

**INVESTIGATING EVOLUTIONARY PHENOTYPIC TRANSITIONS USING  
INTEGRATIVE BIOLOGICAL NETWORK FRAMEWORKS**

By

Sandarage Pasan Chinthana Fernando

B.S., University of Colombo, 2013

A Dissertation Submitted in Partial Fulfillment of  
the Requirements for the Degree of  
Doctor of Philosophy

---

Department of Biology  
Biology Program  
Bioinformatics Specialization  
In the graduate school  
The University of South Dakota,  
December 2018

Copyright by

SANDARAGE PASAN CHINTHANA FERNANDO

2018

All Rights Reserved



The members of the committee appointed to examine the Dissertation of Sandarage Pasan

Chinthana Fernando find it satisfactory and recommend that it be accepted



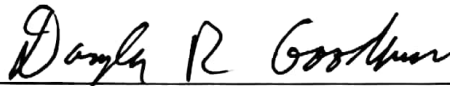
---

Chairperson, Dr. Paula Mabee



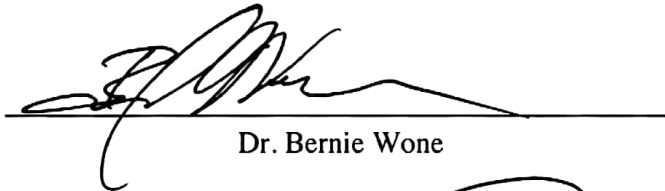
---

Dr. Erliang Zeng



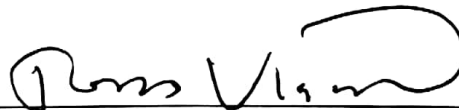
---

Dr. Doug Goodman



---

Dr. Bernie Wone



---

Dr. Todd Vision

## DISSERTATION ABSTRACT

The evolution of species is accompanied by phenotypic changes, such as the fin to limb transition, which led to the phenotypic diversity observed among species today. Studying and understanding these phenotypic transitions and the underlying genetic changes are important topics in evolutionary biology. With the advent of next-generation experimental methods, large volumes of biological data, such as protein-protein interaction (PPI) data, have accumulated and big data analyses have become the norm in bioinformatics studies. This dissertation attempts to use such publicly available large-scale data sets to study the evolutionary phenotypic transitions, which gives a new perspective to evolutionary biological studies. Here, the focus is on biological network data because they represent complex biological relationships, such as the interactions between proteins and relationships between different taxa that are important when studying the phenotypic transitions. This work builds the computational framework to integrate large-scale biological network data such as PPI networks, anatomy ontology data, and phylogenetic trees and solves the challenges associated with the integrations. The first objective focuses on solving the challenge of poor PPI network data quality by integrating with anatomy ontology data, which significantly improved the accuracy of network-based candidate gene prediction. The second objective uses the improved integrated networks to study the gene module changes associated with the fin to limb transition, which was the selected use case. This enabled the identification of crucial conserved and module-specific genes and formulate important evolutionary hypotheses regarding fin to limb transition. The integrative framework developed for the first and second objectives is general and can be adapted to study any other phenotypic comparison given sufficient data. The final objective attempts to solve challenges associated with integrating large-scale phylogenetic trees with large anatomical trait matrices, which required the development of a bioinformatics pipeline. In summary, the computational frameworks developed for this dissertation enables the study of the evolutionary history of a desired anatomical character in a phylogenetic tree and the associated changes in the gene modules which led to the phenotypic changes of the anatomical character. This is greatly beneficial for future evolutionary biology studies.



---

Chairperson, Dr. Paula Mabee

## **ACKNOWLEDGMENTS**

I am thankful to my co-supervisor, Dr. Paula Mabee, for her constant encouragement, financial support, and advice, which was immensely helpful for the completion of this dissertation. Her knowledge and advice regarding evolutionary phenotypic transitions and phylogenetics aided me when performing the evolutionary biological analyses. I also like to thank my co-supervisor, Dr. Erliang Zeng, for his advice and financial assistance. His knowledge and support were crucial for designing and implementing the computational algorithms and the bioinformatics pipelines. Furthermore, I like to thank Dr. Todd Vision, Dr. Bernie Wone, and Dr. Doug Goodman for their valuable inputs and advice for the improvement of my dissertation. I am thankful to Laura Jackson for her support on designing and testing the PhenTree pipeline. I also like to thank Zhixiu Lu for designing the web interface for the PhenTree pipeline. Moreover, I am thankful to Dr. Jim Balhoff for providing the anatomical profiles for the model organisms and his help with the Phenoscope Knowledgebase. The High Performance Computing (HPC) cluster of the University of South Dakota was essential for the implementation of several algorithms, and I would like to thank the HPC support team for their assistance. I am thankful to Dr. Kaius Helenurm for his career guidance and assistance during the completion of this dissertation. I especially thank Suchini Fernando for her continuous encouragement and technical edits. This dissertation was made possible by funding from the National Science Foundation: DBI-1062404, DBI-1062542, DGE-1633213, EPSCoR RII-Track1 Grants and the University of South Dakota Foundation: Raymond D. Dillion Memorial Travel Grant and the Nelson Research Assistantship.

## TABLE OF CONTENTS

DISSERTATION ABSTRACT.....	iv
ACKNOWLEDGMENTS .....	v
LIST OF TABLES.....	viii
LIST OF FIGURES.....	xi
LIST OF SUPPLEMENTARY TABLES.....	xix
LIST OF SUPPLEMENTARY FIGURES.....	xxv
INTRODUCTION.....	1
CHAPTER 1: DEVELOP AN INTEGRATIVE FRAMEWORK BY COMBINING PPI NETWORKS WITH ANATOMY ONTOLOGY DATA AND TEST THE HYPOTHESIS THAT THE INTEGRATION ENHANCED THE ACCURACY OF CANDIDATE GENE PREDICTION ASSOCIATED WITH ANATOMICAL PHENOTYPES.	
Abstract.....	5
1.1 Background.....	6
1.2 Methods.....	13
1.3 Results.....	32
1.4 Discussion.....	41
1.5 Conclusion.....	51
CHAPTER 2: STUDY THE MODULAR STRUCTURE OF THE PHENOTYPIC CHANGES USING THE INTEGRATED NETWORKS	
Abstract.....	108
2.1 Background.....	109
2.2 Methods.....	113

2.3 Results.....	120
2.4 Discussion.....	126
2.5 Conclusion.....	145
<b>CHAPTER 3: INTEGRATE LARGE-SCALE TRAIT DATA WITH LARGE PHYLOGENIES</b>	
<b>BY COMPUTATIONALLY SOLVING THE CHALLENGES ASSOCIATED WITH BIG</b>	
<b>DATA INTEGRATION</b>	
Abstract.....	283
3.1 Background.....	284
3.2 Methods.....	287
3.3 Results.....	296
3.4 Discussion.....	300
3.5 Conclusion.....	307
REFERENCES.....	341

## LIST OF TABLES

<b>Table 1.1.</b> The statistics for the reconciliation of gene names between the anatomical profiles and the STRING PPI networks for zebrafish and mouse.....	52
<b>Table 1.2.</b> The statistics for the unfiltered and filtered anatomy-based gene networks for zebrafish.....	53
<b>Table 1.3.</b> The statistics for the unfiltered and filtered anatomy-based gene networks for mouse.....	54
<b>Table 1.4.</b> The statistics for the unfiltered and filtered integrated networks for zebrafish.....	55
<b>Table 1.5.</b> The statistics for the unfiltered and filtered integrated networks for mouse.....	56
<b>Table 2.1.</b> The statistics for the number of genes with original annotations to each anatomical entity.....	147
<b>Table 2.2.</b> The statistics for the extracted modules.....	148
<b>Table 2.3.</b> Comparison of the 37 genes that are common to the pectoral fin and forelimb modules (conserved genes) (Fig. 2.8 and Fig. 2.9). The genes are ordered according to the rank in the zebrafish module.....	149
<b>Table 2.4.</b> Comparison of the 81 genes that are common to the pelvic fin and hindlimb modules (conserved genes) (Fig. 2.11 and Fig. 2.12). The genes are ordered according to the rank in the zebrafish module.....	152
<b>Table 2.5.</b> The 45 predicted genes of the pectoral fin module ranked and ordered according to the weighted degree of each gene. NA indicates ‘not available’ due to the ortholog not found in the mouse.....	158

**Table 2.6.** The top 50 predicted genes of the pelvic fin module ranked and ordered according to the weighted degree of each gene. NA indicates ‘not available’ due to the ortholog not found in the mouse. The full predicted gene list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles).....160

**Table 2.7.** The 18 predicted genes of the forelimb module ranked and ordered according to the weighted degree of each gene. NA indicates ‘not available’ due to the ortholog not found in the zebrafish.....162

**Table 2.8.** The 32 predicted genes of the hindlimb module ranked and ordered according to the weighted degree of each gene. NA indicates ‘not available’ due to the ortholog not found in the zebrafish.....163

**Table 2.9.** Some of the enriched Biological Process terms from the Gene Ontology and Uberon terms for the mouse orthologs of the pectoral fin module-specific genes that are related to novel anatomical entities emerged in tetrapods during fin to limb transition. The terms are organized into specific anatomical regions.....165

**Table 2.10.** Some of the enriched Biological Process terms from the Gene Ontology and Uberon terms for the mouse orthologs of the pelvic fin module-specific genes that are related to novel anatomical entities emerged in tetrapods during fin to limb transition. The terms are organized into specific anatomical regions.....166

**Table 2.11.** Some of the enriched Biological Process terms from the Gene Ontology and Uberon terms that are related to fin to limb transition for the zebrafish orthologs of the forelimb module-specific genes. The terms are organized into specific anatomical regions.....167

**Table 2.12.** Some of the enriched Biological Process terms from the Gene Ontology and Uberon terms that are related to fin to limb transition for the zebrafish orthologs of the hindlimb module-specific genes. The terms are organized into specific anatomical regions.....168

**Table 3.1.** Percentage of missing data before and after data propagation. The table contains the change in the percentage of missing data before propagation in the pre-processed matrix compared to after propagation in the final output matrix. Missing percentages are relative to the total number of species in the final output matrix (12,582 species; 25,164 cells).....309



## LIST OF FIGURES

**Figure 1.1.** Representation of protein-protein interactions in a graph. The nodes (gray) represent proteins and the edges (black) represent their interactions.....57

**Figure 1.2.** A representation of how Uberon entities are related. Uberon classes are shown in gray boxes. The ‘is\_a’ relationships are represented by full arrows and ‘part\_of’ relationships are represented by dashed arrows. This figure is adapted from the Mungall, et al. (2012) publication.....58

**Figure 1.3.** A hypothetical scenario that compares candidate gene predictions based on a (a) PPI network and an (b) anatomy-based gene network. The nodes A, B, and C in both networks represent three genes known to be associated with a certain phenotype, which can be denoted as phe1. Because their phenotype is known, they are colored in black. In the PPI network (a), genes D and F are predicted to be associated with phe1 based on their interactions with known genes. In contrast, the anatomy-based gene network (b) only predicts D as a potential candidate because the gene F does not have any interaction with other genes. The absence of interactions of gene F can be due to two reasons: (1) it is not annotated with any anatomical terms, (2) it is not annotated with terms that are similar to the anatomy terms associated with genes: A, B, and C. The anatomy-based gene network (b) is built entirely on anatomy ontology information, thus it provides a different interaction structure. Hypothetically, the gene F could be a false positive interaction in the PPI network, and the integrative use of the anatomy-based network may reduce the false positives by filtering them.....59

**Figure 1.4.** The general workflow for generating anatomy-based gene networks. The genes are represented by  $G_1, G_2$ , etc., and their Uberon annotations are represented by  $t_{a1}, t_{b1}$ , etc. In the gene

similarity matrix, the similarity scores between genes are represented by  $s_{11}, s_{12},$   
 etc.....60

**Figure 1.5.** The general evaluation workflow used for evaluating the networks. If single-function evaluation method is selected, distribution of the AUC values is compared, whereas the direct AUC values for the ROC and precision-recall curves are compared in the multi-function method.....61

**Figure 1.6.** Gene similarity score/combined score distributions for (a) zebrafish and (b) mouse unfiltered PPI networks.....62

**Figure 1.7.** The gene similarity score distributions for the zebrafish unfiltered anatomy-based gene networks constructed by (a) Lin method, (b) Resnik method, (c) Schlicker method, and (d) Wang method.....63

**Figure 1.8.** The gene similarity score distributions for the mouse unfiltered anatomy-based gene networks constructed by (a) Lin method, (b) Resnik method, (c) Schlicker method, and (d) Wang method.....64

**Figure 1.9.** The gene similarity score distributions for the zebrafish unfiltered integrated networks constructed by (a) Lin method, (b) Resnik method, (c) Schlicker method, and (d) Wang method.....65

**Figure 1.10.** The gene similarity score distributions for the mouse unfiltered integrated networks constructed by (a) Lin method, (b) Resnik method, (c) Schlicker method, and (d) Wang method.....66

**Figure 1.11.** The comparison of (a) ROC curves and (b) precision-recall curves for different network-based candidate gene prediction methods. These curves were generated for filtered zebrafish PPI networks using the multi-function evaluation method.....67

**Figure 1.12.** The comparison of (a) ROC curves and (b) precision-recall curves for different filtered anatomy-based gene networks for the zebrafish. These curves were generated using the multi-function evaluation method.....68

**Figure 1.13.** The comparison of (a) ROC curves and (b) precision-recall curves for different filtered anatomy-based gene networks for the mouse. These curves were generated using the multi-function evaluation method.....69

**Figure 1.14.** The comparison of (a) ROC curves and (b) precision-recall curves for different filtered integrated networks for the zebrafish. These curves were generated using the multi-function evaluation method.....70

**Figure 1.15.** The comparison of (a) ROC curves and (b) precision-recall curves for different filtered integrated networks for the mouse. These curves were generated using the multi-function evaluation method.....71

**Figure 1.16.** The comparison of ROC curves for the filtered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods in the zebrafish.....72

**Figure 1.17.** The comparison of precision-recall curves for the filtered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods in the zebrafish.....73

**Figure 1.18.** The comparison of ROC curves for the filtered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods in the mouse.....74

**Figure 1.19.** The comparison of precision-recall curves for the filtered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods in the mouse.....75

**Figure 1.20.** The comparison of (a) ROC and (b) precision-recall curves for the filtered non-randomized anatomy-based gene network (blue), random profile anatomy-based gene network (green), and fully random anatomy-based gene network (red) and the comparison of (c) ROC and (d) precision-recall curves for the filtered non-randomized integrated network (blue), random profile integrated network (green), and fully random integrated network for the Wang method for the zebrafish.....76

**Figure 1.21.** The comparison of (a) ROC and (b) precision-recall curves for the filtered integrated network (green), PPI network (red), and anatomy-based gene network (blue) for the Wang method for the zebrafish. The integrated network and the anatomy-based gene network were created using the zebrafish anatomy profile after randomly removing 30 Uberon terms, which had at least 10 gene annotations for each term. The same 30 terms were used for the evaluation to generate the above curves.....77

**Figure 1.22.** The comparison of ROC (a) and precision-recall (b) curves of the filtered integrated network (green), PPI network (red), and anatomy-based gene network (blue) for the Wang method for the zebrafish. The networks were evaluated using annotation profiles that contain Biological Process terms of the Gene Ontology for zebrafish genes.....78

**Figure 2.1.** The ROC curves for the four anatomical entities that were generated during network-based candidate gene prediction evaluations.....169

**Figure 2.2.** The precision-recall curves for the four anatomical entities that were generated during network-based candidate gene prediction evaluations.....170

**Figure 2.3.** Visualization of the pectoral fin module including genes with direct annotations to the pectoral fin (green), genes annotated only to the pectoral fin parts or developmental precursors (blue), and predicted genes (red). Node size is proportional to the degree (number of interactions) of the gene. An interactive version of this module is available in [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) as a Cytoscape network file.....171

**Figure 2.4.** Visualization of the pelvic fin module including genes with direct annotations to the pelvic fin (green), genes annotated only to the pelvic fin parts or developmental precursors (blue), and predicted genes (red). Node size is proportional to the degree (number of interactions) of the gene. An interactive version of this module is available in [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) as a Cytoscape network file.....172

**Figure 2.5.** Visualization of the forelimb module including genes with direct annotations to the forelimb (green), genes annotated only to the forelimb parts or developmental precursors (blue), and predicted genes (red). Node size is proportional to the degree (number of interactions) of the gene. An interactive version of this module is available in [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) as a Cytoscape network file.....173

**Figure 2.6.** Visualization of the hindlimb module including genes with direct annotations to the hindlimb (green), genes annotated only to the hindlimb parts or developmental precursors (blue), and predicted genes (red). Node size is proportional to the degree (number of interactions) of the gene. An interactive version of this module is available in [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) as a Cytoscape network file.....174

**Figure 2.7.** Boxplot comparisons of the distributions of module gene counts in the immediate neighborhood of module genes *versus* network background genes for each anatomical entity. In the boxplots, the red line and the square represent the median and mean, respectively.....175

**Figure 2.8.** Venn diagram showing the number of pectoral fin module-specific genes, conserved genes, and forelimb module-specific genes.....176

**Figure 2.9.** Extractions of the 37 conserved genes from (a) the pectoral fin module and (b) the forelimb module. Node size is proportional to the degree (number of interactions) of the gene. The arrow represents the direction of modular evolution.....177

**Figure 2.10.** Boxplot comparison of normalized weighted degree distributions for (a) pectoral fin module-specific genes, (b) pectoral fin conserved genes, (c) forelimb conserved genes, and (d) forelimb module-specific genes. In the boxplots, the red line and the square represent the median and mean, respectively.....178

**Figure 2.11.** Venn diagram showing the number of pelvic fin module-specific genes, conserved genes, and hindlimb module-specific genes.....179

**Figure 2.12.** Extractions of the 81 conserved genes from (a) the pelvic fin module and (b) the hindlimb module. Node size is proportional to the degree (number of interactions) of the gene. The arrow represents the direction of modular evolution.....180

**Figure 2.13.** Boxplot comparison of normalized weighted degree distributions for (a) pelvic fin module-specific genes, (b) pelvic fin conserved genes, (c) hindlimb conserved genes, and (d) hindlimb module-specific genes. In the boxplots, the red line and the square represent the median and mean, respectively.....181

**Figure 2.14.** The boxplot comparisons of the weighted degree distributions for the predicted genes *versus* genes with original annotations for each module. In the boxplots, the red line and the square represent the median and mean, respectively.....182

**Figure 2.15.** Visualization of the enriched Biological Process terms from the Gene Ontology generated using the REVIGO online tool (<http://revigo.irb.hr/>). The interactions are based on Resnik semantic similarity between the terms.....183

**Figure 3.1.** Ontology-based inference of presence of an anatomical structure. The presence of a structure (pectoral fin) is inferred from a quality (unbranched) of its part (pectoral fin ray), as seen in *Neocyttus rhomboidalis* (Tyler, 1980) The arrows represent the direction of ontological inference, and the ‘X’s represent relationships that are not inferred through ontological reasoning. This figure was adapted from Jackson, et al. (2018).....310

**Figure 3.2.** A schematic representation of the propagation algorithm. During the first iteration (red arrow), data are propagated from genera to corresponding species. For instance, state 1 from Genus A1 is propagated to Species A11, which initially lacked data. During the second iteration (green arrow), data are propagated from families to the remaining species with missing data (Species A22). The character states of the species with existing data are not modified by the propagation during each iteration. For example, the character states of species that had original data (Species A12 and Species A21) are not replaced during first iteration and character states propagated from genera during the first iteration (Species A11) are not replaced during the second iteration.....311

**Figure 3.3.** The general workflow for integrating a synthetic morphological supermatrix retrieved from the Phenoscape Knowledgebase with a species-level tree obtained from the Open Tree of Life to be used for ancestral state reconstruction. The PhenTree pipeline (shown in blue) converts the supermatrix step by step to a version that can be merged with the species-level tree. This figure was adapted from Jackson, et al. (2018).....312

**Fig. 3.4.** A snapshot of the user interface of the PhenTree web tool. A detailed tutorial is available in the tutorial tab. The tool can be accessed using the following link:

<http://phentree.biocombs.org/>.....313

**Fig. 3.5.** Combined usage of inference and propagation extends morphological data. The bar charts show the number of species with asserted (light gray), inferred only (medium gray), and propagated (dark grey) data for the pectoral fin and pelvic fin. Increase in the number of species with data after inference and then propagation demonstrate the importance of these steps in reducing missing data. \*Of the 8,798 species for which pectoral fin data are propagated from family and genus-level data, 5,077 are propagated from asserted data, and 3,721 are propagated from inferred data. \*\*Of the 4,072 species for which pelvic fin data are propagated from family and genus-level data, 2,906 are propagated from asserted data, and 1,166 are propagated from inferred data.....314

**Figure 3.6.** The distribution of asserted (dark blue), inferred only (light blue), and propagated data (green) after performing ancestral state reconstructions for the pectoral and pelvic fins.....315



## LIST OF SUPPLEMENTARY TABLES

<b>Supplementary Table S2.1.</b> The genes with original annotations that were lost due to network cutoff or isolation in the network.....	184
<b>Supplementary Table S2.2.</b> The top 50 genes of the pectoral fin module ranked and ordered based on the weighted degree. The full gene list is available at <a href="https://github.com/pasanfernando/Chapter2_datafiles">https://github.com/pasanfernando/Chapter2_datafiles</a> repository.....	188
<b>Supplementary Table S2.3.</b> The top 50 genes of the pelvic fin module ranked and ordered based on the weighted degree. The full gene list is available at <a href="https://github.com/pasanfernando/Chapter2_datafiles">https://github.com/pasanfernando/Chapter2_datafiles</a> repository.....	190
<b>Supplementary Table S2.4.</b> The top 50 genes of the forelimb module ranked and ordered based on the weighted degree. The full gene list is available at <a href="https://github.com/pasanfernando/Chapter2_datafiles">https://github.com/pasanfernando/Chapter2_datafiles</a> repository.....	192
<b>Supplementary Table S2.5.</b> The top 50 genes of the hindlimb module ranked and ordered based on the weighted degree. The full gene list is available at <a href="https://github.com/pasanfernando/Chapter2_datafiles">https://github.com/pasanfernando/Chapter2_datafiles</a> repository.....	194
<b>Supplementary Table S2.6.</b> The top 100 enriched Biological Process terms from the Gene Ontology for the pectoral fin module-specific genes. The full enriched term list is available at <a href="https://github.com/pasanfernando/Chapter2_datafiles">https://github.com/pasanfernando/Chapter2_datafiles</a> repository.....	196
<b>Supplementary Table S2.7.</b> The top 100 enriched Biological Process terms from the Gene Ontology for the pectoral fin conserved genes. The full enriched term list is available at <a href="https://github.com/pasanfernando/Chapter2_datafiles">https://github.com/pasanfernando/Chapter2_datafiles</a> repository.....	199

**Supplementary Table S2.8.** The top 100 enriched Biological Process terms from the Gene Ontology for the forelimb conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....202

**Supplementary Table S2.9.** The top 100 enriched Biological Process terms from the Gene Ontology for the forelimb module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....205

**Supplementary Table S2.10.** The top 100 enriched Uberon terms for the pectoral fin module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....208

**Supplementary Table S2.11.** The top 100 enriched Uberon terms for the pectoral fin conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....211

**Supplementary Table S2.12.** The top 100 enriched Uberon terms for the forelimb conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....214

**Supplementary Table S2.13.** The top 100 enriched Uberon terms for the forelimb module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....217

**Supplementary Table S2.14.** The enriched Biological Process terms from the Gene Ontology for the mouse orthologs of the pectoral fin module-specific genes.....220

**Supplementary Table S2.15.** The top 100 enriched Uberon terms for the mouse orthologs of the pectoral fin module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....223

**Supplementary Table S2.16.** The enriched Biological Process terms from the Gene Ontology for the zebrafish orthologs of the forelimb module-specific genes.....226

**Supplementary Table S2.17.** The enriched Uberon terms for the zebrafish orthologs of the forelimb module-specific genes.....228

**Supplementary Table S2.18.** The top 100 enriched Biological Process terms from the Gene Ontology for the pelvic fin module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....229

**Supplementary Table S2.19.** The top 100 enriched Biological Process terms from the Gene Ontology for the pelvic fin conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....232

**Supplementary Table S2.20.** The top 100 enriched Biological Process terms from the Gene Ontology for the hindlimb conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....235

**Supplementary Table S2.21.** The top 100 enriched Biological Process terms from the Gene Ontology for the hindlimb module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....238

**Supplementary Table S2.22.** The enriched Uberon terms for the pelvic fin module-specific genes.....241

**Supplementary Table S2.23.** The top 100 enriched Uberon terms for the pelvic fin conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....243

**Supplementary Table S2.24.** The top 100 enriched Uberon terms for the hindlimb conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....246

**Supplementary Table S2.25.** The top 100 enriched Uberon terms for the hindlimb module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....249

**Supplementary Table S2.26.** The top 100 enriched Biological Process terms from the Gene Ontology for the mouse orthologs of the pelvic fin module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....252

**Supplementary Table S2.27.** The top 100 Uberon terms for the mouse orthologs of the pelvic fin module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....255

**Supplementary Table S2.28.** The enriched Biological Process terms from the Gene Ontology for the zebrafish orthologs of the hindlimb module-specific genes.....258

**Supplementary Table S2.29.** The enriched Uberon terms for the zebrafish orthologs of the hindlimb module-specific genes.....261

**Supplementary Table S2.30.** The enriched Biological Process terms from the Gene Ontology that are common to the predicted genes and genes with original annotations for the pectoral fin. The enriched terms are sorted based on the p-value of those terms for the predicted genes.....263

**Supplementary Table S2.31.** The enriched Biological Process terms from the Gene Ontology that are common to the predicted genes and genes with original annotations for the pelvic fin.

The enriched terms are sorted based on the p-value of those terms for the predicted genes.....265

**Supplementary Table S2.32.** The enriched Biological Process terms from the Gene Ontology that are common to the predicted genes and genes with original annotations for the forelimb. The enriched terms are sorted based on the p-value of those terms for the predicted genes.....266

**Supplementary Table S2.33.** The enriched Biological Process terms from the Gene Ontology that are common to the predicted genes and genes with original annotations for the hindlimb. The enriched terms are sorted based on the p-value of those terms for the predicted genes.....269

**Supplementary Table S2.34.** The enriched Uberon terms that are common to the predicted genes and genes with original annotations for the pectoral fin. The enriched terms are sorted based on the p-value of those terms for the predicted genes.....274

**Supplementary Table S2.35.** The enriched Uberon terms that are common to the predicted genes and genes with original annotations for the pelvic fin. The enriched terms are sorted based on the p-value of those terms for the predicted genes.....276

**Supplementary Table S2.36.** The top 100 enriched Uberon terms that are common to the predicted genes and genes with original annotations for the forelimb. The enriched terms are sorted based on the p-value of those terms for the predicted genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....277

**Supplementary Table S2.37.** The top 100 enriched Uberon terms that are common to the predicted genes and genes with original annotations for the hindlimb. The enriched terms are sorted based on the p-value of those terms for the predicted genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.....280

**Supplementary Table S3.1.** Propagation statistics for the matrix that contained parts of the paired fins (115 characters) for Teleostei. The propagation algorithm was implemented on one character at a time; therefore, the statistics are recorded separately for each character.....316

**Supplementary Table S3.2.** The species that were mismatched between the final output matrix that contains Vertebrate Taxonomy Ontology (VTO) taxon names and the Teleostei species-level tree from the Open Tree of Life that contains taxon names based on the NCBI taxonomy system.....323

## LIST OF SUPPLEMENTARY FIGURES

**Supplementary Figure S1.1.** The boxplot comparisons of the AUC distributions for (a) ROC curves and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for the different network-based candidate gene prediction methods for the zebrafish filtered PPI network. In the boxplots, the red line and the square represent the median and mean, respectively.....79

**Supplementary Figure S1.2.** The boxplot comparisons of the AUC distributions for (a) ROC curves and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for different filtered anatomy-based gene networks for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.....80

**Supplementary Figure S1.3.** The boxplot comparisons of the AUC distributions for (a) ROC curves and (a) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for the different filtered anatomy-based gene networks for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.....81

**Supplementary Figure S1.4.** The boxplot comparisons of the AUC distributions for (a) ROC curves and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for the different filtered integrated networks for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.....82

**Supplementary Figure S1.5.** The boxplot comparisons of the AUC distributions for (a) ROC curves and (b) precision-recall curves and the histogram comparisons of the AUC distributions

for (c) ROC curves and (d) precision-recall curves for the different filtered integrated networks for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.....83

**Supplementary Figure S1.6.** The boxplot comparisons for the AUC distributions of ROC curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.....84

**Supplementary Figure S1.7.** The boxplot comparisons for the AUC distributions of precision-recall curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.....85

**Supplementary Figure S1.8.** The histogram comparisons for the AUC distributions of ROC curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish.....86

**Supplementary Figure S1.9.** The histogram comparisons for the AUC distributions of precision-recall curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish.....87

**Supplementary Figure S1.10.** The boxplot comparisons for the AUC distributions of ROC curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.....88

**Supplementary Figure S1.11.** The boxplot comparisons for the AUC distributions of precision-recall curves for filtered anatomy-based gene networks, integrated networks, and PPI networks



for the four semantic similarity calculation methods for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.....89

**Supplementary Figure S1.12.** The histogram comparisons for the AUC distributions of ROC curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse.....90

**Supplementary Figure S1.13.** The histogram comparisons for the AUC distributions of precision-recall curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse.....91

**Supplementary Figure S1.14.** The comparison of ROC curves for the unfiltered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods for the zebrafish.....92

**Supplementary Figure S1.15.** The comparison of precision-recall curves for the unfiltered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods for the zebrafish.....93

**Supplementary Figure S1.16.** The comparison of ROC curves for the unfiltered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods for the mouse.....94

**Supplementary Figure S1.17.** The comparison of precision-recall curves for the unfiltered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods for the mouse.....95

**Supplementary Figure S1.18.** The boxplot comparisons for the AUC distributions of ROC curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for

the four semantic similarity calculation methods for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.....96

**Supplementary Figure S1.19.** The boxplot comparisons for the AUC distributions of precision-recall curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.....97

**Supplementary Figure S1.20.** The histogram comparisons for the AUC distributions of ROC curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish.....98

**Supplementary Figure S1.21.** The histogram comparisons for the AUC distributions of precision-recall curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish.....99

**Supplementary Figure S1.22.** The boxplot comparisons for the AUC distributions of ROC curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.....100

**Supplementary Figure S1.23.** The boxplot comparisons for the AUC distributions of precision-recall curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.....101

**Supplementary Figure S1.24.** The histogram comparisons for the AUC distributions of ROC curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse.....102

**Supplementary Figure S1.25.** The histogram comparisons for the AUC distributions of precision-recall curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse.....103

**Supplementary Figure S1.26.** The boxplot comparisons of the AUC distributions for (a) ROC and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC and (d) precision-recall curves for the filtered non-randomized anatomy-based gene network (blue), random profile anatomy-based gene network (green), and fully random anatomy-based gene network (red) for the Wang method for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.....104

**Supplementary Figure S1.27.** The boxplot comparisons of the AUC distributions for (a) ROC and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC and (d) precision-recall curves for the filtered non-randomized integrated network (blue), random profile integrated network (green), and fully random integrated network (red) for the Wang method for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.....105

**Supplementary Figure S1.28.** The boxplot comparisons of the AUC distributions for (a) ROC and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC and (d) precision-recall curves for the filtered integrated network (green), PPI network (red), and anatomy-based gene network (blue) for the Wang method for the zebrafish. The integrated network and the anatomy-based gene network were generated using the zebrafish anatomy profile after randomly removing 30 Uberon terms, which had at least 10 gene annotations. The same 30 terms were used for the evaluation to generate the above distributions.

In the boxplots, the red line and the square represent the median and mean, respectively.....106

**Supplementary Figure S1.29.** The boxplot comparisons of the AUC distributions for (a) ROC and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC and (d) precision-recall curves for the filtered integrated network (green), PPI network (red), and anatomy-based gene network (blue) for the Wang method in zebrafish. The networks were evaluated using the annotation profiles containing Biological Process terms of Gene Ontology (GO) for the zebrafish genes. In the boxplots, the red line and the square represent the median and mean, respectively.....107

## INTRODUCTION

The process of evolution is correlated with constant changes in phenotypes of species, which lead to the wealth of phenotypic diversity observed among different species today. Studying such phenotypic changes is a cornerstone in evolutionary biology and has assisted in developing modern disciplines, such as the evolutionary developmental biology (evo–devo), which is focused on studying the relationships between the processes of individual development and phenotypic changes during evolution (Austin, 2017; Müller, 2007; Raff, 2000). One major concern of such disciplines is to link the changes in phenotypes to their corresponding genotypic changes. It is important to understand what genes and genetic interactions were lost and what new interactions were gained during an evolutionary transition, such as the fin to limb transition. This research attempts to use protein-protein interaction (PPI) networks and anatomy ontology data to study such phenotypic transitions by building integrative algorithms to better identify genetic patterns and study the evolution of phenotypic traits.

Traditionally, wet lab methods, such as gene knockout (Erard, et al., 2017; Hall, et al., 2009), gene knockdown (Huang, et al., 2013), overexpression (Amatruda, et al., 1992; Gu, et al., 2014), natural mutation (Albalat, et al., 2010), and ectopic expression (Alexandre, et al., 1996; Joos, et al., 2018), are used to predict gene associations to phenotypes. These methods alter the expression of a gene or a gene set and observe its effect on desired phenotypes to unravel the phenotypic function of the genes. These wet lab methods are accurate in their predictions but suffer from high resource and time consumption. Therefore, faster computational candidate gene prediction methods are required to predict gene candidates for desired phenotypes.

Computational candidate gene prediction methods (Cowen, et al., 2017; Zhang, et al., 2017) have gained a reputation in the last few decades due to the wealth of genomic and

proteomic data accumulated using next-generation sequencing methods. They are much faster than the wet lab methods and do not require experimental resources. Computational candidate gene prediction algorithms work on different biological principles, such as the sequence similarity (Zhang, et al., 2017) and PPI (Cowen, et al., 2017; Sharan, et al., 2007), to predict new gene candidates. The use of PPI networks has become widespread for candidate gene prediction, especially for predicting disease-causing genes, due to the availability of large PPI datasets in public databases.

The PPI networks are computationally stored using the graph data structure, which enables the use of graph theory and other network/graph analysis algorithms to identify network patterns and modules (Cormen, 2009; Cowen, et al., 2017). The goal here is to use biological network analysis algorithms to study the phenotypic transitions during the evolution. This also includes building novel integrative network frameworks to enhance the efficiency of the analysis. Generally, traditional wet lab methods and other computational prediction methods can predict the phenotypic function of a gene, but they fail to identify the gene interactions that determine the phenotype. In addition to the individual roles of the genes, their interactions with each other have a significant effect in determining the phenotype (Sharan, et al., 2007). Analyzing biological networks, such as PPI networks, allows identifying important gene interactions and modules that determine phenotypes, which is a significant advantage of the biological network analysis (Cowen, et al., 2017; Yu, et al., 2013); hence, the use of biological networks in this work.

This dissertation contains three objectives that use network algorithms in a sequential manner to understand the genotypic basis behind evolutionary transitions. These involve building novel integrative frameworks, integration of large networks, and computationally solving the

challenges when integrating and analyzing large biological datasets. The fin to limb transition (Amaral and Schneider, 2018; Coates, 1994; Don, et al., 2013) is an important event that is associated with aquatic to terrestrial transformation of vertebrates (Clack, 2012; Long and Gordon, 2004); hence, it was selected as the biological use case to demonstrate the computational methods.

The main challenge with biological network analysis is the low network quality and the low prediction accuracy associated with the candidate gene predictions (Sharan, et al., 2007; von Mering, et al., 2002). Before using PPI networks for studying evolutionary transitions, their quality must be enhanced for analyzing anatomical phenotypes. The first objective of this dissertation focuses on integrating experimental knowledge about gene-anatomical phenotype relationships *via* anatomy ontology, which includes building a novel integrative framework that uses anatomy ontology and PPI networks and evaluating the networks to assess whether the quality has been improved.

The second objective focuses on using the quality-enhanced PPI networks to analyze the genetic changes behind phenotypic transitions during evolution. PPI networks allow the identification of gene modules associated with anatomical phenotypes, such as the pectoral fin and the forelimb, and observe the changes in those modules over the course of evolution. For instance, the zebrafish pectoral fin module can be compared with the mouse forelimb module to discover the modular evolution of genes including the changes in the genes, gene interactions, and also the importance of the genes in the module. These modular changes can be associated with the fin to limb transition, which is extremely valuable and cannot be directly analyzed using wet lab methods.

The first two objectives focus on the gene-to-phenotype relationship when analyzing evolutionary transitions. Another aspect when studying evolution is mapping anatomical phenotypes to large phylogenetic trees, which enables the understanding of how anatomical characters, such as the pectoral fin, evolved in different species during the course of evolution. When studying the phenotypic changes during evolution, large-scale phylogenetic trees and trait matrices are required for a better understanding of those changes, which introduces various challenges in large-scale data integration (Dececchi, et al., 2016; Didier, 2017; Jackson, et al., 2018). The third objective focuses on identifying these challenges and solving them using graph/network algorithms to efficiently integrate large phylogenetic trees with trait matrices. This completes the three-way relationship between genes, phenotypes, and taxonomy for a better understanding about evolutionary transitions. In summary, this dissertation demonstrates how large-scale biological networks, large data integration, and graph/network algorithms can be used to computationally study phenotypic changes associated with evolutionary transitions, which will be beneficial for evolutionary biology in the post-genomic era.



# **CHAPTER 1: DEVELOP AN INTEGRATIVE FRAMEWORK BY COMBINING PPI NETWORKS WITH ANATOMY ONTOLOGY DATA AND TEST THE HYPOTHESIS THAT THE INTEGRATION ENHANCED THE ACCURACY OF CANDIDATE GENE PREDICTION ASSOCIATED WITH ANATOMICAL PHENOTYPES.**

## **Abstract**

This objective focuses on developing an integrative framework using protein-protein interaction (PPI) networks and anatomy ontology data to improve the candidate gene prediction accuracy. The integration was expected to improve the quality of the PPI networks by reducing false positive and false negative interactions. The hypothesis was that the integrated networks have a better candidate gene prediction accuracy compared to the original PPI networks due to the increase in network quality. To test this hypothesis, the candidate gene prediction performance of the integrated networks and the original PPI networks were compared using the mouse and the zebrafish anatomical profiles. According to the results, integrated networks outperformed the PPI networks and confirmed that the integration of anatomy ontology data with PPI networks improves the quality of the interactions. The increased candidate gene prediction accuracy of the integrated networks was observed under diverse computational settings, including four semantic similarity calculation methods (Lin, Resnik, Schlicker, and Wang), two evaluation curves (ROC and precision-recall curves), two evaluation modes (single-function evaluation method *versus* multi-function evaluation method), two model organisms (mouse and zebrafish), and two network conditions (filtered vs. unfiltered), which showed their robustness. The results were further validated to confirm the biological significance and to negate the effect

of the circular use of the anatomy profiles, which also confirmed the significance of the integrative networks proving their usability in biological network analyses.

## **1.1 Background**

Unraveling the molecular and phenotypic functions of proteins is a cornerstone in biology as it improves our understanding of the biological system, which leads to a plethora of practical implications, such as drug development and agricultural improvements. However, a large proportion of gene sequences are still not functionally annotated due to experimental constraints (Erdin, et al., 2011; Zhang, et al., 2017). Therefore, computational candidate gene prediction has become widespread in bioinformatics to improve the pace of the gene functional annotations. There are different candidate gene prediction algorithms that are based on different biological principles, such as sequence similarity (Jones and Swindells, 2002; Zhang, et al., 2017), phylogenetic profile-based similarity (Pellegrini, et al., 1999), protein structure-based similarity (Yachdav, et al., 2014), expression profile-based similarity (Wang, et al., 2017), and PPI networks (Cowen, et al., 2017; Sharan, et al., 2007). Some of these methods have limitations; for instance, sequence similarity-based candidate gene prediction methods assign functions based on sequence homology. However, in some instances, proteins having very similar sequences are shown to have different functions (Erdin, et al., 2011). Alternatively, PPI network-based candidate gene prediction has become widespread because it is based on experimental interactions, and proteins are more likely to share the same function when they are physically interacting with each other.

### *1.1.1 Protein-protein interaction (PPI) networks*

PPI networks represent the physical interactions of proteins for a given organism as a computational graph (Cowen, et al., 2017; Sharan, et al., 2007; Zickenrott, et al., 2017). The proteins in the graph are represented by the nodes and the interactions between them are represented by the edges (Fig. 1.1). Protein interaction data are experimentally elucidated by methods, such as the yeast two-hybrid assay, high-throughput mass-spectrometric protein complex identification (HMS-PCI), and X-ray crystallography (Shoemaker and Panchenko, 2007). There are databases such as the Database of Interacting Proteins (DIP) that include only experimental protein interactions (Xenarios, et al., 2001). Initially, the number of experimental protein interactions was not adequate for large-scale analyses of biological systems. Therefore, some experiments, such as the yeast two-hybrid assay and high-throughput mass-spectrometry are now conducted at a large-scale using next-generation methods (Shoemaker and Panchenko, 2007). However, the experimental interaction data are still incomplete, and a portion of the interactions generated using next-generation methods are known to be false positives (Szkarczyk, et al., 2017; von Mering, et al., 2002; Zeng, et al., 2008). To solve these issues, protein-protein interactions were computationally predicted using algorithms that use additional biological properties of genes, such as the co-expression (Shoemaker and Panchenko, 2007). As a result, databases such as STRING (Szkarczyk, et al., 2017) began to emerge, which contain both experimentally and computationally predicted protein interactions.

The STRING database (<https://string-db.org/>) has gained a rapid popularity because of the abundance, the higher coverage, and the quality control of the PPI data it stores (Franceschini, et al., 2013; Szkarczyk, et al., 2015; Szkarczyk, et al., 2017; von Mering, et al., 2005). Currently, it has PPI data for 2031 organisms, which is the highest number of taxa in

comparison to any other PPI database (Szklarczyk, et al., 2017). In addition to the traditional protein interaction data elucidated by experimental methods, STRING also includes protein interaction data derived from literature mining and computational predictions. As a result, the traditional term: “protein-protein interaction”, which was used to describe direct physical interactions between proteins, has been updated to “protein-protein association”, which is based on functional association not limited to direct interactions (von Mering, et al., 2005). This accounts for the higher abundance of the PPI data for a large number of organisms in the STRING database, and because it is also quality controlled, it is the prime source for PPI network analyses.

The central application of the PPI networks in bioinformatics is to predict the molecular or phenotypic function of unknown/unannotated proteins. Usually, in a PPI network, there is a proportion of proteins with known functions, and the functions of others are unknown, i.e., unannotated. The proteins with known functions can be used to predict the function of unannotated proteins based on the assumption that the interacting proteins are more likely to share common functions (Cowen, et al., 2017; Sharan, et al., 2007). There are several computational algorithms developed for this task that use graph theory or other network analysis techniques. Some algorithms are generally applied in other domains, such as social networks, physical computer networks, trade networks, chemical networks, etc. (Martínez-López, et al., 2009; Salter-Townshend, et al., 2012).

The algorithms for PPI network-based candidate gene prediction can be separated into two major groups based on their execution: module assisted methods and direct methods (Sharan, et al., 2007). The module assisted methods first look for a functional module in the network and then assign the function or functions of the known proteins in the module to all the

other unannotated members (Bader and Hogue, 2003; Pizzuti and Rombo, 2014; Tripathi, et al., 2016). Direct methods do not have a module detection step, instead, they directly assign the functions to unannotated proteins based on their interactions to known neighbors (Chua, et al., 2006; Kourmpetis, et al., 2010). Neighborhood counting is a widespread and a simple direct method category that has been used for candidate gene prediction for years. The most frequently used and the simplest neighborhood counting method is the Majority Voting method, which obtains the majority vote from the known neighbor proteins when assigning the function to unannotated proteins (Schwikowski, et al., 2000; Zeng, et al., 2012). Performance comparisons have shown that the direct methods have a higher prediction accuracy compared to the module assisted methods (Sharan, et al., 2007); therefore, a neighborhood counting method that uses chi-square-like scores to assign gene function (Hishigaki, et al., 2001) was used for this analysis.

As with any computational prediction method, the biggest challenge for the network-based candidate gene prediction algorithms is to increase the accuracy of the predictions (Cowen, et al., 2017; Sharan, et al., 2007; Zeng, et al., 2012). There are two main approaches taken to fulfill this goal: (1) developing more advanced algorithms to improve the prediction accuracy, (2) improving the quality of the PPI networks. Although new algorithms are added to the candidate gene prediction algorithm arsenal regularly, the quality of the PPI networks must be improved to make better use of those algorithms. Current PPI data are generated at a large-scale using next-generation PPI detection methods, but they also accumulate network noise or false positive interactions (Sharan, et al., 2007; Szklarczyk, et al., 2017; von Mering, et al., 2002). Furthermore, the PPI networks are still incomplete and may not represent all possible interactions within a cell (Hart, et al., 2006). Therefore, improving the quality of the PPI networks using experimental and computational methods is a hot topic in bioinformatics

(Cowen, et al., 2017; Shoemaker and Panchenko, 2007; Szklarczyk, et al., 2017; von Mering, et al., 2005).

The computational methods that improve the PPI network quality involve integrating information from various other data sources, such as literature text data and gene expression profiles, with experimental PPI networks, of which the STRING database is the prime example (Szklarczyk, et al., 2017; von Mering, et al., 2005). However, there is always a room for improvement by using newly available data sources. Particularly, when using PPI networks for predicting the phenotype of the gene instead of the molecular function, additional challenges arise because a phenotype is a result of numerous proteins participating in several biological pathways. The prediction of a phenotype is not as straightforward as predicting a molecular function of a protein (Cowen, et al., 2017). Therefore, computational methods are essential to improve the quality of the PPI networks with the emphasis on their usage in predicting the phenotypes of the proteins. Experimental knowledge about gene-to-phenotype relationships, such as diseases and development of anatomical structures, is continuously accumulated in literature. However, there is a lack of methods to integrate this information directly with PPI networks and use it for candidate gene prediction. One way to use this information is through biological ontologies, which is the focus of this chapter.

### *1.1.2 Biological Ontologies*

Ontologies conceptualize the knowledge of certain domains by representing them as a computable, hierarchical set of terms that can be easily used for computational analysis (Gene Ontology, 2016; Pesquita, et al., 2009). In simple terms, ontologies can be understood as vocabularies that aid computer algorithms to understand subject domains, such as biological

functions, anatomical systems, environments, web information, etc. In bioinformatics, the Gene Ontology (GO) is the leading ontology, which has been used for biological analyses focused on the molecular function, the cellular location, and the biological processes of genes (Ashburner, et al., 2000; Consortium, 2004; Gene Ontology, 2016). The GO has three domain subontologies to aid in these areas: cellular component (GO-CC), biological process (GO-BP), and molecular function (GO-MF). The GO has been integrated with numerous biological tools and has become the standard functional annotation system in several biological databases, such as the NCBI GenBank (Benson, et al., 2008). Recently, various other biological ontologies have emerged focusing on specific biological aspects, such as the Human Phenotype Ontology (HPO) for human phenotypes (Köhler, et al., 2016), the Vertebrate Taxonomy Ontology (VTO) to store taxon information on vertebrates (Midford, et al., 2013), and the Disease Ontology to store human disease information (Kibbe, et al., 2015). The focus of this research is anatomical phenotypes; hence, the most suitable ontology to use is the Uberon anatomy ontology (Elhanan, et al., 2017; Haendel, et al., 2014; Mungall, et al., 2012).

Uberon is an integrated cross-species anatomy ontology, which constitutes over 14,000 classes representing anatomical entities (Elhanan, et al., 2017; Haendel, et al., 2014; Mungall, et al., 2012). This ontology integrates species-specific anatomical ontologies, such as the Mouse Anatomy Ontology (MA), Xenopus Anatomy Ontology (XAO) and Zebrafish Anatomy Ontology (ZFA) in a species-neutral way, along with multi-species anatomical ontologies such as Mammalian Phenotype Ontology (MP) and Teleost Anatomy Ontology (TAO) (Dahdul, et al., 2012; Dahdul, et al., 2010). This integration provides the required bridge between different taxa, which can be used for computational analyses involving multiple species, another characteristic

that makes Uberon a suitable candidate for the scope of this particular research. A basic example of how anatomical entities are related in Uberon is demonstrated in Fig. 1.2.

The goal of this work is to improve the candidate gene prediction accuracy of PPI networks by improving their network quality. For this purpose, the Uberon anatomy ontology is used to capture the experimental knowledge stored in the literature regarding gene to anatomical phenotype relationships, such as genes involved with the development of the pectoral fin, forelimb, and axial skeleton. A novel integrative framework is proposed to integrate Uberon anatomy ontology data with PPI networks. The first step of the integrative framework is to generate a gene network based on the anatomy ontology (anatomy-based gene network), which is constructed by calculating the semantic similarity between Uberon terms annotated to genes. Semantic similarity is used to represent the similarity between two ontology terms based on their relationship and the proximity in the ontology structure (Pesquita, et al., 2009; Wang, et al., 2007) or by calculating the information content using the number of gene annotations for each term (Resnik, 1995; Schlicker, et al., 2006).

The second step is to integrate the anatomy-based gene network with the PPI network using an accuracy-based weighting framework. The hypothesis is that the integration of the PPI network with the anatomy-based gene network will enhance the accuracy of the candidate gene prediction by improving the quality of the PPI network. This may result through either removing false positive interactions in the PPI network or by introducing new interactions to the PPI network by the incorporation of anatomy-based gene network interactions. This hypothesis can be further explained using the hypothetical scenario represented in Fig. 1.3. The candidate gene prediction performance of the integrated network and the PPI network can be compared to confirm the validity of the hypothesis.



The concept of constructing gene networks by calculating the semantic similarity between ontology terms has been presented before (Jiang, et al., 2011; Le and Dang, 2016; Zeng, et al., 2012), but they were only applied to GO. For example, Jiang, et al. (2011) constructed a gene network using the GO-BP component to infer disease genes in human. However, the network was not integrated with an existing PPI network, instead, it was used directly for disease gene prediction and the results were compared with a human PPI network. In the proposed integrative framework, the anatomy-based gene network constructed using anatomy ontology data is integrated with the PPI network, which will only keep the interactions that are supported by both the source networks. The integration is expected to enhance the quality of the source networks by reducing the false positive and false negative interactions. Furthermore, to my knowledge, a gene similarity network has not been constructed before using anatomy ontology. Therefore, the proposed integrative framework will be beneficial for studies that use anatomical phenotypes, such as evolutionary analyses studying the anatomical changes in organisms during the evolutionary history, which will be a valuable addition to the computational candidate gene prediction arsenal that mostly focuses on disease genes in humans.

## **1.2 Methods**

### *1.2.1 Retrieval and pre-processing of the protein-protein interaction (PPI) datasets from the STRING database*

The proposed integrative framework can be applied to any PPI dataset if the Uberon phenotypic annotations are available; but for this work, the focus is on phenotypic changes associated with the aquatic to terrestrial vertebrate transition; therefore, the zebrafish and the mouse were selected to represent aquatic and terrestrial vertebrates, respectively. The zebrafish is

the most frequently used model organism to represent fishes, and the mouse is an extensively studied model organism to represent terrestrial vertebrates. The PPI datasets for the zebrafish and the mouse were downloaded from the STRING database (<https://string-db.org/>). The proteins in the PPI datasets are represented by a unique STRING identifier (id). However, to facilitate the integration at later stages, the STRING id was converted into the gene symbol during the pre-processing step. The PPI datasets contain a combined score for each interaction to represent its strength (Szkarczyk, et al., 2017; von Mering, et al., 2005). Usually, the full PPI networks are too large for downstream analyses; therefore, the networks must be filtered based on a cutoff. The recommended cutoff is 0.7 to select only the high-quality interactions (von Mering, et al., 2005), which was applied to the mouse and zebrafish networks.

### 1.2.2 Construction of the anatomical profiles

In order to construct the anatomy-based gene networks for the mouse and the zebrafish, initially, their anatomical profiles must be constructed. Anatomical profiles represent the multiple anatomical term annotations for each gene. Usually, a single gene is annotated with multiple Uberon terms, which can be represented in the following format: consider two genes represented by  $G_1$  and  $G_2$ . If their associated Uberon terms are represented by  $(t_{a1}, t_{a2} \dots t_{am})$  and  $(t_{b1}, t_{b2} \dots t_{bn})$ , respectively, their anatomical profiles are given below.

$$G_1: (t_{a1}, t_{a2} \dots t_{am})$$

$$G_2: (t_{b1}, t_{b2} \dots t_{bn})$$

To obtain the anatomical profiles, the associations between genes and Uberon terms were retrieved from the Monarch Initiative repository (<https://monarchinitiative.org/>) via a script written in Scala (Odersky, 2008). Then the redundant associations were deduplicated and the

associations were converted into anatomical profiles that are in the format as shown above. The Monarch Initiative retrieves genes and their anatomical phenotype annotations from the zebrafish (<https://zfin.org/>) and the mouse (<http://www.informatics.jax.org/>) model organism databases and associate them with the corresponding Uberon terms. Moreover, the annotations available in the Monarch Initiative are pre-processed and cross-checked with other model organism annotations to remove uncertain gene-phenotype associations that are resulted when the expression of multiple genes are disrupted at one time to observe the effect on a given phenotype (Mungall, et al., 2017).

### *1.2.3 Gene name/symbol reconciliation between the PPI networks and the anatomical profiles*

Before the integration, it is important to reconcile the gene names/symbols between the two data sources. The STRING database obtains data from various data sources, such as Entrez Gene database (Maglott, et al., 2005) and UniProt knowledgebase (Apweiler, et al., 2004), and the Monarch Initiative data repository obtains data from model organism databases.

Occasionally, the gene names do not match, and a manual inspection indicated that some STRING gene names are outdated compared to data from the Monarch Initiative. A computational method was developed to match the genes between the two sources, which works in three rounds: (1) match the genes directly using their names/symbols, (2) match using their Ensembl identifiers, and (3) match the gene names in the anatomical profiles to the synonyms available in the STRING database. Each round attempted to sequentially minimize the number of gene mismatches.

After the reconciliation, the outdated gene names in the PPI networks were replaced by the up-to-date gene names from the anatomical profiles. However, there were still genes in the

anatomical profiles that were mismatched with those in the PPI networks. Therefore, reconciled anatomical profiles, which only contain the matched genes, were generated for mouse and zebrafish to use for network evaluations in section 1.2.7. Furthermore, these profiles contain a majority of Uberon terms with only few genes annotated to each term. For instance, evaluations that are based on Uberon terms that have only two or three genes are not reliable. In reality, those genes may have more gene-to-phenotype associations that are still not discovered by experimental gene function/phenotype prediction methods. The network-based candidate gene prediction algorithm could predict those associations correctly, but since they are unknown, they will be considered as false positives. Therefore, the reconciled mouse and zebrafish anatomical profiles were filtered to contain only the Uberon terms that have at least 10 gene annotations. These filtered profiles were used for the evaluations in section 1.2.7.

#### *1.2.4 Generation of anatomy-based gene networks*

To generate gene networks based on anatomy ontology, similarity scores between gene pairs must be calculated. This can be achieved by first calculating semantic similarity scores between Uberon anatomical terms annotated to a particular gene pair, and then, aggregating those scores to calculate a single similarity score between the two genes. The general workflow for generating anatomy-based gene networks is represented in Fig. 1.4.

To calculate semantic similarity between Uberon anatomical terms, four methods were used: Wang method (Wang, et al., 2007), Resnik method (Resnik, 1995), Lin method (Lin, 1998), and Schlicker method (Schlicker, et al., 2006). The latter three methods (Resnik, Lin, and Schlicker) are based on calculating the information content (IC) of each node in the ontology, and their equations are given below (equations 1.1, 1.2, and 1.3).

Resnik:

$$sim_R(t_1, t_2) = \max_{t \in S(t_1, t_2)} \{IC(t)\} \quad (1.1)$$

Lin:

$$sim_L(t_1, t_2) = \max_{t \in S(t_1, t_2)} \left\{ \frac{2IC(t)}{IC(t_1) + IC(t_2)} \right\} \quad (1.2)$$

Schlicker:

$$sim_S(t_1, t_2) = \max_{t \in S(t_1, t_2)} \left\{ \frac{2IC(t)}{IC(t_1) + IC(t_2)} (1 + IC(t)) \right\} \quad (1.3)$$

In the above equations,  $t_1$  and  $t_2$  represent the ontology terms of which the similarity is calculated, whereas  $S$  denotes the set of common ancestors for the two terms. The information content for a given term  $t$  is represented by  $IC(t)$ , which is calculated based on the number of genes annotated to the term  $t$  as illustrated below (equations 1.4 and 1.5).

$$IC(t) = -\log(p(t)) \quad (1.4)$$

$$P(t) = \frac{\text{Number of genes associated with the term } t + \text{constant}}{\text{Total number of genes associated with the entire ontology}} \quad (1.5)$$

The Wang method is another widespread semantic similarity calculation method but it does not involve the information content. It only depends on the ontology structure and the relationships between the terms. The equation for the Wang semantic similarity calculation between two terms  $t_1$  and  $t_2$  is given below.

Wang:

$$sim_W(t_1, t_2) = \frac{\sum_{t \in T_1 \cap T_2} (S_{t_1}(t) + S_{t_2}(t))}{SV(t_1) + SV(t_2)} \quad (1.6)$$

In the above equation,  $S_{t_i}(t)$  represents the semantic contribution of term ' $t$ ' on term  $t_i$ , when ' $t$ ' is an ancestor of  $t_i$ . The term  $SV(t_i)$  represents the semantic contribution of all the

ancestors of term  $t_i$  on itself. This method has proven to yield better results compared to other information content-based methods (Wang, et al., 2007).

The above four semantic similarity calculation methods were used to calculate the similarity between gene pairs using the method explained below. Let us assume that the gene  $G_1$  is annotated with the Uberon terms:  $(t_{a1}, t_{a2}... t_{am})$ , and the gene  $G_2$  is annotated with the Uberon terms:  $(t_{b1}, t_{b2}...t_{bn})$ , then the similarity between the two genes,  $sim(G_1, G_2)$ , can be calculated using the following equation (1.7).

$$sim(G_1, G_2) = \frac{\sum_{1 \leq i \leq m} sim(t_{ai}, t(G_2)) + \sum_{1 \leq j \leq n} sim(t_{bj}, t(G_1))}{m+n} \quad (1.7)$$

The  $t(G_1) = (t_{a1}, t_{a2}... t_{am})$  and  $t(G_2) = (t_{b1}, t_{b2}...t_{bn})$  represent Uberon term anatomical profiles for gene  $G_1$  and gene  $G_2$ , respectively, and the  $sim(t_{ai}, t(G_2))$  represents the maximum semantic similarity between term  $t_{ai}$  and any of the terms in  $t(G_2)$ , which can be calculated using the equation (1.8) below.

$$sim(t_{ai}, t(G_2)) = \max_{t_b \in t(G_2)} sim(t_{ai}, t_b) \quad (1.8)$$

Using the equation (1.7), a similarity score for each gene pair in the anatomical profiles of the mouse and the zebrafish was calculated. This generates a pairwise gene similarity matrix for each semantic similarity calculation method (Lin, Resnik, Schlicker, and Wang) per organism. All the methods were implemented using Python (Van Rossum and Drake, 2011) scripts, and the Uberon ontology was downloaded from: <http://uberon.github.io/downloads.html> (01/10/2018).

After obtaining a gene similarity matrix, the final step of the anatomy-based gene network construction was to connect each pair of genes with an edge. This can be done with or

without applying a cutoff. Without the cutoff, all the gene pairs with a similarity score are retained in the anatomy-based gene network (unfiltered network). With the cutoff, if the pairwise gene similarity score between two genes is higher than the cutoff, an edge will be placed to connect these two genes, otherwise, these two genes will not be connected (filtered network). Suitable cutoffs for gene networks based on different semantic similarity methods were obtained by analyzing their similarity score distributions. Cutoffs were selected to keep the number of interactions/gene pairs approximately similar to that of the STRING PPI networks with the 0.7 cutoff for each model organism. Finally, four filtered anatomy-based gene networks and four unfiltered (without applying the cutoff) networks were generated using the four semantic similarity calculation methods (Lin, Resnik, Schlicker, and Wang) for the mouse and the zebrafish. Both filtered and unfiltered networks were evaluated using the methods explained in section 1.2.7 to compare the performance of different semantic similarity methods for anatomy-based gene network generation for each model organism.

### *1.2.5 Integration of the anatomy-based gene networks with the STRING PPI networks*

The integration was achieved by an accuracy-based weighting method that uses the combined scores for each gene interaction from the STRING PPI networks and the gene similarity scores calculated during the previous section (1.2.4) for the anatomy-based gene networks. This accuracy-based weighting method was used to integrate multiple networks in a previous work (Fraser and Marcotte, 2004), but in this scenario, it was used to integrate only two network types: PPI and anatomy-based gene networks.

