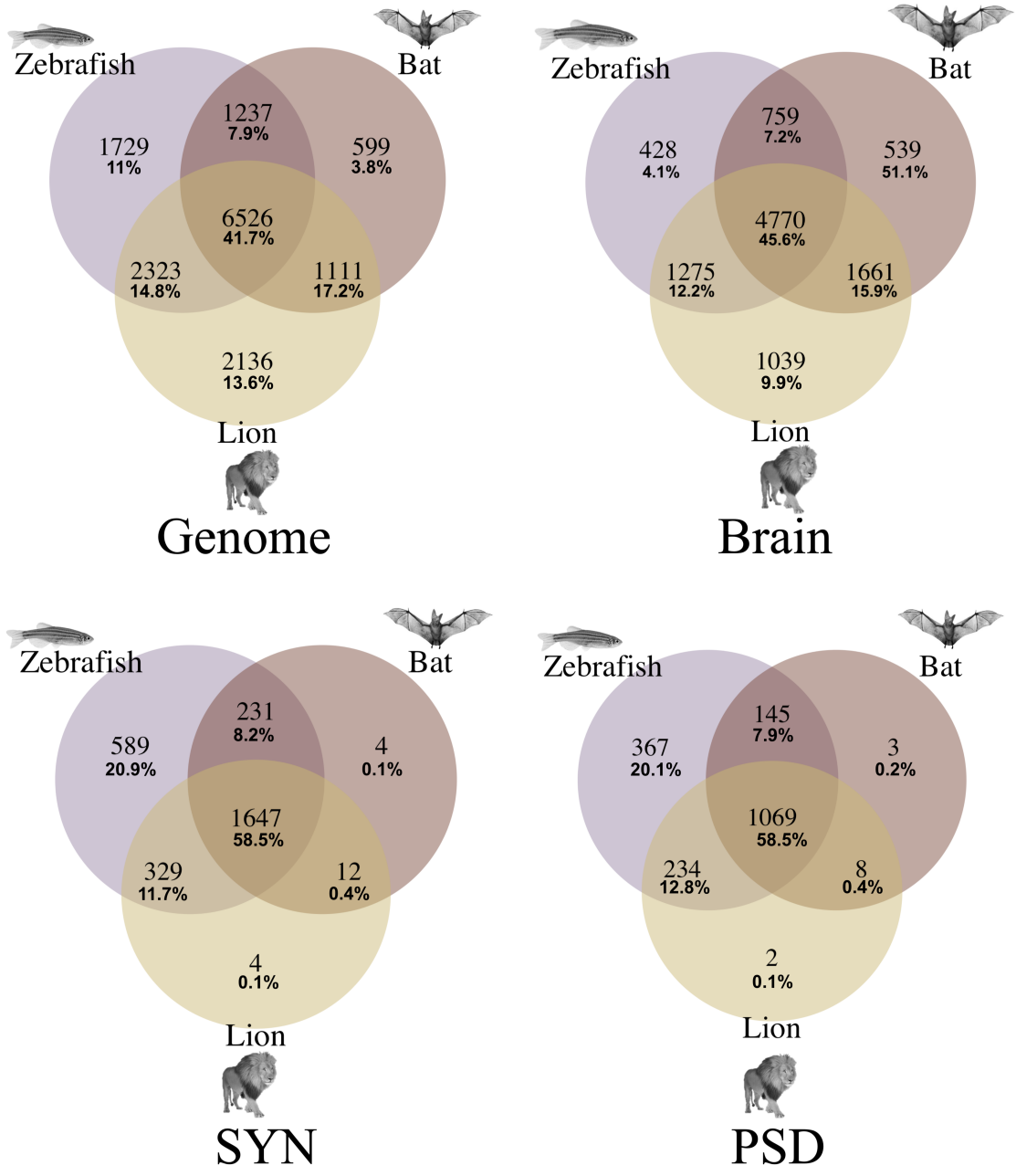


Figure 7.1: **Venn diagram depicting orthologs in 3 species.** Comparison of orthologs identified in the mouse genome, brain, SYN and PSD in the in the 3 *de novo* assembled species; zebrafish (purple), bat (brown) and lion (yellow).

A greater amount of shared orthologs were found in the SYN and PSD in comparison with the brain and genome, indicating an increased protein conservation in these proteomes. A high degree of sequence conservation has been reported in proteins ex-

pressed in the synapse, and particularly in the vertebrate PSD. These proteins have shown high levels of molecular complexity, but also numerous neurological disorders have been associated (Bayés et al., 2011). Very few proteins were found uniquely expressed in the bat and lion PSD (3 and 2, respectively).

1,069 PSD proteins were found expressed in all species. GO terms were analysed using Biomart and R tools (Figure 7.2). Most of the significantly enriched ontologies play a major role in the synaptic structure, neurotransmission and protein complexes. Yet, the foremost enriched terms included; ATP binding (GO:0005524), nucleotide binding (GO:0000166) and protein binding (GO:0005515), emphasizing the paramount synaptic ability, i.e., signal communication. The binding of a multiplex number of proteins and other molecules shapes sophisticated complexes that are capable of receiving and communicating signals from the surroundings, triggering a cascade of intracellular actions and pathways (Emes and Grant, 2012).

Given that the primary goal of the synaptic proteome is the rapid transfer of molecular signals to the postsynaptic terminal, which also underlies a substantial repertoire of behaviours and cognitive processes (Bayés et al., 2012), a large amount of Panther pathways were associated with the shared PSD protein set (n=96). Amongst the most enriched were those involved in signalling pathways, various of which play a main role in neurotrasmittion and plasticity (e.g., metabotropic and ionotropic glutamate receptors, β-adrenergic receptors).

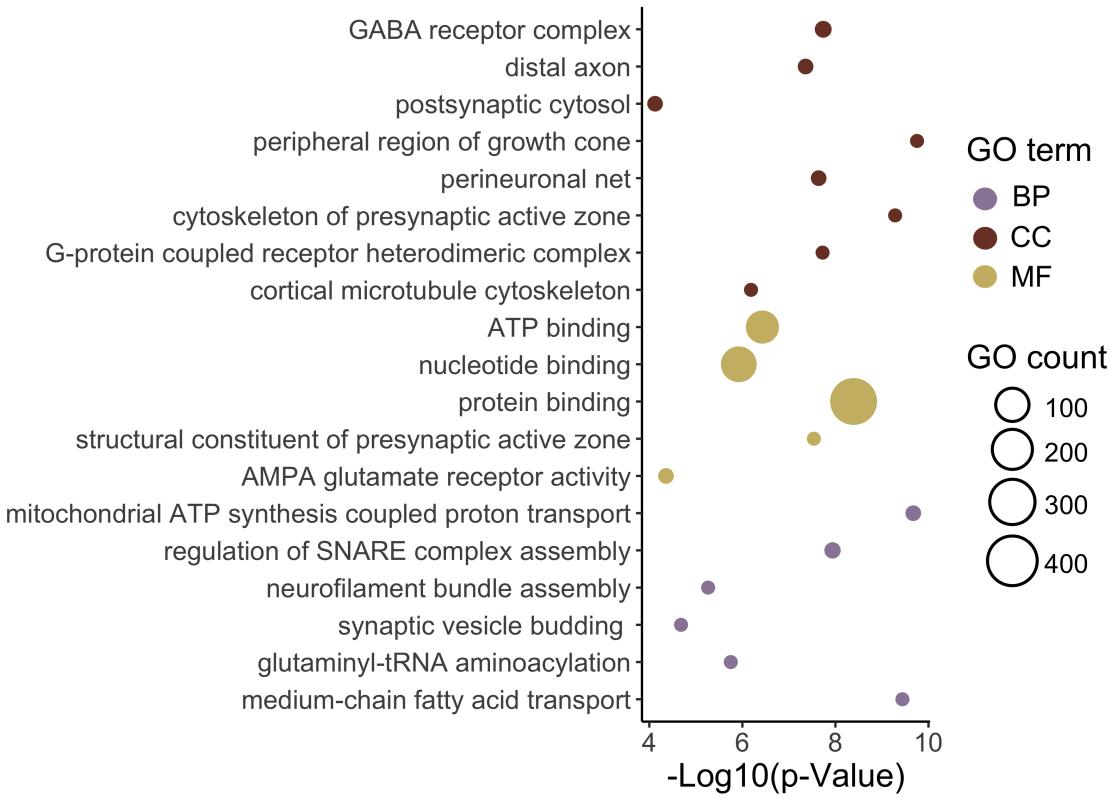


Figure 7.2: **GO analysis for shared PSD proteins** 1,069 PSD proteins were identified to be shared among the 3 *de novo* species; zebrafish, bat and lion. GO terms were explored using R tools. P-values were estimated based on hypergeometric distribution and adjusted FDR, applying a threshold of 0.05. Highly significant enriched terms were selected and plotted against their negative log10 P-value.

Further highly enriched pathways included; EGF (epidermal growth factor) receptor signalling, which are contained in cell-adhesion neuregulin proteins. This complex not only promotes normal development of the nervous system, but also directly interacts with neurotransmitter receptors and neurotransmitter-gated ion channels (Neddens and Buonanno, 2011). For instance, it has been reported the interaction of EGF receptor signalling complex with MAGUKs proteins, such as PSD-95 (DLG4), DLG2 and DLG3 through PDZ domains, thereby suggesting a role in synaptic plasticity (Garcia et al., 2000).

Notably was the direct link of important brain diseases with the most enriched pathways in the PSD. For example, the Wnt signalling pathway, is central for multiple synap-

togenesis processes, such as; cell proliferation, fate differentiation and migration, axon pathfinding, dendrite growth and synapse formation. Hence, its alteration promotes the development of a number of neural diseases, including; schizophrenia, autism and bipolar disorder (Okerlund and Cheyette, 2011; De Ferrari and Moon, 2006). Moreover, the inflammation-mediated signalling pathway is directly involved in the orchestration of inflammatory responses, consequence of brain disease. As a result, molecules involved, such as cytokines and chemokines have been associated to CNS disorders, including; HIV-associated dementia, AD and multiple sclerosis, and therefore targeted as therapeutic prospects (Banisadr et al., 2005; Tran and Miller, 2003).

7.2 Top 20 enriched PSD homologs

In order to perform an accurate comparison across the three different species, one-to-one PSD orthologs were identified. Although in this research it has been use a generalised filtering of a minimal expression of 0.5TPM in all samples and species, to avoid the bias for low expressed genes, and therefore carry out a more robust homology identification, transcripts with expression levels of <1 TPM were filtered out. In this way 205 1:1 homologs were captured in the zebrafish, bat and lion. However, because each specie was obtained from a different RNA seq experiment, a standardized procedure was required. Gene expression across all samples were normalised based on ranking and scaling between 0 and 1. It was noted that other normalisation techniques directed to comparable results (i.e., TMM normalisation and a scaling procedure). This normalised dataset was used for the rest of the analyses carried out in the present chapter.

The 20 most enriched PSD proteins in the three species were obtained by calculating the mean expression of each homolog in all samples (Table 7.1). Prominent was the enrichment of NADH dehydrogenase ubiquinone mitochondrial proteins. Studies un-

covered an immense importance of the subunit 2 (ND2) of such enzyme. The main excitatory amino acid receptor in the CNS is the NMDA receptor, which function depends upon tyrosine phosphorylation (Yu et al., 1997).
Src, the most abundant tyrosine kinase has shown to directly interact with ND2 outside mitochondria, but at excitatory synapses (Gingrich et al., 2004). The fact that these proteins were widely enriched in the mammalian and fish transcriptomes might suggest relevancy upon synaptic plasticity.

An assortment of ribosomal proteins were also found enriched in the PSD homologs. Synthesis of dendrite proteins is vital for constant synaptic adjustments, namely; LTP and LTD. Ribosomal proteins are the horsepower for the protein synthesis machinery. The presence of ribosomes at single synapses is a constraint, therefore ribosomal proteins effectively fulfil the synaptic protein synthesis demands locally at dendrites (Schuman et al., 2006). Defective ribosomal proteins are linked to AD, Huntington, PD, sclerosis and dementia, together with various neurodevelopmental alterations (Slomnicki et al., 2016).

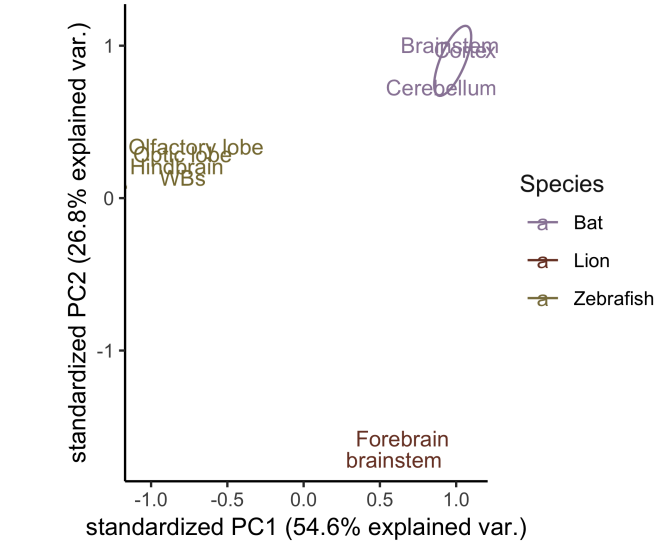
7.3 Comparison of PSD orthologs across species and tissues

Variability among samples was investigated via principal component analysis. PC1 and PC2 showed a strong separation or inverse correlation between the three species without any partition among the tissues (Figure 7.3). Moreover, although the bat's and zebrafish's samples are grouped together within PC3 and PC4, these dimensions showed a clear separation of the brain tissues. Particularly, within PC3, the lion's forebrain and bat's cortex are under an adjacent area (eigenvector close to -1) and inversely correlated with the lion's brainstem and bat's cerebellum (eigenvector of 1.5). On the contrary, PC4 depicted a closer affinity between the lion's forebrain, the zebrafish's olfactory lobe

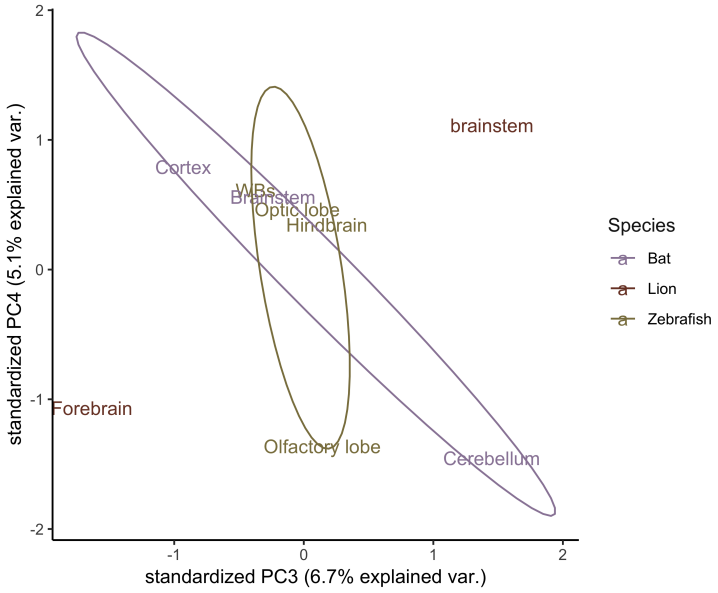
Table 7.1: **Top 20 enriched PSD homologs in zebrafish, bat and lion.**

Gene	Function	Protein class	Zebrafish Mean TPM	Bat Mean TPM	Lion Mean TPM
Rpl18a	RNA binding	nucleic acid binding	568	861	784
Rpl3	RNA binding	nucleic acid binding	734	711	553
Ywhae	regulation of signalling pathways	-	330	818	621
Ndufb9	NADH dehydrogenase	oxidoreductase	354	784	268
Gabarapl2	binding	cytoskeletal protein	326	546	422
Rpl6	binding	nucleic acid binding	1085	257	396
Ndufa8	NADH dehydrogenase	oxidoreductase	215	456	234
Ndufb7	NADH dehydrogenase	oxidoreductase	312	391	164
Ndufa6	NADH dehydrogenase	oxidoreductase	178	376	260
Cyc1	metal ion binding	-	210	286	205
Cct7	protein binding	chaperone	200	339	122
Atp6v1b2	ATP binding	hydrolase	114	565	311
Psmb4	protease	-	181	151	238
Suclg1	binding	hydrolase	123	213	29
Ndufb5	NADH dehydrogenase	oxidoreductase	267	187	76
Pdhb	pyruvate dehydrogenase	-	246	207	73
Arpc1a	actin binding	cytoskeletal protein	103	229	148
Ddx5	binding	-	110	172	196
Dctn2	protein binding	cytoskeletal protein	104	185	169
Ndufb6	NADH dehydrogenase	oxidoreductase	151	197	75

and bat's cerebellum (eigenvector ≤-1) from the zebrafish's whole brains, optic lobe and hindbrain, bat's cortex and lion's brainstem.



(a)



(b)

Figure 7.3: **Principal component analysis (PCA) of PSD homologs.** a) PC1 vs PC2 b) PC3 vs PC4. Each colour represents a different specie.

Concurrent with the PCA analysis, a phylogeny generated from the expression distance matrices separated species (Figure 7.4). Yet, an evident clustering of the bat and lion

samples led to a wide separation between mammals and fish. Lastly, a Pearson correlation matrix showed a perfect negative correlation coefficients (of 1) of all zebrafish brain samples over the bat and lion (Figure 7.5). Whilst at the same time, no linear correlation was found between the bat and lion samples (correlation coefficient close to zero). Thereby, together the PCA, phylogeny-based dendogram and correlation coefficient matrix implies the presence of a strong divergence of expression between species with little impact among the different brain tissues.

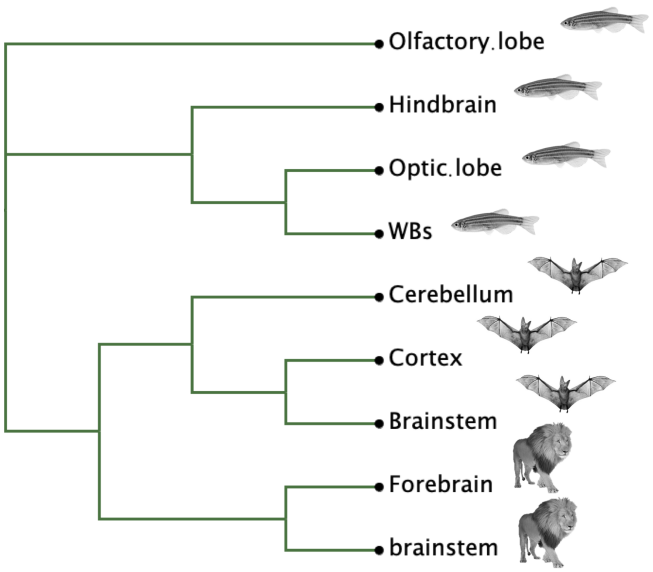


Figure 7.4: **PSD expression phylogeny.** Expression phylogeny dendogram was inferred based upon expression distance matrix by neighbor-joining methodology using R tools (pvclust package).

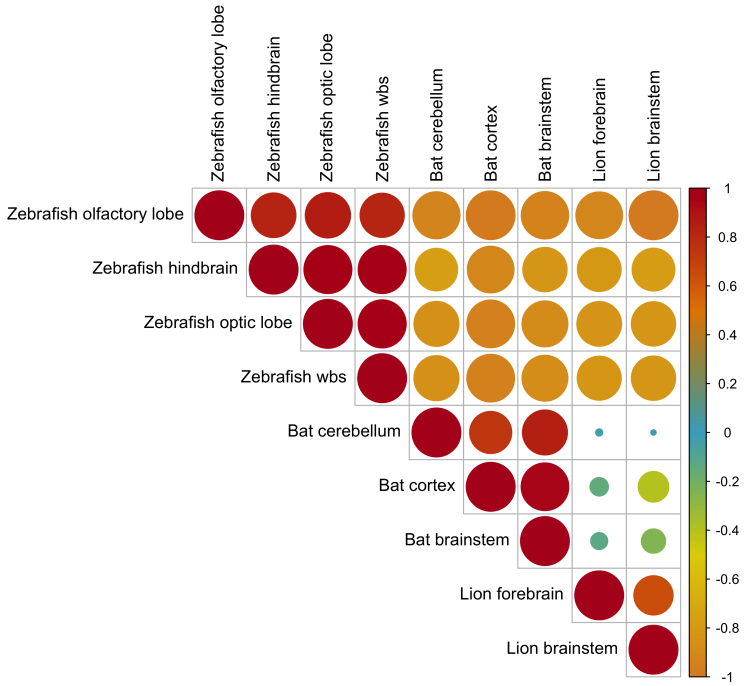


Figure 7.5: **Correlation coefficient matrix based on PSD homologs.** Positive correlations are depicted as red circles, while negative correlations are represented in orange and non-linear relation are in blue. The intensity of the colour and size of the circles are relative to the correlation coefficient.

7.4 Expression distribution of PSD homologs

As an effort to accurately explore and visualize the normalised expression of PSD homologs between species and tissues, a clustered heatmap was generated (Figure 7.6). A clear similarity between the bat and lion homologs was observed, which clustered together and were separated from the zebrafish tissues. Yet, it was also noted a generally conserved expression patterns in all the species. For example, the top 20 enriched proteins are distinctly represented in the heatmap (Figure 7.6 with red squares). Similarly, large clusters of genes with an average high, mean and low expression levels in all the species are distinguished with different colours (orange, yellow and blue squares, respectively).

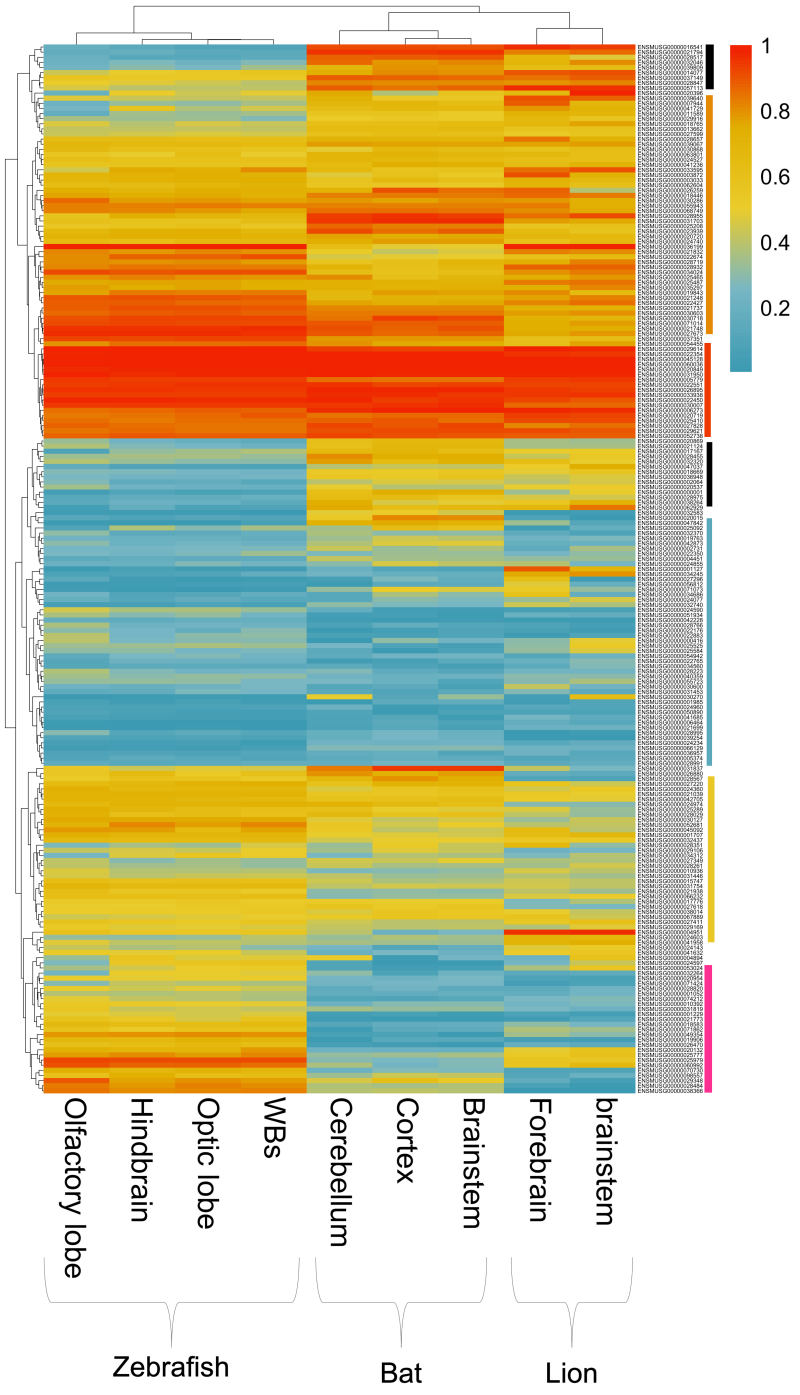


Figure 7.6: **Clustered heatmap representation of PSD homologs.** The normalised expression from 205 identified homologs in the zebrafish, bat and lion *de novo* assemblies is depicted. Dendogram clustering on the X-axis indicates sample similarity, while Y-axis dendogram clustering groups transcripts with similar expression. Expression dissimilarities are denoted with coloured squres; red representing the highest expressed genes, followed by orange, yellow and blue, which depicted genes with low expression levels. Black squares delineate enriched genes in the bat and lion, and pink squares represent enriched only in the zebrafish.

Small clusters of homologs that were determined with a high expression in the bat and lion but not in the zebrafish (Figure 7.6; depicted with black squares), holds a molecular function involved in binding (trafficking protein particle complex), catalytic activity (glutamate dehydrogenase, phospholipid phosphatase) and receptor activity (glutamatergic neurotransmitter receptors). Enriched expression of PSD genes in the zebrafish with low expression in the mammalian species are clustered (Figure 7.6; pink square). Molecular functions were identified, including binding (ras GTPase-activating protein-binding, striatin, golgi SNAP receptor complex, syntaxin), catalytic activity (dipeptidyl peptidase), receptor activity (glutamate receptor). Signal transducer proteins and transporters.

PSD homologs with highly variable expression between samples were identified. Based on the squared cofficient of variation (CV2), 11 proteins were determined as highly variable (p-Value ≤1e-3, chi-squared distribution) (Figure 7.7). GO annotations of these showed an enrichment of proteins involved in the interaction of multiple molecules (protein, enzyme, GTP, actin and actin filament and nucleotide) and reversible phosphorylation processes. Some of these proteins represent specific species adaptations. For example *Ataxin* (CV2 of 5.3e-01, qval of 5.7e-07), which is enriched in bat and lion mediates long-term olfactory habituation, an event that causes a reduced behavioural response due to a long period of odor exposure (McCann et al., 2011). *Cofilin 2*, also enriched in bat and lion, is an actin-binding protein muscle, which aberrations results in congenital myopathies (Agrawal et al., 2007).

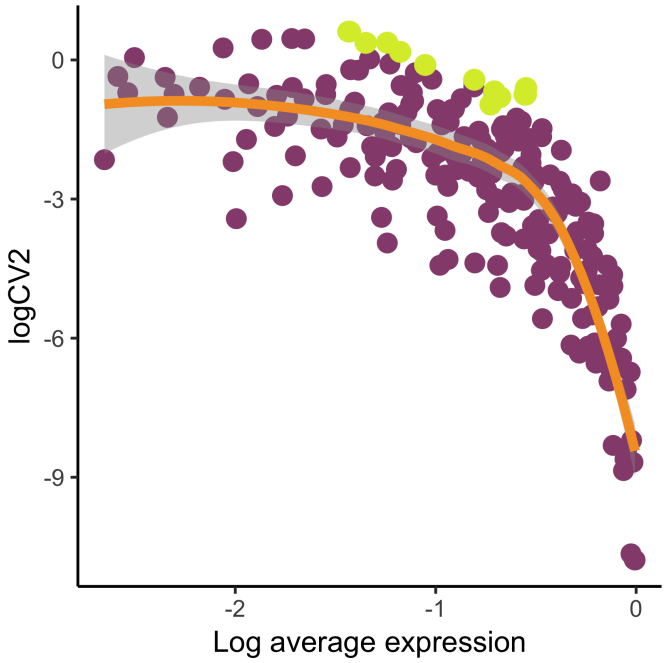


Figure 7.7: **Squared coefficient of variation (CV2).** For each PSD homolog it was estimated their CV2. The x-axis shows the log mean expression of the 3 tissues; the y-axis represents the log transformed CV2. The solid orange curve represent the fitted variance-mean dependence; the dashed lines reflects a 95% confidence interval; green dots correspond to transcripts which CV2 is significantly higher than 50% (CV2 > 0.25) .

7.5 Species-enriched PSDs

PSD orthologs that were enriched (≥1.23 foldchange) in a single species were determined. A greater number of zebrafish-enriched PSDs were revealed (n=43 orthologs), followed by bat-enriched (n=30) and lion-enriched (n=24). Although the functional similarity of PSD proteins across species is generally conserved, a number of functional groups were particularly enriched in determined species. For example, using Panther tools, the term "Protein Class" was used to classify proteins according to their gene function.

The zebrafish-enriched PSD were highly expressed in membrane traffic proteins, such as; members of the SNARE complex (*Syntaxin 6* and Golgi SNAP receptor complex

member 1-*GOSR1*) and vesicle coat proteins, along with nucleic acid binding proteins, namely; cleavage and polyadenylation specificity factor (Nudt21), paraspeckle component (Pspc1), splicing factor, proline- and glutamine-rich (Sfpq) and ras GTPase-activating protein-binding protein (G3bp1). These proteins are pivotal for post-trans criptionally regulation of mRNA levels and all brain development and synaptic function (synaptogenesis and axon guidance) (Su et al., 2018). Enzyme modulator and transporter protein class were also enriched in the zebrafish, the former included the protein kinase MOB family member 4 (Mob4) and the GTPases ras-related protein RAP-1b (Rap1b) and importin-7 (Ipo7), while the later involved glutamate receptor (Grid2) and vacuolar protein sorting (Vps45).

On the other hand, the bat and lion-enriched PSD proteins were involved in very similar Panther protein classes, mainly in catalytic activity (i.e., phosphatases and kinases) and binding proteins. The most enriched bat protein class were involved in catalytic activity and binding, including; GTP-binding protein Di-Ras2 (Diras2), eerine/threonine-protein kinase mTOR (Mtor), phospholipid phosphatase (Plpp3), cyclin-dependent kinase (Cdk17) and guanine nucleotide-binding protein G (Gnai3). Among the most PSD proteins enriched in lion, comprised; girdin (Ccdc88a), Bardet-Biedl syndrom (Bbs1), histone deacetylase 11 (Hdac11), Rho-related GTP-binding (RhoQ), Striatin (Strn) and 39S ribosomal protein L12 (Mrpl12), many of which are enzyme modulators, membrane traffic proteins and nucleic acid binding (Table 7.2).

Table 7.2: **Species-enriched Panther protein classes of PSD orthologs.**

Protein class	No. proteins	Percentage	Specie
Transporter (PC00227)	2	8.0	Zebrafish
Membrane traffic protein	5	20.0	Zebrafish
Hydrolase	2	8.0	Zebrafish
Oxidoreductase	1	4.0	Zebrafish
Cell adhesion molecule	1	4.0	Zebrafish
Cell junction protein	1	4.0	Zebrafish
Enzyme modulator	3	12.0	Zebrafish
Transferase	1	4.0	Zebrafish
Transcription factor	1	4.0	Zebrafish
Nucleic acid binding	4	16.0	Zebrafish
Receptor (PC00197)	1	4.0	Zebrafish
Cytoskeletal protein	1	4.0	Zebrafish
Signaling molecule	2	8.0	Zebrafish
Binding	5	29.4	Bat
Receptor activity	1	5.9	Bat
Structural molecule activity	2	11.8	Bat
Signal transducer activity	2	11.8	Bat
Catalytic activity	6	35.3	Bat
Transporter activity	1	5.9	Bat
Binding	7	50.0	Lion
Structural molecule activity	1	7.1	Lion
Signal transducer activity	1	7.1	Lion
Catalytic activity	4	28.6	Lion
Transporter activity	1	7.1	Lion

Significantly enriched gene ontology terms were also explored for each species-enriched PSD orthologs (Figure 7.8). Notably, enriched molecular function GO terms associated with "binding" were observed in all species-enriched PSD proteins. In general, all three species showed a larger number of proteins involved in cellular component, compared to the other two GO terms. Similarly as the Panther protein class analy-

sis, the zebrafish-enriched PSD were mostly associated with the vesicle-mediated protein transport, which is mediated by SNARE proteins and Golgi-derived structures. GO terms associated with bat-enriched PSDs showed an enrichment in mitochondria, kinases and protein phosphorylation. Studies have identified a key role in reversible protein phosphorylation in animals during the transition to and from torpor, allowing control of several enzymes and protein stabilization (also found enriched) (Eddy et al., 2005). The mitochondria have a pivotal importance during torpor, which mediates oxidative metabolism and reduces the overall metabolic rate (Storey, 1997). Additionally, as observed in Chapter 5, an adaptation in mitochondrial genes has been found in bats to fulfill the heavy energy consumption that the flight activity demands. Within the lion-enriched GO terms, more ontologies related to "synapse" were found compared with the other two species. The term "cytoskeleton" was also seen in various ontologies. Actin cytoskeleton organization is the primordial component of the cellular scaffold for molding and keeping the pre- and postsynaptic terminal shape. Moreover, synaptic transmission is often accompanied by changes in the cytoskeleton driving to new synaptic connections, and thereby contributing short and long-term memory (Lamprecht and LeDoux, 2004; Cingolani and Goda, 2008).