Initially, the PPI and anatomy-based gene networks had to be evaluated separately, then accuracy scores were determined for each network type, which were used to decide the weights

for the integration of gene similarity scores. For instance, if the accuracy values for the PPI and anatomy-based gene networks are  $AC_1$  and  $AC_2$  respectively, the weights for the PPI network ( $W_1$ ) and the anatomy-based gene network ( $W_2$ ) are calculated using the equation (1.9) and equation (1.10). In this instance, the accuracy values for each network are equivalent to their area under the curve (AUC) values of the ROC curves generated during the evaluation workflow (explained in section 1.2.7)

$$W_1 = \frac{AC_1}{AC_1 + AC_2} \quad (1.9)$$

$$W_2 = \frac{AC_2}{AC_1 + AC_2} \quad (1.10)$$

Then, the weights are used to calculate the gene similarity scores for the integrated network, based on the gene similarity scores of the original two networks. For instance, consider the similarity between two genes:  $G_a$  and  $G_b$ . If the similarity scores from the PPI network and the anatomy-based gene network for those two genes are given by  $sim_1(G_a, G_b)$  and  $sim_2(G_a, G_b)$ , respectively, the similarity score in the integrated network:  $sim_3(G_a, G_b)$  is calculated by the equation (1.11) below.

$$sim_3(G_a, G_b) = W_1 sim_1(G_a, G_b) + W_2 sim_2(G_a, G_b) \quad (1.11)$$

If an interaction is not found in an original network, the similarity score will be zero for that network; for instance, if  $G_a$  and  $G_b$  are not interacting in the PPI network,  $sim_1(G_a, G_b)$  will be zero. During the calculation of weights, the most accurate original network will obtain a higher weight, and the integrated network will be weighted towards the more accurate network. This method was used to integrate PPI and anatomy-based gene networks for the mouse and the zebrafish. The integration was performed separately for the different anatomy-based gene networks for the four different semantic similarity calculation methods (Lin, Resnik, Schlicker,



and Wang). Eventually, four integrated networks were generated for each model organism. Then a gene similarity score cutoff was applied as explained in section 1.2.4 to keep only the high-quality interactions. As proposed by the hypothesis, the integrated networks are expected to be more reliable and accurate because applying a proper cutoff will only keep the interactions that are common to both the original networks. Therefore, false positive interactions in the STRING PPI network will be filtered out if they are not found in the anatomy-based gene network and false negative interactions in the STRING PPI networks will be added to the integrated network only if they have a very high similarity score in the anatomy-based gene network.

#### *1.2.6 Network-based candidate gene prediction*

To evaluate the performance of the networks, a network-based candidate gene prediction algorithm must be implemented. Majority Voting (MV) is a frequently used algorithm where a function is assigned to a gene, based on counting the number of genes with that function in its immediate neighborhood (Schwikowski, et al., 2000). Although this method is widespread, it only works well for functions/phenotypes with a higher number of gene annotations. A function/phenotype with fewer annotations will always get low prediction scores because it is not frequently found in the network (Sharan, et al., 2007; Zeng, et al., 2012). To solve this issue, a chi-square based method was introduced (Hishigaki, et al., 2001), which considers the expected frequency for each function/phenotype and calculates the difference between the expected and observed number of genes for that function in the neighborhood. This method reduces the bias towards frequent functions/phenotypes because the expected frequency mitigates the disadvantage for functions/phenotypes with a low number of gene annotations (Sharan, et al., 2007). Therefore, the Hishigaki method was selected for the candidate gene prediction.

When using the Hishigaki method, one function/phenotype is considered at a time, and for each unknown gene without the function/phenotype, a prediction score is calculated according to the equation (1.12) below.

$$prediction\ score = \frac{(n_{f(u)} - e_f)^2}{e_f} \quad (1.12)$$

In the equation (1.12),  $n_{f(u)}$  denotes the number of genes with the considered function/phenotype ( $f$ ) in the neighborhood of the gene in interest ( $u$ ). Generally, the length of the neighborhood can be defined by the user but the immediate neighborhood (a length of one edge from the gene  $u$ ) is shown to yield better results (Hishigaki, et al., 2001); therefore, only the immediate neighborhood of a gene was considered for predictions. The expected frequency for the function/phenotype is given by  $e_f$ , which is calculated according to the equation (1.13) below.

$$e_f = \frac{tot_f * n(u)}{tot_N} \quad (1.13)$$

In the equation (1.13),  $tot_f$  denotes the total number of genes annotated with the given function/phenotype ( $f$ ) in the network and  $tot_N$  indicates the total number of genes in the network. The total number of genes in the immediate neighborhood of the gene of interest ( $u$ ) is denoted by  $n(u)$ .

The conventional Hishigaki method does not consider the interaction weights/gene similarity scores between gene pairs; hence; it is usually implemented on networks that are filtered with a similarity score cutoff. To evaluate the full/unfiltered gene networks, the Hishigaki method was modified to consider the interaction weights/gene similarity scores (denoted as weighted-Hishigaki). In the modified version, instead of directly counting the number of genes in the neighborhood, the interaction weights for each neighborhood gene are

summed. The updated equations (1.14 and 1.15) for the weighted-Hishigaki method are given below.

$$prediction\ score = \frac{(\sum_{v \in n_f(u)} sim(v,u) - e_f)^2}{e_f} \quad (1.14)$$

$$e_f = \frac{tot_f * \sum_{v \in n(u)} sim(v,u)}{tot_N} \quad (1.15)$$

In the equation 1.14,  $n_{f(u)}$  denotes the genes with the considered function/phenotype ( $f$ ) in the neighborhood of the gene in interest ( $u$ ) and  $v$  iterates through those neighbors that are annotated with the given function/phenotype ( $f$ ). In the equation 1.15,  $v$  iterates through all the neighbors of gene  $u$ . In both the equations,  $sim(v,u)$  represents the interaction weight/gene similarity score for the interaction between genes  $v$  and  $u$ .

To test the performance of the different network prediction methods, Hishigaki method, weighted-Hishigaki method, Majority Voting, and weighted-majority method were implemented on the zebrafish STRING PPI network filtered with 0.7 score cutoff (refer to section 1.2.1 for details) and evaluated with the workflow explained in section 1.2.7.

During the evaluations, Hishigaki method was used on filtered gene networks and weighted-Hishigaki method was used to compare the full/unfiltered networks. When using the unfiltered networks, weighted-Hishigaki method must be used to consider the different weights of the network because they contain all the interactions with a gene similarity score (even low-quality interactions). Furthermore, when comparing different networks during the evaluation, the prediction scores were normalized using the min-max normalization (Han, et al., 2011; Witten, et al., 2016) to facilitate the evaluation. The equation (1.16) for the min-max normalization, which

transforms a value  $v$  of a numeric attribute  $A$  to  $v'$  is given below. The minimum value and the maximum value of the attribute  $A$  is represented by  $min_A$  and  $max_A$ , respectively.

$$v' = \frac{v - min_A}{max_A - min_A} \quad (1.16)$$

### 1.2.7 Evaluation of the network performance

Evaluation is an important step in network-based candidate gene prediction because it allows assessing the performance of a network or the prediction algorithm. In this work, the focus is on evaluating the performance of different networks (integrated *versus* PPI). The filtered anatomical profiles for the mouse and the zebrafish that were reconciled during the network reconciliation step (section 1.2.3) were used for the evaluation because all the genes in these profiles can be matched to all the network types (anatomy-based gene networks, PPI, and integrated networks). Furthermore, these profiles only contain Uberon terms that have at least 10 gene annotations, which reduces the problems caused by evaluating terms with low number of gene annotations.

Usually, during the evaluation, the genes in the profiles are separated into two groups: the training and the validation set, where the functions/phenotypes of the training set are considered as known and the functions/phenotypes of the validation set are treated as unknown. Then, the model built from the training set is used to predict the functions/phenotypes of the validation set, and a suitable evaluation metric (e.g., accuracy, precision, receiver operating characteristic curve, etc.) is used to assess the proportion of genes whose functions/phenotypes are correctly predicted (Han, et al., 2011; Witten, et al., 2016).

There are different methods to separate the dataset into a training and a validation set; the simplest one is separating them according to a pre-defined percentage. For instance, if there are 100 genes in the profile, 60 genes can be included in the training set and 40 genes can be

included in the validation set (Han, et al., 2011). However, in this method, one particular gene would only be placed in either the training set or the validation set, hence, the evaluation is only performed on the genes in the validation set and not performed on genes that were selected to the training set. The leave-one-out cross-validation is another method, where each gene in the profile will be left out as the validation set iteratively at a time, and the rest of the genes will be considered as the training set (Han, et al., 2011; Witten, et al., 2016). For example, if there are 100 genes, one gene will be selected as the validation set and the remaining 99 will be in the training set, which will be used to predict the functions of the gene in the validation set. This process is repeated another 99 times until each gene gets selected as the validation gene. In this method, each gene gets an equal chance to be in the validation and the training set; hence, the bias is removed. Moreover, this method is suited for smaller datasets as the iterative process multiplies the number of times the predictions are made, and each iteration only selects one gene in the validation, leaving enough genes in the training set to be used by the model or the network. Considering these reasons, the leave-one-out cross-validation was selected for the evaluation of the networks.

Different evaluation metrics can be used to assess the predictions. Conventionally, accuracy and error rate (Han, et al., 2011; Witten, et al., 2016) are used as the evaluation metrics, but they are not ideal for unbalanced datasets, such as the gene anatomical profiles, where few genes are annotated to a given anatomical term (positives for that anatomical term) and many genes are not annotated to it (negatives). As an alternative solution, the precision and recall are widely used in network-based candidate gene prediction (Han, et al., 2011). The precision only depends on the number of true positives and the predicted positives for a given anatomical term (equation 1.17); therefore, it is ideal for unbalanced samples because having a large number of

negatives or positives does not affect the final precision value. The recall is also known as the true positive rate (equation 1.18), which depends on the number of true positives as well as the false positives.

$$precision = \frac{\textit{number of true positives}}{\textit{number of true positives} + \textit{number of false positives}} \quad (1.17)$$

$$recall = \frac{\textit{number of true positives}}{\textit{number of true positives} + \textit{number of false negatives}} \quad (1.18)$$

Another option is to use curves, such as the receiver operating characteristic (ROC) curve and the precision-recall curve. These curves are widely used because instead of comparing a single number, such as the precision and the recall, two metrics can be calculated to different prediction score thresholds and plotted in a curve in a pairwise manner. Therefore, the evaluation is performed on a series of prediction score thresholds, which makes the comparison more comprehensive (Han, et al., 2011; Witten, et al., 2016). The most common evaluation curve is the ROC curve, which plots the recall (true positive rate; equation 1.18) to the false positive rate (equation 1.19).

$$false\ positive\ rate = \frac{\textit{number of false positives}}{\textit{number of false positives} + \textit{number of true negatives}} \quad (1.19)$$

The precision-recall curve is another curve type, which can be to evaluate the network-based candidate gene predictions. Because of the inclusion of the precision, it is used in evaluating network-based candidate gene predictions (Zeng, et al., 2012). For this work, both ROC and precision-recall curves were used for the evaluations.

Although a gene has multiple Uberon terms annotated to it in the annotation profile, typically, the leave-one-out cross-validation is applied to a single Uberon term at a time and a single ROC curve or a precision-recall curve is generated for each term. If there are 100 Uberon terms in all the profiles, 100 curves will be generated (Han, et al., 2011). Finally, the distribution

of the area under the curve (AUC) values for the ROC curves can be plotted and compared among different networks or prediction methods to evaluate their performance (Kourmpetis, et al., 2010). This method will be referred to as the ‘single-function evaluation method’ from herein. The single-function evaluation method was used to assess the performance of different networks by comparing the AUC distribution of ROC curves and precision-recall curves using box plots and histograms.

However, it is easier to generate and compare a single ROC curve or a precision-recall curve for each network rather than comparing the distribution of AUC values. To achieve this, the single-function evaluation method was modified to consider multiple terms/phenotypes at a time. This method will be referred to as the ‘multi-function evaluation method’ from herein. During the multi-function evaluation, all the Uberon terms annotated to the gene of interest will be considered at the same time when calculating the true positives and false positives. For instance, if the Hishigaki method was able to predict 5 Uberon terms correctly as compared to the original annotations, the number of true positives will be 5; if it predicted 2 Uberon terms/functions that are not in the original annotation list for the gene, there will be 2 false positives. With the multi-function method, one single ROC or precision-recall curve will be generated for each network, making it convenient for the network comparison. Both the single-function and multi-function evaluation methods were used in this research to increase the robustness of the conclusions. The general workflow for evaluating multiple networks is given in Fig. 1.5.

First, the evaluations were performed on the anatomy-based gene networks (section 1.2.4) to find the best semantic similarity method for the network generation for the mouse and the zebrafish. During the evaluations, the Hishigaki method was used for candidate gene

prediction on filtered networks. Then, the same approach was used on the integrated networks (section 1.2.5) to select the best performing integrated network from Lin, Resnik, Schlicker, and Wang methods for the mouse and the zebrafish.

According to the proposed hypothesis, the integrated networks are expected to outperform the STRING PPI networks. The evaluation workflow (Fig. 1.5) was implemented on STRING PPI networks for the mouse and the zebrafish and compared with the results of the integrated networks for each semantic similarity method separately. Although anatomy-based gene network performance is not included in the hypothesis, their results were compared with the STRING PPI and the integrated networks as well. In summary, for each model organism, a comparison was done using the single and multi-function evaluation methods on the integrated network, the anatomy-based gene network, and the STRING PPI network for each semantic similarity calculation method (Lin, Resnik, Schlicker, and Wang). For example, during the multi-function evaluation, three ROC curves were generated each for the integrated network, the anatomy-based gene network, and the STRING PPI network for each semantic similarity calculation method and their AUC values were compared to find the best performing network. However, to support the hypothesis, the interest was on comparing the AUC values of the PPI vs. the integrated networks. Even if an anatomy-based gene network performed better than an integrated network, it could not be selected for network-based candidate gene prediction in practice because it only contains genes in the anatomical profile. On the other hand, the integrated networks contain not only the genes from the anatomical profiles but also the genes from the STRING PPI networks; therefore, they are more practical for predicting new candidates for unknown phenotypes. Again, Hishigaki method was used on filtered networks and weighted-



Hishigaki method was used on unfiltered networks to evaluate whether filtering a network would impact the results.

### *1.2.8 Validating the evaluation results*

It is important to understand the biological significance of the evaluation results. For instance, during the previous step, if the integrated networks performed better than the STRING PPI networks, it must be confirmed that the increased performance is due to the biological significance of integrating experimental phenotype data *via* anatomy ontology and not due to random error/noise. Furthermore, the same anatomical profiles from mouse and zebrafish that were used for the evaluation of the networks were used during the construction of the anatomy-based gene networks (section 1.2.4) as well. Therefore, an argument can be made that the increase in the performance of the integrated networks is merely due to using the same anatomical profiles for both network construction and evaluation, which can be considered as a circular use of the anatomical profiles. To test these issues, further experiments were required to validate the biological significance of the evaluations. Due to the computational intensity and time constraints, the anatomy-based gene network and the integrated network for the Wang method from zebrafish were selected based on their better performance during the evaluations (section 1.3.6).

First, to confirm whether the increase in the performance of the anatomy-based gene network and the integrated networks is purely due to the random error, they can be compared with fully random networks. Here, a fully random network is defined as a network of the same size in terms of the number of nodes and the interactions as the original network (Hishigaki, et al., 2001). Using a Python script, fully random networks were generated for the zebrafish

integrated network and the anatomy-based gene network for the Wang method. The number of genes and the number of interactions of the fully random networks were the same for each network type; only the arrangement of the interactions was randomized. Furthermore, another type of random network was generated by only randomizing the anatomical profiles. Here, the zebrafish anatomical profile, which was used to generate the anatomy-based gene network and the integrated network, was randomized by randomly assigning Uberon terms to each gene to match the original number of annotations. Then, the random profiles were used to create a random anatomy-based gene network using the Wang method, and it was integrated with the zebrafish PPI network to generate the random integrated network. The second method is only a partial randomization because only the profiles were randomized, and the number of genes and the interactions are different from the original networks. From herein, the first random network type will be referred to as ‘fully random networks’, and the second type will be referred to as ‘random profile networks’.

The evaluation workflow (Fig. 1.5) was implemented on the fully random anatomy-based gene network and the random profile anatomy-based gene network, and the results were compared with the original zebrafish anatomy-based gene network for the Wang method. This was repeated for fully random and random profile integrated networks as well. If there is a biological significance in the original anatomy-based gene network and the integrated network, the original networks should outperform the random networks.

To evaluate the effect of the circular use of the zebrafish anatomical profiles, 30 Uberon terms were randomly removed from the reconciled zebrafish anatomical profile file, and an anatomy-based gene network was generated using the remaining Uberon terms in the profile using the Wang method. Then, the generated anatomy-based gene network was integrated with

the zebrafish PPI network and the networks were evaluated using a new anatomical profile file that only contains the removed 30 Uberon terms and their gene annotations. The zebrafish PPI network was also evaluated using the same new anatomical profile, and its performance was compared with the anatomy-based gene network and the integrated network. Because the 30 Uberon terms were removed during the network generation, if the anatomy-based gene network and the integrated network outperformed the PPI network, there is no effect from the circular use of anatomical terms in the profile for the evaluation. If the performance increase in the integrated network is only due to re-using the same Uberon terms for the evaluation, the integrated network should not show a performance increase compared to the PPI network when those 30 Uberon terms were used for the evaluation because they were not involved in the network construction.

Another experiment to measure if there is an effect from the circular use of the anatomical profiles is not to use anatomical profiles for evaluation at all. In this scenario, the Gene Ontology (GO) profiles can be used for the evaluation. The zebrafish GO annotations were downloaded from the GO consortium downloads page (<http://www.geneontology.org/page/download-go-annotations>) and they were pre-processed to keep only the GO-BP annotations (Ashburner, et al., 2000). Then, GO profiles were constructed for the zebrafish using the GO-BP annotations. These profiles were reconciled using the method explained in section 1.2.3, and genes that are not in the zebrafish networks (PPI, anatomy-based gene network, and integrated network) for the Wang method were removed from the profiles. Then, the reconciled GO profile file was used to evaluate the aforementioned three networks using the evaluation workflow (Fig. 1.5). Here, GO profiles for the zebrafish were not used to construct the anatomy-based gene network, the integrated network, and the PPI network; if the integrated network outperformed the STRING PPI network, it indicates that the integrated

network is higher in quality irrespective of the gene profiles (Uberon anatomy profiles or GO-BP profiles) used for the evaluation.

This integrative framework was developed in Python programming language (Van Rossum and Drake, 2011) and the scripts are available in a GitHub repository, which can be accessed using the following link:

[https://github.com/pasanfernando/Integrative\\_network\\_analysis.git](https://github.com/pasanfernando/Integrative_network_analysis.git).

## **1.3 Results**

### *1.3.1 The STRING PPI networks*

The zebrafish STRING PPI network contained 23,018 genes and 12,558,675 interactions; the distribution of the gene similarity scores/combined scores of the PPI network is shown in Fig. 1.6.a. The mouse STRING PPI network contained 21,052 genes and 6,262,253 interactions of which the distribution is shown in Fig. 1.6.b. Most of the gene similarity scores are distributed between 0.1 to 0.3 region. Further, there is a slightly higher distribution between 0.9 and 1 region, especially for the mouse PPI network.

After applying the 0.7 gene similarity score cutoff to keep only the high-quality interactions, most of the interactions were removed from the networks. The filtered zebrafish PPI network contained 14,677 genes and 501,704 interactions, and the filtered mouse PPI network contained 13,866 genes and 414,667 interactions.

### *1.3.2 Construction of anatomical profiles and reconciliation of the gene names*

The original zebrafish anatomical profiles file retrieved from the Monarch Initiative repository contained 5,405 genes annotated to 960 Uberon anatomical terms, and the mouse

profiles file contained 14,652 genes annotated to 1,537 Uberon terms (Table 1.1). Not all of these genes were found in the STRING PPI networks, owing to differences in the data sources. After implementing the gene reconciliation algorithm that contained three rounds (direct name matching, Ensembl id matching, and gene synonym matching), the number of original matches has increased from 2,527 (direct name matching) to 3,048 for the zebrafish and from 8,166 to 8,607 for the mouse. The detailed reconciliation statistics are shown in Table 1.1.

The extra 521 genes for the zebrafish and 441 genes for the mouse that were matched during round 2 (using Ensembl ids) and round 3 (using gene synonyms) contained outdated gene names in the PPI networks. Therefore, they were updated to the correct names that were used in the anatomical profiles. After the reconciliation, the original anatomical profiles were filtered to contain only the genes that were matched with the PPI networks and to contain only the Uberon terms that have at least 10 gene annotations (Table 1.1). These reconciled and filtered anatomical profiles were used during the evaluation of different network types because it is important to evaluate the networks using the genes that are found in all the three types of networks (PPI, anatomy-based gene networks, and integrated networks) for the zebrafish and the mouse.

### *1.3.3 Generation of the anatomy-based gene networks*

When constructing the anatomy-based gene networks, original anatomical profiles (before the reconciliation) were used to retain all of their genes in the networks. The reconciled anatomical profiles were only used for the evaluation of the networks. The gene similarity score distributions for the four types of unfiltered anatomy-based gene networks (Lin, Resnik, Schlicker, and Wang) for the zebrafish and the mouse are shown in Figs. 1.7 and 1.8, respectively. The gene similarity scores for Lin, Resnik, and Schlicker methods are distributed

between a range from 0 to 0.40. In contrast, the distribution for the Wang method is symmetrical around the 0.50 region.

Obtaining these distributions are critical to determine the score cutoff applied to each network. For example, applying 0.7 as the cutoff for the Wang anatomy-based gene network for the zebrafish, generated a filtered network with 5,386 genes and 789,282 interactions; if the same 0.7 cutoff was applied to the zebrafish Resnik network, the filtered network will only have 30 genes and 31 interactions. If these two networks were evaluated, the changes in the number of genes and the number of interactions would have a significant effect on their performance. Therefore, a cutoff must be applied to keep the network size relatively constant among the different networks. However, it is practically difficult to apply cutoffs to keep the exact number of genes and the interactions among the networks. Therefore, using the trial and error method, different cutoffs were applied to anatomy-based gene networks to keep the number of interactions between 500,000 and 750,000. The statistics for the network sizes of filtered and unfiltered networks and their cutoffs are shown in Tables 1.2 and 1.3 for the zebrafish and the mouse, respectively. The unfiltered mouse anatomy-based gene networks are significantly larger than the zebrafish networks; for instance, the zebrafish anatomy-based gene network for the Wang method has 14,604,258 interactions, whereas its mouse counterpart has 107,324,905 interactions. This is due to the difference between number of genes in the original anatomical profiles for the zebrafish and the mouse. The mouse profile has 14,652 genes compared to the 5,405 genes in the zebrafish profile (Table 1.1). When calculating the gene similarity scores during the anatomy-based network generation, the number of pairwise comparisons is significantly high in the mouse compared to the zebrafish, which causes the higher number of interactions in the mouse anatomy-based gene networks.

#### *1.3.4 Integration of the anatomy-based gene networks with the STRING PPI networks*

During the integration, unfiltered anatomy-based gene networks for the zebrafish and the mouse were integrated with the corresponding STRING PPI networks. When selecting the cutoffs for filtering the integrated networks, their gene similarity score distributions were considered as explained before (section 1.3.3). The statistics for the filtered and unfiltered network sizes are shown in Table 1.4 for the zebrafish and Table 1.5 for the mouse. The generated integrated networks are larger than the anatomy-based gene networks in terms of the number of genes and the interactions. For instance, the Wang anatomy-based gene network for the zebrafish has 5,405 genes and 14,604,258 interactions (Table 1.2), whereas the zebrafish integrated network for the Wang method has 25,375 genes and 26,821,274 interactions (Table 1.4). During the integration, the 5,405 genes in the anatomy-based gene network were integrated with the 23,018 genes in the zebrafish PPI network, which caused an increase in the network size. The common genes and interactions were retained according to the integration formula (equation 1.11), and also the genes and the interactions that are unique to one network were included in the integrated network. Therefore, the integrated networks are more complete in terms of the number of genes and the information contained.

The gene similarity score distributions for the integrated networks for the zebrafish and the mouse are shown in Figs. 1.9 and 1.10, respectively. When compared to the distributions of the corresponding anatomy-based gene networks as shown in Figs. 1.7 and 1.8, the distributions of the integrated networks are slightly skewed to the right; especially, the gene similarity scores of the Wang anatomy-based gene networks were symmetrical and distributed around 0.5; in contrast, the distributions for the Wang integrated networks are shifted to 0-0.50 region. This is

due to the effect of the integration. Only the interactions that have high similarity scores in the anatomy-based gene network and the PPI network receive higher scores in the integrated network. Most of the interactions in the anatomy-based gene network received low support from the PPI network, thus the gene similarity score distribution of the integrated network is skewed to right. By applying the cutoffs as shown in Tables 1.4 and 1.5, the interactions with the highest similarity scores, which received support from both the PPI and the anatomy-based gene networks, could be selected.

### *1.3.5 Network-based candidate gene prediction*

The performance of the Hishigaki prediction method, weighted-Hishigaki prediction method, Majority Voting (MV) method, and weighted-Majority Voting method was evaluated using the zebrafish STRING PPI filtered network. For the multi-function evaluation method, the ROC curve comparison and the precision-recall curve comparison are shown in Figs. 1.11.a and 1.11.b, respectively. For the single-function evaluation method, the comparison of the AUC distributions of ROC curves and the precision-recall curves is shown in Supplementary Fig. S1.1. According to all the figures, the performances of different network-based candidate gene prediction methods are approximately similar, except for the precision-recall curve comparison (Fig 1.11.b). All other comparisons indicate that the Hishigaki and weighted Hishigaki methods slightly outperform the Majority Voting and weighted Majority Voting methods. Even in the precision-recall curve comparisons (Fig. 1.11.b), the weighted-Hishigaki method outperforms other methods in the initial region. Considering these results and the advantages discussed in section 1.2.6, the Hishigaki method and the weighted-Hishigaki method were selected as the network-based candidate gene prediction methods. The Hishigaki method was used for the



filtered networks and the weighted-Hishigaki method was used for the unfiltered/full networks because it considers the interaction weights of the unfiltered networks.

### *1.3.6 Evaluation of the network performances*

First, the anatomy-based gene networks were evaluated to compare the performance of different semantic similarity calculation methods (Lin, Resnik, Schlicker, and Wang) used during the network generation. For the zebrafish, the comparison of the ROC curves and the precision-recall curves generated using the multi-function evaluation method for the different filtered anatomy-based gene networks are shown in Figs. 1.12.a and 1.12.b, respectively. For the mouse, the comparison of the ROC curves and the precision-recall curves for the different filtered anatomy-based gene networks are shown in Figs. 1.13.a and 1.13.b, respectively. For the single-function evaluation method, the AUC distribution comparisons for the zebrafish and the mouse are shown in Supplementary Figs. S1.2 and S1.3, respectively. According to these results, all four semantic similarity calculation methods have very similar performance, but the anatomy-based gene networks generated from the Wang method slightly outperformed other methods, especially in precision-recall curve comparisons for the zebrafish (Fig. 1.12.b) and the mouse (Fig. 1.13.b).

The integrated networks for the four semantic similarity methods (Lin, Resnik, Schlicker, and Wang) were compared separately to select the best integrated network for the zebrafish and the mouse. For the zebrafish, the comparison of the ROC curves and the precision-recall curves generated using the multi-function evaluation method for the different filtered integrated networks are shown in Figs. 1.14.a and 1.14.b, and for the mouse, the comparison of the ROC curves and the precision-recall curves are shown in Figs. 1.15.a and 1.15.b, respectively. For the

single-function evaluation method, the AUC distribution comparisons for the zebrafish and the mouse are shown in Supplementary Figs. S1.4 and S1.5, respectively. For the zebrafish, the Lin method has the best AUC for both the curve types and the Wang method comes as a close second. These results are confirmed by the AUC distribution comparisons (Supplementary Fig. S1.4). For the mouse, the performance of the four integrated networks are very similar, but the Schlicker method slightly outperforms the other three methods (Figs. 1.15.a and 1.15.b).

According to the proposed hypothesis, the integrated networks are expected to perform better than the PPI networks. To test this hypothesis, the comparisons of ROC and precision-recall curves for the filtered integrated networks, the anatomy-based gene networks, and the PPI networks for the zebrafish are shown in Figs. 1.16 and 1.17, respectively. In each figure, four comparisons are shown for the four semantic similarity calculation methods. The comparisons for ROC and precision-recall curves for the mouse are shown in Figs. 1.18 and 1.19, respectively. According to these results, the AUC values of the curves for the integrated networks (green) are significantly higher than the PPI networks. This indicates that the filtered integrated networks for all the semantic similarity calculation methods have higher candidate gene prediction accuracy than the corresponding PPI networks in both the model organisms. Although the anatomy-based gene networks (blue curves) have the highest AUC values, the integrated networks have similar AUC values in most comparisons; in some instances, such as the zebrafish ROC curve comparison for the Lin method (Fig. 1.16.a), the integrated network (green curve) even outperforms the anatomy-based gene network (blue curve). For the proposed hypothesis, only the integrated networks are compared with the PPI networks. The anatomy-based gene networks are not practical for candidate gene predictions because they only contain a limited number of proteins/genes with known Uberon annotations. The integrated networks have

unknown genes coming from the PPI networks, which make them better networks for gene predictions.

Above results are confirmed once again by the results of the single-function evaluation method for the zebrafish (Supplementary Figs. S1.6, S1.7, S1.8, and S1.9) and the mouse (Supplementary Figs. S1.10, S1.11, S1.12, and S1.13). The boxplot and histogram distributions for the AUC values of ROC and precision-recall curves for the filtered integrated networks are higher than the filtered PPI networks for both the model organisms. This indicates that after the integration, even when the networks are evaluated for each anatomical term separately, the performance is generally better than the original PPI networks.

Furthermore, the ROC (Supplementary Fig. S1.14) and precision-recall curve (Supplementary Fig. S1.15) comparisons for the zebrafish and the ROC (Supplementary Fig. S1.16) and precision-recall curve (Supplementary Fig. S1.17) comparisons for the mouse unfiltered networks that were evaluated using the multi-function evaluation method also show that the integrated networks have better candidate gene prediction performance than the PPI networks. This is further confirmed by the single-function evaluation method results for the same unfiltered networks for the zebrafish (Supplementary Figs. S1.18, S1.19, S1.20, and S1.21) and the mouse (Supplementary Figs. S1.22, S1.23, S1.24, and S1.25). As can be seen, even in the unfiltered networks, integrated networks for the four semantic similarity methods have a better candidate gene prediction than the PPI networks. This proves that filtering the networks by a gene similarity score cutoff does not affect the final conclusions.

### *1.3.7 Validating the evaluation results*

The ROC and precision-recall curve comparisons for the non-randomized anatomy-based gene network and the integrated network for the Wang method with their fully random and random profile counterparts are shown in Fig. 1.20. For the single-function evaluation method, the comparisons of the AUC distributions for the same networks are shown in Supplementary Figs. S1.26 and S1.27. According to the comparisons, non-randomized networks (blue) have a higher performance compared to the randomized networks in both the network types. When comparing the two randomization methods, random profile networks (green), which were constructed by only randomizing the anatomy profiles have a higher performance than the fully random networks (red). This comparison with randomized networks indicates that the high performance observed by the anatomy-based gene network and the integrated network is due to their biological significance, not merely due to random error.

To test the effect of the circular use of the same anatomy profiles for anatomy-based gene network construction and for the evaluation of those networks, 30 Uberon terms with at least 10 gene annotations were randomly removed from the zebrafish anatomy profile, which was used for anatomy-based gene network and integrated network construction using the Wang method. Then, those networks were evaluated using only the removed 30 Uberon terms, and their performance was compared with the filtered zebrafish STRING PPI network. The ROC and precision-recall curve comparisons are shown in Figs. 1.21.a and 1.21.b, respectively. The comparison of the AUC distributions for the single-function evaluation method is shown in Supplementary Fig. S1.28. According to the comparisons, the integrated network (green) performs better than the PPI network (red), even when being evaluated using the 30 Uberon terms that were not used for the network construction. Here, the integrated network (green) even

has a better performance than the anatomy-based gene network (blue). This proves that using the same anatomy profile for network construction and evaluation does not affect the final conclusion of the integrated network having a better performance than the PPI network.

To prove this point further, the filtered PPI network, anatomy-based gene network, and integrated network for the zebrafish were evaluated using GO-BP profiles for the zebrafish. Because the anatomy-based gene network and the integrated network were constructed using the zebrafish anatomy profile using the Wang method, the GO-BP annotations, which were used for the evaluation, do not have a direct influence on the network construction. The ROC and precision-recall curve comparisons for this evaluation are shown in Figs. 1.22.a and 1.22.b, respectively, and the comparison of the AUC distributions for the single-function evaluation method is shown in Supplementary Fig. S1.29. According to the comparisons, the integrated network (green) performs better than both the PPI (red) and anatomy-based gene (blue) networks, even when evaluated by the GO-BP profiles. Therefore, irrespective of the profile type (anatomy or GO-BP) integration has improved the quality of the network.

## **1.4 Discussion**

The goal of this work was to test whether the integration of anatomy ontology data with PPI networks enhances the network-based candidate gene prediction accuracy. The PPI networks are used to predict gene candidates for certain phenotypes or functions, but they suffer from low prediction accuracy due to the low quality of the networks. One way to increase the quality of the PPI networks is to incorporate experimental knowledge gathered in the biological literature about gene-phenotype relationships. Here, using biological ontologies and semantic similarity calculation methods, a computational framework was developed to integrate experimental

knowledge about gene-phenotype associations with PPI interaction data. First, anatomy-based gene networks were constructed by calculating pairwise similarity scores between genes using the semantic similarity between respective anatomical profiles, which represents a network of genes based on their Uberon annotations. In this network, if two genes receive a high gene similarity score, these genes are regulating similar anatomical phenotypes. Next, the anatomy-based gene networks for the mouse and the zebrafish were integrated with the corresponding PPI networks downloaded from the STRING database. In the integrated networks, the gene interactions that receive a higher support from both the PPI and anatomy-based gene networks will receive higher similarity scores. This filters out false positive interactions in the PPI networks if those interactions are not supported by the anatomy-based gene networks, that is, if they receive low scores from the anatomy-based gene networks. On the other hand, gene interactions with low scores in the PPI networks will be enhanced if the interactions are supported by the anatomy-based gene networks, that is, if they have high similarity scores from the anatomy-based gene networks. In cases where the gene similarity score is zero in anatomy-based gene networks due to lack of anatomical term annotations, the support from the PPI network should be extremely high for those interactions to receive a moderate similarity score in the integrated networks and potentially be selected after application of the gene similarity cutoff. Furthermore, if two proteins are not interacting in a PPI network, they need a higher support from the anatomy-based gene network to receive a moderate gene similarity score in the integrated network and to be selected after the application of the gene similarity score cutoff. Finally, a hypothesis was formulated that the integrated networks should have a higher candidate prediction accuracy than the PPI networks because of the increased quality of the network interactions.

According to the network performance evaluations, the integrated networks performed better than the PPI networks, proving that the integration has increased the quality of the networks. The evaluations were performed under different settings to improve the robustness of the final conclusion, which yielded the same result. For example, both ROC curves and precision-recall curves were used during the evaluation. Furthermore, the conventional single-function method where curves are generated for each function at a time and a modified multi-function evaluation method where one curve is generated for the entire network were used for the evaluation. These evaluation methods were tested on anatomy-based gene networks and integrated networks constructed using four semantic similarity methods (Lin, Resnik, Schlicker, and Wang) for two model organisms (zebrafish and mouse). Usually, biological networks are filtered by a score threshold before their usage, but for this evaluation, both filtered and unfiltered networks were used to confirm that the filtering of the networks does not affect the consistency of the results. Under a number of various experimental settings as explained above, the integrated networks performed better than the PPI networks, strengthening the conclusion that the integration of the literature knowledge using anatomy ontology increases the candidate gene prediction accuracy of the PPI networks.

To test the biological significance of the results, the integrated and the anatomy-based gene networks for the Wang method for the zebrafish were compared with the randomized versions of these networks (Fig. 1.20). The results demonstrated that the higher candidate gene prediction performance observed in the integrated network has a biological significance and it was not due to random error. Out of the two randomization procedures used, the random profile networks (green) that were generated by only randomizing the anatomical profiles have a better performance than the fully random networks (red) that were constructed by completely

randomizing the entire networks. The AUC values for the ROC curves of fully random networks are closer to 0.5, which is expected in a randomized prediction (Han, et al., 2011). When only the anatomical profiles are randomized, the original number of annotations per gene was kept constant even after the randomization, which may lower the randomization effect by including closely related Uberon terms for the same gene. This may explain their higher performance compared to the fully random networks. However, the non-randomized networks (blue) have a significantly higher performance than the two randomized network types, especially, in precision-recall curves, due to their biological significance.

Another challenge faced during the analysis is the concern of circular use of the same anatomical profiles for the construction of the networks (anatomy-based gene networks and then the integrated networks) and for their evaluation. The increased performance observed in the integrated networks may be due to using the same anatomical profiles for the evaluation. Two experiments were conducted to assess whether the circular use of the anatomical profiles affects the observed results. First, 30 Uberon terms were randomly selected and their annotations were removed from the zebrafish anatomy profile and the networks constructed using the Wang method were evaluated using the same 30 terms (Fig. 1.21). Second, the networks were evaluated using the GO-BP profiles for the zebrafish (Fig. 1.22). In both experiments, the annotations used for the network construction were not used for the evaluation, and in both the occasions the integrated network (green) outperformed the PPI network (red). This indicates the increased performance observed in the integrated networks is not due to the circular use of the same anatomical profiles.

The anatomy-based gene networks that were constructed by calculating the pairwise semantic similarity between genes are used to generate the integrated networks in the zebrafish



and the mouse. According to the network performance evaluations (Fig. 1.16, Fig. 1.17, Fig. 1.18, and Fig. 1.19), the anatomy-based gene networks (blue) performed slightly better than the integrated networks (green). However, the anatomy-based gene networks contain a limited number of genes that only have Uberon annotations, thus they are not practical for candidate gene prediction as the integrated networks. For example, the zebrafish filtered anatomy-based gene network constructed using the Schlicker method contains 5,386 genes (Table 1.2), whereas the corresponding integrated network contains 12,755 genes (Table 1.3). Furthermore, the anatomy-based gene networks represent the gene organization in the network based only on their phenotypic annotations and do not include the molecular interactions coming from experimental sources, which is achieved after the integration. During the integration, the anatomy-based gene networks were populated with new genes coming from the PPI networks, which included some unknown genes. Even with the highly populated network structure, the integrated networks outperformed the PPI networks. In some cases, such as the zebrafish networks for the Lin method (Fig. 1.16.a), the integrated network even outperforms the anatomy-based gene network.

Moreover, the extra validation experiments conducted to check the concern of the circular use of the same anatomy profiles (Fig. 1.21 and Fig. 1.22) show that the integrated network performs better than the anatomy-based gene network when the same anatomical profile was not used for the construction of the networks and for the evaluation. When the GO-BP profiles were used for the evaluation (Fig. 1.22), the integrated network (green) comfortably outperforms the anatomy-based gene network. This indicates that there is some bias in using the same anatomical profiles for the construction and the evaluation of the anatomy-based gene networks, which may have affected their slightly higher performance than the integrated networks. However, the integration process reduces this bias by including the interactions from

the PPI networks. Because the PPI networks are mainly based on empirical data about direct PPI interactions, the interactions that receive higher support in the integrated networks incorporates information from these empirical sources, such as mass spectrometry, which was not captured by anatomy-based gene networks. This may be a potential reason for the higher performance of the integrated networks when the networks were evaluated using the Uberon or GO terms that were not included for the network construction. Therefore, the integrated networks are suited for candidate gene prediction, although the anatomy-based gene networks have a slightly higher candidate gene prediction in the original evaluations. Anatomy-based gene networks were only constructed as an intermediate step for the construction of the integrated networks and the goal was to investigate the effect of the integration of the anatomical annotations with the PPI networks. The anatomy-based gene networks only contain the genes with anatomical annotations, hence, they will miss a large portion of unknown genes (added by the PPI networks) if they are used for candidate gene prediction. Also, the anatomy-based gene networks do not include the physical interactions of the proteins coming from empirical sources. The interaction scores in the integrated networks captures information from both PPI networks and anatomy-based gene networks. Because of these reasons, the integrated networks are better suited for discovering new gene candidates than the anatomy-based gene networks.

This research involves the integration of large-scale datasets including the PPI data in the STRING database and the anatomical annotations retrieved from the Monarch Initiative repository. Integrating such data in large quantities has been a challenge, which requires computational methods to facilitate the integrations. During the integration, matching gene identities from the two data sources (STRING and Monarch Initiative repository) was a challenge. For example, some of the gene names were outdated in the STRING database.

Furthermore, there were several synonyms for some of the genes, which can be challenging when using only the gene name matching. Therefore, as explained in the Methods (section 1.2.3), a gene name reconciliation algorithm was developed to first, match the genes using the gene name/symbol, then, using their Ensembl ids, and finally, using synonyms matching. This improved the number of matchings and provided an efficient integration (Table 1.1). Integrating data from different data sources still remains a challenge; although data repositories such as Monarch Initiative have been developed, this requires special attention and development of bioinformatics algorithms and tools to achieve a better integration.

Large-scale integration also introduces computational challenges, especially, when working with biological networks. Usually, the network analysis algorithms are computationally intensive and require specialized modifications depending on the problem for a better implementation (Cormen, 2009). This work required the construction of large networks (anatomy-based gene networks and integrated networks), which is a computationally challenging task, especially, for a model organism such as the mouse with a large number of annotations. Furthermore, there is a lack of developed bioinformatics tools and codes to address large-scale network integration problems. Software packages, such as Cytoscape (Smoot, et al., 2011) are valuable for network visualization and general analysis tasks but are not suitable for specific problems, such as network integration. Moreover, built-in codes and libraries for semantic similarity calculations are not readily available for the Python programming language; there are several packages, such as Owl Sim in Java (Horridge and Bechhofer, 2011; Washington, et al., 2009) and GoSemSim in R (Yu, 2010) that can load ontologies in Owl format and perform basic semantic similarity calculations, but they lack the implementation of specified methods, such as the Wang method. Furthermore, GoSemSim is only focused on the GO and could not be used on

the Uberon anatomy ontology. The existing libraries implemented on Java and R were not suitable for large-scale network construction and integration. Therefore, all the scripts that were required for this work were newly developed in Python focusing on efficient running time and memory usage. These scripts can be generally implemented on any ontology, including GO, Uberon, etc., which could be very useful for the community interested in biological ontologies and networks.

For this research, the proposed integrative method was implemented using anatomy ontology data because the focus was to study changes in anatomical phenotypes associated with evolutionary changes. The methodology used for the integration can be generally applied to any type of ontology and network data given the nature of the biological question. For instance, if the focus is to study changes in the molecular function, the molecular function component of the GO (Ashburner, et al., 2000; Consortium, 2004) can be used instead of the Uberon anatomy ontology. Furthermore, this approach can be extended for any model organism if the PPI network and ontology data are available, although this work focuses on only the mouse and the zebrafish. For instance, this work can be extended to human anatomical entities using Uberon as it is a species-independent ontology. To facilitate the usability, the scripts used for the integrated framework will be organized into a Python package, which can be easily used with any ontology or a domain.

An interesting future challenge would be to include quality terms using Phenotype And Trait Ontology (PATO) to analyze the genes associated with certain qualities of anatomical entities, such as the size or the presence and absence of an anatomical entity (Gkoutos, et al., 2009). For this task, a computational framework must be established to include composite entity-quality terms. Alternatively, phenotype ontologies, which already include quality of an

anatomical entity, such as the Human Phenotype Ontology (Köhler, et al., 2016) and the Mammalian Phenotype Ontology (Smith and Eppig, 2012) can be directly used for the integrative framework. However, according to Manda, et al. (2016), incorporation of entity-quality terms does not significantly increase the amount of semantic information captured from phenotype profiles based on taxon-phenotype annotations. Nonetheless, incorporation of additional phenotype information into gene networks, which was not tested before, may increase the performance and is a potential method to improve the network performance. This integration of quality terms for anatomical structures may also help to identify condition-specific gene interaction changes. Currently, large-scale PPI networks retrieved from databases such as STRING represent all possible interactions that occur within the cell for an organism in any condition type. However, the protein interactions may differ depending on the condition or the phenotype (e.g., pectoral fin presence *versus* absence, a human disease presence *versus* absence, etc.) (Creixell, et al., 2015; Georgii, et al., 2009; Greene, et al., 2015). Presently, the best method to compare the protein interaction changes associated with specific conditions and phenotypes is to generate experimental protein interactions through methods such as mass spectrometry for each condition or the phenotype, which is time-consuming and expensive (Rao, et al., 2014; Shoemaker and Panchenko, 2007). An alternative computational method is to compare condition-specific co-expression networks (Creixell, et al., 2015; Ficklin, et al., 2017). By including quality terms, such as presence/absence and size, with anatomical terms for calculating semantic similarity, these phenotype networks can be used to filter out the condition-specific interactions from the large-scale PPI networks available in the STRING. This can further improve condition-specific PPI network generation, which is extremely beneficial for future molecular biology studies.

The proposed integrative approach will be extremely useful for disease phenotypes in human and other model organisms. Predicting disease genes using biological networks is extremely widespread due to the associated medical implications (Peng, et al., 2017; Wang, et al., 2011; Zickenrott, et al., 2017), and the goal is to improve the accuracy of the predictions. Using the proposed integrative method, experimental knowledge regarding known gene to disease associations can be integrated with the human gene networks. For this purpose, Human Disease Ontology (Kibbe, et al., 2015; Osborne, et al., 2009) can be used instead the Uberon to semantically capture the gene to disease annotations. Therefore, the integrative framework proposed in this work is adaptable for a broad number of research questions and is a powerful tool for the bioinformatics community, who thrives on improving the accuracy of biological analyses.

This work focuses on improving the quality of the PPI networks computationally using biological ontologies to capture gene-anatomical phenotype associations, which was not captured by PPI networks. However, the quality of the empirical PPI networks must also improve to make better biological conclusions. This is an active research area which focuses on improving the quality of the experimental methods, such as yeast two-hybrid assay and high-throughput mass-spectrometry and adding novel experimental methods, such as electron microscopy (Beck and Baumeister, 2016) and synthetic lethality (Wang, et al., 2017; Ye, et al., 2005). Furthermore, novel computational techniques to predict PPIs have also emerged, such as using post-translational modifications (e.g., phosphorylation and glycosylation patterns of proteins) for the predictions (Duan and Walther, 2015; Minguéz, et al., 2014). The experimental and computational PPI data generation methods must be continuously improved and complement each other to achieve the overall goal of generating high-quality PPI networks. In this context,

the integrative method proposed in this work is a timely addition to improve PPI network quality as biological conclusions that are made based on the networks highly depend on their quality.

## **1.5 Conclusion**

This work focuses on improving the quality of the PPI networks by integrating anatomy ontology data. The networks were evaluated under different settings: four semantic similarity calculation methods (Lin, Resnik, Schlicker, and Wang), two evaluation curves (ROC and precision-recall curves), two evaluation modes (single-function evaluation method vs. multi-function evaluation method), two model organisms (mouse and zebrafish), and two network conditions (filtered vs. unfiltered); in all the settings, the integrated networks significantly outperformed the PPI networks, which reflects in robustness. The results were further validated to confirm the biological significance and to negate the effect of the circular use of the anatomy profiles, which also confirmed the significance of the integrative framework. Together, these results support the hypothesis that the integration of the experimental knowledge *via* anatomy ontology increases the quality of the PPI networks, therefore, improving their candidate gene prediction accuracy. The integrated networks are now optimized to detect gene modules associated with anatomical phenotypes with a higher accuracy (the focus of the second Chapter). Furthermore, this approach can be applied to diverse biological questions, such as predicting disease genes or evolutionary novelties, which makes it an important tool in bioinformatics to unravel candidate genes.

## Tables

Table 1.1. The statistics for the reconciliation of gene names between the anatomical profiles and the STRING PPI networks for zebrafish and mouse.

	Zebrafish	Mouse
Number of genes in the original anatomical profile	5,405	14,652
Number of Uberon terms/functions in the original anatomical profile	960	1,537
Number of genes directly matched to the PPI network using the gene name (round 1)	2,527	8,166
Number of genes matched using the ensemble IDs (round 2)	402	378
Number of genes matched using synonyms in the STRING database (round 3)	119	63
Number of final gene matches to the PPI network (in all 3 rounds)/ Number of genes in the reconciled anatomical profile	3,048	8,607
Number of Uberon terms in the reconciled anatomical profile	943	1,524
Number of final gene mismatches	2,357	6,045
Number of genes kept in the profile file after filtering the Uberon terms with less than 10 gene annotations	3,037	8,606
Number of Uberon terms kept in the profile file after filtering the terms with less than 10 gene annotations	294	850



Table 1.2. The statistics for the unfiltered and filtered anatomy-based gene networks for zebrafish

	Lin method	Resnik method	Schlicker method	Wang method
Number of genes in the unfiltered network	5,405	5,405	5,405	5,405
Number of interactions in the unfiltered network	14,534,897	14,534,897	14,534,897	14,604,258
The gene similarity score cutoff	0.55	0.18	0.24	0.70
Number of genes in the filtered network	5,387	4,909	5,401	5,386
Number of interactions in the filtered network	700,138	712,286	692,539	789,282

Table 1.3. The statistics for the unfiltered and filtered anatomy-based gene networks for mouse

	Lin method	Resnik method	Schlicker method	Wang method
Number of genes in the unfiltered network	14,652	14,652	14,652	14,652
Number of interactions in the unfiltered network	107,094,117	107,094,117	107,094,117	107,324,905
The gene similarity score cutoff	0.9	0.32	0.41	0.95
Number of genes in the filtered network	9,784	10,081	12,755	9,126
Number of interactions in the filtered network	588,359	536,602	522,183	510,139

Table 1.4. The statistics for the unfiltered and filtered integrated networks for zebrafish

	Lin method	Resnik method	Schlicker method	Wang method
Number of genes in the unfiltered network	25,375	25,375	25,375	25,375
Number of interactions in the unfiltered network	26,753,086	26,753,086	26,753,086	26,821,274
The gene similarity score cutoff	0.33	0.23	0.24	0.4
Number of genes in the filtered network	17,394	20,066	20,929	13,940
Number of interactions in the filtered network	730,855	726,589	690,208	744,519

Table 1.5. The statistics for the unfiltered and filtered integrated networks for mouse

	Lin method	Resnik method	Schlicker method	Wang method
Number of genes in the unfiltered network	27,097	27,097	27,097	27,097
Number of interactions in the unfiltered network	111,461,010	111,461,010	111,461,010	111,690,355
The gene similarity score cutoff	0.50	0.27	0.30	0.53
Number of genes in the filtered network	13,125	17,898	18,002	12,916
Number of interactions in the filtered network	653,848	661,619	613,671	712,720

## Figures

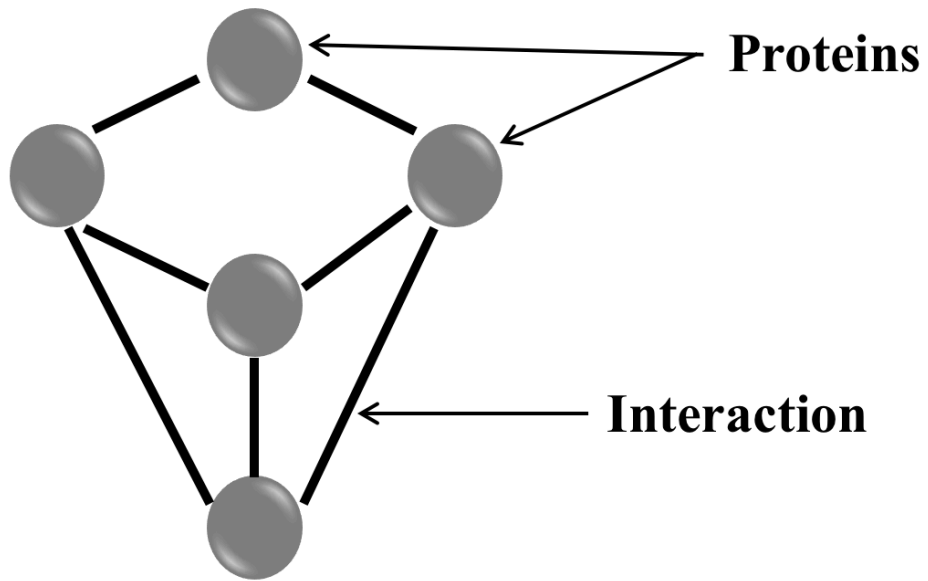


Figure 1.1. Representation of protein-protein interactions in a graph. The nodes (gray) represent proteins and the edges (black) represent their interactions



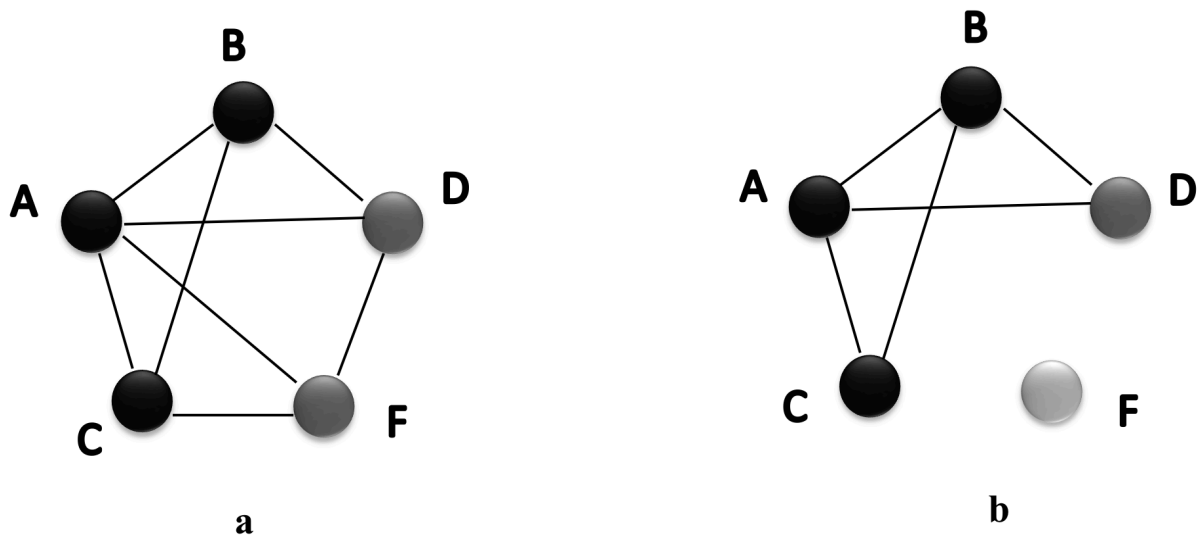


Figure 1.3. A hypothetical scenario that compares candidate gene predictions based on a (a) PPI network and an (b) anatomy-based gene network. The nodes A, B, and C in both networks represent three genes known to be associated with a certain phenotype, which can be denoted as phe1. Because their phenotype is known, they are colored in black. In the PPI network (a), genes D and F are predicted to be associated with phe1 based on their interactions with known genes. In contrast, the anatomy-based gene network (b) only predicts D as a potential candidate because the gene F does not have any interaction with other genes. The absence of interactions of gene F can be due to two reasons: (1) it is not annotated with any anatomical terms, (2) it is not annotated with terms that are similar to the anatomy terms associated with genes: A, B, and C. The anatomy-based gene network (b) is built entirely on anatomy ontology information, thus it provides a different interaction structure. Hypothetically, the gene F could be a false positive interaction in the PPI network, and the integrative use of the anatomy-based network may reduce the false positives by filtering them.

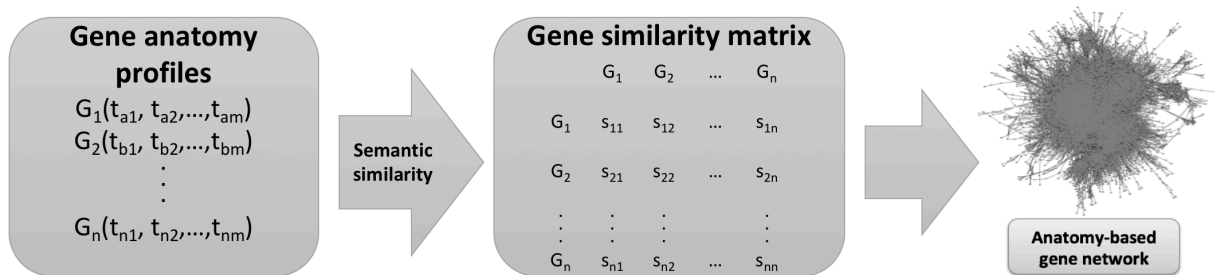


Figure 1.4. The general workflow for generating anatomy-based gene networks. The genes are represented by  $G_1, G_2$ , etc., and their Uberon annotations are represented by  $t_{a1}, t_{b1}$ , etc. In the gene similarity matrix, the similarity scores between genes are represented by  $s_{11}, s_{12}$ , etc.



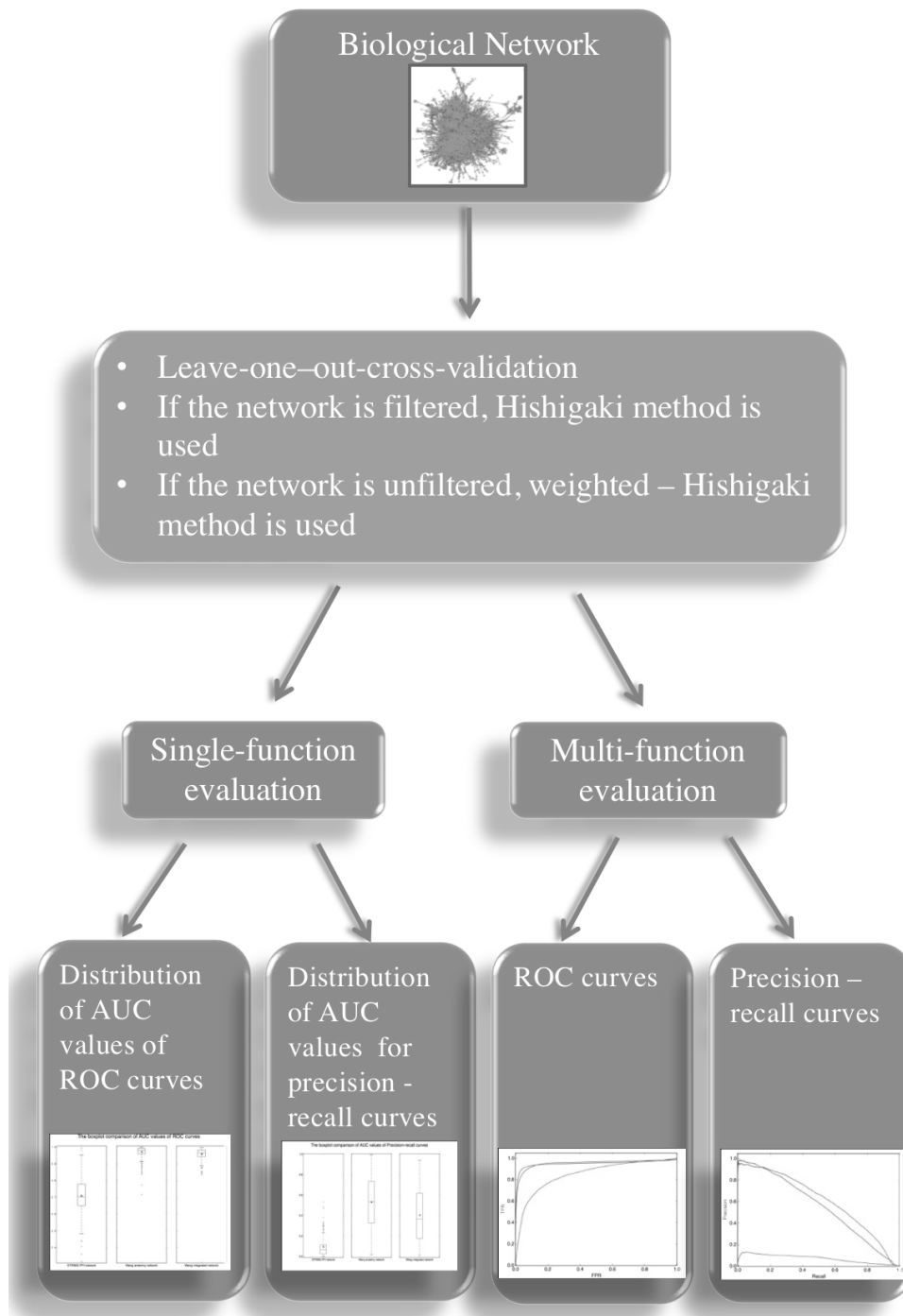


Figure 1.5. The general evaluation workflow used for evaluating the networks. If the single-function evaluation method is selected, the distributions of the area under the curve (AUC) values are compared, whereas the direct AUC values for the ROC and precision-recall curves are compared in the multi-function method.