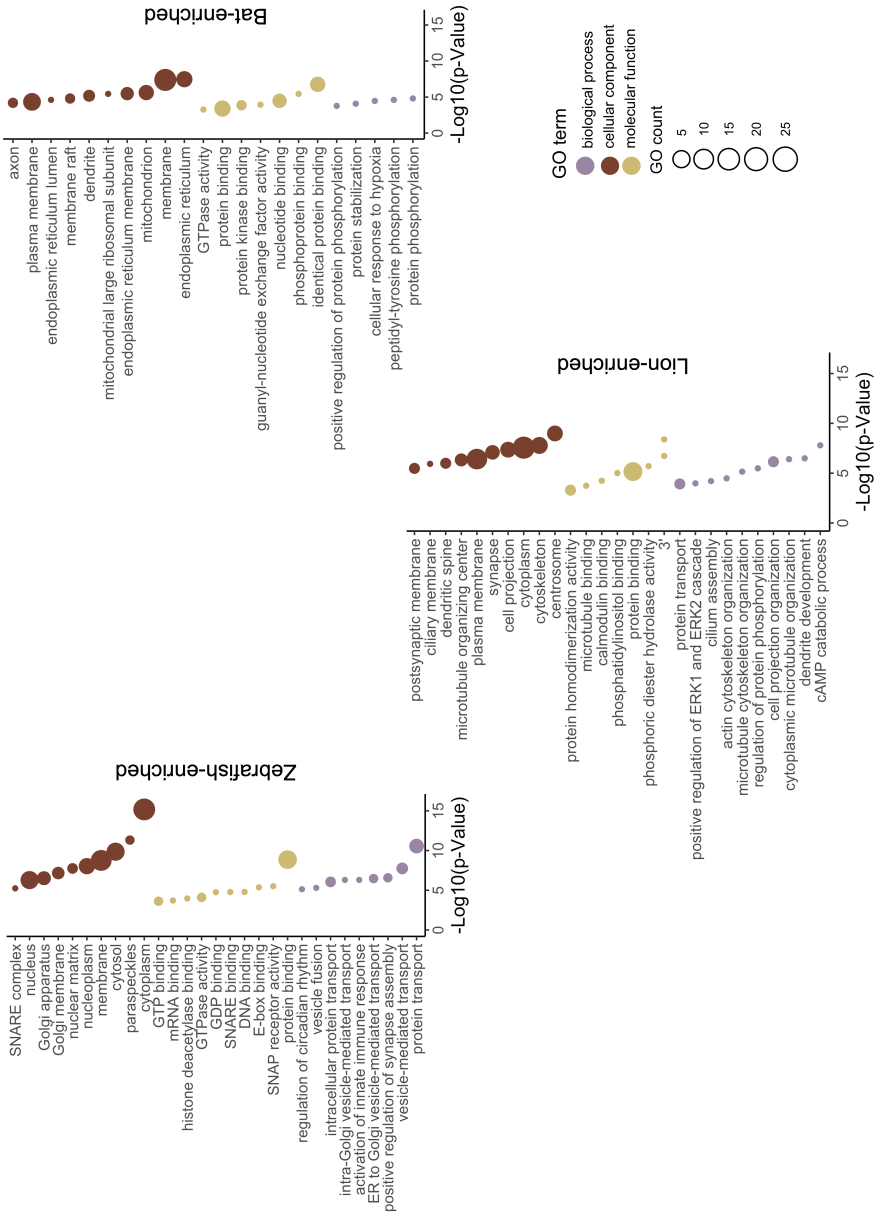


Figure 7.8: **GO analysis of PSD specie-enrichment.** Expression enrichment in a single specie was obtained, and proteins were analysed using GO terms. The top most significantly enriched GO terms were plotted against the negative log10 P-value and coloured based to their functional category. P values were estimated based on hypergeometric distribution and adjusted by false discovery rate (FDR) control procedure applying a threshold of 0.05.

7.6 Summary and comments

This chapter provided a general comparison for the transcriptomes of the *de novo* assembled species. Within the zebrafish transcriptome, it was found a larger number of transcripts, together with a greater amount of mouse orthologs, whereas the bat transcriptome contained the smallest number of both, respectively. From the mouse orthologs expressed in the genome, brain, SYN and PSD, it was noted a higher number of shared orthologs for the SYN and PSD across the species than with the brain. In a similar way, within the brain proteome, more orthologs were found compared to the genome. Therefore, a substantial evolutionary constraint in the SYN and PSD proteins since the divergence of fish and mammals was suggested. It was observed that the overall molecular function of the shared PSD are involved in elemental forms of environmental stimuli apparatus, which have been highly conserved through various mutations, duplications and deletion processes. Noteworthy, was the large number of proteins involved in molecular binding but also in signalling pathways among the shared PSD proteins aimed to create macromolecular complexes. Amongst the most enriched PSD 1:1 orthologs, ribosomal and NADH dehydrogenase ubiquinone were particularly enriched. Thereby a paramount functioning in synaptogenesis and plasticity for these molecules, respectively was proposed.

Little evidence indicating either homogeneity or dissimilarity among brain tissues was found (e.g., tissues from the forebrain being more similar to the cortex or optic lobe than tissues from the cerebellum or brainstem). However, a solid separation of the zebrafish transcriptome from the lion and bat, based on PCA analysis, correlation matrix and gene expression clustering was evident, defining a strong lineage-specific differences in gene expression. Specific PSD proteins were identified with high expression in the zebrafish and low expression expression in the bat and lion, which might underlie the few distinctions of brain structures and synapse types (e.g., size, the presence of cortex in mammals and general complexity).

Chapter 8

Conclusions and Future perspectives

“One, remember to look up at the stars and not down at your feet. Two, never give up work. Work gives you meaning and purpose and life is empty without it. Three, if you are lucky enough to find love, remember it is there and don’t throw it away."

- Stephen Hawking -

The evolution of the synapse proteome precedes the existence of complex organisms such as eukaryotes. Approximately, 2,662 mya (Emes and Grant, 2012) an integration of signalling pathways, e.g., cell-cell communication, chemosensory, cell-differentiation and cell-adhesion allowed ancient bacteria to adapt to changing environments. Therefore, primordial attributes of the mammalian synaptic transition were already present before the emergence of multicellularity. For example, neurosecretory SNARE proteins, MAGUKs and PSD (postsynaptic density) scaffolds were expressed before the emergence of synapses in the unicellular choanoflagellates around 1,450 mya. Together, these molecules were key feature for the expansion and diversification of protein families to complex macromolecular circuits that characterize the origin of the brain. To this extent, the PSD is undoubtedly a master piece written in the evolutionary history of metazoans. This extraordinarily complex sub-organelle allows the gathering, assimilation and circulation of information in the shape of learning and memory. Proteomic studies have enormously contributed to the characterization of more than 1,500 PSD proteins that make up the immense pool of behaviours observed in the animal kingdom, but also given its high intricacy, a manifold of PSDs have been directly linked to a substantial number of neurological disorders.

With the recent progress of NGS, the transcriptome of any species can be rapidly explored. *De novo* transcriptome assembly is a key method that assembles millions of short-reads into a full-length transcriptome with the absence of a reference genome. This is fundamentally important since very few species possess a high-quality reference genome. Yet, several bioinformatics issues challenges the accuracy of the assembly and annotation of the transcriptome. This thesis developed a robust bioinformatics pipeline that tackles various of such difficulties, and assesses the validity of the *de novo* transcriptome, yielding its precise understanding. Therefore, this work exploited the recent high throughput technologies to *de novo* assembly the core brain transcriptome of three species, and explores gene expression in the PSD, SYN (synaptosome), brain and genome.

The general results in this work sustained the "synapse first model" (Ryan and Grant (2009)), in which basic synaptic components evolved a million years ago before the existence of a nervous system. Notable was the stronger conservation of PSD and SYN genes compared with those expressed in the brain and genome in all species. The most highly enriched PSDs were receptors, cell-adhesion and signalling enzymes. These genes encode proteins comprising the underpinning foundation to the adaptive apparatus of the postsynaptic membrane (Emes and Grant, 2012).

With the assembly and annotation of the zebrafish transcriptome, a frame of reference for testing the effectiveness of the pipeline was provided by the comparison of the *de novo* assembled transcriptome to the available *D. rerio* reference genome. As it is described in Bayés et al. (2017) evidence of the TSGD (teleost-specific genome duplication) was found in various PSD genes, which also displayed a highly variable expression. On the other hand, the bat showed levels of gene constraint, likely a consequence of high metabolic demands. Moreover, species-specific adaptations were observed. For example the zebrafish exhibited various GO terms involved in neurogenesis. On the other hand bats showed evidence of convergent evolution in *TMC1* (transmembrane channel like 1 protein), aberrations of which have been linked to deafness in humans. Highly expressed genes in the bat brain are foremost during turpor. Carnivore adaptation was found in the lion, such as responses to carbohydrate, immune responses, night vision and similar to bats, higher metabolic rates.

Additionally, highly expressed genes in all species were directly associated with neurodegenerative diseases (Alzheimer's disease, Huntington disease and Parkinson's disease) and psychiatric disorders (schizophrenia). The implication of a highly conserved population of synaptic proteins in disease brings to light the weighty necessity to further comprehend the mechanisms of synaptic dysfunction to determine prospective targets for therapeutic aid. Until recently, largely proteomic technologies have charac-

terized hundreds of synaptic proteins, however, advances in transcriptome approaches has provided valuable information of how a gene is expressed and regulated in various tissues, conditions and time points, thereby, substantial in the understanding of disease. Proposed further studies are the integration of proteomic and transcriptomic tools. This can grant advantageous insights, which might not be immediately clear with a single analysis. Moreover, in the scope of transcriptomics, single-cell RNA sequencing is a relatively novel tool that might unveil regulatory networks and pathways of complex unknown dendritic cell-populations in disease and examine the precise expression profiles (Hwang et al., 2018).

Bibliography

M. Abedin. *Cadherin evolution and the origin of animals*. PhD thesis, UC Berkeley, 2010.

P. B. Agrawal, R. S. Greenleaf, K. K. Tomczak, V.-L. Lehtokari, C. Wallgren-Pettersson, W. Wallefeld, N. G. Laing, B. T. Darras, S. K. Maciver, P. R. Dormitzer, et al. Nemaline myopathy with minicores caused by mutation of the cfl2 gene encoding the skeletal muscle actin–binding protein, cofilin-2. *The American Journal of Human Genetics*, 80(1):162–167, 2007.

Z. M. Ahmed, R. Goodyear, S. Riazuddin, A. Lagziel, P. K. Legan, M. Behra, S. M. Burgess, K. S. Lilley, E. R. Wilcox, S. Riazuddin, et al. The tip-link antigen, a protein associated with the transduction complex of sensory hair cells, is protocadherin-15. *Journal of Neuroscience*, 26(26):7022–7034, 2006.

M. Aksenov, M. Aksenova, D. A. Butterfield, and W. R. Markesbery. Oxidative modification of creatine kinase bb in alzheimerâĂŹs disease brain. *Journal of neurochemistry*, 74(6):2520–2527, 2000.

A. Alié and M. Manuel. The backbone of the post-synaptic density originated in a unicellular ancestor of choanoflagellates and metazoans. *BMC evolutionary biology*, 10 (1):34, 2010.

A. Alunni, J.-M. Hermel, A. Heuzé, F. Bourrat, F. Jamen, and J.-S. Joly. Evidence for neural stem cells in the medaka optic tectum proliferation zones. *Developmental neurobiology*, 70(10):693–713, 2010.

F. J. Alvarez-Leefmans and E. Delpire. *Physiology and pathology of chloride transporters and channels in the nervous system*. Elsevier, 2009.

A. Amores, A. Force, Y. Yan, C. Amemiya, A. Fritz, R. Ho, L. Joly, J. Langeland, V. Prince, Y. Wang, et al. Genome duplications in vertebrate evolution: evidence from zebrafish hox clusters. *Science*, 282:1711–1714, 1998.

C. B. Andrews, S. A. Mackenzie, and T. R. Gregory. Genome size and wing parameters in passerine birds. *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1654):55–61, 2009.

I. Anjum, S. S. Jaffery, M. Fayyaz, Z. Samoo, and S. Anjum. The role of vitamin d in brain health: a mini literature review. *Cureus*, 10(7), 2018.

A. Antunes, J. L. Troyer, M. E. Roelke, J. Pecon-Slattery, C. Packer, C. Winterbach, H. Winterbach, G. Hemson, L. Frank, P. Stander, et al. The evolutionary dynamics of the lion panthera leo revealed by host and viral population genomics. *PLoS genetics*, 4(11):e1000251, 2008.

J. Arikkath and K. P. Campbell. Auxiliary subunits: essential components of the voltage-gated calcium channel complex. *Current opinion in neurobiology*, 13(3):298–307, 2003.

G. Astarita, K.-M. Jung, V. Vasilevko, N. V. DiPatrizio, S. K. Martin, D. H. Cribbs, E. Head, C. W. Cotman, and D. Piomelli. Elevated stearoyl-coa desaturase in brains of patients with alzheimer's disease. *PLoS One*, 6(10):e24777, 2011.

C. Astorga, R. A. Jorquera, M. Ramírez, A. Kohler, E. López, R. Delgado, A. Córdova, P. Olguín, and J. Sierralta. Presynaptic dlg regulates synaptic function through the localization of voltage-activated ca 2+ channels. *Scientific reports*, 6:32132, 2016.

S. B. Bagatharia, M. N. Joshi, R. V. Pandya, A. S. Pandit, R. P. Patel, S. M. Desai, A. Sharma, O. Panchal, F. P. Jasmani, and A. K. Saxena. Complete mitogenome of asiatic lion resolves phylogenetic status within panthera. *BMC genomics*, 14(1):572, 2013.

G. Banisadr, W. Rostène, P. Kitabgi, and S. M. Parsadaniantz. Chemokines and brain functions. *Current Drug Targets-Inflammation & Allergy*, 4(3):387–399, 2005.

A. Barnett and G. J. Brewer. Autophagy in aging and alzheimer's disease: pathologic or protective? *Journal of Alzheimer's Disease*, 25(3):385–394, 2011.

R. Barnett, N. Yamaguchi, I. Barnes, and A. Cooper. The origin, current diversity and future conservation of the modern lion (panthera leo). *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1598):2119–2125, 2006.

H. Bauer and S. Van Der Merwe. Inventory of free-ranging lions panthera leo in africa. *Oryx*, 38(1):26–31, 2004.

H. Bauer, K. Nowell, and C. Packer. Panthera leo. iucn red list of threatened species, version 2011., 2012.

H. Bauer, G. Chapron, K. Nowell, P. Henschel, P. Funston, L. T. Hunter, D. W. Macdonald, and C. Packer. Lion (panthera leo) populations are declining rapidly across africa, except in intensively managed areas. *Proceedings of the National Academy of Sciences*, 112(48):14894–14899, 2015.

A. Bayés and S. G. Grant. Neuroproteomics: understanding the molecular organization and complexity of the brain. *Nature Reviews Neuroscience*, 10(9):635, 2009.

À. Bayés, L. N. Van De Lagemaat, M. O. Collins, M. D. Croning, I. R. Whittle, J. S. Choudhary, and S. G. Grant. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature neuroscience*, 14(1):19, 2011.

À. Bayés, M. O. Collins, M. D. Croning, L. N. van de Lagemaat, J. S. Choudhary, and S. G. Grant. Comparative study of human and mouse postsynaptic proteomes finds high compositional conservation and abundance differences for key synaptic proteins. *PloS one*, 7(10):e46683, 2012.

A. Bayés, M. O. Collins, R. Reig-Viader, G. Gou, D. Goulding, A. Izquierdo, J. S. Choudhary, R. D. Emes, and S. G. Grant. Evolution of complexity in the zebrafish synapse proteome. *Nature communications*, 8:14613, 2017.

C. G. Becker and T. Becker. Gradients of ephrin-a2 and ephrin-a5b mrna during retinotopic regeneration of the optic projection in adult zebrafish. *Journal of Comparative Neurology*, 427(3):469–483, 2000.

C. G. Becker, J. Schweitzer, J. Feldner, M. Schachner, and T. Becker. Tenascin-r as a repellent guidance molecule for newly growing and regenerating optic axons in adult zebrafish. *Molecular and Cellular Neuroscience*, 26(3):376–389, 2004.

N. C. Berchtold, P. D. Coleman, D. H. Cribbs, J. Rogers, D. L. Gillen, and C. W. Cotman. Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and alzheimer's disease. *Neurobiology of aging*, 34(6):1653–1661, 2013.

J. Bernal. Thyroid hormone receptors in brain development and function. *Nature Reviews Endocrinology*, 3(3):249, 2007.

M. J. Berridge. Calcium hypothesis of alzheimer's disease. *Pflügers Archiv-European Journal of Physiology*, 459(3):441–449, 2010.

C. Berthelot, F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadie, A. Alberti, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature communications*, 5:3657, 2014.

E. Birney, S. Kumar, and A. R. Krainer. Analysis of the rna-recognition motif and rs and rgg domains: conservation in metazoan pre-mrna splicing factors. *Nucleic acids research*, 21(25):5803–5816, 1993.

P. Blay, C. Nilsson, C. Owman, A. Aldred, and G. Schreiber. Transthyretin expression in the rat brain: effect of thyroid functional state and role in thyroxine transport. *Brain research*, 632(1-2):114–120, 1993.

T. Bliss, G. Collingridge, and R. Morris. Synaptic plasticity in health and disease: introduction and overview, 2014.

T. M. Böckers. The postsynaptic density. *Cell and tissue research*, 326(2):409–422, 2006.

T. M. Böckers, M. G. Mameza, M. R. Kreutz, J. Bockmann, C. Weise, F. Buck, D. Richter, E. D. Gundelfinger, and H.-J. Kreienkamp. Synaptic scaffolding proteins in rat brain ankyrin repeats of the multidomain shank protein family interact with the cytoskeletal protein α-fodrin. *Journal of Biological Chemistry*, 276(43):40104–40112, 2001.

J. Boggs. Myelin basic protein: a multifunctional protein. *Cellular and Molecular Life Sciences CMLS*, 63(17):1945–1961, 2006.

J. W. Boughman. Vocal learning by greater spear–nosed bats. *Proceedings of the Royal Society of London B: Biological Sciences*, 265(1392):227–233, 1998.

I. Braasch and J. H. Postlethwait. The teleost agouti-related protein 2 gene is an ohnolog gone missing from the tetrapod genome. *Proceedings of the National Academy of Sciences*, 108(13):E47–E48, 2011.

D. Braida, F. Guerini, L. Ponzoni, I. Corradini, S. De Astis, L. Pattini, E. Bolognesi, R. Benfante, D. Fornasari, M. Chiappedi, et al. Association between snap-25 gene polymorphisms and cognition in autism: functional consequences and potential therapeutic strategies. *Translational psychiatry*, 5(1):e500, 2016.

N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic rna-seq quantification. *Nat Biotech*, 34(5):525–527, 05 2016.

C. E. Brook and A. P. Dobson. Bats as âĂŸspecialâĂŹreservoirs for emerging zoonotic pathogens. *Trends in microbiology*, 23(3):172–180, 2015.

A. K. Brunet-Rossinni and S. N. Austad. Ageing studies on bats: a review. *Biogerontology*, 5(4):211–222, 2004.

J. S. Buell and B. Dawson-Hughes. Vitamin d and neurocognitive dysfunction: preventing âĂIJdâĂİ ecline? *Molecular aspects of medicine*, 29(6):415–422, 2008.

R. Buffenstein and M. Pinto. Endocrine function in naturally long-living small mammals. *Molecular and cellular endocrinology*, 299(1):101–111, 2009.

P. Burkhardt. The origin and evolution of synaptic proteins–choanoflagellates lead the way. *Journal of Experimental Biology*, 218(4):506–514, 2015.

P. Burkhardt, C. M. Stegmann, B. Cooper, T. H. Kloepper, C. Imig, F. Varoqueaux, M. C. Wahl, and D. Fasshauer. Primordial neurosecretory apparatus identified in the choanoflagellate monosiga brevicollis. *Proceedings of the National Academy of Sciences*, page 201106189, 2011.

P. Burkhardt, M. Grønborg, K. McDonald, T. Sulur, Q. Wang, and N. King. Evolutionary insights into premetazoan functions of the neuronal protein homer. *Molecular biology and evolution*, 31(9):2342–2355, 2014.

J. H. Byrne, R. Heidelberger, and M. N. Waxham. *From molecules to networks: an introduction to cellular and molecular neuroscience*. Academic Press, 2014.

C. H. Calisher, J. E. Childs, H. E. Field, K. V. Holmes, and T. Schountz. Bats: important reservoir hosts of emerging viruses. *Clinical microbiology reviews*, 19(3):531–545, 2006.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):421, 2009.

M. Carr, B. S. Leadbeater, R. Hassan, M. Nelson, and S. L. Baldauf. Molecular phylogeny of choanoflagellates, the sister group to metazoa. *Proceedings of the National Academy of Sciences*, 2008.

J. M. Catchen, J. S. Conery, and J. H. Postlethwait. Automated identification of conserved synteny after whole-genome duplication. *Genome research*, 2009.

J. César Rosa and M. de Cerqueira César. Role of hexokinase and vdac in neurological disorders. *Current molecular pharmacology*, 9(4):320–331, 2016.

Y. F. Chan, M. E. Marks, F. C. Jones, G. Villarreal, M. D. Shapiro, S. D. Brady, A. M. Southwick, D. M. Absher, J. Grimwood, J. Schmutz, et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitx1 enhancer. *science*, 327(5963): 302–305, 2010.

K. Chandrasekaran, T. Giordano, D. R. Brady, J. Stoll, L. J. Martin, and S. I. Rapoport. Impairment in mitochondrial cytochrome oxidase gene expression in alzheimer disease. *Molecular brain research*, 24(1-4):336–340, 1994.

C. T. Chasapis, A. C. Loutsidou, C. A. Spiliopoulou, and M. E. Stefanidou. Zinc and human health: an update. *Archives of toxicology*, 86(4):521–534, 2012.

T. E. Chater and Y. Goda. The role of ampa receptors in postsynaptic mechanisms of synaptic plasticity. *Frontiers in cellular neuroscience*, 8:401, 2014.

X. Chen, J. M. Levy, A. Hou, C. Winters, R. Azzam, A. A. Sousa, R. D. Leapman, R. A. Nicoll, and T. S. Reese. Psd-95 family maguks are essential for anchoring ampa and nmda receptor complexes at the postsynaptic density. *Proceedings of the National Academy of Sciences*, 112(50):E6983–E6992, 2015.

Y. A. Chen and R. H. Scheller. Snare-mediated membrane fusion. *Nature reviews Molecular cell biology*, 2(2):98, 2001.

D. Cheng, C. C. Hoogenraad, J. Rush, E. Ramm, M. A. Schlager, D. M. Duong, P. Xu, S. R. Wijayawardana, J. Hanfelt, T. Nakagawa, et al. Relative and absolute quantification of postsynaptic density proteome isolated from rat forebrain and cerebellum. *Molecular & cellular proteomics*, 5(6):1158–1170, 2006.

L. K. Chico, L. J. Van Eldik, and D. M. Watterson. Targeting protein kinases in central nervous system disorders. *Nature Reviews Drug Discovery*, 8(11):892, 2009.

D. M. Church, L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein, X. She, C. J. Bult, R. Agarwala, J. L. Cherry, M. DiCuccio, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*, 7(5):e1000112, 2009.

L. A. Cingolani and Y. Goda. Actin in action: the interplay between the actin cytoskeleton and synaptic efficacy. *Nature Reviews Neuroscience*, 9(5):344, 2008.

D. E. Clapham. Calcium signaling. *Cell*, 131(6):1047–1058, 2007.

M. O. Collins, H. Husi, L. Yu, J. M. Brandon, C. N. Anderson, W. P. Blackstock, J. S. Choudhary, and S. G. Grant. Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *Journal of neurochemistry*, 97(s1):16–23, 2006.

C. Conaco, D. S. Bassett, H. Zhou, M. L. Arcila, S. M. Degnan, B. M. Degnan, and K. S. Kosik. Functionalization of a protosynaptic gene expression network. *Proceedings of the National Academy of Sciences*, 109(Supplement 1):10612–10618, 2012.

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13, 2016.

E. Conibear and N. G. Davis. Palmitoylation and depalmitoylation dynamics at a glance. *J Cell Sci*, 123(23):4007–4010, 2010.

E. Corder, A. M. Saunders, N. Risch, W. Strittmatter, D. Schmechel, P. Gaskell, J. Rimmler, P. Locke, P. Conneally, K. Schmader, et al. Protective effect of apolipoprotein e type 2 allele for late onset alzheimer disease. *Nature genetics*, 7(2):180–184, 1994.

E. Covey and J. H. Casseday. Timing in the auditory system of the bat. *Annual review of physiology*, 61(1):457–476, 1999.

A. M. Craig and Y. Kang. Neurexin–neuroligin signaling in synapse development. *Current opinion in neurobiology*, 17(1):43–52, 2007.

C. J. Creevey, J. Muller, T. Doerks, J. D. Thompson, D. Arendt, and P. Bork. Identifying single copy orthologs in metazoa. *PLoS computational biology*, 7(12):e1002269, 2011.

M. D. Croning, M. C. Marshall, P. McLaren, J. D. Armstrong, and S. G. Grant. G2cdb: the genes to cognition database. *Nucleic acids research*, 37(suppl_1):D846–D851, 2008.

D. D'agostino, M. Bertelli, S. Gallo, S. Cecchin, E. Albiero, P. G. Garofalo, A. Gambardella, J.-M. S. Hilaire, H. Kwiecinski, E. Andermann, et al. Mutations and polymorphisms of the clcn2 gene in idiopathic epilepsy. *Neurology*, 63(8):1500–1502, 2004.

C. Davey, A. Tallafuss, and P. Washbourne. Differential expression of neuroligin genes in the nervous system of zebrafish. *Developmental Dynamics*, 239(2):703–714, 2010.

K. T. Davies, I. Maryanto, and S. J. Rossiter. Evolutionary origins of ultrasonic hearing and laryngeal echolocation in bats inferred from morphological analyses of the inner ear. *Frontiers in zoology*, 10(1):2, 2013.

C. M. de Arrieta, B. Morte, A. Coloma, and J. Bernal. The human rc3 gene homolog, nrgn contains a thyroid hormone-responsive element located in the first intron. *Endocrinology*, 140(1):335–343, 1999.

G. De Ferrari and R. Moon. The ups and downs of wnt signaling in prevalent neurological disorders. *Oncogene*, 25(57):7545, 2006.

A. De Haas, H. Van Weering, E. De Jong, H. Boddeke, and K. Biber. Neuronal chemokines: versatile messengers in central nervous system cell interaction. *Molecular neurobiology*, 36(2):137–151, 2007.

A. de Mendoza, H. Suga, and I. Ruiz-Trillo. Evolution of the maguk protein gene family in premetazoan lineages. *BMC evolutionary biology*, 10(1):93, 2010.

J. de Wit and A. Ghosh. Control of neural circuit formation by leucine-rich repeat proteins. *Trends in neurosciences*, 37(10):539–550, 2014.

C. Dean, F. G. Scholl, J. Choih, S. DeMaria, J. Berger, E. Isacoff, and P. Scheiffele. Neurexin mediates the assembly of presynaptic terminals. *Nature neuroscience*, 6(7):708, 2003.

P. Dehal and J. L. Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology*, 3(10):e314, 2005.

V. Demartsev, A. Ilany, A. Kershenbaum, Y. Geva, O. Margalit, I. Schnitzer, A. Barocas, E. Bar-Ziv, L. Koren, and E. Geffen. The progression pattern of male hyrax songs and the role of climactic ending. *Scientific reports*, 7(1):2794, 2017.

A. Di Rita, A. Peschiaroli, D. Pasquale, D. Strobbe, Z. Hu, J. Gruber, M. Nygaard, M. Lambrughi, G. Melino, E. Papaleo, et al. Huwe1 e3 ligase promotes pink1/parkin-independent mitophagy by regulating ambra1 activation via ikkα. *Nature communications*, 9(1):3755, 2018.

T. Dickmeis. Glucocorticoids and the circadian clock. *Journal of Endocrinology*, 200(1): 3, 2009.

A. Diot, K. Morten, and J. Poulton. Mitophagy plays a central role in mitochondrial ageing. *Mammalian Genome*, 27(7-8):381–395, 2016.

A. Dosemeci, A. J. Makusky, E. Jankowska-Stephens, X. Yang, D. J. Slotta, and S. P. Markey. Composition of the synaptic psd-95 complex. *Molecular & Cellular Proteomics*, 6(10):1749–1760, 2007.

S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184, 2009.

S. F. Eddy, J. McNally, and K. B. Storey. Up-regulation of a thioredoxin peroxidase-like protein, proliferation-associated gene, in hibernating bats. *Archives of biochemistry and biophysics*, 435(1):103–111, 2005.

S. R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9): 755–763, 1998.

G. M. Elias and R. A. Nicoll. Synaptic trafficking of glutamate receptors by maguk scaffolding proteins. *Trends in cell biology*, 17(7):343–352, 2007.

B. Ellenbroek and J. Youn. Rodent models in neuroscience research: is it a rat race? *Disease models & mechanisms*, 9(10):1079–1087, 2016.

R. O. Emerson and J. H. Thomas. Adaptive evolution in zinc finger transcription factors. *PLoS genetics*, 5(1):e1000325, 2009.

R. D. Emes and S. G. Grant. The human postsynaptic density shares conserved elements with proteomes of unicellular eukaryotes and prokaryotes. *Frontiers in neuroscience*, 5:44, 2011.

R. D. Emes and S. G. Grant. Evolution of synapse complexity and diversity. *Annual review of neuroscience*, 35:111–131, 2012.

R. D. Emes, A. J. Pocklington, C. N. Anderson, A. Bàyes, M. O. Collins, C. A. Vickers, M. D. Croning, B. R. Malik, J. S. Choudhary, J. D. Armstrong, et al. Evolutionary expansion and anatomical specialization of synapse proteome complexity. *Nature neuroscience*, 11(7):799–806, 2008.

E. C. Engle. The genetic basis of complex strabismus. *Pediatric research*, 59(3):343, 2006.

E. C. Engle. Oculomotility disorders arising from disruptions in brainstem motor neuron development. *Archives of neurology*, 64(5):633–637, 2007.

J. Fang, X. Wang, S. Mu, S. Zhang, and D. Dong. Bgd: A database of bat genomes. *PloS one*, 10(6):e0131296, 2015.

M. S. Fanselow. Neural organization of the defensive behavior system responsible for fear. *Psychonomic bulletin & review*, 1(4):429–438, 1994.

A. Fedotova, A. Bonchuk, V. Mogila, and P. Georgiev. C2h2 zinc finger proteins: the largest but poorly explored family of higher eukaryotic transcription factors. *Acta Naturae*, 9(2 (33)), 2017.

E. Fernández, M. O. Collins, R. T. Uren, M. V. Kopanitsa, N. H. Komiyama, M. D. Croning, L. Zografos, J. D. Armstrong, J. S. Choudhary, and S. G. Grant. Targeted tandem affinity purification of psd-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Molecular systems biology*, 5(1):269, 2009.

M. A. Ferreira, M. C. O'Donovan, Y. A. Meng, I. R. Jones, D. M. Ruderfer, L. Jones, J. Fan, G. Kirov, R. H. Perlis, E. K. Green, et al. Collaborative genome-wide association analysis supports a role for ank3 and cacna1c in bipolar disorder. *Nature genetics*, 40(9): 1056, 2008.

R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 2016. doi: 10.1093/nar/gkv1344. URL +http://dx.doi.org/10.1093/nar/gkv1344.

M. Francesconi and B. Lehner. The effects of genetic variation on gene expression dynamics during development. *Nature*, 505(7482):208, 2014.

H. B. Fraser, D. P. Wall, and A. E. Hirsh. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC evolutionary biology*, 3(1): 11, 2003.

R. W. Friedrich, C. J. Habermann, and G. Laurent. Multiplexing using synchrony in the zebrafish olfactory bulb. *Nature neuroscience*, 7(8):862, 2004.

G. Froyen, M. Corbett, J. Vandewalle, I. Jarvela, O. Lawrence, C. Meldrum, M. Bauters, K. Govaerts, L. Vandeleur, H. Van Esch, et al. Submicroscopic duplications of the hydroxysteroid dehydrogenase hsd17b10 and the e3 ubiquitin ligase huwe1 are associated with mental retardation. *The American Journal of Human Genetics*, 82(2): 432–443, 2008.

L. Funke, S. Dakoji, and D. S. Bredt. Membrane-associated guanylate kinases regulate adhesion and plasticity at cell junctions. *Annu. Rev. Biochem.*, 74:219–245, 2005.

T. Gabaldón and E. V. Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366, 2013.

G. Gabella. Autonomic nervous system. *e LS*, 2001.

J. Ganz and M. Brand. Adult neurogenesis in fish. *Cold Spring Harbor perspectives in biology*, 8(7):a019018, 2016.

R. A. Garcia, K. Vasudevan, and A. Buonanno. The neuregulin receptor erbb-4 interacts with pdz-containing proteins at neuronal synapses. *Proceedings of the National Academy of Sciences*, 97(7):3596–3601, 2000.

C. Gautier, B. Bothorel, D. Ciocca, D. Valour, A. Gaudeau, C. Dupré, G. Lizzo, C. Brasseur, I. Riest-Fery, J.-P. Stephan, et al. Gene expression profiling during hibernation in the european hamster. *Scientific reports*, 8(1):13167, 2018.

B. Georgi, B. F. Voight, and M. Bućan. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS genetics*, 9(5):e1003484, 2013.

W. H. Gharib and M. Robinson-Rechavi. When orthologs diverge between human and mouse. *Briefings in bioinformatics*, 12(5):436–441, 2011.

J. R. Gingrich, K. A. Pelkey, S. R. Fam, Y. Huang, R. S. Petralia, R. J. Wenthold, and M. W. Salter. Unique domain anchoring of src to synaptic nmda receptors via the mitochondrial protein nadh dehydrogenase subunit 2. *Proceedings of the National Academy of Sciences*, 101(16):6237–6242, 2004.