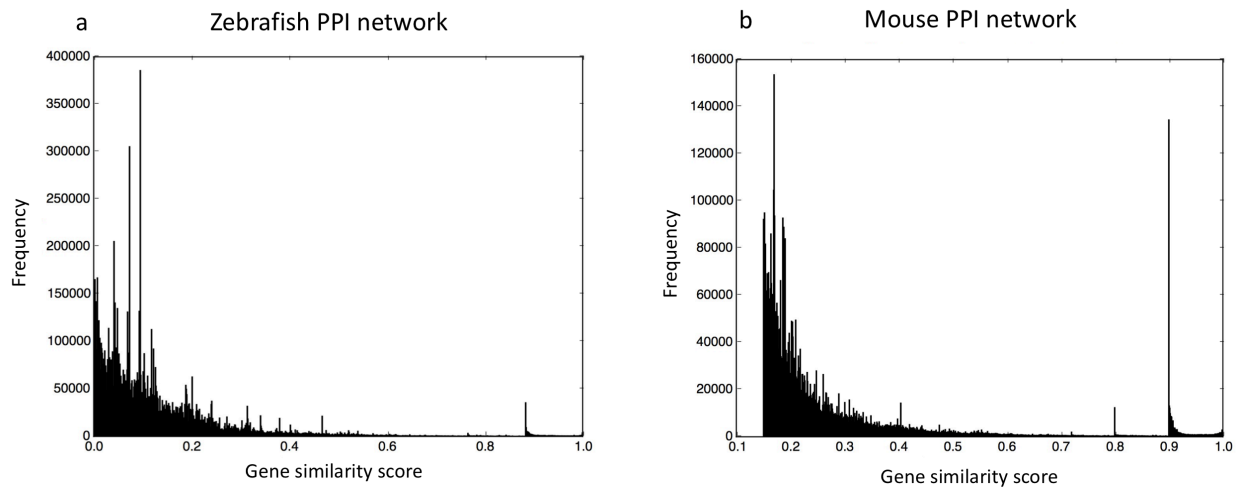


Figure 1.6. Gene similarity score/combined score distributions for (a) zebrafish and (b) mouse unfiltered PPI networks.

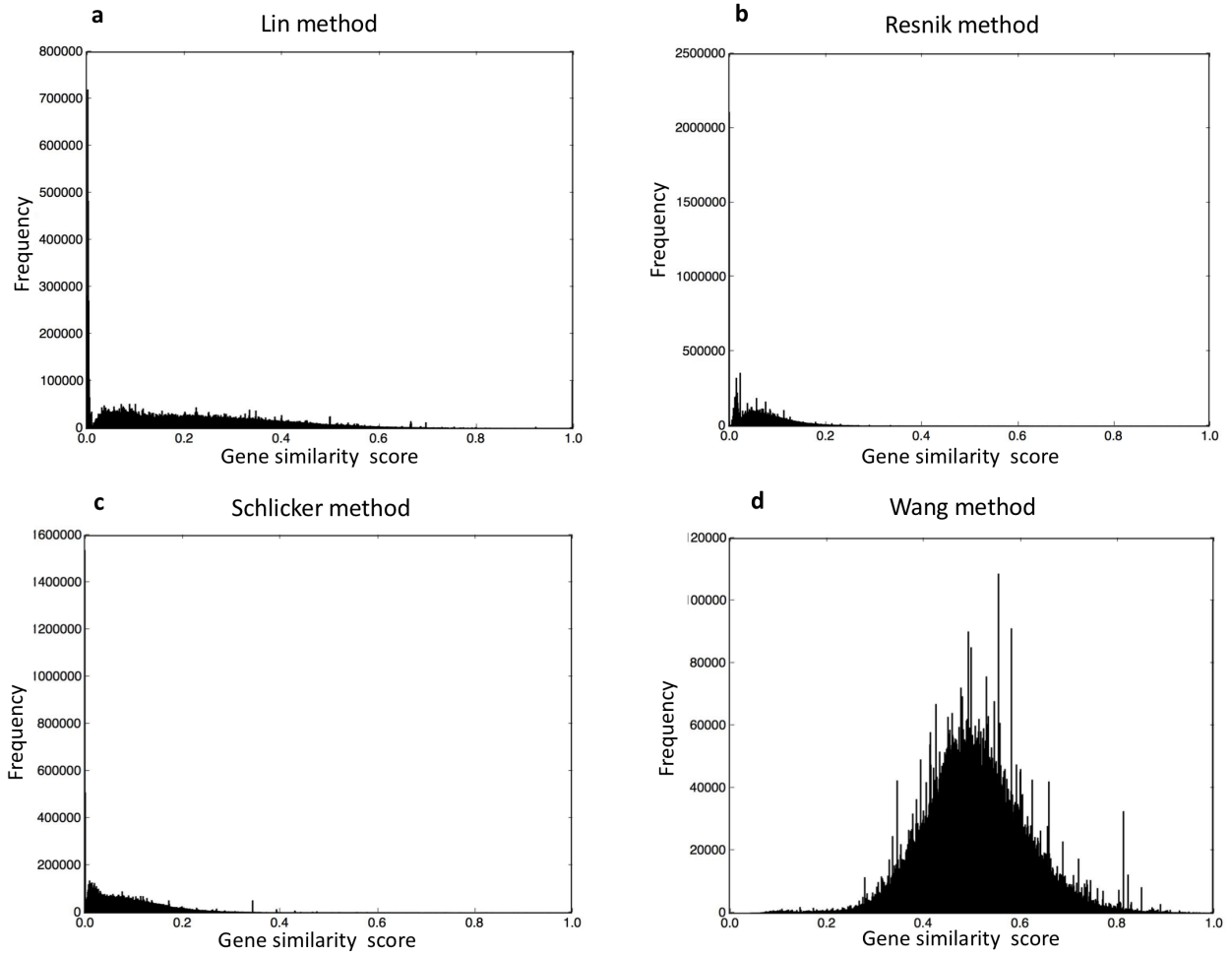


Figure 1.7. The gene similarity score distributions for the zebrafish unfiltered anatomy-based gene networks constructed by (a) Lin method, (b) Resnik method, (c) Schlicker method, and (d) Wang method.

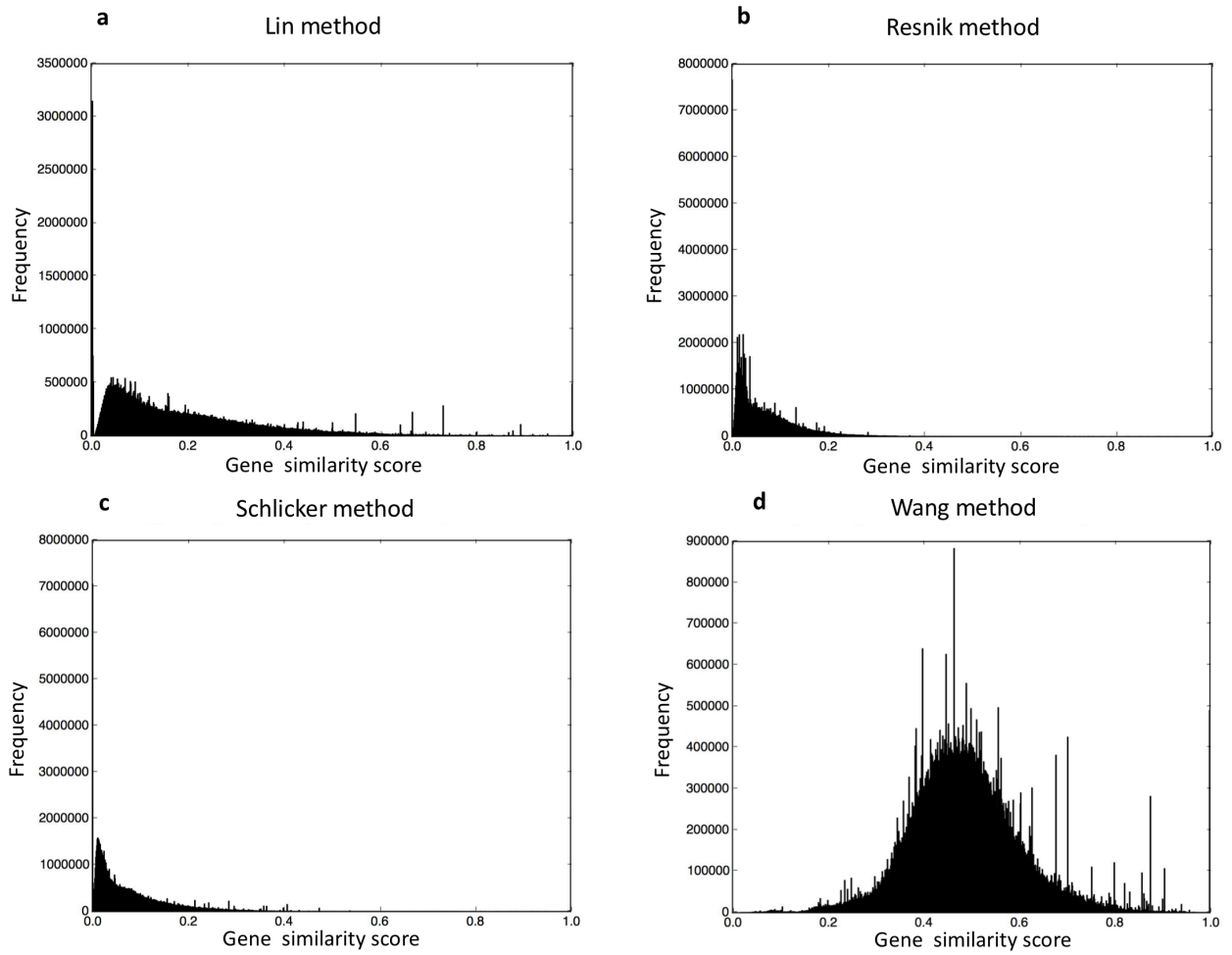


Figure 1.8. The gene similarity score distributions for the mouse unfiltered anatomy-based gene networks constructed by (a) Lin method, (b) Resnik method, (c) Schlicker method, and (d) Wang method.

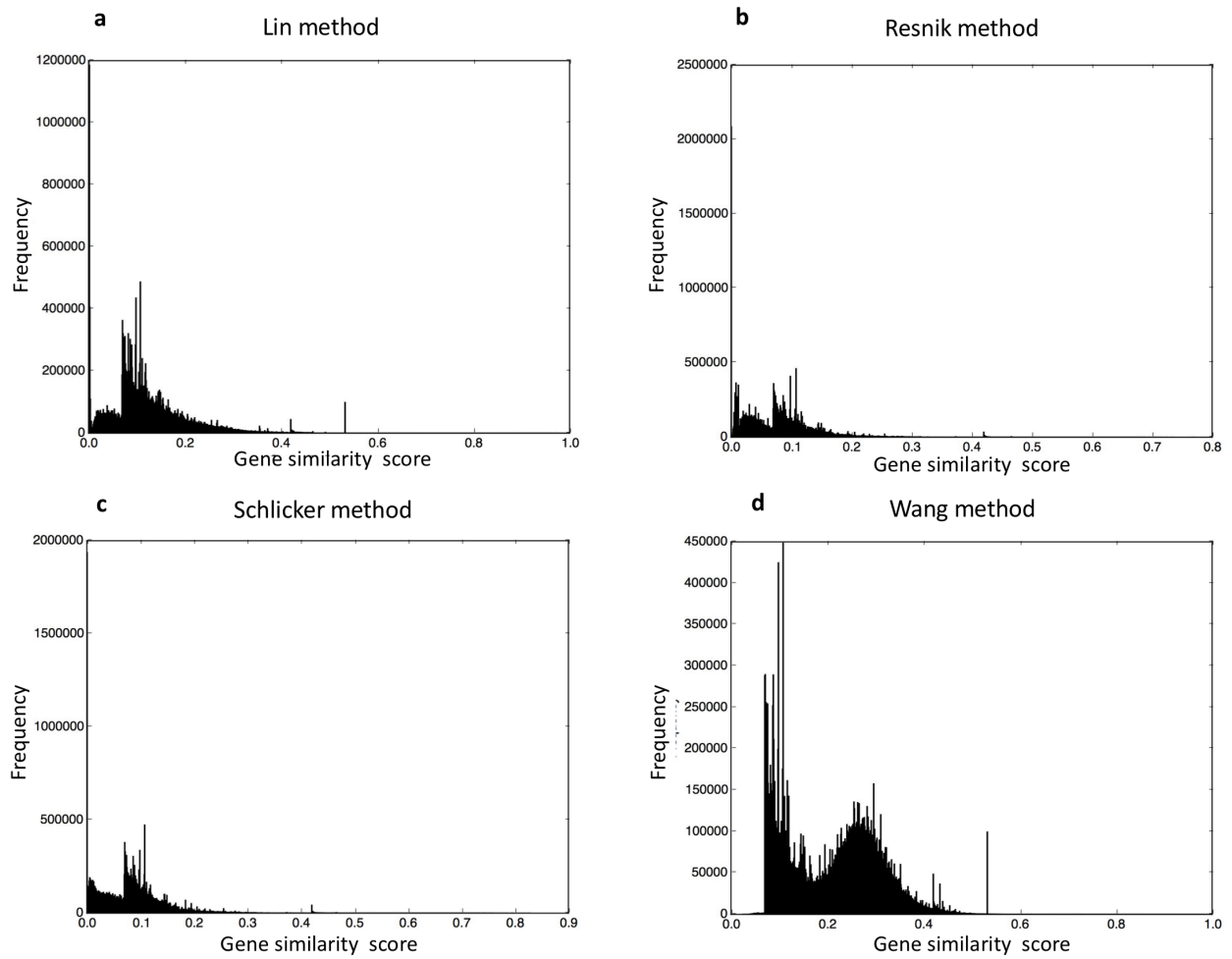


Figure 1.9. The gene similarity score distributions for the zebrafish unfiltered integrated networks constructed by (a) Lin method, (b) Resnik method, (c) Schlicker method, and (d) Wang method.

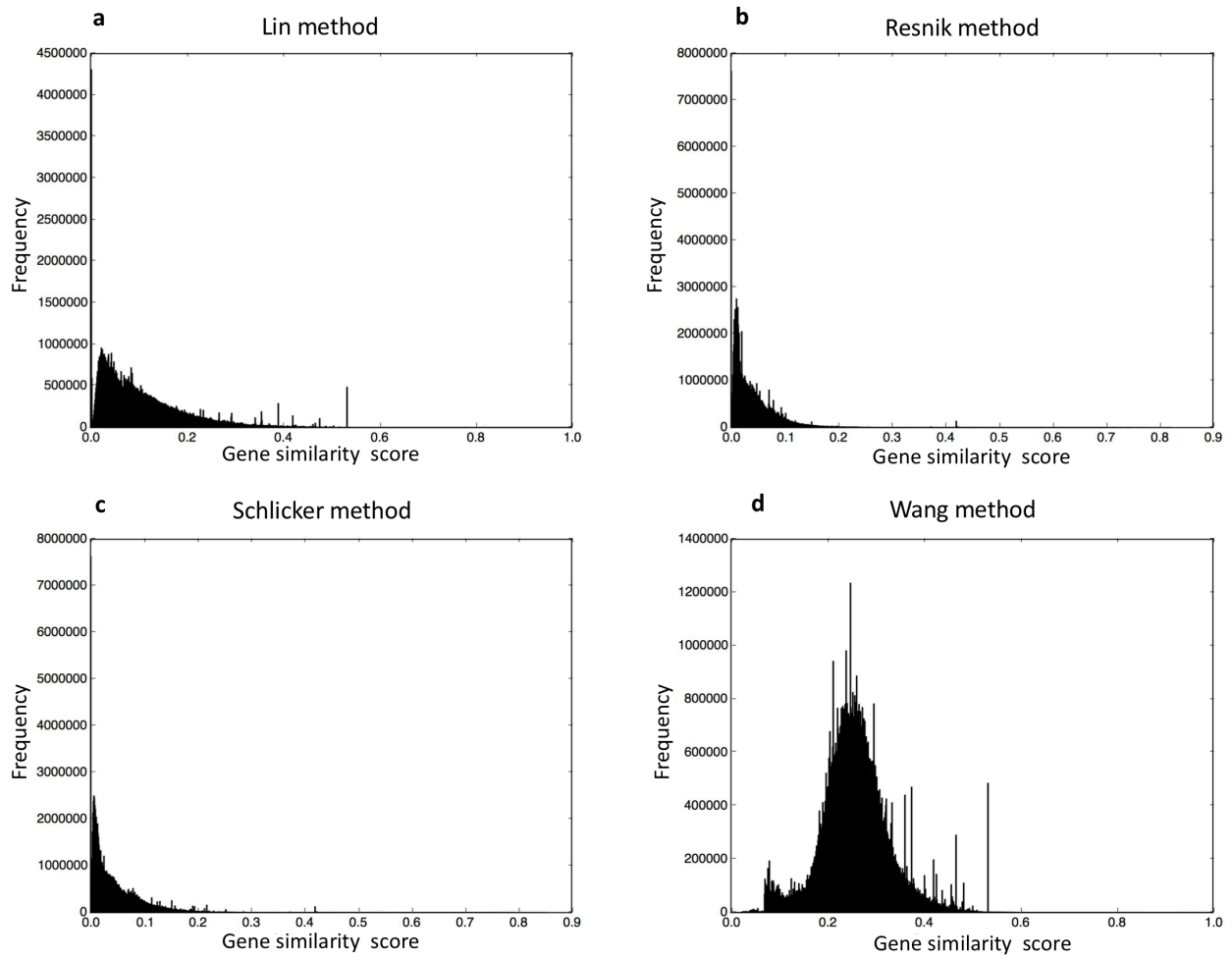


Figure 1.10. The gene similarity score distributions for the mouse unfiltered integrated networks constructed by (a) Lin method, (b) Resnik method, (c) Schlicker method, and (d) Wang method.

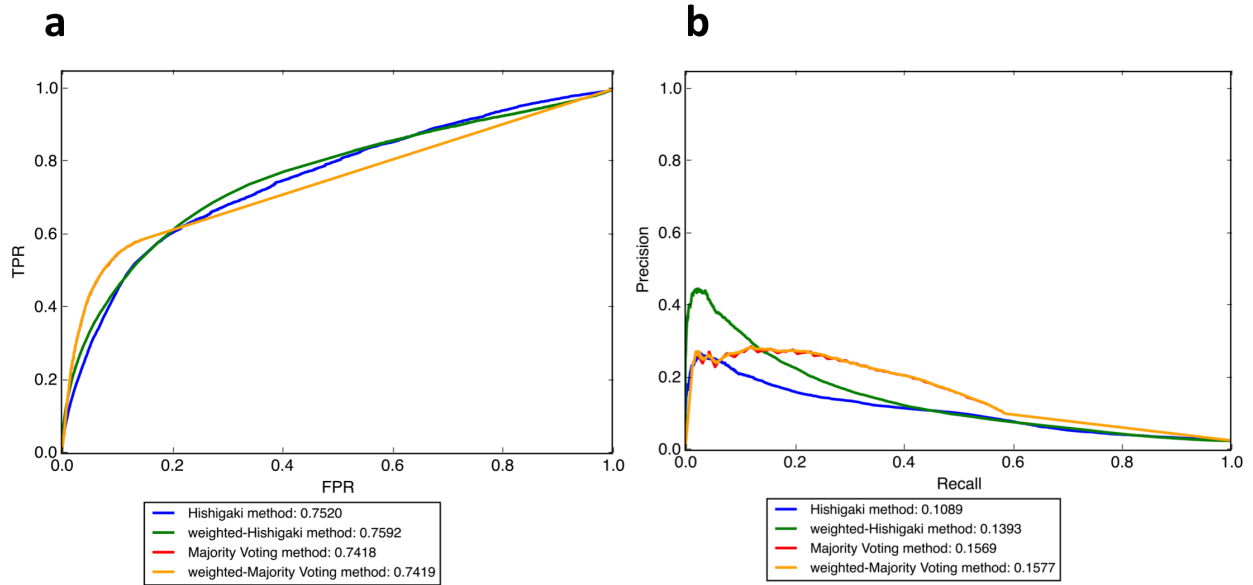


Figure 1.11. The comparison of (a) ROC curves and (b) precision-recall curves for different network-based candidate gene prediction methods. These curves were generated for filtered zebrafish PPI networks using the multi-function evaluation method.

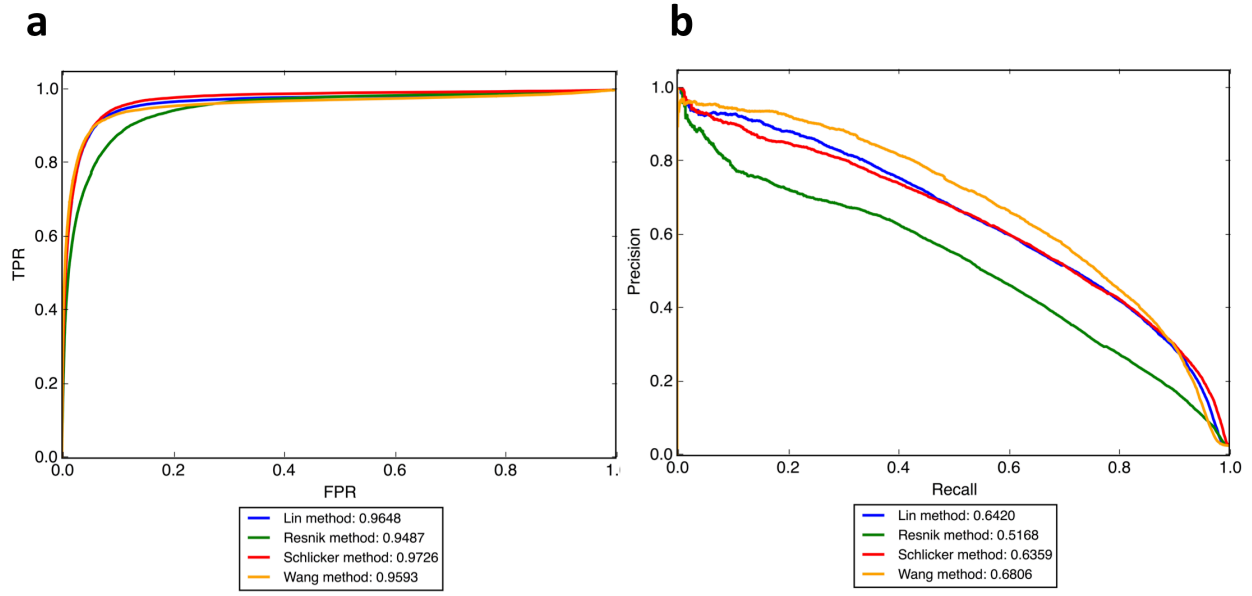


Figure 1.12. The comparison of (a) ROC curves and (b) precision-recall curves for different filtered anatomy-based gene networks for the zebrafish. These curves were generated using the multi-function evaluation method.



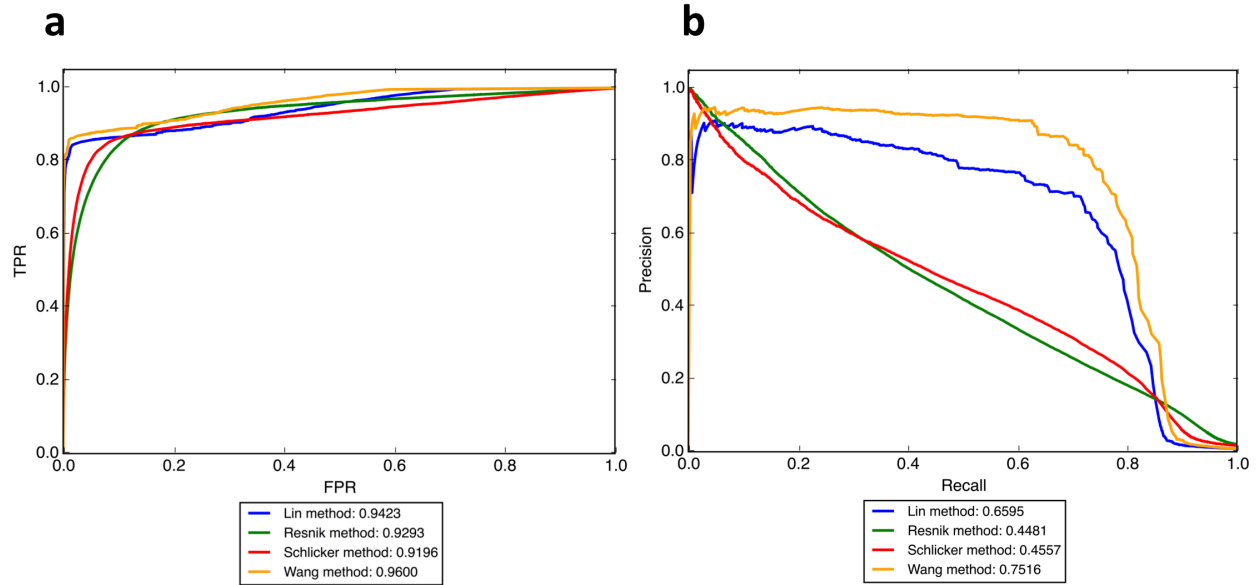


Figure 1.13. The comparison of (a) ROC curves and (b) precision-recall curves for different filtered anatomy-based gene networks for the mouse. These curves were generated using the multi-function evaluation method.

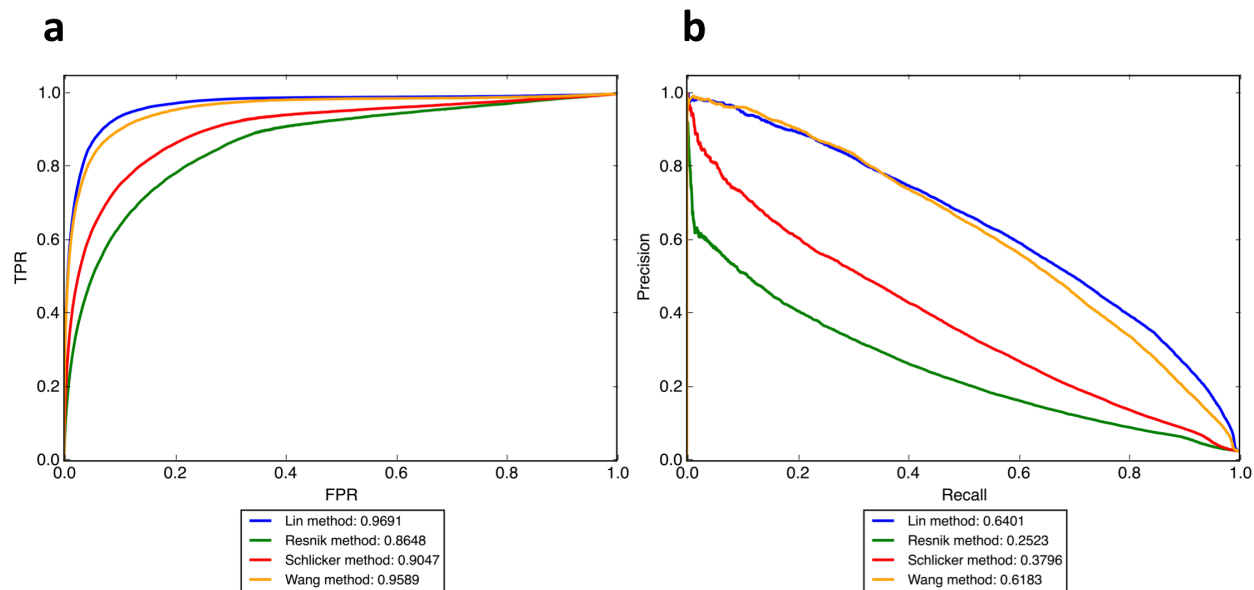


Figure 1.14. The comparison of (a) ROC curves and (b) precision-recall curves for different filtered integrated networks for the zebrafish. These curves were generated using the multi-function evaluation method.

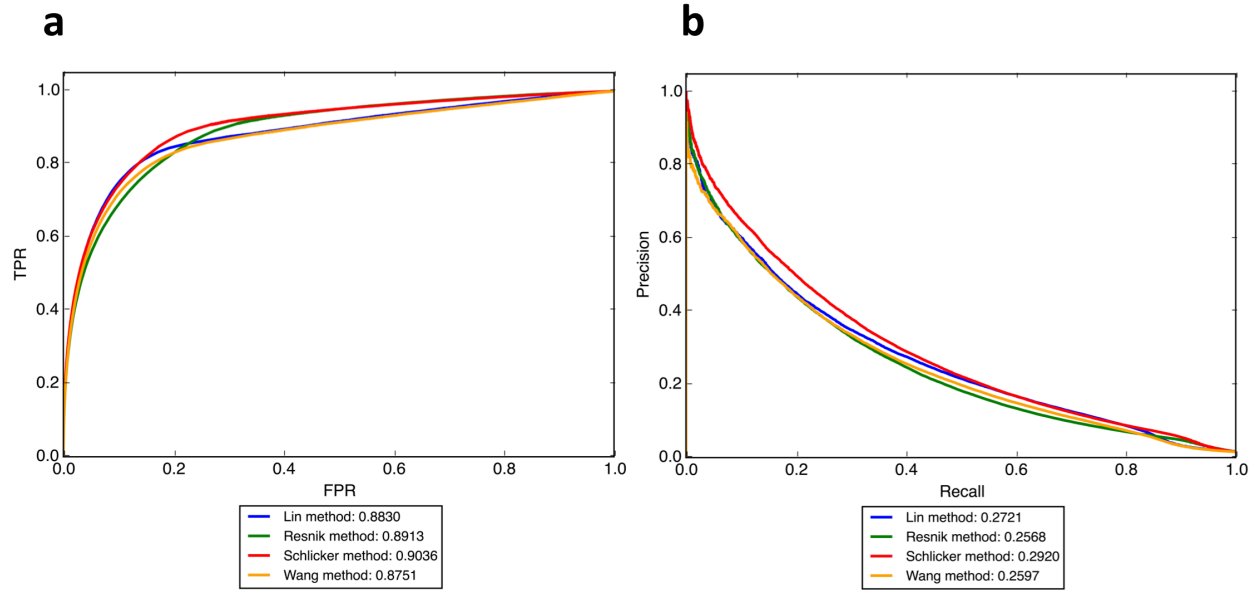


Figure 1.15. The comparison of (a) ROC curves and (b) precision-recall curves for different filtered integrated networks for the mouse. These curves were generated using the multi-function evaluation method.

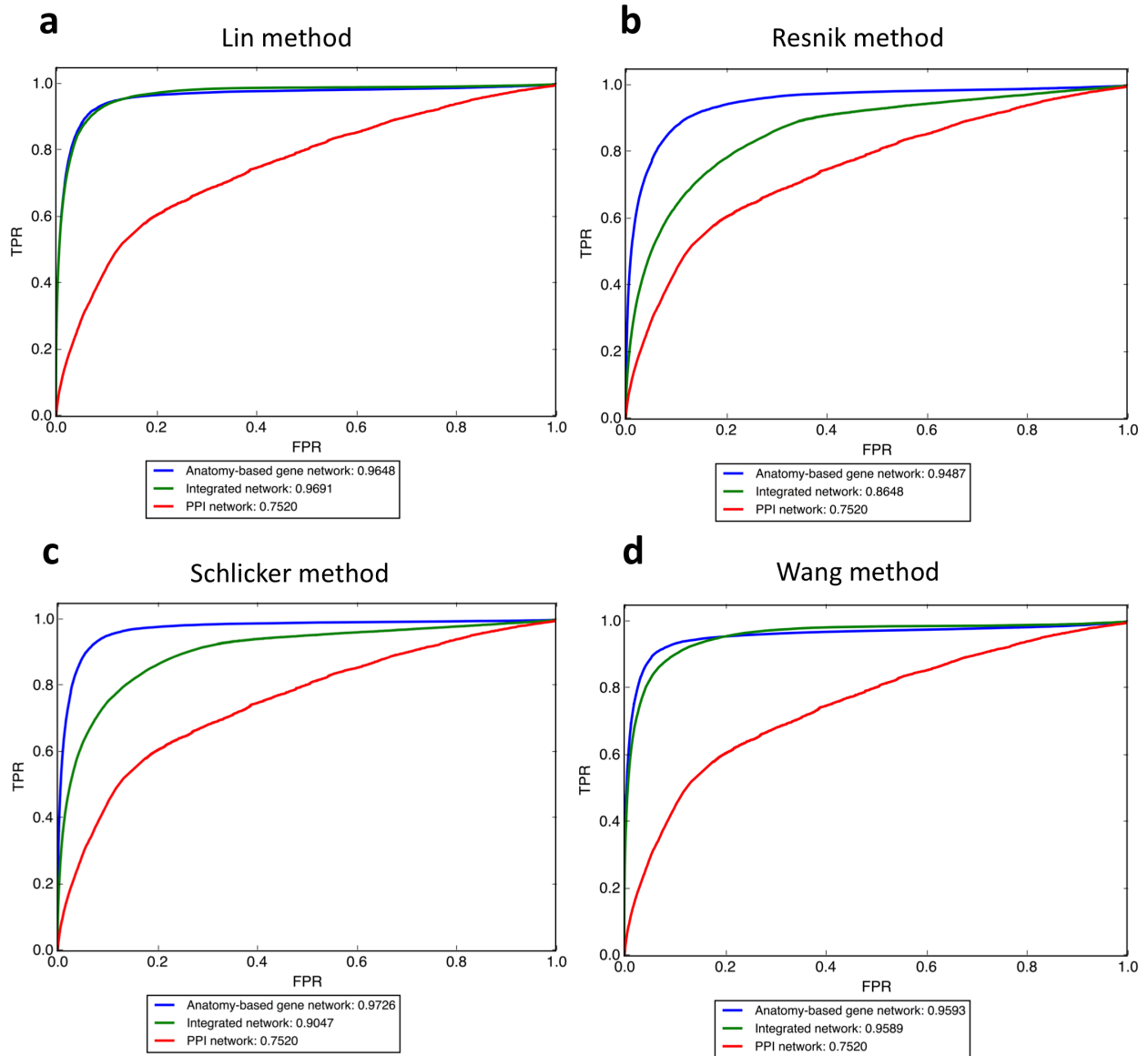


Figure 1.16. The comparison of ROC curves for the filtered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods in the zebrafish.

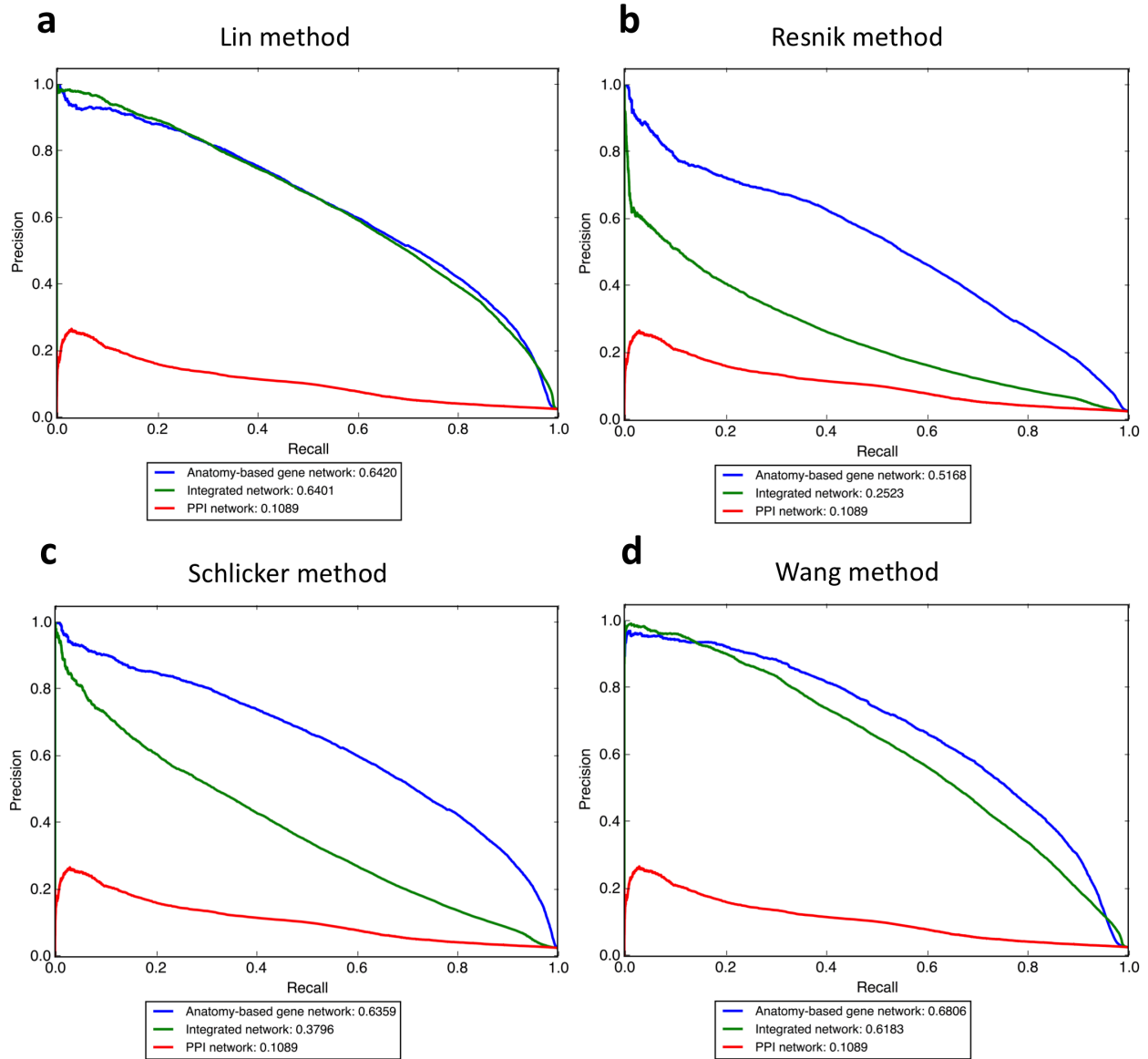


Figure 1.17. The comparison of precision-recall curves for the filtered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods in the zebrafish.

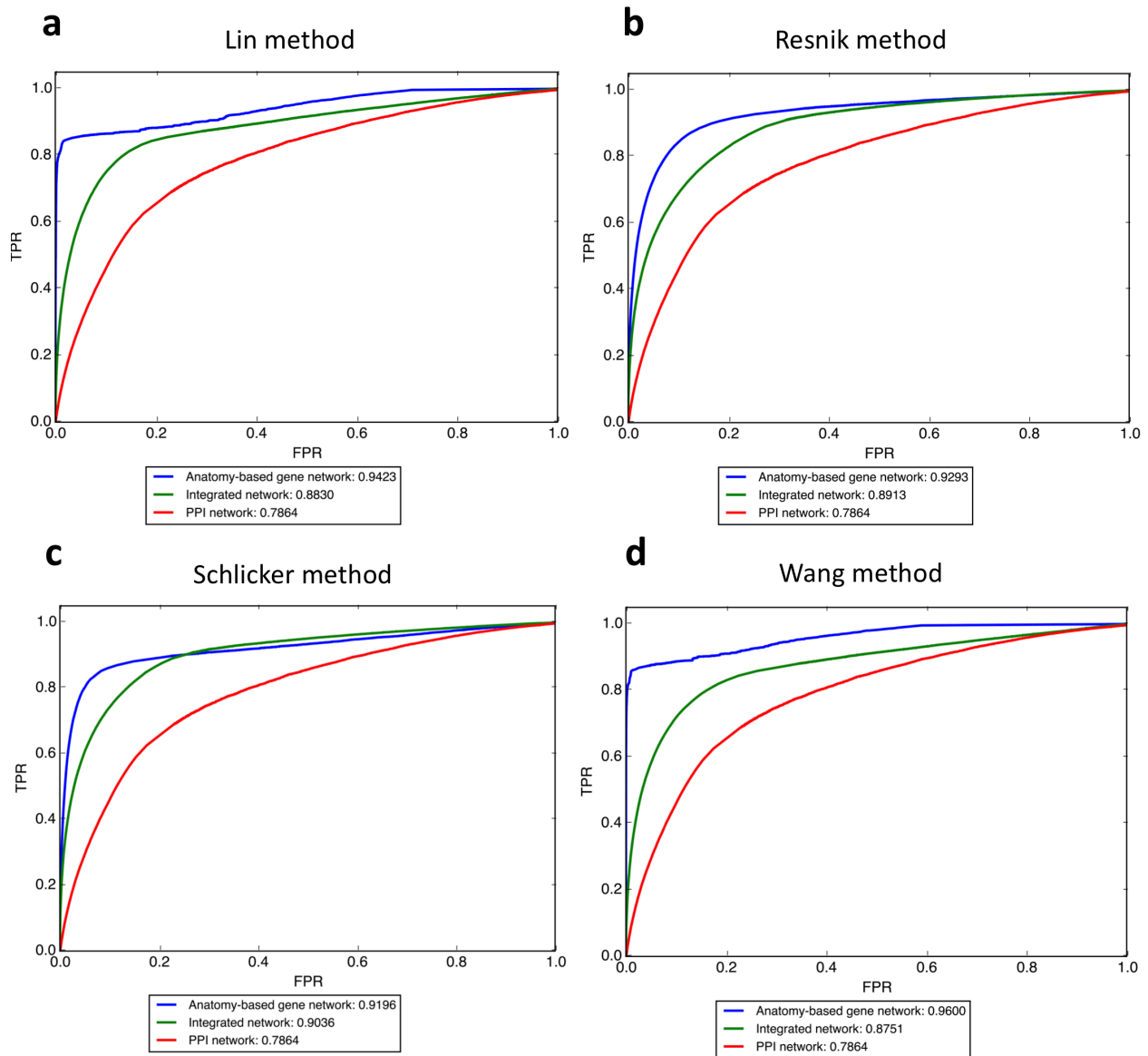


Figure 1.18. The comparison of ROC curves for the filtered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods in the mouse.

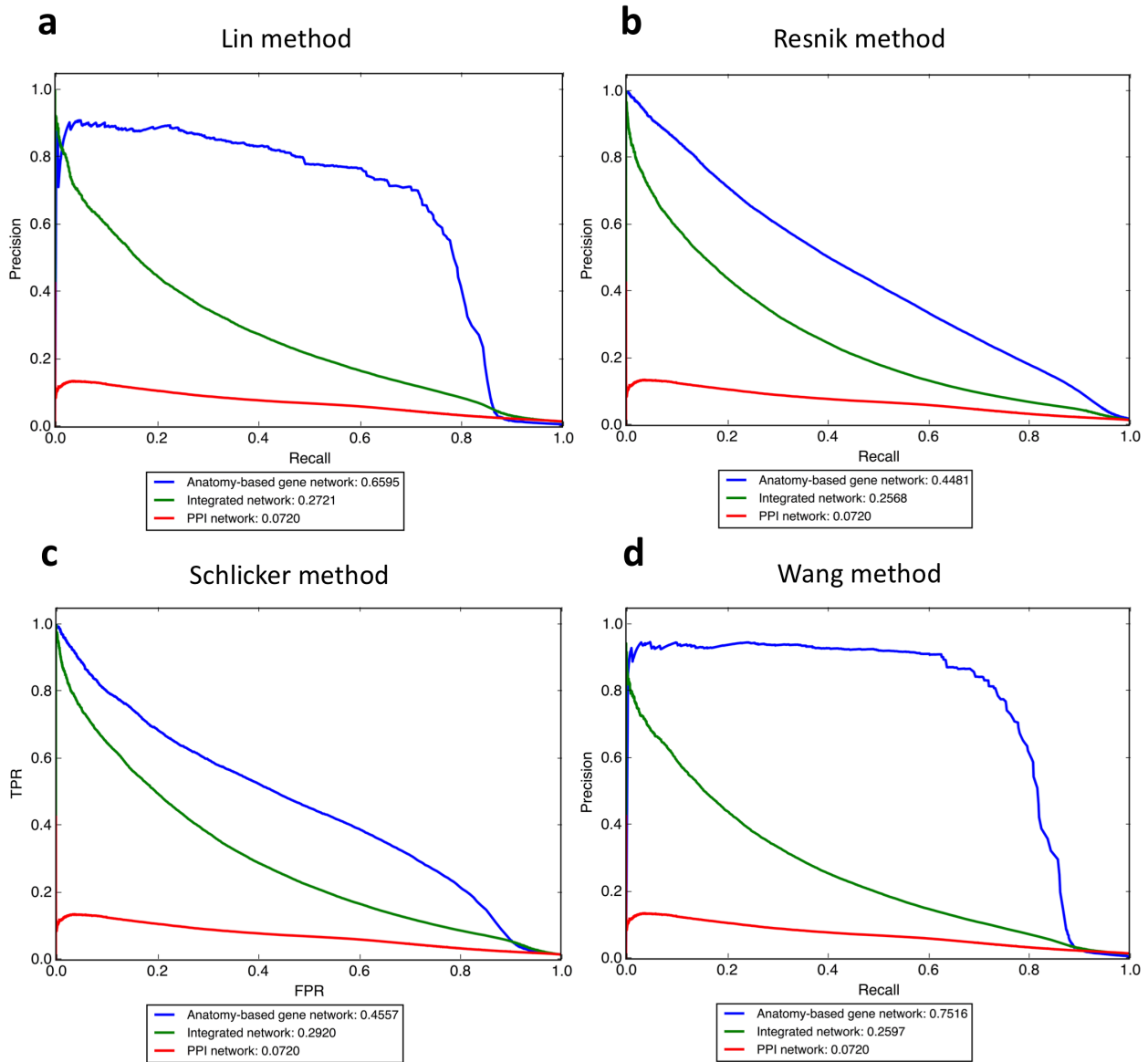


Figure 1.19. The comparison of precision-recall curves for the filtered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods in the mouse.

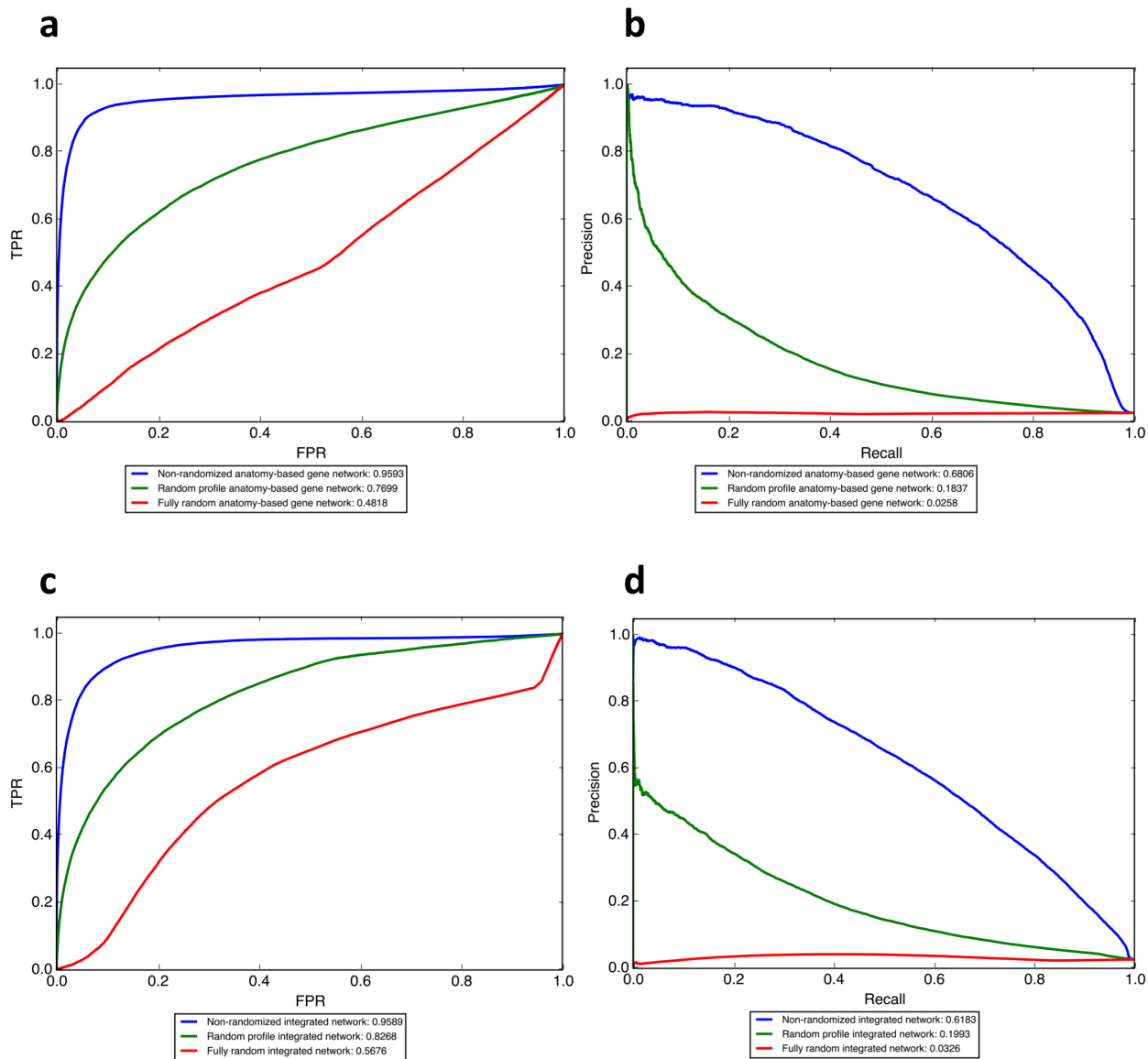


Figure 1.20. The comparison of (a) ROC and (b) precision-recall curves for the filtered non-randomized anatomy-based gene network (blue), random profile anatomy-based gene network (green), and fully random anatomy-based gene network (red) and the comparison of (c) ROC and (d) precision-recall curves for the filtered non-randomized integrated network (blue), random profile integrated network (green), and fully random integrated network for the Wang method for the zebrafish.



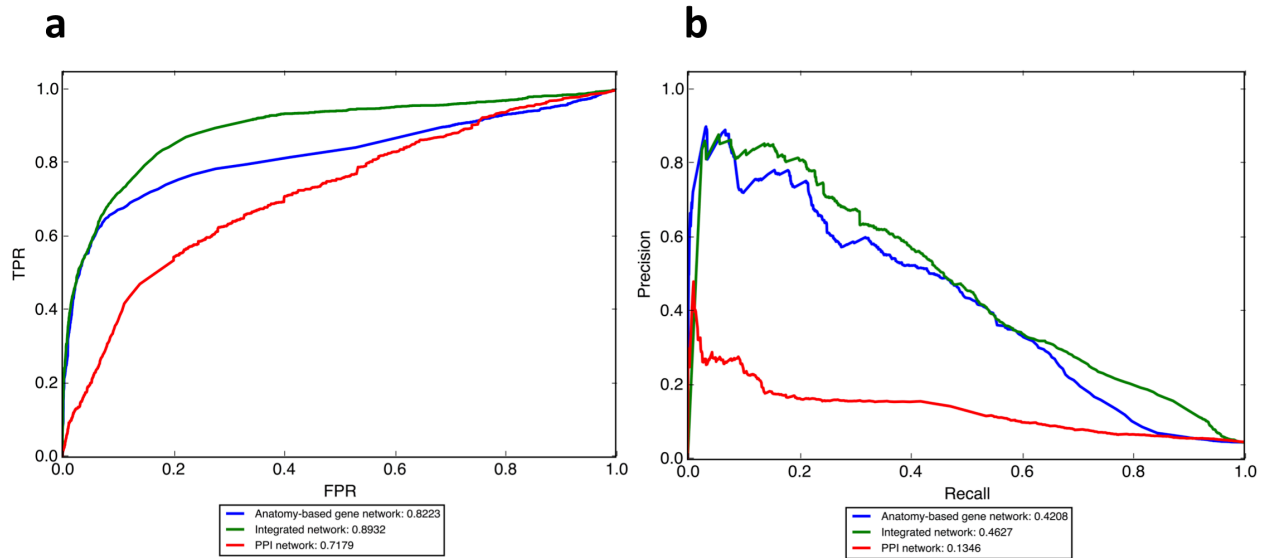


Figure 1.21. The comparison of (a) ROC and (b) precision-recall curves for the filtered integrated network (green), PPI network (red), and anatomy-based gene network (blue) for the Wang method for the zebrafish. The integrated network and the anatomy-based gene network were created using the zebrafish anatomy profile after randomly removing 30 Uberon terms, which had at least 10 gene annotations for each term. The same 30 terms were used for the evaluation to generate the above curves.

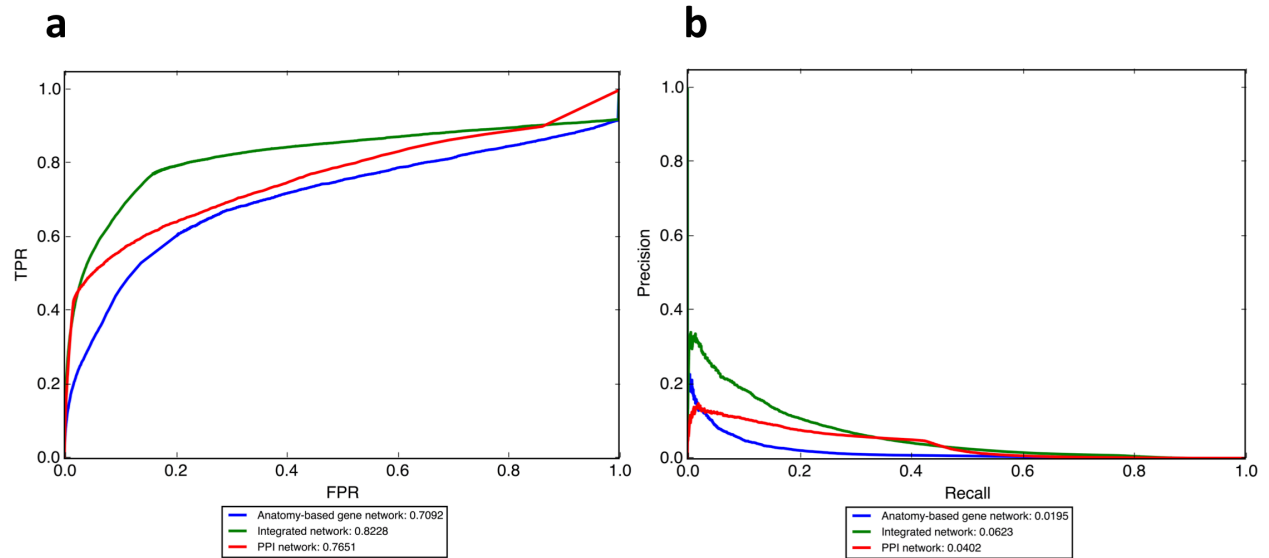
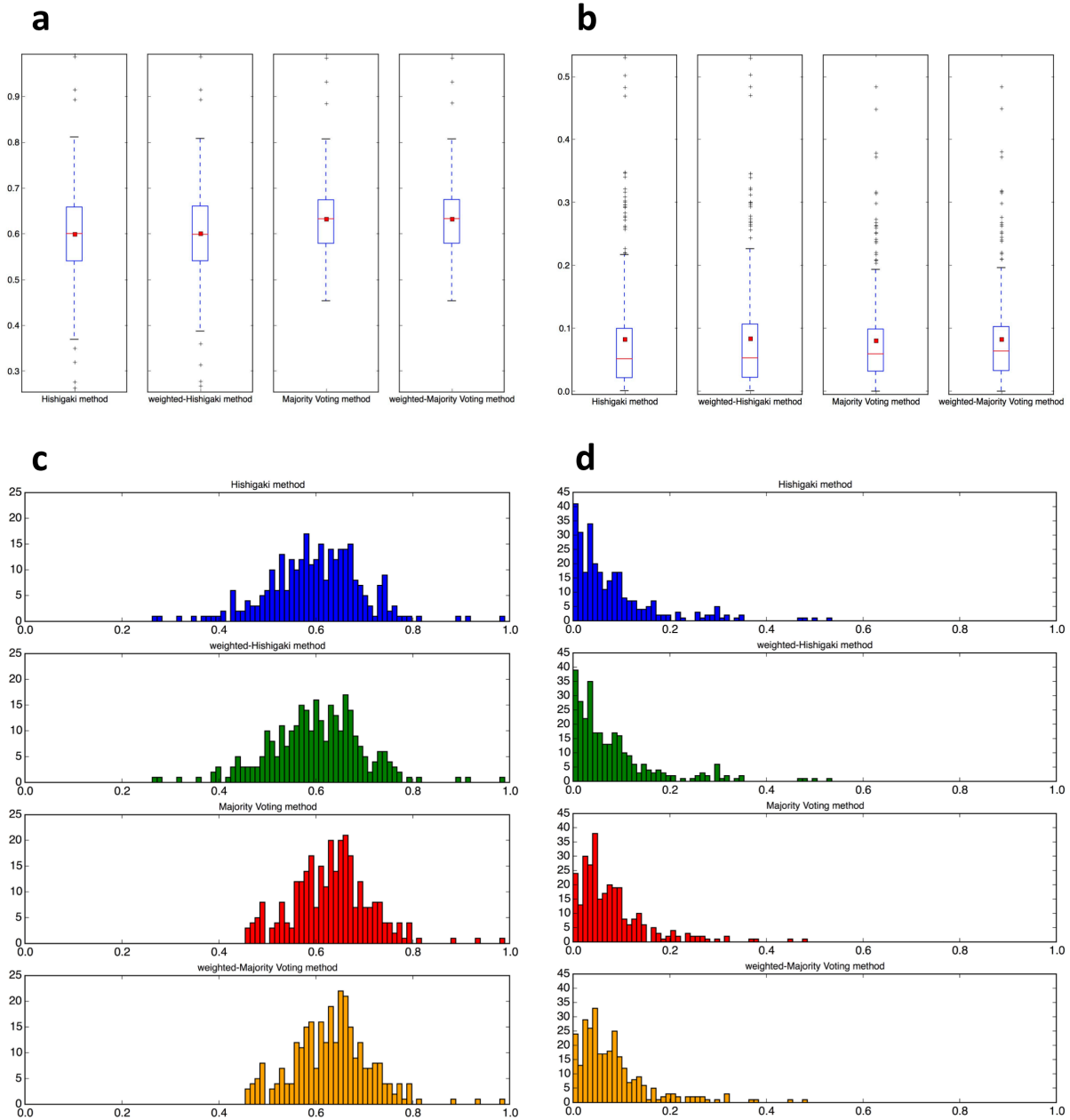
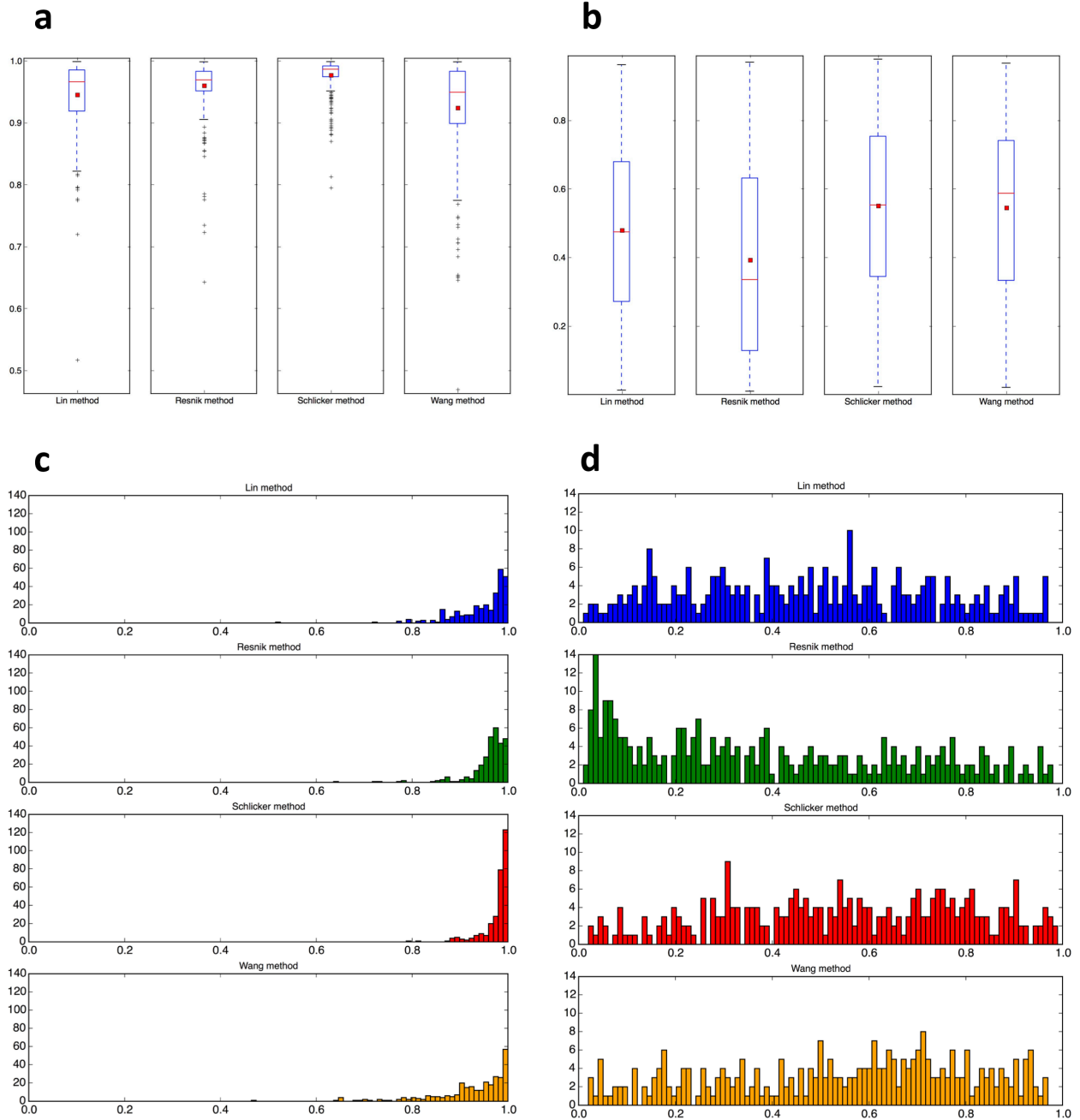


Figure 1.22. The comparison of ROC (a) and precision-recall (b) curves of the filtered integrated network (green), PPI network (red), and anatomy-based gene network (blue) for the Wang method for the zebrafish. The networks were evaluated using annotation profiles that contain Biological Process terms of the Gene Ontology for zebrafish genes.

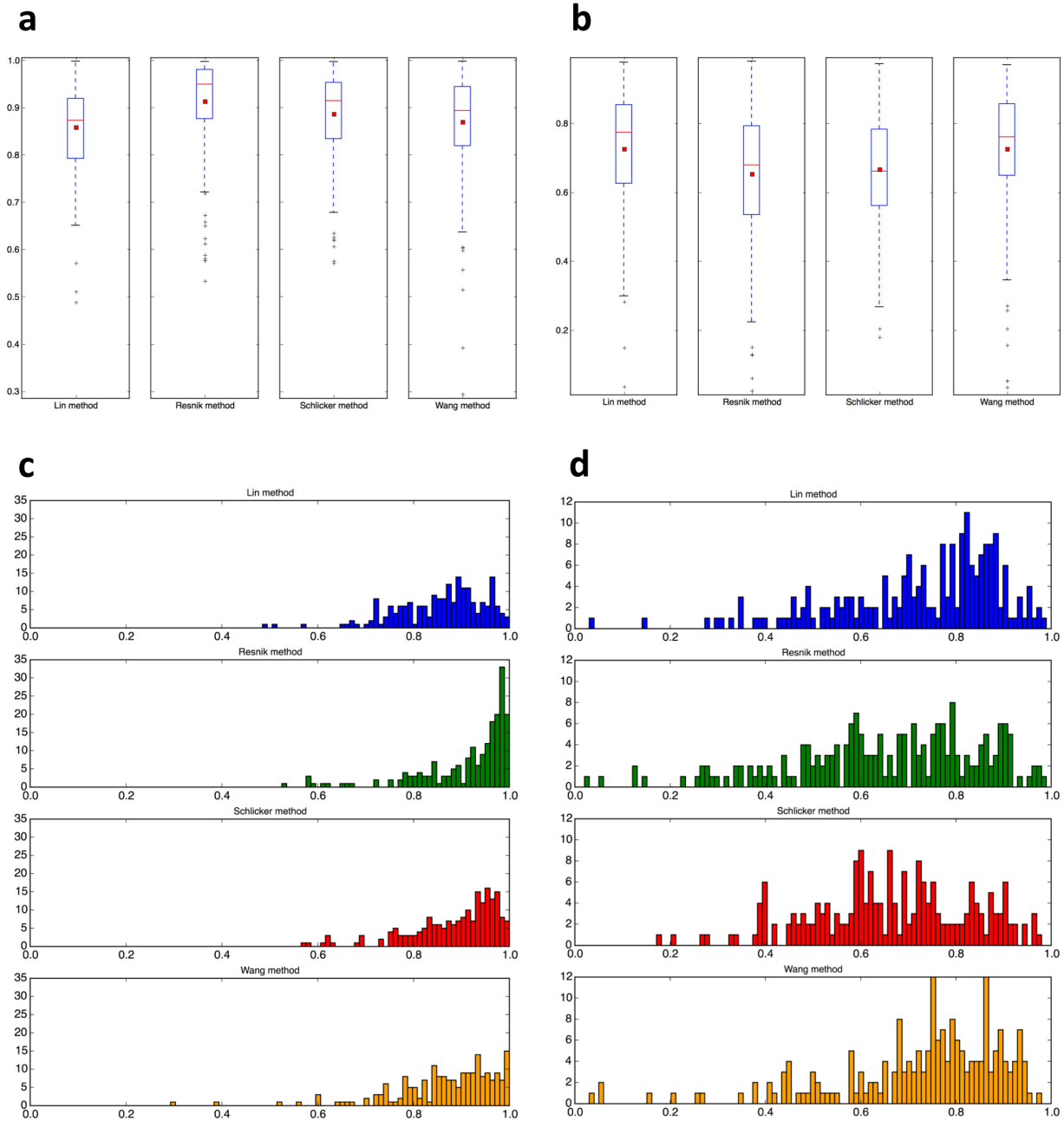
## Supplementary Figures



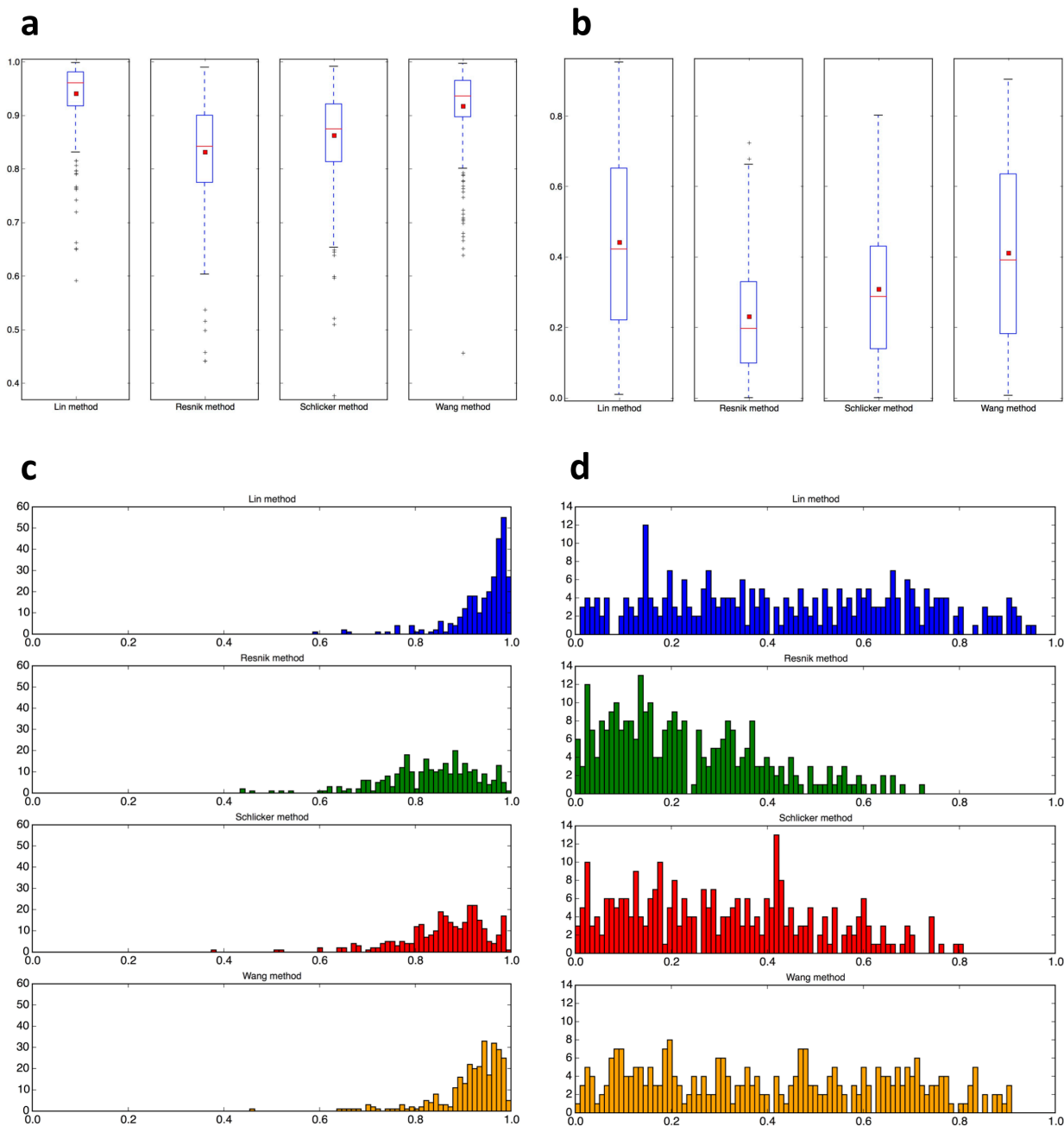
Supplementary Figure S1.1. The boxplot comparisons of the AUC distributions for (a) ROC curves and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for the different network-based candidate gene prediction methods for the zebrafish filtered PPI network. In the boxplots, the red line and the square represent the median and mean, respectively.



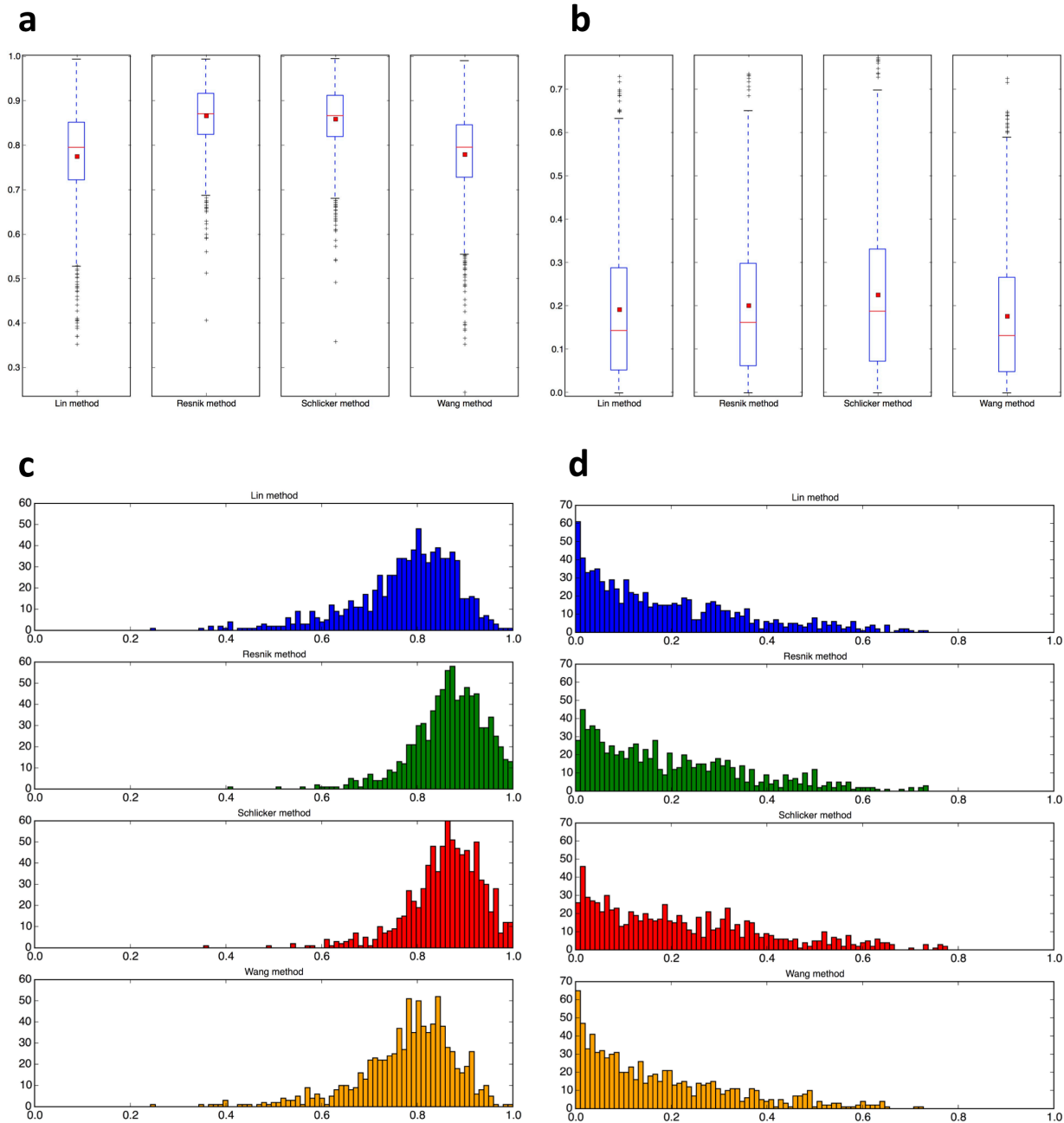
Supplementary Figure S1.2. The boxplot comparisons of the AUC distributions for (a) ROC curves and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for different filtered anatomy-based gene networks for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.



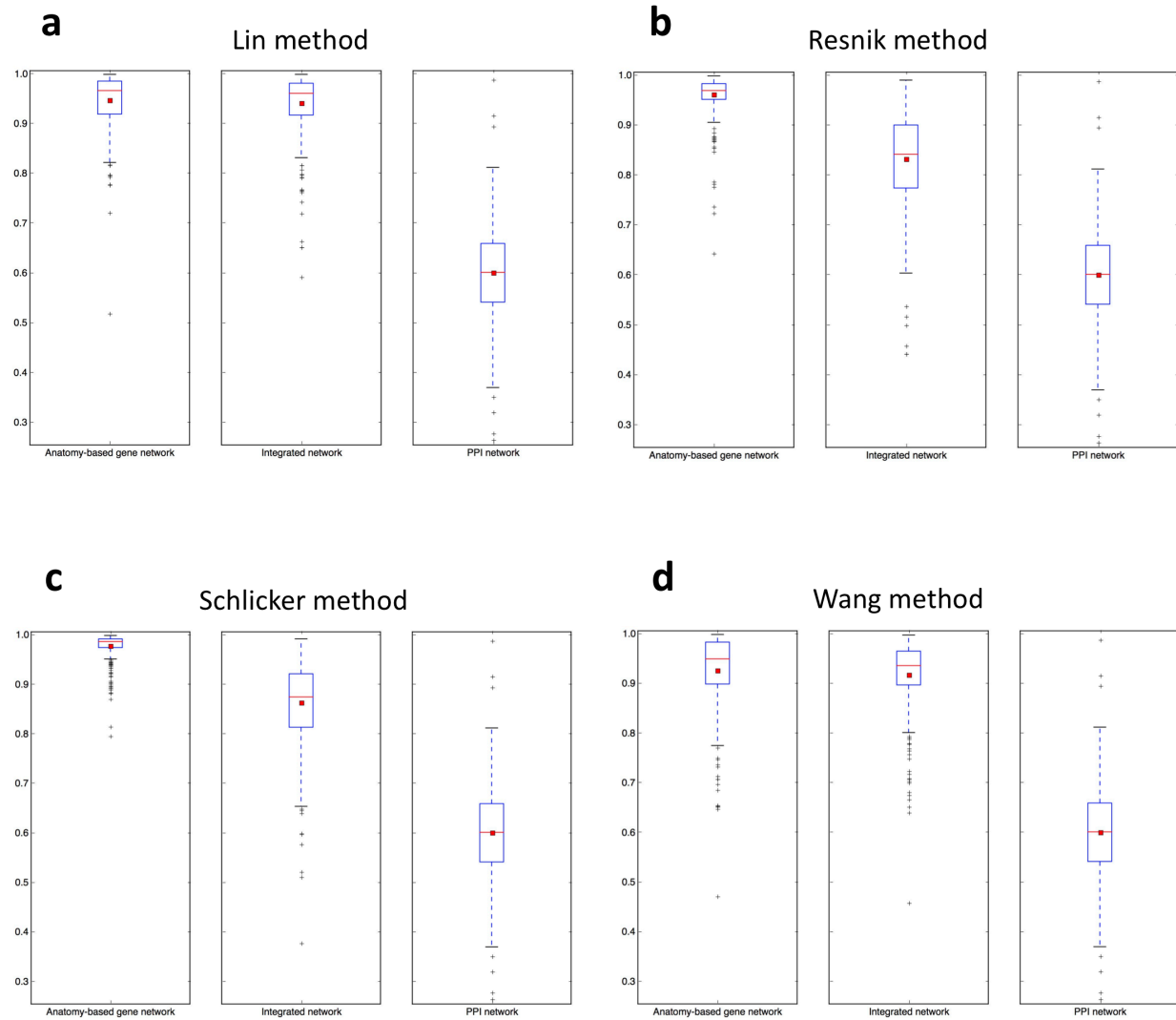
Supplementary Figure S1.3. The boxplot comparisons of the AUC distributions for (a) ROC curves and (a) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for the different filtered anatomy-based gene networks for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.



Supplementary Figure S1.4. The boxplot comparisons of the AUC distributions for (a) ROC curves and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for the different filtered integrated networks for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.

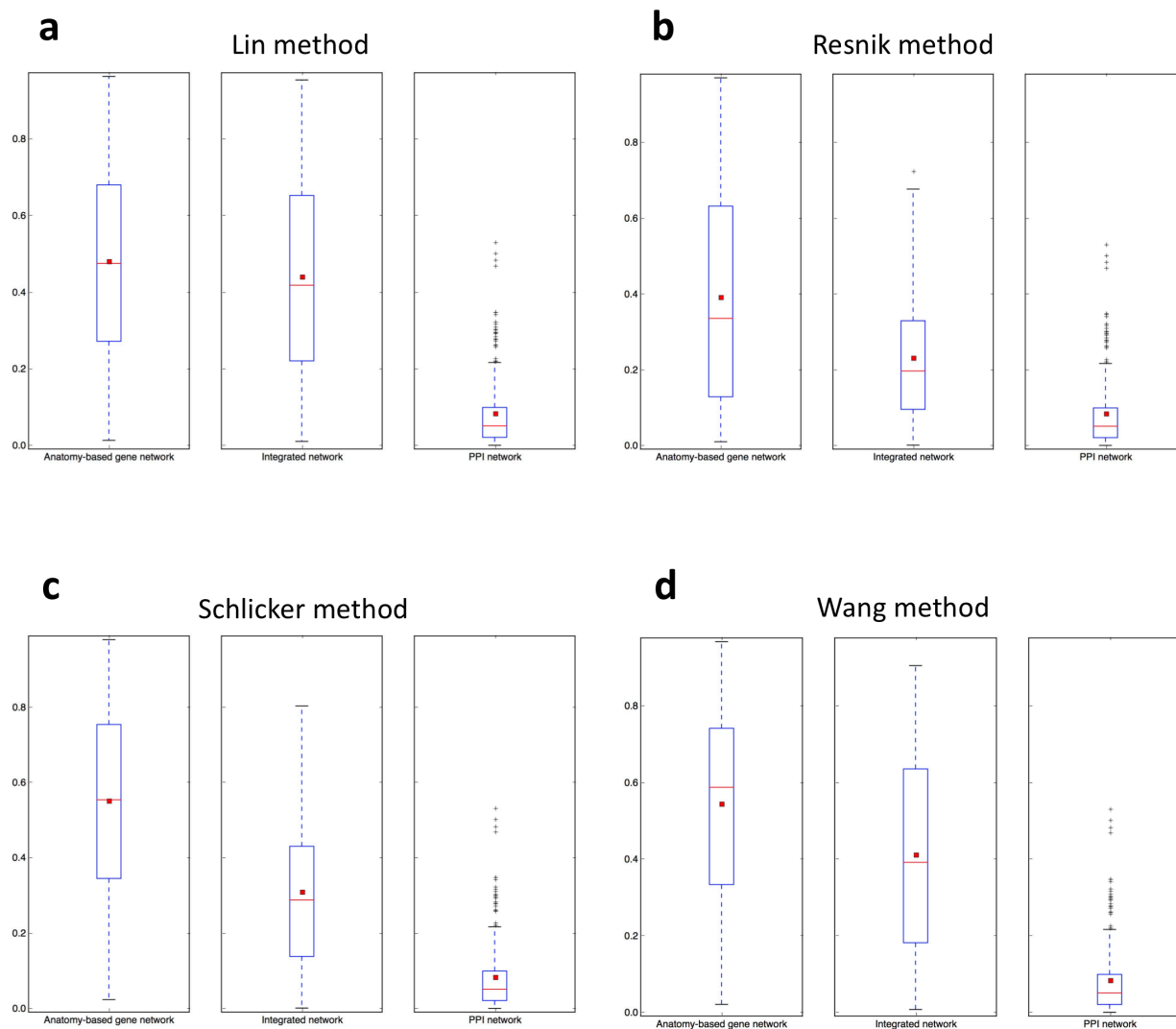


Supplementary Figure S1.5. The boxplot comparisons of the AUC distributions for (a) ROC curves and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC curves and (d) precision-recall curves for the different filtered integrated networks for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.

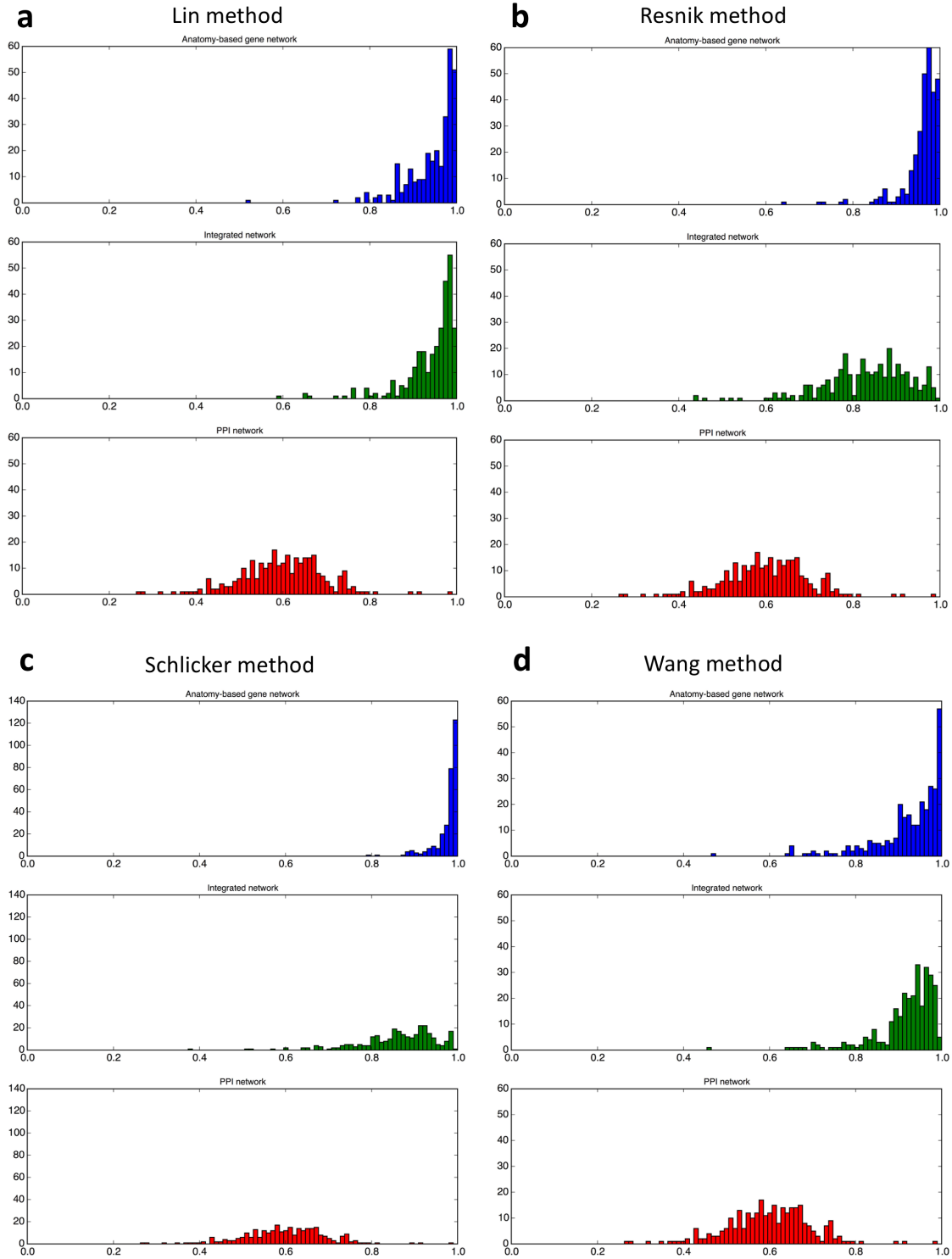


Supplementary Figure S1.6. The boxplot comparisons for the AUC distributions of ROC curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.

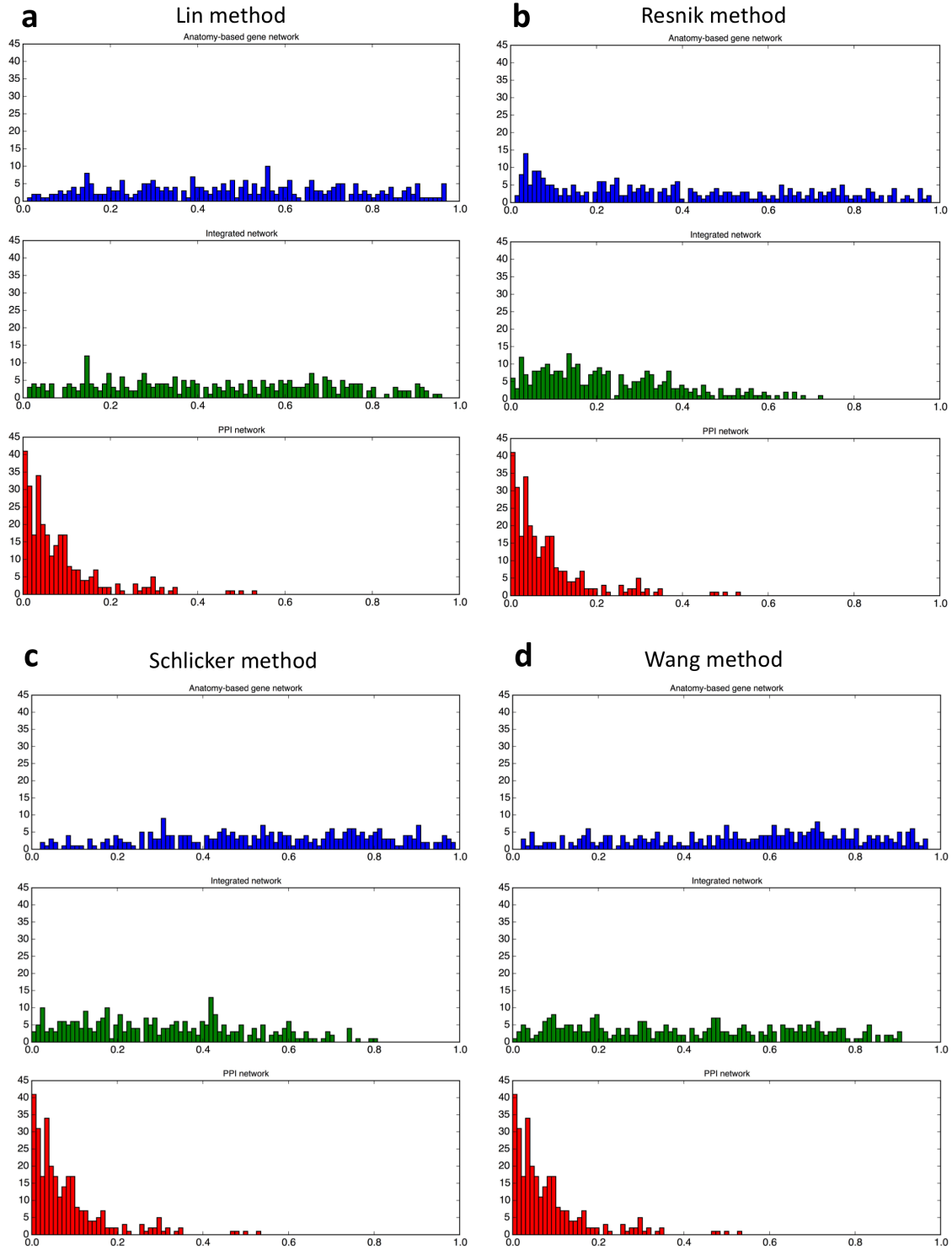




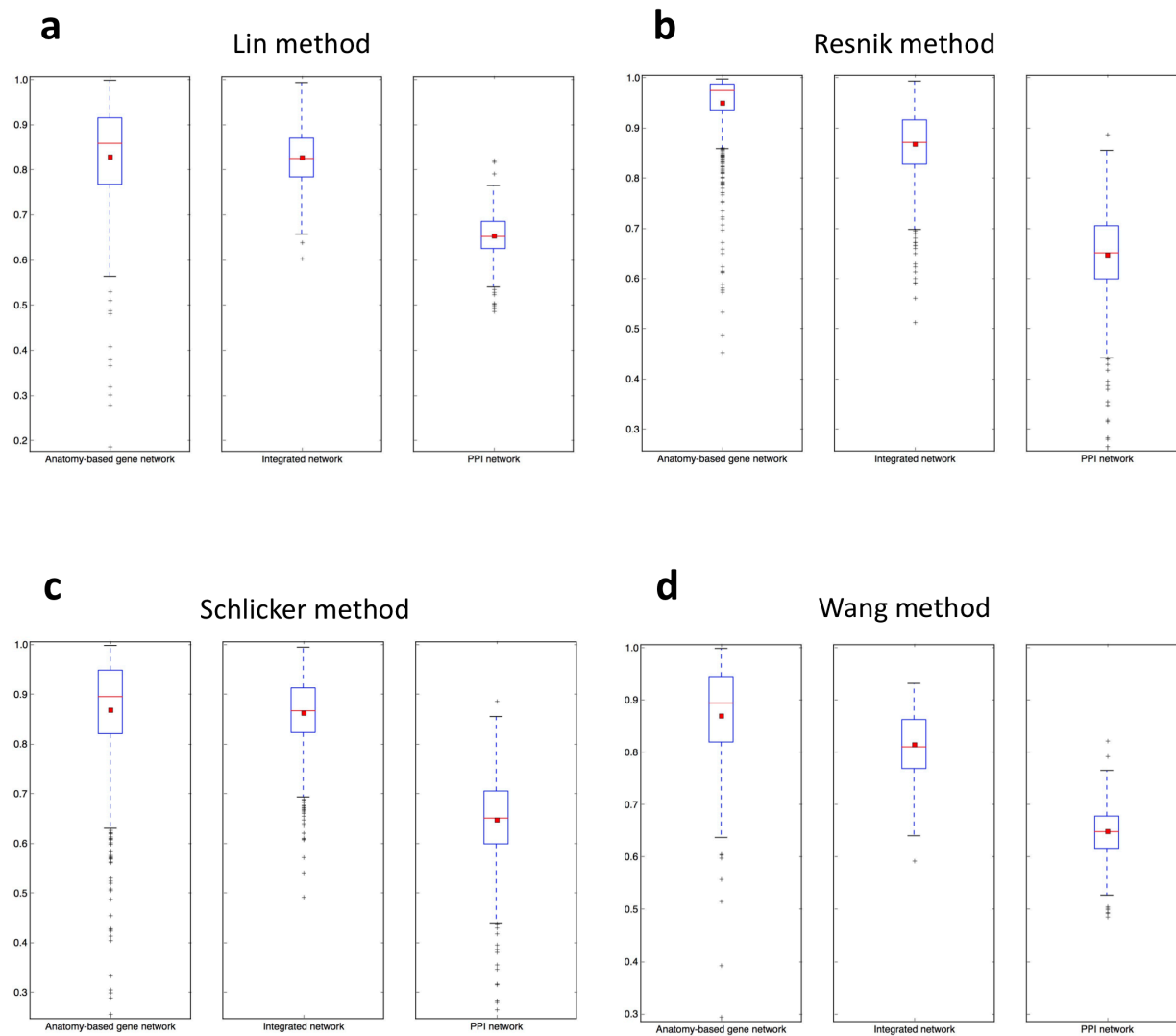
Supplementary Figure S1.7. The boxplot comparisons for the AUC distributions of precision-recall curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.



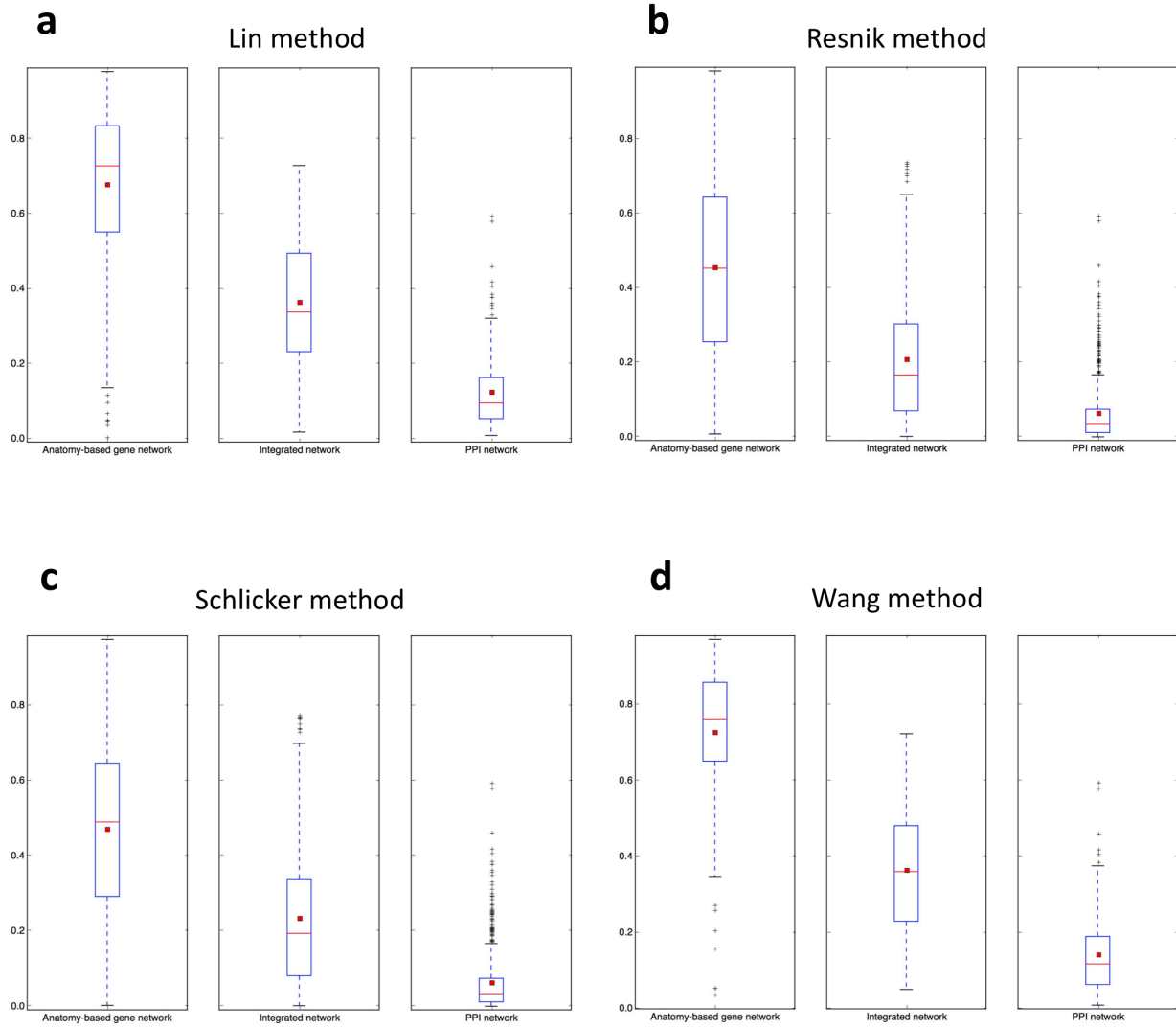
Supplementary Figure S1.8. The histogram comparisons for the AUC distributions of ROC curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish.



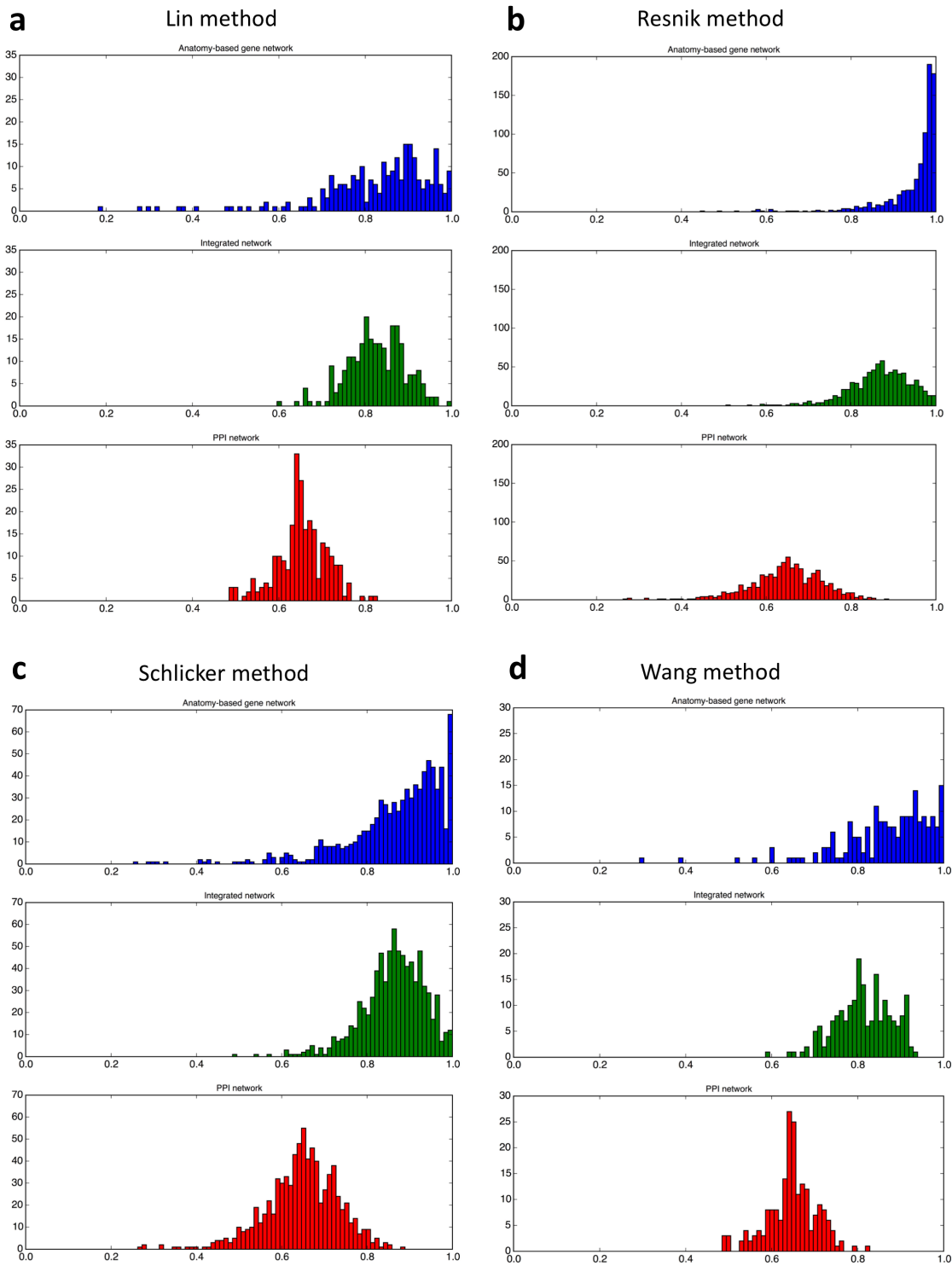
Supplementary Figure S1.9. The histogram comparisons for the AUC distributions of precision-recall curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish.



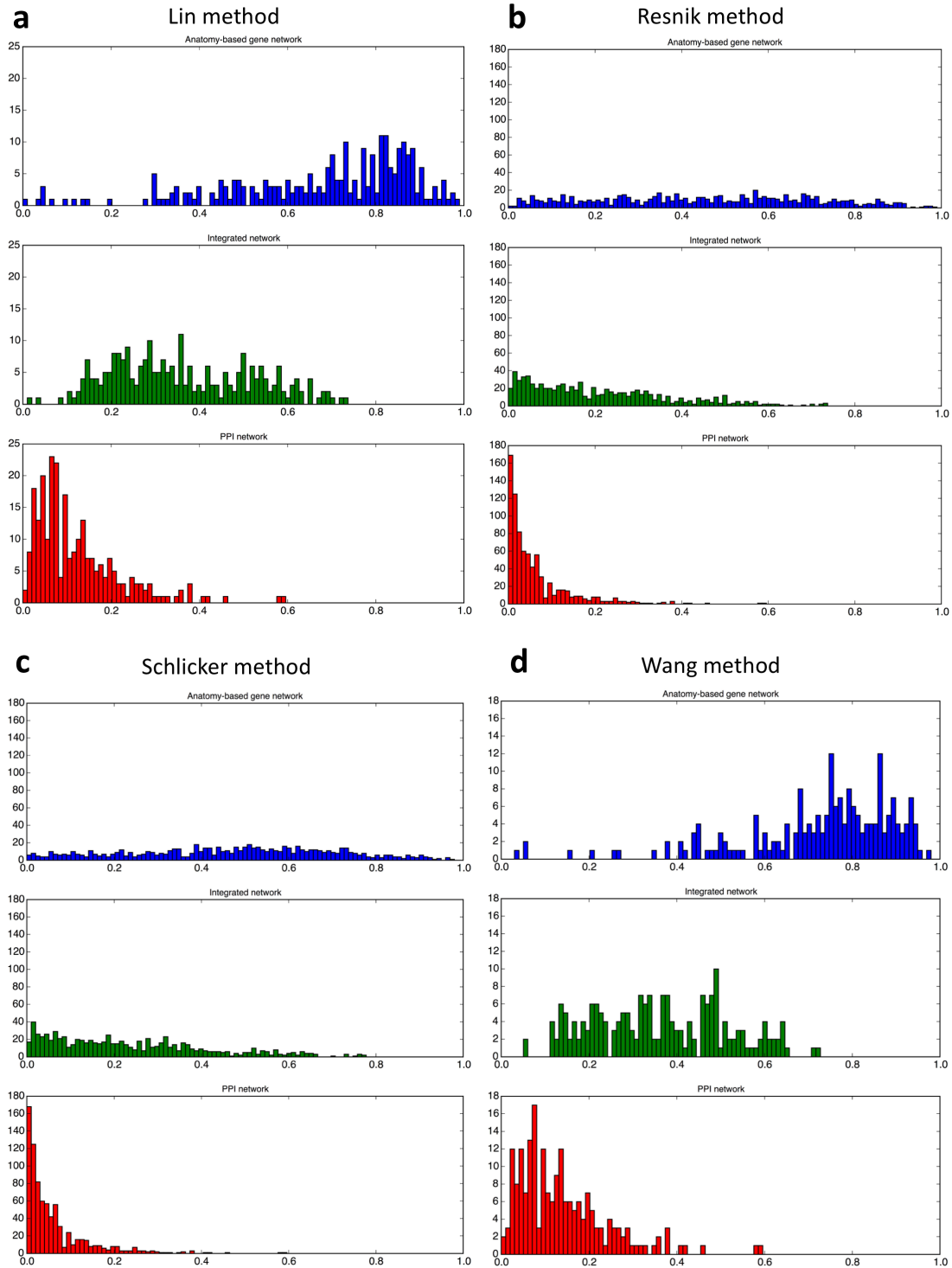
Supplementary Figure S1.10. The boxplot comparisons for the AUC distributions of ROC curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.



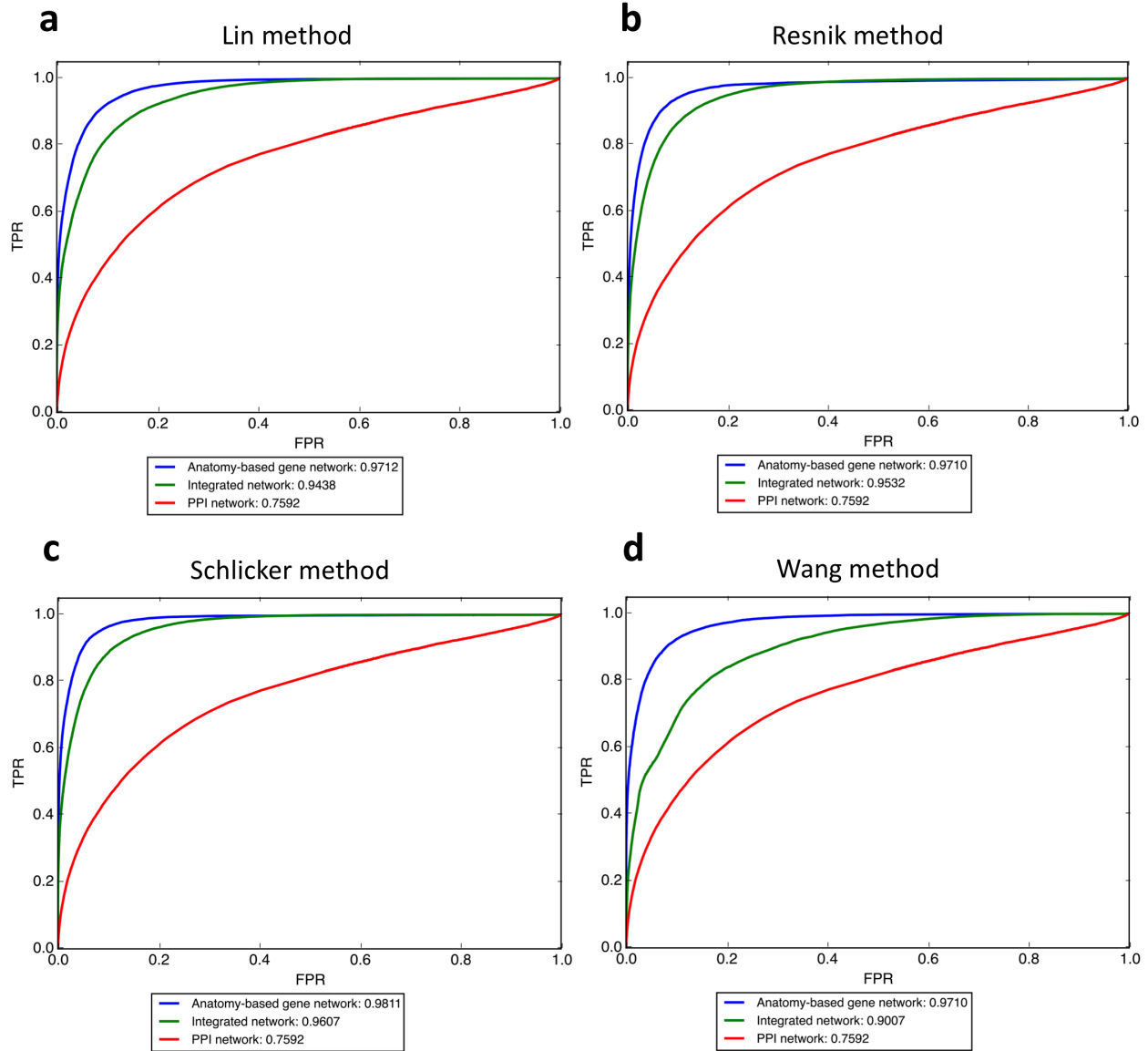
Supplementary Figure S1.11. The boxplot comparisons for the AUC distributions of precision-recall curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.



Supplementary Figure S1.12. The histogram comparisons for the AUC distributions of ROC curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse.

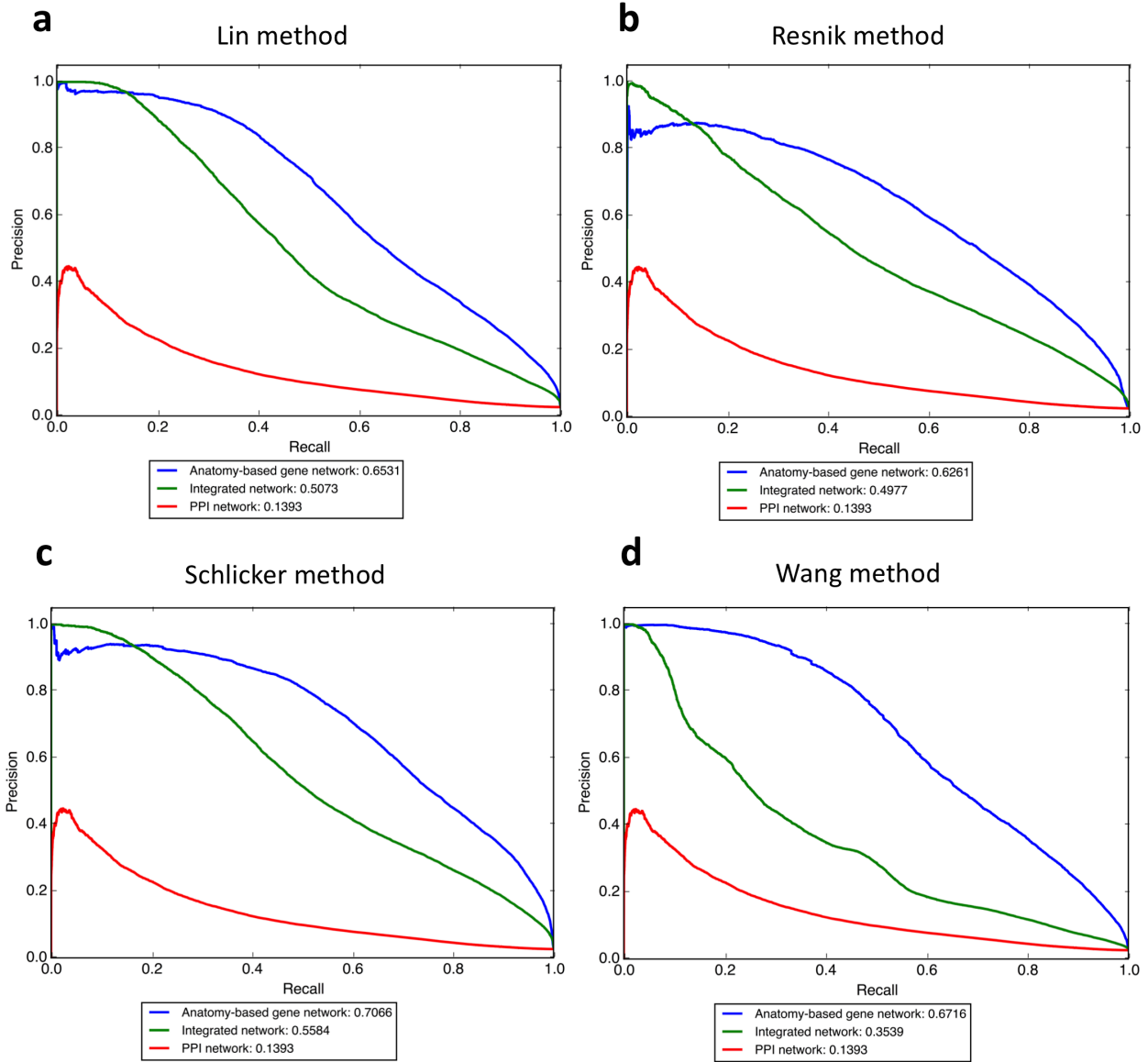


Supplementary Figure S1.13. The histogram comparisons for the AUC distributions of precision-recall curves for filtered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse.

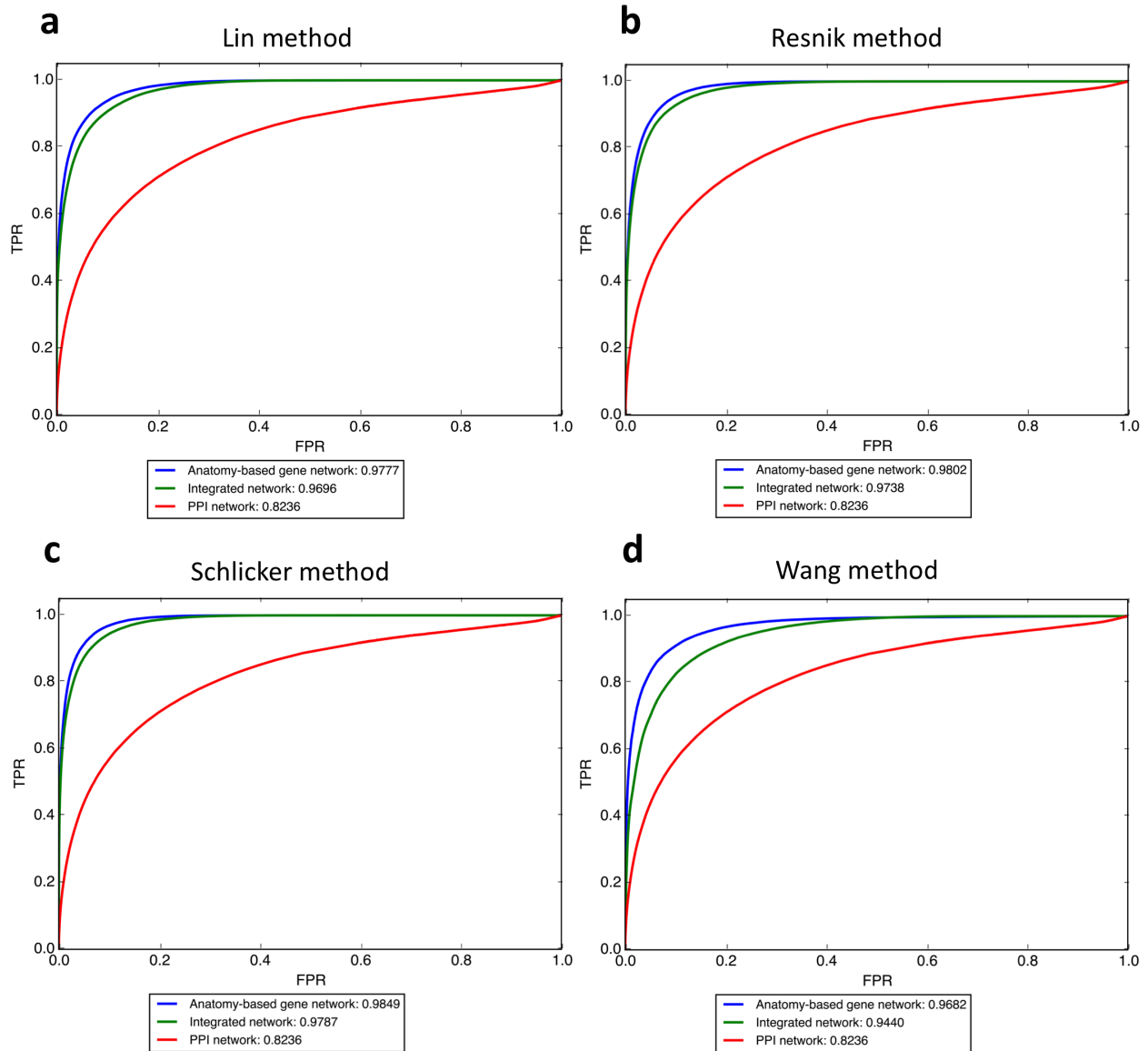


Supplementary Figure S1.14. The comparison of ROC curves for the unfiltered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods for the zebrafish.

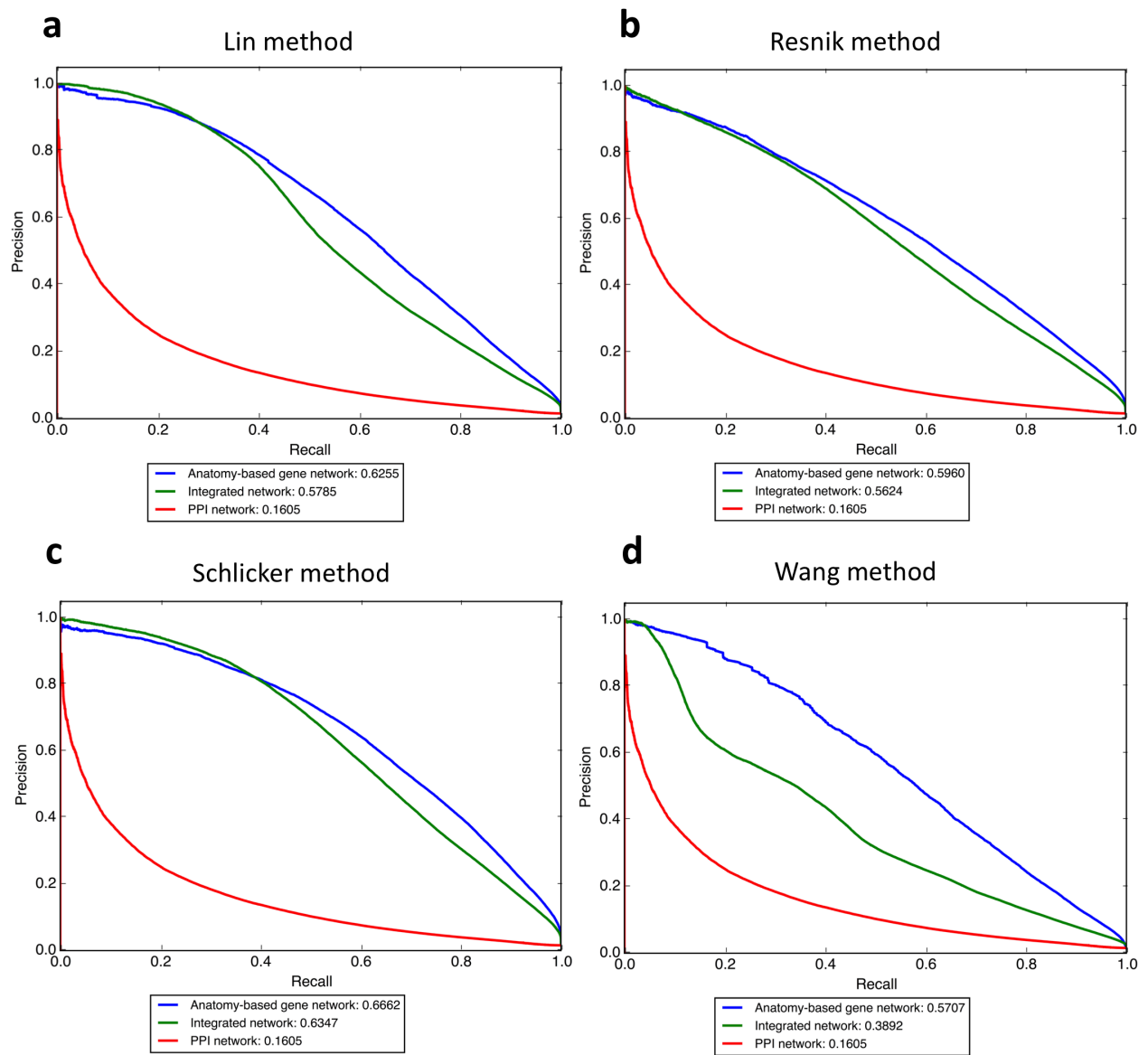




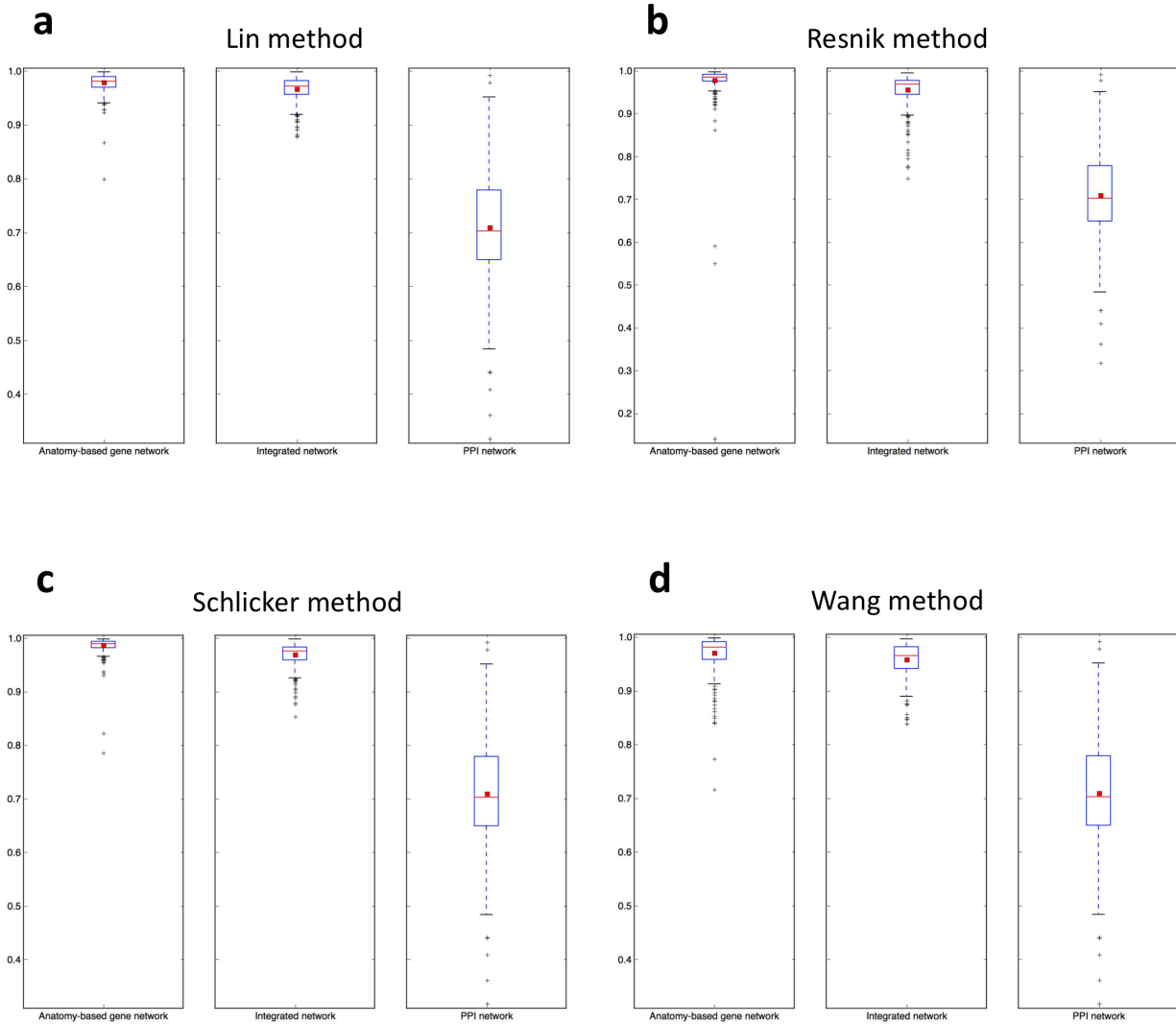
Supplementary Figure S1.15. The comparison of precision-recall curves for the unfiltered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods for the zebrafish.