M. R. Gleason, A. Nagiel, S. Jamet, M. Vologodskaia, H. López-Schier, and A. Hudspeth. The transmembrane inner ear (tmie) protein is essential for normal hearing and balance in the zebrafish. *Proceedings of the National Academy of Sciences*, 106(50): 21347–21352, 2009.

T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss. Rna-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986, 2008.

M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, and A. Brazma. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome biology*, 14(7):R70, 2013.

J. L. Goodson. The vertebrate social behavior network: evolutionary themes and variations. *Hormones and behavior*, 48(1):11–22, 2005.

M. Gouy, S. Guindon, and O. Gascuel. Seaview version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27(2):221–224, 2009.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. *Nature biotechnology*, 29(7):644, 2011.

S. Grant. The synapse proteome and phosphoproteome: a new paradigm for synapse biology, 2006.

S. G. Grant. Synaptopathies: diseases of the synaptome. *Current opinion in neurobiology*, 22(3):522–529, 2012.

E. K. Green, D. Grozeva, I. Jones, L. Jones, G. Kirov, S. Caesar, K. Gordon-Smith, C. Fraser, L. Forty, E. Russell, et al. The bipolar disorder risk allele at cacna1c also confers risk of recurrent major depression and of schizophrenia. *Molecular psychiatry*, 15 (10):1016, 2010.

T. R. Gregory. A bird's-eye view of the c-value enigma: genome size, cell size, and metabolic rate in the class aves. *Evolution*, 56(1):121–130, 2002.

A. Grimm, A. G. Mensah-Nyagan, and A. Eckert. Alzheimer, mitochondria and gender. *Neuroscience & Biobehavioral Reviews*, 67:89–101, 2016.

C. J. Grimmelikhuijzen and J. Westfall. The nervous systems of cnidarians. In *The nervous systems of invertebrates: an evolutionary and comparative approach*, pages 7–24. Springer, 1995.

Z. Gu, S. A. Rifkin, K. P. White, and W.-H. Li. Duplicate genes increase gene expression diversity within and between species. *Nature genetics*, 36(6):577, 2004.

B. Guo. Complex genes are preferentially retained after whole-genome duplication in teleost fish. *Journal of molecular evolution*, 84(5-6):253–258, 2017.

B. Haas and A. Papanicolaou. Transdecoder, 2012.

B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, et al. De novo transcript sequence reconstruction from rna-seq: reference generation and analysis with trinity. *Nature protocols*, 8(8), 2013.

K. F. Haas and K. Broadie. Roles of ubiquitination at the synapse. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1779(8):495–506, 2008.

M. K. Hayashi, C. Tang, C. Verpelli, R. Narayanan, M. H. Stearns, R.-M. Xu, H. Li, C. Sala, and Y. Hayashi. The postsynaptic density proteins homer and shank form a polymeric network structure. *Cell*, 137(1):159–171, 2009.

G. F. Hédou, K. Koshibu, M. Farinelli, E. Kilic, C. E. Gee, U. Kilic, K. Baumgärtel, D. M. Hermann, and I. M. Mansuy. Protein phosphatase 1-dependent bidirectional synaptic plasticity controls ischemic recovery in the adult brain. *Journal of Neuroscience*, 28(1):154–162, 2008.

S. K. Heinrich, B. Wachter, O. H. Aschenborn, S. Thalwitzer, J. Melzheimer, H. Hofer, and G. Á. Czirják. Feliform carnivores have a distinguished constitutive innate immune response. *Biology open*, 5(5):550–555, 2016.

H. Heuer and C. A. Mason. Thyroid hormone induces cerebellar purkinje cell dendritic development via the thyroid hormone receptor $\alpha 1$. *Journal of Neuroscience*, 23(33): 10604–10612, 2003.

B. R. Hopiavuori, L. D. Bennett, R. S. Brush, M. J. Van Hook, W. B. Thoreson, and R. E. Anderson. Very long-chain fatty acids support synaptic structure and function in the mammalian retina. *OCL*, 23(1):D113, 2016.

J. Horikawa and N. Suga. Biosonar signals and cerebellar auditory neurons of the mustached bat. *Journal of neurophysiology*, 55(6):1247–1267, 1986.

Y. Horiuchi, R. Kimura, N. Kato, T. Fujii, M. Seki, T. Endo, T. Kato, and K. Kawashima. Evolutional study on acetylcholine expression. *Life sciences*, 72(15):1745–1756, 2003.

S. M. Houten and R. J. Wanders. A general introduction to the biochemistry of mitochondrial fatty acid β-oxidation. *Journal of inherited metabolic disease*, 33(5):469–477, 2010.

M. A. Howard, G. M. Elias, L. A. Elias, W. Swat, and R. A. Nicoll. The role of sap97 in synaptic glutamate receptor dynamics. *Proceedings of the National Academy of Sciences*, 107(8):3805–3810, 2010.

K. Howe, M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, L. Matthews, et al. The zebrafish reference genome se-

quence and its relationship to the human genome.pdf. *Nature*, 496(7446):498–503, 2013a.

K. Howe, M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, L. Matthews, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446):498, 2013b.

L.-S. Hsu and C.-Y. Tseng. Zebrafish calcium/calmodulin-dependent protein kinase ii (cam-kii) inhibitors: Expression patterns and their roles in zebrafish brain development. *Developmental Dynamics*, 239(11):3098–3105, 2010.

S. Huggenberger, M. André, and H. H. Oelschläger. The nose of the sperm whale: overviews of functional design, structural homologies and evolution. *Journal of the Marine Biological Association of the United Kingdom*, 96(4):783–806, 2016.

A. L. Hughes and M. K. Hughes. Small genomes for better flyers. *Nature*, 377(6548): 391–391, 1995.

B. Hughes, Mark. *Nervous system.* Crash course. Edinburgh, 3rd. ed. / mark hughes, thomas miller. edition, 2007. ISBN 0723434298.

H. Husi, M. A. Ward, J. S. Choudhary, W. P. Blackstock, and S. G. Grant. Proteomic analysis of nmda receptor–adhesion protein signaling complexes. *Nature neuroscience*, 3(7):661, 2000.

B. Hwang, J. H. Lee, and D. Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):96, 2018.

Y. Ito, H. Tanaka, H. Okamoto, and T. Ohshima. Characterization of neural stem cells and their progeny in the adult zebrafish optic tectum. *Developmental biology*, 342 (1):26–38, 2010.

R. Jahn and D. Fasshauer. Molecular machines governing exocytosis of synaptic vesicles. *Nature*, 490(7419):201, 2012.

S. Jamesdaniel, B. Hu, M. H. Kermany, H. Jiang, D. Ding, D. Coling, and R. Salvi. Noise induced changes in the expression of p38/mapk signaling proteins in the sensory epithelium of the inner ear. *Journal of proteomics*, 75(2):410–424, 2011.

J. Y. Jason and M. D. Ehlers. Ubiquitin and protein turnover in synapse function. *Neuron*, 47(5):629–632, 2005.

P. H.-S. Jen and P. A. Schlegel. Neurons in the cerebellum of echolocating bats respond to acoustic signals. *Brain research*, 196(2):502–507, 1980.

T. P. Jensen, A. G. Filoteo, T. Knopfel, and R. M. Empson. Presynaptic plasma membrane ca2+ atpase isoform 2a regulates excitatory synaptic transmission in rat hippocampal ca3. *The Journal of physiology*, 579(1):85–99, 2007.

J. Ji, A. M. F. Salapatek, H. Lau, G. Wang, H. Y. Gaisano, and N. E. Diamant. Snap-25, a snare protein, inhibits two types of k+ channels in esophageal smooth muscle. *Gastroenterology*, 122(4):994–1006, 2002.

A. R. Jones, C. C. Overly, and S. M. Sunkin. The allen brain atlas: 5 years and beyond. *Nature Reviews Neuroscience*, 10(11):821, 2009.

G. Jones and E. C. Teeling. The evolution of echolocation in bats. *Trends in Ecology & Evolution*, 21(3):149–156, 2006.

I. K. Jordan, Y. I. Wolf, and E. V. Koonin. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC evolutionary biology*, 4(1):22, 2004.

Y. Kaminsky and l. Kosenko. Brain purine metabolism and xanthine dehydrogenase/oxidase conversion in hyperammonemia are under control of nmda receptors and nitric oxide. *Brain research*, 1294:193–201, 2009.

E. R. Kandel, J. H. Schwartz, T. M. Jessell, D. of Biochemistry, M. B. T. Jessell, S. Siegelbaum, and A. Hudspeth. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

C. Kanduri, P. Raijas, M. Ahvenainen, A. K. Philips, L. Ukkola-Vuoti, H. Lähdesmäki, and I. Järvelä. The effect of listening to music on human transcriptome. *PeerJ*, 3: e830, 2015.

J. Khwanmunee, L. Leelawatwattana, and P. Prapunpoj. Gene structure and evolution of transthyretin in the order chiroptera. *Genetica*, 144(1):71–83, 2016.

D. Kim, B. Langmead, and S. L. Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357–360, 2015.

E. Kim and M. Sheng. Pdz domain proteins of synapses. *Nature Reviews Neuroscience*, 5(10):771, 2004.

S. Kim, Y. S. Cho, H.-M. Kim, O. Chung, H. Kim, S. Jho, H. Seomun, J. Kim, W. Y. Bang, C. Kim, et al. Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly. *Genome biology*, 17(1):211, 2016.

S. H. Kim, R. Vlkolinsky, N. Cairns, M. Fountoulakis, and G. Lubec. The reduction of nadh: Ubiquinone oxidoreductase 24-and 75-kda subunits in brains of patients with down syndrome and alzheimer's disease. *Life sciences*, 68(24):2741–2750, 2001.

N. King. The unicellular ancestry of animal development. *Developmental cell*, 7(3): 313–325, 2004.

C. Klee, T. Crouch, and M. Krinks. Calcineurin: a calcium-and calmodulin-binding protein of the nervous system. *Proceedings of the National Academy of Sciences*, 76 (12):6270–6273, 1979.

T. Kleefstra, H. Yntema, A. Oudakker, M. Banning, V. M. Kalscheuer, J. Chelly, C. Moraine, H.-H. Ropers, J.-P. Fryns, I. Janssen, et al. Zinc finger 81 (znf81) mutations associated with x-linked mental retardation. *Journal of medical genetics*, 41 (5):394–399, 2004.

D. J. Kliebenstein. A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PloS one*, 3(3):e1838, 2008.

M. Knörnschild, M. Nagy, M. Metz, F. Mayer, and O. von Helversen. Complex vocal imitation during ontogeny in a bat. *Biology letters*, 6(2):156–159, 2010.

M. Kobayashi, T. Nakatani, T. Koda, K.-i. Matsumoto, R. Ozaki, N. Mochida, K. Takao, T. Miyakawa, and I. Matsuoka. Absence of brinp1 in mice causes increase of hippocampal neurogenesis and behavioral alterations relevant to human psychiatric disorders. *Molecular brain*, 7(1):12, 2014.

B. Kolb and I. Q. Whishaw. *An introduction to brain and behavior*. Worth Publishers, 2001.

L. L. Kordonowy and M. D. MacManes. Characterization of a male reproductive transcriptome for peromyscus eremicus (cactus mouse). *PeerJ*, 4:e2617, 2016.

L. Koren and E. Geffen. Individual identity is communicated through multiple pathways in male rock hyrax (procavia capensis) songs. *Behavioral Ecology and sociobiology*, 65(4):675–684, 2011.

H.-C. Kornau, L. T. Schenker, M. B. Kennedy, and P. H. Seeburg. Domain interaction between nmda receptor subunits and the postsynaptic density protein psd-95. *Science*, 269(5231):1737–1740, 1995.

E. Koropouli and A. L. Kolodkin. Semaphorins and the dynamic regulation of synapse assembly, refinement, and function. *Current opinion in neurobiology*, 27:1–7, 2014.

T. Koshiba, N. Bashiruddin, and S. Kawabata. Mitochondria and antiviral innate immunity. *International journal of biochemistry and molecular biology*, 2(3):257, 2011.

K. S. Kosik. Exploring the early origins of the synapse by comparative genomics. *Biology letters*, 5(1):108–111, 2009.

E. Kratz, J. C. Dugas, and J. Ngai. Odorant receptor gene regulation: implications from genomic organization. *TRENDS in Genetics*, 18(1):29–34, 2002.

R. Lamprecht and J. LeDoux. Structural plasticity and memory. *Nature Reviews Neuroscience*, 5(1):45, 2004.

R. J. Langham, J. Walsh, M. Dunn, C. Ko, S. A. Goff, and M. Freeling. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics*, 166(2):935–945, 2004.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357, 2012.

S. L. Leber, I. C. Llenos, C. L. Miller, J. R. Dulay, J. Haybaeck, and S. Weis. Homer1a protein expression in schizophrenia, bipolar disorder, and major depression. *Journal of Neural Transmission*, 124(10):1261–1273, 2017.

A. Lee, Y. Hirabayashi, S.-K. Kwon, T. L. Lewis, and F. Polleux. Emerging roles of mitochondria in synaptic transmission and neurodegeneration. *Current opinion in physiology*, 2018.

A. K. Lee, K. A. Kulcsar, O. Elliott, H. Khiabanian, E. R. Nagle, M. E. Jones, B. R. Amman, M. Sanchez-Lockhart, J. S. Towner, G. Palacios, et al. De novo transcriptome reconstruction and annotation of the egyptian rousette bat. *BMC genomics*, 16(1):1033, 2015.

M. Lei, D. Dong, S. Mu, Y.-H. Pan, and S. Zhang. Comparison of brain transcriptome of the greater horseshoe bats (rhinolophus ferrumequinum) in active and torpid episodes. *PLoS One*, 9(9):e107746, 2014.

G. Li, J. Wang, S. J. Rossiter, G. Jones, J. A. Cotton, and S. Zhang. The hearing gene prestin reunites echolocating bats. *Proceedings of the National Academy of Sciences*, 2008.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

J. Li, Y. Liu, T. Kim, R. Min, and Z. Zhang. Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS computational biology*, 6(8):e1000910, 2010a.

W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

Y. Li, Z. Liu, P. Shi, and J. Zhang. The hearing gene prestin unites echolocating bats and whales. *Current Biology*, 20(2):R55–R56, 2010b.

M. C. Liberman, J. Gao, D. Z. He, X. Wu, S. Jia, and J. Zuo. Prestin is required for electromotility of the outer hair cell and for the cochlear amplifier. *Nature*, 419(6904):300, 2002.

H.-y. Lin, Q. Liu, X. Li, J. Yang, S. Liu, Y. Huang, M. J. Scanlon, D. Nettleton, and P. S. Schnable. Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by erd-gwas. *Genome biology*, 18 (1):192, 2017.

M. G. Lin and J. H. Hurley. Structure and function of the ulk1 complex in autophagy. *Current opinion in cell biology*, 39:61–68, 2016.

M. M. Lipinski, B. Zheng, T. Lu, Z. Yan, B. F. Py, A. Ng, R. J. Xavier, C. Li, B. A. Yankner, C. R. Scherzer, et al. Genome-wide analysis reveals mechanisms modulating autophagy in normal brain aging and in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 107(32):14164–14169, 2010.

J. Lisman, H. Schulman, and H. Cline. The molecular basis of camkii function in synaptic and behavioural memory. *Nature Reviews Neuroscience*, 3(3):175, 2002.

X. Liu, R. Heidelberger, and R. Janz. Phosphorylation of syntaxin 3b by camkii regulates the formation of t-snare complexes. *Molecular and Cellular Neuroscience*, 60:53–62, 2014.

Y. Liu, N. Han, L. F. Franchini, H. Xu, F. Pisciottano, A. B. Elgoyhen, K. E. Rajan, and S. Zhang. The voltage-gated potassium channel subfamily kqt member 4 (kcnq4) displays parallel evolution in echolocating bats. *Molecular biology and evolution*, 29 (5):1441–1450, 2011.

C. Lüscher, R. A. Nicoll, R. C. Malenka, and D. Muller. Synaptic plasticity and dynamic modulation of the postsynaptic membrane. *Nature neuroscience*, 3(6):545, 2000.

C. V. Ly and P. Verstreken. Mitochondria at the synapse. *The Neuroscientist*, 12(4):291–299, 2006.

W.-R. Ma and J. Zhang. Jag1b is essential for patterning inner ear sensory cristae by regulating anterior morphogenetic tissue separation and preventing posterior cell death. *Development*, 142(4):763–773, 2015.

C. MacKintosh and D. E. Ferrier. Recent advances in understanding the roles of whole genome duplications in evolution. *F1000Research*, 6, 2017.

O. Madsen, M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. Parallel adaptive radiations in two major clades of placental mammals. *Nature*, 409(6820):610, 2001.

P. Madsen, M. Johnson, N. A. De Soto, W. Zimmer, and P. Tyack. Biosonar performance of foraging beaked whales (mesoplodon densirostris). *Journal of Experimental Biology*, 208(2):181–194, 2005.

J. Mamrot, R. Legaie, S. J. Ellery, T. Wilson, T. Seemann, D. R. Powell, D. K. Gardner, D. W. Walker, P. Temple-Smith, A. T. Papenfuss, et al. De novo transcriptome assembly for the spiny mouse (acomys cahirinus). *Scientific Reports*, 7, 2017.

I. M. Mansuy. Calcineurin in memory and bidirectional plasticity. *Biochemical and biophysical research communications*, 311(4):1195–1208, 2003.

J. C. Mar, N. A. Matigian, A. Mackay-Sim, G. D. Mellick, C. M. Sue, P. A. Silburn, J. J. McGrath, J. Quackenbush, and C. A. Wells. Variance of gene expression identifies altered network constraints in neurological disease. *PLoS genetics*, 7(8):e1002207, 2011.

W. Marcotti, A. Erven, S. L. Johnson, K. P. Steel, and C. J. Kros. Tmc1 is necessary for normal functional maturation and survival of inner and outer hair cells in the mouse cochlea. *The Journal of physiology*, 574(3):677–698, 2006.

L. Marino, J. K. Rilling, S. K. Lin, and S. H. Ridgway. Relative volume of the cerebellum in dolphins and comparison with anthropoid primates. *Brain, Behavior and Evolution*, 56(4):204–211, 2000.

A. G. Marshal. Bats, flowers and fruit: evolutionary relationships in the old world. *Biological journal of the Linnean Society*, 20(1):115–135, 1983.

V. Martella, M. Campolo, E. Lorusso, P. Cavicchio, M. Camero, A. L. Bellacicco, N. Decaro, G. Elia, G. Greco, M. Corrente, et al. Norovirus in captive lion cub (panthera leo). *Emerging infectious diseases*, 13(7):1071, 2007.

M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

I. Maurer, S. Zierz, and H.-J. Möller. A selective defect of cytochrome c oxidase is present in brain of alzheimer disease patients. *Neurobiology of aging*, 21(3):455–462, 2000.

C. McCann, E. E. Holohan, S. Das, A. Dervan, A. Larkin, J. A. Lee, V. Rodrigues, R. Parker, and M. Ramaswami. The ataxin-2 protein is required for microrna function and synapse-specific long-term olfactory habituation. *Proceedings of the National Academy of Sciences*, 108(36):E655–E662, 2011.

L. K. McCorry. Physiology of the autonomic nervous system. *American journal of pharmaceutical education*, 71(4):78, 2007.

R. R. McWilliams, E. Gidey, L. Fouassier, R. B. DOCTOR, et al. Characterization of an ankyrin repeat-containing shank2 isoform (shank2e) in liver epithelial cells. *Biochemical Journal*, 380(1):181–191, 2004.

A. Merched, J.-M. Serot, S. Visvikis, D. Aguillon, G. Faure, and G. Siest. Apolipoprotein e, transthyretin and actin in the csf of alzheimer's patients: relation with the senile plaques and cytoskeleton biochemistry. *Febs letters*, 425(2):225–228, 1998.

A. Meyer. Developmental biology: Hox gene variation and evolution. *Nature*, 391 (6664):225, 1998.

M. Miljkovic-Licina, D. Gauchat, and B. Galliot. Neuronal evolution: analysis of regulatory genes in a first-evolved nervous system, the hydra nervous system. *Biosystems*, 76(1-3):75–87, 2004.

N. Miyasaka, A. A. Wanner, J. Li, J. Mack-Bucher, C. Genoud, Y. Yoshihara, and R. W. Friedrich. Functional development of the olfactory system in zebrafish. *Mechanisms of development*, 130(6-8):336–346, 2013.

N. Miyasaka, I. Arganda-Carreras, N. Wakisaka, M. Masuda, U. Sümbül, H. S. Seung, and Y. Yoshihara. Olfactory projectome in the zebrafish forebrain revealed by genetic single-neuron labelling. *Nature Communications*, 5:3639, 2014.

N. Mizushima. The role of the atg1/ulk1 complex in autophagy regulation. *Current opinion in cell biology*, 22(2):132–139, 2010.

C. B. Moens and V. E. Prince. Constructing the hindbrain: insights from the zebrafish. *Developmental dynamics*, 224(1):1–17, 2002.

W. Morishita, J. H. Connor, H. Xia, E. M. Quinlan, S. Shenolikar, and R. C. Malenka.

Regulation of synaptic strength by protein phosphatase 1. *Neuron*, 32(6):1133–1148, 2001.

L. K. Mosavi, T. J. Cammett, D. C. Desrosiers, and Z.-y. Peng. The ankyrin repeat as molecular architecture for protein recognition. *Protein Science*, 13(6):1435–1448, 2004.

F. J. Moss, P. Viard, A. Davies, F. Bertaso, K. M. Page, A. Graham, C. Cantí, M. Plumpton, C. Plumpton, J. J. Clare, et al. The novel product of a five-exon stargazin-related gene abolishes cav2. 2 calcium channel expression. *The EMBO journal*, 21(7):1514–1523, 2002.

M. Motomiya, M. Kobayashi, N. Iwasaki, A. Minami, and I. Matsuoka. Activity-dependent regulation of brinp family genes. *Biochemical and biophysical research communications*, 352(3):623–629, 2007.

A. Muñoz-Garcia and J. B. Williams. Basal metabolic rate in carnivores is associated with diet after controlling for phylogeny. *Physiological and biochemical Zoology*, 78 (6):1039–1056, 2005.

R. P. Munton, S. Vizi, and I. M. Mansuy. The role of protein phosphatase-1 in the modulation of synaptic and structural plasticity. *FEBS letters*, 567(1):121–128, 2004.

W. J. Murphy, E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'brien. Molecular phylogenetics and the origins of placental mammals. *Nature*, 409(6820):614, 2001.

T. Nagata, M. Koyanagi, R. Lucas, and A. Terakita. An all-trans-retinal-binding opsin peropsin as a potential dark-active and light-inactivated g protein-coupled receptor. *Scientific reports*, 8(1):3535, 2018.

S. Naisbitt, E. Kim, J. C. Tu, B. Xiao, C. Sala, J. Valtschanoff, R. J. Weinberg, P. F. Worley, and M. Sheng. Shank, a novel family of postsynaptic density proteins that binds

to the nmda receptor/psd-95/gkap complex and cortactin. *Neuron*, 23(3):569–582, 1999.

H. S. Najafabadi, S. Mnaimneh, F. W. Schmitges, M. Garton, K. N. Lam, A. Yang, M. Albu, M. T. Weirauch, E. Radovani, P. M. Kim, et al. C2h2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature biotechnology*, 33(5):555, 2015.

T. Nakatani, S. Ueno, N. Mori, and I. Matsuoka. Role of nrsf/rest in the molecular mechanisms regulating neural-specific expression of trkc/neurotrophin-3 receptor gene. *Molecular brain research*, 135(1-2):249–259, 2005.

E. P. Nawrocki, S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones, J. Tate, and R. D. Finn. Rfam 12.0: updates to the rna families database. *Nucleic Acids Research*, 43(D1):D130–D137, 2015. doi: 10.1093/nar/gku1063. URL +http://dx.doi.org/10.1093/nar/gku1063.

J. Neddens and A. Buonanno. Expression of the neuregulin receptor erbb4 in the brain of the rhesus monkey (macaca mulatta). *PLoS One*, 6(11):e27337, 2011.

J. H. Ng, M. Tachedjian, J. Deakin, J. W. Wynne, J. Cui, V. Haring, I. Broz, H. Chen, K. Belov, L.-F. Wang, et al. Evolution and comparative analysis of the bat mhc-i region. *Scientific reports*, 6:21256, 2016.

J. H. Ng, M. Tachedjian, L.-F. Wang, and M. L. Baker. Insights into the ancestral organisation of the mammalian mhc class ii region from the genome of the pteropid bat, pteropus alecto. *BMC genomics*, 18(1):388, 2017.

D. G. Nicholls and S. L. Budd. Mitochondria and neuronal survival. *Physiological reviews*, 80(1):315–360, 2000.

L. Nie. Mutations of kcnq4 channels associated with nonsyndromic progressive sensorineural hearing loss. *Current opinion in otolaryngology & head and neck surgery*, 16(5):441, 2008.

M. Niethammer, E. Kim, and M. Sheng. Interaction between the c terminus of nmda receptor subunits and multiple members of the psd-95 family of membrane-associated guanylate kinases. *Journal of Neuroscience*, 16(7):2157–2163, 1996.

R. Nieuwenhuys, J. Hans, and C. Nicholson. *The central nervous system of vertebrates*. Springer, 2014.

J. Nithianantharajah, N. H. Komiyama, A. McKechanie, M. Johnstone, D. H. Blackwood, D. St Clair, R. D. Emes, L. N. Van De Lagemaat, L. M. Saksida, T. J. Bussey, et al. Synaptic scaffold evolution generated components of vertebrate cognitive complexity. *Nature neuroscience*, 16(1):16, 2013.

R. A. Nixon. The role of autophagy in neurodegenerative disease. *Nature medicine*, 19 (8):983, 2013.

U. M. Norberg and J. M. Rayner. Ecological morphology and flight in bats (mammalia; chiroptera): wing adaptations, flight performance, foraging strategy and echolocation. *Phil. Trans. R. Soc. Lond. B*, 316(1179):335–427, 1987.

R. Nouvian, J. Neef, A. V. Bulankina, E. Reisinger, T. Pangršič, T. Frank, S. Sikorra, N. Brose, T. Binz, and T. Moser. Exocytosis at the hair cell ribbon synapse apparently operates without neuronal snare proteins. *Nature neuroscience*, 14(4):411, 2011.

K. P. O'Brien, M. Remm, and E. L. Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research*, 33(suppl_1):D476–D480, 2005.

D. H. O'Day and M. A. Myre. Calmodulin-binding domains in alzheimer's disease proteins: extending the calcium hypothesis. *Biochemical and biophysical research communications*, 320(4):1051–1054, 2004.

T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, and M. Aono. Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nature materials*, 10(8):591, 2011.

N. D. Okerlund and B. N. Cheyette. Synaptic wnt signalingâĂŤa contributor to major psychiatric disorders? *Journal of neurodevelopmental disorders*, 3(2):162, 2011.

C. Oliva, P. Escobedo, C. Astorga, C. Molina, and J. Sierralta. Role of the maguk protein family in synapse formation and function. *Developmental neurobiology*, 72(1):57–72, 2012.

N. A. O'Rourke, N. C. Weiler, K. D. Micheva, and S. J. Smith. Deep molecular diversity of mammalian synapses: why it matters and how to measure it. *Nature Reviews Neuroscience*, 13(6):365, 2012.

C. Packer, R. Kock, and M. J. Appel'IITI. lions (panthera leo). *Nature*, 379:1, 1996.

B. Pan, G. S. Géléoc, Y. Asai, G. C. Horwitz, K. Kurima, K. Ishikawa, Y. Kawashima, A. J. Griffith, and J. R. Holt. Tmc1 and tmc2 are components of the mechanotransduction channel in hair cells of the mammalian inner ear. *Neuron*, 79(3):504–515, 2013.

A. Panov, Z. Orynbayeva, V. Vavilin, and V. Lyakhovich. Fatty acids in energy metabolism of the central nervous system. *BioMed research international*, 2014, 2014.

A. Pascual-Leone, C. Freitas, L. Oberman, J. C. Horvath, M. Halko, M. Eldaief, S. Bashir, M. Vernet, M. Shafi, B. Westover, et al. Characterizing brain cortical plasticity and network dynamics across the age-span in health and disease with tms-eeg and tms-fmri. *Brain topography*, 24(3-4):302, 2011.

R. Patro, G. Duggal, and C. Kingsford. Salmon: accurate, versatile and ultrafast quantification from rna-seq data using lightweight-alignment. *bioRxiv*, page 021592, 2015.

F. Paumet, J. Wesolowski, A. Garcia-Diaz, C. Delevoye, N. Aulner, H. A. Shuman, A. Subtil, and J. E. Rothman. Intracellular bacteria encode inhibitory snare-like proteins. *PloS one*, 4(10):e7375, 2009.

I. Pelassa, C. Zhao, M. Pasche, B. Odermatt, and L. Lagnado. Synaptic vesicles are

âĂIJprimedâĂİ for fast clathrin-mediated endocytosis at the ribbon synapse. *Frontiers in molecular neuroscience*, 7:91, 2014.

M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, 33(3):290–295, 2015.

M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg. Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nature protocols*, 11(9):1650–1667, 2016.

A. E. Peters, A. Bavishi, H. Cho, and M. Choudhary. Evolutionary constraints and expression analysis of gene duplications in rhodobacter sphaeroides 2.4. 1. *BMC research notes*, 5(1):192, 2012.

M. Picard and B. S. McEwen. Mitochondria impact brain function and cognition. *Proceedings of the National Academy of Sciences*, 111(1):7–8, 2014.

A. Pocklington, J. Armstrong, and S. Grant. Organization of brain complexityâĂŤsynapse proteome form and function. *Briefings in Functional Genomics*, 5(1):66–73, 2006.

F. Polleux, G. Ince-Dunn, and A. Ghosh. Transcriptional regulation of vertebrate axon guidance and synapse formation. *Nature Reviews Neuroscience*, 8(5):331, 2007.

J. U. Pontius, J. C. Mullikin, D. R. Smith, A. S. Team, K. Lindblad-Toh, S. Gnerre, M. Clamp, J. Chang, R. Stephens, B. Neelam, et al. Initial sequence and comparative analysis of the cat genome. *Genome research*, 17(11):1675–1689, 2007.

Y. Prat, M. Taub, and Y. Yovel. Vocal learning in a social mammal: Demonstrated by isolation and playback experiments in bats. *Science Advances*, 1(2):e1500019, 2015.

G. R. Prescott and L. H. Chamberlain. Regional and developmental brain expression patterns of snap25 splice variants. *BMC neuroscience*, 12(1):35, 2011.

S. R. Proulx. Multiple routes to subfunctionalization and gene duplicate specialization. *Genetics*, pages genetics–111, 2011.

A. L. Purcell and T. J. Carew. Tyrosine kinases, synaptic plasticity and memory: insights from vertebrates and invertebrates. *Trends in neurosciences*, 26(11):625–630, 2003.

J. S. Racine. Rstudio: A platform-independent ide for r and sweave. *Journal of Applied Econometrics*, 27(1):167–172, 2012.

M. Raiteri. Presynaptic autoreceptors: Mini-review. *Journal of neurochemistry*, 78(4): 673–675, 2001.

A. Raj, S. A. Rifkin, E. Andersen, and A. Van Oudenaarden. Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913, 2010.

G. M. Ramakers, P. Pasinelli, J. J. Hens, W. H. Gispen, and P. N. De Graan. Protein kinase c in synaptic plasticity: changes in the in situ phosphorylation state of identified pre- and postsynaptic substrates. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 21(3):455–486, 1997.

N. A. Ramakrishnan, M. J. Drescher, and D. G. Drescher. Direct interaction of otoferlin with syntaxin 1a, snap-25, and the l-type voltage-gated calcium channel cav1. 3. *Journal of Biological Chemistry*, 284(3):1364–1372, 2009.

A. Rambaut and A. Drummond. Figtree: Tree figure drawing tool, version 1.2. 2. *Institute of Evolutionary Biology, University of Edinburgh*, 2008.

D. Ran and Z. J. Daye. Gene expression variability and the analysis of large-scale rna-seq studies with the mdseq. *Nucleic acids research*, 45(13):e127–e127, 2017.

V. Rangaraju, S. tom Dieck, and E. M. Schuman. Local translation in neuronal compartments: how local is local? *EMBO reports*, 18(5):693–711, 2017.

M. L. Reese, S. Dakoji, D. S. Bredt, and V. Dötsch. The guanylate kinase domain of the maguk psd-95 binds dynamically to a conserved motif in map1a. *Nature Structural and Molecular Biology*, 14(2):155, 2007.

A. Reichelt, R. Rodgers, and S. Clapcote. The role of neurexins in schizophrenia and autistic spectrum disorder. *Neuropharmacology*, 62(3):1519–1526, 2012.

A. Represa, J. C. Deloulme, M. Sensenbrenner, Y. Ben-Ari, and J. Baudier. Neurogranin: immunocytochemical localization of a brain-specific protein kinase c substrate. *Journal of Neuroscience*, 10(12):3782–3792, 1990.

A. Rissone, M. Monopoli, M. Beltrame, F. Bussolino, F. Cotelli, and M. Arese. Comparative genome analysis of the neurexin gene family in danio rerio: insights into their functions and evolution. *Molecular biology and evolution*, 24(1):236–252, 2006.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26 (1):139–140, 2010.

P. Rodenas-Cuadrado, X. S. Chen, L. Wiegrebe, U. Firzlaff, and S. C. Vernes. A novel approach identifies the first transcriptome networks in bats: a new genetic model for vocal communication. *BMC genomics*, 16(1):836, 2015.

S. C. Rothschild, J. A. Lister, and R. M. Tombes. Differential expression of camk-ii genes during early zebrafish embryogenesis. *Developmental Dynamics*, 236(1):295–305, 2007.