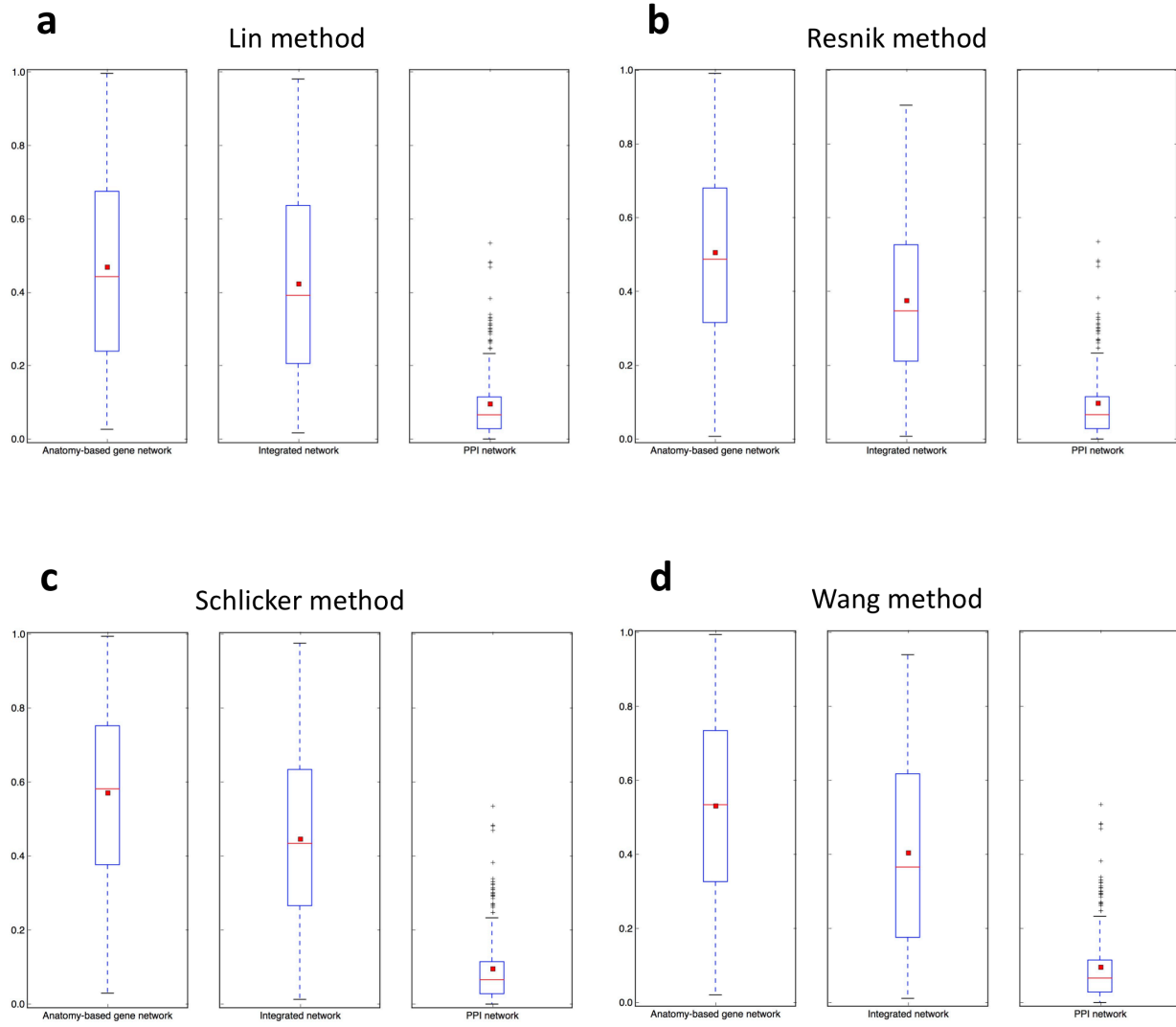
Supplementary Figure S1.16. The comparison of ROC curves for the unfiltered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods for the mouse.



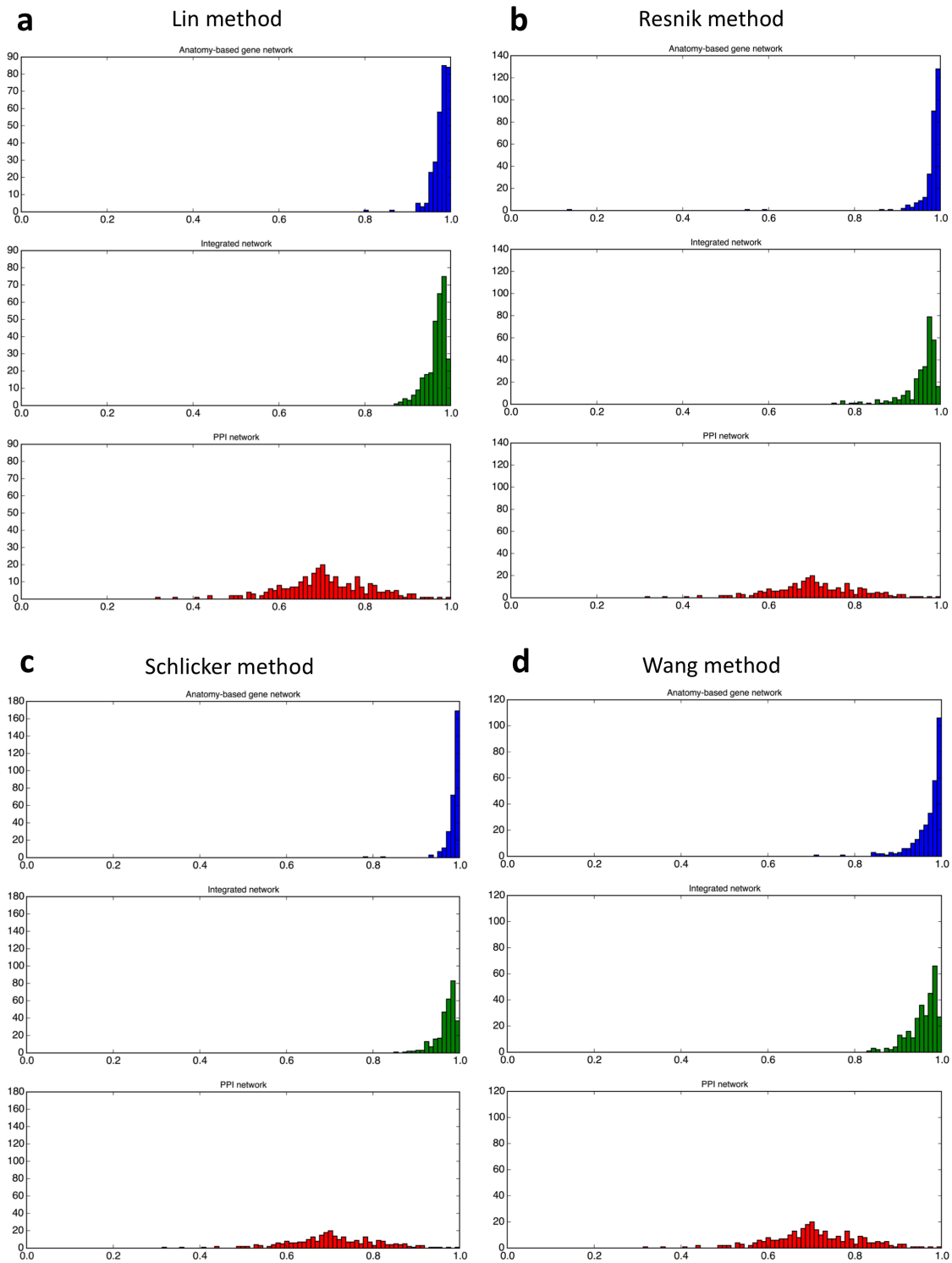
Supplementary Figure S1.17. The comparison of precision-recall curves for the unfiltered integrated networks (green), PPI networks (red), and anatomy-based gene networks (blue) for the four semantic similarity calculation methods for the mouse.



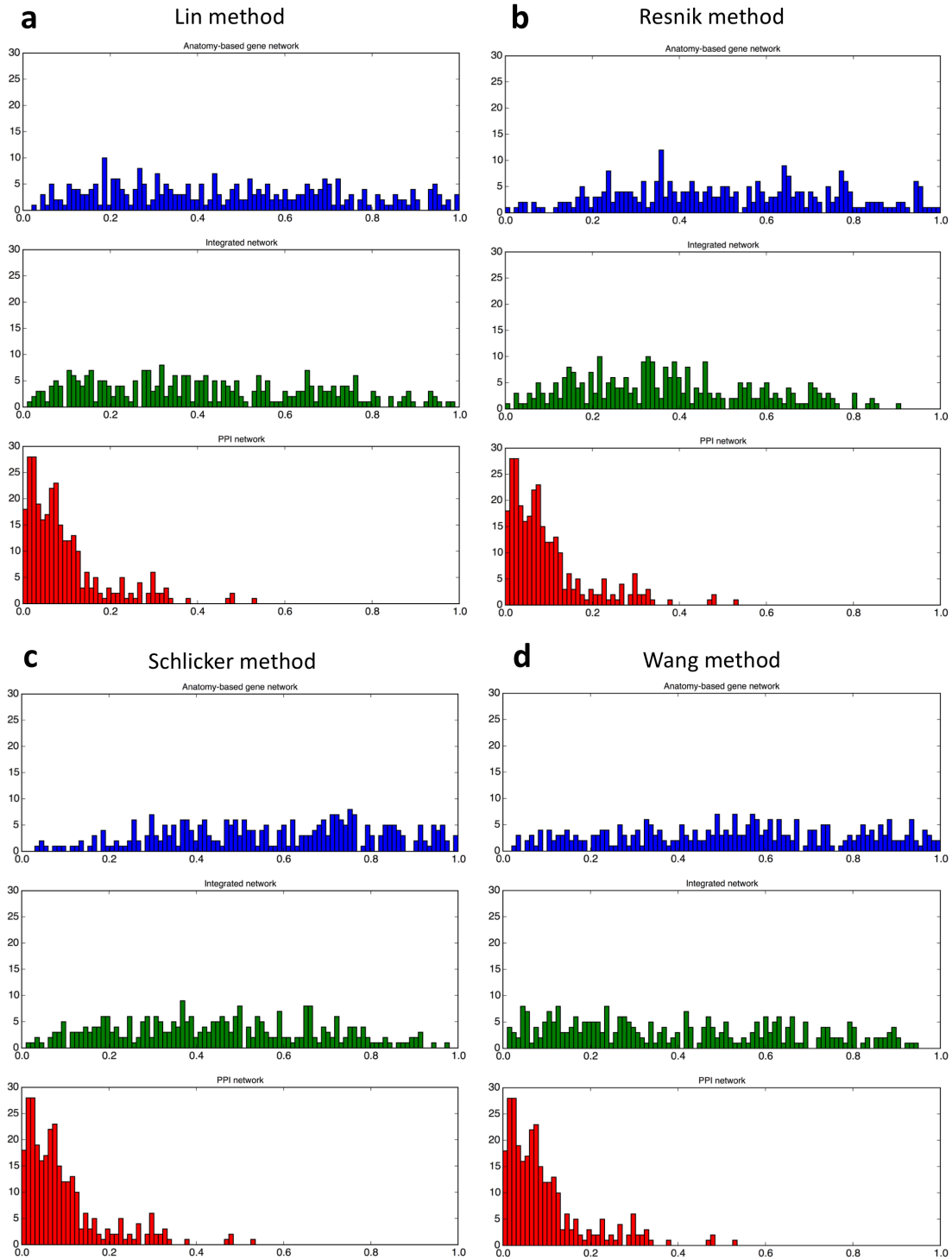
Supplementary Figure S1.18. The boxplot comparisons for the AUC distributions of ROC curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.



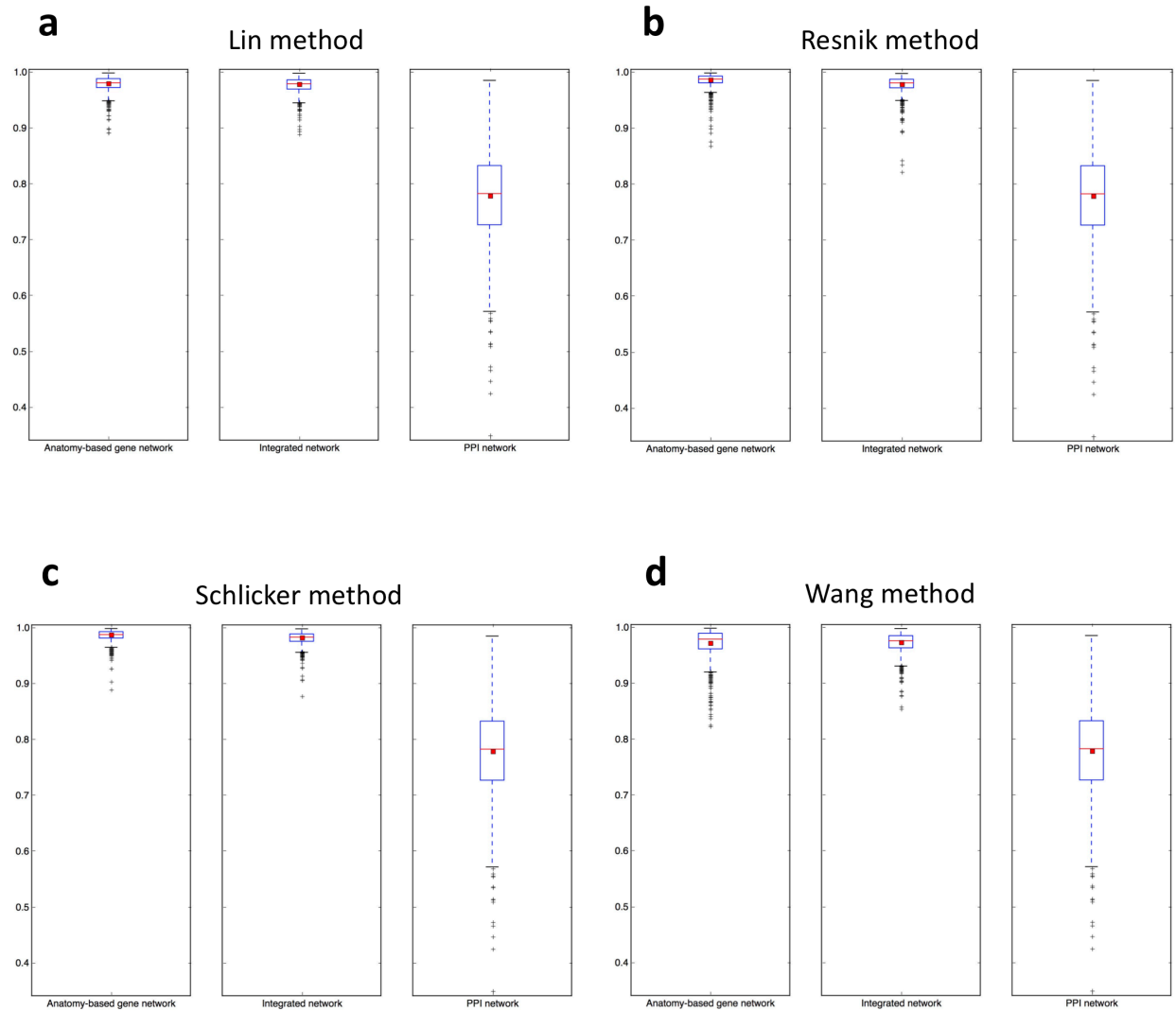
Supplementary Figure S1.19. The boxplot comparisons for the AUC distributions of precision-recall curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.



Supplementary Figure S1.20. The histogram comparisons for the AUC distributions of ROC curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish.

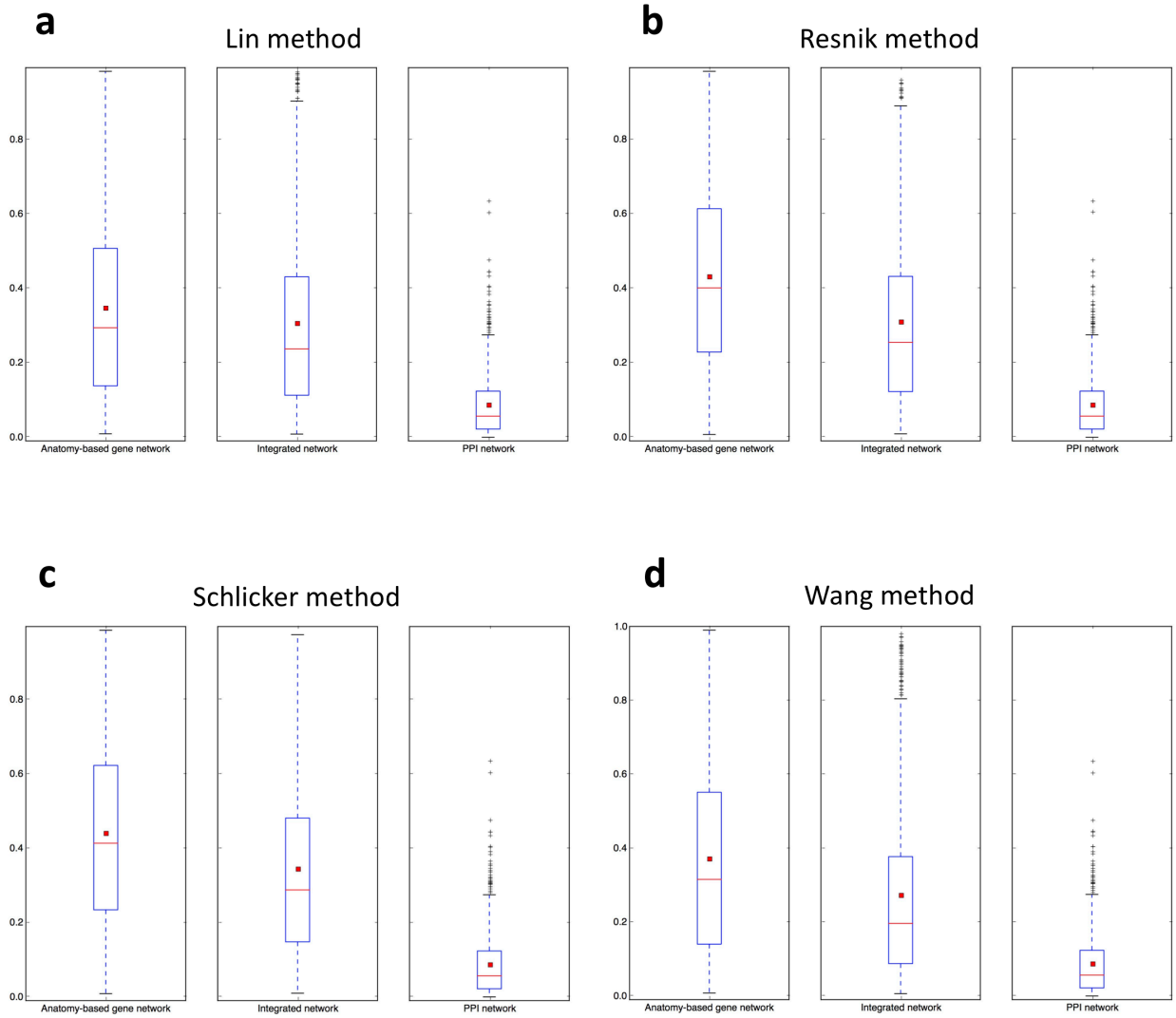


Supplementary Figure S1.21. The histogram comparisons for the AUC distributions of precision-recall curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the zebrafish.

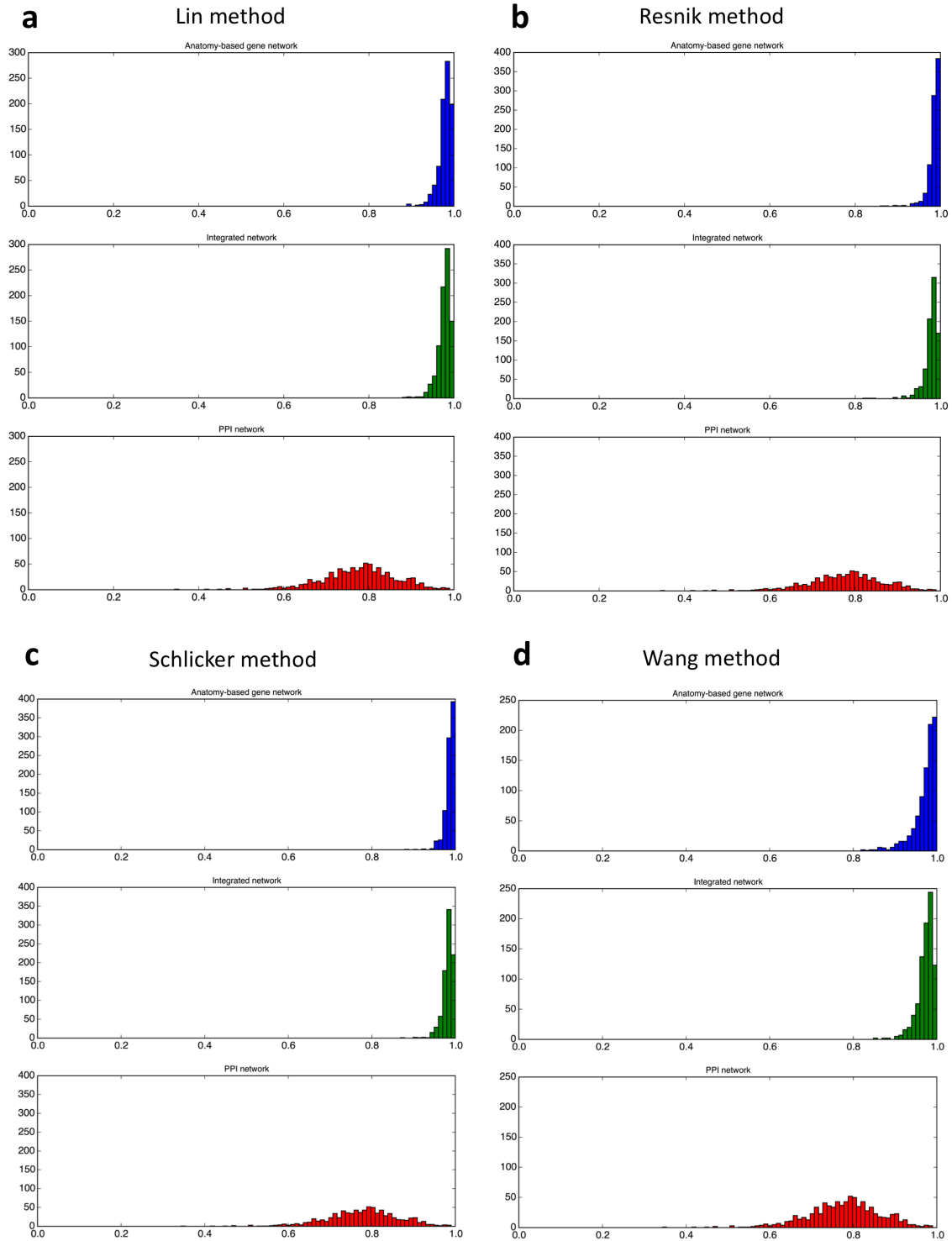


Supplementary Figure S1.22. The boxplot comparisons for the AUC distributions of ROC curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.

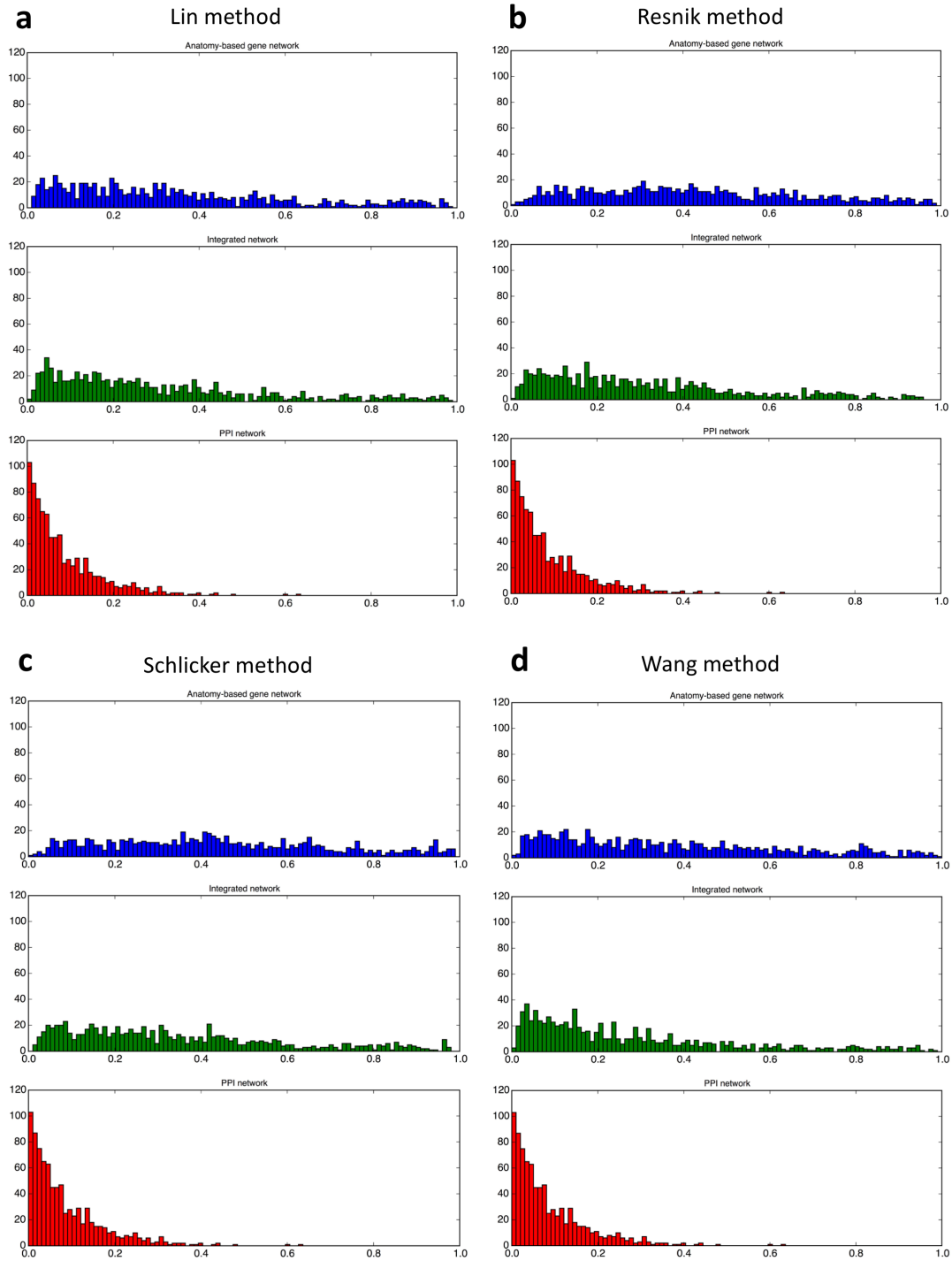




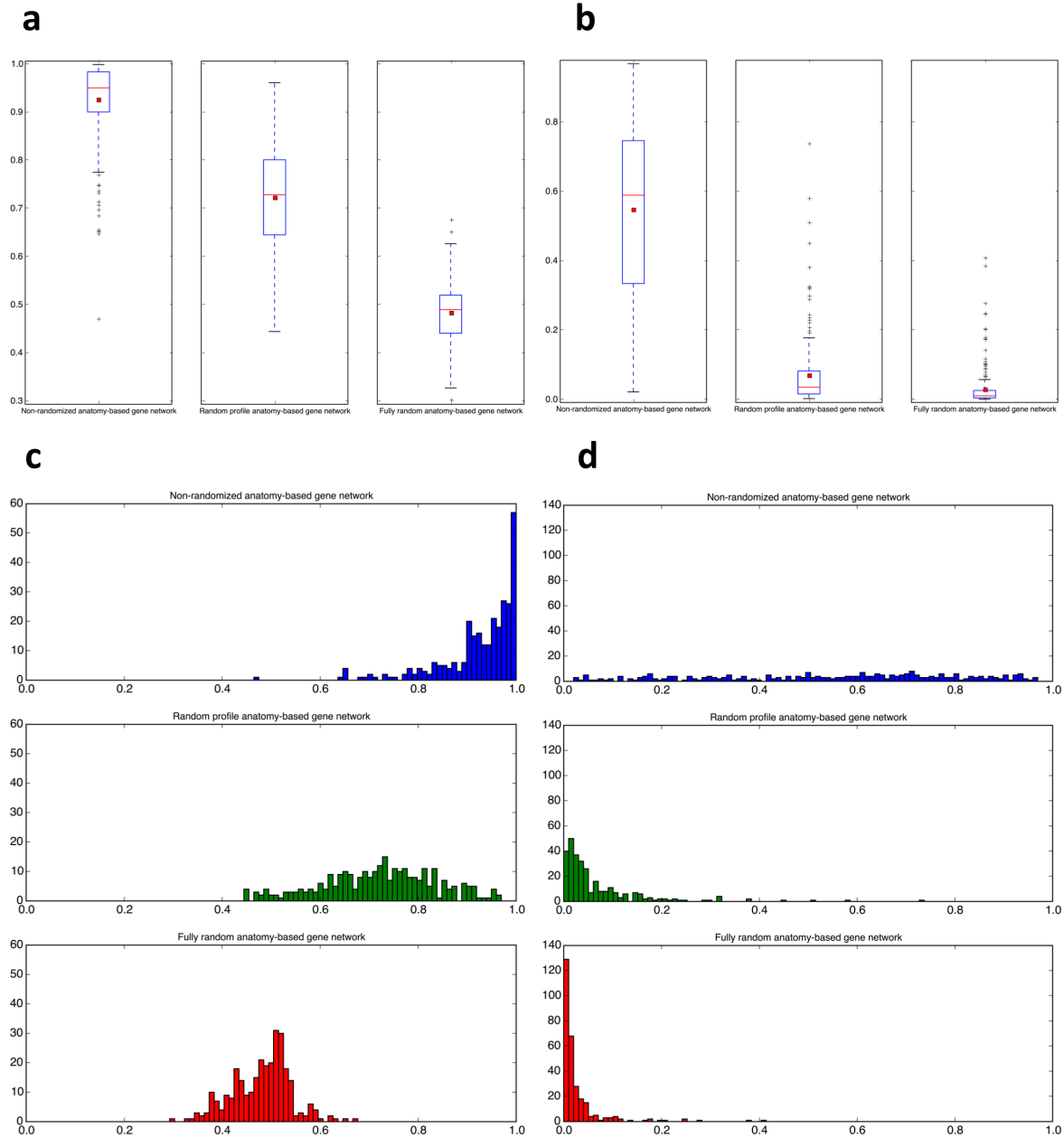
Supplementary Figure S1.23. The boxplot comparisons for the AUC distributions of precision-recall curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse. In the boxplots, the red line and the square represent the median and mean, respectively.



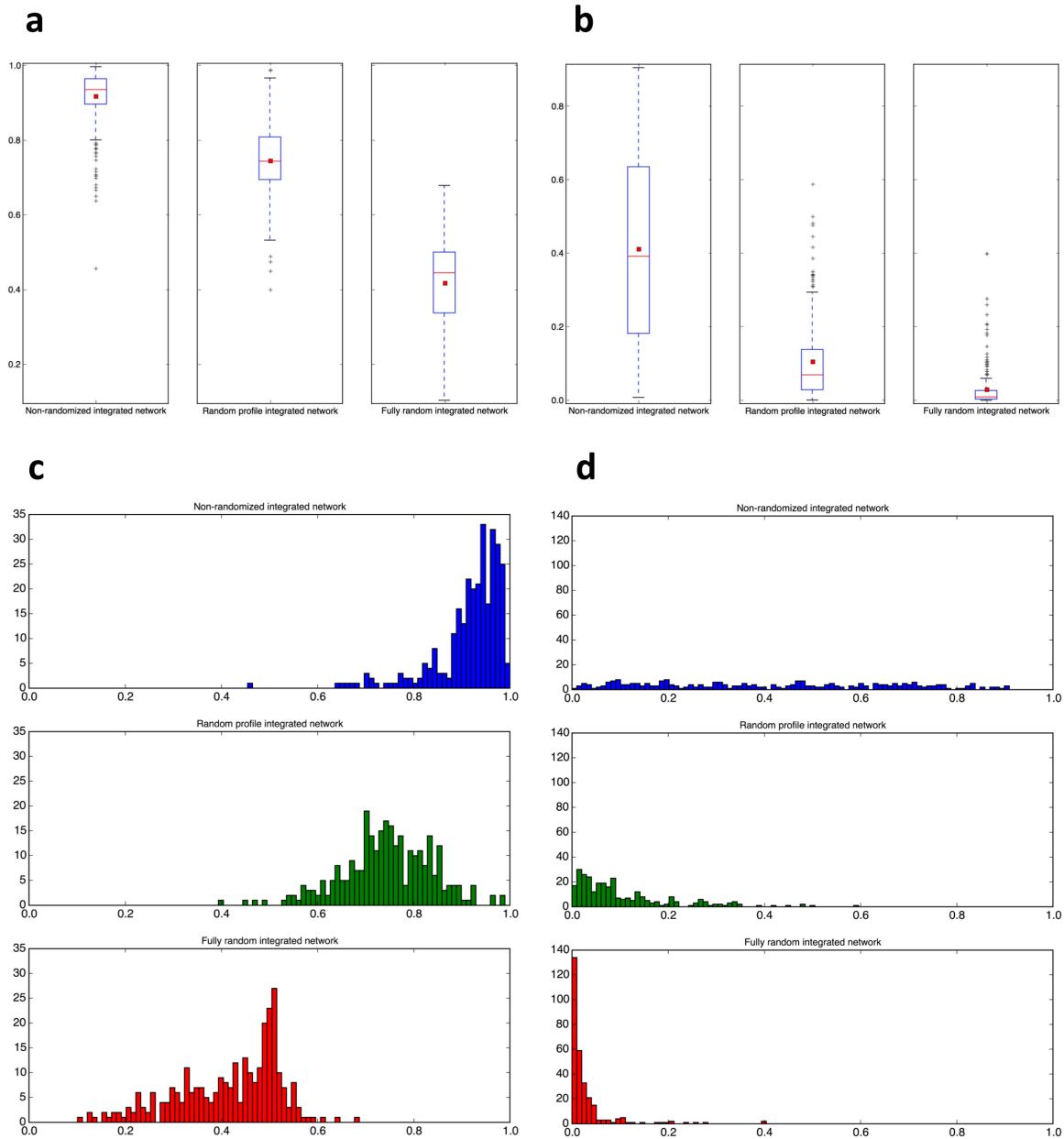
Supplementary Figure S1.24. The histogram comparisons for the AUC distributions of ROC curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse.



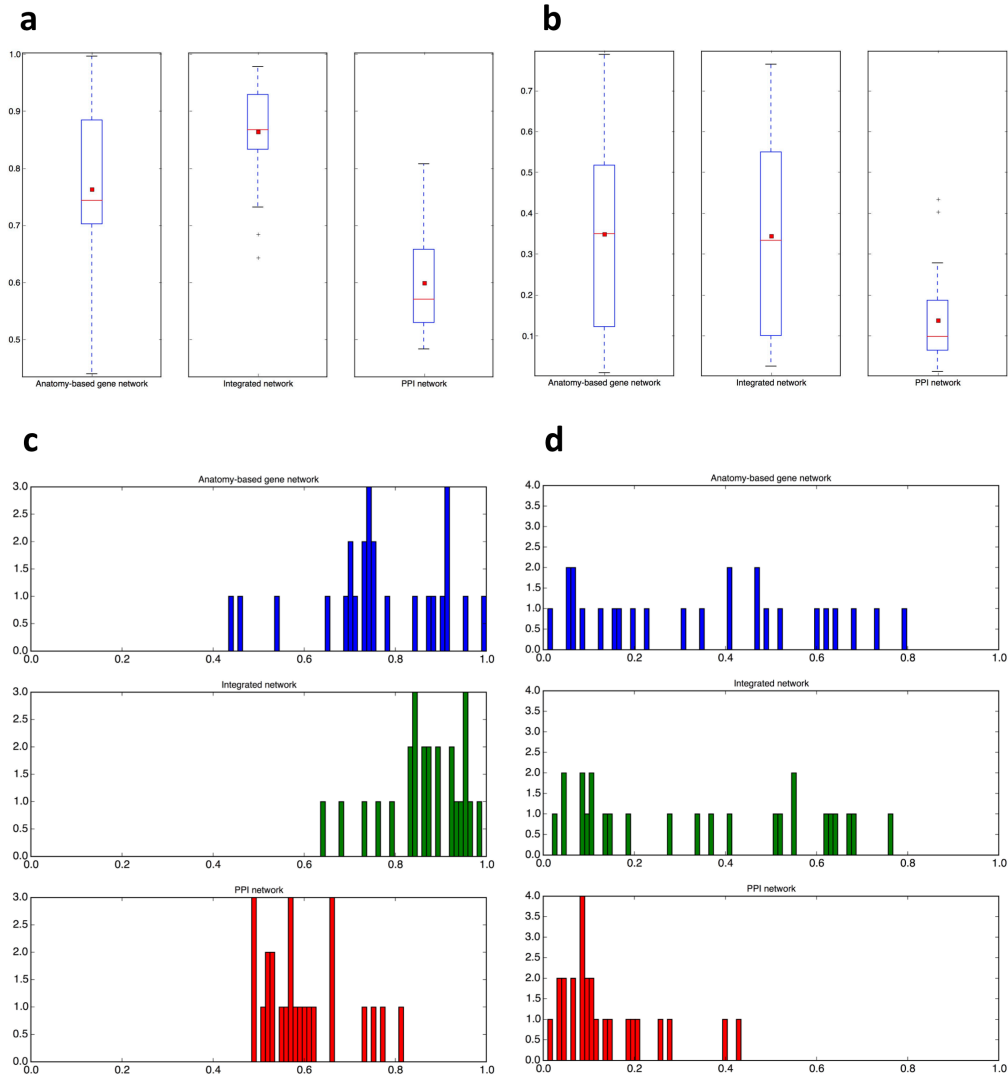
Supplementary Figure S1.25. The histogram comparisons for the AUC distributions of precision-recall curves for unfiltered anatomy-based gene networks, integrated networks, and PPI networks for the four semantic similarity calculation methods for the mouse.



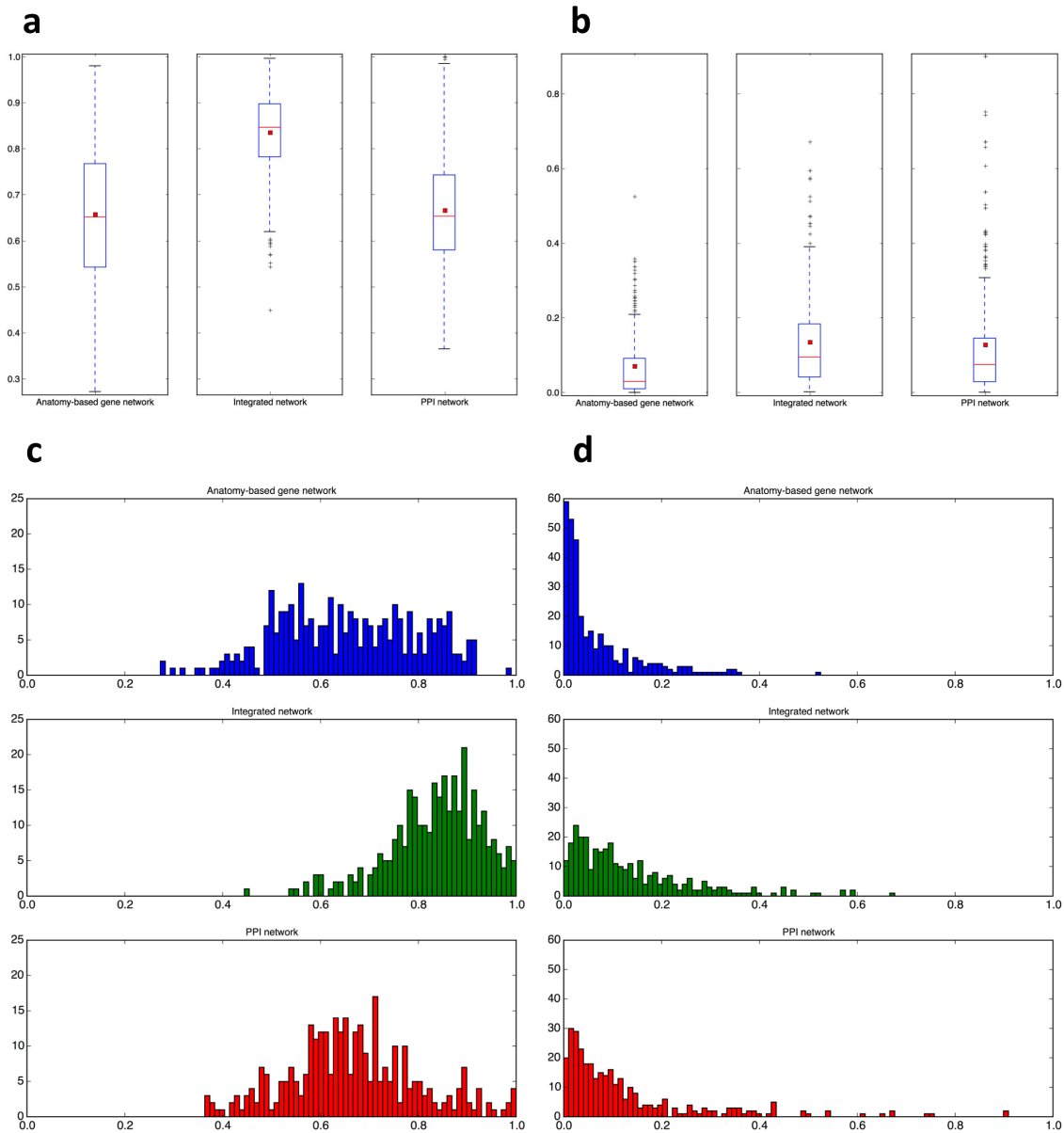
Supplementary Figure S1.26. The boxplot comparisons of the AUC distributions for (a) ROC and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC and (d) precision-recall curves for the filtered non-randomized anatomy-based gene network (blue), random profile anatomy-based gene network (green), and fully random anatomy-based gene network (red) for the Wang method for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.



Supplementary Figure S1.27. The boxplot comparisons of the AUC distributions for (a) ROC and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC and (d) precision-recall curves for the filtered non-randomized integrated network (blue), random profile integrated network (green), and fully random integrated network (red) for the Wang method for the zebrafish. In the boxplots, the red line and the square represent the median and mean, respectively.



Supplementary Figure S1.28. The boxplot comparisons of the AUC distributions for (a) ROC and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC and (d) precision-recall curves for the filtered integrated network (green), PPI network (red), and anatomy-based gene network (blue) for the Wang method for the zebrafish. The integrated network and the anatomy-based gene network were generated using the zebrafish anatomy profile after randomly removing 30 Uberon terms, which had at least 10 gene annotations. The same 30 terms were used for the evaluation to generate the above distributions. In the boxplots, the red line and the square represent the median and mean, respectively.



Supplementary Figure S1.29. The boxplot comparisons of the AUC distributions for (a) ROC and (b) precision-recall curves and the histogram comparisons of the AUC distributions for (c) ROC and (d) precision-recall curves for the filtered integrated network (green), PPI network (red), and anatomy-based gene network (blue) for the Wang method in zebrafish. The networks were evaluated using the annotation profiles containing Biological Process terms of Gene Ontology for the zebrafish genes. In the boxplots, the red line and the square represent the median and mean, respectively.

## **Chapter 2: Study the modular structure of the phenotypic changes using the integrated networks**

### **Abstract**

Evolutionary phenotypic transitions, such as the fin to limb transition, result from changes in associated genes and their interactions. Identifying and analyzing these important changes in gene modules are vital to achieve a better understanding of evolutionary phenotypic changes. When performing network module analyses, the quality of the biological networks and the accuracy of the predictions are important for the conclusions. Therefore, this chapter focuses on using the integrated networks constructed in Chapter 1, which were shown to have a better quality, to detect and compare the gene modules for paired fins (pectoral fin, pelvic fin) and paired limbs (forelimb, hindlimb) to identify modular changes associated with fin to limb transition. The fin to limb transition is the most widely studied evolutionary phenotypic transition, yet current studies focus on one gene or few genes at a time. Therefore, it is the ideal use case to study modular changes at a larger scale for this work. During module detection, candidate genes for each module were predicted. The genes in each module were ranked according to their weighted-degree, and important hub genes were identified. For each comparison, the conserved genes were discovered and compared, and discoveries regarding their role in fin to limb transition were made. Furthermore, the fate of zebrafish module-specific genes in the mouse was investigated and an evolutionary hypothesis was formulated regarding their involvement in newly emerged anatomical structures during the aquatic to terrestrial vertebrate transition. The role of mouse module-specific genes in the zebrafish was investigated to discover evidence for their involvement with anatomical entities that were lost during the aquatic to



terrestrial vertebrate transition. The module analysis results were organized and tabulated in a way that can be used by evolutionary developmental biologists for further investigations. The computational framework developed for this work can be applied to study evolutionary phenotypic transitions involving diverse model organisms and anatomical entities with sufficient data, which is valuable for future evolutionary bioinformatics studies.

## **2.1 Background**

Phenotypes are the result of multiple genes working together in complex biological pathways. Modifications in phenotypes due to environmental changes require rewiring of these gene interactions and their involvement in biological pathways (Defoort, et al., 2018; Robertson and Lovell, 2009; Yamada and Bork, 2009). Most often, it is the interaction of multiple proteins rather than the contribution of a single protein that determines the resulting phenotypes, such as fin development and limb development (Cowen, et al., 2017; Sharan, et al., 2007; Vespignani, 2003). Therefore, investigating the collection of genes and their interactions, i.e., modular gene structure (Cowen, et al., 2017), underlying the phenotypes is important in evolutionary biology to understand the evolutionary mechanisms that drives phenotypic changes. Hence, gene module analysis has become widespread in bioinformatics and the concept of modular evolution has emerged to explain the changes in gene groups rather than focusing on single genes when studying the evolution of organisms (Tang, et al., 2011; Tripathi, et al., 2016).

Protein-protein interaction (PPI) networks are represented as computational graphs, which enables the use of graph theoretic methods to study the network structure. In graph theory, a module is defined as a set of nodes that are highly connected with nodes inside the module and sparsely connected with outside nodes (Cowen, et al., 2017; Gagneur, et al., 2004). These graph

modules usually correspond to biological functions and phenotypes; hence, they are identified as functional modules in the biological vocabulary (Tang, et al., 2011; Ulitsky and Shamir, 2007).

There are several module detection algorithms that can be used to detect modules in a graph (Pereira-Leal, et al., 2004; Tripathi, et al., 2016). Some methods, such as graph partitioning (Kernighan and Lin, 1970), partitional clustering (Lloyd, 1982), and clique percolation (Zhang, et al., 2006) do not require any prior information about known gene functions or phenotypes; they only consider the network structure. These methods are more suited for identifying smaller protein complexes associated with molecular functions or biological processes. Complex phenotypes such as pectoral fin development, involve large number of genes (over 100). Thus, there are computational limitations for using the aforementioned network structure-based module detection methods. In such situations, the genes that are known to be associated with the phenotype can be used as prior information to detect the modules (Cowen, et al., 2017). These methods start from the set of known genes for the given phenotype and expand the module based on network structure. For example, one of the simplest ways to isolate a functional module by expansion is to assume all the immediate neighbors of the genes with the known phenotype are included in the module (Cowen, et al., 2017). However, this method has proven to yield a lot of false positives as not all immediate neighbors of the gene of interest should have the same phenotype (Cowen, et al., 2017). Therefore, network-based candidate gene prediction algorithms, such as Hishigaki method (Hishigaki, et al., 2001) and label propagation algorithm (Gregory, 2010; Liu and Murata, 2010) are often used to predict the most likely candidates that should be included in a module, which have been shown to be more accurate (Cowen, et al., 2017; Hishigaki, et al., 2001; Sharan, et al., 2007).

One of the advantages of performing module analysis is the ability to identify hub genes. Hub genes are defined as important genes that are central to the stability of the module (Liu, et al., 2016; Taylor, et al., 2009). They have a higher number of interactions with other module members, and their removal is most likely to destroy the module organization, and thus the biological function or the phenotype that is governed by the module. By performing network analysis, a simple set of genes for a function or a phenotype can be transformed into a ranked list that is sorted by their importance in the module. Usually, the number of interactions of a gene within the module (degree) is used for the ranking (Tang, et al., 2014).

When studying the evolution of gene modules corresponding to the evolution of species, most studies focus on identifying the genes that are retained during evolution, i.e., conserved genes, and their organization in the respective modules (Kelley, et al., 2003; Sharan, et al., 2005; Shui and Cho, 2016; Wuchty, et al., 2003). It is hypothesized that gradual modular changes are associated with evolution while maintaining the basic modular structure (Liu, et al., 2011; Vespignani, 2003) because dramatic changes in gene interactions may destroy the proper function of an organism. To support this hypothesis, the conserved genes are observed to play an important role in maintaining the stability of the gene modules during species evolution (Shui and Cho, 2016; Vespignani, 2003; Wuchty, et al., 2003). The recruitment and the removal of other genes and the rewiring of biological pathways are often held together by the conserved genes. Performing module analysis allows identification of these important conserved genes, which are often identified as hub genes (Vespignani, 2003; Wuchty, et al., 2003).

This work focuses not only on identifying the conserved genes, but also on the module-specific genes that have been recruited or removed during the evolution. Although conserved genes are typically considered as important evolutionary hub genes, there can be exceptions to

this observation (Liang, et al., 2006). Therefore, it is important to identify the role of module-specific genes in the respective modules and their evolutionary fate if they are removed from the module.

The use case for this work is the fin to limb transition, which is an important phenotypic change associated with the aquatic to terrestrial vertebrate transformation (Amaral and Schneider, 2018; Clack, 2012; Shubin, 2008). According to fossil record, the transformation of lobe-finned fishes into land vertebrates occurred in the Devonian, 365-408 million years ago (Shubin, 2008; Sordino and Duboule, 1996). It is believed that Devonian fishes belonging to the Panderichthyidae are the closest relatives of earliest land vertebrates such as *Ichthyostega* and *Acanthostega* (Vorobyeva and Schultze, 1991). This transformation is associated with many phenotypic changes including the fin to limb transition and changes in the cranial and axial skeleton (Clack, 2012; Shubin, 2008). The relationship between homologous anatomical structures of land and aquatic vertebrates is evident by several similar characteristics. For instance, the pectoral fin endoskeleton of panderichthyid fishes shows significant similarities with limbs, such as the presence of a proximal humerus and two distal bones (Amaral and Schneider, 2018; Sordino and Duboule, 1996). This indicates that forelimbs and hindlimbs of tetrapods are homologous to pectoral and pelvic fins of fish, respectively.

Identifying the genetic changes associated with the fin to limb transition is a prominent topic of ongoing studies in evolutionary biology. Many wet lab experiments have been performed showing the evolutionary importance of genes such as *shh* and *hoxd* (Akimenko and Ekker, 1995; Freitas, et al., 2012; Zhang, et al., 2010). Few computational studies, however, have been targeted on the fin to limb transition, including studies such as Onimaru, et al. (2016) and Sheth, et al. (2012). The recent availability of large PPI networks and the ability to perform

module analysis through the advancement of novel network algorithms provide an opportunity for a new perspective on the genetic changes associated with the fin to limb transition, and this is the motivation for this study. The quality of PPI network data has been a major issue in previous network analyses. Thus, the improved quality of the PPI networks that are integrated with anatomy ontology data (Chapter 1) is the key to achieve a higher accuracy for the module analyses performed here.

In this work, new gene candidates that did not contain original annotations to paired fins or paired limbs are predicted for each module and conserved genes are identified to understand their role in the modular evolution. Moreover, fin and limb module-specific genes are identified, and their evolutionary roles are investigated. This work suggests some evolutionary hypotheses regarding the role of conserved and module-specific genes in fin to limb transition. Finally, this study provides a general computational platform to perform gene module comparisons for any phenotypic transition given sufficient data for the analysis.

## **2.2 Methods**

### *2.2.1 Selection of the integrated networks for module detection*

Based on the comparison of network-based candidate gene prediction performances for the filtered integrated networks (sections 1.2.4 and 1.2.5) that were generated by four semantic similarity methods (Lin, Resnik, Schlicker, and Wang), the best performing networks were selected for the zebrafish and the mouse. Because these networks have the highest performance for each model organism, the detected modules will be of the highest accuracy. These networks were filtered based on a gene similarity score cutoff (section 1.2.5); therefore, only the gene interactions with a high reliability were retained in the networks.

### 2.2.2 Detection of network modules

The pectoral fin (UBERON:0000151) module identified from the zebrafish integrated network was compared with the forelimb (UBERON:0002102) module from the mouse, and the pelvic fin (UBERON:0000152) module from the zebrafish was compared with the hindlimb (UBERON:0002103) module from the mouse. For module detection, genes with direct annotations to aforementioned anatomical entities were extracted from the anatomical profiles that were constructed using data from the Monarch Initiative repository (<https://monarchinitiative.org/>) (section 1.2.2). In the anatomical profiles, some genes are directly annotated to aforementioned anatomical entities, and others are annotated to parts of the entities. The genes that were annotated to the parts but did not have direct annotations to the parent term were extracted using a Python script that uses the Uberon anatomy ontology relationships. These were added to the list of genes with direct annotations for the particular anatomical entity. The genes that were annotated to developmental precursors (buds of the pectoral fin, pelvic fin, forelimb, and hindlimb) but did not have direct annotations to the parent term were also extracted and added to the gene lists. The genes that were directly annotated to the particular anatomical entity or to a part or a developmental precursor are collectively identified as ‘genes with original annotations’ herein. Then, the network-based candidate gene prediction performance for each anatomical entity was evaluated, and ROC and precision-recall curves were generated for each (section 1.2.7).

For predicting the candidate genes, the Hishigaki network-based candidate gene prediction method (Hishigaki, et al., 2001) was used (section 1.2.6), and the precision values for each anatomical entity generated during the evaluation were used to select a precision threshold

for the candidate gene prediction. The precision (equation 1.17 in section 1.2.7) is identified as a suitable evaluation metric for network-based candidate gene predictions (Sharan, et al., 2007) because it only uses the known gene annotations (positives) for a specific function/anatomical entity. Furthermore, it is suited for unbalanced samples such as gene functions where the number of genes annotated to a function/anatomical entity (positives) is low compared to the remaining genes without annotations (negatives). The trial and error method was used to select the best precision threshold for each module focusing on the resulting module size. When comparing the pectoral fin module with the forelimb module and the pelvic fin module with the hindlimb module, it was assumed that the equivalent modules are approximately similar in size (number of genes) to perform an efficient comparison. Finally, after predicting the candidate genes, the modules for the pectoral fin and the pelvic fin were extracted from the zebrafish integrated network and the modules for the forelimb and the hindlimb were extracted from the mouse integrated network. Some of the genes with direct annotations to the term, a part, or the developmental precursor could not be extracted due to two reasons: (1) the gene was lost due to the application of the network cutoff and (2) the gene did not have any interaction with other genes in the module, i.e., it was isolated.

The extracted modules were visualized using the Cytoscape software (Shannon, et al., 2003; Smoot, et al., 2011). For the visualizations, different colors were assigned to distinguish predicted genes, genes with direct annotations to the anatomical entity, and genes that were only annotated to the parts or the developmental precursor of the anatomical entity.

### *2.2.3 Assessing the biological significance of the modules*

It is important to assess whether the module genes for each anatomical entity (pectoral fin, pelvic fin, forelimb, and hindlimb) are clustered together in the network and forming a cohesive module or just randomly scattered throughout the network. Genes in a module (module genes) must be clustered together compared to genes that were randomly picked from the rest of the genes in the network, i.e., the network background (Cowen, et al., 2017). An algorithm was developed to count the number of module genes in the immediate neighborhood of each gene in the network. Then, the distributions of these counts were compared between the module genes and the network background genes. If the module genes are densely clustered together, they should have a higher number of module genes in their neighborhoods compared to the network background genes. Therefore, the module gene count distribution should be higher for the module genes compared to the network background genes.

### *2.2.4 Comparison of the network modules*

Relevant to the fin to limb transition use case as previously described, the pectoral fin module of the zebrafish was compared with the forelimb module of the mouse, and the pelvic fin module of the zebrafish was compared with the hindlimb module of the mouse to identify the modular changes.

Teleosts, such as the zebrafish, have more genes compared to tetrapods, such as the mouse, and a whole genome duplication event is proposed to have occurred at the origin of actinopterygian fishes, i.e., the teleost genome duplication (Braasch, et al., 2014; Meyer and Schartl, 1999). Therefore, most of the mouse genes have duplicated copies in the zebrafish. To perform the module comparison, the orthology gene mappings between mouse and zebrafish



genes were downloaded (06/26/2018) from the ZFIN (Bradford, et al., 2011; Westerfield, et al., 1998): the zebrafish model organism database (<https://zfin.org/downloads>). During the comparison, if multiple zebrafish orthologs were present in a zebrafish module for a single mouse gene, all zebrafish orthologs were retained for that mouse gene. There can be multiple mouse orthologs for a single zebrafish gene, but for the selected four modules, there were no such genes. By performing the module comparison, conserved genes (genes that are common to the two modules), zebrafish module-specific genes, and mouse module-specific genes were identified.

In PPI network analysis, the degree of a node/gene (the number of interactions for the particular node/gene) is often used as an important metric (Aittokallio and Schwikowski, 2006; Jeong and Chen, 2013). Genes with higher degrees in a module, i.e., hub genes, are considered as more important because they have more interactions with other module genes and removal of such a gene from the module may significantly affect the integrity of the module (Taylor, et al., 2009). When analyzing networks with weights assigned for interactions (weighted networks), weighted degree is preferred over the degree because it considers the different interaction weights rather than just counting the number of interactions for a specific node (Rubinov and Sporns, 2010; Tang, et al., 2014). The equation for weighted degree calculation (equation 2.1) is given below.

$$\textit{Weighted degree} = \sum_{v \in n(u)} \textit{sim}(v, u) \quad (2.1)$$

In the equation 2.1,  $n(u)$  is the neighborhood of the gene of interest ( $u$ ) and  $v$  iterates through all the neighbors of gene  $u$  and obtains the summation of each interaction weight (gene similarity score) for the interaction between genes  $v$  and  $u$ , which is represented by  $\textit{sim}(v,u)$ .

The weighted degree for each gene in a module was calculated and the genes were ranked accordingly. The weighted degree of each zebrafish module gene was also compared with the corresponding mouse ortholog. However, due to the size differences of the zebrafish and mouse modules, the weighted degree of each gene was normalized by dividing by the total number of genes in each module. Calculating the weighted degrees for module genes enables to compare the weighted degree distributions between selected gene portions, such as conserved genes, zebrafish module-specific genes, and mouse module-specific genes. Boxplots were generated for the comparison of normalized weighted degree distributions for conserved genes, zebrafish module-specific genes, and mouse module-specific genes for the comparison of pectoral fin *versus* forelimb and pelvic fin *versus* hindlimb modules.

Enrichment analyses were performed using the biological process component of the Gene Ontology (GO-BP) and Uberon anatomy ontology (section 2.2.6) for each group (conserved genes, zebrafish module-specific genes, and mouse module-specific genes) for the module comparisons. This enabled the identification of the enriched biological processes and anatomical entities for the module-specific genes *versus* conserved genes. The fate of the zebrafish module-specific genes in mouse was investigated by extracting mouse orthologs for the pectoral fin and pelvic fin module-specific genes and performing enrichment analyses using Uberon and GO-BP terms. The fate of the mouse module-specific genes in zebrafish was investigated by extracting zebrafish orthologs for the forelimb and hindlimb module-specific genes and performing enrichment analyses using Uberon and GO-BP terms.