S. C. Rothschild, J. Lahvic, L. Francescatto, J. J. McLeod, S. M. Burgess, and R. M. Tombes. Camk-ii activation is essential for zebrafish inner ear development and acts through delta–notch signaling. *Developmental biology*, 381(1):179–188, 2013.

E. A. Rowland, C. K. Snowden, and I. M. Cristea. Protein lipoylation: an evolutionarily conserved metabolic regulator of health and disease. *Current opinion in chemical biology*, 42:76–85, 2018.

M. Roy, O. Sorokina, N. Skene, C. Simonnet, F. Mazzo, R. Zwart, E. Sher, C. Smith, J. D. Armstrong, and S. G. Grant. Proteomic analysis of postsynaptic proteins in regions of the human neocortex. *Nature neuroscience*, 21(1):130, 2018.

I. Ruiz-Trillo, G. Burger, P. W. Holland, N. King, B. F. Lang, A. J. Roger, and M. W. Gray. The origins of multicellularity: a multi-taxon genome initiative. *TRENDS in Genetics*, 23(3):113–118, 2007.

T. J. Ryan and S. G. Grant. The origin and evolution of synapses. *Nature Reviews Neuroscience*, 10(10):701, 2009.

P. C. Sabeti, S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. Mikkelsen, D. Altshuler, and E. Lander. Positive natural selection in the human lineage. *science*, 312(5780):1614–1620, 2006.

O. Sakarya, K. A. Armstrong, M. Adamska, M. Adamski, I.-F. Wang, B. Tidor, B. M. Degnan, T. H. Oakley, and K. S. Kosik. A post-synaptic scaffold at the origin of the animal kingdom. *PloS one*, 2(6):e506, 2007.

C. Sala, V. Piëch, N. R. Wilson, M. Passafaro, G. Liu, and M. Sheng. Regulation of dendritic spine morphology and synaptic function by shank and homer. *Neuron*, 31(1): 115–130, 2001.

P. Scheiffele, J. Fan, J. Choih, R. Fetter, and T. Serafini. Neuroligin expressed in nonneuronal cells triggers presynaptic development in contacting axons. *Cell*, 101(6): 657–669, 2000.

T. Schermerhorn. Normal glucose metabolism in carnivores overlaps with diabetes pathology in non-carnivores. *Frontiers in endocrinology*, 4:188, 2013.

E. Schnell, M. Sizemore, S. Karimzadegan, L. Chen, D. S. Bredt, and R. A. Nicoll. Direct interactions between psd-95 and stargazin control synaptic ampa receptor number. *Proceedings of the National Academy of Sciences*, 99(21):13902–13907, 2002.

P. Schönfeld and G. Reiser. Why does brain metabolism not favor burning of fatty acids to provide energy?-reflections on disadvantages of the use of free fatty acids as fuel for brain. *Journal of Cerebral Blood Flow & Metabolism*, 33(10):1493–1499, 2013.

G. Schuller and S. Radtke-Schuller. Neural control of vocalization in bats: mapping of brainstem areas with electrical microstimulation eliciting species-specific echolocation calls in the rufous horseshoe bat. *Experimental brain research*, 79(1):192–206, 1990.

E. M. Schuman, J. L. Dynes, and O. Steward. Synaptic regulation of translation of dendritic mrnas. *Journal of Neuroscience*, 26(27):7143–7146, 2006.

C. P. Schwartz and M. S. Smotherman. Mapping vocalization-related immediate early gene expression in echolocating bats. *Behavioural brain research*, 224(2):358–368, 2011.

C. Scott. dammit: an open and accessible de novo transcriptome annotator. *in prep.*, 2016. URL www.camillescott.org/dammit.

I. Seim, X. Fang, Z. Xiong, A. V. Lobanov, Z. Huang, S. Ma, Y. Feng, A. A. Turanov, Y. Zhu, T. L. Lenz, et al. Genome analysis reveals insights into physiology and longevity of the brandtâĂŹs bat myotis brandtii. *Nature communications*, 4:2212, 2013.

Y. J. Seo, H. M. Ju, S. H. Lee, S. H. Kwak, M. J. Kang, J.-H. Yoon, C.-H. Kim, and H.-J. Cho. Damage of inner ear sensory hair cells via mitochondrial loss in a murine model of sleep apnea with chronic intermittent hypoxia. *Sleep*, 40(9), 2017.

C. F. Sephton and G. Yu. The function of rna-binding proteins at the synapse: implications for neurodegeneration. *Cellular and molecular life sciences*, 72(19):3621–3635, 2015.

J. Serot, D. Christmann, T. Dubost, and M. Couturier. Cerebrospinal fluid transthyretin: aging and late onset alzheimerâĂŹs disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 63(4):506–508, 1997.

Y. Sha, J. H. Phan, and M. D. Wang. Effect of low-expression gene filtering on detection of differentially expressed genes in rna-seq data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6461–6464. IEEE, 2015.

M. D. Shapiro, M. E. Marks, C. L. Peichel, B. K. Blackman, K. S. Nereng, B. Jónsson, D. Schluter, and D. M. Kingsley. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, 428(6984):717, 2004.

K. Sharma, S. Schmitt, C. G. Bergner, S. Tyanova, N. Kannaiyan, N. Manrique-Hoyos, K. Kongi, L. Cantuti, U.-K. Hanisch, M.-A. Philips, et al. Cell type–and brain region–resolved mouse brain proteome. *Nature neuroscience*, 18(12):1819, 2015.

T. I. Shaw, A. Srivastava, W.-C. Chou, L. Liu, A. Hawkinson, T. C. Glenn, R. Adams, and T. Schountz. Transcriptome sequencing and annotation for the jamaican fruit bat (artibeus jamaicensis). *PloS one*, 7(11):e48472, 2012.

Y.-Y. Shen, L. Liang, Z.-H. Zhu, W.-P. Zhou, D. M. Irwin, and Y.-P. Zhang. Adaptive evolution of energy metabolism genes and the origin of flight in bats. *Proceedings of the National Academy of Sciences*, 107(19):8666–8671, 2010.

Y.-Y. Shen, L. Liang, G.-S. Li, R. W. Murphy, and Y.-P. Zhang. Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genetics*, 8(6):e1002788, 2012.

M. Sheng and E. Kim. The shank family of scaffold proteins. *J Cell Sci*, 113(11):1851–1856, 2000.

M. Sheng and M. J. Kim. Postsynaptic signaling and plasticity mechanisms. *Science*, 298(5594):776–780, 2002.

Y. Shi and D. M. Holtzman. Interplay between innate immunity and alzheimer disease: Apoe and trem2 in the spotlight. *Nature Reviews Immunology*, page 1, 2018.

Y. Shiraishi-Yamaguchi and T. Furuichi. The homer family proteins. *Genome biology*, 8 (2):206, 2007.

V. Shoshan-Barmatz, V. De Pinto, M. Zweckstetter, Z. Raviv, N. Keinan, and N. Arbel. Vdac, a multi-functional mitochondrial protein regulating cell life and death. *Molecular aspects of medicine*, 31(3):227–285, 2010.

J. Siemens, C. Lillo, R. A. Dumont, A. Reynolds, D. S. Williams, P. G. Gillespie, and U. Müller. Cadherin 23 is a component of the tip link in hair-cell stereocilia. *Nature*, 428(6986):950, 2004.

F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015.

B. A. Simms and G. W. Zamponi. Neuronal voltage-gated calcium channels: structure, function, and dysfunction. *Neuron*, 82(1):24–45, 2014.

L. P. Slomnicki, M. Pietrzak, A. Vashishta, J. Jones, N. Lynch, S. Elliot, E. Poulos, D. Malicote, B. E. Morris, J. Hallgren, et al. Requirement of neuronal ribosome synthesis for growth and maintenance of the dendritic tree. *Journal of Biological Chemistry*, pages jbc–M115, 2016.

D. H. Small. Dysregulation of calcium homeostasis in alzheimerâĂŹs disease. *Neurochemical research*, 34(10):1824–1829, 2009.

J. D. Smith and T. R. Gregory. The genome sizes of megabats (chiroptera: Pteropodidae) are remarkably constrained. *Biology letters*, 5(3):347–351, 2009.

R. Smith-Unna, C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly. Transrate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, 26 (8):1134–1144, 2016.

S. G. Snowden, A. A. Ebshiana, A. Hye, Y. An, O. Pletnikova, R. OâĂŹBrien, J. Troncoso, C. Legido-Quigley, and M. Thambisetty. Association between fatty acid metabolism in the brain and alzheimer disease neuropathology and cognitive performance: A nontargeted metabolomic study. *PLoS medicine*, 14(3):e1002266, 2017.

S. F. Sorrells, M. F. Paredes, A. Cebrian-Silla, K. Sandoval, D. Qi, K. W. Kelley, D. James, S. Mayer, J. Chang, K. I. Auguste, et al. Human hippocampal neurogenesis drops sharply in children to undetectable levels in adults. *Nature*, 555(7696):377, 2018.

J. C. Sousa, I. Cardoso, F. Marques, M. J. Saraiva, and J. A. Palha. Transthyretin and alzheimer's disease: where in the brain? *Neurobiology of aging*, 28(5):713–718, 2007.

N. C. Spitzer. Electrical activity in early neuronal development. *Nature*, 444(7120):707, 2006.

R. Sprengel, B. Suchanek, C. Amico, R. Brusa, N. Burnashev, A. Rozov, Ø. Hvalby, V. Jensen, O. Paulsen, P. Andersen, et al. Importance of the intracellular domain of nr2 subunits for nmda receptor function in vivo. *Cell*, 92(2):279–289, 1998.

D. G. Stathakis, K. B. Hoover, Z. You, and P. J. Bryant. Human postsynaptic density-95 (psd95): location of the gene (dlg4) and possible function in nonneural as well as in neural tissues. *Genomics*, 44(1):71–82, 1997.

A. Stengel, H. Karasawa, and Y. Taché. The role of brain somatostatin receptor 2 in the regulation of feeding and drinking behavior. *Hormones and behavior*, 73:15–22, 2015.

K. B. Storey. Metabolic regulation in mammalian hibernation: enzyme and protein adaptations. *Comparative Biochemistry and Physiology Part A: Physiology*, 118(4): 1115–1124, 1997.

S. Strack, S. Choi, D. M. Lovinger, and R. J. Colbran. Translocation of autophosphorylated calcium/calmodulin-dependent protein kinase ii to the postsynaptic density. *Journal of Biological Chemistry*, 272(21):13467–13470, 1997.

A. D. Strand, A. K. Aragaki, Z. C. Baquet, A. Hodges, P. Cunningham, P. Holmans, K. R. Jones, L. Jones, C. Kooperberg, and J. M. Olson. Conservation of regional gene expression in mouse and human brain. *PLoS genetics*, 3(4):e59, 2007.

C.-H. Su, W.-Y. Tarn, et al. Alternative splicing in neurogenesis and brain development. *Frontiers in molecular biosciences*, 5:12, 2018.

E. C. Suárez-Castillo and J. E. García-Arrarás. Molecular evolution of the ependymin protein family: a necessary update. *BMC evolutionary biology*, 7(1):23, 2007.

T. C. Südhof. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature*, 455(7215):903, 2008.

D. Sulzer and D. J. Surmeier. Neuronal vulnerability, pathogenesis, and parkinson's disease. *Movement Disorders*, 28(6):715–724, 2013.

Y. Sun, D. Hu, J. Liang, Y.-P. Bao, S.-Q. Meng, L. Lu, and J. Shi. Association between variants of zinc finger genes and psychiatric disorders: systematic review and meta-analysis. *Schizophrenia research*, 162(1-3):124–137, 2015.

B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007. doi: 10.1093/bioinformatics/btm098. URL +http://dx.doi.org/10.1093/bioinformatics/btm098.

E. C. Teeling. Hear, hear: the convergent evolution of echolocation in bats? *Trends in Ecology & Evolution*, 24(7):351–354, 2009.

E. C. Teeling, G. Jones, and S. J. Rossiter. Phylogeny, genes, and hearing: Implications for the evolution of echolocation in bats. In *Bat Bioacoustics*, pages 25–54. Springer, 2016.

H. Teng, W. Cai, L. Zhou, J. Zhang, Q. Liu, Y. Wang, W. Dai, M. Zhao, and Z. Sun. Evolutionary mode and functional divergence of vertebrate nmda receptor subunit 2 genes. *PloS one*, 5(10):e13342, 2010.

J. Thiagavel, C. Cechetto, S. E. Santana, L. Jakobsen, E. J. Warrant, and J. M. Ratcliffe. Auditory opportunity and visual constraint enabled the evolution of echolocation in bats. *Nature communications*, 9(1):98, 2018.

G. M. Thomas and R. L. Huganir. Mapk cascade signalling and synaptic plasticity. *Nature Reviews Neuroscience*, 5(3):173, 2004.

S. N. Thomas and A. J. Yang. Mass spectrometry analysis of lysine posttranslational modifications of tau protein from alzheimerâĂŹs disease brain. In *Tau Protein*, pages 161–177. Springer, 2017.

P. B. Tran and R. J. Miller. Chemokine receptors: signposts to brain development and disease. *Nature Reviews Neuroscience*, 4(6):444, 2003.

C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

A. Turner. *The big cats and their fossil relatives: an illustrated guide to their evolution and natural history*. Columbia University Press, 2000.

N. Tuteja. Signaling through g protein coupled receptors. *Plant signaling & behavior*, 4(10):942–947, 2009.

H. Vacher, D. P. Mohapatra, and J. S. Trimmer. Localization and targeting of voltage-dependent ion channels in mammalian central neurons. *Physiological reviews*, 88 (4):1407–1447, 2008.

J. G. Valtschanoff and R. J. Weinberg. Laminar organization of the nmda receptor complex within the postsynaptic density. *Journal of Neuroscience*, 21(4):1211–1217, 2001.

M. Van Duijn, F. Keijzer, and D. Franken. Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adaptive Behavior*, 14(2):157–170, 2006.

R. Varga, P. M. Kelley, B. J. Keats, A. Starr, S. M. Leal, E. Cohn, and W. J. Kimberling. Non-syndromic recessive auditory neuropathy is the result of mutations in the otoferlin (otof) gene. *Journal of medical genetics*, 40(1):45–50, 2003.

C. Verpelli, M. J. Schmeisser, C. Sala, and T. M. Boeckers. Scaffold proteins at the post-synaptic density. In *Synaptic Plasticity*, pages 29–61. Springer, 2012.

J. Vesa, E. Hellsten, L. A. Verkruyse, L. A. Camp, J. Rapola, P. Santavuori, S. L. Hofmann, and L. Peltonen. Mutations in the palmitoyl protein thioesterase gene causing infantile neuronal ceroid lipofuscinosis. *Nature*, 376(6541):584, 1995.

D. Villela, C. K. Suemoto, C. A. Pasqualucci, L. T. Grinberg, and C. Rosenberg. Do copy number changes in cacna2d2, cacna2d3, and cacna1d constitute a predisposing risk factor for alzheimerâĂŹs disease? *Frontiers in genetics*, 7:107, 2016.

N. Volfovsky, T. K. Oleksyk, K. C. Cruz, A. L. Truelove, R. M. Stephens, and M. W. Smith. Genome and gene alterations by insertions and deletions in the evolution of human and chimpanzee chromosome 22. *BMC genomics*, 10(1):51, 2009.

G. P. Wagner, K. Kin, and V. J. Lynch. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory in biosciences*, 131(4): 281–285, 2012.

R. S. Walikonis, O. N. Jensen, M. Mann, D. W. Provance, J. A. Mercer, and M. B. Kennedy. Identification of proteins in the postsynaptic density fraction by mass spectrometry. *Journal of Neuroscience*, 20(11):4069–4080, 2000.

R. J. Wanders, J. P. Ruiter, L. IJlst, H. R. Waterham, and S. M. Houten. The enzymology of mitochondrial fatty acid beta-oxidation and its application to follow-up analysis of positive neonatal screening results. *Journal of inherited metabolic disease*, 33(5): 479–494, 2010.

H.-Y. Wang, H.-C. Chien, N. Osada, K. Hashimoto, S. Sugano, T. Gojobori, C.-K. Chou, S.-F. Tsai, C.-I. Wu, and C.-K. J. Shen. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS biology*, 5(2):e13, 2006.

W. Wang and T. Lufkin. Hmx homeobox gene function in inner ear and nervous system cell-type specification and development. *Experimental cell research*, 306(2):373–379, 2005.

W. Wang, X. Chen, H. Xu, and T. Lufkin. Msx3: a novel murine homologue of the drosophila msh homeoÉą gene restricted to the dorsal embryonic central nervous system. *Mechanisms of development*, 58(1-2):203–215, 1996.

X. Wang, G. Lu, L. Li, J. Yi, K. Yan, Y. Wang, B. Zhu, J. Kuang, M. Lin, S. Zhang, et al. Huwe1 interacts with brca1 and promotes its degradation in the ubiquitin–proteasome pathway. *Biochemical and biophysical research communications*, 444 (3):290–295, 2014.

Z. Y. Wang, L. Jin, H. Tan, and D. M. Irwin. Evolution of hepatic glucose metabolism: liver-specific glucokinase deficiency explained by parallel loss of the gene for glucokinase regulatory protein (gckr). *PloS one*, 8(4):e60896, 2013.

A. J. Watt, H. Cuntz, M. Mori, Z. Nusser, P. J. Sjöström, and M. Häusser. Traveling waves in developing cerebellar cortex mediated by asymmetrical purkinje cell connectivity. *Nature neuroscience*, 12(4):463, 2009.

K. White, M.-J. Kim, C. Han, H.-J. Park, D. Ding, K. Boyd, L. Walker, P. Linser, Z. Meneses, C. Slade, et al. Loss of idh2 accelerates age-related hearing loss in male mice. *Scientific reports*, 8(1):5039, 2018.

D. E. Wilson and D. M. Reeder. *Mammal species of the world: a taxonomic and geographic reference*, volume 2. JHU Press, 2005.

J. Woo, S.-K. Kwon, and E. Kim. The ngl family of leucine-rich repeat-containing synaptic adhesion molecules. *Molecular and Cellular Neuroscience*, 42(1):1–10, 2009.

K. M. Woolfrey and M. L. Dell'Acqua. Coordination of protein phosphorylation and dephosphorylation in synaptic plasticity. *Journal of Biological Chemistry*, 290(48): 28604–28612, 2015.

G. J. Wright and P. Washbourne. Neurexins, neuroligins and lrrtms: synaptic adhesion getting fishy. *Journal of neurochemistry*, 117(5):765–778, 2011.

S. H. Wu, J. C. Arévalo, F. Sarti, L. Tessarollo, W.-B. Gan, and M. V. Chao. Ankyrin repeat-rich membrane spanning/kidins220 protein regulates dendritic branching and spine stability in vivo. *Developmental neurobiology*, 69(9):547–557, 2009.

Z. Xia and D. R. Storm. The role of calmodulin as a signal integrator for synaptic plasticity. *Nature Reviews Neuroscience*, 6(4):267, 2005.

N. Yamaguchi, A. Cooper, L. Werdelin, and D. W. Macdonald. Evolution of the mane and group-living in the lion (panthera leo): a review. *Journal of Zoology*, 263(4):329–342, 2004.

S. Yamamori, M. Itakura, D. Sugaya, O. Katsumata, H. Sakagami, and M. Takahashi. Differential expression of snap-25 family proteins in the mouse brain. *Journal of Comparative Neurology*, 519(5):916–932, 2011.

T. Yamauchi. Neuronal ca2+/calmodulin-dependent protein kinase iiâĂŤdiscovery, progress in a quarter of a century, and perspective: implication for learning and memory. *Biological and Pharmaceutical Bulletin*, 28(8):1342–1354, 2005.

L.-B. Yang, R. Li, S. Meri, J. Rogers, and Y. Shen. Deficiency of complement defense protein cd59 may contribute to neurodegeneration in alzheimer's disease. *Journal of Neuroscience*, 20(20):7505–7509, 2000.

S. Yasunaga, M. Grati, M. Cohen-Salmon, A. El-Amraoui, M. Mustapha, N. Salem, E. El-Zir, J. Loiselet, and C. Petit. A mutation in otof, encoding otoferlin, a fer-1-like protein, causes dfnb9, a nonsyndromic form of deafness. *Nature genetics*, 21(4):363, 1999.

S. Yoo and S. Blackshaw. Regulation and function of neurogenesis in the adult mammalian hypothalamus. *Progress in neurobiology*, 170:53–66, 2018.

M. Yoshida, S. Koyanagi, A. Matsuo, T. Fujioka, H. To, S. Higuchi, and S. Ohdo. Glucocorticoid hormone regulates the circadian coordination of μ-opioid receptor expression in mouse brainstem. *Journal of Pharmacology and Experimental Therapeutics*, 315(3):1119–1124, 2005.

X.-M. Yu, R. Askalan, G. J. Keil, and M. W. Salter. Nmda channel regulation by channel-associated protein tyrosine kinase src. *Science*, 275(5300):674–678, 1997.

E. M. Zdobnov, F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simão, P. Ioannidis, M. Seppey, A. Loetscher, and E. V. Kriventseva. Orthodb v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic acids research*, 45(D1):D744–D749, 2016.

H. Zeng, S. Chattarji, M. Barbarosie, L. Rondi-Reig, B. D. Philpot, T. Miyakawa, M. F. Bear, and S. Tonegawa. Forebrain-specific calcineurin knockout selectively impairs bidirectional synaptic plasticity and working/episodic-like memory. *Cell*, 107(5): 617–629, 2001.

F. Zhang, Y. Y. Shugart, W. Yue, Z. Cheng, G. Wang, Z. Zhou, C. Jin, J. Yuan, S. Liu, and Y. Xu. Increased variability of genomic transcription in schizophrenia. *Scientific reports*, 5:17995, 2015.

G. Zhang, C. Cowled, Z. Shi, Z. Huang, K. A. Bishop-Lilly, X. Fang, J. W. Wynne, Z. Xiong, M. L. Baker, W. Zhao, et al. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science*, page 1230835, 2012a.

Z. Zhang, J. Li, X.-Q. Zhao, J. Wang, G. K.-S. Wong, and J. Yu. Kaks_calculator: calculating ka and ks through model selection and model averaging. *Genomics, proteomics & bioinformatics*, 4(4):259–263, 2006.

Z. Zhang, J. Xiao, J. Wu, H. Zhang, G. Liu, X. Wang, and L. Dai. Paraat: a parallel tool for constructing multiple protein-coding dna alignments. *Biochemical and biophysical research communications*, 419(4):779–781, 2012b.

Q.-Y. Zhao, Y. Wang, Y.-M. Kong, D. Luo, X. Li, and P. Hao. Optimizing de novo transcriptome assembly from short-read rna-seq data: a comparative study. *BMC bioinformatics*, 12(14):S2, 2011.

J. Zhu, Y. Shang, C. Xia, W. Wang, W. Wen, and M. Zhang. Guanylate kinase domains of the maguk family scaffold proteins as specific phospho-protein-binding modules. *The EMBO journal*, 30(24):4986–4997, 2011.

E. B. Ziff. Enlightening the postsynaptic density. *Neuron*, 19(6):1163–1174, 1997.

R. S. Zucker. Calcium-and activity-dependent synaptic plasticity. *Current opinion in neurobiology*, 9(3):305–313, 1999.

G. K. Zupanc. Adult neurogenesis in teleost fish. In *Neurogenesis in the Adult Brain I*, pages 137–167. Springer, 2011.

G. K. Zupanc, K. Hinsch, and F. H. Gage. Proliferation, migration, neuronal differentiation, and long-term survival of new cells in the adult zebrafish brain. *Journal of Comparative Neurology*, 488(3):290–319, 2005.

Appendices

Appendix A

frontiers in Genetics

FOCUSED REVIEW
published: 11 January 2016
doi: 10.3389/fgene.2015.00361

Assembly, Assessment, and Availability of *De novo* Generated Eukaryotic Transcriptomes

Joanna Moreton[1,2]*, *Abril Izquierdo*[2] and *Richard D. Emes*[1,2]

[1] Advanced Data Analysis Centre, Sutton Bonington Campus, University of Nottingham, Leicestershire, UK, [2] School of Veterinary Medicine and Science, Sutton Bonington Campus, University of Nottingham, Leicestershire, UK

OPEN ACCESS

Edited by:
Chun Liang,
Miami University, USA

Reviewed by:
Shizuka Uchida,
Goethe University Frankfurt, Germany
Madavan Vasudevan,
Bionivid Technology Pvt. Ltd., India

***Correspondence:**

***Joanna Moreton** is a data analyst at the University of Nottingham Advanced Data Analysis Centre. She initially studied Mathematics at the University of Essex and then stayed on to complete a PhD in Bioinformatics. Her first post-doctoral position was at the University of Nottingham's Next Generation Sequencing Facility "Deep Seq." During this time she was involved in many different research projects including genome re-sequencing, RNA-Seq, ChIP-Seq, MeDIP-Seq, and RIP-Seq. Her current research interest is in the analysis of Next Generation Sequencing data.*
Joanna.Moreton@nottingham.ac.uk

***Received:** 20 September 2015*
***Accepted:** 19 December 2015*
***Published:** 11 January 2016*

Citation:
Moreton J, Izquierdo A and Emes RD (2016) Assembly, Assessment, and Availability of De novo Generated Eukaryotic Transcriptomes. Front. Genet. 6:361. doi: 10.3389/fgene.2015.00361

De novo assembly of a complete transcriptome without the need for a guiding reference genome is attractive, particularly where the cost and complexity of generating a eukaryote genome is prohibitive. The transcriptome should not however be seen as just a quick and cheap alternative to building a complete genome. Transcriptomics allows the understanding and comparison of spatial and temporal samples within an organism, and allows surveying of multiple individuals or closely related species. *De novo* assembly in theory allows the building of a complete transcriptome without any prior knowledge of the genome. It also allows the discovery of alternate splice forms of coding RNAs and also non-coding RNAs, which are often missed by proteomic approaches, or are incompletely annotated in genome studies. The limitations of the method are that the generation of a truly complete assembly is unlikely, and so we require some methods for the assessment of the quality and appropriateness of a generated transcriptome. Whilst no single consensus pipeline or tool is agreed as optimal, various algorithms, and easy to use software do exist making transcriptome generation a more common approach. With this expansion of data, questions still exist relating to how do we make these datasets fully discoverable, comparable and most useful to understand complex biological systems?

Keywords: *de novo* transcriptome assembly, high-throughput sequencing, assessment, availability, annotation

INTRODUCTION

It is desirable to fully understand the complexity of an organism and the diversity of cell types arising from a single genome, or to compare the compliment of genes between evolutionary groups. This requires a capability to view and catalog the changes in gene expression of a cell or tissue. The transcriptome is the complete set of transcripts (RNA molecules) within a cell including protein-coding and non-coding RNAs. Additionally, the transcriptome encompasses all alternative splice forms, alternatively polyadenylated, and RNA-edited transcripts. Together, these reflect the genes that are actively expressed in a particular tissue (Grobe et al., 2002; Lu et al., 2013). Understanding the complete transcriptome is a technical challenge requiring technologies for capturing an accurate representation of the RNA in a cell or tissue. The dominant technology for the assessment of gene expression was microarrays which use printed or synthesized probes corresponding to mRNAs (Fu et al., 2009). Whilst these technologies are robust and offer a more mature framework for data analysis, they require an already annotated complete genome to design the probes. Microarrays are also limited by inaccurate hybridization of sequences to probes, which is difficult to model and hence account for (Wang et al., 2009; Compeau et al., 2011). In the case of model organisms,

microarrays are still hugely useful to measure and compare gene expression. However, where high quality **annotation** and appropriate arrays do not exist, DNA sequencing offers the best method to understand the transcriptome. With the advent of Next Generation Sequencing (NGS) technologies and improved extraction methods to accurately purify RNA from smaller amounts of tissue or even single cells (Islam et al., 2011), the possibility to catalog and measure gene expression from a wider range of organisms has become possible.

KEY CONCEPT 1 | Annotation
The process of assigning functional information to transcripts, such as gene ontology terms, in order to characterize the sequences and allow understanding of the system studied.

Transcriptome assembly is the process of identifying transcripts and their variants that are expressed in a determined sample (Lu et al., 2013). The simple premise is to reconstruct the complete sequences of all transcripts in the transcriptome. It is uncommon to achieve this in practice as most of the time the sequencing depth is not sufficient to cover all full-length transcripts, particularly the ones of low abundance. A transcriptome is therefore a set of contiguous (contig) sequences that represent transcript regions (Li et al., 2014). Generally the strategies for transcriptome assembly fall into two categories: reference-based and *de novo* (**Figure 1**), although a combination of both can be used (Chen et al., 2011; Garber et al., 2011; Martin and Wang, 2011; Haas et al., 2013; Lu et al., 2013). Whilst a comprehensive set of tools is unrealistic, we have compiled a set of commonly used, freely available tools for *de novo* assembly and assessment (Supplementary Table 1).

TRANSCRIPTOME ASSEMBLY METHODS

Reference-Based Transcriptome Assembly Method

Reference-based transcriptome assembly is widely used when a model organism, with a sequenced genome for the target transcriptome, is accessible. Thus, the transcriptome is reconstructed by mapping to previously known sequences (Martin and Wang, 2011). The short reads are aligned to the reference genome allowing the overlapping regions to be assembled into transcripts. Where a good quality reference exists, the reference-based strategy is highly sensitive and it has become the basic method for many RNA sequencing (RNA-seq) studies. However, the accuracy of reference-based transcriptome assembly depends on correct read alignment, and issues such as alternative splicing and sequencing errors increase the difficulty of this task (Grabherr et al., 2011). In a referenced-based assembly approach, the sequence reads are aligned to the genome using a tool such as TopHat2 (Kim et al., 2013), which takes splicing into consideration. This is necessary as copies of mature spliced RNA have been sequenced, but these need to be mapped to a genome containing introns. All alternative splicing events are then captured in a graph for each given locus. Different paths are traversed in the graph to find transcript variants (Martin and Wang, 2011). Two transcriptome assemblers that are commonly used for graph building and traversal are Cufflinks (Trapnell et al., 2010) and Scripture (Guttman et al., 2010). The computational requirements of reference-based transcriptome assembly are significantly less compared to ***de novo* transcriptome assembly**. Furthermore, the presence of artifacts or sequencing contamination does not represent a major issue since these can often be resolved when aligning the reads to the genome. However, the quality of the results depends largely on the quality of the genome model used.

KEY CONCEPT 2 | Reference-based transcriptome assembly
A method which is used to reconstruct transcript sequences by aligning RNA sequencing reads to a reference genome.

KEY CONCEPT 3 | *De novo* transcriptome assembly
A process by which overlapping RNA sequencing reads are combined without a reference genome to reconstruct transcript sequences.

The transcriptome assembly can also be complicated by reads that align to multiple sites in the genome; these are known as multi-mapped reads. This problem is increased if the reads are short, therefore large complex transcriptomes are not easily assembled from very short reads (Martin and Wang, 2011). If there is insufficient unique information in the read sequences, then it is difficult to assign the reads to the correct location during alignment to the reference genome. If multi-mapped reads are discarded, then information for non-unique regions will be lost including gene families where gene sequences can be highly similar (Robert and Watson, 2015). If they are retained, it can be a challenge to accurately estimate gene or transcript abundances (Patro et al., 2014). Recently, Robert and Watson (2015) proposed a method for dealing with multi-mapped reads. They suggest taking all of the reads that cannot not be aligned to a unique gene and instead allocating them to a "multi-mapped group." These groups are determined from the RNA-seq data rather than relying on existing annotation. By performing differential expression analysis on multi-mapped gene groups, rather than individual genes, important biological information can be examined that would have otherwise been filtered out (Robert and Watson, 2015).