### 2.2.5 Validating the predicted genes

Detected modules contain a portion of predicted genes that can be potential candidates for the given anatomical entity (section 2.2.2). These predicted genes can be validated using either experimental methods, such as gene knockdown (Huang, et al., 2013) or computational methods that are the focus of this dissertation. First, an enrichment analysis can be performed to confirm whether the predicted genes share similar ontology terms (e.g., biological process and anatomy ontology terms) as the genes with original annotations to the anatomical entity of interest. Here, enrichment analyses were performed on the predicted genes *versus* the genes with original annotations using the GO-BP terms and the Uberon anatomy ontology terms (section 2.2.6). The enriched terms were then compared to detect common terms enriched in the predicted gene set.

A second way to validate a predicted gene is to compare it with orthologs from another organism to determine whether it has retained the gene function or annotated to a homologous anatomical entity during the evolution of species. Here, the predicted genes for the pectoral fin and pelvic fin modules in zebrafish were compared with the orthologous genes in the forelimb and hindlimb modules in mouse, respectively, and the predicted genes for the forelimb and hindlimb modules in mouse were compared with orthologous genes in the pectoral fin and pelvic fin modules in zebrafish.

As the third way to further confirm the importance of the predicted genes in each module, the weighted degree distributions for the predicted genes were compared with the weighted degree distributions for the genes with original annotations for each module. If the predicted genes have a higher weighted degree distribution, it can be concluded that the predicted genes have a similar or higher importance as known genes that contribute to the anatomical phenotype.

### *2.2.6 Functional enrichment analysis*

Functional enrichment analysis is used to identify ontology terms that are enriched for a given set of genes (Huang, et al., 2009; Kuleshov, et al., 2016). Gene Ontology (GO) terms are commonly used for this purpose. For this work, the DAVID (<https://david.ncifcrf.gov/>) online functional enrichment analysis tool was used to perform gene set enrichment analysis using GO-BP component. DAVID uses Fisher's exact test (Routledge, 2005) to perform enrichment analyses. Although the GO is widely used for enrichment analysis, anatomy ontologies are rarely used for functional enrichment analysis. To my knowledge, the only published tool that uses anatomy ontology to perform enrichment analysis is the Plant Ontology Enrichment Analysis Server (POEAS) (Shameer, et al., 2014). To perform enrichment analysis using Uberon anatomy ontology, a Python program (Uberon enrichment analysis program) was developed, which uses Fisher's exact test. All the gene sets (sections 2.2.4 and 2.2.5) were used for GO and Uberon enrichment analyses. Terms with p-values less than 0.05 were considered as enriched terms.

## **2.3 Results**

### *2.3.1 Selection of the integrated networks for module detection*

For the zebrafish, the filtered integrated network generated using the Lin method was selected for module detection because it outperformed all other integrated networks (Schlicker, Wang, and Resnik) during network-based candidate gene predictions (section 1.3.6; Fig. 1.14.a, Fig. 1.14.b, and Supplementary Fig. S1.4). For the mouse, the filtered integrated network generated using the Schlicker method was selected based on the network-based candidate gene prediction performance (section 1.3.6; Fig. 1.15.a, Fig. 1.15.b, and Supplementary Fig. S1.5).

The zebrafish Lin integrated network contained 17,394 genes and 730,855 interactions and was filtered using gene similarity score cutoff 0.33 (Table 1.4). The mouse Schlicker integrated network contained 18,002 genes and 613,671 interactions and was filtered using gene similarity score cutoff 0.30 (Table 1.5).

### 2.3.2 *Detection of network modules*

The statistics for the number of genes in the module with original annotations to each anatomical entity is given in Table 2.1. The total number of genes for the pectoral fin (198) and the forelimb (267) are not substantially different, but the total number of genes for the pelvic fin (15) and the hindlimb (777) are substantially different.

The ROC curves generated for each anatomical entity during the network-based candidate gene prediction evaluations are given in Fig. 2.1 and the precision-recall curves are given in Fig. 2.2. According to the curves, all anatomical entities except the pelvic fin have high accuracies for network-based candidate gene predictions. The reason for the low accuracy for the pelvic fin is because it has fewer genes (15) with original annotations, and the network-based candidate gene prediction may have predicted correct gene candidates, which could have been considered as false positives because they were not included in the original annotations.

The statistics for the extracted gene modules including the precision thresholds used for candidate gene prediction, the number of predicted genes, the number of genes with original annotations, the total number of genes, and the number of genes lost due to network cutoff or isolation are given in Table 2.2. The genes with original annotations that were lost during the module extraction are listed in Supplementary Table S2.1. A high precision threshold of 0.7 was used for candidate gene predictions for pectoral fin, forelimb, and hindlimb modules. The

precision threshold for the pelvic fin was lowered to 0.05 to obtain an approximately similar number of genes in the pelvic fin and the hindlimb modules.

The visualizations of the resulting modules for the pectoral fin, pelvic fin, forelimb, and hindlimb are given in Figs. 2.3, 2.4, 2.5, and 2.6, respectively. The companion Cytoscape network files for these modules are available online at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository. The top 50 genes in the pectoral fin, pelvic fin, forelimb, and hindlimb modules ranked based on the weighted degree are listed in Supplementary Tables S2.2, S2.3, S2.4, and S2.5, respectively. The full gene lists for the above modules are available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

### 2.3.3 Assessing the biological significance of the modules

The boxplot comparisons of the distribution of module gene counts in the immediate neighborhoods of module genes *versus* network background genes for each module is given in Fig. 2.7. It is clear that module gene counts are higher in the immediate neighborhood of the module genes compared to the network background genes for all the modules. This indicates that the module genes are clustered together and not randomly scattered throughout the network. Therefore, there is a biological significance in the pectoral fin, the pelvic fin, the forelimb, and the hindlimb modules.

### 2.3.4 Comparison of the network modules

#### 2.3.4.1 Pectoral fin and forelimb comparison

According to the comparison, 183 genes are specific to the pectoral fin module, 207 genes are specific to the forelimb module, and 37 genes are shared (conserved genes) by the

pectoral fin and forelimb (Fig. 2.8). The conserved genes are listed in Table 2.3. To observe the relationship of conserved genes, the subnetworks networks of the 37 conserved genes from the pectoral fin and the forelimb modules are visualized in Fig. 2.9. This figure represents how the conserved genes are interacting with each other within the respective modules. A boxplot comparison of normalized weighted degree distributions for pectoral fin module-specific genes, pectoral fin conserved genes (genes of the pectoral fin in common with forelimb), forelimb conserved genes (genes of the forelimb in common with pectoral fin), and forelimb module-specific genes is given in Fig. 2.10. It is evident that conserved genes in both modules have higher normalized weighted degree distributions compared to the respective module-specific genes. This indicates the conserved genes are more important to the stability of the modules.

The enriched GO-BP terms for the pectoral fin module-specific genes, pectoral fin conserved genes, forelimb conserved genes, and forelimb module-specific genes are listed in Supplementary Tables S2.6, S2.7, S2.8, and S2.9, respectively. The enriched Uberon anatomy ontology terms for the same gene groups are listed in S2.10, S2.11, S2.12, and S2.13. For the enrichment analyses, the terms with p-values less than 0.05 were selected as enriched terms.

To understand the fate of the pectoral fin module-specific genes in the mouse, the enriched BP and Uberon terms for the mouse orthologs for the pectoral fin module-specific genes are given in Supplementary Tables S2.14 and S2.15, respectively. To identify the role of forelimb module-specific genes in the zebrafish, the enriched BP and Uberon terms for their zebrafish orthologs are given in Supplementary Tables S2.16 and S2.17, respectively.

#### 2.3.4.2 Pelvic fin and hindlimb comparison

According to the comparison, 536 genes are specific to the pelvic fin module, 601 genes are specific to the hindlimb module, and 81 genes are conserved genes in both the pelvic fin and hindlimb (Fig. 2.11) modules. The conserved genes are listed in Table 2.4. To observe the relationships among conserved genes, subnetworks of the 81 conserved genes from the pectoral fin and the forelimb modules are visualized in Fig. 2.12. This figure represents how the conserved genes are interacting with each other within the respective modules. A boxplot comparison of normalized weighted degree distributions for pelvic fin module-specific genes, pelvic fin conserved genes, hindlimb conserved genes, and hindlimb module-specific genes are given in Fig. 2.13. According to the figure, it is evident that conserved genes in the hindlimb module have a higher normalized weighted degree distribution compared to the hindlimb module-specific genes. For the pelvic fin module, the normalized weighted degree distribution of the conserved genes is marginally higher than the module-specific genes. This means the conserved genes are more important to the stability of the modules, especially in the forelimb module.

The top 100 enriched GO-BP terms for the pelvic fin module-specific genes, pelvic fin conserved genes, hindlimb conserved genes, and hindlimb module-specific genes are listed in Supplementary Tables S2.18, S2.19, S2.20, and S2.21, respectively. The enriched Uberon anatomy ontology terms for the same gene groups are listed in Supplementary Tables S2.22, S2.23, S2.24, and S2.25, respectively. For the enrichment analyses, the terms with p-values less than 0.05 were selected as enriched terms.

To understand the fate of the pelvic fin module-specific genes in the mouse, the enriched GO-BP and Uberon terms for their mouse orthologs are given in Supplementary Tables S2.26



and S2.27, respectively. To identify the role of the hindlimb module-specific genes in the zebrafish, the enriched BP and Uberon terms for their zebrafish orthologs are given in Supplementary Tables S2.28 and S2.29, respectively.

### 2.3.5 Validating the predicted genes

The list of predicted genes for pectoral fin, pelvic fin, forelimb, and hindlimb modules are given in Tables 2.5, 2.6, 2.7, and 2.8, respectively. In these, the predicted genes are ordered according to the rank based on weighted degree and a column that details whether the ortholog of each predicted gene is associated with the equivalent anatomical entity of the other organism is also included. For instance, in Table 2.5, the top predicted gene for the pectoral fin, *bmp4*, is directly annotated to the forelimb in the mouse. This provides a certain validation for the prediction of the *bmp4* gene. Out of the 45 predicted genes for the pectoral fin, 14 genes have mouse orthologs that are associated with the forelimb (9 direct annotations, 2 annotations only to the parts or the developmental precursors, and 3 predicted genes). Out of the 605 predicted genes for the pelvic fin (Table 2.6), 78 genes have mouse orthologs that are associated with the hindlimb (46 direct annotations, 20 annotations only to the parts or the developmental precursors, and 12 predicted genes). Out of the 18 predicted genes for the forelimb (Table 2.7), 6 genes have mouse orthologs that are associated with the pectoral fin (2 direct annotations, 1 annotation only to the parts or the developmental precursors, and 3 predicted genes). Out of the 32 predicted genes for the hindlimb (Table 2.8), 12 genes have mouse orthologs that are associated with the pelvic fin (all 12 are predicted genes).

The enriched GO-BP terms that are common to the predicted gene set and genes with original annotations to pectoral fin, pelvic fin, forelimb, and hindlimb are listed in

Supplementary Tables S2.30, S2.31, S2.32, and S2.33, respectively. The enriched Uberon terms that are common to the predicted genes and genes with original annotations to pectoral fin, pelvic fin, forelimb, and hindlimb are listed in Supplementary Tables S2.34, S2.35, S2.36, and S2.37, respectively. In these, the enriched terms are sorted based on the p-value of those terms for the predicted genes. There are several common enriched terms for all the modules, some of which are related to the anatomical entity of the module.

The boxplot comparisons of the weighted degree distributions for the predicted genes *versus* genes with original annotations for the pectoral fin, pelvic fin, forelimb, and hindlimb modules are shown in Fig. 2.14. In all the modules, the weighted degree distributions of the predicted genes are higher than the genes with original annotations. This indicates that predicted genes as a group have a higher number of interactions and central to the stability of the modules. This could be due to selecting high precision thresholds for the predictions, the accurate performance of the network-based candidate gene predictions and using high quality integrated networks for the predictions. This provides further evidence that the predicted genes are important genes for the anatomical phenotypes of the paired fins and limbs.

## **2.4 Discussion**

This chapter focuses on identifying gene modules related to evolutionary phenotypic transitions and comparing them to identify the modular changes associated with those transitions. The fin to limb transition was selected as the use case for this purpose, and the modules for the pectoral fin, pelvic fin, forelimb, and hindlimb were extracted and the homologous modules were compared to identify the changes. The integrated networks (Chapter 1) were used to ensure that

the detected modules have accurate predicted genes and the interactions within the modules are more reliable.

#### 2.4.1 Pectoral fin and forelimb module comparison

The pectoral fin (220 genes) and forelimb modules (243 genes) are approximately similar in the module size (Table 2.2) and contain a majority of genes that have direct annotations to the anatomical entity or only to its parts or the developmental precursors (Figs. 2.3 and 2.5). In the pectoral fin module, the top-ranked hub gene based on the weighted degree is the *shha* (sonic hedgehog a) gene (Supplementary Table 2.2) that is well-known to be associated with pectoral fin development (Grandel, et al., 2000; Letelier, et al., 2018). Not only in the fishes, the *shh* gene is also important in the development and morphogenesis of limbs in tetrapods including humans (Abbasi, 2011; Lopez-Rios, 2016). The loss or gain of activity in the sonic hedgehog signaling pathway in tetrapods results in gained, lost, or malformed limbs (Lopez-Rios, 2016). The *shh* gene is also considered as an important gene associated with fin to limb transition because it is important in the morphological patterning of paired fins and limbs (Amaral and Schneider, 2018; Coates and Cohn, 1998). It is also an important hub gene in the forelimb module, which is ranked 4<sup>th</sup> according to the weighted degree (Supplementary Table S2.4).

The top-ranked gene in the forelimb module is *bmp4* (bone morphogenetic protein 4), which is another important gene associated with limb formation and morphogenesis in tetrapods (Bakrania, et al., 2008; Bandyopadhyay, et al., 2006). As with the *shh* signaling pathway, mutations in *bmp4* affect the *bmp4* signaling pathway, which causes abnormalities in limb and digit formation in tetrapods (Bakrania, et al., 2008). *Bmp4* is ranked 2<sup>nd</sup> in the pectoral fin module (S2.2, Table 2.3) and was predicted during module detection. *Bmp4* does not have direct

annotations to the pectoral fin or its parts or developmental precursors, but it is involved with caudal fin development in zebrafish (Smith, et al., 2008).

When comparing modules from different species networks, a main goal is to identify the genes in common, which are interpreted as the conserved genes (Sharan, et al., 2005; Shui and Cho, 2016). The extractions of conserved genes for the pectoral fin and forelimb modules are shown in Fig. 2.9. Some of the important hub genes in the pectoral fin module, such as *shha*, *bmp4*, *bmp2b*, and *bmp7a*, have retained their importance in the forelimb module and ranked high based on the weighted degree (Table 2.3). Other genes, such as *sox9*, have changed its importance (based on ranking) during the transition from pectoral fin to forelimb. For instance, *sox9a* and *sox9b* genes are ranked 83<sup>rd</sup> and 104<sup>th</sup>, respectively, in the pectoral fin module (Supplementary Table 2.2, Fig. 2.9), while their mouse ortholog, *sox9*, has ranked 15<sup>th</sup> and has become more important in the forelimb module (Table 2.3, Fig. 2.9). The *sox9* gene is known to be involved with digit patterning in the limbs of tetrapods due to its participation in the a *bmp-sox9-wnt* Turing network (Raspopovic, et al., 2014). In the zebrafish, *sox9a* and *sox9b* genes are involved with pectoral fin development (Yan, et al., 2005). Because digits emerged after the transition from fins to limbs, hypothetically, the involvement of *sox9* in a digit patterning pathway in the mouse may have increased its number of interactions with other module genes in the forelimb module, and hence, the importance.

The conserved genes in both the modules have higher normalized weighted degree distributions compared to the respective module-specific genes (Fig. 2.10). This means that conserved genes share more interactions within the module and are more central to modular stability. From an evolutionary point of view, during the transition from the pectoral fin to the forelimb, it appears that more important genes with higher degrees in the pectoral fin module,

such as *shha* and *bmp4*, were conserved in the forelimb module and new forelimb module-specific genes were recruited surrounding those conserved genes.

A comparison of enriched GO-BP terms and Uberon terms between module-specific genes and conserved genes indicates that conserved genes are more enriched in general biological processes and anatomical entities than the module-specific genes. For example, for the pectoral fin module-specific genes, five of the top six enriched BP terms are specifically related to fin development (fin development, pectoral fin development, fin morphogenesis, embryonic pectoral fin morphogenesis, and pectoral fin morphogenesis) (Supplementary Table S2.6), whereas for the conserved genes in the pectoral fin module, only two of the top six enriched BP terms are related to fin development (pectoral fin development and embryonic pectoral fin morphogenesis) (Supplementary Table S2.7). The remaining four include terms such as embryonic viscerocranium morphogenesis, cardioblast differentiation, and inner ear development, which are more diverse biological processes. Furthermore, some biological processes, such as heart morphogenesis and swim bladder development are only enriched for the conserved pectoral fin genes. This is further confirmed when comparing enriched Uberon terms between pectoral module-specific genes and conserved genes (Supplementary Tables S2.10 and S2.11). For instance, Uberon entities such as anal fin, dorsal fin, and pelvic fin are only enriched for pectoral fin module-specific genes, whereas several heart-related terms (epicardium and heart rudiment) and digestive system-related terms (liver and intestine) are only enriched in the conserved genes of the pectoral fin module. From these results, it appears that pectoral fin module-specific genes are involved with fin-related pathways as a group and pectoral fin conserved genes are involved with a number of general anatomical entities, such as heart, digestive system, ear, and related pathways as a group.

Similar results were observed when comparing enriched terms for forelimb module-specific genes and forelimb conserved genes in mouse. For instance, comparison of enriched BP terms between forelimb module-specific genes (Supplementary Table S2.9) and forelimb conserved genes (Supplementary Table S2.8) indicates that there are five limb-related biological processes (embryonic digit morphogenesis, embryonic limb morphogenesis, embryonic forelimb morphogenesis, and limb development) within the top 15 enriched BP terms for the module-specific genes (Supplementary Table S2.9). However, for forelimb conserved genes, there are only 3 limb-related biological processes (embryonic limb morphogenesis, embryonic forelimb morphogenesis, and embryonic hindlimb morphogenesis) within the top 15 enriched BP terms, and they are ranked lower compared to forelimb module-specific enriched terms. Furthermore, as observed for the pectoral fin, general biological processes, such as heart development, lung development, and kidney development have achieved higher ranks within the top enriched BP terms for the forelimb conserved genes.

Based on higher weighted degree distributions and enrichment analyses for pectoral fin conserved genes, it can be speculated that conserved genes, as a group, are more central to the stability of the module. Further, they are not only limited to fin-related anatomical entities and biological processes, but also to more diverse biological processes. Hypothetically, this may be a reason for them to be conserved in the forelimb module during the evolutionary transition, as they are central genes involved in the functioning of the whole organism. The forelimb module-specific genes may have been recruited for the forelimb development while forming interactions with the conserved genes. This may explain the high weighted degree distributions observed for the conserved genes in both pectoral fin and forelimb modules.

One important point to note is that the enrichment analyses are performed for the group of genes as a whole and the results indicate the enrichment for the whole group. Therefore, higher weighted degree distributions observed for the conserved genes and the speculation that they are involved in diverse biological processes and the lower weighted degree distributions for the module-specific genes and the speculation that they are involved in more limb or fin-related biological processes are for the whole groups of genes. Thus, although it appears as there are exceptions to this general observation, e.g., module-specific genes such as *lef1* and *hdac1* for the pectoral fin (Supplementary Table S2.2) have higher weighted degrees and are involved with diverse biological processes similar to the conserved genes, the generalization here pertains to the whole gene set.

#### 2.4.2 Pelvic fin and hindlimb module comparison

Detection of the pelvic fin module was challenging because only 15 genes had original annotations (Table 2.1). However, for the hindlimb, there were 777 genes with original annotations (Table 2.1). Unlike the limb development in the mouse, where forelimb and hindlimb buds emerge at the same timepoint, the pelvic fin buds emerge at a much later stage than the pectoral fin bud (Grandel and Schulte-Merker, 1998). This may be a potential reason for the lack of annotations to the pelvic fin, as many of the studied gene disruptions kill the larval zebrafish before the pelvic fin develops or the larvae are sacrificed at a pre-determined early stage. To perform a meaningful comparison, more genes had to be predicted for the pelvic fin by lowering the precision threshold during the network-based candidate gene prediction (Table 2.2). The resulting pelvic fin module contains a majority of predicted genes (Fig. 2.4), which may not be as reliable as other modules, but it was necessary to perform the comparison. Lack of

annotations for certain anatomical and GO terms, such as pelvic fin, is a major issue in bioinformatics analyses (Baumgartner Jr, et al., 2007; Kim, et al., 2003). Ideally, more wet lab experiments should be focused on those anatomical entities.

In the pelvic fin module, the top ranked gene (predicted) is *hsp90ab1* (Supplementary Table S2.3). Although it is a heat shock protein and does not have known effects on the pelvic fin, studies show that the inhibition of its expression causes defects in zebrafish, especially in eye development (Yeyati, et al., 2007). Furthermore, according to Yeyati, et al. (2007), the disruption of *hsp90ab1* expression is associated with caudal fin fold defects in the zebrafish; which is not recorded in ZFIN or Monarch Initiative repository. Therefore, there is evidence to suggest that *hsp90ab1* may also be associated with the pelvic fin.

The top ranked gene in the hindlimb module based on weighted degree is *trp53* (Supplementary Table S2.5), which is associated with embryonic hindlimb development (Im, et al., 2016) in the mouse. It appears that function of *trp53* is crucial for limb development in the mouse, as radiation-induced apoptosis that disrupts *trp53* expression interferes with limb development and causes deformed limbs (Vares, et al., 2011; Wang, et al., 2000). The *trp53* gene is also found in the pelvic fin module (predicted gene) but has a lower rank (24) based on the weighted degree (Supplementary Table S2.3, Table 2.4, and Fig. 2.12).

When comparing the conserved genes in the pelvic fin and the hindlimb modules (Table 2.4 and Fig. 2.12), several important genes, which are central to the modular stability, can be identified. For example, the *ctnnb1* gene, which is predicted and ranked 4<sup>th</sup> in the pelvic fin module, is ranked 3<sup>rd</sup> in the forelimb module (Table 2.4 and Fig. 2.12). The *ctnnb1* is essential for the  $\beta$ -catenin pathway, which is necessary for the hindlimb initiation in the mouse (Kawakami, et al., 2011). Although it does not have known associations to the paired fins in the zebrafish, it is



known to be essential for the development of the fish embryo (Li, et al., 2014; Zhang, et al., 2012); hence, it is potentially an essential gene for the development of the pelvic fin.

The conserved genes in the hindlimb module have a higher weighted degree distribution compared to the module-specific genes (Fig. 2.13). This implies that conserved genes are more important to the stability of the hindlimb module as a group. However, the weighted degree distribution of pelvic fin conserved genes is marginally higher than the module-specific genes (Fig. 2.13). Majority of genes in the pelvic fin is predicted, which may be a reason for this observation. Using enrichment analysis to compare the enriched GO-BP terms and Uberon terms for pelvic fin module-specific genes and conserved genes is challenging, as there are limited original annotations to the pelvic fin. However, the top enriched BP terms for the pelvic fin module-specific genes include fin development and fin morphogenesis (Supplementary Table S2.18); hence, the majority of those predicted genes are associated with fin development. The top 10 enriched Uberon terms for the pelvic fin module-specific genes include five fin-related entities (pectoral fin, anal fin, median fin fold, pelvic fin, and dorsal fin) (Supplementary Table S2.22), which means those genes are mainly involved with fin development as a group. However, a similar observation can be seen for the conserved genes in the pelvic fin module, as there are four fin-related terms (pectoral fin, median fin fold, ventral fin fold, and dorsal fin) in the top 10 enriched Uberon terms (Supplementary Table S2.23). There is no clear evidence to hypothesize that conserved genes in pelvic fin are generally involved with more diverse biological processes and anatomical terms compared to module-specific genes as seen for the pectoral fin (section 2.4.1). Having a large number of predicted genes in the pelvic fin module is a potential reason for the pelvic fin results to deviate from the other three modules.

For the hindlimb module-specific genes, the top 10 enriched BP terms include biological processes such as embryonic limb morphogenesis, skeletal system development, ossification, embryonic digit morphogenesis, and cartilage development that could be related to limb development (Supplementary Table S2.21), whereas for the hindlimb conserved genes, there is only one limb-related enriched term (embryonic limb morphogenesis) in the top 10 BP terms (Supplementary Table S2.20). However, there are more diverse biological processes such as heart development and lung development in the top 10 BP terms. Similarly, nine (hindlimb, appendicular skeleton, skeletal system, hindlimb zeugopod, tibia, hindlimb stylopod, femur, forelimb, and trabecular bone tissue) out of the top 10 enriched Uberon terms for the hindlimb module-specific genes are related to the limb (Supplementary Table S2.25), whereas only three (appendicular skeleton, hindlimb, forelimb) out of the top 10 enriched Uberon terms for hindlimb conserved genes are related to the limb (Supplementary Table S2.24). The remaining terms are mostly related to the face and head (jaw skeleton, facial skeleton, face, head, mouth, and cranium) (Supplementary Table S2.24).

Although the top enriched terms for the pelvic fin module-specific and conserved genes are not significantly different, the enriched terms for the hindlimb module-specific genes are more specific towards the limb development and morphogenesis as seen with the pectoral fin and forelimb comparison. Furthermore, the conserved genes for both pelvic fin and hindlimb modules have higher weighted degree distributions and thus are more important to the function of individual modules than the module-specific genes. Therefore, as explained in the section 2.4.1, during the transition from the pelvic fin to the hindlimb, the conserved genes that were retained in the hindlimb were potentially involved with diverse biological processes and

anatomical entities and thus more central to the function of the modules. The hindlimb module-specific genes may have been recruited surrounding the important conserved genes.

#### 2.4.3 The fate of the zebrafish module-specific genes in the mouse

For both pectoral fin-forelimb and pelvic fin-hindlimb comparisons, there were large number of zebrafish module-specific genes that were not included in the limb modules (Fig. 2.8 and Fig. 2.11). It is interesting to investigate the new roles of those zebrafish module-specific genes in the mouse. The aquatic to terrestrial vertebrate transition resulted in several anatomical changes and introduced new anatomical entities such as lungs and neck to tetrapods that helped them to overcome the terrestrial environment (Clack, 2012; Shubin, 2008). A closer inspection of enriched BP and Uberon terms for the mouse orthologs of the pectoral fin (Supplementary Tables S2.14 and S2.15) and pelvic fin (Supplementary Tables S2.26 and S2.27) module-specific genes indicate that those genes are enriched for a number of novel anatomical entities that occurred in tetrapods and related biological processes. These enriched anatomical entities are listed in Table 2.9 for the pectoral fin module-specific genes and in Table 2.10 for the pelvic fin module-specific genes.

For instance, the pectoral fin module-specific gene *lef1*, which is an important gene (ranked 7<sup>th</sup>) in the pectoral fin module, is involved with palate development, trachea gland development BP terms and associated with neck-related phenotypes (Duan, et al., 1999; Nawshad and Hay, 2003). Neck is an important anatomical entity emerged in tetrapods, which allowed them to support the head that was crucial for their success in the land. Of the pelvic fin module-specific genes, the *mapk1* gene is an example that is associated with neck-related phenotypes and biological processes, such as thymus development and trachea formation

(Boucherat, et al., 2014; Hoffman, et al., 2006). The *mapk1* gene is ranked 12<sup>th</sup> in the pelvic fin module and is important for the stability of the module. It is also involved with the lung phenotypes and the development of the lung (Boucherat, et al., 2014), which is another important structure emerged in tetrapods that enabled them to breath and thrive in terrestrial environments. In both the pectoral fin and pelvic fin modules, *lama5* gene is an example for a module-specific gene which is involved with lung development in the mouse (Nguyen, et al., 2002). Furthermore, it is also involved with hair follicle development and hair-related phenotypes (Gao, et al., 2008), which is another anatomical entity that is specific for mammals such as the mouse. The *egfra* is another important pelvic fin module-specific gene that is not only involved with the lung (Threadgill, et al., 1995) and hair (Laisney, et al., 2010) in the mouse but also in the development of more specific structures such as placenta (Lee and Threadgill, 2008) and eyelid (Laisney, et al., 2010). With these examples and the enriched terms listed in Table 2.9 and Table 2.10, it can be hypothesized that most of the zebrafish module-specific genes were recruited for tetrapod specific anatomical entities that helped them to thrive in a terrestrial environment.

#### 2.4.4 *The fate of the mouse module-specific genes in the zebrafish*

There is a large number of module-specific genes for the forelimb (Fig. 2.8) and hindlimb (Fig. 2.11) in the mouse. These genes do not appear in pectoral fin or pelvic fin modules, and the question of their developmental function in the zebrafish arises. Enrichment analyses showed that most of the mouse module-specific genes are enriched in the head of the zebrafish, specifically, to the jaw skeleton and post-hyoid pharyngeal arch skeleton (Supplementary Table S2.16, S2.17, S2.28, S2.29 and Tables 2.11, 2.12). The latter region covers the gill chamber and contains parts such as gill rakers (Gillis, et al., 2013). This suggests that at least some of the

mouse module-specific genes may have been associated with the operculogular series and gill specific structures that are absent in mouse and more generally lost in tetrapods. For instance, *fst* is a crucial forelimb module-specific gene (Supplementary Table S2.4), which has a zebrafish ortholog (*fsta*) with phenotypes related to splanchnocranium (Dalcq, et al., 2012) and post-hyoid pharyngeal arch skeleton (Dal-Pra, et al., 2006) that supports the gill chamber. Furthermore, *twist1* is module-specific for both the forelimb and hindlimb, and it has two zebrafish orthologs (*twist1a* and *twist1b*) that are involved with pharyngeal system development (Das and Crump, 2012).

There are also some mouse module-specific genes that are involved with the caudal fin of the zebrafish. For instance, *tgfbr3*, which is module-specific for both the forelimb and the hindlimb, is involved with the development of the caudal fin (Kamaid, et al., 2015). Furthermore, hindlimb module-specific genes such as *wnt5b* are involved with pectoral fin morphogenesis (Yokoi, et al., 2003) although they are not included in the pelvic fin module. Another example is *lep*, which is module-specific for both forelimb and hindlimb; its ortholog in zebrafish, *lepa*, is associated with otolith development (Liu, et al., 2012). In fact, otoliths were enriched in the zebrafish orthologs of both forelimb (Table 2.11) and hindlimb (Table 2.12) module-specific genes. Otoliths are located in the inner ear cavity of all teleost fishes where they aid in hearing and serve as balance organs (Rodríguez Mendoza, 2006). According to these results, it can be speculated that some mouse module-specific genes were involved with fish specific anatomical entities such as operculogular series and caudal fin that were lost during the transition to tetrapods, and then, those genes were recruited for the limb modules.

#### 2.4.4 Validating the predicted genes

One important application of PPI networks is that they can be used to predict candidate genes during module detection (Cowen, et al., 2017; Sharan, et al., 2007). There can be several important genes that were missed by wet lab gene prediction methods, such as gene knockout for anatomical entities such as pelvic fin (section 2.4.2). However, validating the predictions is important in any bioinformatics analysis. The best way to validate is to use wet lab methods, but within the scope of the computational analysis, several steps were taken to ensure that the best predictions were made. First, before the detection of the four modules, the candidate gene prediction accuracy for each anatomical entity was evaluated (Fig. 2.1 and Fig. 2.2). According to the ROC and precision-recall curves, pectoral fin, forelimb, and hindlimb show high accurate predictions (the AUC values of ROC curves for all three entities are higher than 0.85) (Fig. 2.1). The problem with the pelvic fin is the low number of original annotations (15 genes) (Table 2.1) due to experimental restrictions (section 2.4.2); hence, the network-based candidate gene prediction method may have predicted correct gene candidates that were treated as false positives due to lack of knowledge regarding those genes. Therefore, relying only on evaluation methods such as ROC and precision-recall curves is not adequate.

A second way to validate the predictions is to confirm whether the orthologs of the predicted genes are involved with the homologous anatomical entities in other organisms. For example, the predicted genes of the zebrafish modules were checked with their mouse counterparts, and vice versa. The orthologs of several predicted genes were annotated to the homologous anatomical entity. For instance, 78 of the predicted genes in the pelvic fin module were associated with the hindlimb in mouse. This provides a certain level of validation for the predicted genes. However, not all the predicted genes have to be associated with the homologous

anatomical entity in the other organism. As discussed in the sections 2.4.3 and 2.4.4, it appears that some zebrafish paired fin module genes have lost the association with the paired limbs during evolution and new genes have been recruited. Furthermore, there is a possibility that some predicted genes in the zebrafish fin modules are associated with the mouse limb modules, but the associations are not yet discovered by wet lab methods.

A third method used to validate the predicted genes is to perform enrichment analysis for the predicted genes and analyze the enriched terms to identify terms that are related to the anatomical entity in question (Huang, et al., 2009; Kuleshov, et al., 2016). For this work, the enriched GO-BP and Uberon terms for predicted genes and genes with original annotations for each module were compared to identify the common terms for both the groups. This confirms whether the predicted genes are regulating the same biological processes and involved with the same anatomical entities as the genes with original annotations. For the pectoral fin, the common enriched biological processes (Supplementary Table S2.30) include terms such as skeletal system development and Wnt signaling pathway, which are related to pectoral fin development. For instance, Wnt signaling pathway is known to be associated with paired fin and caudal fin development in the zebrafish (Stoick-Cooper, et al., 2007; Wehner, et al., 2014). When considering the common enriched Uberon terms for the pectoral fin (Supplementary Table S2.34), there are anatomical entities such as median fin fold, ventral fin fold, and caudal fin, which are related to the pectoral fin.

For the pelvic fin, the only biological process that was common to the predicted genes and the original genes is fin development (Supplementary Table S2.31). When considering the enriched Uberon terms (Supplementary Table S2.35), all terms that were common to the predicted and original genes, such as median fin fold and lepidotrichium were related to the fins.

This is useful evidence that suggests although the number of predicted genes for the pelvic fin is high (605) (Table 2.2), as a group, they are highly involved with fin-related development processes and anatomical entities. Therefore, there is a high potential for those genes to be involved with the pelvic fin.

For the forelimb module, the common BP terms for predicted genes and genes with original annotations (Supplementary Table S2.32) include processes such as BMP signaling pathway, embryonic limb morphogenesis, embryonic digit morphogenesis, and Wnt signaling pathway, and common Uberon terms (Supplementary Table S2.36) include anatomical entities such as appendicular skeleton and skeletal system. For the hindlimb module, common BP terms (Supplementary Table S2.33) include embryonic limb morphogenesis, BMP signaling pathway, and embryonic digit morphogenesis, and common Uberon terms (Supplementary Table S2.37) include entities such as appendicular skeleton and limb, which are related to the hindlimb. Therefore, from the results of enrichment analyses, it is clear that for all four modules, there are enriched biological processes and anatomical entities that are related to the limb or the fin in question that are shared by predicted genes and genes with original annotations.

When predicting genes for the modules, a relatively high precision threshold (0.7) was used for all the modules (Table 2.2) except the pelvic fin (due to lack of original annotations). Therefore, the weighted degree distributions for the predicted genes are high for all the modules (even the pelvic fin) compared to the genes with original annotations (Fig. 2.14). This indicates that the predicted genes in all the modules are important for the stability of the module and have high number of interactions with other members as a group. For example, the *bmp4* gene is a predicted gene in the pectoral fin module, which is ranked 2<sup>nd</sup> (section 2.4.1) (Supplementary Table S2.2). The *wnt3a* (ranked 4<sup>th</sup>) and *wnt5b* (ranked 10<sup>th</sup>) are another two important predicted



genes in the pectoral fin module (Supplementary Table S2.2), which potentially are involved with the pectoral fin development. They both are associated with the Wnt signaling pathway, which is crucial for paired fin development in fish (Mercader, 2007; Stoick-Cooper, et al., 2007). According to Supplementary Table S2.2, there are several predicted genes that have a higher rank, which could be involved with the pectoral fin.

For the pelvic fin module, nearly all the top ranked genes are predictions (Supplementary Table S2.3). For instance, the *ctnnb1* gene is an important predicted hub gene, which ranked 4<sup>th</sup> in the pelvic fin module (Supplementary Table S2.3). Although it does not have any known associations with fin development in fish in literature, it is known to be associated with limb development in the mouse (Browne, et al., 2012). Therefore, it is found in the hindlimb module as an important hub gene (ranked 3<sup>rd</sup>) (Table 2.4). Therefore, there is a high possibility that it is also associated with fin development in fish. The *cad* gene is another predicted hub gene, which is ranked 8<sup>th</sup> in the pelvic fin module. Although it does not have any direct associations with the pelvic fin, it is expressed in fin buds during zebrafish development (Willer, et al., 2005). Therefore, it can be a potential important gene for pelvic fin development that needs more investigation.

The mouse is better studied than zebrafish, but there is room for improvement regarding unravelling genes associated with limb development. For example, in the forelimb module, some crucial genes such as *smad4* (rank 6<sup>th</sup>) and *bmp7* (rank 7<sup>th</sup>) were predicted (Supplementary Table S2.4). Although *smad4* does not have direct annotations to paired limbs, there is literature support for its role in early limb development (Zuzarte-Luis, et al., 2004). The *bmp7* is an important gene not only for limb development in tetrapods (Bandyopadhyay, et al., 2006) but also for fin development in the zebrafish (Mullins, et al., 1996). It has direct annotations to the

limb term but not specifically to the forelimb and hindlimb; hence, not included in those modules. Not only the *bmp7* was accurately predicted for the forelimb (rank 7<sup>th</sup>) (Supplementary Table S2.4) and hindlimb modules (rank 26<sup>th</sup>) (Supplementary Table S2.5), network analysis results show that it is an important hub gene in both the modules. Interestingly, the *hras* and *kras* predicted genes ranked 2<sup>nd</sup> and 5<sup>th</sup>, respectively in the hindlimb module. They are both oncogenes that are involved in development of tumors, such as lung cancer in the mouse (To, et al., 2008). There is no literature evidence to support their involvement in limb development; therefore, these are good candidates to validate using wet lab experiments.

These validation results show that the predicted genes are important to the stability of the individual modules. They also indicate the practical advantage of the integrated networks developed in Chapter 1 for accurately unravelling new candidate genes for anatomical entities. As a future step, some predicted genes that are important in the modules can be validated using wet lab methods, such as gene knockout, to confirm their role in fin or limb development.

#### *2.4.1 Challenges and future directions*

When performing big data analyses, such as large-scale network analyses, the final results depend on the original input data, which are retrieved from public databases. The completeness of these datasets is an important factor that determines the accuracy of the predictions and conclusions. When considering gene-phenotype associations, the human and mouse models are extensively studied, and more complete datasets are available for them than other model organisms such as the zebrafish and the *Xenopus*. When studying the aquatic to terrestrial vertebrate transition, *Xenopus* is an important model organism that could have been useful for the analysis. Unfortunately, limited anatomical phenotype annotations are available for

Xenopus genes in public databases; hence, it had to be excluded from the analysis. Even the zebrafish annotations are not complete; anatomical entities, such as pelvic fin needs more focus in wet lab experiments. Ultimately, the predictions of computational analyses, such as this work can be used as a foundation for those wet lab experiments, which is valuable for the improvement of evolutionary bioinformatics.

When comparing large-scale gene sets and PPI networks across different model organisms, identifying the correct gene orthologs is important. When comparing mouse and zebrafish genes, this becomes complicated due to the extra duplication event in the teleost fishes (Braasch, et al., 2014; Meyer and Schartl, 1999). As a result, if multiple zebrafish orthologs are present for a single mouse gene, all orthologs were kept during the analysis. For instance, the *sox9* gene is a member of the forelimb module in mouse and it has two zebrafish orthologs, *sox9a* and *sox9b*, which are present in the pectoral fin module (Table 2.3).

When performing network analyses on large PPI networks and modules, the visualization of these modules is a challenge. Typically, visualizing smaller protein complexes are more clear and easier to understand than visualizing large gene modules of phenotypic entities such as hindlimb. Although a static figure can be used to distinguish the distribution of predicted *versus* genes with original annotations, to focus on individual genes, it has to be opened in an interactive network viewer, such as Cytoscape (Smoot, et al., 2011). Therefore, the Cytoscape network files for the four modules were included in the repository at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles), which enable the user to zoom in, interact, and observe the arrangement of desired genes in the module.

For this work, enrichment analysis was used extensively to identify the enriched Uberon and GO-BP terms for gene sets. There are several enrichment tools developed for the GO

(Huang, et al., 2009), but no published tools are available for the Uberon ontology. Therefore, a Python program (Uberon enrichment analysis program) that uses the Fisher's exact test was developed to perform enrichment analysis using the Uberon ontology (Section 2.2.6).

Furthermore, most of the current tools and publications list the enriched terms as a table sorted according to the p-values (e.g., Supplementary Table S2.6). The tables typically contain related ontological terms in a tabulated form and have to be read one by one. Tools such as REVIGO (Supek, et al., 2011) (<http://revigo.irb.hr/>) and Gorilla (Eden, et al., 2009) have been developed to visualize the enriched GO terms in a graph structure. This enables to cluster related ontology terms together based on semantic similarity between the terms. For instance, the graph visualization of the top 100 enriched BP terms for the pectoral fin module-specific genes (listed in Supplementary Table S2.6) that was generated using REVIGO tool is given in Fig. 2.15. As a future step, this functionality will be added to the Uberon enrichment analysis program.

One of the main advantages of performing network analysis is to identify important hub genes that are crucial for the anatomical entity/function. To identify them, each gene in the module was assigned a rank based on its weighted degree (Supplementary Tables S2.2, S2.3, S2.4, and S2.5). This information is valuable for biologists not only to discover that a specific gene is involved with the paired fin or limb but also to understand its importance and the role in the gene module. Tables showing the comparison of conserved genes for module comparisons (Table 2.3 and Table 2.4) allow to identify how the importance of conserved genes changed after the transition. Overall, the results generated in this work, are extremely valuable for evolutionary developmental biologists to shed light on fin to limb transition. All the scripts, used for this Chapter and the integrated network generation in Chapter 1 are openly available and written in a generalized way to work with any PPI network for any organism. Therefore, this analysis can be

extended to study evolutionary transitions involving a broad range of anatomical entities and model organisms if sufficient data is available.

## **2.5 Conclusion**

The goal of this chapter was to study the modular changes associated with evolutionary phenotypic transitions using the integrated networks generated during Chapter 1. Employing the integrated networks ensured that the module detections, gene predictions, and identification of important genes in the modules are more accurate than using conventional PPI networks. During the analysis of paired fin and paired limb modules, important hub genes that are crucial for the stability of the modules were identified. All the genes of the modules were ranked based on the weighted degree. Some of the important genes were predicted during the module detection and they have strong evidence to confirm their involvement with the respective fins or limbs.

Furthermore, it was found that the conserved genes have a higher potential to be hub genes than the module-specific genes due to their higher weighted degree distribution. It could be speculated that during the fin to limb transition, most of the crucial hub genes of fin modules were conserved in limb modules and module-specific limb genes were recruited surrounding those conserved genes. These conserved genes, such as the *shh* gene, as a group, are involved not only with fin or limb development, but also with more diverse developmental processes compared to module-specific genes. Moreover, further speculations were made regarding the fate of zebrafish and mouse module-specific genes. There was evidence to suggest that zebrafish fin module-specific genes are employed in anatomical structures, such as lung and neck that emerged after the aquatic to terrestrial vertebrate transition. Furthermore, there was evidence to speculate that mouse limb module-specific genes were involved with anatomical structures, such

operculogular series in the zebrafish that were lost during the transition. The network analysis results of this work provide the groundwork for evolutionary developmental biologists to investigate into aforementioned hypotheses. The computational framework was established to perform large-scale network analyses to study evolutionary transitions involving any model organism and anatomical entity with sufficient data, which is valuable for the improvement of evolutionary biology.

## Tables

Table 2.1. The statistics for the number of genes with original annotations to each anatomical entity

	Pectoral fin	Pelvic fin	Forelimb	Hindlimb
Number of genes with direct annotations to the anatomical entity	192	13	216	530
Number of genes annotated only to the parts of the anatomical entity	3	2	44	239
Number of genes annotated only to the bud of the anatomical entity	3	0	7	8
The total number of genes used for the module detection	198	15	267	777

Table 2.2. The statistics for the extracted modules

	Pectoral fin	Pelvic fin	Forelimb	Hindlimb
The precision threshold used for candidate gene prediction	0.7	0.05	0.7	0.7
Number of predicted genes in the module	45	605	18	32
Number of genes with original annotations to the anatomical entity, a part, or the bud in the module	175	12	225	639
Total number of genes in the module	220	617	243	671
Number of genes with original annotations that were lost due to the network cutoff	17	3	25	101
Number of genes with original annotations that were lost due to isolation in the network	6	0	17	37



Table 2.3. Comparison of the 37 genes that are common to the pectoral fin and forelimb modules (conserved genes) (Fig. 2.8 and Fig. 2.9). The genes are ordered according to the rank in the zebrafish module.

Zebrafish gene name	Annotation type	Weighted degree	Rank in the module	Mouse gene name	Annotation type	Weighted degree	Rank in the module
<i>shha</i>	direct annotation to the pectoral fin	36.6983	1	<i>shh</i>	direct annotation to the forelimb	37.6249	4
<i>bmp4</i>	predicted	35.5652	2	<i>bmp4</i>	direct annotation to the forelimb	47.2641	1
<i>bmp2b</i>	predicted	34.3109	3	<i>bmp2</i>	direct annotation to the forelimb	31.3004	13
<i>wnt3a</i>	predicted	31.4927	4	<i>wnt3a</i>	predicted	30.9124	14
<i>fgf8a</i>	predicted	31.0094	5	<i>fgf8</i>	direct annotation to the forelimb	34.0672	8
<i>gli2a</i>	predicted	27.2862	8	<i>gli2</i>	direct annotation to the forelimb	33.0563	11
<i>bmp7a</i>	direct annotation to the pectoral fin	25.8141	11	<i>bmp7</i>	predicted	34.2132	7
<i>fgfr1a</i>	predicted	25.6442	12	<i>fgfr1</i>	direct annotation to the forelimb	33.4360	10
<i>fgf10a</i>	direct annotation to the pectoral fin	24.2622	17	<i>fgf10</i>	direct annotation to the forelimb	22.7474	33
<i>smo</i>	direct annotation to the pectoral fin	23.6222	19	<i>smo</i>	direct annotation to the forelimb	28.7826	20
<i>wnt4a</i>	predicted	20.4183	29	<i>wnt4</i>	predicted	27.9668	23
<i>ptch1</i>	direct annotation to the pectoral fin	18.8304	30	<i>ptch1</i>	predicted	27.3374	24
<i>ihha</i>	predicted	18.1064	32	<i>ihh</i>	direct annotation	29.6653	17

					to the forelimb		
<i>tbx5a</i>	direct annotation to the pectoral fin	16.7968	35	<i>tbx5</i>	direct annotation to the forelimb	8.6229	86
<i>aldh1a2</i>	direct annotation to the pectoral fin	16.6865	36	<i>aldh1a2</i>	annotated to a part or bud	7.1983	100
<i>hand2</i>	direct annotation to the pectoral fin	16.4984	38	<i>hand2</i>	direct annotation to the forelimb	13.8461	62
<i>zic2a</i>	predicted	15.2976	40	<i>zic2</i>	annotated to a part or bud	7.3701	99
<i>tcf7</i>	direct annotation to the pectoral fin	15.0392	42	<i>tcf7</i>	direct annotation to the forelimb	7.7495	93
<i>msx1a</i>	predicted	14.6308	45	<i>msx1</i>	predicted	23.4172	31
<i>met</i>	direct annotation to the pectoral fin	14.5234	46	<i>met</i>	direct annotation to the forelimb	13.9973	60
<i>esr2a</i>	predicted	13.2000	51	<i>esr2</i>	direct annotation to the forelimb	7.6748	95
<i>apc</i>	direct annotation to the pectoral fin	9.3808	61	<i>apc</i>	direct annotation to the forelimb	12.6243	67
<i>tfap2a</i>	predicted	9.2007	62	<i>tfap2a</i>	direct annotation to the forelimb	9.3891	78
<i>dlx5a</i>	annotated to a part or bud	8.9299	81	<i>dlx5</i>	predicted	16.1115	51
<i>sox9a</i>	direct annotation to the pectoral fin	8.8290	83	<i>sox9</i>	direct annotation to the forelimb	29.7370	15
<i>displ</i>	direct annotation to the pectoral fin	8.6768	85	<i>displ</i>	direct annotation to the forelimb	1.8115	183
<i>mecom</i>	predicted	8.4715	89	<i>mecom</i>	annotated to a part or bud	3.3452	155
<i>sox9b</i>	predicted	7.2558	104	<i>sox9</i>	direct annotation	29.7370	15

					to the forelimb		
<i>thraa</i>	direct annotation to the pectoral fin	7.0911	105	<i>thra</i>	direct annotation to the forelimb	3.2473	157
<i>cyp26b1</i>	direct annotation to the pectoral fin	6.5056	111	<i>cyp26b1</i>	direct annotation to the forelimb	3.5223	151
<i>ctgfa</i>	direct annotation to the pectoral fin	5.3072	124	<i>ctgf</i>	direct annotation to the forelimb	20.3372	41
<i>wls</i>	direct annotation to the pectoral fin	4.8573	133	<i>wls</i>	direct annotation to the forelimb	6.6655	107
<i>osr1</i>	direct annotation to the pectoral fin	4.6429	136	<i>osr1</i>	annotated to a part or bud	0.3046	241
<i>sparc</i>	direct annotation to the pectoral fin	3.7395	148	<i>sparc</i>	direct annotation to the forelimb	12.4871	69
<i>chsy1</i>	direct annotation to the pectoral fin	1.9657	176	<i>chsy1</i>	annotated to a part or bud	1.0702	210
<i>rspo2</i>	direct annotation to the pectoral fin	0.7583	200	<i>rspo2</i>	direct annotation to the forelimb	3.6481	149
<i>pax1b</i>	direct annotation to the pectoral fin	0.3795	209	<i>pax1</i>	direct annotation to the forelimb	6.2887	111

Table 2.4. Comparison of the 81 genes that are common to the pelvic fin and hindlimb modules (conserved genes) (Fig. 2.11 and Fig. 2.12). The genes are ordered according to the rank in the zebrafish module.

Zebrafish gene name	Annotation type	Weighted degree	Rank in the module	Mouse gene name	Annotation type	Weighted degree	Rank in the module
<i>ctnbl</i>	predicted	106.2749	4	<i>ctnbl</i>	direct annotation to the hindlimb	72.1808	3
<i>smarca4a</i>	predicted	97.4336	11	<i>smarca4</i>	direct annotation to the hindlimb	36.4034	56
<i>kras</i>	predicted	93.403	15	<i>kras</i>	predicted	70.7748	5
<i>rac1a</i>	predicted	92.9107	16	<i>rac1</i>	direct annotation to the hindlimb	24.2559	103
<i>mapk14a</i>	predicted	92.1789	17	<i>mapk14</i>	direct annotation to the hindlimb	71.7856	4
<i>src</i>	predicted	91.7256	18	<i>src</i>	direct annotation to the hindlimb	70.3811	6
<i>fgfr1a</i>	predicted	91.2553	19	<i>fgfr1</i>	direct annotation to the hindlimb	53.223	16
<i>met</i>	predicted	90.8735	20	<i>met</i>	direct annotation to the hindlimb	34.8637	64
<i>tp53</i>	predicted	83.8035	24	<i>trp53</i>	direct annotation to the hindlimb	82.9324	1
<i>mtor</i>	predicted	80.8121	28	<i>mtor</i>	annotated to a part or bud	38.6216	49
<i>ptenb</i>	predicted	78.531	32	<i>pten</i>	annotated to a part or bud	43.6208	35
<i>bmp4</i>	predicted	76.215	36	<i>bmp4</i>	direct annotation to the hindlimb	63.9989	10
<i>bmp2b</i>	predicted	68.2429	45	<i>bmp2</i>	predicted	45.0361	32

<i>rac1b</i>	predicted	67.0359	50	<i>rac1</i>	direct annotation to the hindlimb	24.2559	103
<i>notch2</i>	predicted	66.3735	54	<i>notch2</i>	direct annotation to the hindlimb	39.0854	47
<i>fgfr2</i>	predicted	63.0559	67	<i>fgfr2</i>	direct annotation to the hindlimb	48.7644	24
<i>ptena</i>	predicted	60.7765	80	<i>pten</i>	annotated to a part or bud	43.6208	35
<i>mapk14b</i>	predicted	59.24	89	<i>mapk14</i>	direct annotation to the hindlimb	71.7856	4
<i>smad2</i>	predicted	58.5213	92	<i>smad2</i>	predicted	45.6514	30
<i>wnt3a</i>	predicted	57.9542	95	<i>wnt3a</i>	direct annotation to the hindlimb	42.6564	39
<i>ret</i>	predicted	57.2523	98	<i>ret</i>	annotated to a part or bud	23.5326	110
<i>vegfaa</i>	predicted	57.0152	100	<i>vegfa</i>	annotated to a part or bud	50.7476	20
<i>junba</i>	predicted	54.8789	109	<i>junb</i>	direct annotation to the hindlimb	15.5932	205
<i>fgf8a</i>	predicted	52.5593	122	<i>fgf8</i>	direct annotation to the hindlimb	46.6654	29
<i>gli2a</i>	predicted	52.1448	130	<i>gli2</i>	direct annotation to the hindlimb	42.8396	38
<i>dicer1</i>	predicted	50.473	140	<i>dicer1</i>	direct annotation to the hindlimb	18.3287	187
<i>smad3a</i>	predicted	49.5403	150	<i>smad3</i>	annotated to a part or bud	49.4434	22
<i>hrasb</i>	predicted	48.0853	162	<i>hras</i>	predicted	72.8342	2
<i>shha</i>	predicted	48.0501	163	<i>shh</i>	direct annotation to the hindlimb	44.4006	33

<i>bmp2a</i>	predicted	47.8327	165	<i>bmp2</i>	predicted	45.0361	32
<i>fgf4</i>	predicted	47.5358	167	<i>fgf4</i>	annotated to a part or bud	24.3609	102
<i>bmp7a</i>	predicted	46.1288	178	<i>bmp7</i>	predicted	48.2316	26
<i>vegfab</i>	predicted	45.8605	183	<i>vegfa</i>	annotated to a part or bud	50.7476	20
<i>mysm1</i>	predicted	43.0713	209	<i>mysm1</i>	direct annotation to the hindlimb	4.6408	394
<i>fgf10a</i>	predicted	42.2153	212	<i>fgf10</i>	direct annotation to the hindlimb	33.825	67
<i>fgf10b</i>	predicted	40.7886	219	<i>fgf10</i>	direct annotation to the hindlimb	33.825	67
<i>sirt1</i>	predicted	40.1693	232	<i>sirt1</i>	annotated to a part or bud	25.4897	94
<i>bmpr1aa</i>	predicted	36.7892	251	<i>bmpr1a</i>	direct annotation to the hindlimb	35.5242	61
<i>lef1</i>	predicted	35.665	259	<i>lef1</i>	predicted	36.7074	53
<i>gli2b</i>	predicted	33.6578	277	<i>gli2</i>	direct annotation to the hindlimb	42.8396	38
<i>wnt4a</i>	predicted	33.0982	286	<i>wnt4</i>	predicted	35.4014	62
<i>smo</i>	predicted	32.2549	293	<i>smo</i>	direct annotation to the hindlimb	34.086	66
<i>sod2</i>	predicted	30.6061	308	<i>sod2</i>	annotated to a part or bud	18.8196	178
<i>gdf11</i>	direct annotation to the pelvic fin	29.3173	321	<i>gdf11</i>	direct annotation to the hindlimb	8.6569	283
<i>stat1b</i>	predicted	28.9557	323	<i>stat1</i>	direct annotation to the hindlimb	33.3038	72
<i>smad3b</i>	predicted	27.6166	333	<i>smad3</i>	annotated to a part or bud	49.4434	22
<i>casp3a</i>	predicted	27.2208	340	<i>casp3</i>	predicted	49.0774	23

<i>map3k14b</i>	predicted	26.1706	350	<i>map3k14</i>	annotated to a part or bud	7.1186	312
<i>aldh1a2</i>	predicted	25.4504	354	<i>aldh1a2</i>	annotated to a part or bud	9.8933	263
<i>gli3</i>	predicted	25.0939	359	<i>gli3</i>	direct annotation to the hindlimb	38.3084	52
<i>bmpr1ba</i>	predicted	24.8781	361	<i>bmpr1b</i>	annotated to a part or bud	20.0326	131
<i>ihha</i>	predicted	24.0223	366	<i>ihh</i>	direct annotation to the hindlimb	33.7444	68
<i>ptch1</i>	predicted	23.8507	369	<i>ptch1</i>	predicted	38.8191	48
<i>bmpr1ab</i>	predicted	22.5899	386	<i>bmpr1a</i>	direct annotation to the hindlimb	35.5242	61
<i>vdrb</i>	predicted	22.4408	388	<i>vdr</i>	direct annotation to the hindlimb	15.8149	202
<i>shhb</i>	predicted	20.6403	404	<i>shh</i>	direct annotation to the hindlimb	44.4006	33
<i>bmpr1bb</i>	predicted	20.1116	408	<i>bmpr1b</i>	annotated to a part or bud	20.0326	131
<i>fsta</i>	predicted	17.8232	422	<i>fst</i>	annotated to a part or bud	27.0851	86
<i>hand2</i>	predicted	15.7932	437	<i>hand2</i>	direct annotation to the hindlimb	24.7719	101
<i>msx1a</i>	predicted	15.3589	442	<i>msx1</i>	predicted	26.6468	88
<i>cx43</i>	predicted	15.0787	444	<i>gjal</i>	predicted	28.3709	84
<i>vdra</i>	predicted	13.7382	454	<i>vdr</i>	direct annotation to the hindlimb	15.8149	202
<i>pax3a</i>	predicted	13.4122	458	<i>pax3</i>	direct annotation to the hindlimb	31.1909	76
<i>ihhb</i>	predicted	13.389	460	<i>ihh</i>	direct annotation to the hindlimb	33.7444	68

<i>lamc1</i>	predicted	13.3297	461	<i>lamc1</i>	direct annotation to the hindlimb	7.8573	299
<i>desma</i>	predicted	13.1643	462	<i>des</i>	annotated to a part or bud	14.3164	217
<i>coll1a1a</i>	predicted	12.9905	465	<i>coll1a1</i>	direct annotation to the hindlimb	43.3734	36
<i>lmx1bb</i>	predicted	11.9354	481	<i>lmx1b</i>	annotated to a part or bud	9.1733	271
<i>col2a1a</i>	predicted	11.8949	483	<i>col2a1</i>	direct annotation to the hindlimb	43.3479	37
<i>tbx4</i>	direct annotation to the pelvic fin	10.3408	496	<i>tbx4</i>	direct annotation to the hindlimb	10.3871	259
<i>hspg2</i>	predicted	9.6841	500	<i>hspg2</i>	direct annotation to the hindlimb	14.1006	220
<i>fras1</i>	predicted	9.0362	507	<i>fras1</i>	annotated to a part or bud	14.535	212
<i>coll1a2</i>	predicted	6.6976	539	<i>coll1a2</i>	direct annotation to the hindlimb	28.4794	83
<i>scube2</i>	predicted	6.4974	543	<i>scube2</i>	direct annotation to the hindlimb	3.4888	434
<i>coll1a1b</i>	predicted	6.2965	545	<i>coll1a1</i>	direct annotation to the hindlimb	14.3918	215
<i>coll1a1a</i>	predicted	5.9988	550	<i>coll1a1</i>	direct annotation to the hindlimb	14.3918	215
<i>coll1a1b</i>	predicted	5.9171	554	<i>coll1a1</i>	direct annotation to the hindlimb	43.3734	36
<i>rspo2</i>	annotated to a part or bud	4.0927	570	<i>rspo2</i>	direct annotation to the hindlimb	3.6481	429



<i>frem2a</i>	predicted	3.5366	577	<i>frem2</i>	annotated to a part or bud	1.5047	537
<i>pitx1</i>	predicted	2.8035	592	<i>pitx1</i>	direct annotation to the hindlimb	9.5149	267
<i>pax3b</i>	predicted	2.4374	595	<i>pax3</i>	direct annotation to the hindlimb	31.1909	76

Table 2.5. The 45 predicted genes of the pectoral fin module ranked and ordered according to the weighted degree of each gene. NA indicates ‘not available’ due to the ortholog not found in the mouse.