Once reads are mapped and transcripts are identified, there are tools that can be used to quantitate gene expression such as Cufflinks (Trapnell et al., 2010), DESeq2 (Love et al., 2014), or EdgeR (Robinson et al., 2010). Thus, for organisms with an accurate, complete and well annotated genome, the measurement of genes expressed in a sample is becoming commonplace with robust methods for mapping transcript fragments to the genome and measuring the transcriptome content. However, where an annotated genome does not exist, or the number of alternate transcript isoforms is high, the problem of generating an accurate representation of the complete transcriptome remains. It is in these situations that *de novo* transcriptome assembly is particularly attractive as it provides an alternative option for assessing a non-model transcriptome (Zhao et al., 2011). *De novo* transcriptome assembly works without a reference to attempt to directly reconstruct overlapping reads into transcripts

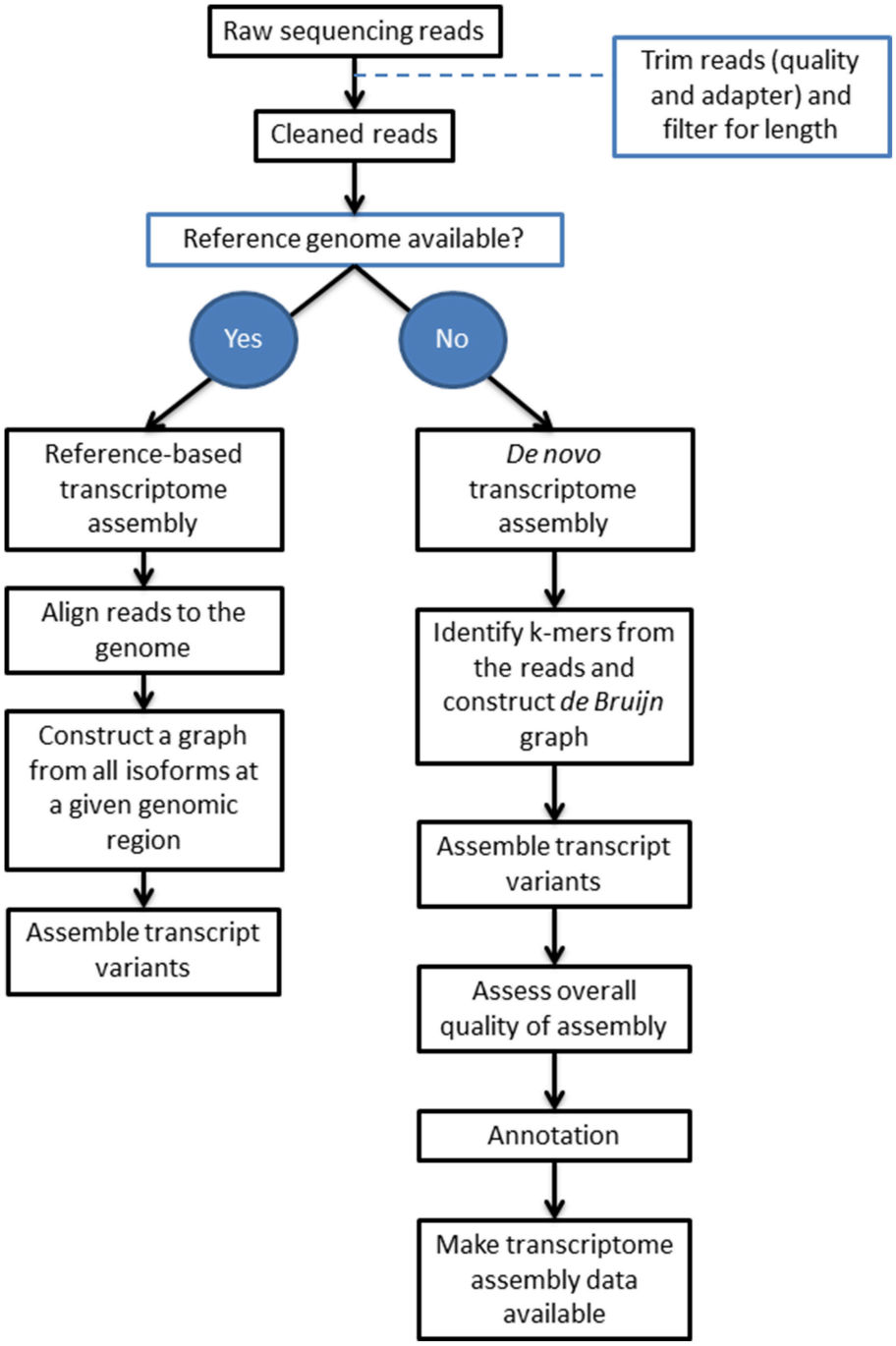


FIGURE 1 | An overview of the two transcriptome assembly pipelines. The key parts of two transcriptome assembly pipelines are shown depending on whether a reference genome is available. This review is focused on *de novo* transcriptome assembly; more information on the pipeline for reference-based transcriptome assembly can be found in review papers such as Martin and Wang (2011).

(Grabherr et al., 2011; Martin and Wang, 2011; Clarke et al., 2013; Lu et al., 2013). The complexities of this approach make it more computationally demanding, however a range of software tools exist including Oases (Schulz et al., 2012), Trans-ABySS (Robertson et al., 2010), MIRA (Chevreux et al., 2004), and Trinity (Grabherr et al., 2011). Several studies have been carried out to evaluate the execution of transcript assemblers (e.g., Clarke et al., 2013), and although they all differ in performance, currently there is no single transcriptome assembler categorized to be the best option for every condition (Grabherr et al., 2011; Clarke

et al., 2013; Góngora-Castillo and Buell, 2013; Lu et al., 2013). With these specialist comparisons of performance available, it is not the objective of this review to describe nuances of different approaches or to promote a single method as optimal. In many cases the use of multiple approaches and subsequent merging of assemblies to generate a consensus single or set of assemblies might be appropriate. For example, incorporating sequences from different assemblers and parameters to generate a consensus transcriptome, by using transcripts present in multiple original transcriptome assemblies (Moreton et al., 2014).

De novo Transcriptome Assembly Method

De novo transcriptome assemblers commonly use a strategy which involves constructing *de Bruijn* graphs (e.g., Grabherr et al., 2011; Schulz et al., 2012). In this approach all subsequences of length k are found in the reads and these are known as "**k-mers**." A *de Bruijn* graph is created using all unique k-mers as nodes, with connecting edges representing immediately overlapping k-mers (**Figure 2**). That is if a k-mer substring is shifted by one sequence base, and it overlaps another k-mer (by k-1 bases), then an edge is drawn between the nodes associated with those k-mers (Martin and Wang, 2011). A linear chain of k-mer nodes is compressed into a single node where possible (where the two nodes are joined by a single unique edge). Transcript variants can then be assembled by traversing the paths of the graph. **Figure 2** shows a toy example of a *de Bruijn* graph constructed from two 7 bp sequence reads and k-mers of length 5. In this example two paths can be found from the graph representing two possible transcript isoforms.

KEY CONCEPT 4 | k-mers
A subsequence of specified length k. They are often used by *de novo* assemblers to allow sequence information to be compacted, which makes reconstruction of transcripts easier computationally.

Before the introduction of *de Bruijn* graphs, assemblers used the overlap-layout-consensus algorithm where overlap information between read sequences is added to a mathematical graph to find a consensus sequence (Li et al., 2012b). In this strategy, each graph node corresponds to a read and if two reads overlap, their nodes are joined by an edge on the graph. The overlap-layout-consensus alignment step is computationally intensive when assembling a huge number of short reads, so a *de Bruijn* graph algorithm is preferred for generating *de novo* assemblies. By compacting the sequence information into k-mers, the graph theory method for finding a path in the graph becomes easier computationally (Pevzner et al., 2001; Li et al., 2012b). One disadvantage in using the *de Bruijn* graph approach is the generation of misassembled contigs which occurs because of the use of k-mers (Clarke et al., 2013). If two transcripts from different genes have the same k-mer sequence they could be erroneously connected. The computational proficiency of the *de Bruijn* graph strategy is clearly beneficial, but it is an ongoing problem to balance this with assembly accuracy (Clarke et al., 2013).

There are a number of difficulties that are encountered by the *de novo* transcriptome assembly strategy. For example, it is challenging to discriminate between transcript variants that are produced from processes such as alternative splicing or sequences transcribed from paralogous genes (Grabherr et al., 2011; Vijay et al., 2013). These sorts of sequences will share k-mer sequences and hence it is difficult to tease them apart into separate transcripts. Software tools have been designed to distinguish transcript variants using paired-end read data and read coverage (Góngora-Castillo and Buell, 2013). For instance, the Trinity assembler (Grabherr et al., 2011) reconstructs alternatively spliced transcripts and paralogous sequences by clustering overlapping contigs and generating a *de Bruijn* graph for each cluster of sequences independently. These graphs are then supplemented with the read and paired-end information to generate all possible transcript variants. Despite the challenges, the transcriptomes of many different organisms have been assembled using the *de novo* approach (e.g., Kumar and Blaxter, 2010; Robertson et al., 2010; Zhao et al., 2011; Price et al., 2015). These complexities are additionally compounded when mixed samples are included, for example in pathogen and host, or when transcripts may not form distinct entities due to dense or overlapping transcripts, as seen in prokaryote organisms. In the case of bacterial *de novo* assembly, tools such as Rockhopper (McClure et al., 2013; Tjaden, 2015) have been specifically developed.

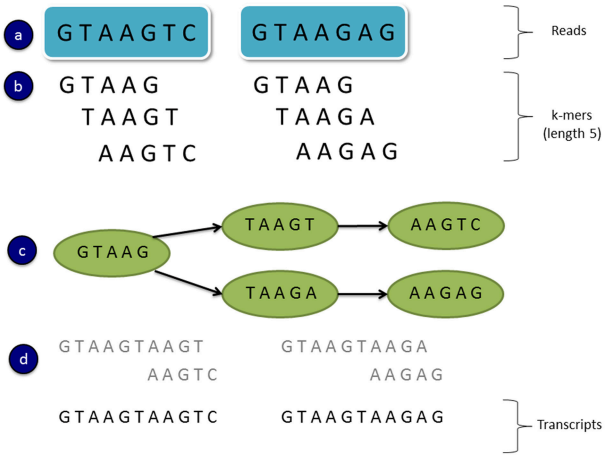


FIGURE 2 | An example of a simple *de Bruijn* graph. (A) Read sequences **(B)** All subsequence k-mers of length 5 from the reads **(C)** A *de Bruijn* graph constructed from unique k-mers as the nodes and overlapping k-mers connected by edges (a k-mer shifted by one base overlaps another k-mer by k-1 bases) **(D)** Assembled transcripts by traversing the two paths in the graph.

ASSESSMENT OF GENERATED *DE NOVO* ASSEMBLIES

Whilst a number of studies have focused on transcriptome assembly, the assessment of the overall quality of the derived assemblies is less well defined. A number of different measures are commonly used to evaluate assembled transcriptomes. Commonly used metrics when there is no close reference include the number of contigs (transcripts) assembled, summed contig

length, mean transcript length, N50 value, and the proportion of reads that could be mapped back to the assembled transcripts (RMBT; e.g., Zhao et al., 2011). These measures can be used to compare and select optimal assemblies, for example the N50 value can be maximized whilst keeping the total assembly length as long as possible (Zerbino, 2010). It is also important to consider the time taken to generate the assemblies (Kumar and Blaxter, 2010). When reference sequences of closely related species are available, the assembled contigs can be compared using a sequence similarity tool such as BLAST (McGinnis and Madden, 2004) to assess the validity of the assembly (e.g., Arun-Chinnappa and McCurdy, 2015; Ghaffari et al., 2015). However, this approach is biased by the appropriateness of the choice of related species for comparison and will be biased toward available "model" genomes.

Assessment of the completeness of an assembled transcriptome is more problematic. This is due to the impossibility of knowing *a priori* what the complete transcriptome for a previously unsequenced cell, or collection of cells, at a particular time point is. However, the theoretical completeness can also be assessed, using methods to determine the assembly of transcripts that are expected to be present in all cells at all times, such as the Core Eukaryotic Genes Mapping Approach (CEGMA) tool by Parra et al. (2007). Although not developed specifically for this purpose, many studies have used this approach to determine if a collection of newly assembled transcripts encode one or more of a set of core genes conserved across a wide range of eukaryotic species, thus providing a percentage "completeness" score (e.g., Chauhan et al., 2014; Moreton et al., 2014; Frías-López et al., 2015; Powell et al., 2015; Price et al., 2015). A recent web-based tool "TRUFA," developed by Kornobis et al. (2015), incorporates CEGMA into its pipeline as part of the assessment stage of *de novo* assemblies. As of May 2015 CEGMA is no longer being supported, however a new tool "BUSCO" has been published by Simão et al. (2015), to assess assembly and annotation completeness using sets of Benchmarking Universal Single-Copy Orthologs (BUSCO), selected from OrthoDB (Kriventseva et al., 2015). When comparing the completeness of genome assemblies and gene sets across 40 species, the BUSCO assessments were more consistent than CEGMA, the run-times were much faster and the software can also be used to assess gene sets and transcriptomes (Simão et al., 2015).

Some authors have suggested that evaluation measures such as N50 might be misleading and uninformative for evaluating transcriptome assemblies (e.g., O'Neil and Emrich, 2013; Li et al., 2014; Chen et al., 2015). For example, Chen et al. (2015) found that the transcriptome assemblies with the highest N50 values, did not make a significant contribution to the best assembled transcript set based on coding potential. Li et al. (2014) developed the "DETONATE" (DE novo TranscriptOme rNa-seq Assembly with or without the Truth Evaluation) software, which includes both reference-free (RSEM-EVAL) and reference-based (REF-EVAL) methods. The reference-free approach is based on a probabilistic model that uses only the read and assembly data. When reference transcripts are available, the REF-EVAL component can be used to generate scores based on different reference-based measures. DETONATE is currently only designed to evaluate assemblies generated from Illumina data, although there are plans to update the package to handle data from other sequencing platforms. O'Neil and Emrich (2013) assessed a number of metrics for *de novo* transcriptome assemblies including unique annotations and "ortholog hit ratio" from their earlier work (O'Neil et al., 2010). The correlation between the REF-EVAL score and the ortholog hit ratio measure was found to be low, although the number of unique proteins matched had good correlation to REF-EVAL (Li et al., 2014).

There are a number of errors that can occur in *de novo* transcriptome assembly, for example two transcripts may be combined into a single false chimeric transcript, or contigs might be incomplete or mis-assembled (Smith-Unna et al., 2015). These errors can be detected using read evidence. The *TransRate* tool (Smith-Unna et al., 2015) aligns the paired-end reads that were used to generate the assembly, back to the assembled contigs. The alignments are then evaluated and each contig is assigned a score based on properties such as how well the nucleotides in the aligned reads matched to the assembled contigs, the coverage of the contig nucleotides, and the order of the contig nucleotides based on the paired-end read orientations. *TransRate* also calculates an assembly score which is generated from the individual contig scores, and the proportion of input reads that were incorporated into the *de novo* assembly. As mentioned before, RSEM-EVAL is another reference-free evaluation method; however it does not focus on the evaluation of individual contigs. The RSEM-EVAL tool is also limited to assemblies generated from Illumina data, but *TransRate* is not restricted in this way. The *TransRate* tool is also useful because it allows the filtration of individual contigs based on their scores. Furthermore, the authors used 155 previously published *de novo* assemblies in a meta-analysis to allow users to analyze their assemblies in comparison with others. In summary, assembly assessments are essential and will be increasingly important for evaluation of new methods, or in the combination of assemblies as part of optimization strategies.

ANNOTATION OF TRANSCRIPTOME ASSEMBLY

Annotation of function is required to characterize transcripts and allow understanding of the system studied. Most approaches to annotation of protein coding transcripts use one or more homology based approaches to identify related sequences of known function, and hence transfer this annotation to the new transcript (Emes, 2008). There are however limitations to these approaches. The problem of transfer of inappropriate or inaccurate annotation from one dataset to another, leading to the propagation of annotation error, is the most concerning. A preferred method is the use of protein domain architecture to drive the annotation. Searching for conserved domains using hidden Markov model search tools, such as HMMER3 (Finn et al., 2011), is a relatively simple process. These tools search comprehensive libraries of domains such as Pfam (Finn et al., 2014) or InterPro (Mitchell et al., 2015). Databases

such as Pfam2GO, from the gene ontology consortium (Gene Ontology Consortium, 2015), allow the domain content to generate restricted descriptors of each transcript. Pipeline tools to automate this process using both sequence similarity and domain composition, such as the Trinotate pipeline (https://trinotate.github.io/), are available but are currently relatively slow or computationally intense to use. Another consideration for the annotation process is searching for repeat elements using programs such as RepeatMasker (http://www.repeatmasker.org) or the Tandem Repeats Finder (Benson, 1999). For example, RepeatMasker can be used with the Repbase database (Bao et al., 2015) to identify transposable elements and other types of repeats (Gillard et al., 2014; Kumar et al., 2014; Cokus et al., 2015; Richardson and Sherman, 2015).

DE NOVO TRANSCRIPTOME ASSEMBLY AVAILABILITY

Whilst most journals require raw sequencing reads to be made publicly available in a database such as the Sequence Read Archive (SRA; Kodama et al., 2012), often the assembled transcripts and annotations are not made available. This results in lack of clarity and wasted effort to redo the analysis. The SRA is part of the International Nucleotide Sequence Database Collaboration (Kodama et al., 2012). This repository is available at the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov/sra), European Bioinformatics Institute (EBI, www.ebi.ac.uk/ena), and DNA Data Bank of Japan (DDBJ, http://trace.ddbj.nig.ac.jp/dra). There are support pages and handbooks to help with submitting data, and these are available at the NCBI, EBI, and DDBJ websites. As well as raw sequence data, alignment files in BAM (Li et al., 2009) format can also be submitted to the SRA. With reducing costs of sequencing and availability of software for transcriptome assembly, the making of transcriptome assembly open and available is a key problem in bioinformatics. Often generic genome browsers are difficult to set up and are not well-suited for transcriptome data (Jones and Blaxter, 2013), and so a number of software solutions to host and visualize transcriptome assemblies have been developed. Jones and Blaxter (2013) developed the web application "afterParty" which enables users to make a transcriptome publicly available. The application can take as input either Roche 454 reads, or assembled contigs (putative transcripts) from any platform. If raw 454 sequencing reads are used as an input, then afterParty can assemble them using MIRA (Chevreux et al., 2004) and then annotate the resulting contigs using BLASTX (Altschul et al., 1997), UniProt (Uniprot Consortium, 2012), and InterProScan (Zdobnov and Apweiler, 2001). In the other afterParty workflows, contigs generated by the user from any sequencing platform can be uploaded with or without annotation. AfterParty can also be used to browse transcriptomes and visualize data sets in a web browser. For example, all contigs with annotation matching a particular search term can be used to generate a scatter plot of GC content against coverage in a comparison to the full assembly (Jones and Blaxter, 2013). Different contig sets, chart types, and displays can be selected. In addition to filtering by annotation, a DNA or protein sequence can be used to find contigs with sequence similarity. The contigs can also be searched by properties such as length, quality, coverage, and GC content. A number of studies have already used the afterParty website as a means of hosting and distributing transcriptome data (e.g., Heitlinger et al., 2014; Short et al., 2014; McTaggart et al., 2015). For users running afterParty locally, the source code, and dependencies can be installed. However, the more convenient method would be to use the virtual disk image (available on GitHub), which contains all the required dependencies to run the software using a virtual machine. Alternatively, afterParty is also available through a public server.

RNAbrowse is an alternative package with a web interface that can be used to store and visualize *de novo* transcriptome data (Mariette et al., 2014). It is based on the BioMart (Smedley et al., 2015) software and in addition to the web interface it includes a command line tool for administration which requires a unix server and MySQL database. The project introduction page of the web interface contains useful information such as the software and parameters used to generate the alignment, annotation, assembly, and variant analysis. The contig and variant overview pages show general statistics and related figures such as a bar chart of contig length distribution. There is a blast query form to search the contigs using an input sequence, and the BioMart search page can also be used to filter the data based on criteria such as contig name, length, or annotation. In the sequence view, the longest open reading frame can be identified. It is also possible to view the sequences and annotations in JBrowse (Skinner et al., 2009) and compare read coverage between samples in the contig depth view. The figures produced using the interface can be easily printed or downloaded and there is also a dedicated download page to enable users to save some or all of the data (Mariette et al., 2014). In its simplest form, RNAbrowse can be set up using the assembled contig sequences (FASTA format) alongside the annotation and alignment files. Again, installation requires a number of prerequisite tools and the setup process can be quite time consuming (Mariette et al., 2014). This may therefore be better attempted in collaboration with a bioinformatics group or local support. However, there is a project website with lots of information about RNAbrowse including guides, demonstrations, example datasets and a configuration file template for larger projects. Different schedulers can also be selected to address any time issues (Mariette et al., 2014). As an example of a practical use, RNAbrowse has been used to display and distribute beech tree *de novo* transcriptome data (Lesur et al., 2015).

Apart from more complete packages such as afterParty and RNAbrowse, there are limited tools with web interfaces that are available for analysis of transcriptome data. CBrowse (Li et al., 2012a) is a web browser which takes assembled contig sequences and BAM/SAM alignment files as input, and enables the user to identify polymorphisms and view the contigs in the web interface. Its focus is not on annotation, however CBrowse can be used to disseminate assembled transcriptome data (Li et al., 2012a). As a less permanent solution, some research groups have used individual online resources to make their data available. For example, Aya et al. (2015) developed a transcriptome database as

a public web resource for downloading and browsing fern *de novo* transcriptome assembly data, where both BLAST and keyword searches can be performed. Another research group released their axolotl read and transcriptome assembly data on a website with a keyword search facility (Stewart et al., 2013). However, the risk of non-specialist solutions is that repositories are not maintained or, with the movement of personnel, that the skill to maintain repositories is lost. As an interim solution, we and others have simply made transcriptome assembly data available to download by partnering with appropriate journals (Moreton et al., 2014; Ghaffari et al., 2015). Given these considerations, and the enhanced ability to query, filter and visualize transcriptome data, tools like afterParty, and RNAbrowse make the most ideal options.

CONCLUSION

As the desire to catalog and compare the varied transcriptomes of complex organisms continues, *de novo* transcriptome assembly is an important tool in the bioinformatician's arsenal. Whilst rapid progress in single molecule sequencing is being made, it is currently not mature and so assembly, annotation and assessment of transcriptomes from relatively short reads will continue to be essential. To make these methods truly useful, assemblies that are accurately assembled and annotated are essential, but also the availability and openness of assembled transcriptomes not simply raw data must become expected practice.

AUTHOR CONTRIBUTIONS

JM, AI, and RE wrote the paper, prepared figures, and reviewed drafts of the paper.

ACKNOWLEDGMENTS

AI was supported by an international PhD studentship from Consejo Nacional de Ciencia y Tecnologia (CONACYT) Mexico. JM and RE were funded by the University of Nottingham.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2015.00361

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Arun-Chinnappa, K. S., and McCurdy, D. W. (2015). *De novo* assembly of a genome-wide transcriptome map of *Vicia faba* (L.) for transfer cell research. *Front. Plant Sci.* 6:217. doi: 10.3389/fpls.2015.00217

Aya, K., Kobayashi, M., Tanaka, J., Ohyanagi, H., Suzuki, T., Yano, K., et al. (2015). *De novo* transcriptome assembly of a fern, *Lygodium japonicum*, and a web resource database, Ljtrans DB. *Plant Cell Physiol.* 56, e5. doi: 10.1093/pcp/pcu184

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6:11. doi: 10.1186/s13100-015-0041-9

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Chauhan, P., Hansson, B., Kraaijeveld, K., De Knijff, P., Svensson, E. I., and Wellenreuther, M. (2014). *De novo* transcriptome of *Ischnura elegans* provides insights into sensory biology, colour and vision genes. *BMC Genomics* 15:808. doi: 10.1186/1471-2164-15-808

Chen, G., Wang, C., and Shi, T. (2011). Overview of available methods for diverse RNA-Seq data analyses. *Sci. China Life Sci.* 54, 1121–1128. doi: 10.1007/s11427-011-4255-x

Chen, S., Mcelroy, J. S., Dane, F., and Peatman, E. (2015). Optimizing transcriptome assemblies for leaf and seedling by combining multiple assemblies from three *de novo* assemblers. *Plant Genome* 8:1. doi: 10.3835/plantgenome2014.10.0064

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E., Wetter, T., et al. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14, 1147–1159. doi: 10.1101/gr.1917404

Clarke, K., Yang, Y., Marsh, R., Xie, L., and Zhang, K. K. (2013). Comparative analysis of *de novo* transcriptome assembly. *Sci. China Life Sci.* 56, 156–162. doi: 10.1007/s11427-013-4444-x

Cokus, S. J., Gugger, P. F., and Sork, V. L. (2015). Evolutionary insights from *de novo* transcriptome assembly and SNP discovery in California white oaks. *BMC Genomics* 16:552. doi: 10.1186/s12864-015-1761-4

Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29, 987–991. doi: 10.1038/nbt.2023

Emes, R. D. (2008). Inferring function from homology. *Methods Mol. Biol.* 453, 149–168. doi: 10.1007/978-1-60327-429-6_6

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Frías-López, C., Almeida, F. C., Guirao-Rico, S., Vizueta, J., Sánchez-Gracia, A., Arnedo, M. A., et al. (2015). Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). *PeerJ* 3:e1064. doi: 10.7717/peerj.1064

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., et al. (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 10:161. doi: 10.1186/1471-2164-10-161

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477. doi: 10.1038/nmeth.1613

Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056. doi: 10.1093/nar/gku1179

Ghaffari, N., Arshad, O. A., Jeong, H., Thiltges, J., Criscitiello, M. F., Yoon, B.-J., et al. (2015). Examining *De Novo* transcriptome assemblies via a quality assessment pipeline. *Comput. Biol. Bioinformatics IEEE/ACM Trans.* 99:1. doi: 10.1109/TCBB.2015.2446478

Gillard, G. B., Garama, D. J., and Brown, C. M. (2014). The transcriptome of the NZ endemic sea urchin Kina (*Evechinus chloroticus*). *BMC Genomics* 15:45. doi: 10.1186/1471-2164-15-45

Góngora-Castillo, E., and Buell, C. R. (2013). Bioinformatics challenges in *de novo* transcriptome assembly using short read sequences in the absence of a reference genome sequence. *Nat. Prod. Rep.* 30, 490–500. doi: 10.1039/c3np20099j

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Grobe, K., Esko, J., Aikawa, J., Grobe, K., Tsujimoto, M., and Esko, J. (2002). Analysis of the mouse transcriptome based on functional annotation. *Nature* 420, 563–573. doi: 10.1038/nature01266

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510. doi: 10.1038/nbt.1633

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Heitlinger, E., Taraschewski, H., Weclawski, U., Gharbi, K., and Blaxter, M. (2014). Transcriptome analyses of *Anguillicola crassus* from native and novel hosts. *PeerJ* 2:e684. doi: 10.7717/peerj.684

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., et al. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21, 1160–1167. doi: 10.1101/gr.110882.110

Jones, M., and Blaxter, M. (2013). afterParty: turning raw transcriptomes into permanent resources. *BMC Bioinformatics* 14:301. doi: 10.1186/1471-2105-14-301

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Kodama, Y., Shumway, M., and Leinonen, R. (2012). The Sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* 40, D54–D56. doi: 10.1093/nar/gkr854

Kornobis, E., Cabellos, L., Aguilar, F., Frías-López, C., Rozas, J., Marco, J., et al. (2015). TRUFA: a user-friendly web server for *de novo* RNA-seq analysis using cluster computing. *Evol. Bioinform. Online* 11, 97–104. doi: 10.4137/EBO.S23873

Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simão, F. A., Pozdnyakov, I. A., et al. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.* 43, D250–D256. doi: 10.1093/nar/gku1220

Kumar, M., Gantasala, N. P., Roychowdhury, T., Thakur, P. K., Banakar, P., Shukla, R. N., et al. (2014). *De Novo* transcriptome sequencing and analysis of the cereal cyst nematode, *Heterodera avenae*. *PLoS ONE* 9:e96311. doi: 10.1371/journal.pone.0096311

Kumar, S., and Blaxter, M. L. (2010). Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics* 11:571. doi: 10.1186/1471-2164-11-571

Lesur, I., Bechade, A., Lalanne, C., Klopp, C., Noirot, C., Leple, J. C., et al. (2015). A unigene set for European beech (*Fagus sylvatica* L.) and its use to decipher the molecular mechanisms involved in dormancy regulation. *Mol. Ecol. Resour.* 15, 1192–1204. doi: 10.1111/1755-0998.12373

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., et al. (2014). Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. *Genome Biol.* 15, 553. doi: 10.1186/s13059-014-0553-5

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, P., Ji, G., Dong, M., Schmidt, E., Lenox, D., Chen, L., et al. (2012a). CBrowse: a SAM/BAM-based contig browser for transcriptome assembly visualization and analysis. *Bioinformatics* 28, 2382–2384. doi: 10.1093/bioinformatics/bts443

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., et al. (2012b). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Brief. Funct. Genomics* 11, 25–37. doi: 10.1093/bfgp/elr035

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Lu, B., Zeng, Z., and Shi, T. (2013). Comparative study of *de novo* assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Sci. China Life Sci.* 56, 143–155. doi: 10.1007/s11427-013-4442-z

Mariette, J., Noirot, C., Nabihoudine, I., Bardou, P., Hoede, C., Djari, A., et al. (2014). RNAbrowse: RNA-Seq *de novo* assembly results browser. *PLoS ONE* 9:e96821. doi: 10.1371/journal.pone.0096821

Martin, J. A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12, 671–682. doi: 10.1038/nrg3068

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., et al. (2013). Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res.* 41, e140. doi: 10.1093/nar/gkt444

McGinnis, S., and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32, W20–W25. doi: 10.1093/nar/gkh435

McTaggart, S. J., Hannah, T., Bridgett, S., Garbutt, J. S., Kaur, G., and Boots, M. (2015). Novel insights into the insect trancriptome response to a natural DNA virus. *BMC Genomics* 16:310. doi: 10.1186/s12864-015-1499-z

Mitchell, A., Chang, H. Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 43, D213–D221. doi: 10.1093/nar/gku1243

Moreton, J., Dunham, S. P., and Emes, R. D. (2014). A consensus approach to vertebrate *de novo* transcriptome assembly from RNA-seq data: assembly of the duck (*Anas platyrhynchos*) transcriptome. *Front. Genet.* 5:190. doi: 10.3389/fgene.2014.00190

O'Neil, S. T., Dzurisin, J. D., Carmichael, R. D., Lobo, N. F., Emrich, S. J., and Hellmann, J. J. (2010). Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. *BMC Genomics* 11:310. doi: 10.1186/1471-2164-11-310

O'Neil, S. T., and Emrich, S. J. (2013). Assessing *de novo* transcriptome assembly metrics for consistency and utility. *BMC Genomics* 14:465. doi: 10.1186/1471-2164-14-465

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. doi: 10.1093/bioinformatics/btm071

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32, 462–464. doi: 10.1038/nbt.2862

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A.* 98, 9748–9753. doi: 10.1073/pnas.171285098

Powell, D., Knibb, W., Remilton, C., and Elizur, A. (2015). *De-novo* transcriptome analysis of the banana shrimp (*Fenneropenaeus merguiensis*) and identification of genes associated with reproduction and development. *Mar. Genomics* 22, 71–78. doi: 10.1016/j.margen.2015.04.006

Price, S. J., Garner, T. W. J., Balloux, F., Ruis, C., Paszkiewicz, K. H., Moore, K., et al. (2015). A *de novo* Assembly of the Common Frog (*Rana temporaria*) transcriptome and comparison of transcription following exposure to *Ranavirus* and *Batrachochytrium dendrobatidis*. *PLoS ONE* 10:e0130500. doi: 10.1371/journal.pone.0130500

Richardson, M. F., and Sherman, C. D. H. (2015). *De Novo* assembly and characterization of the invasive northern pacific seastar transcriptome. *PLoS ONE* 10:e0142003. doi: 10.1371/journal.pone.0142003

Robert, C., and Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* 16, 1–16. doi: 10.1186/s13059-015-0734-x

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). *De novo* assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912. doi: 10.1038/nmeth.1517

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28, 1086–1092. doi: 10.1093/bioinformatics/bts094

Short, S., Yang, G., Guler, Y., Green Etxabe, A., Kille, P., and Ford, A. T. (2014). Crustacean intersexuality is feminization without demasculinization: implications for environmental toxicology. *Environ. Sci. Technol.* 48, 13520–13529. doi: 10.1021/es5050503

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness

with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., and Holmes, I. H. (2009). JBrowse: a next-generation genome browser. *Genome Res.* 19, 1630–1638. doi: 10.1101/gr.094607.109

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 43, W589–W598. doi: 10.1093/nar/gkv350

Smith-Unna, R. D., Boursnell, C., Patro, R., Hibberd, J. M., and Kelly, S. (2015). TransRate: reference free quality assessment of *de-novo* transcriptome assemblies. *BioRxiv* 021626. doi: 10.1101/021626

Stewart, R., Rascón, C. A., Tian, S., Nie, J., Barry, C., Chu, L.-F., et al. (2013). Comparative RNA-seq Analysis in the unsequenced axolotl: the oncogene burst highlights early gene expression in the blastema. *PLoS Comput. Biol.* 9:e1002936. doi: 10.1371/journal.pcbi.1002936

Tjaden, B. (2015). *De novo* assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol.* 16, 1. doi: 10.1186/s13059-014-0572-2

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Uniprot Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 40, D71–D75. doi: 10.1093/nar/gkr981

Vijay, N., Poelstra, J. W., Künstner, A., and Wolf, J. B. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. *Mol. Ecol.* 22, 620–634. doi: 10.1111/mec.12014

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847

Zerbino, D. R. (2010). Using the Velvet *de novo* assembler for short-read sequencing technologies. *Curr. Protoc. Bioinformatics* Chapter 11, Unit 11:15. doi: 10.1002/0471250953.bi1105s31

Zhao, Q. Y., Wang, Y., Kong, Y. M., Luo, D., Li, X., and Hao, P. (2011). Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* (12 Suppl. 14):S2. doi: 10.1186/1471-2105-12-S14-S2