Zebrafish gene name	ZFIN identifier	Weighted degree	Rank in the module	Mouse ortholog name	Mouse ortholog status
<i>bmp4</i>	zdb-gene-980528-2059	35.5652	2	<i>bmp4</i>	direct annotation to the forelimb
<i>bmp2b</i>	zdb-gene-980526-474	34.3109	3	<i>bmp2</i>	direct annotation to the forelimb
<i>wnt3a</i>	zdb-gene-001106-1	31.4927	4	<i>wnt3a</i>	predicted
<i>fgf8a</i>	zdb-gene-990415-72	31.0094	5	<i>fgf8</i>	direct annotation to the forelimb
<i>gli2a</i>	zdb-gene-990706-8	27.2862	8	<i>gli2</i>	direct annotation to the forelimb
<i>wnt5b</i>	zdb-gene-980526-87	25.9169	10	<i>wnt5b</i>	not associated with the forelimb
<i>fgfr1a</i>	zdb-gene-980526-255	25.6442	12	<i>fgfr1</i>	direct annotation to the forelimb
<i>smad5</i>	zdb-gene-990603-9	25.5690	13	<i>smad5</i>	not associated with the forelimb
<i>gli1</i>	zdb-gene-030321-1	25.2858	14	<i>gli1</i>	not associated with the forelimb
<i>tcf7l1a</i>	zdb-gene-980605-30	24.8101	15	<i>tcf7l1</i>	not associated with the forelimb
<i>foxd3</i>	zdb-gene-980526-143	24.7462	16	<i>foxd3</i>	not associated with the forelimb
<i>ta</i>	zdb-gene-980526-437	23.7992	18	NA	not associated with the forelimb
<i>cdx4</i>	zdb-gene-980526-330	23.1331	20	<i>cdx4</i>	not associated with the forelimb
<i>pax2a</i>	zdb-gene-990415-8	22.2653	22	<i>pax2</i>	not associated with the forelimb
<i>ctnnb2</i>	zdb-gene-040426-2575	22.1606	24	NA	not associated with the forelimb
<i>ptch2</i>	zdb-gene-980526-44	21.7001	25	<i>ptch2</i>	not associated with the forelimb
<i>isl1</i>	zdb-gene-980526-112	21.5418	26	<i>isl1</i>	not associated with the forelimb
<i>fgf3</i>	zdb-gene-980526-178	20.9767	28	<i>fgf3</i>	not associated with the forelimb
<i>wnt4a</i>	zdb-gene-980526-352	20.4183	29	<i>wnt4</i>	predicted
<i>gpc4</i>	zdb-gene-011119-1	18.5224	31	<i>gpc4</i>	not associated with the forelimb
<i>ihha</i>	zdb-gene-051010-1	18.1064	32	<i>ihh</i>	direct annotation to the forelimb
<i>wnt11</i>	zdb-gene-990603-12	17.1389	34	<i>wnt11</i>	not associated with the forelimb
<i>zic2a</i>	zdb-gene-000710-4	15.2976	40	<i>zic2</i>	annotated to a part or bud
<i>nkx2.2a</i>	zdb-gene-980526-403	15.0866	41	<i>nkx2-2</i>	not associated with the forelimb

<i>dlx2a</i>	zdb-gene-980526-212	14.8123	43	<i>dlx2</i>	not associated with the forelimb
<i>pitx2</i>	zdb-gene-990714-27	14.7898	44	<i>pitx2</i>	not associated with the forelimb
<i>msx1a</i>	zdb-gene-980526-312	14.6308	45	<i>msx1</i>	predicted
<i>myf5</i>	zdb-gene-000616-6	13.7077	48	<i>myf5</i>	not associated with the forelimb
<i>esr2a</i>	zdb-gene-030116-2	13.2000	51	<i>esr2</i>	direct annotation to the forelimb
<i>dharma</i>	zdb-gene-990415-22	9.6753	59	NA	not associated with the forelimb
<i>tfap2a</i>	zdb-gene-011212-6	9.2007	62	<i>tfap2a</i>	direct annotation to the forelimb
<i>mecom</i>	NA	8.4715	89	<i>mecom</i>	annotated to a part or bud
<i>sox9b</i>	zdb-gene-001103-2	7.2558	104	<i>sox9</i>	direct annotation to the forelimb
<i>grem2b</i>	zdb-gene-030911-9	6.8164	108	<i>grem2</i>	not associated with the forelimb
<i>unm t31148</i>	zdb-gene-070117-1894	6.6863	109	NA	not associated with the forelimb
<i>dzip1</i>	zdb-gene-040526-1	6.4409	112	<i>dzip1</i>	not associated with the forelimb
<i>b3gat3</i>	zdb-gene-020419-3	6.0316	115	NA	not associated with the forelimb
<i>scube2</i>	zdb-gene-050302-80	5.3349	123	<i>scube2</i>	not associated with the forelimb
<i>hot</i>	zdb-gene-070117-2108	2.7511	160	NA	not associated with the forelimb
<i>frem2b</i>	zdb-gene-081119-4	2.6947	163	<i>frem2</i>	not associated with the forelimb
<i>unm s273</i>	zdb-gene-070117-864	2.4459	167	NA	not associated with the forelimb
<i>unm s245</i>	zdb-gene-070117-865	2.4459	167	NA	not associated with the forelimb
<i>eda</i>	zdb-gene-050107-6	2.3936	169	<i>eda</i>	not associated with the forelimb
<i>zmp:0000001138</i>	zdb-gene-140106-98	2.1176	172	NA	not associated with the forelimb
<i>mgt</i>	zdb-gene-070117-2188	1.0103	197	NA	not associated with the forelimb

Table 2.6. The top 50 predicted genes of the pelvic fin module ranked and ordered according to the weighted degree of each gene. NA indicates ‘not available’ due to the ortholog not found in the mouse. The full predicted gene list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles).

Zebrafish gene name	ZFIN identifier	Weighted degree	Rank in the module	Mouse ortholog name	Mouse ortholog status
<i>hsp90ab1</i>	zdb-gene-990415-95	129.9415	1	<i>hsp90ab1</i>	not associated with the hindlimb
<i>mapk3</i>	zdb-gene-040121-1	121.5328	2	<i>mapk3</i>	not associated with the hindlimb
<i>rhoab</i>	zdb-gene-040322-2	108.0837	3	<i>rhoa</i>	not associated with the hindlimb
<i>ctnmb1</i>	zdb-gene-980526-362	106.2749	4	<i>ctnmb1</i>	direct annotation to the hindlimb
<i>hsp90aa1.2</i>	zdb-gene-031001-3	105.1273	5	<i>hsp90aa1</i>	not associated with the hindlimb
<i>paics</i>	zdb-gene-030131-9762	104.4687	6	<i>paics</i>	not associated with the hindlimb
<i>gsk3b</i>	zdb-gene-990714-4	101.2293	7	<i>gsk3b</i>	not associated with the hindlimb
<i>cad</i>	zdb-gene-021030-4	97.8373	8	<i>cad</i>	not associated with the hindlimb
<i>cdc42</i>	zdb-gene-030131-8783	97.7752	9	NA	not associated with the hindlimb
<i>acta1b</i>	zdb-gene-030131-55	97.6682	10	<i>acta1</i>	not associated with the hindlimb
<i>smarca4a</i>	zdb-gene-030605-1	97.4336	11	<i>smarca4</i>	direct annotation to the hindlimb
<i>mapk1</i>	zdb-gene-030722-2	96.3956	12	<i>mapk1</i>	not associated with the hindlimb
<i>jupa</i>	zdb-gene-991207-22	95.0796	13	<i>jup</i>	not associated with the hindlimb
<i>cdk1</i>	zdb-gene-010320-1	93.9169	14	<i>cdk1</i>	not associated with the hindlimb
<i>kras</i>	NA	93.4030	15	<i>kras</i>	predicted
<i>rac1a</i>	zdb-gene-030131-5415	92.9107	16	<i>rac1</i>	direct annotation to the hindlimb
<i>mapk14a</i>	zdb-gene-010202-2	92.1789	17	<i>mapk14</i>	direct annotation to the hindlimb
<i>src</i>	zdb-gene-030131-3809	91.7256	18	<i>src</i>	direct annotation to the hindlimb
<i>fgfr1a</i>	zdb-gene-980526-255	91.2553	19	<i>fgfr1</i>	direct annotation to the hindlimb
<i>met</i>	zdb-gene-041014-1	90.8735	20	<i>met</i>	direct annotation to the hindlimb
<i>insrb</i>	zdb-gene-020503-4	90.6827	21	<i>insr</i>	not associated with the hindlimb
<i>si:ch211-163m16.1</i>	NA	84.6841	22	NA	not associated with the hindlimb
<i>actl6a</i>	zdb-gene-020419-36	83.9023	23	<i>actl6a</i>	not associated with the hindlimb
<i>tp53</i>	zdb-gene-990415-270	83.8035	24	<i>trp53</i>	direct annotation to the hindlimb
<i>pak2a</i>	zdb-gene-021011-2	83.2637	25	<i>pak2</i>	not associated with the hindlimb
<i>ehmt2</i>	zdb-gene-010501-6	82.9897	26	<i>ehmt2</i>	not associated with the hindlimb

<i>kdrl</i>	zdb-gene-000705-1	81.0991	27	NA	not associated with the hindlimb
<i>mtor</i>	zdb-gene-030131-2974	80.8121	28	<i>mtor</i>	annotated to a part or bud
<i>pkn2</i>	zdb-gene-061207-42	79.5943	29	<i>pkn2</i>	not associated with the hindlimb
<i>prkcb</i>	zdb-gene-040426-1178	79.5529	30	<i>prkcb</i>	not associated with the hindlimb
<i>hsp90aa1.1</i>	zdb-gene-990415-94	78.6325	31	<i>hsp90aa1</i>	not associated with the hindlimb
<i>pten</i>	zdb-gene-030616-47	78.5310	32	<i>pten</i>	annotated to a part or bud
<i>kita</i>	zdb-gene-980526-464	77.2229	33	<i>kit</i>	not associated with the hindlimb
<i>akt2</i>	zdb-gene-031007-5	77.0797	34	<i>akt2</i>	not associated with the hindlimb
<i>igflra</i>	zdb-gene-020503-1	76.5270	35	<i>igflr</i>	not associated with the hindlimb
<i>bmp4</i>	zdb-gene-980528-2059	76.2150	36	<i>bmp4</i>	direct annotation to the hindlimb
<i>igflrb</i>	zdb-gene-020503-2	75.0886	37	<i>igflr</i>	not associated with the hindlimb
<i>rap1b</i>	zdb-gene-030131-9662	73.3502	38	<i>rap1b</i>	not associated with the hindlimb
<i>rac2</i>	zdb-gene-040625-27	73.3136	39	<i>rac2</i>	not associated with the hindlimb
<i>hsa9</i>	zdb-gene-030828-12	72.3324	40	<i>hsa9</i>	not associated with the hindlimb
<i>hdac1</i>	zdb-gene-020419-32	69.4428	41	<i>hdac1</i>	not associated with the hindlimb
<i>pola1</i>	zdb-gene-030114-9	69.0229	42	<i>pola1</i>	not associated with the hindlimb
<i>actc1a</i>	zdb-gene-040520-4	68.6027	43	<i>actc1</i>	not associated with the hindlimb
<i>top2b</i>	zdb-gene-041008-136	68.4413	44	<i>top2b</i>	not associated with the hindlimb
<i>bmp2b</i>	zdb-gene-980526-474	68.2429	45	<i>bmp2</i>	predicted
<i>insra</i>	zdb-gene-020503-3	68.1565	46	<i>insr</i>	not associated with the hindlimb
<i>flt1</i>	zdb-gene-050407-1	68.0761	47	<i>flt1</i>	not associated with the hindlimb
<i>ralbb</i>	zdb-gene-040625-121	67.8323	48	<i>ralb</i>	not associated with the hindlimb
<i>btk</i>	zdb-gene-070531-1	67.4762	49	NA	not associated with the hindlimb
<i>rac1b</i>	zdb-gene-060312-45	67.0359	50	<i>rac1</i>	direct annotation to the hindlimb

Table 2.7. The 18 predicted genes of the forelimb module ranked and ordered according to the weighted degree of each gene. NA indicates ‘not available’ due to the ortholog not found in the zebrafish.

Mouse gene name	MGI identifier	Weighted degree	Rank in the module	Zebrafish ortholog names	Zebrafish ortholog status
<i>smad4</i>	mgi:894293	34.5294	6	<i>smad4a, smad4b</i>	not associated with the pectoral fin
<i>bmp7</i>	mgi:103302	34.2132	7	<i>bmp7a</i>	direct annotation to the pectoral fin
<i>wnt3a</i>	mgi:98956	30.9124	14	<i>wnt3a</i>	predicted
<i>nog</i>	mgi:104327	29.3757	19	NA	not associated with the pectoral fin
<i>wnt4</i>	mgi:98957	27.9668	23	<i>wnt4a</i>	predicted
<i>ptch1</i>	mgi:105373	27.3374	24	<i>ptch1</i>	direct annotation to the pectoral fin
<i>wnt1</i>	mgi:98953	26.4361	25	<i>wnt1</i>	not associated with the pectoral fin
<i>bmpr1a</i>	mgi:1338938	26.0359	26	<i>bmpr1ab, bmp1aa</i>	not associated with the pectoral fin
<i>chrd</i>	mgi:1313268	24.6081	28	<i>chrd</i>	not associated with the pectoral fin
<i>msx1</i>	mgi:97168	23.4172	31	<i>msx1a</i>	predicted
<i>msx2</i>	mgi:97169	22.8361	32	<i>msx2b, msx2a</i>	not associated with the pectoral fin
<i>fst</i>	mgi:95586	18.1062	46	<i>fsta, fstb</i>	not associated with the pectoral fin
<i>wnt7b</i>	mgi:98962	17.7777	47	<i>wnt7ba, wnt7bb</i>	not associated with the pectoral fin
<i>dlx5</i>	mgi:101926	16.1115	51	<i>dlx5a</i>	annotated to a part or bud
<i>wnt9a</i>	mgi:2446084	15.9927	52	<i>wnt9a</i>	not associated with the pectoral fin
<i>foxc2</i>	mgi:1347481	13.9395	61	NA	not associated with the pectoral fin
<i>nkx3-2</i>	mgi:108015	13.6063	63	<i>nkx3.2</i>	not associated with the pectoral fin
<i>tbx4</i>	mgi:102556	8.7400	84	<i>tbx4</i>	not associated with the pectoral fin

Table 2.8. The 32 predicted genes of the hindlimb module ranked and ordered according to the weighted degree of each gene. NA indicates ‘not available’ due to the ortholog not found in the zebrafish.

Mouse gene name	MGI identifier	Weighted degree	Rank in the module	Zebrafish ortholog names	Zebrafish ortholog status
<i>hras</i>	mgi:96224	72.8342	2	<i>hrasb</i>	predicted
<i>kras</i>	mgi:96680	70.7748	5	<i>kras</i>	predicted
<i>myc</i>	mgi:97250	66.8072	7	<i>mycb, myca</i>	not associated with the pelvic fin
<i>fos</i>	mgi:95574	64.7904	9	<i>fosab, fosaa</i>	not associated with the pelvic fin
<i>tgfb1</i>	mgi:98725	54.2763	15	<i>tgfb1b, tgfb1a</i>	not associated with the pelvic fin
<i>igf1</i>	mgi:96432	52.9341	17	<i>igf1</i>	not associated with the pelvic fin
<i>fgf2</i>	mgi:95516	51.2651	19	<i>fgf2</i>	not associated with the pelvic fin
<i>casp3</i>	mgi:107739	49.0774	23	<i>casp3a</i>	predicted
<i>bmp7</i>	mgi:103302	48.2316	26	<i>bmp7a</i>	predicted
<i>wnt5a</i>	mgi:98958	48.1588	27	<i>wnt5a</i>	not associated with the pelvic fin
<i>smad2</i>	mgi:108051	45.6514	30	<i>smad2</i>	predicted
<i>pdgfra</i>	mgi:97530	45.2074	31	<i>pdgfra</i>	not associated with the pelvic fin
<i>bmp2</i>	mgi:88177	45.0361	32	<i>bmp2a, bmp2b</i>	predicted
<i>nog</i>	mgi:104327	39.7063	46	NA	not associated with the pelvic fin
<i>ptch1</i>	mgi:105373	38.8191	48	<i>ptch1</i>	predicted
<i>lef1</i>	mgi:96770	36.7074	53	<i>lef1</i>	predicted
<i>wnt1</i>	mgi:98953	35.9468	60	<i>wnt1</i>	not associated with the pelvic fin
<i>wnt4</i>	mgi:98957	35.4014	62	<i>wnt4a</i>	predicted
<i>mmp2</i>	mgi:97009	35.3982	63	<i>mmp2</i>	not associated with the pelvic fin
<i>spp1</i>	mgi:98389	33.6944	70	NA	not associated with the pelvic fin
<i>dcn</i>	mgi:94872	33.0020	73	<i>dcn</i>	not associated with the pelvic fin
<i>mmp14</i>	mgi:101900	30.9303	77	<i>mmp14a, mmp14b</i>	not associated with the pelvic fin
<i>chrd</i>	mgi:131326 8	29.8136	81	<i>chrd</i>	not associated with the pelvic fin
<i>gjal</i>	mgi:95713	28.3709	84	<i>cx43</i>	predicted
<i>msx1</i>	mgi:97168	26.6468	88	<i>msx1a</i>	predicted
<i>twist1</i>	mgi:98872	25.5266	93	<i>twist1a, twist1b</i>	not associated with the pelvic fin
<i>bglap</i>	mgi:88156	24.8621	99	NA	not associated with the pelvic fin
<i>bglap2</i>	mgi:88157	24.8447	100	NA	not associated with the pelvic fin
<i>sp7</i>	mgi:215356 8	21.0093	119	<i>sp7</i>	not associated with the pelvic fin
<i>foxc2</i>	mgi:134748 1	17.4167	192	NA	not associated with the pelvic fin

<i>nkx3-2</i>	mgi:108015	15.3314	207	<i>nkx3.2</i>	not associated with the pelvic fin
<i>hapln1</i>	mgi:133700 6	8.9023	279	<i>hapln1a</i> , <i>hapln1b</i>	not associated with the pelvic fin



Table 2.9. Some of the enriched Biological Process terms from the Gene Ontology and Uberon terms for the mouse orthologs of the pectoral fin module-specific genes that are related to novel anatomical entities generated in tetrapods during fin to limb transition. The terms are organized into specific anatomical regions.

Anatomical region	Term Identifier	Term name	P-value
Lung	uberon 0002167	right lung	0.00023623
	uberon 0002168	left lung	0.00120309
	uberon 0003512	lung blood vessel	0.04982657
	GO:0030324	lung development	0.00446791
Neck and lower jaw	uberon 0001708	jaw skeleton	2.93E-05
	uberon 0003451	lower jaw incisor	0.00017578
	uberon 0002413	cervical vertebra	0.00380992
	uberon 0003216	hard palate	0.00029064
	uberon 0001716	secondary palate	0.00151298
	GO:0060021	palate development	0.00113011
	GO:0061153	trachea gland development	0.01600806
Face and hair	uberon 0001456	face	5.43E-08
	uberon 0005600	crus commune	9.42E-05
	uberon 0001690	ear	0.00026285
	uberon 0000004	nose	0.00090415
	uberon 0001711	eyelid	2.75E-06
	uberon 0034772	margin of eyelid	0.03232832
	uberon 0002073	hair follicle	0.02808325
	uberon 0010514	strand of duvet hair	0.02594653
	GO:0001942	hair follicle development	0.03174493
	GO:0060789	hair follicle placode formation	0.03176245
	GO:0061029	eyelid development in camera-type eye	0.07754317
Other	uberon 0002544	digit	0.00066645

Table 2.10. Some of the enriched Biological Process terms from the Gene Ontology and Uberon terms for the mouse orthologs of the pelvic fin module-specific genes that are related to novel anatomical entities generated in tetrapods during fin to limb transition. The terms are organized into specific anatomical regions.

Anatomical region	Term Identifier	Term name	P-value
Lung	uberon 0001004	respiratory system	3.10E-06
	uberon 0002012	pulmonary artery	4.49E-06
	uberon 0002048	lung	1.66E-05
	uberon 0003512	lung blood vessel	6.31E-05
	uberon 0006524	alveolar system	0.00031454
	uberon 0008870	pulmonary alveolar parenchyma	0.00034866
	uberon 0000117	respiratory tube	0.00063848
	uberon 0004785	respiratory system mucosa	0.00465107
	GO:0030324	lung development	3.55E-04
GO:0060425	lung morphogenesis	0.0083659	
Neck and lower jaw	uberon 0003216	hard palate	0.00044958
	uberon 0002370	thymus	0.00057905
	GO:0048538	thymus development	4.03E-04
	GO:0060021	palate development	0.00391519
	GO:0060440	trachea formation	0.00614484
	GO:0060017	parathyroid gland development	0.00614484
Face and hair	uberon 0001818	tarsal gland	0.00026935
	uberon 0001711	eyelid	0.00043278
	uberon 0000004	nose	0.00707171
	uberon 0001681	nasal bone	0.01712347
	uberon 0002073	hair follicle	0.00052472
	uberon 0010512	strand of guard hair	0.00311381
	GO:0061029	eyelid development in camera-type eye	0.00211434
	GO:0001942	hair follicle development	3.29E-06
Other	uberon 0001987	placenta	2.69E-05
	uberon 0003946	placenta labyrinth	0.00010402

Table 2.11. Some of the enriched Biological Process terms from the Gene Ontology and Uberon terms that are related to fin to limb transition for the zebrafish orthologs of the forelimb module-specific genes. The terms are organized into specific anatomical regions.

Anatomical region	Term Identifier	Term name	P-value
Fin related	uberont 4000164	caudal fin	0.00325469
	GO:0035118	embryonic pectoral fin morphogenesis	1.60E-05
Other	uberont 0000033	head	2.96E-07
	uberont 0005886	post-hyoid pharyngeal arch skeleton	0.03545473
	uberont 0001708	jaw skeleton	0.0342243
	uberont 0008895	splanchnocranium	0.01593721
	uberont 0002280	otolith	0.01730733
	GO:0060037	pharyngeal system development	0.02837219

Table 2.12. Some of the enriched Biological Process terms from the Gene Ontology and Uberon terms that are related to fin to limb transition for the zebrafish orthologs of the hindlimb module-specific genes. The terms are organized into specific anatomical regions.

Anatomical region	Term Identifier	Term name	P-value
Fin related	uberon 4000164	caudal fin	0.0046263
	uberon 0012438	blastema of regenerating fin/limb	0.00591299
	uberon 2001456	pectoral fin endoskeletal disc	0.0328624
	GO:0035118	embryonic pectoral fin morphogenesis	0.00553237
	GO:0031101	fin regeneration	0.00239174
Other	uberon 0000033	head	2.90E-15
	uberon 0005886	post-hyoid pharyngeal arch skeleton	3.02E-05
	uberon 0001708	jaw skeleton	0.00287797
	uberon 0005884	hyoid arch skeleton	0.01978595
	uberon 0011611	ceratohyal bone	0.0328624
	uberon 0002280	otolith	0.01008715
	GO:0060037	pharyngeal system development	0.02837219
	GO:0048701	embryonic cranial skeleton morphogenesis	1.03E-05
	GO:0048703	embryonic viscerocranium morphogenesis	9.92E-04
	GO:0060021	palate development	0.01953552
	GO:0060037	pharyngeal system development	0.01986392
	GO:0048840	otolith development	0.03165263

## Figures

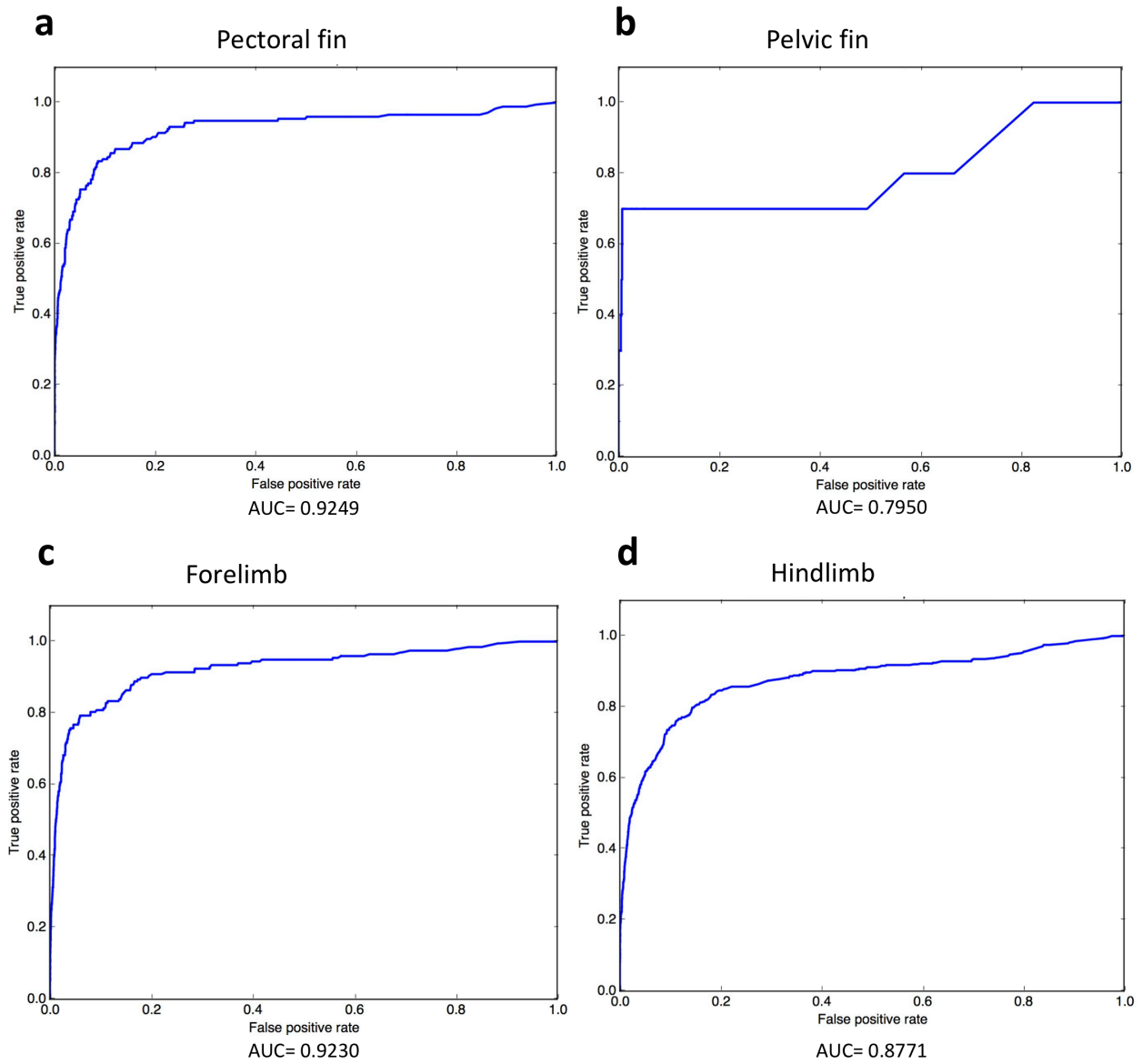


Figure 2.1. The ROC curves for the four anatomical entities that were generated during network-based candidate gene prediction evaluations.

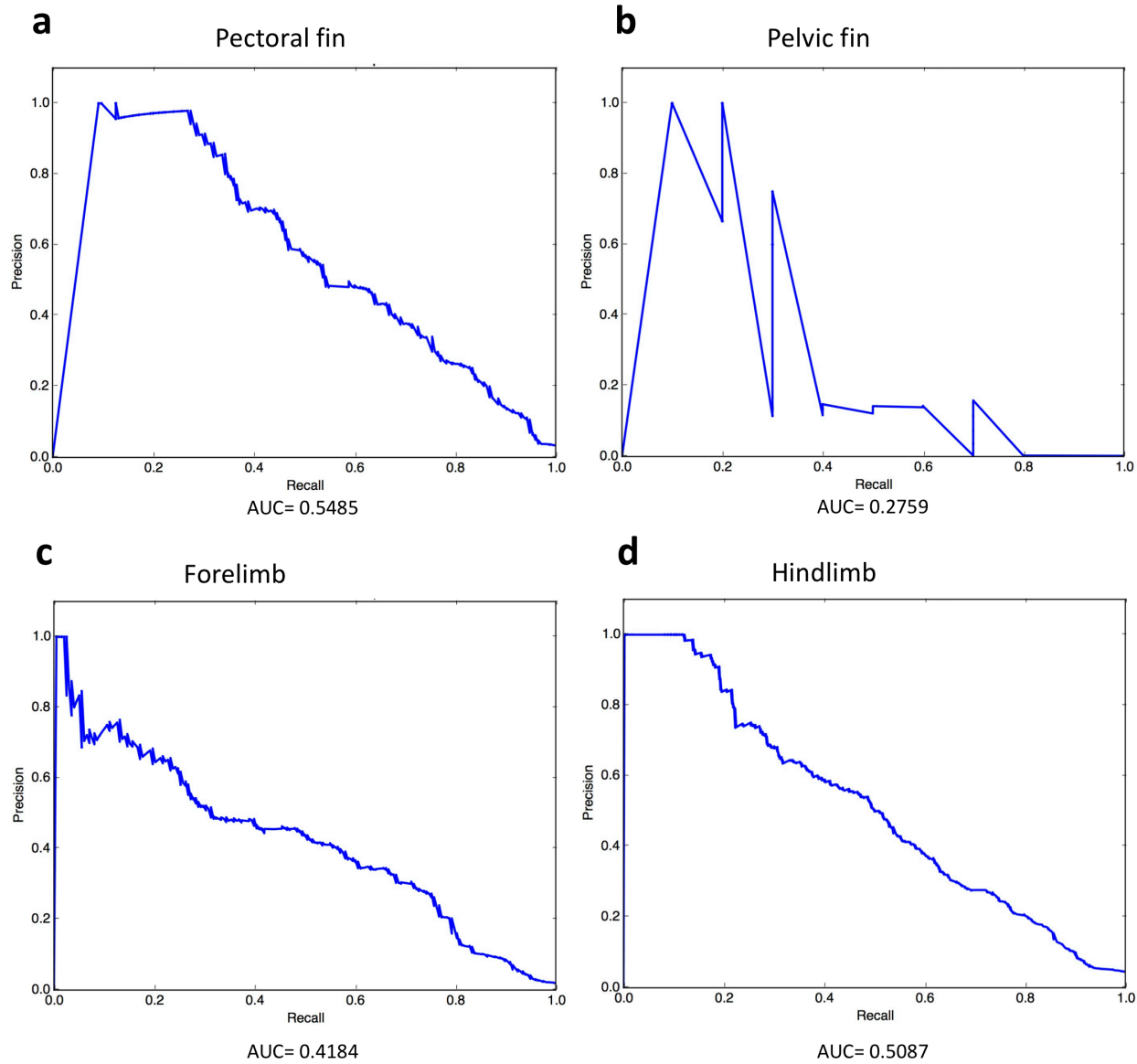


Figure 2.2. The precision-recall curves for the four anatomical entities that were generated during network-based candidate gene prediction evaluations.







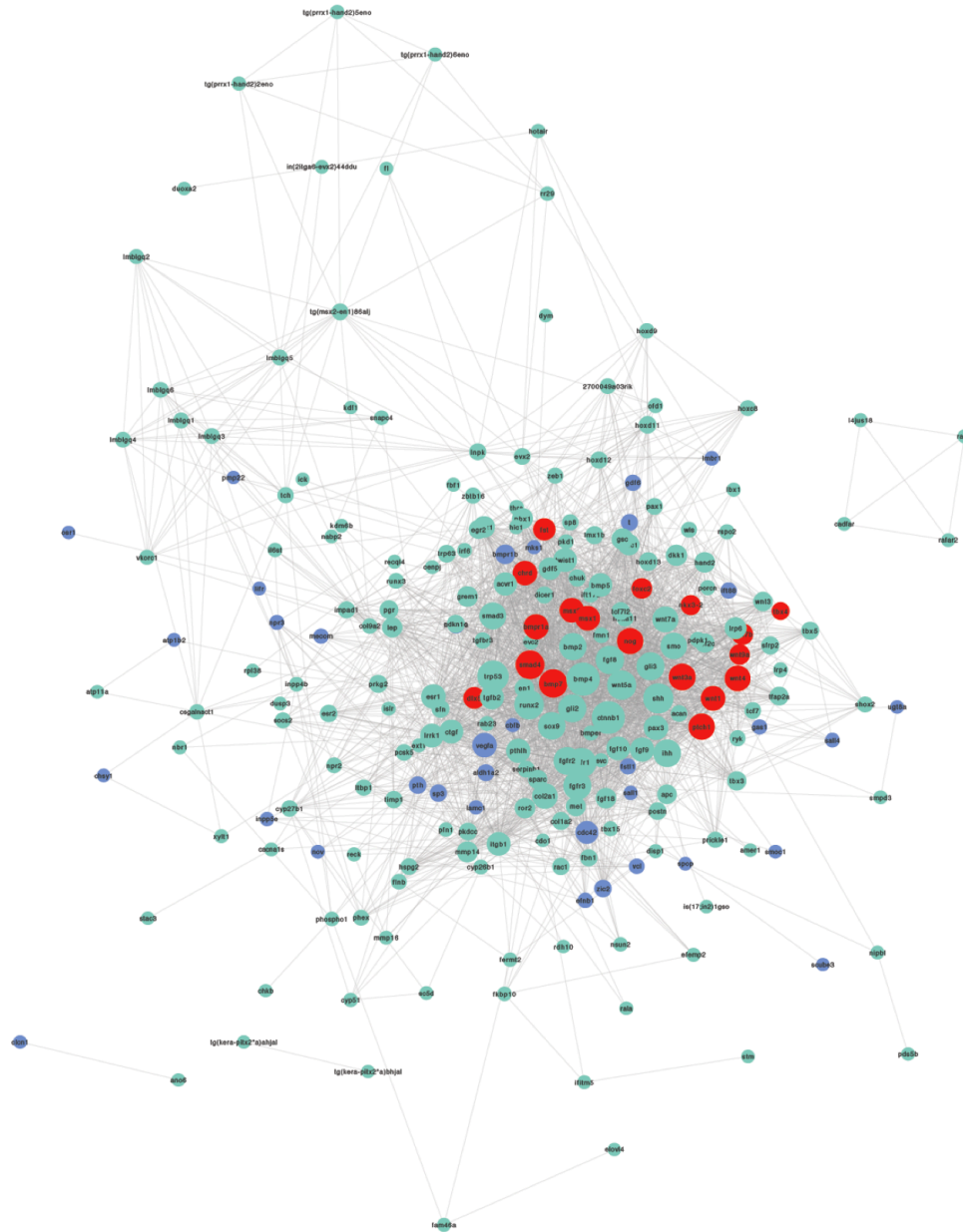


Figure 2.5. Visualization of the forelimb module including genes with direct annotations to the forelimb (green), genes annotated only to the forelimb parts or developmental precursors (blue), and predicted genes (red). Node size is proportional to the degree (number of interactions) of the gene. An interactive version of this module is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) as a Cytoscape network file.

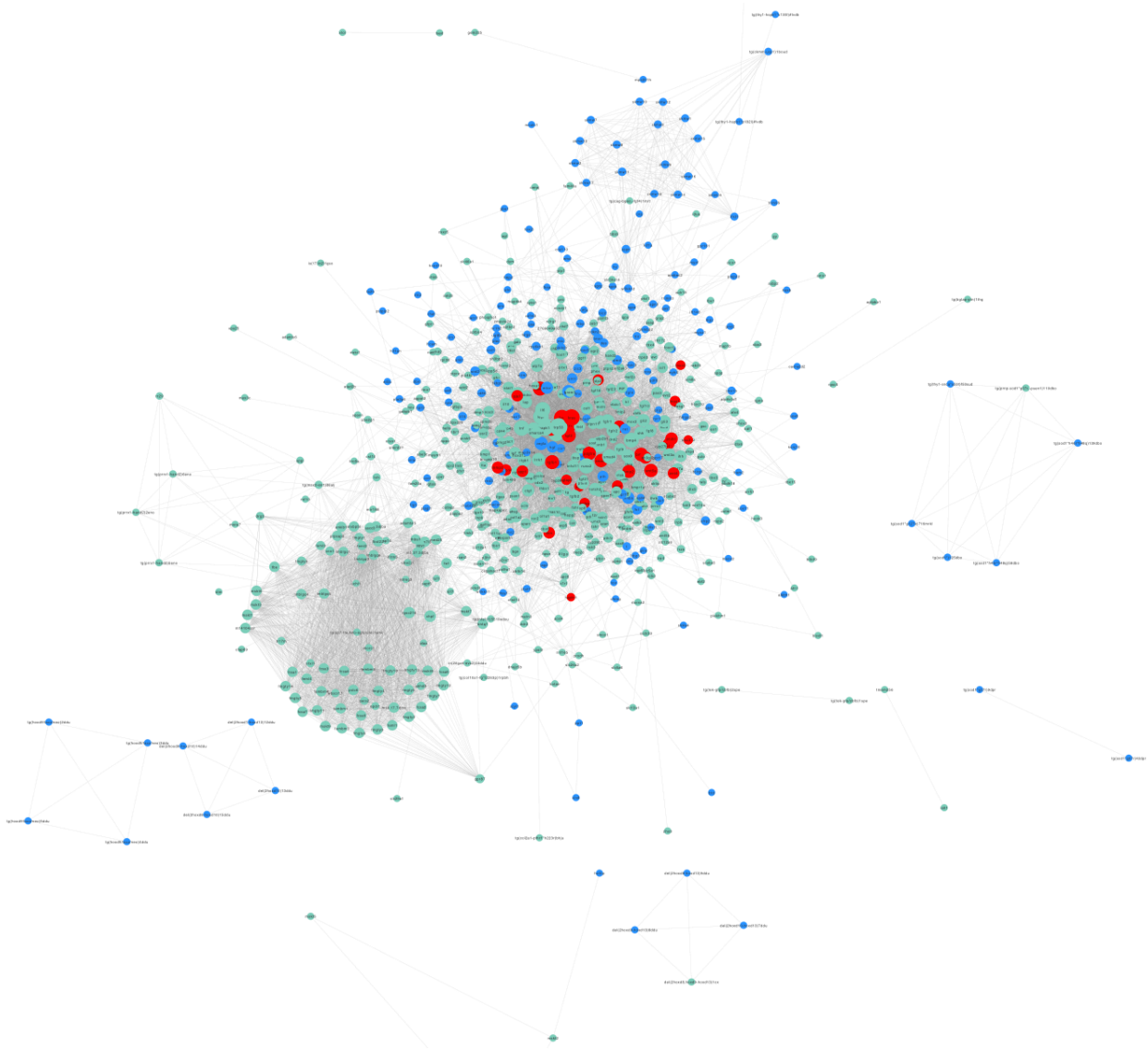


Figure 2.6. Visualization of the hindlimb module including genes with direct annotations to the hindlimb (green), genes annotated only to the hindlimb parts or developmental precursors (blue), and predicted genes (red). Node size is proportional to the degree (number of interactions) of the gene. An interactive version of this module is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) as a Cytoscape network file.

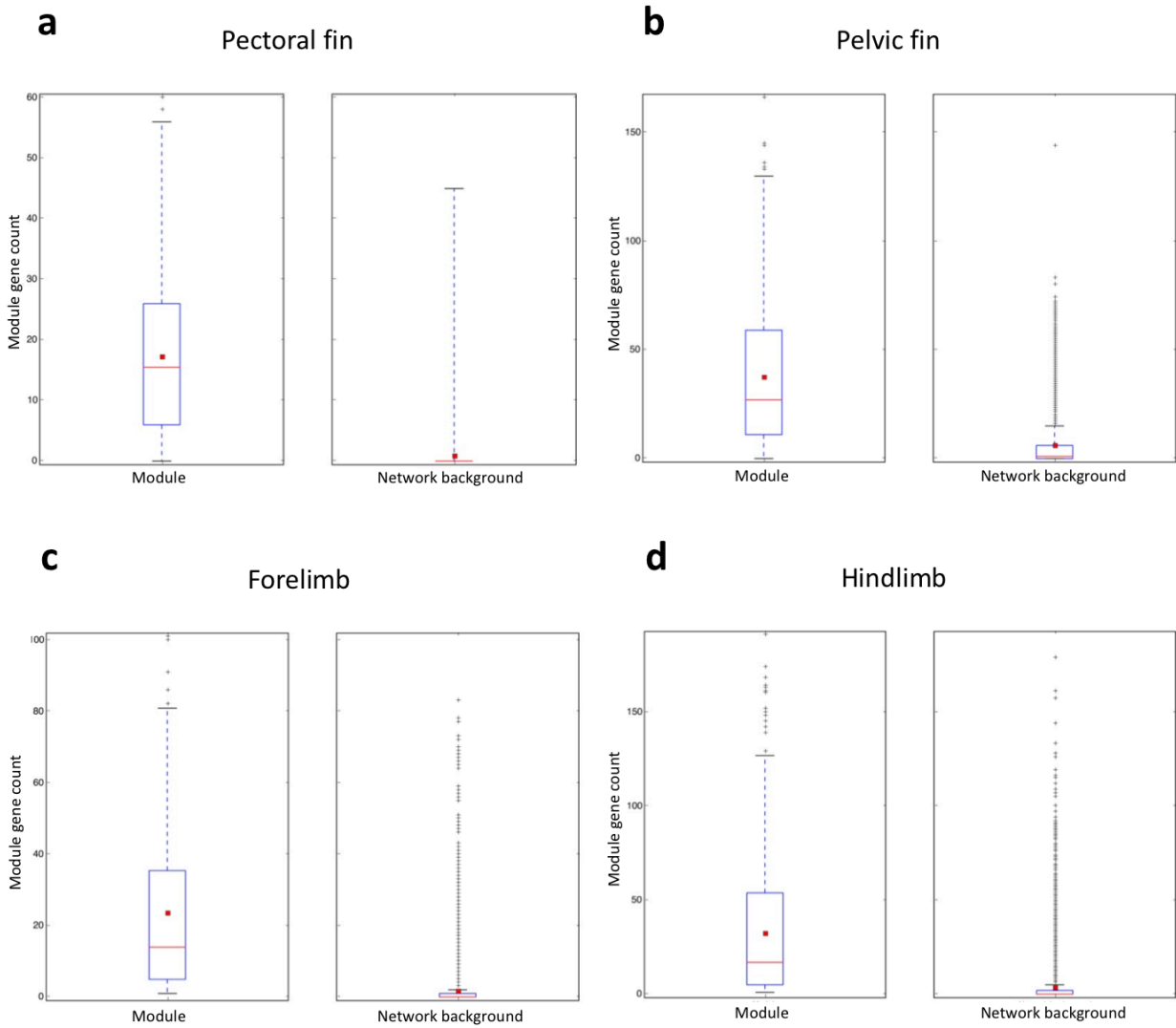


Figure 2.7. Boxplot comparisons of the distributions of module gene counts in the immediate neighborhood of module genes *versus* network background genes for each anatomical entity. In the boxplots, the red line and the square represent the median and mean, respectively.



Figure 2.8. Venn diagram showing the number of pectoral fin module-specific genes, conserved genes, and forelimb module-specific genes.

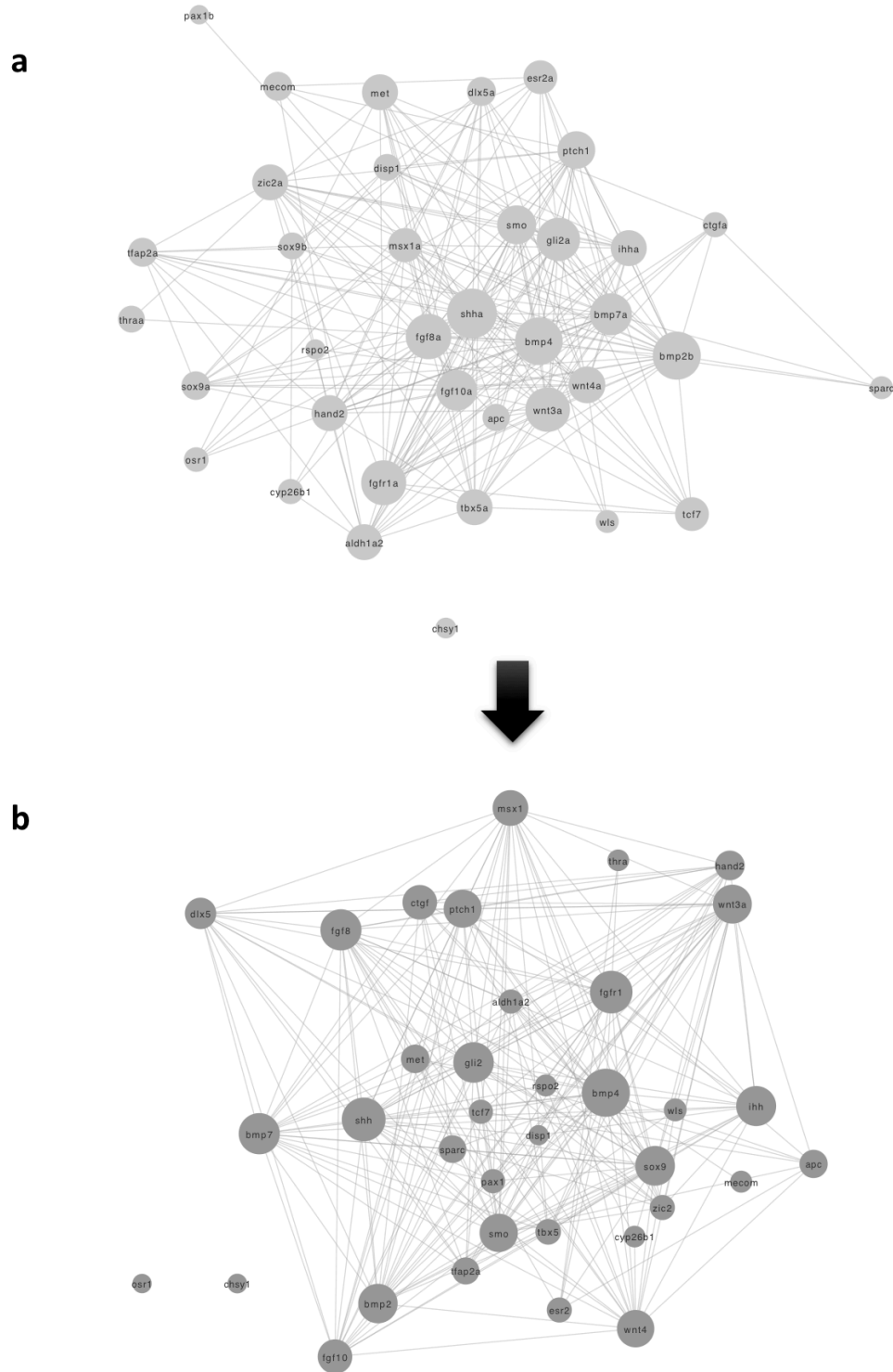


Figure 2.9. Extractions of the 37 conserved genes from (a) the pectoral fin module and (b) the forelimb module. Node size is proportional to the degree (number of interactions) of the gene. The arrow represents the direction of modular evolution.

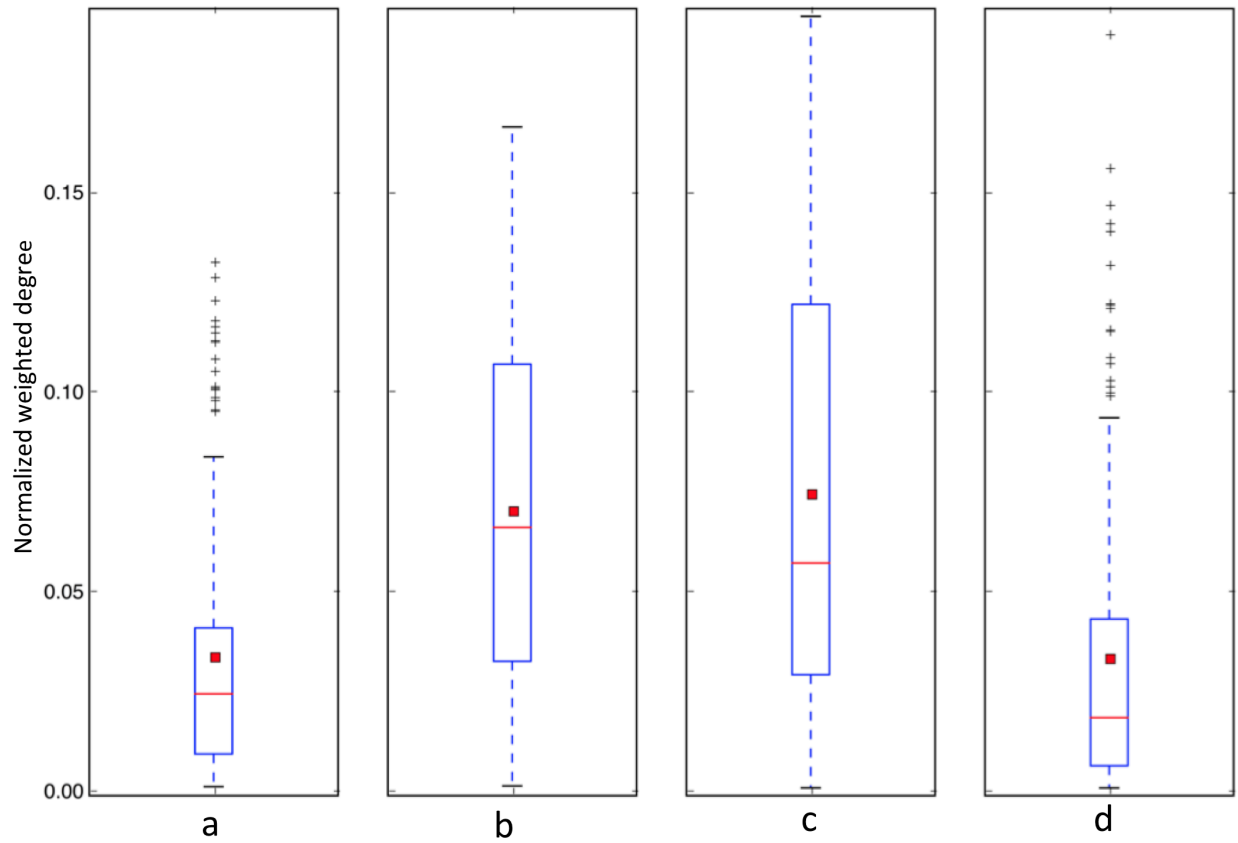


Figure 2.10. Boxplot comparison of normalized weighted degree distributions for (a) pectoral fin module-specific genes, (b) pectoral fin conserved genes, (c) forelimb conserved genes, and (d) forelimb module-specific genes. In the boxplots, the red line and the square represent the median and mean, respectively.

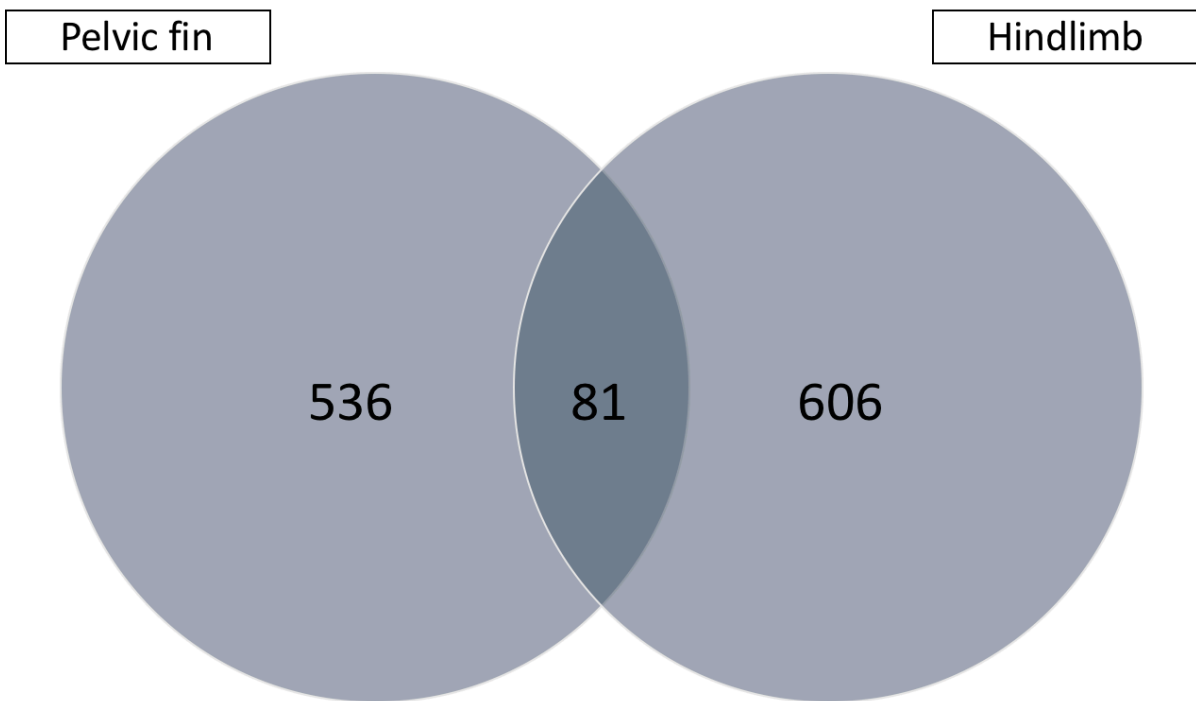


Figure 2.11. Venn diagram showing the number of pelvic fin module-specific genes, conserved genes, and hindlimb module-specific genes.

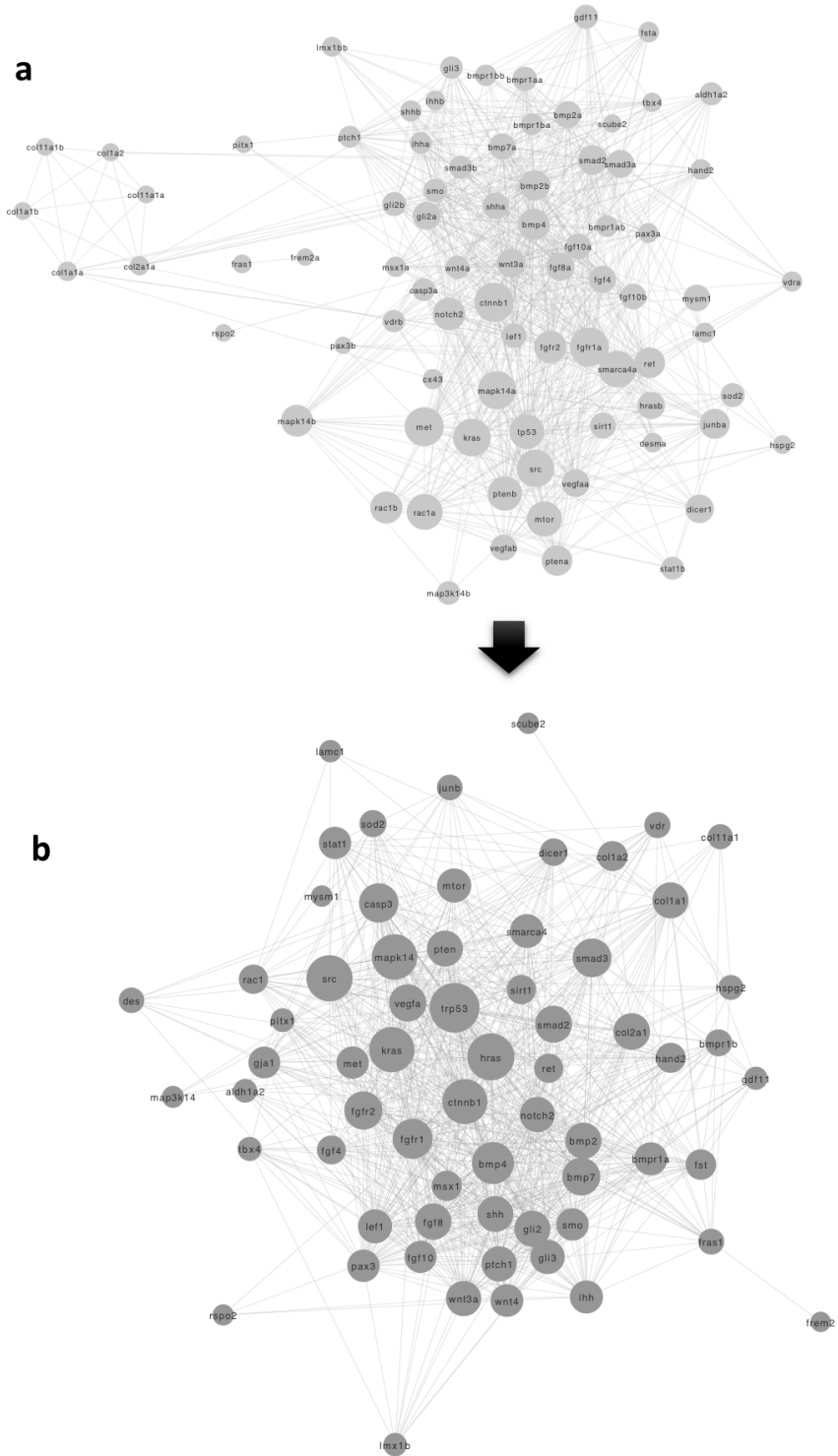


Figure 2.12. Extractions of the 81 conserved genes from (a) the pelvic fin module and (b) the hindlimb module. Node size is proportional to the degree (number of interactions) of the gene. The arrow represents the direction of modular evolution.



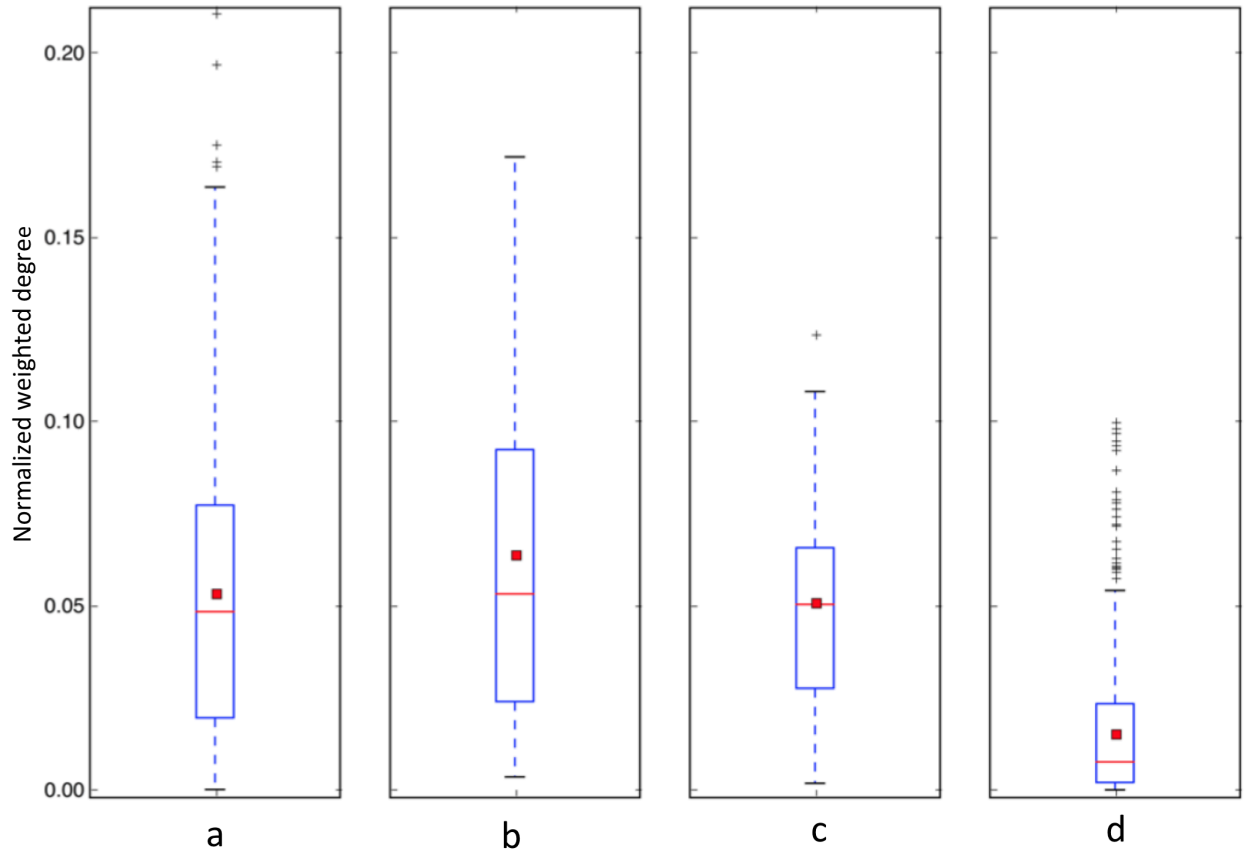


Figure 2.13. Boxplot comparison of normalized weighted degree distributions for (a) pelvic fin module-specific genes, (b) pelvic fin conserved genes, (c) hindlimb conserved genes, and (d) hindlimb module-specific genes. In the boxplots, the red line and the square represent the median and mean, respectively.