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix B

```
#!/usr/bin/perl
use warnings;
use strict;
use Getopt::Long;

my $usage = "
    ================================================================================

USAGE:
-f    fasta file
-s    Salmon TPM file from Transrate e.g xxx.quant.sf
-t    Minimum TPM e.g. 1
-l    Minimum transcript length in bp
================================================================================

" ;

my ($fasta, $salmon, $min_tpm, $min_length);

GetOptions(
    'f|fasta:s'     => \$fasta,
    's|salmon:s'     => \$salmon,
  't|tpm:s'     => \$min_tpm,
  'l|length:s'     => \$min_length,
  );
```

```
23
24 if( ! defined $fasta) {
25 print "$usage\nWARNING: Cannot proceed without input file fasta file\n\n"; exit;
26 }
27 if( ! defined $salmon) {
28 print "$usage\nWARNING: Cannot proceed without input salmon quantification file\
       n\n"; exit;
29 }
30 if( ! defined $min_tpm) {
31 print "$usage\nWARNING: Cannot proceed without minimum tpm value\n\n"; exit;
32 }
33 if( ! defined $min_length) {
34 print "$usage\nWARNING: Cannot proceed without minimum transcript length\n\n";
       exit;
35 }
36
37
38 open TPM, $salmon;
39 my %tpm_lookup;
40 while (<TPM>)
41   {
42   chomp $_;
43   unless ($_ =~ /^\#/)
44     {
45     my @data = split '\t', $_;
46     if ($data[2] >= $min_tpm && $data[1] >= $min_length)
47       {
48       $tpm_lookup{$data[0]} = $data[2];
49       }
50     }
51   }
52
53
54
55 open FASTA, $fasta;
```

```
56 open OUT, ">$fasta\.tpm$min_tpm\.minlength$min_length\.fa";
57 {
58 local $/ = '>';
59 <FASTA>;                                          # throw away the first line
     'cos will only contain ">"
60 while (<FASTA>)
61   {
62       chomp $_;
63       my ($seq_id, @sequence) = split "\n";          # split the fasta input
     into Id and sequence
64   my $fasta_sequence = join '',@sequence;          # reassembles the sequence
65       if (exists $tpm_lookup{$seq_id})
66     {
67     print OUT "\>$seq_id\n$fasta_sequence\n";
68     }
69   }
70 }
71 close FASTA;
72 close OUT;
```

```
#!/usr/bin/perl
use warnings;
use strict;
use Getopt::Long;

my $usage = "
    ================================================================================

USAGE:
-f  fasta file
-c  cd-est-hit cluster xxx.est.clstr
-m  Minimum number of sequences in cluster to retain
-o  Out file name
-g  should minimum cluster size be from different input files (Y/N) Assumes that
     sequences are named xxx_yyy where yyy is a number
================================================================================

" ;

my ($fasta, $cluster, $min_seq, $output, $groups);

GetOptions(
    'f|fasta:s'     => \$fasta,
    'c|cluster:s'     => \$cluster,
  'm|min:s'     => \$min_seq,
  'o|output:s'     => \$output,
  'g|group:s'     => \$groups,
  );

if( ! defined $fasta) {
print "$usage\nWARNING: Cannot proceed without input file fasta file\n\n"; exit;
}
if( ! defined $cluster) {
print "$usage\nWARNING: Cannot proceed without cd-est-hit cluster file\n\n";
    exit;
```

```
}
if( ! defined $min_seq) {
print "$usage\nWARNING: Cannot proceed without minimum number in cluster\n\n";
    exit;
}
if( ! defined $output) {
print "$usage\nWARNING: Cannot proceed without output file name\n\n"; exit;
}
if( ! defined $groups) {
print "$usage\nWARNING: Cannot proceed without group (-g) should minimum cluster
     size be from different input files (Y/N)\n\n"; exit;
}
$groups = uc($groups);
unless ($groups =~ /[YN]/){print "$usage\nWARNING: Cannot proceed without group
    (-g) should minimum cluster size be from different input files (Y/N)\n\n";
    exit;}



my $array_size_needed = $min_seq -1 ; # array size needed will be 1 less than
    minimum due to array index starting at 0

my %cluster_lookup;

if ($groups eq "N")
  {
  open CLUSTER, $cluster;
  {
  local $/ = '>Cluster';
  <CLUSTER>;                                          # throw away the first
    line 'cos will only contain ">"
  while (<CLUSTER>)
    {
    chomp $_;
    my ($cluster_id, @sequences) = split "\n";          # split the fasta
```

```
    input into Id and sequence
60
61      if (exists $sequences[$array_size_needed])
62        {
63        foreach(@sequences)
64          {
65          chomp $_;
66          if ($_ =~ /.*?\>(.*?)\.\.\.\s\*/)
67            {
68            $cluster_lookup{$1} = 1;
69            }
70          }
71        }
72      }
73    }
74    close CLUSTER;
75  }
76
77
78
79  if ($groups eq "Y")
80    {
81    open CLUSTER, $cluster;
82    {
83    local $/ = '>Cluster';
84    <CLUSTER>;                                    # throw away the first
      line 'cos will only contain ">"
85    while (<CLUSTER>)
86      {
87      chomp $_;
88      my ($cluster_id, @sequences) = split "\n";      # split the fasta
      input into Id and sequence
89
90      if (exists $sequences[$array_size_needed])
91        {
```

```
92        my $longest = "NA";
93        my @groups = ();
94        foreach(@sequences)
95          {
96          chomp $_;
97          if ($_ =~ /.*?\>(.*?)\_\d+\.\.\./)
98            {
99            push @groups, $1;
100           }
101         if ($_ =~ /.*?\>(.*?)\.\.\.\s\*/)
102           {
103           $longest = $1;
104           }
105         }
106         my @uniq_groups = uniq_array(@groups);
107
108         if (exists $uniq_groups[$array_size_needed])
109           {
110           $cluster_lookup{$longest} = 1;
111           }
112       }
113     }
114   }
115   close CLUSTER;
116 }
117
118
119
120
121 open FASTA, $fasta;
122 open OUT, ">$output";
123 {
124 local $/ = '>';
125 <FASTA>;                                    # throw away the first line
      'cos will only contain ">"
```

```
126 while (<FASTA>)
127   {
128       chomp $_;
129       my ($seq_id, @sequence) = split "\n";              # split the fasta input
      into Id and sequence
130     my $fasta_sequence = join '',@sequence;           # reassembles the sequence
131       if (exists $cluster_lookup{$seq_id})
132     {
133     print OUT "\>$seq_id\n$fasta_sequence\n";
134     }
135   }
136 }
137 close FASTA;
138 close OUT;
139
140
141
142 sub uniq_array{
143 ##### make a unique list from the @genes array
144 my @in = @_;
145 my %seen = ();
146 my @uniq = ();
147 foreach (@in)
148 {chomp $_;
149 unless ($seen{$_}){
150 $seen{$_} = 1;
151 push (@uniq, $_);
152 }
153 }
154
155 return @uniq;
156 }
```

```
#!/usr/bin/perl
use warnings;
use strict;


unless (exists $ARGV[1]) {print "\n[file] [seq prefix]\n";exit;}

open FASTA, $ARGV[0];
my $count = 1;

{
local $/ = '>';
<FASTA>;                                        # throw away the first line
     'cos will only contain ">"

while (<FASTA>)
  {
      chomp $_;
      my ($seq_id, @sequence) = split "\n";          # split the fasta input
    into Id and sequence
      my $fasta_sequence = join '',@sequence;        # reassembles the
    sequence
  print "\>$ARGV[1]\_$count\n$fasta_sequence\n";
  $count++;
  }
}
close FASTA;
```

```
#!/usr/bin/env perl

use strict;
use warnings;

use FindBin;
use lib ("$FindBin::Bin/../PerlLib");
use Fasta_reader;
use Data::Dumper;

=ExampleCommands
# make blastable
makeblastdb \
  -in refTranscripts.fasta\
      -out refTranscripts -dbtype nucl
# run blast+
blastn -query Trinity.fasta -db refTranscripts -out blastn.fmt6.txt \
    -evalue 1e-20 -dust no -task megablast -num_threads 2 -max_target_seqs 1 -
    outfmt 6
# analyze results
analyze_blastPlus_topHit_coverage.pl blastn.fmt6.txt refTranscripts.fasta
    Trinity.fasta
=cut

    ;


my $usage = "usage: $0 blast+.outfmt6.txt query.fasta search_db.fasta [
    output_prefix=NameOfBlastFileHere] [verbose=0]\n\n";

my $blast_out = $ARGV[0] or die $usage;
my $fasta_file_A = $ARGV[1] or die $usage;
my $fasta_file_B = $ARGV[2] or die $usage; # the fasta files don't have to be in
     any special order.
my $output_prefix = $ARGV[3] || "$blast_out";
```

```
32 my $verbose = $ARGV[4] || 0;
33
34 main: {
35
36
37     my $counter = 0;
38
39     my %query_to_top_hit; # only storing the hit with the greatest blast score.
40
41     # outfmt6:
42     # qseqid sseqid pident length mismatch gapopen qstart qend sstart send
       evalue bitscore
43
44     print STDERR "-parsing blast output: $blast_out\n" if $verbose;
45     open (my $fh, $blast_out) or die "Error, cannot open file $blast_out";
46     while (<$fh>) {
47         chomp;
48         my $line = $_;
49         my @x = split(/\t/);
50         my $query_id = $x[0];
51         my $db_id = $x[1];
52         my $percent_id = $x[2];
53         my $query_start = $x[6];
54         my $query_end = $x[7];
55         my $db_start = $x[8];
56         my $db_end = $x[9];
57
58         my $evalue = $x[10];
59         my $bitscore = $x[11];
60
61         if ( (! exists $query_to_top_hit{$query_id}) || ($bitscore >
       $query_to_top_hit{$query_id}->{bitscore}) ) {
62
63             $query_to_top_hit{$query_id} = { query_id => $query_id,
64                                              db_id => $db_id,
```

```
65                                          percent_id => $percent_id,
66                                          query_start => $query_start,
67                                          query_end => $query_end,
68                                          db_start => $db_start,
69                                          db_end => $db_end,
70                                          evalue => $evalue,
71                                          bitscore => $bitscore,
72
73                                          query_match_len => abs($query_end -
    $query_start) + 1,
74                                          db_match_len => abs($db_end -
   $db_start) + 1,
75
76                                          line => $line,
77                                      };
78
79          }
80
81          $counter++;
82          if ($counter % 100 == 0) {
83              print STDERR "\r[$counter]     " if $verbose;
84          }
85
86
87      }
88      close $fh;
89      $counter = 0;
90      print STDERR "\n" if $verbose;
91
92      ## identify those entries we need sequence length info for.
93      my %seq_lengths;
94      my %seq_headers;
95      {
96          foreach my $entry (values %query_to_top_hit) {
97              my $query_id = $entry->{query_id};
```

```
98            my $db_id = $entry->{db_id};
99
100           $seq_lengths{$query_id} = undef;
101           $seq_lengths{$db_id} = undef;
102       }
103
104       ## get sequence length info
105       foreach my $fasta_file ($fasta_file_A, $fasta_file_B) {
106
107           print STDERR "-parsing seq length info from file: $fasta_file\n" if
    $verbose;
108
109           my $fasta_reader = new Fasta_reader($fasta_file);
110
111           while (my $seq_obj = $fasta_reader->next()) {
112
113               $counter++;
114               if ($counter % 100 == 0) {
115                   print STDERR "\r[$counter]    " if $verbose;
116               }
117
118
119               my $acc = $seq_obj->get_accession();
120               if (exists $seq_lengths{$acc}) {
121
122                   my $sequence = $seq_obj->get_sequence();
123                   $seq_lengths{$acc} = length($sequence);
124
125                   my $header = $seq_obj->get_header();
126                   # remove the accession
127                   my @header_pieces = split(/\s+/, $header);
128                   shift @header_pieces;
129                   $header = join(" ", @header_pieces);
130                   $seq_headers{$acc} = $header;
131               }
```

```
132             }
133
134             $counter = 0;
135             print STDERR "\n" if $verbose;
136
137         }
138
139     }
140
141
142     ## analyze the results.
143     ## make this hit-centric, only retain the longest-coverage query hit for
    each database sequence.
144
145     print STDERR "-analyzing hits.\n" if $verbose;
146     my %db_id_to_greatest_pct_cov; # ties broken by bitscore
147     {
148
149
150         open (my $ofh, ">$output_prefix.w_pct_hit_length") or die $!;
151
152         print $ofh join("\t", "#qseqid", "sseqid", "pident", "length", "mismatch
    ",
153                         "gapopen", "qstart", "qend", "sstart", "send", "evalue",
     "bitscore",
154                         "db_hit_len", "pct_hit_len_aligned", "hit_descr") . "\n"
    ;
155
156         foreach my $entry (values %query_to_top_hit) {
157
158             my $db_id = $entry->{db_id};
159             my $db_match_len = $entry->{db_match_len};
160             my $db_seq_len = $seq_lengths{$db_id} or die "Error, no length found
     for $db_id, with hit: " . Dumper($entry);
161
```

```
162         my $percent_length_matched = sprintf("%.2f", $db_match_len /
    $db_seq_len * 100);
163
164         my $line = $entry->{line};
165         my $header = $seq_headers{$db_id};
166
167         print $ofh join("\t", $line, $db_match_len, $percent_length_matched,
     $header) . "\n";
168
169         $entry->{db_hit_pct_cov} = $percent_length_matched;
170
171
172         if ( ! exists $db_id_to_greatest_pct_cov{$db_id} ) {
173             $db_id_to_greatest_pct_cov{$db_id} = $entry;
174         }
175         else {
176             my $prev_entry = $db_id_to_greatest_pct_cov{$db_id};
177             if ($percent_length_matched > $prev_entry->{db_hit_pct_cov}
178                 ||
179                 ($percent_length_matched == $prev_entry->{db_hit_pct_cov}
180                  &&
181                  $entry->{bitscore} > $prev_entry->{bitscore})
182                 ) {
183                 $db_id_to_greatest_pct_cov{$db_id} = $entry;
184             }
185         }
186
187         $counter++;
188         if ($counter % 100 == 0) {
189             print STDERR "\r[$counter]     " if $verbose;
190         }
191
192     }
193     close $ofh;
194
```

```
195     }
196
197     $counter = 0;
198     print STDERR "\n" if $verbose;
199
200     ## histogram summary
201
202     my @bins = qw(10 20 30 40 50 60 70 80 90 100);
203     my %bin_counts;
204
205     open (my $ofh, ">$output_prefix.hist") or die "Error, cannot write to
        $output_prefix.hist";
206     open (my $list_ofh, ">$output_prefix.hist.list") or die $!;
207     {
208
209
210         foreach my $entry (values %db_id_to_greatest_pct_cov) {
211
212             my $pct_cov = $entry->{db_hit_pct_cov};
213
214             my $prev_bin = 0;
215             foreach my $bin (@bins) {
216                 if ($pct_cov > $prev_bin && $pct_cov <= $bin) {
217                     $bin_counts{$bin}++;
218                     print $list_ofh join("\t", "Bin_$bin", $entry->{line}) . "\n
        ";
219                 }
220                 $prev_bin = $bin;
221             }
222
223
224         }
225     }
226     close $list_ofh;
227
```

```
228     ## Report counts per bin
229     print "#hit_pct_cov_bin\tcount_in_bin\t>bin_below\n";
230     print $ofh "#hit_pct_cov_bin\tcount_in_bin\t>bin_below\n";
231
232     my $cumul = 0;
233     foreach my $bin (reverse(@bins)) {
234         my $count = $bin_counts{$bin} || 0;
235         $cumul += $count;
236         print join("\t", $bin, $count, $cumul) . "\n";
237         print $ofh join("\t", $bin, $count, $cumul) . "\n";
238
239     }
240     close $ofh;
241
242
243     exit(0);
244
245
246 }
```

```
#!/usr/bin/perl
use warnings;
use strict;


use Getopt::Long;
# Richard Emes University of Nottingham 2016

my $usage = "
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

R D Emes University of Nottingham 2016
pull genes from a reference genome GFF file and a fasta file of resequenced DNA.
copes with soft clipping meaning reads map over ends of transcript/genome [in this case will return to end of reference fasta]
USAGE:
-f  fasta file of genomic DNA or transcriptome
-g  gtf file where columns 4/5 are start end points
-o  output
-n  name of entity to parse e.g \"transcript\" \"exon\" etc must match column 3 of GTF
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

" ;


my ($file , $gtf , $name, $out);

GetOptions(
    'f|fasta:s'     => \$file ,
    'g|gtf:s'   => \$gtf ,
  'n|name:s'   => \$name,
  'o|output:s'   => \$out,
                 );

```

```
32
33 if( ! defined $file) {
34 print "$usage\nWARNING: Cannot proceed without fasta file to process\n\n"; exit;
35 }
36 if( ! defined $gtf) {
37 print "$usage\nWARNING: Cannot proceed without genomic DNA file\n\n"; exit;
38 }
39 if( ! defined $name) {
40 print "$usage\nWARNING: Cannot proceed without name of entity to parse\n\n";
       exit;
41 }
42 if( ! defined $out) {
43 print "$usage\nWARNING: Cannot proceed without output file\n\n"; exit;
44 }
45 ##################################################
46
47 my %lookup; # hash of arrays key is fasta sequence name array contains details
       in single string
48 ## read GTF
49 open FILE, "<$gtf";
50 while (<FILE>)
51 {
52 chomp $_;
53 unless ($_ =~ /^\#/)
54   {
55   my $line = $_;
56   my @data = split '\t', $line;
57
58   if ($data[2] eq $name)
59     {
60     my $name = $data[0];
61     my $start = $data[3];
62     my $end = $data[4];
63     my $details = $data[8];
64
```

```
65      $details =~ s/\"//g;
66      $details =~ s/\;//g;
67      $details =~ s/\'//g;
68      if ($details =~ /(.*?)\s$/) {$details = $1};
69      my $hash_details = $start."@".$end."@".$details;
70      push(@{$lookup{$name}}, $hash_details);
71      }
72    }
73  }
74  close FILE;
75
76
77
78  open OUT, ">$out";
79  # read in fasta file
80  my $fasta_sequence;
81
82  {
83    open FASTA, "<$file";
84    {
85    local $/ = '>';
86    <FASTA>;                    # throw away the first line 'cos will only contain "
        >"
87
88    while (<FASTA>)
89      {
90      chomp $_;
91      my ($seq_id, @sequence) = split "\n"; # split the fasta input into Id and
        sequence
92      $fasta_sequence = join '',@sequence;  # reassembles the sequence
93      my $seq_length = length $fasta_sequence;
94
95      if (exists $lookup{$seq_id})
96        {
97        foreach (@{$lookup{$seq_id}})
```

```
98          {
99          chomp $_;
100         my ($start,$end,$details) = split '\@', $_;
101
102         if ($end > $seq_length){$end = $seq_length;} # in case of soft clipping
    of reads map over the end of contig
103         if ($start < 1){$start = 1} # in case of soft clipping of reads map over
     the end of contig
104         $details =~ s/gene_id //g;
105         $details =~ s/ transcript_id /_/g;
106         my $length = ($end-$start)+1;   # to account for substr start and end
107         my $sub_start = $start-1 ;    # to account for substr start and end
108         my $seq = substr($fasta_sequence, $sub_start, $length); #$seq, start,
    length of substring)
109         print OUT "\>$details\n$seq\n";
110         }
111       }
112     }
113   close FASTA;
114 }
115 }
116 close OUT;
```

```
# NIPA a robust set of tools for analyis of gene lists.

#biocLite("biomaRt")
#biocLite("GOstats")
#biocLite("ReactomePA")
#biocLite("gage")
#biocLite("pathview")
#biocLite("gageData")
#biocLite("ggplot2")
#biocLite("stringr")
#biocLite("dplyr")


source("http://www.bioconductor.org/biocLite.R")
library(DBI)
library(GOstats)
library(biomaRt)
library(pathview)
library(gage)
library(gageData)
library(ReactomePA)
library(ggplot2)
library(stringr)
library(dplyr)

################################################################################
## Input Variables -- USER TO CHANGE [START]
################################################################################
goi.column = 1 # if results are from analysis and are a column of a larger table
     give input column else will assume is column 1 or a single column assumes
    tab delimited
goi.header = "no" # "yes" or "no" if header on file

goi.list <- "~/Desktop/" # change to input gene list
working.directory = "~/Desktop/"  # change to working directory where you want
```

```
    output
34
35 species = "mouse"   #currently one of "mouse", "human", "rat", "pig", "zebrafish
    "
36 outfile.prefix <- "specific" # prefix attached to output files. Ap
37
38 # if not installed you will need to download the appropriate species
    bioconductor package below.
39 biocLite("org.Mm.eg.db") # for Mouse
40 #biocLite("org.Hs.eg.db") # for Human
41 # biocLite("org.Rn.eg.db") # for Rat
42 # biocLite("org.Ss.eg.db") # for Pig
43 #biocLite("org.Dr.eg.db") # for Zebrafish
44
45 id.type = "ENSG"      # one of
46 # "ENSG" (ensembl gene),
47 # "ENST" (ensembl trasncript),
48 # "ENSP" (ensembl peptide),
49 # "Entrez"
50 # "Uniprot" (UniProt/SwissProt Accession)
51 # "Unigene"
52 # "Refseq_mrna" (RefSeq mRNA [e.g. NM_001195597])
53 # "Refseq_peptide" (RefSeq Protein ID [e.g. NP_001005353])
54
55
56 # set variables for hypergeometric cutoff enrichment qval less than this and
    with greater or equal to minimum number of genes in pathway or GO term will
    be drawn
57 kegg.qval.cutoff = 0.1
58 GO.cutoff = 0.05
59 min.genes.cutoff = 2
60
61 # change below to determine which test to conduct.
62 doGO = "yes" # yes or no.      Run GoStats hypergeometric test to find enriched
     GO terms in BP, MF and CC category
```

```
63 doReactome = "yes" # yes or no. Run ReactomePA to find enriched pathways in
       Reactomedb -- BIT SLOWER
64 doKEGG = "yes" # yes or no.      Run hypergeometric test to find and plot
       enriched KEGG pathways and visualise using PathView
65
66
67 # colour pathways by expression fold change?
68 keggFC = "yes" # yes or no. will colour enriched KEGG pathways by FC data [
       specify column below]
69 keggFC.col = 9 # if keggFC = yes specify column of input table with FC values
       assumes tab delimited
70 ##############################################################################
71 ## Input Variables -- USER TO CHANGE [END]
72 ##############################################################################
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87 ##############################################################################
88 # Dont alter below this line
89 ##############################################################################
90
91 ##############################################################################
92 ## set variables based on species given
93 ##############################################################################
```

```
94  if (species == "mouse")
95  {
96
97    library(org.Mm.eg.db)
98    ensembl.spp <- "mmusculus_gene_ensembl"
99    species.ens.code = "Mm"
100   species.kegg.code = "mmu"
101   kegg.data.code = "mm"
102   reactome.spp = "mouse" #one of "human", "rat", "mouse", "celegans", "yeast", "
        zebrafish", "fly".
103   kegg.gsets.spp <- kegg.gsets(species = "mmu", id.type = "kegg")
104 }
105
106 if (species == "human")
107 {
108   library(org.Hs.eg.db)
109   ensembl.spp <- "hsapiens_gene_ensembl"
110   species.ens.code = "Hs"
111   species.kegg.code = "hsa"
112   kegg.data.code = "hsa"
113   reactome.spp = "human" #one of "human", "rat", "mouse", "celegans", "yeast", "
        zebrafish", "fly".
114   kegg.gsets.spp <- kegg.gsets(species = "hsa", id.type = "kegg")
115 }
116
117 if (species == "rat")
118 {
119   library(org.Rn.eg.db)
120   ensembl.spp <- "rnorvegicus_gene_ensembl"
121   species.ens.code = "Rn"
122   species.kegg.code = "rno"
123   kegg.data.code = "rno"
124   reactome.spp = "rat" #one of "human", "rat", "mouse", "celegans", "yeast", "
        zebrafish", "fly".
125   kegg.gsets.spp <- kegg.gsets(species = "rno", id.type = "kegg")
```

```
126 }
127
128 if (species == "pig")
129 {
130   library(org.Ss.eg.db)
131   ensembl.spp <- "sscrofa_gene_ensembl"
132   species.ens.code = "Ss"
133   species.kegg.code = "ssc"
134   kegg.data.code = "ssc"
135   #reactome.spp = "rat" #one of "human", "rat", "mouse", "celegans", "yeast", "
        zebrafish", "fly".
136   doReactome = "no"
137   kegg.gsets.spp <- kegg.gsets(species = "ssc", id.type = "kegg")
138 }
139
140 if (species == "zebrafish")
141 {
142   library(org.Dr.eg.db)
143   ensembl.spp <- "drerio_gene_ensembl"
144   species.ens.code = "Dr"
145   species.kegg.code = "dre"
146   kegg.data.code = "dre"
147   #reactome.spp = "rat" #one of "human", "rat", "mouse", "celegans", "yeast", "
        zebrafish", "fly".
148   doReactome = "no"
149   kegg.gsets.spp <- kegg.gsets(species = "dre", id.type = "kegg")
150 }
151
152 ##############################################################################
153 # Build kegg sets
154 ##############################################################################
155 kegg.sets.test <- kegg.gsets.spp$kg.sets
156 kegg.sets.spp = kegg.gsets.spp$sigmet.idx
157
158 ##############################################################################
```

```
# Get Data
#############################################################################
setwd(working.directory)

if (goi.header == "yes") {my.data.in <- read.table(goi.list,sep='\t',header =
    TRUE)}
if (goi.header == "no") {my.data.in <- read.table(goi.list,sep='\t',header =
    FALSE)}
myInterestingGenes <- as.vector(unlist(my.data.in[goi.column]))
myInterestingGenes <- unique(myInterestingGenes)

species.db <- paste("org",species.ens.code,"eg.db",sep=".")
ensembl = useEnsembl(biomart="ensembl", dataset=ensembl.spp)



#############################################################################
# Convert IDs to Entrez IDs and match to gene input list
#############################################################################

if (id.type =="ENSG")
{
  all.genes <- getBM(attributes=c('ensembl_gene_id', 'entrezgene', 'external_
    gene_name'), mart = ensembl)
  colnames(all.genes) <- c("ID","Entrez","Name")
  all.genes.entrez <- na.omit(all.genes)
  all.genes.entrez <- all.genes.entrez[all.genes.entrez$ID!="",]
  goi.entrez <-unique(as.character(all.genes.entrez[all.genes.entrez$ID %in%
    myInterestingGenes,2]))
}

if (id.type =="ENSP")
{
  all.genes <- getBM(attributes=c('ensembl_peptide_id', 'entrezgene', 'external_
    gene_name'), mart = ensembl)
```

```
189   colnames(all.genes) <- c("ID","Entrez","Name")
190   all.genes.entrez <- na.omit(all.genes)
191   all.genes.entrez <- all.genes.entrez[all.genes.entrez$ID!="",]
192   goi.entrez <-unique(as.character(all.genes.entrez[all.genes.entrez$ID %in%
        myInterestingGenes,2]))
193 }
194
195 if (id.type =="ENST")
196 {
197   all.genes <- getBM(attributes=c('ensembl_transcript_id', 'entrezgene', '
        external_gene_name'), mart = ensembl)
198   colnames(all.genes) <- c("ID","Entrez","Name")
199   all.genes.entrez <- na.omit(all.genes)
200   all.genes.entrez <- all.genes.entrez[all.genes.entrez$ID!="",]
201   goi.entrez <-unique(as.character(all.genes.entrez[all.genes.entrez$ID %in%
        myInterestingGenes,2]))
202 }
203
204 if (id.type == "Entrez")
205 {
206   all.genes <- getBM(attributes=c('entrezgene', 'entrezgene', 'external_gene_
        name'), mart = ensembl)
207   colnames(all.genes) <- c("ID","Entrez","Name")
208   all.genes.entrez <- na.omit(all.genes)
209   all.genes.entrez <- all.genes.entrez[all.genes.entrez$ID!="",]
210   goi.entrez <-unique(as.character(all.genes.entrez[all.genes.entrez$ID %in%
        myInterestingGenes,2]))
211 }
212
213 if (id.type == "Refseq_mrna")
214 {
215   all.genes <- getBM(attributes=c('refseq_mrna', 'entrezgene', 'external_gene_
        name'), mart = ensembl)
216   colnames(all.genes) <- c("ID","Entrez","Name")
217   all.genes.entrez <- na.omit(all.genes)
```

```
218   all.genes.entrez <- all.genes.entrez[all.genes.entrez$ID!="",]
219   goi.entrez <-unique(as.character(all.genes.entrez[all.genes.entrez$ID %in%
        myInterestingGenes,2]))
220 }
221
222 if (id.type == "Refseq_peptide")
223 {
224   all.genes <- getBM(attributes=c('refseq_peptide', 'entrezgene', 'external_gene
        _name'), mart = ensembl)
225   colnames(all.genes) <- c("ID","Entrez","Name")
226   all.genes.entrez <- na.omit(all.genes)
227   all.genes.entrez <- all.genes.entrez[all.genes.entrez$ID!="",]
228   goi.entrez <-unique(as.character(all.genes.entrez[all.genes.entrez$ID %in%
        myInterestingGenes,2]))
229 }
230
231 if (id.type == "Unigene")
232 {
233   all.genes <- getBM(attributes=c('unigene', 'entrezgene', 'external_gene_name')
        , mart = ensembl)
234   colnames(all.genes) <- c("ID","Entrez","Name")
235   all.genes.entrez <- na.omit(all.genes)
236   all.genes.entrez <- all.genes.entrez[all.genes.entrez$ID!="",]
237   goi.entrez <-unique(as.character(all.genes.entrez[all.genes.entrez$ID %in%
        myInterestingGenes,2]))
238 }
239
240 if (id.type == "Uniprot")
241 {
242   all.genes <- getBM(attributes=c('uniprot_swissprot', 'entrezgene', 'external_
        gene_name'), mart = ensembl)
243   colnames(all.genes) <- c("ID","Entrez","Name")
244   all.genes.entrez <- na.omit(all.genes)
245   all.genes.entrez <- all.genes.entrez[all.genes.entrez$ID!="",]
246   goi.entrez <-unique(as.character(all.genes.entrez[all.genes.entrez$ID %in%
```

```
    myInterestingGenes,2]))
}


# if keggFC = yes create foldchanges named list of log fold change values
if (keggFC == "yes")
{
  entrez.FC.match <- merge(all.genes.entrez,my.data.in,by.x="ID",by.y=names(my.
    data.in[goi.column]))
  #foldchanges = unlist(entrez.FC.match[keggFC.col+2])
  #foldchanges = unlist(entrez.FC.match$Entrez)
  foldchanges = apply(entrez.FC.match,2,unlist)
  names(foldchanges) = entrez.FC.match$Entrez
}


############################################################
# Set gene "universse" of all genes
universe <- unique(as.character(all.genes.entrez$Entrez))
############################################################


#############################################################################
# start report and set up variables to catch failing sections.
#############################################################################
run.report = paste(outfile.prefix,"NIPA.report.txt",sep=".")

cat(c("-----------------------------------------------------------","The NIPA
    run has initiated: Any warnings will appear below.","
    -----------------------------------------------------------"), file=run.
    report, append=FALSE, sep = "\n")

if (length(goi.entrez)==0 )
{
  cat(c("The run has terminated","Conversion of gene/peptide list to entrez
```

```
    failed", "Are the IDs properly formatted or possibly too few IDs"),
277       file=run.report, append=TRUE, sep='\n')
278   stop("Run terminated, see NIPA.report.txt")
279 }
280
281 # set flags to capture failed sections.
282 fail.GO.MF = 0
283 fail.GO.BP = 0
284 fail.GO.CC = 0
285 fail.reactome = 0
286 fail.KEGG = 0
287 stats.KEGG.fail = 0
288
289 ##############################################################################
290 #
291 # part 1 GO analysis
292 #
293 ##############################################################################
294 ##############################################################################
295 ##########################################################
296
297 if (doGO == "yes")
298 {
299   # Biological Process
300   params.BP <- new('GOHyperGParams',
301                    geneIds=goi.entrez,
302                    universeGeneIds=universe,
303                    ontology='BP',
304                    pvalueCutoff=GO.cutoff,
305                    conditional=F,
306                    testDirection='over',
307                    annotation=species.db
308   )
309   hgOver.BP <- hyperGTest(params.BP)
310   result.BP <- summary(hgOver.BP)
```

```
311
312
313 if (nrow(result.BP)==0 )
314 {
315   fail.GO.BP = 1
316   cat(c("GO Biological process search identified no enriched terms","Probably
    too few IDs"),
317       file=run.report, append=TRUE, sep='\n')
318 }
319 if (fail.GO.BP !=1)
320 {
321   result.BP <- result.BP[result.BP$Count >= min.genes.cutoff,] # filter those
    with < cut off count
322   result.BP <- result.BP[order(result.BP$Pvalue),] # order by Pvalue
323
324   top.result.BP <- head(result.BP,10)
325   top.result.BP$Term <- as.factor(top.result.BP$Term)
326   top.result.BP$Term <- factor(top.result.BP$Term, levels = top.result.BP$Term
    )
327
328   if (nrow(top.result.BP) > 0)
329   {
330     top.result.BP$Pvalue[top.result.BP$Pvalue == 0 ] <- 1e-10 # catches any
    where p value = 0
331     max.y.plot = 1.2*(max(-log10(top.result.BP$Pvalue)))
332     sig.BP.plot <-
333       ggplot(data = top.result.BP,
334              aes(x = as.factor(Term), y = -log10(top.result.BP$Pvalue),
335                  colour = Count,
336                  scale_colour_gradient(low="blue"),
337                  size = Count))+
338       geom_point() +
339       scale_color_continuous("GOI count")+
340       scale_size_continuous(range = c(5,20), guide=FALSE)+
341       scale_x_discrete(labels = function(x) str_wrap(x, width = 30))+
```

```
342         geom_hline(yintercept=1.30103,lty=2, color="grey") + # equivalent of p =
    0.05
343         geom_hline(yintercept=2,lty=4, color="grey") + # equivalent of p = 0.01
344         geom_hline(yintercept=3,lty=3, color="grey") + # equivalent of p = 0.001
345         coord_flip()+
346         geom_point(stat = "identity") +
347         theme_bw() +
348         theme(axis.text.x = element_text(colour = "black"),
349               panel.grid.major = element_blank(),
350               panel.grid.minor = element_blank(),
351               panel.background = element_rect(fill = "white")) +
352         ylim(-0.5,max.y.plot)+
353         xlab("") +
354         ylab("Enrichment (-log10 pvalue)")
355
356       BP.plot.out = paste(outfile.prefix,"GO.BP.Significant.enrichment.plot.pdf"
    ,sep=".")
357       pdf(BP.plot.out)
358       print(sig.BP.plot)
359       dev.off()
360     }
361
362     # Add gene names to results table
363     allgos.BP <- geneIdUniverse(hgOver.BP)
364     output.BP.match <- NULL
365     for (i in 1:nrow(result.BP))
366     {
367
368       go.holding = result.BP$GOBPID[i]
369       all.entrez.in.GO <- as.vector(unlist(allgos.BP[go.holding]))
370       goi.entrez.in.GO <- intersect(all.entrez.in.GO,goi.entrez)
371       input.in.GO.IDs <- all.genes[all.genes$Entrez %in% goi.entrez.in.GO, 1]
372       input.in.GO.IDs <- unique(input.in.GO.IDs[input.in.GO.IDs != ""])
373       input.in.GO.IDs <- paste(input.in.GO.IDs, collapse = " ")
374       input.in.GO.external <- unique(all.genes[all.genes$Entrez %in% goi.entrez.
```

```
    in.GO, 3])
375     input.in.GO.external <- paste(input.in.GO.external, collapse = " ")
376     temp <- cbind(go.holding,input.in.GO.IDs,input.in.GO.external)
377     output.BP.match <- rbind(output.BP.match,temp)
378   }
379   result.BP <- merge(result.BP, output.BP.match, by.x ="GOBPID", by.y="go.
    holding", all.x=TRUE)
380   BP.table.out = paste(outfile.prefix,"GO.BP.table",sep=".")
381   result.BP <- result.BP[order(result.BP$Pvalue),] # order by Pvalue
382   write.table(result.BP, file=BP.table.out, row.names = FALSE, col.names=TRUE,
    sep = '\t', quote=FALSE)
383
384 }
385
386
387
388 # Molecular Function
389 params.MF <- new('GOHyperGParams',
390                  geneIds=goi.entrez,
391                  universeGeneIds=universe,
392                  ontology='MF',
393                  pvalueCutoff=GO.cutoff,
394                  conditional=F,
395                  testDirection='over',
396                  annotation=species.db
397 )
398 hgOver.MF <- hyperGTest(params.MF)
399 result.MF <- summary(hgOver.MF)
400
401 if (nrow(result.MF)==0 )
402 {
403   fail.GO.MF = 1
404   cat(c("GO Molecular Function search identified no enriched terms","Probably
    too few IDs"),
405       file=run.report, append=TRUE, sep='\n')
```

```
406 }
407 if (fail.GO.MF !=1)
408 {
409   result.MF <- result.MF[result.MF$Count >= min.genes.cutoff,] # filter those
    with < cut off count
410   result.MF <- result.MF[order(result.MF$Pvalue),] # order by Pvalue
411
412   top.result.MF <- head(result.MF,10)
413   top.result.MF$Term <- as.factor(top.result.MF$Term)
414   top.result.MF$Term <- factor(top.result.MF$Term, levels = top.result.MF$Term
    )
415
416   if (nrow(top.result.MF) > 0)
417   {
418     top.result.MF$Pvalue[top.result.MF$Pvalue == 0 ] <- 1e-10 # catches any
    where p value = 0
419     max.y.plot = 1.2*(max(-log10(top.result.MF$Pvalue)))
420     sig.MF.plot <-
421       ggplot(data = top.result.MF,
422              aes(x = as.factor(Term), y = -log10(top.result.MF$Pvalue),
423                  colour = Count,
424                  scale_colour_gradient(low="blue"),
425                  size = Count))+
426       geom_point() +
427       scale_color_continuous("GOI count")+
428       scale_size_continuous(range = c(5,20), guide=FALSE)+
429       scale_x_discrete(labels = function(x) str_wrap(x, width = 30))+
430       geom_hline(yintercept=1.30103,lty=2, color="grey") + # equivalent of p =
     0.05
431       geom_hline(yintercept=2,lty=4, color="grey") + # equivalent of p = 0.01
432       geom_hline(yintercept=3,lty=3, color="grey") + # equivalent of p = 0.001
433       coord_flip()+
434       geom_point(stat = "identity") +
435       theme_bw() +
436       theme(axis.text.x = element_text(colour = "black"),
```

```
437             panel.grid.major = element_blank(),
438             panel.grid.minor = element_blank(),
439             panel.background = element_rect(fill = "white")) +
440       ylim(-0.5,max.y.plot)+
441       xlab("") +
442       ylab("Enrichment (-log10 pvalue)")
443
444     MF.plot.out = paste(outfile.prefix,"GO.MF.Significant.enrichment.plot.pdf"
    ,sep=".")
445     pdf(MF.plot.out)
446     print(sig.MF.plot)
447     dev.off()
448   }
449
450   # Add gene names to results table
451   allgos.MF <- geneIdUniverse(hgOver.MF)
452   output.MF.match <- NULL
453   for (i in 1:nrow(result.MF))
454   {
455     go.holding = result.MF$GOMFID[i]
456     all.entrez.in.GO <- as.vector(unlist(allgos.MF[go.holding]))
457     goi.entrez.in.GO <- intersect(all.entrez.in.GO,goi.entrez)
458     input.in.GO.IDs <- all.genes[all.genes$Entrez %in% goi.entrez.in.GO, 1]
459     input.in.GO.IDs <- unique(input.in.GO.IDs[input.in.GO.IDs != ""])
460     input.in.GO.IDs <- paste(input.in.GO.IDs, collapse = " ")
461     input.in.GO.external <- unique(all.genes[all.genes$Entrez %in% goi.entrez.
    in.GO, 3])
462     input.in.GO.external <- paste(input.in.GO.external, collapse = " ")
463     temp <- cbind(go.holding,input.in.GO.IDs,input.in.GO.external)
464     output.MF.match <- rbind(output.MF.match,temp)
465   }
466   result.MF <- merge(result.MF, output.MF.match, by.x ="GOMFID", by.y="go.
    holding", all.x=TRUE)
467   MF.table.out = paste(outfile.prefix,"GO.MF.table",sep=".")
468   result.MF <- result.MF[order(result.MF$Pvalue),] # order by Pvalue
```

```
469     write.table(result.MF, file=MF.table.out, row.names = FALSE, col.names=TRUE,
        sep = '\t', quote=FALSE)
470   }
471
472   # Cellular Compartment
473   params.CC <- new('GOHyperGParams',
474                    geneIds=goi.entrez,
475                    universeGeneIds=universe,
476                    ontology='CC',
477                    pvalueCutoff=GO.cutoff,
478                    conditional=F,
479                    testDirection='over',
480                    annotation=species.db
481   )
482   hgOver.CC <- hyperGTest(params.CC)
483   result.CC <- summary(hgOver.CC)
484
485   if (nrow(result.CC)==0 )
486   {
487     fail.GO.CC = 1
488     cat(c("GO Cellular location search identified no enriched terms","Probably
        too few IDs"),
489         file=run.report, append=TRUE, sep='\n')
490   }
491
492   if (fail.GO.CC !=1)
493   {
494     result.CC <- result.CC[result.CC$Count >= min.genes.cutoff,] # filter those
        with < cut off count
495     result.CC <- result.CC[order(result.CC$Pvalue),] # order by Pvalue
496     top.result.CC <- head(result.CC,10)
497     top.result.CC$Term <- as.factor(top.result.CC$Term)
498     top.result.CC$Term <- factor(top.result.CC$Term, levels = top.result.CC$Term
        )
499
```

```
500   if (nrow(top.result.CC) > 0)
501   {
502     top.result.CC$Pvalue[top.result.CC$Pvalue == 0 ] <- 1e-10 # catches any
    where p value = 0
503     max.y.plot = 1.2*(max(-log10(top.result.CC$Pvalue)))
504     sig.CC.plot <-
505       ggplot(data = top.result.CC,
506              aes(x = as.factor(Term), y = -log10(top.result.CC$Pvalue),
507                  colour = Count,
508                  scale_colour_gradient(low="blue"),
509                  size = Count))+
510       geom_point() +
511       scale_color_continuous("GOI count")+
512       scale_size_continuous(range = c(5,20), guide=FALSE)+
513       scale_x_discrete(labels = function(x) str_wrap(x, width = 30))+
514       geom_hline(yintercept=1.30103,lty=2, color="grey") + # equivalent of p =
     0.05
515       geom_hline(yintercept=2,lty=4, color="grey") + # equivalent of p = 0.01
516       geom_hline(yintercept=3,lty=3, color="grey") + # equivalent of p = 0.001
517       coord_flip()+
518       geom_point(stat = "identity") +
519       theme_bw() +
520       theme(axis.text.x = element_text(colour = "black"),
521             panel.grid.major = element_blank(),
522             panel.grid.minor = element_blank(),
523             panel.background = element_rect(fill = "white")) +
524       ylim(-0.5,max.y.plot)+
525       xlab("") +
526       ylab("Enrichment (-log10 pvalue)")
527     CC.plot.out = paste(outfile.prefix,"GO.CC.Significant.enrichment.plot.pdf"
    ,sep=".")
528     pdf(CC.plot.out)
529     print(sig.CC.plot)
530     dev.off()
531   }
```

```
532
533
534     # Add gene names to results table
535     allgos.CC <- geneIdUniverse(hgOver.CC)
536     output.CC.match <- NULL
537     for (i in 1:nrow(result.CC))
538     {
539       go.holding = result.CC$GOCCID[i]
540       all.entrez.in.GO <- as.vector(unlist(allgos.CC[go.holding]))
541       goi.entrez.in.GO <- intersect(all.entrez.in.GO,goi.entrez)
542       input.in.GO.IDs <- all.genes[all.genes$Entrez %in% goi.entrez.in.GO, 1]
543       input.in.GO.IDs <- unique(input.in.GO.IDs[input.in.GO.IDs != ""])
544       input.in.GO.IDs <- paste(input.in.GO.IDs, collapse = " ")
545       input.in.GO.external <- unique(all.genes[all.genes$Entrez %in% goi.entrez.
    in.GO, 3])
546       input.in.GO.external <- paste(input.in.GO.external, collapse = " ")
547       temp <- cbind(go.holding,input.in.GO.IDs,input.in.GO.external)
548       output.CC.match <- rbind(output.CC.match,temp)
549     }
550     result.CC <- merge(result.CC, output.CC.match, by.x ="GOCCID", by.y="go.
    holding", all.x=TRUE)
551     CC.table.out = paste(outfile.prefix,"GO.CC.table",sep=".")
552     result.CC <- result.CC[order(result.CC$Pvalue),] # order by Pvalue
553     write.table(result.CC, file=CC.table.out, row.names = FALSE, col.names=TRUE,
    sep = '\t', quote=FALSE)
554   }
555 }
556
557
558 ##############################################################################
559 ##############################################################################
560 #
561 # part 2 Pathway analysis
562 #
563 ##############################################################################
```

```
564 ##########################################################################
565
566 if (doReactome == "yes")
567 {
568   reactome.out <- enrichPathway(gene=goi.entrez,
569                                 #pvalueCutoff=0.05,
570                                 readable=T,
571                                 organism = reactome.spp,
572                                 pAdjustMethod = "BH",
573                                 qvalueCutoff = 0.01,
574                                 universe = universe
575   )
576
577   reactome.writeout <- (as.data.frame(reactome.out))
578
579   if (nrow(reactome.writeout) == 0)
580   {
581     fail.reactome = 1
582     cat(c("Reactome analysis identified no enriched pathways","Probably too few
      IDs"),
583         file=run.report, append=TRUE, sep='\n')
584   }
585
586   if (fail.reactome==0)
587   {
588     reactome.table.out = paste(outfile.prefix,"reactome.pathway.enrichment.table
      ",sep=".")
589     write.table(reactome.writeout, file=reactome.table.out, row.names = FALSE,
      col.names = TRUE, quote=FALSE, sep='\t')
590
591     reactome.dot <- dotplot <- dotplot(
592       reactome.out,
593       showCategory=15,
594       font.size = 12
595     )
```

```
596     reactome.plot.out = paste(outfile.prefix,"reactome.pathway.enrichment.
        dotplot.tiff",sep=".")
597     tiff(filename=reactome.plot.out,
598           width = 320,
599           height = 240,
600           units = "mm",
601           res=800
602     )
603     print(reactome.dot)
604     dev.off()
605 
606     reactome.map.out = paste(outfile.prefix,"reactome.pathway.enrichment.
        enrichmap.tiff",sep=".")
607     tiff(filename=reactome.map.out,
608           width = 320,
609           height = 240,
610           units = "mm",
611           res=800,
612           type = "Xlib",
613           pointsize = 12
614     )
615     enrichMap(reactome.out,
616                 layout=igraph::layout.kamada.kawai,
617                 vertex.label.cex = 0.8
618     )
619     dev.off()
620   }
621 }
622 
623 ####
624 ##############################################################################
625 ##############################################################################
626 #
627 # part 3 KEGG analysis
628 #
```

```
629 ##############################################################################
630 ##############################################################################
631
632 if (doKEGG == "yes")
633 {
634   pathview.goi.entrez <- rep.int(1, length(goi.entrez))
635
636   names(pathview.goi.entrez) = goi.entrez
637
638   keggres = gage(pathview.goi.entrez, gsets=kegg.sets.test, same.dir=TRUE) #
        determine kegg membership of all genes.
639
640   keggres.pathways <- as.data.frame(keggres)
641   keggres.pathways.out <- keggres.pathways[keggres.pathways$greater.set.size >
        0, ]
642
643   if (nrow(keggres.pathways.out) ==0)
644   {
645     fail.KEGG = 1
646     cat(c("KEGG analysis identified no enriched pathways","Probably too few IDs"
        ),
647         file=run.report, append=TRUE, sep='\n')
648   }
649
650   if (fail.KEGG ==0)
651   {
652     keggres.pathways.out$KEGGpathways <- rownames(keggres.pathways.out)
653     matching.kegg.sets.spp <- kegg.sets.test[c(keggres.pathways.out$KEGGpathways
        )] # named list of matched pathways
654     matching.kegg.sets.spp.total.size <- lengths(matching.kegg.sets.spp, use.
        names = TRUE) # named list of the number of total number of genes in matched
        pathway.
655
656
657     matching.kegg.sets.spp.df <- as.data.frame(unlist(matching.kegg.sets.spp,
```

```
    use.names = TRUE))
658 matching.kegg.sets.spp.df$kegg.id <-  gsub("\\d+$", "", rownames(matching.
    kegg.sets.spp.df))
659 row.names(matching.kegg.sets.spp.df) <- NULL
660 colnames(matching.kegg.sets.spp.df) <- c("entrez.id","kegg.id")
661
662 # make subset of matching.kegg.sets.spp.df with just genes of interest in it
663 goi.matching.kegg.sets.spp.df <- matching.kegg.sets.spp.df[matching.kegg.
    sets.spp.df$entrez.id %in% goi.entrez, ]
664
665 #
    ##############################################################################
666 # Stats details
667 #
    ##############################################################################
668 # for each pathway with > 0 goi in it, conduct a hypergeometric test using
    phyper
669 # phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)
670 # x, q vector of quantiles representing the number of white balls drawn
671 # without replacement from an urn which contains both black and white
672 # balls.
673 # m the number of white balls in the urn.
674 # n the number of black balls in the urn.
675 # k the number of balls drawn from the urn.
676 # if
677 # pop size : 5260 # total number of entrez gene in all pathways
678 # sample size : 131 # total goi
679 # Number of items in the pop that are classified as successes : 1998 #
    entrez in a particular pathway
680 # Number of items in the sample that are classified as successes : 62 # goi
    in a particular pathway
681 #
682 # phyper(62,1998,5260-1998,131)
```

```
683   # e.g pathway 100 genes 10 are in goi list of size 400 universe = 20,000
684   # phyper(1,100,20000-100,400, lower.tail=FALSE) = 0.597 = probability of
      finding this many or greater goi in pathway
685   # phyper(80,100,20000-100,400, lower.tail=FALSE) = 4.603708e-122 =
      probability of finding this many or greater goi in pathway
686   #
      ##########################################################################################

687   #
688   #
      ##########################################################################################

689
690
691   universe.size = as.numeric(length(universe))
692   total.goi.size = as.numeric(length(goi.entrez))
693
694
695
696   # do for each pathway in list and generate table of pathways passing cut off
       after FDR qvalue calculation
697   working.pathways <- unique(matching.kegg.sets.spp.df$kegg.id)
698
699   pathways.hypergeometric.results <- data.frame("Pathway"= character(0),"p.val
      "= numeric(0),"FDR q.val"= numeric(0),"ID"= character(0), "entrez.ids"=
      numeric(0), "external.ids"= character(0))
700   pathways.hypergeometric.results.sig <- data.frame("Pathway"= character(0),"p
      .val"= numeric(0),"FDR q.val"= numeric(0), "goi.count"= numeric(0))
701
702   detach("package:dplyr") # to overcome occasional issues of pathview clashing
       with dplyr
703
704   for (i in 1:length(working.pathways))
705   {
706     current.pathway = working.pathways[i]
```

```
707     goi.in.pathway <- as.numeric(nrow(goi.matching.kegg.sets.spp.df[goi.
    matching.kegg.sets.spp.df$kegg.id == current.pathway, ]))
708     total.genes.in.pathway <- as.numeric(nrow(matching.kegg.sets.spp.df[
    matching.kegg.sets.spp.df$kegg.id == current.pathway, ]))
709
710     pval <- phyper(goi.in.pathway,total.genes.in.pathway,(universe.size-total.
    genes.in.pathway),total.goi.size, lower.tail=FALSE)
711     qval <- p.adjust(pval, method = "fdr", n = nrow(keggres.pathways.out))
712
713     current.goi <- goi.matching.kegg.sets.spp.df[goi.matching.kegg.sets.spp.df
    $kegg.id == current.pathway, ]
714     current.goi <- current.goi[1]
715     current.goi.entrez.ids <- as.numeric(as.character(current.goi$entrez.id))
716
717     current.goi.ens <- all.genes.entrez[all.genes.entrez$Entrez %in% current.
    goi.entrez.ids ,]
718
719     current.goi.ens.ids <- unique(current.goi.ens$ID)
720     current.goi.ext.ids <- unique(current.goi.ens$Name)
721
722
723     current.goi.ens.ids <- paste(current.goi.ens.ids, collapse=", ")
724     current.goi.entrez.ids <- paste(current.goi.entrez.ids, collapse=", ")
725     current.goi.ext.ids <- paste(current.goi.ext.ids, collapse=", ")
726
727     current.out <- as.data.frame(cbind(current.pathway,pval,qval,current.goi.
    ens.ids,current.goi.entrez.ids,current.goi.ext.ids))
728     current.sig.out <- as.data.frame(cbind(current.pathway,pval,qval,goi.in.
    pathway))
729
730     pathways.hypergeometric.results <- rbind(pathways.hypergeometric.results,
    current.out)
731
732
733
```

```
734       if (qval < kegg.qval.cutoff & goi.in.pathway >= min.genes.cutoff)
735       {
736         pid <- substr(current.pathway, start=1, stop=8) # get kegg ids
737
738         if (keggFC == "yes")
739         {
740           pathview(gene.data=foldchanges, pathway.id=pid, species=species.kegg.
    code)
741         }
742
743         if (keggFC == "no")
744         {
745           pathview(gene.data=pathview.goi.entrez, pathway.id=pid, species=
    species.kegg.code)
746         }
747
748         pathways.hypergeometric.results.sig <- rbind(pathways.hypergeometric.
    results.sig, current.sig.out)
749       }
750
751
752     }
753     library(dplyr)
754     colnames(pathways.hypergeometric.results) <- c("Pathway","p.val","FDR q.val"
    ,"Ensembl.ids","Entrez.ids","External.ids")
755
756     # make FDR q.val numeric and sort
757     pathways.hypergeometric.results$`FDR q.val` <- as.numeric(as.character(
    pathways.hypergeometric.results$`FDR q.val`))
758     pathways.hypergeometric.results <-  pathways.hypergeometric.results[with(
    pathways.hypergeometric.results, order(pathways.hypergeometric.results$`FDR q
    .val`)), ]
759
760     kegg.table.out = paste(outfile.prefix,"kegg.pathway.enrichment.table",sep=".
    ")
```

```
761   write.table(pathways.hypergeometric.results,file=kegg.table.out, row.names =
      FALSE, col.names = TRUE, quote = FALSE, sep ='\t')
762
763   #
      ##############################################################################################
      ##############################################################################################
764   # draw plot of enriched pathways
765   #
      ##############################################################################################
      ##############################################################################################
766   colnames(pathways.hypergeometric.results.sig) <- c("Pathway","p.val","FDR q.
      val","goi.count")
767
768   if (nrow(pathways.hypergeometric.results.sig)>0)
769   {
770
771
772     pathways.hypergeometric.results.sig$`FDR q.val` <- as.numeric(as.character
      (pathways.hypergeometric.results.sig$`FDR q.val`))
773     pathways.hypergeometric.results.sig <-  pathways.hypergeometric.results.
      sig[with(pathways.hypergeometric.results.sig, order(pathways.hypergeometric.
      results.sig$`FDR q.val`)), ]
774
775     pathways.hypergeometric.results.sig$goi.count <- as.numeric(as.character(
      pathways.hypergeometric.results.sig$goi.count))
776     pathways.hypergeometric.results.sig <-  pathways.hypergeometric.results.
      sig[with(pathways.hypergeometric.results.sig, order(pathways.hypergeometric.
      results.sig$`FDR q.val`)), ]
777
778
779     top.pathways.hypergeometric.results.sig <- head(pathways.hypergeometric.
      results.sig,10)
780     top.pathways.hypergeometric.results.sig$Pathway <- factor(top.pathways.
      hypergeometric.results.sig$Pathway, levels = top.pathways.hypergeometric.
      results.sig$Pathway)
```

```
781     max.y.plot = 1.2*(max(-log10(top.pathways.hypergeometric.results.sig$`FDR
    q.val`)))
782
783     sig.kegg.plot <-
784       ggplot(data = top.pathways.hypergeometric.results.sig,
785              aes(x = as.factor(Pathway), y = -log10(top.pathways.
    hypergeometric.results.sig$`FDR q.val`),
786                  colour = goi.count,
787                  scale_colour_gradient(low="blue"),
788                  size = goi.count))+
789       geom_point() +
790       scale_color_continuous("GOI count")+
791       scale_size_continuous(range = c(5,20), guide=FALSE)+
792       scale_x_discrete(labels = function(x) str_wrap(x, width = 30))+
793       geom_hline(yintercept=1.30103,lty=2, color="grey") + # equivalent of p =
     0.05
794       geom_hline(yintercept=2,lty=4, color="grey") + # equivalent of p = 0.01
795       geom_hline(yintercept=3,lty=3, color="grey") + # equivalent of p = 0.001
796       coord_flip()+
797       geom_point(stat = "identity") +
798       theme_bw() +
799       theme(axis.text.x = element_text(colour = "black"),
800             panel.grid.major = element_blank(),
801             panel.grid.minor = element_blank(),
802             panel.background = element_rect(fill = "white")) +
803       ylim(-0.5,max.y.plot)+
804       xlab("") +
805       ylab("Enrichment (-log10 pvalue)")
806
807     kegg.pdf.out = paste(outfile.prefix,"KEGG.Significant.enrichment.plot.pdf"
    ,sep=".")
808     pdf(kegg.pdf.out)
809     print(sig.kegg.plot)
810     dev.off()
811     stats.KEGG.fail = 1
```

```
812     }
813     if (stats.KEGG.fail == 0)
814     {
815       cat(c("KEGG analysis no terms pass statistical cutoff"),
816           file=run.report, append=TRUE, sep='\n')
817     }
818 
819   }
820 }
```

Appendix C

Table C.1: Gene symbol and function of 153 PSD genes found to be expressed in the mouse but not in the *de novo* zebrafish assembly.

	Gene symbol	Function
1	Brk1	protein complex binding
2	Brinp2	cellular response to retinoic acid
3	Hnrnpa2b1	nucleic acid binding
4	Tomm40l	porin activity
5	Ppp2ca	hydrolase activity
6	Rasgef1a	guanyl-nucleotide exchange factor activity
7	Copg1	structural molecule activity
8	Tuba8	GTP binding
9	Zwint	protein binding
10	Cst3	cysteine-type endopeptidase inhibitor activity
11	Arf5	GTP binding
12	Dydc2	histone methyltransferase activity (H3-K4 specific)
13	Akap5	calmodulin binding
14	Psma7	endopeptidase activity
15	Kras	GTP binding
16	Utrn	zinc ion binding
17	Tsc22d4	DNA binding transcription factor activity

Continued on next page

Table C.1 – *Continued from previous page*

	Ensembl Gene ID	Function
18	Hap1	protein binding
19	Fyco1	metal ion binding
20	Grk2	ATP binding
21	Mthfd1	ATP binding
22	Chgb	protein binding
23	Acta1	protein binding
24	Adgrl1	protein binding
25	Fam162a	regulation of apoptosis
26	Cpsf7	nucleic acid binding
27	Tmem109	protein binding
28	Emd	actin binding
29	Coro1b	protein binding
30	Syt5	metal ion binding
31	Atp5mpl	mitochondrial proteolipid
32	Pitpna	phospholipid transporter activity
33	Ywhaz	protein domain specific binding
34	Atp9a	ATP binding
35	Apoc3	lipid binding
36	Numb	protein binding
37	Nova1	RNA binding
38	Pmch	melanin-concentrating hormone activity
39	Bcas1	protein homodimerization activity
40	Pzp	endopeptidase inhibitor activity
41	Sec11c	serine-type peptidase activity
42	Crip2	metal ion binding
43	Nsg1	clathrin light chain binding

Continued on next page

Table C.1 – *Continued from previous page*

	Ensembl Gene ID	Function
44	Scrn1	dipeptidase activity
45	Lypd1	protein binding
46	Ehd1	protein binding
47	Pisd	phosphatidylserine decarboxylase activity
48	Ap3s1	protein transporter activity
49	Ap2a2	binding
50	Lsp1	signal transducer activity
51	Rasgrp1	calcium ion binding
52	Cpne6	protein binding
53	Gng4	signal transducer activity
54	Anxa7	calcium ion binding
55	Homer2	protein binding
56	Lin7c	protein binding
57	Arc	actin binding
58	Astn2	protein binding
59	Mast4	ATP binding
60	Gpsm3	GTPase regulator activity
61	Lima1	actin filament binding
62	Pak1	ATP binding
63	Rnf112	GTP binding
64	Il9r	protein binding
65	Map4k4	ATP binding
66	Rps20	structural constituent of ribosome
67	Pde1b	metal ion binding
68	Fam81a	FAM81A
69	Agpat1	transferase activity, transferring acyl groups

Continued on next page

Table C.1 – *Continued from previous page*

	Ensembl Gene ID	Function
70	Ermn	actin binding
71	Med13l	RNA polymerase II transcription cofactor activity
72	Nkiras1	GTP binding
73	Mtch1	cell death
74	Tpm1	actin binding
75	Unc79	protein binding
76	Sec61a2	ribosome binding
77	Actr1a	nucleotide binding
78	Jph4	formation of junctional membrane complexes
79	Pip4k2b	phosphatidylinositol phosphate kinase activity
80	Cetn2	calcium ion binding
81	Myl12b	calcium ion binding
82	Atp6v1g2	ATPase activity
83	Thy1	integrin binding
84	ATP8	hydrogen ion transmembrane transporter activity
85	Lsamp	cell adhesion
86	Cd47	protein binding
87	Hook3	protein binding
88	S100a16	calcium ion binding
89	Vgf	neuropeptide hormone activity
90	Prr36	proline rich 36
91	Robo2	protein binding
92	Rps27rt	structural constituent of ribosome
93	Lrrc8b	protein binding
94		RIKEN cDNA
95	Synpo	actin binding

Continued on next page

Table C.1 – *Continued from previous page*

	Ensembl Gene ID	Function
96	Gypa	protein homodimerization activity
97	Ppp1r12b	protein binding
98	Sept11	GTP binding
99	Ddn	RNA polymerase binding
100	Ahnak2	cytoplasmic vesicle
101	Smim20	cellular component organization
102	Arf1	GTP binding
103	Pcdhac2	calcium ion binding
104	Fxyd1	ion channel activity
105	Smim1	small integral membrane protein 1
106	Tubb4b	GTP binding
107	Lgi4	protein binding
108	Kcnb1	protein binding
109	Hadhb	catalytic activity
110	Plekhg1	Rho guanyl-nucleotide exchange factor activity
111	Ube2v1	protein binding
112	Crocc	protein binding
113	Slc25a40	transmembrane transporter activity
114	Irgm1	GTP binding
115	Cnn3	protein binding
116	Gprin1	phosphoprotein binding
117	2410002F23Rik	visual system
118	Ppfia1	protein binding
119	Cbln3	protein binding
120	Dusp15	phosphatase activity
121	Cend1	protein binding

Continued on next page

Table C.1 – *Continued from previous page*

	Ensembl Gene ID	Function
122	Tubb2a	GTP binding
123	Gpx1	glutathione peroxidase activity
124	Tubb2b	GTP binding
125	Basp1	protein binding
126	Gpr162	G-protein coupled receptor activity
127	Acad12	flavin adenine dinucleotide binding
128	Ankrd63	protein binding
129	Tubb4a	GTP binding
130	Rpl23	structural constituent of ribosome
131	Cdh20	calcium ion binding
132	Rpl37a	structural constituent of ribosome
133	Mrpl50	Component of the mitochondrial ribosome (39S)
134	Cfl1	actin binding
135	Hist1h4b	DNA binding
136	Kcnc3	protein binding
137	Fgf1	growth factor activity
138	Arhgap26	GTPase activator activity
139	Csnk1g3	ATP binding
140	Eif4b	RNA binding
141	Cltb	structural molecule activity
142	Exog	metal ion binding
143	Hist1h2bp	DNA binding
144	Ctnna3	cadherin binding
145	Lrrtm3	protein binding
146	Hist1h2aa	DNA binding
147	Usp46	thiol-dependent ubiquitinyl hydrolase activity

Continued on next page

Table C.1 – *Continued from previous page*

	Ensembl Gene ID	**Function**
148	Cisd3	2 iron, 2 sulfur cluster binding
149	Rpl27a	structural constituent of ribosome
150	Shisa6	receptor binding
151	Ly6g5b	external side of plasma membrane
152	Pakap	anatomical structure development
153	Aldoa	integral component membrane

Table C.2: List of Key Synaptic genes found expressed in the *de novo* zebrafish assembly

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
1	ablim1b	UoN.zebrafish.38152.1	ENSDARG00000045064	many:1	6.79	5.02	4.47
2	ablim1b	UoN.zebrafish.38153.1	ENSDARG00000045064	many:1	6.86	5.46	4.26
3	ablim1b	UoN.zebrafish.17553.1	ENSDARG00000045064	many:1	7.84	6.06	9.10
4	baiap2a	UoN.zebrafish.31091.1	ENSDARG00000062799	many:1	4.07	4.55	3.55
5	baiap2a	UoN.zebrafish.5393.1	ENSDARG00000062799	many:1	8.65	6.57	5.61
6	baiap2a	UoN.zebrafish.5394.1	ENSDARG00000062799	many:1	10.83	6.77	8.04
7	baiap2l1a	UoN.zebrafish.34697.1	ENSDARG00000029305	many:1	0.63	2.21	1.44
8	baiap2l1a	UoN.zebrafish.6965.1	ENSDARG00000029305	many:1	1.89	7.05	2.26
9	baiap2l1a	UoN.zebrafish.7098.1	ENSDARG00000029305	many:1	3.51	7.43	4.82
10	bdnf	UoN.zebrafish.29146.1	ENSDARG00000018817	1:1	10.33	26.48	13.59
11	bdnf	UoN.zebrafish.35077.1	ENSDARG00000018817	1:1	24.97	36.21	26.45
12	bdnf	UoN.zebrafish.35078.1	ENSDARG00000018817	1:1	2.14	5.21	4.27
13	bdnf	UoN.zebrafish.14468.1	ENSDARG00000018817	1:1	19.11	21.87	18.23
14	cacng2a	UoN.zebrafish.37829.1	ENSDARG00000032565	many:1	38.67	28.11	39.87
15	cacng2a	UoN.zebrafish.4053.1	ENSDARG00000032565	many:1	9.06	8.21	13.06
16	cacng2b	UoN.zebrafish.18935.1	ENSDARG00000102376	many:1	44.84	51.47	54.32

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
17	cacng2b	UoN.zebrafish.13464.1	ENSDARG00000102376	many:1	11.50	20.29	16.07
18	camk2a	UoN.zebrafish.15400.1	ENSDARG00000053617	1:1	123.88	223.16	182.77
19	camk2b1	UoN.zebrafish.41473.1	ENSDARG00000011065	many:1	79.29	60.91	77.68
20	camk2b1	UoN.zebrafish.41474.1	ENSDARG00000011065	many:1	26.41	34.42	45.32
21	camk2d1	UoN.zebrafish.25095.1	ENSDARG00000043010	1:1	4.01	1.94	2.81
22	camk2d1	UoN.zebrafish.11764.1	ENSDARG00000043010	1:1	4.72	4.38	5.30
23	camk2d1	UoN.zebrafish.18225.1	ENSDARG00000043010	1:1	24.39	29.22	22.08
24	camk2d2	UoN.zebrafish.32254.1	ENSDARG00000014273	unique	55.76	20.18	63.73
25	camk2d2	UoN.zebrafish.32255.1	ENSDARG00000014273	unique	76.90	60.67	85.38
26	camk2d2	UoN.zebrafish.32256.1	ENSDARG00000014273	unique	125.08	197.19	124.09
27	camk2g1	UoN.zebrafish.12366.1	ENSDARG00000071395	many:1	27.26	20.11	27.52
28	camk2g1	UoN.zebrafish.12368.1	ENSDARG00000071395	many:1	18.71	13.52	23.06
29	camk2g1	UoN.zebrafish.12368.2	ENSDARG00000071395	many:1	18.71	13.52	23.06
30	camk2g1	UoN.zebrafish.12367.1	ENSDARG00000071395	many:1	20.35	14.04	24.24
31	camk2g2	UoN.zebrafish.31602.1	ENSDARG00000056206	many:1	26.63	19.29	36.26
32	camk2g2	UoN.zebrafish.25514.1	ENSDARG00000056206	many:1	24.81	15.08	27.47
33	camk2g2	UoN.zebrafish.25515.1	ENSDARG00000056206	many:1	18.17	6.99	22.45

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
34	camk2g2	UoN.zebrafish.41476.1	ENSDARG00000056206	many:1	1.06		5.15
35	camk2g2	UoN.zebrafish.1264.1	ENSDARG00000056206	many:1	11.08	12.41	12.46
36	camk2n1a	UoN.zebrafish.39488.1	ENSDARG00000025855	many:1	26.03	38.15	38.01
37	camk2n1a	UoN.zebrafish.16373.1	ENSDARG00000025855	many:1	332.69	471.07	413.78
38	cntnap2a	UoN.zebrafish.31422.1	ENSDARG00000058969	1:1	51.58	29.06	63.06
39	cntnap2a	UoN.zebrafish.21266.1	ENSDARG00000058969	1:1	3.71	3.43	4.25
40	cntnap2a	UoN.zebrafish.42314.1	ENSDARG00000058969	1:1	15.68	12.58	20.54
41	cntnap2b	UoN.zebrafish.27366.1	ENSDARG00000074558	unique	6.69	4.96	5.87
42	cntnap2b	UoN.zebrafish.13581.1	ENSDARG00000074558	unique	11.86	9.22	11.93
43	dlg1	UoN.zebrafish.23004.1	ENSDARG00000009677	many:1	10.49	13.84	13.47
44	dlg1	UoN.zebrafish.12550.1	ENSDARG00000009677	many:1	16.59	23.38	22.86
45	dlg1	UoN.zebrafish.2123.1	ENSDARG00000009677	many:1		14.67	
46	dlg1	UoN.zebrafish.2123.2	ENSDARG00000009677	many:1		14.67	
47	dlg1	UoN.zebrafish.2124.1	ENSDARG00000009677	many:1	76.20	18.16	55.96
48	dlg1	UoN.zebrafish.2125.1	ENSDARG00000009677	many:1	16.45	7.80	9.98
49	dlg1	UoN.zebrafish.2125.2	ENSDARG00000009677	many:1	16.45	7.80	9.98
50	dlg1	UoN.zebrafish.2125.3	ENSDARG00000009677	many:1	18.07	8.49	11.71

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
51	dlg1	UoN.zebrafish.29565.1	ENSDARG00000009677	many:1	5.14	6.41	3.58
52	dlg1	UoN.zebrafish.35333.1	ENSDARG00000009677	many:1	12.18	16.00	15.64
53	dlg1	UoN.zebrafish.36634.1	ENSDARG00000009677	many:1		0.83	0.63
54	dlg1	UoN.zebrafish.36634.2	ENSDARG00000009677	many:1		0.83	0.63
55	dlg1	UoN.zebrafish.36634.3	ENSDARG00000009677	many:1		0.77	
56	dlg1	UoN.zebrafish.36632.1	ENSDARG00000009677	many:1	5.40		3.08
57	dlg1l	UoN.zebrafish.14520.1	ENSDARG00000102216	many:1	1.18		1.50
58	dlg1l	UoN.zebrafish.14546.1	ENSDARG00000102216	many:1		3.51	6.31
59	dlg1l	UoN.zebrafish.14546.2	ENSDARG00000102216	many:1	6.74		2.45
60	dlg1l	UoN.zebrafish.2386.1	ENSDARG00000102216	many:1	3.89	3.33	4.67
61	dlg1l	UoN.zebrafish.2386.2	ENSDARG00000102216	many:1	3.89	3.33	4.67
62	dlg1l	UoN.zebrafish.2387.1	ENSDARG00000102216	many:1	6.97	1.38	4.07
63	dlg1l	UoN.zebrafish.2387.2	ENSDARG00000102216	many:1	1.34	1.97	6.20
64	dlg1l	UoN.zebrafish.36631.1	ENSDARG00000102216	many:1	4.70	1.63	3.73
65	dlg1l	UoN.zebrafish.35535.1	ENSDARG00000102216	many:1	2.38	0.57	3.41
66	dlg2	UoN.zebrafish.21371.1	ENSDARG00000099323	1:1	0.89	1.45	1.81
67	dlg2	UoN.zebrafish.21371.2	ENSDARG00000099323	1:1	0.89	1.45	1.81

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
68	dlg2	UoN.zebrafish.42826.1	ENSDARG00000099323	1:1	6.90	5.82	7.45
69	dlg2	UoN.zebrafish.5503.1	ENSDARG00000099323	1:1	5.07	2.80	6.12
70	dlg2	UoN.zebrafish.5503.2	ENSDARG00000099323	1:1	5.07	2.80	6.12
71	dlg2	UoN.zebrafish.5503.3	ENSDARG00000099323	1:1	5.07	2.80	6.12
72	dlg3	UoN.zebrafish.32361.1	ENSDARG00000076796	1:1	37.22	58.00	46.22
73	dlg3	UoN.zebrafish.15984.1	ENSDARG00000076796	1:1	16.44	12.58	18.92
74	dlg5a	UoN.zebrafish.11073.1	ENSDARG00000074059	many:1	10.48	8.42	11.16
75	dlg5a	UoN.zebrafish.30209.1	ENSDARG00000074059	many:1	2.85		1.07
76	dlgap1a	UoN.zebrafish.9767.1	ENSDARG00000014280	many:1	5.93	9.62	6.90
77	dlgap3	UoN.zebrafish.24054.1	ENSDARG00000055459	1:1	1.03	1.38	2.92
78	dlgap3	UoN.zebrafish.19986.1	ENSDARG00000055459	1:1	2.07	1.01	2.49
79	dlgap3	UoN.zebrafish.33162.1	ENSDARG00000055459	1:1	6.69	1.57	3.54
80	dlgap3	UoN.zebrafish.33997.1	ENSDARG00000055459	1:1	2.84	2.48	3.75
81	dlgap4b	UoN.zebrafish.9751.1	ENSDARG00000012823	many:1	2.43	1.70	4.69
82	dlgap4b	UoN.zebrafish.27426.1	ENSDARG00000012823	many:1	4.63	2.08	3.66
83	dlgap4b	UoN.zebrafish.10974.1	ENSDARG00000012823	many:1	2.75	1.40	3.62
84	dlgap4b	UoN.zebrafish.2870.1	ENSDARG00000012823	many:1	1.04		1.14

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
85	dlgap4b	UoN.zebrafish.2871.1	ENSDARG00000012823	many:1	2.27	0.66	0.96
86	dlgap4b	UoN.zebrafish.30636.1	ENSDARG00000012823	many:1	3.04		1.06
87	dlgap4b	UoN.zebrafish.35353.1	ENSDARG00000012823	many:1	7.19	3.83	6.28
88	gabbr1b	UoN.zebrafish.30522.1	ENSDARG00000016667	many:1	9.92	7.98	11.42
89	gabbr1b	UoN.zebrafish.17965.1	ENSDARG00000016667	many:1	5.51	4.93	5.26
90	gabbr1b	UoN.zebrafish.17968.1	ENSDARG00000016667	many:1	1.17		9.30
91	gabbr1b	UoN.zebrafish.17968.2	ENSDARG00000016667	many:1	9.54	12.04	0.93
92	gabra1	UoN.zebrafish.39807.1	ENSDARG00000068989	unique	33.29	21.22	40.75
93	gabra1	UoN.zebrafish.17987.1	ENSDARG00000068989	unique	37.05	21.93	47.64
94	gabra2a	UoN.zebrafish.22743.1	ENSDARG00000091459	many:many	12.57	12.10	11.51
95	gabra2a	UoN.zebrafish.22743.1	ENSDARG00000091459	many:many	12.57	12.10	11.51
96	gabra2a	UoN.zebrafish.37176.1	ENSDARG00000091459	many:many	2.25	3.48	2.18
97	gabra2a	UoN.zebrafish.37176.1	ENSDARG00000091459	many:many	2.25	3.48	2.18
98	gabra2a	UoN.zebrafish.24455.1	ENSDARG00000091459	many:many	4.70	5.10	5.33
99	gabra2a	UoN.zebrafish.24455.1	ENSDARG00000091459	many:many	4.70	5.10	5.33
100	gabra2a	UoN.zebrafish.45152.1	ENSDARG00000091459	many:many	4.21	2.80	5.57
101	gabra2a	UoN.zebrafish.45152.1	ENSDARG00000091459	many:many	4.21	2.80	5.57

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
102	gabra4	UoN.zebrafish.30677.1	ENSDARG00000013389	1:1	13.39	5.50	15.20
103	gabra5	UoN.zebrafish.32677.1	ENSDARG00000070730	1:1	6.08	13.34	5.52
104	gabra5	UoN.zebrafish.28779.1	ENSDARG00000070730	1:1	20.50	44.11	17.13
105	gabra5	UoN.zebrafish.35981.1	ENSDARG00000070730	1:1	5.12	14.48	7.20
106	gabra5	UoN.zebrafish.7850.1	ENSDARG00000070730	1:1	4.94	8.93	3.72
107	gabra6b	UoN.zebrafish.28934.1	ENSDARG00000058736	unique	14.90	2.52	10.21
108	gria1b	UoN.zebrafish.21068.1	ENSDARG00000032714	many:1	9.01	12.33	12.11
109	gria2b	UoN.zebrafish.26051.1	ENSDARG00000052765	many:1	44.73	100.61	29.62
110	gria2b	UoN.zebrafish.26052.1	ENSDARG00000052765	many:1	45.19	93.56	63.12
111	gria2b	UoN.zebrafish.30296.1	ENSDARG00000052765	many:1	63.61	118.47	65.89
112	gria3a	UoN.zebrafish.37869.1	ENSDARG00000032737	many:1	26.04	12.57	33.13
113	gria3a	UoN.zebrafish.7162.1	ENSDARG00000032737	many:1	12.77	12.56	17.39
114	gria3b	UoN.zebrafish.25135.1	ENSDARG00000037498	many:1	22.62	29.19	28.70
115	gria3b	UoN.zebrafish.29618.1	ENSDARG00000037498	many:1	7.15	4.99	9.03
116	gria3b	UoN.zebrafish.29618.2	ENSDARG00000037498	many:1	7.15	4.99	9.03
117	gria4a	UoN.zebrafish.20176.1	ENSDARG00000037496	many:1	3.03	4.45	6.44
118	gria4a	UoN.zebrafish.3071.1	ENSDARG00000037496	many:1	19.14	2.57	14.20

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
119	gria4a	UoN.zebrafish.759.1	ENSDARG00000037496	many:1	17.45	8.93	15.03
120	gria4b	UoN.zebrafish.20177.1	ENSDARG00000059368	many:1	12.52	19.54	12.02
121	gria4b	UoN.zebrafish.760.1	ENSDARG00000059368	many:1	11.89	11.65	15.20
122	grik1a	UoN.zebrafish.23940.1	ENSDARG00000069139	many:1	4.30	5.49	6.01
123	grik1a	UoN.zebrafish.30025.1	ENSDARG00000069139	many:1	3.38	3.22	4.10
124	grik1a	UoN.zebrafish.41437.1	ENSDARG00000069139	many:1	1.63	2.69	1.53
125	grik4	UoN.zebrafish.20611.1	ENSDARG00000026753	1:1	6.79	7.50	7.78
126	grik4	UoN.zebrafish.41789.1	ENSDARG00000026753	1:1	1.23	1.53	2.10
127	grik4	UoN.zebrafish.1233.1	ENSDARG00000026753	1:1	2.38		2.26
128	grik4	UoN.zebrafish.1235.1	ENSDARG00000026753	1:1	5.97	3.84	8.11
129	grik4	UoN.zebrafish.14358.1	ENSDARG00000026753	1:1	1.22	2.25	1.64
130	grik5	UoN.zebrafish.19720.1	ENSDARG00000101449	many:1	6.25	8.46	8.80
131	grik5	UoN.zebrafish.19723.1	ENSDARG00000101449	many:1	5.05	3.74	4.36
132	grin2aa	UoN.zebrafish.30925.1	ENSDARG00000034493	many:1	8.45	1.15	10.79
133	grin2ab	UoN.zebrafish.17522.1	ENSDARG00000070543	many:1	16.89	27.33	24.24
134	grin2bb	UoN.zebrafish.39148.1	ENSDARG00000030376	many:1	11.36	34.66	22.08
135	grin2bb	UoN.zebrafish.17408.1	ENSDARG00000030376	many:1	8.77	22.15	12.51

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
136	grin2bb	UoN.zebrafish.26047.1	ENSDARG00000030376	many:1	4.81	9.45	5.08
137	grin2bb	UoN.zebrafish.26048.1	ENSDARG00000030376	many:1	1.94	1.51	2.48
138	grin2bb	UoN.zebrafish.18719.1	ENSDARG00000030376	many:1	15.88	39.36	18.47
139	grin2bb	UoN.zebrafish.30586.1	ENSDARG00000030376	many:1	0.51	0.81	0.69
140	grin2bb	UoN.zebrafish.30586.2	ENSDARG00000030376	many:1	1.42	2.04	2.04
141	grin2bb	UoN.zebrafish.9255.1	ENSDARG00000030376	many:1	7.08	14.84	7.95
142	grin2cb	UoN.zebrafish.37953.1	ENSDARG00000077560	many:1	3.39	1.21	1.81
143	grin2da	UoN.zebrafish.12090.1	ENSDARG00000086207	many:1	2.97	1.86	3.43
144	grin2da	UoN.zebrafish.10975.1	ENSDARG00000086207	many:1	3.26	4.78	3.70
145	grin2da	UoN.zebrafish.4986.1	ENSDARG00000086207	many:1	2.65	2.80	2.94
146	grin2da	UoN.zebrafish.29518.1	ENSDARG00000086207	many:1	5.89	3.15	4.80
147	grin2da	UoN.zebrafish.40959.1	ENSDARG00000086207	many:1	1.14	0.70	1.79
148	grin2da	UoN.zebrafish.40960.1	ENSDARG00000086207	many:1	4.85	5.88	5.59
149	grin2da	UoN.zebrafish.35625.1	ENSDARG00000086207	many:1	4.10	2.31	4.05
150	grin2da	UoN.zebrafish.17894.1	ENSDARG00000086207	many:1	0.55	1.14	1.35
151	grin2da	UoN.zebrafish.14872.1	ENSDARG00000086207	many:1	6.47	9.33	9.51
152	homer1b	UoN.zebrafish.31446.1	ENSDARG00000101759	1:1	20.27	39.20	19.89

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
153	homer1b	UoN.zebrafish.9488.1	ENSDARG00000101759	1:1	53.14	113.05	59.69
154	homer1b	UoN.zebrafish.21213.1	ENSDARG00000101759	1:1	7.31	3.20	9.50
155	homer1b	UoN.zebrafish.44718.1	ENSDARG00000101759	1:1	2.25		2.16
156	homer3b	UoN.zebrafish.43017.1	ENSDARG00000010789	many:1	6.83	1.32	3.45
157	homer3b	UoN.zebrafish.36392.1	ENSDARG00000010789	many:1	9.50	4.88	4.40
158	iqsec1b	UoN.zebrafish.30096.1	ENSDARG00000016551	1:1	24.14	10.76	32.05
159	iqsec1b	UoN.zebrafish.21950.1	ENSDARG00000016551	1:1		6.00	
160	iqsec1b	UoN.zebrafish.21950.2	ENSDARG00000016551	1:1	17.56		14.16
161	iqsec2b	UoN.zebrafish.17435.1	ENSDARG00000077709	many:1	4.43	1.63	8.59
162	iqsec2b	UoN.zebrafish.17435.2	ENSDARG00000077709	many:1	3.06	6.69	2.20
163	iqsec2b	UoN.zebrafish.17434.1	ENSDARG00000077709	many:1	6.17	6.53	7.63
164	iqsec2b	UoN.zebrafish.20450.1	ENSDARG00000077709	many:1	6.65	3.24	2.01
165	iqsec2b	UoN.zebrafish.20450.2	ENSDARG00000077709	many:1	5.91	11.85	13.94
166	iqsec2b	UoN.zebrafish.41282.1	ENSDARG00000077709	many:1	9.23	12.13	9.01
167	iqsec3b	UoN.zebrafish.9772.1	ENSDARG00000093091	many:1	1.63	0.95	2.77
168	iqsec3b	UoN.zebrafish.43449.1	ENSDARG00000093091	many:1	3.37	2.69	4.27
169	iqsec3b	UoN.zebrafish.33769.1	ENSDARG00000093091	many:1	2.52	0.61	3.46

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
170	iqsec3b	UoN.zebrafish.33951.1	ENSDARG00000093091	many:1	3.62	0.92	2.92
171	magi1b	UoN.zebrafish.2719.1	ENSDARG00000003169	1:1	15.62	12.02	18.46
172	magi1b	UoN.zebrafish.30285.1	ENSDARG00000003169	1:1	26.51	32.80	39.91
173	magi1b	UoN.zebrafish.41702.1	ENSDARG00000003169	1:1	9.68	16.76	11.58
174	magi2a	UoN.zebrafish.12658.1	ENSDARG00000021590	many:1	2.09	0.70	4.26
175	magi2a	UoN.zebrafish.29442.1	ENSDARG00000021590	many:1	3.65	2.01	4.01
176	magi2a	UoN.zebrafish.33865.1	ENSDARG00000021590	many:1	0.69		1.35
177	magi2a	UoN.zebrafish.33864.1	ENSDARG00000021590	many:1	2.40	2.61	3.13
178	magi3a	UoN.zebrafish.38105.1	ENSDARG00000101869	many:1	5.38	1.78	6.62
179	magi3a	UoN.zebrafish.2698.1	ENSDARG00000101869	many:1	6.96	3.12	7.62
180	magi3a	UoN.zebrafish.17814.1	ENSDARG00000101869	many:1	0.42	0.68	0.56
181	magi3a	UoN.zebrafish.17815.1	ENSDARG00000101869	many:1			1.10
182	magixa	UoN.zebrafish.31629.1	ENSDARG00000025108	unique	0.90	1.38	1.33
183	magixa	UoN.zebrafish.27534.1	ENSDARG00000025108	unique	3.53	0.89	1.97
184	magixa	UoN.zebrafish.34039.1	ENSDARG00000025108	unique	2.21	1.63	2.79
185	magixa	UoN.zebrafish.42773.1	ENSDARG00000025108	unique	3.65	1.70	2.90
186	magixa	UoN.zebrafish.1.1	ENSDARG00000025108	unique	2.82	1.06	2.27

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
187	mapk1	UoN.zebrafish.2701.1	ENSDARG00000027552	1:1	1.68		
188	mapk1	UoN.zebrafish.2702.1	ENSDARG00000027552	1:1	92.45	138.04	93.65
189	mapk10	UoN.zebrafish.27155.1	ENSDARG00000102730	1:1	6.08	2.53	5.85
190	mapk10	UoN.zebrafish.27154.1	ENSDARG00000102730	1:1	7.58	4.78	7.10
191	mapk10	UoN.zebrafish.20382.1	ENSDARG00000102730	1:1			14.78
192	mapk10	UoN.zebrafish.20382.2	ENSDARG00000102730	1:1	9.23	15.87	
193	mapk10	UoN.zebrafish.20382.3	ENSDARG00000102730	1:1	7.17	1.98	
194	mapk10	UoN.zebrafish.15708.1	ENSDARG00000102730	1:1	3.84	9.05	6.89
195	mapk11	UoN.zebrafish.12270.1	ENSDARG00000045836	1:1	9.55	3.77	8.74
196	mapk11	UoN.zebrafish.30572.1	ENSDARG00000045836	1:1	24.52	27.90	14.35
197	mapk12b	UoN.zebrafish.1566.1	ENSDARG00000006409	many:1	6.42	1.40	3.25
198	mapk13	UoN.zebrafish.16035.1	ENSDARG00000058470	1:1	11.26	16.28	10.79
199	mapk14a	UoN.zebrafish.20628.1	ENSDARG00000000857	many:1	15.77	10.11	16.13
200	mapk14a	UoN.zebrafish.12271.1	ENSDARG00000000857	many:1	6.48	6.53	9.88
201	mapk14b	UoN.zebrafish.44827.1	ENSDARG00000028721	many:1	9.87	7.20	8.68
202	mapk14b	UoN.zebrafish.1005.1	ENSDARG00000028721	many:1	7.73	5.35	9.21
203	mapk3	UoN.zebrafish.19212.1	ENSDARG00000070573	1:1	0.33	0.87	0.42

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
204	mapk3	UoN.zebrafish.12817.1	ENSDARG00000070573	1:1	90.58	117.39	91.39
205	mapk4	UoN.zebrafish.27200.1	ENSDARG00000017681	1:1	9.44	6.14	7.76
206	mapk6	UoN.zebrafish.16811.1	ENSDARG00000032103	1:1	43.88	25.07	46.42
207	mapk7	UoN.zebrafish.32839.1	ENSDARG00000023110	1:1	0.98		0.64
208	mapk7	UoN.zebrafish.18468.1	ENSDARG00000023110	1:1	6.77	4.80	6.97
209	mapk7	UoN.zebrafish.42468.1	ENSDARG00000023110	1:1	3.40	2.46	4.44
210	mapk8a	UoN.zebrafish.42192.1	ENSDARG00000031888	many:1	6.32		
211	mapk8a	UoN.zebrafish.42192.2	ENSDARG00000031888	many:1	10.10	14.06	14.64
212	mapk8b	UoN.zebrafish.20890.1	ENSDARG00000009870	many:1	4.62	3.33	6.12
213	mapk8b	UoN.zebrafish.6619.1	ENSDARG00000009870	many:1	5.28	2.11	5.21
214	mapk8ip1a	UoN.zebrafish.37277.1	ENSDARG00000102229	many:1	55.45	48.56	55.21
215	mapk8ip2	UoN.zebrafish.30592.1	ENSDARG00000063157	1:1	29.50	18.20	31.83
216	mapk9	UoN.zebrafish.19802.1	ENSDARG00000077364	1:1	31.47	20.79	28.04
217	mapkap1	UoN.zebrafish.24825.1	ENSDARG00000091777	1:1	4.97	1.88	3.81
218	mapkap1	UoN.zebrafish.44265.1	ENSDARG00000091777	1:1	4.19	3.34	4.67
219	mapkapk2a	UoN.zebrafish.50.1	ENSDARG00000002552	many:1	15.94	14.00	16.76

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
220	mapkapk5	UoN.zebrafish.27633.1	ENSDARG00000028082	many:1	12.56	13.92	11.89
221	mapkapk5	UoN.zebrafish.27633.1	ENSDARG00000028082	many:1	12.56	13.92	11.89
222	mapkbp1	UoN.zebrafish.2209.1	ENSDARG00000103746	1:1	3.19	1.22	1.79
223	mapkbp1	UoN.zebrafish.29721.1	ENSDARG00000103746	1:1	2.03		1.93
224	mapkbp1	UoN.zebrafish.30249.1	ENSDARG00000103746	1:1	1.94	0.97	1.43
225	mapkbp1	UoN.zebrafish.42803.1	ENSDARG00000103746	1:1	1.96	1.60	2.95
226	mapkbp1	UoN.zebrafish.42804.1	ENSDARG00000103746	1:1	7.44	4.39	7.60
227	mapkbp1	UoN.zebrafish.27679.1	ENSDARG00000103746	1:1	2.73		1.26
228	ncam1a	UoN.zebrafish.40173.1	ENSDARG00000056181	many:1	5.58	7.67	9.98
229	ncam1a	UoN.zebrafish.24473.1	ENSDARG00000056181	many:1	28.15	8.25	19.88
230	ncam1a	UoN.zebrafish.7608.1	ENSDARG00000056181	many:1	5.04	5.31	3.64
231	ncam1a	UoN.zebrafish.14280.1	ENSDARG00000056181	many:1	30.31	21.88	29.62
232	ncam1b	UoN.zebrafish.18443.1	ENSDARG00000007220	many:1	8.47	11.91	7.73
233	ncam1b	UoN.zebrafish.18445.1	ENSDARG00000007220	many:1	0.67	0.63	0.39
234	ncam1b	UoN.zebrafish.18445.2	ENSDARG00000007220	many:1	2.16	4.90	3.54
235	ncam1b	UoN.zebrafish.29555.1	ENSDARG00000007220	many:1	4.34	1.33	4.94
236	ncam1b	UoN.zebrafish.15216.1	ENSDARG00000007220	many:1	4.30	2.66	3.46

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
237	ncam2	UoN.zebrafish.35983.1	ENSDARG00000017466	many:1	37.63	54.50	52.67
238	ncam3	UoN.zebrafish.25070.1	ENSDARG00000089586	unique	6.20	1.46	4.99
239	ncam3	UoN.zebrafish.25071.1	ENSDARG00000089586	unique		0.90	2.09
240	ncam3	UoN.zebrafish.25072.1	ENSDARG00000089586	unique	3.61	2.33	1.03
241	ncam3	UoN.zebrafish.5252.1	ENSDARG00000089586	unique	2.42	3.94	2.07
242	nlgn1	UoN.zebrafish.45220.1	ENSDARG00000077710	1:1	3.61	2.60	4.42
243	nlgn2a	UoN.zebrafish.11229.1	ENSDARG00000077329	many:1	4.54	2.82	7.51
244	nlgn2a	UoN.zebrafish.4750.1	ENSDARG00000077329	many:1	5.55	0.73	1.85
245	nlgn2a	UoN.zebrafish.4750.2	ENSDARG00000077329	many:1	4.40	6.13	6.00
246	nlgn2b	UoN.zebrafish.19844.1	ENSDARG00000079251	many:1	2.89	3.51	3.76
247	nlgn2b	UoN.zebrafish.19844.2	ENSDARG00000079251	many:1	2.89	3.51	3.76
248	nlgn2b	UoN.zebrafish.4848.1	ENSDARG00000079251	many:1	14.24	16.20	15.88
249	nlgn3a	UoN.zebrafish.23065.1	ENSDARG00000104786	many:1	6.22	5.54	7.48
250	nlgn3a	UoN.zebrafish.18840.1	ENSDARG00000104786	many:1	10.61	9.97	9.63
251	nlgn3a	UoN.zebrafish.6845.1	ENSDARG00000104786	many:1	14.84	19.32	14.42
252	nlgn3b	UoN.zebrafish.19192.1	ENSDARG00000062376	many:1	33.70	60.95	42.66
253	nlgn3b	UoN.zebrafish.19193.1	ENSDARG00000062376	many:1	6.87	11.29	8.80

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
254	nlgn4a	UoN.zebrafish.9799.1	ENSDARG00000079455	unique	25.57	22.08	30.55
255	nlgn4b	UoN.zebrafish.32452.1	ENSDARG00000077761	unique	1.80	0.64	2.39
256	nlgn4b	UoN.zebrafish.32453.1	ENSDARG00000077761	unique	1.76	1.45	3.88
257	nlgn4b	UoN.zebrafish.26063.1	ENSDARG00000077761	unique	4.87	1.37	5.43
258	nlgn4b	UoN.zebrafish.20305.1	ENSDARG00000077761	unique	1.67	1.72	1.83
259	nsfa	UoN.zebrafish.9322.1	ENSDARG00000007654	many:1	354.30	377.90	431.15
260	nsfb	UoN.zebrafish.20976.1	ENSDARG00000038991	many:1	3.71	1.73	3.99
261	nsfb	UoN.zebrafish.4626.1	ENSDARG00000038991	many:1	5.94	3.62	5.47
262	nsfb	UoN.zebrafish.43291.1	ENSDARG00000038991	many:1	2.98	2.05	2.15
263	shank3b	UoN.zebrafish.22499.1	ENSDARG00000063054	many:1	4.44	7.02	4.53
264	shank3b	UoN.zebrafish.12175.1	ENSDARG00000063054	many:1	7.70	12.19	10.10
265	snap25a	UoN.zebrafish.36623.1	ENSDARG00000020609	many:1	166.64	100.87	231.18
266	snap25a	UoN.zebrafish.36623.2	ENSDARG00000020609	many:1	3022.59	1630.20	3176.81
267	snap25b	UoN.zebrafish.37177.1	ENSDARG00000058117	many:1	832.81	750.45	714.72
268	stx10	UoN.zebrafish.15256.1	ENSDARG00000075030	unique	15.64	12.60	12.31
269	stx12	UoN.zebrafish.38839.1	ENSDARG00000098813	many:1	45.78	49.38	50.20
270	stx12l	UoN.zebrafish.7242.1	ENSDARG00000044605	many:1	30.51	21.41	32.03

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
271	stx12l	UoN.zebrafish.7242.2	ENSDARG00000044605	many:1	30.51	21.41	32.03
272	stx16	UoN.zebrafish.29030.1	ENSDARG00000003307	1:1	3.40	2.33	4.11
273	stx16	UoN.zebrafish.42477.1	ENSDARG00000003307	1:1	9.99	8.63	10.35
274	stx18	UoN.zebrafish.20852.1	ENSDARG00000035763	1:1	27.18	27.77	25.49
275	stx3a	UoN.zebrafish.25690.1	ENSDARG00000001880	many:1	2.31	1.61	1.81
276	stx3a	UoN.zebrafish.40821.1	ENSDARG00000001880	many:1	1.93	1.77	2.36
277	stx4	UoN.zebrafish.45165.1	ENSDARG00000052518	1:1	15.39	8.46	12.56
278	stx5a	UoN.zebrafish.34292.1	ENSDARG00000025033	many:1	16.49	15.29	17.22
279	stx5al	UoN.zebrafish.22556.1	ENSDARG00000003175	many:1	19.94	23.09	19.99
280	stx6	UoN.zebrafish.22555.1	ENSDARG00000042742	1:1	42.93	39.88	48.19
281	stx8	UoN.zebrafish.30253.1	ENSDARG00000103173	1:1	14.67	13.59	14.40
282	stxbp1a	UoN.zebrafish.2352.1	ENSDARG00000001994	many:1	2.84		
283	stxbp1a	UoN.zebrafish.2352.2	ENSDARG00000001994	many:1	270.64	118.45	278.40
284	stxbp1a	UoN.zebrafish.2351.1	ENSDARG00000001994	many:1	348.70	302.22	408.87
285	stxbp1b	UoN.zebrafish.32422.1	ENSDARG00000056036	many:1	11.72	53.56	8.36
286	stxbp2	UoN.zebrafish.23660.1	ENSDARG00000007603	1:1	2.87	2.46	2.71
287	stxbp2	UoN.zebrafish.33125.1	ENSDARG00000007603	1:1	2.89	3.01	3.14

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
288	stxbp2	UoN.zebrafish.33126.1	ENSDARG00000007603	1:1	4.25	3.43	3.90
289	stxbp2	UoN.zebrafish.36563.1	ENSDARG00000007603	1:1	5.33	3.71	2.73
290	stxbp3	UoN.zebrafish.38806.1	ENSDARG00000008142	1:1	11.07	6.16	11.34
291	stxbp4	UoN.zebrafish.29944.1	ENSDARG00000076997	1:1	1.66	1.53	2.15
292	stxbp4	UoN.zebrafish.29945.1	ENSDARG00000076997	1:1	2.22	1.17	3.94
293	stxbp4	UoN.zebrafish.43279.1	ENSDARG00000076997	1:1	2.28	1.66	5.39
294	stxbp5a	UoN.zebrafish.14571.1	ENSDARG00000002656	many:1	24.81	26.25	30.53
295	stxbp5b	UoN.zebrafish.39057.1	ENSDARG00000029234	many:1	2.52	1.63	2.56
296	stxbp5b	UoN.zebrafish.2250.1	ENSDARG00000029234	many:1	2.76		0.63
297	stxbp5b	UoN.zebrafish.29264.1	ENSDARG00000029234	many:1	0.88	1.15	0.88
298	stxbp5l	UoN.zebrafish.31837.1	ENSDARG00000006383	1:1	11.46	13.57	13.76
299	stxbp5l	UoN.zebrafish.12032.1	ENSDARG00000006383	1:1	13.56	8.79	18.42
300	stxbp5l	UoN.zebrafish.34078.1	ENSDARG00000006383	1:1	2.84	2.69	3.63
301	stxbp5l	UoN.zebrafish.5436.1	ENSDARG00000006383	1:1	3.96	4.50	5.32
302	stxbp5l	UoN.zebrafish.5436.2	ENSDARG00000006383	1:1	3.96	4.50	5.32
303	stxbp6	UoN.zebrafish.18278.1	ENSDARG00000088862	many:1	38.01	16.28	30.04
304	stxbp6l	UoN.zebrafish.42228.1	ENSDARG00000028354	many:1	629.59	115.00	522.99

Continued on next page

Table C.2 – *Continued from previous page*

	Gene	Transcript	ensembl gene ID	O. type[a]	Hb[b]	Olf[c]	Opt[d]
305	syngap1b	UoN.zebrafish.24877.1	ENSDARG00000069765	many:1		1.05	
306	syngap1b	UoN.zebrafish.24877.2	ENSDARG00000069765	many:1	0.75		
307	syngap1b	UoN.zebrafish.11247.1	ENSDARG00000069765	many:1	5.05	5.10	5.24
308	syngap1b	UoN.zebrafish.4844.1	ENSDARG00000069765	many:1	3.18	2.98	4.32
309	syngap1b	UoN.zebrafish.4846.1	ENSDARG00000069765	many:1	4.75	5.00	
310	syngap1b	UoN.zebrafish.4846.2	ENSDARG00000069765	many:1	1.27	1.93	5.43
311	syngap1b	UoN.zebrafish.8769.1	ENSDARG00000069765	many:1	3.40	3.99	3.69
312	syngap1b	UoN.zebrafish.14734.1	ENSDARG00000069765	many:1	5.89	18.32	12.68
313	vamp1	UoN.zebrafish.17662.1	ENSDARG00000031283	many:1	602.16	202.67	646.47
314	vamp2	UoN.zebrafish.21841.1	ENSDARG00000056877	1:1	227.64	262.74	275.45
315	vamp3	UoN.zebrafish.41227.1	ENSDARG00000070161	1:1	61.45	36.18	55.06
316	vamp4	UoN.zebrafish.8581.1	ENSDARG00000043510	1:1	51.73	72.04	50.72
317	vamp5	UoN.zebrafish.37873.1	ENSDARG00000068262	1:1	43.22	41.26	46.03
318	vamp8	UoN.zebrafish.15437.1	ENSDARG00000024116	1:1	8.84	10.17	9.26

[a] Orthology type (zebrafish:mouse); [b] Hindbrain (TPM); [c] Olfactory lobe (TPM); [d] Optic lobe (TPM)