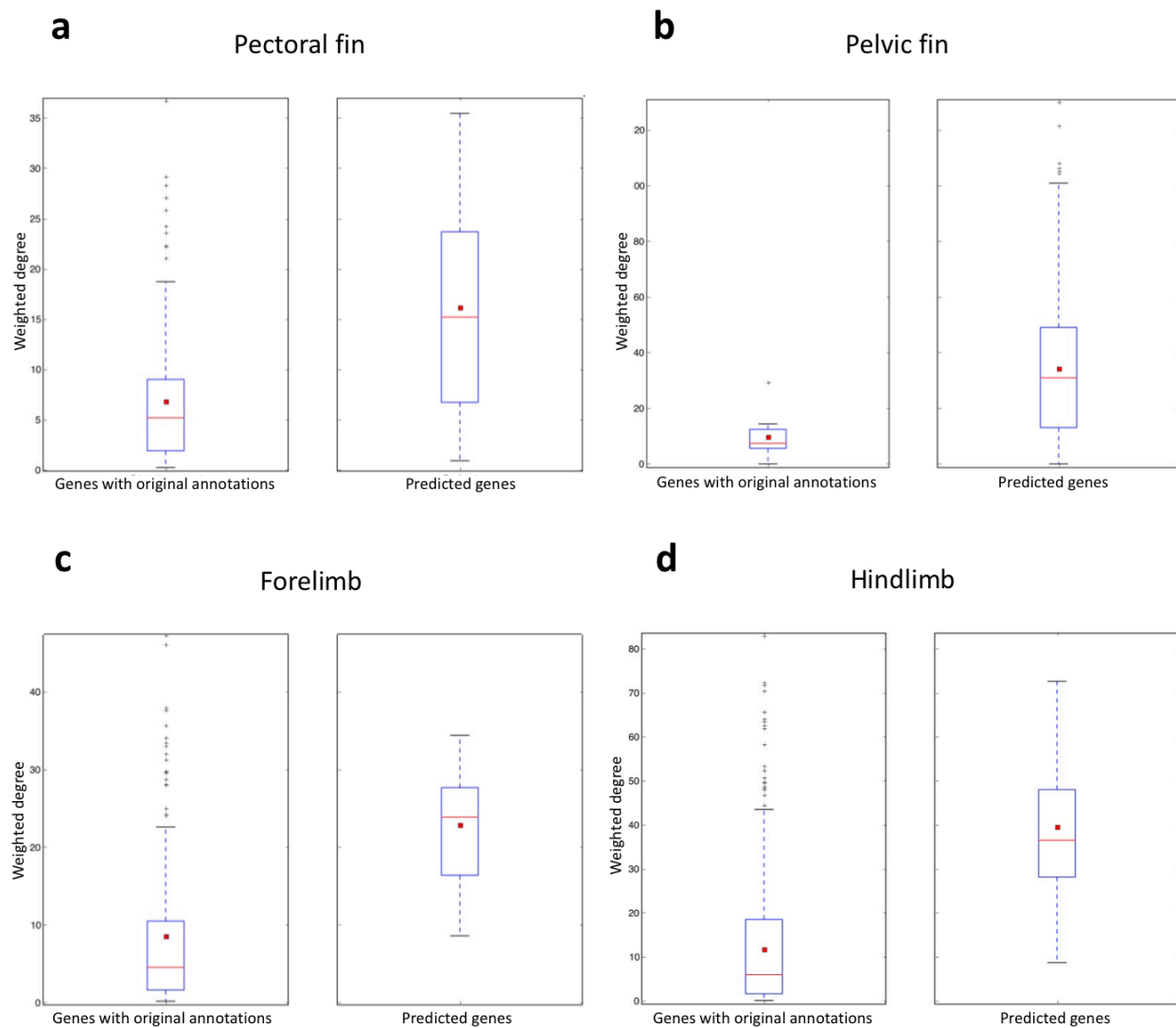


Figure 2.14. The boxplot comparisons of the weighted degree distributions for the predicted genes *versus* genes with original annotations for each module. In the boxplots, the red line and the square represent the median and mean, respectively.



## Supplementary Tables

Supplementary Table S2.1. The genes with original annotations that were lost due to network cutoff or isolation in the network

Pectoral fin	Pelvic fin	Forelimb	Hindlimb
<p><b>Lost due to network cutoff: 17</b></p> <ol style="list-style-type: none"> <li><i>bcl2l16</i></li> <li><i>bot</i></li> <li><i>don</i></li> <li><i>ful</i></li> <li><i>gpatch3</i></li> <li><i>hmx4</i></li> <li><i>mko</i></li> <li><i>not found</i></li> <li><i>si:ch211-185a18.2</i></li> <li><i>tmem260</i></li> <li><i>unm au21</i></li> <li><i>unm m572</i></li> <li><i>unm ti205</i></li> <li><i>unm tm136</i></li> <li><i>unm to219</i></li> <li><i>wan</i></li> <li><i>zon</i></li> </ol> <p><b>Lost due to isolation: 6</b></p> <ol style="list-style-type: none"> <li><i>ap1g1</i></li> <li><i>dul</i></li> <li><i>gaz</i></li> <li><i>nokr2</i></li> <li><i>nrg2a</i></li> <li><i>perp</i></li> </ol>	<p><b>Lost due to network cutoff: 3</b></p> <ol style="list-style-type: none"> <li><i>mir223</i></li> <li><i>sub</i></li> <li><i>wan</i></li> </ol> <p><b>Lost due to isolation: 0</b></p>	<p><b>Lost due to network cutoff: 25</b></p> <ol style="list-style-type: none"> <li><i>am</i></li> <li><i>ccd</i></li> <li><i>cl</i></li> <li><i>dbf</i></li> <li><i>del(2hoxd1-hoxd10)26ddu</i></li> <li><i>del(6dlx6-dlx5)1tlu</i></li> <li><i>etn3</i></li> <li><i>fts</i></li> <li><i>is(in8b2-8b3.1;6c1)1tshir</i></li> <li><i>lgl</i></li> <li><i>mdga2</i></li> <li><i>mhdaali18</i></li> <li><i>mirc1</i></li> <li><i>morc2a</i></li> <li><i>os</i></li> <li><i>pf</i></li> <li><i>t(7;18)50h</i></li> <li><i>tg(cag-mrfp1,-sox9,-egfp)1haak</i></li> <li><i>tg(col2a1-mef2c/vp16)1eno</i></li> <li><i>tg(hand2)#tshir</i></li> <li><i>tg(pgk1-fgf2)15cofn</i></li> <li><i>tg(plp1-lmnbl)1108qsp</i></li> <li><i>tg(prrx1-sox9,-lacz)1haak</i></li> <li><i>tg(tyr,col2a1-trpv4*r594h)#dhco</i></li> <li><i>vsd</i></li> </ol> <p><b>Lost due to isolation: 17</b></p> <ol style="list-style-type: none"> <li><i>btd</i></li> <li><i>cdk20</i></li> <li><i>cpox</i></li> <li><i>del(5d5mit73-d5mit351)5jcs</i></li> <li><i>lgi4</i></li> <li><i>nkx6-1</i></li> <li><i>npat</i></li> <li><i>ostm1</i></li> <li><i>pappa2</i></li> </ol>	<p><b>Lost due to network cutoff: 101</b></p> <ol style="list-style-type: none"> <li><i>4933430i17rik</i></li> <li><i>ano5</i></li> <li><i>aspb</i></li> <li><i>b2b1594clo</i></li> <li><i>b2b2187clo</i></li> <li><i>bolt</i></li> <li><i>cby</i></li> <li><i>cl</i></li> <li><i>clec11a</i></li> <li><i>dbf</i></li> <li><i>del(6dlx6-dlx5)1tlu</i></li> <li><i>dh</i></li> <li><i>dmpy</i></li> <li><i>dp(16cbr1-fam3b)1rhr</i></li> <li><i>fts</i></li> <li><i>gnd</i></li> <li><i>hacd1</i></li> <li><i>hxd</i></li> <li><i>hydro</i></li> <li><i>igf1sl1</i></li> <li><i>inad</i></li> <li><i>is(in8b2-8b3.1;6c1)1tshir</i></li> <li><i>klhl41</i></li> <li><i>lgl</i></li> <li><i>lx</i></li> <li><i>lz</i></li> <li><i>map3k20</i></li> <li><i>mhdaali18</i></li> <li><i>mir140</i></li> <li><i>mir92-1</i></li> <li><i>mirc1</i></li> <li><i>mpc234h</i></li> <li><i>nad</i></li> <li><i>nma</i></li> <li><i>nmf419</i></li> <li><i>not found</i></li> <li><i>ogfod1</i></li> <li><i>os</i></li> <li><i>pf</i></li> <li><i>pl</i></li> <li><i>pma</i></li> <li><i>ps</i></li> </ol>

		<p>10. <i>pex10</i>  11. <i>rnf165</i>  12. <i>scx</i>  13. <i>tg(sod1*g127x)716</i>  <i>mrkl</i>  14. <i>tram2</i>  15. <i>trpv4</i>  16. <i>uchl1</i>  17. <i>vps54</i></p>	<p>43. <i>scarf2</i>  44. <i>skc3</i>  45. <i>srn</i>  46. <i>ssq</i>  47. <i>t(7;18)50h</i>  48. <i>tal</i>  49. <i>tenm4</i>  50. <i>tg(acta1-il15*)11650lsq</i>  51. <i>tg(ar*100q)c25als</i>  52. <i>tg(ar*100q)c32als</i>  53. <i>tg(b19-rnai:il3)241ckn</i>  54. <i>tg(cag-dsred2/rnai:tardbp)6z xu</i>  55. <i>tg(cd4-npm/alk)n1ingh</i>  56. <i>tg(ckmm-cav3)1ysu</i>  57. <i>tg(cnp-gpr17)1qrlu</i>  58. <i>tg(col1a2-npr2*)28keoz</i>  59. <i>tg(colla1-ifitm5*)1brle</i>  60. <i>tg(colla1)73prc</i>  61. <i>tg(ctnnb1)1efu</i>  62. <i>tg(eef1a1-gnas*r201c)184pab i</i>  63. <i>tg(epo*)458mym</i>  64. <i>tg(gfap-il6)g167lms</i>  65. <i>tg(gfap-il6)g16lms</i>  66. <i>tg(gfap-il6)g369lms</i>  67. <i>tg(h2-k-fosl1)1wag</i>  68. <i>tg(igkv3-5*-myc)#plbe</i>  69. <i>tg(lckil4)1315dbl</i>  70. <i>tg(mbp-cadm4*)#pele</i>  71. <i>tg(mt1-hgfsf)19lmb</i>  72. <i>tg(mt2a-tgfb2)4rser</i>  73. <i>tg(myh7-pln)2egk</i>  74. <i>tg(neft*e397k)#milg</i>  75. <i>tg(pgk1-fgf2)15cofn</i>  76. <i>tg(plp1*)4rsj</i>  77. <i>tg(prnp-ar*112q)#deme</i>  78. <i>tg(prnp-fus)wt3cshw</i>  79. <i>tg(prnp-mapt*p301l)jnp13hl mc</i>  80. <i>tg(prnp-snca*a53t)83vle</i>  81. <i>tg(prnp-tardbp*a315t)23jlel</i></p>
--	--	---	---

			<p>82. <i>tg(prnp)c35cwe</i>  83. <i>tg(prnp*)#rgab</i>  84. <i>tg(s100b-v-erbb)4496waw</i>  85. <i>tg(sod1*)df7yaw</i>  86. <i>tg(sod1*g37r)106dpr</i>  87. <i>tg(sod1*g37r)29dpr</i>  88. <i>tg(sod1*g85r)148dwc</i>  89. <i>tg(sod1*g85r)74dwc</i>  90. <i>tg(sod1*g93a)&lt;dl&gt;lgur</i>  91. <i>tg(sod1*g93a)lgur</i>  92. <i>tg(sod1*h46r)lara</i>  93. <i>tg(tcrar28,tcrbr28)krndim</i>  94. <i>tg(teto-htr4*d100a)7niss</i>  95. <i>tg(thyl-mapt*p301l)2vln</i>  96. <i>tg(thyl-mapt*p301s)2541godt</i>  97. <i>tg(thyl-snca*a30p)18pjk</i>  98. <i>tg(thyl-ubqln2*p497s)3mont</i>  99. <i>tg(thyl-ubqln2*p506t)6mont</i>  100. <i>ts(17&lt;16&gt;)65dn</i>  101. <i>vsd</i></p> <p><b>Lost due to isolation: 37</b></p> <ol style="list-style-type: none"> <li>1. <i>adgrf5</i></li> <li>2. <i>akap11</i></li> <li>3. <i>arid5b</i></li> <li>4. <i>col4a3bp</i></li> <li>5. <i>coq9</i></li> <li>6. <i>ctsf</i></li> <li>7. <i>dnase1l2</i></li> <li>8. <i>elmod1</i></li> <li>9. <i>epg5</i></li> <li>10. <i>hhip1l</i></li> <li>11. <i>hr</i></li> <li>12. <i>lgi4</i></li> <li>13. <i>lncpint</i></li> <li>14. <i>ltn1</i></li> <li>15. <i>lyst</i></li> <li>16. <i>mbd5</i></li> <li>17. <i>noto</i></li> <li>18. <i>nxn</i></li> <li>19. <i>pex10</i></li> </ol>
--	--	--	---

			<p> 20. <i>pgam5</i>  21. <i>pole4</i>  22. <i>psph</i>  23. <i>qk</i>  24. <i>rnf13</i>  25. <i>sacs</i>  26. <i>skida1</i>  27. <i>slc38a2</i>  28. <i>sno</i>  29. <i>spata6</i>  30. <i>spg20</i>  31. <i>stard10</i>  32. <i>tg(pmp22)c22clh</i>  33. <i>tg(prnp-fus*r521c)3313ejh</i>  34. <i>tg(sod1*g85r)#roos</i>  35. <i>tg(sod1*g86r)m3jw</i>  <i>g</i>  36. <i>tg(thyl-mapt*)30sched</i>  37. <i>zfp106</i> </p>
--	--	--	--

Supplementary Table S2.2. The top 50 genes of the pectoral fin module ranked and ordered based on the weighted degree. Full gene list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Zebrafish gene name	ZFIN identifier	Annotation type	Module-specific or conserved status	Weighted degree	Rank
<i>shha</i>	zdb-gene-980526-166	direct annotation to the pectoral fin	conserved	36.6983	1
<i>bmp4</i>	zdb-gene-980528-2059	predicted	conserved	35.5652	2
<i>bmp2b</i>	zdb-gene-980526-474	predicted	conserved	34.3109	3
<i>wnt3a</i>	zdb-gene-001106-1	predicted	conserved	31.4927	4
<i>fgf8a</i>	zdb-gene-990415-72	predicted	conserved	31.0094	5
<i>gsk3b</i>	zdb-gene-990714-4	direct annotation to the pectoral fin	module-specific	29.1916	6
<i>lef1</i>	zdb-gene-990714-26	direct annotation to the pectoral fin	module-specific	28.3310	7
<i>gli2a</i>	zdb-gene-990706-8	predicted	conserved	27.2862	8
<i>hdac1</i>	zdb-gene-020419-32	direct annotation to the pectoral fin	module-specific	27.0476	9
<i>wnt5b</i>	zdb-gene-980526-87	predicted	module-specific	25.9169	10
<i>bmp7a</i>	zdb-gene-000208-25	direct annotation to the pectoral fin	conserved	25.8141	11
<i>fgfr1a</i>	zdb-gene-980526-255	predicted	conserved	25.6442	12
<i>smad5</i>	zdb-gene-990603-9	predicted	module-specific	25.5690	13
<i>gli1</i>	zdb-gene-030321-1	predicted	module-specific	25.2858	14
<i>tcf7l1a</i>	zdb-gene-980605-30	predicted	module-specific	24.8101	15
<i>foxd3</i>	zdb-gene-980526-143	predicted	module-specific	24.7462	16
<i>fgf10a</i>	zdb-gene-030715-1	direct annotation to the pectoral fin	conserved	24.2622	17
<i>ta</i>	zdb-gene-980526-437	predicted	module-specific	23.7992	18
<i>smo</i>	zdb-gene-980526-89	direct annotation to the pectoral fin	conserved	23.6222	19
<i>cdx4</i>	zdb-gene-980526-330	predicted	module-specific	23.1331	20
<i>chd</i>	zdb-gene-990415-33	direct annotation to the pectoral fin	module-specific	22.2847	21
<i>pax2a</i>	zdb-gene-990415-8	predicted	module-specific	22.2653	22
<i>smad1</i>	zdb-gene-991119-8	direct annotation to the pectoral fin	module-specific	22.2133	23
<i>ctnnb2</i>	zdb-gene-040426-2575	predicted	module-specific	22.1606	24
<i>ptch2</i>	zdb-gene-980526-44	predicted	module-specific	21.7001	25
<i>isl1</i>	zdb-gene-980526-112	predicted	module-specific	21.5418	26
<i>fgf20a</i>	zdb-gene-060110-1	direct annotation to the pectoral fin	module-specific	21.0334	27



<i>fgf3</i>	zdb-gene-980526-178	predicted	module-specific	20.9767	28
<i>wnt4a</i>	zdb-gene-980526-352	predicted	conserved	20.4183	29
<i>ptch1</i>	zdb-gene-980526-196	direct annotation to the pectoral fin	conserved	18.8304	30
<i>gpc4</i>	zdb-gene-011119-1	predicted	module-specific	18.5224	31
<i>ihha</i>	zdb-gene-051010-1	predicted	conserved	18.1064	32
<i>tbx16</i>	zdb-gene-990615-5	direct annotation to the pectoral fin	module-specific	17.1606	33
<i>wnt11</i>	zdb-gene-990603-12	predicted	module-specific	17.1389	34
<i>tbx5a</i>	zdb-gene-991124-7	direct annotation to the pectoral fin	conserved	16.7968	35
<i>aldh1a2</i>	zdb-gene-011010-3	direct annotation to the pectoral fin	conserved	16.6865	36
<i>fgf24</i>	zdb-gene-030708-1	direct annotation to the pectoral fin	module-specific	16.6315	37
<i>hand2</i>	zdb-gene-000511-1	direct annotation to the pectoral fin	conserved	16.4984	38
<i>szl</i>	zdb-gene-030530-1	direct annotation to the pectoral fin	module-specific	15.9244	39
<i>zic2a</i>	zdb-gene-000710-4	predicted	conserved	15.2976	40
<i>nkx2.2a</i>	zdb-gene-980526-403	predicted	module-specific	15.0866	41
<i>tcf7</i>	zdb-gene-050222-4	direct annotation to the pectoral fin	conserved	15.0392	42
<i>dlx2a</i>	zdb-gene-980526-212	predicted	module-specific	14.8123	43
<i>pitx2</i>	zdb-gene-990714-27	predicted	module-specific	14.7898	44
<i>msx1a</i>	zdb-gene-980526-312	predicted	conserved	14.6308	45
<i>met</i>	zdb-gene-041014-1	direct annotation to the pectoral fin	conserved	14.5234	46
<i>foxl1</i>	zdb-gene-040426-1181	direct annotation to the pectoral fin	module-specific	14.3996	47
<i>myf5</i>	zdb-gene-000616-6	predicted	module-specific	13.7077	48
<i>acvr1l</i>	zdb-gene-990415-9	direct annotation to the pectoral fin	module-specific	13.7041	49
<i>wnt2ba</i>	zdb-gene-030717-2	direct annotation to the pectoral fin	module-specific	13.2992	50

Supplementary Table S2.3. The top 50 genes of the pelvic fin module ranked and ordered based on the weighted degree. The full gene list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Zebrafish gene name	ZFIN identifier	Annotation type	Module-specific or conserved status	Weighted degree	Weighted-rank
<i>hsp90ab1</i>	zdb-gene-990415-95	predicted	module-specific	129.9415	1
<i>mapk3</i>	zdb-gene-040121-1	predicted	module-specific	121.5328	2
<i>rhoab</i>	zdb-gene-040322-2	predicted	module-specific	108.0837	3
<i>ctnbl1</i>	zdb-gene-980526-362	predicted	conserved	106.2749	4
<i>hsp90aa1.2</i>	zdb-gene-031001-3	predicted	module-specific	105.1273	5
<i>paics</i>	zdb-gene-030131-9762	predicted	module-specific	104.4687	6
<i>gsk3b</i>	zdb-gene-990714-4	predicted	module-specific	101.2293	7
<i>cad</i>	zdb-gene-021030-4	predicted	module-specific	97.8373	8
<i>cdc42</i>	zdb-gene-030131-8783	predicted	module-specific	97.7752	9
<i>acta1b</i>	zdb-gene-030131-55	predicted	module-specific	97.6682	10
<i>smarca4a</i>	zdb-gene-030605-1	predicted	conserved	97.4336	11
<i>mapk1</i>	zdb-gene-030722-2	predicted	module-specific	96.3956	12
<i>jupa</i>	zdb-gene-991207-22	predicted	module-specific	95.0796	13
<i>cdk1</i>	zdb-gene-010320-1	predicted	module-specific	93.9169	14
<i>kras</i>	NA	predicted	conserved	93.403	15
<i>rac1a</i>	zdb-gene-030131-5415	predicted	conserved	92.9107	16
<i>mapk14a</i>	zdb-gene-010202-2	predicted	conserved	92.1789	17
<i>src</i>	zdb-gene-030131-3809	predicted	conserved	91.7256	18
<i>fgfr1a</i>	zdb-gene-980526-255	predicted	conserved	91.2553	19
<i>met</i>	zdb-gene-041014-1	predicted	conserved	90.8735	20
<i>insrb</i>	zdb-gene-020503-4	predicted	module-specific	90.6827	21
<i>si:ch211-163m16.1</i>	NA	predicted	module-specific	84.6841	22
<i>actl6a</i>	zdb-gene-020419-36	predicted	module-specific	83.9023	23
<i>tp53</i>	zdb-gene-990415-270	predicted	conserved	83.8035	24
<i>pak2a</i>	zdb-gene-021011-2	predicted	module-specific	83.2637	25
<i>ehmt2</i>	zdb-gene-010501-6	predicted	module-specific	82.9897	26
<i>kdrl</i>	zdb-gene-000705-1	predicted	module-specific	81.0991	27
<i>mtor</i>	zdb-gene-030131-2974	predicted	conserved	80.8121	28
<i>pkn2</i>	zdb-gene-061207-42	predicted	module-specific	79.5943	29
<i>prkcbb</i>	zdb-gene-040426-1178	predicted	module-specific	79.5529	30
<i>hsp90aa1.1</i>	zdb-gene-990415-94	predicted	module-specific	78.6325	31

<i>ptenb</i>	zdb-gene-030616-47	predicted	conserved	78.531	32
<i>kita</i>	zdb-gene-980526-464	predicted	module-specific	77.2229	33
<i>akt2</i>	zdb-gene-031007-5	predicted	module-specific	77.0797	34
<i>igflra</i>	zdb-gene-020503-1	predicted	module-specific	76.527	35
<i>bmp4</i>	zdb-gene-980528-2059	predicted	conserved	76.215	36
<i>igflrb</i>	zdb-gene-020503-2	predicted	module-specific	75.0886	37
<i>rap1b</i>	zdb-gene-030131-9662	predicted	module-specific	73.3502	38
<i>rac2</i>	zdb-gene-040625-27	predicted	module-specific	73.3136	39
<i>hspa9</i>	zdb-gene-030828-12	predicted	module-specific	72.3324	40
<i>hdac1</i>	zdb-gene-020419-32	predicted	module-specific	69.4428	41
<i>pola1</i>	zdb-gene-030114-9	predicted	module-specific	69.0229	42
<i>actc1a</i>	zdb-gene-040520-4	predicted	module-specific	68.6027	43
<i>top2b</i>	zdb-gene-041008-136	predicted	module-specific	68.4413	44
<i>bmp2b</i>	zdb-gene-980526-474	predicted	conserved	68.2429	45
<i>insra</i>	zdb-gene-020503-3	predicted	module-specific	68.1565	46
<i>flt1</i>	zdb-gene-050407-1	predicted	module-specific	68.0761	47
<i>ralbb</i>	zdb-gene-040625-121	predicted	module-specific	67.8323	48
<i>btk</i>	zdb-gene-070531-1	predicted	module-specific	67.4762	49
<i>rac1b</i>	zdb-gene-060312-45	predicted	conserved	67.0359	50

Supplementary Table S2.4. The top 50 genes of the forelimb module ranked and ordered based on the weighted degree. The full gene list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Mouse gene name	MGI identifier	Annotation type	Module-specific or conserved status	Weighted degree	Rank
<i>bmp4</i>	mgi:88180	direct annotation to the forelimb	conserved	47.2641	1
<i>ctnbl1</i>	mgi:88276	direct annotation to the forelimb	module-specific	46.0747	2
<i>trp53</i>	mgi:98834	direct annotation to the forelimb	module-specific	37.8867	3
<i>shh</i>	mgi:98297	direct annotation to the forelimb	conserved	37.6249	4
<i>wnt5a</i>	mgi:98958	direct annotation to the forelimb	module-specific	35.6754	5
<i>smad4</i>	mgi:894293	predicted	module-specific	34.5294	6
<i>bmp7</i>	mgi:103302	predicted	conserved	34.2132	7
<i>fgf8</i>	mgi:99604	direct annotation to the forelimb	conserved	34.0672	8
<i>runx2</i>	mgi:99829	direct annotation to the forelimb	module-specific	34.0628	9
<i>fgfr1</i>	mgi:95522	direct annotation to the forelimb	conserved	33.436	10
<i>gli2</i>	mgi:95728	direct annotation to the forelimb	conserved	33.0563	11
<i>gli3</i>	mgi:95729	direct annotation to the forelimb	module-specific	32.0078	12
<i>bmp2</i>	mgi:88177	direct annotation to the forelimb	conserved	31.3004	13
<i>wnt3a</i>	mgi:98956	predicted	conserved	30.9124	14
<i>sox9</i>	mgi:98371	direct annotation to the forelimb	conserved	29.737	15
<i>fgfr2</i>	mgi:95523	direct annotation to the forelimb	module-specific	29.6869	16
<i>ihh</i>	mgi:96533	direct annotation to the forelimb	conserved	29.6653	17
<i>fgfr3</i>	mgi:95524	direct annotation to the forelimb	module-specific	29.5449	18
<i>nog</i>	mgi:104327	predicted	module-specific	29.3757	19
<i>smo</i>	mgi:108075	direct annotation to the forelimb	conserved	28.7826	20
<i>wnt7a</i>	mgi:98961	direct annotation to the forelimb	module-specific	28.1128	21
<i>smad3</i>	mgi:1201674	direct annotation to the forelimb	module-specific	27.969	22
<i>wnt4</i>	mgi:98957	predicted	conserved	27.9668	23
<i>ptch1</i>	mgi:105373	predicted	conserved	27.3374	24
<i>wnt1</i>	mgi:98953	predicted	module-specific	26.4361	25
<i>bmpr1a</i>	mgi:1338938	predicted	module-specific	26.0359	26
<i>vegfa</i>	mgi:103178	annotated to a part or bud	module-specific	24.9811	27
<i>chrd</i>	mgi:1313268	predicted	module-specific	24.6081	28
<i>col2a1</i>	mgi:88452	direct annotation to the forelimb	module-specific	24.2746	29
<i>fgf9</i>	mgi:104723	direct annotation to the forelimb	module-specific	24.0714	30
<i>msx1</i>	mgi:97168	predicted	conserved	23.4172	31
<i>msx2</i>	mgi:97169	predicted	module-specific	22.8361	32

<i>fgf10</i>	mgi:1099809	direct annotation to the forelimb	conserved	22.7474	33
<i>tgfb2</i>	mgi:98726	direct annotation to the forelimb	module-specific	22.6531	34
<i>pax3</i>	mgi:97487	direct annotation to the forelimb	module-specific	22.5835	35
<i>pthlh</i>	mgi:97800	direct annotation to the forelimb	module-specific	22.3993	36
<i>esr1</i>	mgi:1352467	direct annotation to the forelimb	module-specific	22.0893	37
<i>bmp5</i>	mgi:88181	direct annotation to the forelimb	module-specific	21.7889	38
<i>itgb1</i>	mgi:96610	direct annotation to the forelimb	module-specific	21.6904	39
<i>acvr1</i>	mgi:87911	direct annotation to the forelimb	module-specific	20.5535	40
<i>ctgf</i>	mgi:95537	direct annotation to the forelimb	conserved	20.3372	41
<i>lrrk1</i>	mgi:2142227	direct annotation to the forelimb	module-specific	20.1363	42
<i>cdc42</i>	mgi:106211	annotated to a part or bud	module-specific	19.6953	43
<i>twist1</i>	mgi:98872	direct annotation to the forelimb	module-specific	19.5487	44
<i>lrp6</i>	mgi:1298218	direct annotation to the forelimb	module-specific	18.1913	45
<i>fst</i>	mgi:95586	predicted	module-specific	18.1062	46
<i>wnt7b</i>	mgi:98962	predicted	module-specific	17.7777	47
<i>gdf5</i>	mgi:95688	direct annotation to the forelimb	module-specific	17.6851	48
<i>dkk1</i>	mgi:1329040	direct annotation to the forelimb	module-specific	17.3777	49
<i>gsc</i>	mgi:95841	direct annotation to the forelimb	module-specific	16.3233	50

Supplementary Table S2.5. The top 50 genes of the hindlimb module ranked and ordered based on the weighted degree. The full gene list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Mouse gene name	MGI identifier	Annotation type	Module-specific or conserved status	Weighted degree	Rank
<i>trp53</i>	mgi:98834	direct annotation to the hindlimb	conserved	82.9324	1
<i>hras</i>	mgi:96224	predicted	conserved	72.8342	2
<i>ctnbl1</i>	mgi:88276	direct annotation to the hindlimb	conserved	72.1808	3
<i>mapk14</i>	mgi:1346865	direct annotation to the hindlimb	conserved	71.7856	4
<i>kras</i>	mgi:96680	predicted	conserved	70.7748	5
<i>src</i>	mgi:98397	direct annotation to the hindlimb	conserved	70.3811	6
<i>myc</i>	mgi:97250	predicted	module-specific	66.8072	7
<i>tnf</i>	mgi:104798	direct annotation to the hindlimb	module-specific	65.6905	8
<i>fos</i>	mgi:95574	predicted	module-specific	64.7904	9
<i>bmp4</i>	mgi:88180	direct annotation to the hindlimb	conserved	63.9989	10
<i>lepr</i>	mgi:104993	direct annotation to the hindlimb	module-specific	63.4939	11
<i>il6</i>	mgi:96559	direct annotation to the hindlimb	module-specific	62.6003	12
<i>runx2</i>	mgi:99829	direct annotation to the hindlimb	module-specific	61.843	13
<i>smad4</i>	mgi:894293	direct annotation to the hindlimb	module-specific	58.2585	14
<i>tgfb1</i>	mgi:98725	predicted	module-specific	54.2763	15
<i>fgfr1</i>	mgi:95522	direct annotation to the hindlimb	conserved	53.223	16
<i>igf1</i>	mgi:96432	predicted	module-specific	52.9341	17
<i>esr1</i>	mgi:1352467	direct annotation to the hindlimb	module-specific	52.2754	18
<i>fgf2</i>	mgi:95516	predicted	module-specific	51.2651	19
<i>vegfa</i>	mgi:103178	annotated to a part or bud	conserved	50.7476	20
<i>mmp9</i>	mgi:97011	direct annotation to the hindlimb	module-specific	49.7613	21
<i>smad3</i>	mgi:1201674	annotated to a part or bud	conserved	49.4434	22
<i>casp3</i>	mgi:107739	predicted	conserved	49.0774	23
<i>fgfr2</i>	mgi:95523	direct annotation to the hindlimb	conserved	48.7644	24
<i>itgb1</i>	mgi:96610	direct annotation to the hindlimb	module-specific	48.3952	25
<i>bmp7</i>	mgi:103302	predicted	conserved	48.2316	26
<i>wnt5a</i>	mgi:98958	predicted	module-specific	48.1588	27
<i>fgfr3</i>	mgi:95524	direct annotation to the hindlimb	module-specific	48.0184	28
<i>fgf8</i>	mgi:99604	direct annotation to the hindlimb	conserved	46.6654	29
<i>smad2</i>	mgi:108051	predicted	conserved	45.6514	30
<i>pdgfra</i>	mgi:97530	predicted	module-specific	45.2074	31
<i>bmp2</i>	mgi:88177	predicted	conserved	45.0361	32

<i>shh</i>	mgi:98297	direct annotation to the hindlimb	conserved	44.4006	33
<i>sox9</i>	mgi:98371	direct annotation to the hindlimb	module-specific	43.7136	34
<i>pten</i>	mgi:109583	annotated to a part or bud	conserved	43.6208	35
<i>coll1a1</i>	mgi:88467	direct annotation to the hindlimb	conserved	43.3734	36
<i>col2a1</i>	mgi:88452	direct annotation to the hindlimb	conserved	43.3479	37
<i>gli2</i>	mgi:95728	direct annotation to the hindlimb	conserved	42.8396	38
<i>wnt3a</i>	mgi:98956	direct annotation to the hindlimb	conserved	42.6564	39
<i>lrrk1</i>	mgi:2142227	direct annotation to the hindlimb	module-specific	42.2122	40
<i>il10</i>	mgi:96537	direct annotation to the hindlimb	module-specific	41.3969	41
<i>il4</i>	mgi:96556	direct annotation to the hindlimb	module-specific	41.1804	42
<i>lep</i>	mgi:104663	direct annotation to the hindlimb	module-specific	40.6877	43
<i>pthlh</i>	mgi:97800	direct annotation to the hindlimb	module-specific	40.5325	44
<i>tnfrsf11</i>	mgi:1100089	direct annotation to the hindlimb	module-specific	40.3085	45
<i>nog</i>	mgi:104327	predicted	module-specific	39.7063	46
<i>notch2</i>	mgi:97364	direct annotation to the hindlimb	conserved	39.0854	47
<i>ptch1</i>	mgi:105373	predicted	conserved	38.8191	48
<i>mtor</i>	mgi:1928394	annotated to a part or bud	conserved	38.6216	49
<i>egr1</i>	mgi:95295	direct annotation to the hindlimb	module-specific	38.4933	50

Supplementary Table S2.6. The top 100 enriched Biological Process terms from the Gene Ontology for the pectoral fin module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term Identifier	Term name	P-value
GO:0033333	fin development	5.88E-30
GO:0033339	pectoral fin development	5.97E-24
GO:0051216	cartilage development	4.67E-20
GO:0033334	fin morphogenesis	2.69E-17
GO:0035118	embryonic pectoral fin morphogenesis	3.08E-15
GO:0035138	pectoral fin morphogenesis	5.77E-14
GO:0001947	heart looping	1.90E-13
GO:0007275	multicellular organism development	1.36E-12
GO:0048703	embryonic viscerocranium morphogenesis	6.36E-12
GO:0003143	embryonic heart tube morphogenesis	1.68E-11
GO:0006355	regulation of transcription, DNA-templated	5.96E-11
GO:0014032	neural crest cell development	3.04E-10
GO:0007368	determination of left/right symmetry	3.33E-10
GO:0001756	somitogenesis	1.10E-09
GO:0007507	heart development	1.61E-09
GO:0048701	embryonic cranial skeleton morphogenesis	5.80E-09
GO:2000223	regulation of BMP signaling pathway involved in heart jogging	2.16E-08
GO:0042476	odontogenesis	3.49E-08
GO:0009953	dorsal/ventral pattern formation	2.13E-07
GO:0007422	peripheral nervous system development	3.03E-07
GO:0016055	Wnt signaling pathway	7.25E-07
GO:0030166	proteoglycan biosynthetic process	7.54E-07
GO:0030902	hindbrain development	1.19E-06
GO:0009880	embryonic pattern specification	1.33E-06
GO:0043010	camera-type eye development	2.58E-06
GO:0060028	convergent extension involved in axis elongation	4.11E-06
GO:0030182	neuron differentiation	6.91E-06
GO:0060030	dorsal convergence	8.19E-06
GO:0031290	retinal ganglion cell axon guidance	1.01E-05
GO:0030198	extracellular matrix organization	1.32E-05
GO:0001503	ossification	1.64E-05
GO:0060037	pharyngeal system development	2.01E-05
GO:0006351	transcription, DNA-templated	2.26E-05



GO:0036342	post-anal tail morphogenesis	2.44E-05
GO:0048666	neuron development	4.11E-05
GO:0048665	neuron fate specification	5.57E-05
GO:0015012	heparan sulfate proteoglycan biosynthetic process	5.57E-05
GO:0048066	developmental pigmentation	7.59E-05
GO:0050650	chondroitin sulfate proteoglycan biosynthetic process	7.61E-05
GO:0030900	forebrain development	1.28E-04
GO:0030901	midbrain development	1.28E-04
GO:0060042	retina morphogenesis in camera-type eye	1.44E-04
GO:0048793	pronephros development	1.50E-04
GO:0048709	oligodendrocyte differentiation	1.65E-04
GO:0045879	negative regulation of smoothened signaling pathway	1.65E-04
GO:0048384	retinoic acid receptor signaling pathway	1.65E-04
GO:0006024	glycosaminoglycan biosynthetic process	1.65E-04
GO:0001649	osteoblast differentiation	1.65E-04
GO:0031101	fin regeneration	1.73E-04
GO:0031017	exocrine pancreas development	2.47E-04
GO:0007420	brain development	3.89E-04
GO:0060536	cartilage morphogenesis	4.99E-04
GO:0048264	determination of ventral identity	5.79E-04
GO:0021903	rostrocaudal neural tube patterning	6.11E-04
GO:0001755	neural crest cell migration	6.28E-04
GO:0043049	otic placode formation	7.62E-04
GO:0060348	bone development	9.12E-04
GO:0060351	cartilage development involved in endochondral bone morphogenesis	9.12E-04
GO:0060070	canonical Wnt signaling pathway	0.00113706
GO:0007224	smoothened signaling pathway	0.00122997
GO:0060059	embryonic retina morphogenesis in camera-type eye	0.00122997
GO:0035462	determination of left/right asymmetry in diencephalon	0.00127048
GO:0007417	central nervous system development	0.00205425
GO:0071696	ectodermal placode development	0.0021555
GO:0045743	positive regulation of fibroblast growth factor receptor signaling pathway	0.0021555
GO:0003146	heart jogging	0.00241705
GO:0048596	embryonic camera-type eye morphogenesis	0.00268046
GO:0031076	embryonic camera-type eye development	0.00268046
GO:0031016	pancreas development	0.00308488
GO:0001501	skeletal system development	0.00308488
GO:0030509	BMP signaling pathway	0.00385573

GO:0050769	positive regulation of neurogenesis	0.00389087
GO:0035775	pronephric glomerulus morphogenesis	0.00389087
GO:0021986	habenula development	0.00457459
GO:0021984	adenohypophysis development	0.00457459
GO:0030903	notochord development	0.00473414
GO:0060041	retina development in camera-type eye	0.00521292
GO:0021508	floor plate formation	0.00530953
GO:0061371	determination of heart left/right asymmetry	0.00572415
GO:0048699	generation of neurons	0.00609484
GO:0045892	negative regulation of transcription, DNA-templated	0.00682315
GO:0060026	convergent extension	0.00682935
GO:0009952	anterior/posterior pattern specification	0.00722391
GO:0009948	anterior/posterior axis specification	0.00781325
GO:0045893	positive regulation of transcription, DNA-templated	0.00844084
GO:0042472	inner ear morphogenesis	0.00972333
GO:0008104	protein localization	0.00972333
GO:0048702	embryonic neurocranium morphogenesis	0.00972333
GO:0007498	mesoderm development	0.01074824
GO:0070121	Kupffer's vesicle development	0.0119139
GO:0007154	cell communication	0.0124591
GO:0003140	determination of left/right asymmetry in lateral mesoderm	0.01293391
GO:0060027	convergent extension involved in gastrulation	0.01478156
GO:0071599	otic vesicle development	0.01529561
GO:0070278	extracellular matrix constituent secretion	0.01576409
GO:0055014	atrial cardiac muscle cell development	0.01576409
GO:0071711	basement membrane organization	0.01576409
GO:0048840	otolith development	0.01654059
GO:0030917	midbrain-hindbrain boundary development	0.01915513
GO:0030514	negative regulation of BMP signaling pathway	0.01915513

Supplementary Table S2.7. The top 100 enriched Biological Process terms from the Gene Ontology for the pectoral fin conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term identifier	Term name	P-value
GO:0033339	pectoral fin development	1.06E-19
GO:0048703	embryonic viscerocranium morphogenesis	3.70E-11
GO:0009953	dorsal/ventral pattern formation	1.73E-10
GO:0042664	negative regulation of endodermal cell fate specification	1.99E-09
GO:0010002	cardioblast differentiation	3.57E-09
GO:0035118	embryonic pectoral fin morphogenesis	3.89E-09
GO:0048839	inner ear development	1.62E-08
GO:0051216	cartilage development	3.64E-08
GO:0007275	multicellular organism development	4.86E-08
GO:0060041	retina development in camera-type eye	6.00E-08
GO:0030902	hindbrain development	6.23E-08
GO:0009952	anterior/posterior pattern specification	1.20E-07
GO:0003342	proepicardium development	1.29E-07
GO:0001947	heart looping	2.41E-07
GO:0043049	otic placode formation	2.45E-07
GO:0048793	pronephros development	4.23E-07
GO:0007224	smoothened signaling pathway	4.83E-07
GO:0031016	pancreas development	1.78E-06
GO:0001889	liver development	1.97E-06
GO:0030916	otic vesicle formation	2.80E-06
GO:0030903	notochord development	3.30E-06
GO:0021984	adenohypophysis development	3.63E-06
GO:0007368	determination of left/right symmetry	3.73E-06
GO:0060070	canonical Wnt signaling pathway	1.11E-05
GO:0030182	neuron differentiation	1.18E-05
GO:0048795	swim bladder morphogenesis	3.39E-05
GO:0030917	midbrain-hindbrain boundary development	3.63E-05
GO:0001756	somitogenesis	4.81E-05
GO:0031018	endocrine pancreas development	5.02E-05
GO:0048557	embryonic digestive tract morphogenesis	5.64E-05
GO:0042694	muscle cell fate specification	8.45E-05
GO:0021703	locus ceruleus development	8.45E-05
GO:0031290	retinal ganglion cell axon guidance	9.49E-05

GO:0003007	heart morphogenesis	1.49E-04
GO:0006355	regulation of transcription, DNA-templated	1.76E-04
GO:0016055	Wnt signaling pathway	2.39E-04
GO:0048701	embryonic cranial skeleton morphogenesis	2.49E-04
GO:0031076	embryonic camera-type eye development	2.52E-04
GO:0021536	diencephalon development	3.68E-04
GO:0042476	odontogenesis	4.35E-04
GO:0048709	oligodendrocyte differentiation	5.06E-04
GO:0055002	striated muscle cell development	5.06E-04
GO:0048794	swim bladder development	5.06E-04
GO:0021508	floor plate formation	5.06E-04
GO:0001708	cell fate specification	5.83E-04
GO:0048752	semicircular canal morphogenesis	5.83E-04
GO:0008015	blood circulation	7.53E-04
GO:0048702	embryonic neurocranium morphogenesis	9.44E-04
GO:0060536	cartilage morphogenesis	0.00104762
GO:0048264	determination of ventral identity	0.00115614
GO:0036342	post-anal tail morphogenesis	0.00126982
GO:0009887	organ morphogenesis	0.00126982
GO:0071599	otic vesicle development	0.00151259
GO:0009880	embryonic pattern specification	0.00164163
GO:0045165	cell fate commitment	0.00285445
GO:0003146	heart jogging	0.00302826
GO:0060042	retina morphogenesis in camera-type eye	0.00302826
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	0.00313287
GO:0008543	fibroblast growth factor receptor signaling pathway	0.00320693
GO:0007517	muscle organ development	0.00339044
GO:0001501	skeletal system development	0.00357876
GO:0035050	embryonic heart tube development	0.00357876
GO:0040007	growth	0.00377186
GO:0031017	exocrine pancreas development	0.00396973
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	0.00417235
GO:0060395	SMAD protein signal transduction	0.00459171
GO:0043408	regulation of MAPK cascade	0.00480842
GO:0061131	pancreas field specification	0.00482266
GO:0035777	pronephric distal tubule development	0.00482266
GO:0003303	BMP signaling pathway involved in heart jogging	0.00482266
GO:0060876	semicircular canal formation	0.0072255

GO:0048618	post-embryonic foregut morphogenesis	0.0072255
GO:0003170	heart valve development	0.0072255
GO:0042573	retinoic acid metabolic process	0.0096227
GO:0001957	intramembranous ossification	0.0096227
GO:0060956	endocardial cell differentiation	0.0096227
GO:0035776	pronephric proximal tubule development	0.0096227
GO:0007267	cell-cell signaling	0.01039994
GO:0060788	ectodermal placode formation	0.01201428
GO:0035143	caudal fin morphogenesis	0.01201428
GO:0016539	intein-mediated protein splicing	0.01201428
GO:0021523	somatic motor neuron differentiation	0.01201428
GO:0021587	cerebellum morphogenesis	0.01440024
GO:0071699	olfactory placode morphogenesis	0.01440024
GO:0001839	neural plate morphogenesis	0.01440024
GO:0021628	olfactory nerve formation	0.01440024
GO:0007411	axon guidance	0.01591333
GO:0048339	paraxial mesoderm development	0.01678059
GO:0048385	regulation of retinoic acid receptor signaling pathway	0.01678059
GO:0021953	central nervous system neuron differentiation	0.01915536
GO:0006461	protein complex assembly	0.01915536
GO:0048663	neuron fate commitment	0.01915536
GO:0050935	iridophore differentiation	0.02152455
GO:0045892	negative regulation of transcription, DNA-templated	0.02198396
GO:0021520	spinal cord motor neuron cell fate specification	0.02388818
GO:0045893	positive regulation of transcription, DNA-templated	0.02464145
GO:0021986	habenula development	0.0309458
GO:0032474	otolith morphogenesis	0.0309458
GO:0006351	transcription, DNA-templated	0.03240686
GO:0048384	retinoic acid receptor signaling pathway	0.03328729

Supplementary Table S2.8. The top 100 enriched Biological Process terms from the Gene Ontology for the forelimb conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term identifier	Term name	P-value
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	1.13E-19
GO:0045893	positive regulation of transcription, DNA-templated	7.32E-17
GO:0007389	pattern specification process	1.23E-15
GO:0030326	embryonic limb morphogenesis	2.72E-15
GO:0001503	ossification	1.44E-13
GO:0010628	positive regulation of gene expression	2.30E-13
GO:0042475	odontogenesis of dentin-containing tooth	2.99E-13
GO:0035115	embryonic forelimb morphogenesis	4.31E-13
GO:0007275	multicellular organism development	1.04E-12
GO:0007507	heart development	1.41E-12
GO:0030324	lung development	1.52E-12
GO:0035116	embryonic hindlimb morphogenesis	3.57E-11
GO:0048754	branching morphogenesis of an epithelial tube	6.25E-11
GO:0008285	negative regulation of cell proliferation	9.08E-11
GO:0001822	kidney development	1.15E-10
GO:0021904	dorsal/ventral neural tube patterning	2.30E-10
GO:0001658	branching involved in ureteric bud morphogenesis	3.88E-10
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	4.34E-10
GO:0010629	negative regulation of gene expression	1.38E-09
GO:0001649	osteoblast differentiation	1.80E-09
GO:0001947	heart looping	1.90E-09
GO:0042733	embryonic digit morphogenesis	2.80E-09
GO:0090263	positive regulation of canonical Wnt signaling pathway	3.07E-09
GO:0008284	positive regulation of cell proliferation	3.45E-09
GO:0060349	bone morphogenesis	3.91E-09
GO:0001701	in utero embryonic development	3.97E-09
GO:0007224	smoothened signaling pathway	4.38E-09
GO:0043066	negative regulation of apoptotic process	5.41E-09
GO:0034504	protein localization to nucleus	6.28E-09
GO:0050679	positive regulation of epithelial cell proliferation	6.64E-09
GO:0045892	negative regulation of transcription, DNA-templated	6.85E-09
GO:0030501	positive regulation of bone mineralization	8.41E-09

GO:0051216	cartilage development	1.06E-08
GO:0045595	regulation of cell differentiation	1.11E-08
GO:0030154	cell differentiation	1.27E-08
GO:0060021	palate development	1.31E-08
GO:0030509	BMP signaling pathway	1.51E-08
GO:0060445	branching involved in salivary gland morphogenesis	1.58E-08
GO:0002062	chondrocyte differentiation	2.61E-08
GO:0045597	positive regulation of cell differentiation	2.61E-08
GO:0009953	dorsal/ventral pattern formation	3.62E-08
GO:0007267	cell-cell signaling	4.19E-08
GO:0009887	organ morphogenesis	5.88E-08
GO:0009952	anterior/posterior pattern specification	6.92E-08
GO:0001708	cell fate specification	1.01E-07
GO:0030902	hindbrain development	1.22E-07
GO:0048646	anatomical structure formation involved in morphogenesis	1.22E-07
GO:0060441	epithelial tube branching involved in lung morphogenesis	1.22E-07
GO:0009880	embryonic pattern specification	1.22E-07
GO:0003007	heart morphogenesis	1.42E-07
GO:0009954	proximal/distal pattern formation	1.71E-07
GO:0042476	odontogenesis	2.00E-07
GO:0042127	regulation of cell proliferation	2.23E-07
GO:0042472	inner ear morphogenesis	2.57E-07
GO:0045165	cell fate commitment	2.57E-07
GO:0021983	pituitary gland development	3.11E-07
GO:0007368	determination of left/right symmetry	4.65E-07
GO:0002053	positive regulation of mesenchymal cell proliferation	7.42E-07
GO:0060070	canonical Wnt signaling pathway	7.45E-07
GO:0001656	metanephros development	8.28E-07
GO:0030901	midbrain development	1.02E-06
GO:0032355	response to estradiol	1.40E-06
GO:0001823	mesonephros development	1.44E-06
GO:0090090	negative regulation of canonical Wnt signaling pathway	1.47E-06
GO:0070374	positive regulation of ERK1 and ERK2 cascade	1.47E-06
GO:0045666	positive regulation of neuron differentiation	1.54E-06
GO:0061053	somite development	2.38E-06
GO:0001759	organ induction	2.97E-06
GO:0048538	thymus development	3.74E-06
GO:0014032	neural crest cell development	4.43E-06
GO:0006355	regulation of transcription, DNA-templated	8.35E-06

GO:0048557	embryonic digestive tract morphogenesis	8.63E-06
GO:0045669	positive regulation of osteoblast differentiation	8.77E-06
GO:0050680	negative regulation of epithelial cell proliferation	1.10E-05
GO:0060425	lung morphogenesis	1.31E-05
GO:0032967	positive regulation of collagen biosynthetic process	1.68E-05
GO:0030878	thyroid gland development	1.68E-05
GO:0060129	thyroid-stimulating hormone-secreting cell differentiation	2.18E-05
GO:0021938	smoothened signaling pathway involved in regulation of cerebellar granule cell precursor cell proliferation	2.18E-05
GO:0030900	forebrain development	2.23E-05
GO:0071542	dopaminergic neuron differentiation	2.35E-05
GO:0045880	positive regulation of smoothened signaling pathway	2.60E-05
GO:0006351	transcription, DNA-templated	2.67E-05
GO:0001837	epithelial to mesenchymal transition	2.88E-05
GO:0031016	pancreas development	3.48E-05
GO:0060684	epithelial-mesenchymal cell signaling	3.63E-05
GO:0010463	mesenchymal cell proliferation	3.63E-05
GO:0060513	prostatic bud formation	3.63E-05
GO:0060665	regulation of branching involved in salivary gland morphogenesis by mesenchymal-epithelial signaling	3.63E-05
GO:0071773	cellular response to BMP stimulus	4.17E-05
GO:0030177	positive regulation of Wnt signaling pathway	4.93E-05
GO:0016055	Wnt signaling pathway	5.27E-05
GO:0001501	skeletal system development	5.27E-05
GO:0030879	mammary gland development	5.35E-05
GO:0008543	fibroblast growth factor receptor signaling pathway	7.77E-05
GO:0001525	angiogenesis	9.09E-05
GO:0060485	mesenchyme development	1.01E-04
GO:0021978	telencephalon regionalization	1.01E-04
GO:0042487	regulation of odontogenesis of dentin-containing tooth	1.01E-04
GO:0001657	ureteric bud development	1.08E-04



Supplementary Table S2.9. The top 100 enriched Biological Process terms from the Gene Ontology for the forelimb module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term identifier	Term name	P-value
GO:0001501	skeletal system development	4.17E-37
GO:0042733	embryonic digit morphogenesis	2.51E-35
GO:0030326	embryonic limb morphogenesis	4.31E-34
GO:0035115	embryonic forelimb morphogenesis	4.06E-27
GO:0007275	multicellular organism development	6.29E-23
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	6.62E-22
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	1.17E-19
GO:0051216	cartilage development	1.71E-19
GO:0009953	dorsal/ventral pattern formation	7.65E-19
GO:0060021	palate development	2.57E-16
GO:0060173	limb development	1.17E-15
GO:0009952	anterior/posterior pattern specification	1.39E-15
GO:0035108	limb morphogenesis	1.54E-15
GO:0045893	positive regulation of transcription, DNA-templated	3.23E-15
GO:0001958	endochondral ossification	3.95E-15
GO:0032332	positive regulation of chondrocyte differentiation	5.07E-15
GO:0030509	BMP signaling pathway	9.40E-15
GO:0002062	chondrocyte differentiation	2.31E-14
GO:0009954	proximal/distal pattern formation	4.06E-14
GO:0001503	ossification	4.29E-14
GO:0001843	neural tube closure	5.04E-14
GO:0002053	positive regulation of mesenchymal cell proliferation	5.74E-14
GO:0050680	negative regulation of epithelial cell proliferation	3.21E-13
GO:0008284	positive regulation of cell proliferation	1.23E-12
GO:0001701	in utero embryonic development	3.32E-12
GO:0045669	positive regulation of osteoblast differentiation	3.57E-12
GO:0001822	kidney development	4.12E-12
GO:0048701	embryonic cranial skeleton morphogenesis	4.34E-12
GO:0060348	bone development	8.46E-12
GO:0048706	embryonic skeletal system development	1.62E-11
GO:0035116	embryonic hindlimb morphogenesis	2.86E-11
GO:0042475	odontogenesis of dentin-containing tooth	3.29E-11
GO:0016055	Wnt signaling pathway	5.78E-11

GO:0035136	forelimb morphogenesis	5.83E-11
GO:0021915	neural tube development	7.86E-11
GO:0060070	canonical Wnt signaling pathway	9.68E-11
GO:0030199	collagen fibril organization	1.49E-10
GO:0007507	heart development	1.71E-10
GO:0035137	hindlimb morphogenesis	1.73E-10
GO:0048704	embryonic skeletal system morphogenesis	1.74E-10
GO:0002063	chondrocyte development	2.81E-10
GO:0007389	pattern specification process	3.02E-10
GO:0030324	lung development	3.93E-10
GO:0048589	developmental growth	1.76E-09
GO:0008285	negative regulation of cell proliferation	2.24E-09
GO:0030501	positive regulation of bone mineralization	2.84E-09
GO:0048705	skeletal system morphogenesis	3.29E-09
GO:0003151	outflow tract morphogenesis	3.29E-09
GO:0007179	transforming growth factor beta receptor signaling pathway	4.13E-09
GO:0001756	somitogenesis	4.60E-09
GO:0050679	positive regulation of epithelial cell proliferation	4.72E-09
GO:0090090	negative regulation of canonical Wnt signaling pathway	6.76E-09
GO:0009887	organ morphogenesis	1.37E-08
GO:0045880	positive regulation of smoothed signaling pathway	1.62E-08
GO:0030514	negative regulation of BMP signaling pathway	1.77E-08
GO:0090263	positive regulation of canonical Wnt signaling pathway	2.36E-08
GO:0060065	uterus development	2.41E-08
GO:0001658	branching involved in ureteric bud morphogenesis	2.52E-08
GO:0001502	cartilage condensation	3.41E-08
GO:0007411	axon guidance	3.82E-08
GO:0001568	blood vessel development	3.97E-08
GO:0045165	cell fate commitment	4.50E-08
GO:0031069	hair follicle morphogenesis	5.13E-08
GO:0030282	bone mineralization	5.13E-08
GO:0045778	positive regulation of ossification	8.66E-08
GO:0030154	cell differentiation	8.79E-08
GO:0006355	regulation of transcription, DNA-templated	1.65E-07
GO:0001649	osteoblast differentiation	2.22E-07
GO:0003007	heart morphogenesis	2.58E-07
GO:0071542	dopaminergic neuron differentiation	3.87E-07
GO:1904948	midbrain dopaminergic neuron differentiation	6.89E-07
GO:0060349	bone morphogenesis	7.21E-07

GO:0048566	embryonic digestive tract development	7.74E-07
GO:0032331	negative regulation of chondrocyte differentiation	7.74E-07
GO:0060351	cartilage development involved in endochondral bone morphogenesis	1.23E-06
GO:0001707	mesoderm formation	1.26E-06
GO:0007050	cell cycle arrest	1.62E-06
GO:0006351	transcription, DNA-templated	1.70E-06
GO:0055007	cardiac muscle cell differentiation	1.77E-06
GO:0045879	negative regulation of smoothened signaling pathway	1.79E-06
GO:0008589	regulation of smoothened signaling pathway	2.30E-06
GO:0048568	embryonic organ development	2.85E-06
GO:0030901	midbrain development	2.85E-06
GO:0007492	endoderm development	2.85E-06
GO:0010628	positive regulation of gene expression	3.14E-06
GO:0032330	regulation of chondrocyte differentiation	3.17E-06
GO:0010468	regulation of gene expression	3.18E-06
GO:0042127	regulation of cell proliferation	3.53E-06
GO:0042474	middle ear morphogenesis	3.66E-06
GO:0003148	outflow tract septum morphogenesis	3.66E-06
GO:0036342	post-anal tail morphogenesis	4.53E-06
GO:0060272	embryonic skeletal joint morphogenesis	4.72E-06
GO:0001525	angiogenesis	6.05E-06
GO:0007224	smoothened signaling pathway	7.47E-06
GO:0001657	ureteric bud development	8.54E-06
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	8.54E-06
GO:0045599	negative regulation of fat cell differentiation	8.54E-06
GO:0007267	cell-cell signaling	9.90E-06
GO:0060395	SMAD protein signal transduction	1.28E-05

Supplementary Table S2.10. The top 100 enriched Uberon terms for the pectoral fin module-specific genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0000151	pectoral fin	1.45E-158
uberon 0005886	post-hyoid pharyngeal arch skeleton	5.01E-26
uberon 0003102	surface structure	1.65E-20
uberon 0011610	ceratohyal cartilage	3.13E-19
uberon 2000040	median fin fold	4.47E-17
uberon 0001708	jaw skeleton	4.83E-16
uberon 4000163	anal fin	2.44E-14
uberon 0003107	meckel's cartilage	9.41E-14
uberon 0003097	dorsal fin	2.89E-11
uberon 0008896	post-hyoid pharyngeal arch	4.39E-11
uberon 0011004	pharyngeal arch cartilage	5.36E-11
uberon 4000164	caudal fin	6.11E-11
uberon 4000172	lepidotrichium	1.70E-10
uberon 0000468	multicellular organism	3.18E-10
uberon 0005419	pectoral appendage bud	5.05E-10
uberon 0001840	semicircular canal	4.11E-09
uberon 0000033	head	4.18E-08
uberon 0004752	palatoquadrate cartilage	1.65E-07
uberon 0000152	pelvic fin	1.73E-07
uberon 2001069	ventral fin fold	4.17E-07
uberon 0001846	internal ear	5.16E-07
uberon 0000165	mouth	8.69E-07
uberon 0003278	skeleton of lower jaw	7.86E-06
uberon 0001703	neurocranium	1.01E-05
uberon 2000694	ceratobranchial 5 tooth	1.07E-05
uberon 0001032	sensory system	2.63E-05
uberon 0002533	post-anal tail bud	3.01E-05
uberon 0011242	ethmoid cartilage	3.21E-05
uberon 0009635	parachordal cartilage	4.32E-05
uberon 0011607	hyomandibular cartilage	5.16E-05
uberon 4000174	caudal fin lepidotrichium	5.88E-05
uberon 0002240	spinal cord	8.58E-05
uberon 0004741	cleithrum	9.86E-05

uberont 2001516	ceratobranchial cartilage	0.00012274
uberont 0001016	nervous system	0.00015561
uberont 0000966	retina	0.0001935
uberont 2005317	pectoral fin fold	0.00020788
uberont 0003051	ear vesicle	0.00024665
uberont 0005884	hyoid arch skeleton	0.00026962
uberont 2001239	ceratobranchial 5 bone	0.00029839
uberont 0001894	diencephalon	0.00030376
uberont 0004753	scapulocoracoid	0.00031826
uberont 0006860	swim bladder	0.00034864
uberont 0002028	hindbrain	0.00035814
uberont 0001890	forebrain	0.00036888
uberont 0003099	cranial neural crest	0.00038728
uberont 0007215	trabecula cranii	0.00044491
uberont 0000941	cranial nerve ii	0.00049253
uberont 2001821	notochord posterior region	0.00065456
uberont 0005421	pectoral appendage apical ectodermal ridge	0.000661
uberont 2000250	opercle	0.0006919
uberont 0003931	diencephalic white matter	0.00089287
uberont 0001898	hypothalamus	0.00095346
uberont 0003011	facial motor nucleus	0.00095346
uberont 0003936	postoptic commissure	0.00102597
uberont 2002193	dorsolateral septum	0.00112865
uberont 0005729	pectoral appendage field	0.00112865
uberont 0000044	dorsal root ganglion	0.00116026
uberont 0000019	camera-type eye	0.00154635
uberont 2005222	ventral larval melanophore stripe	0.00176204
uberont 0005598	trunk somite	0.00176204
uberont 0003114	pharyngeal arch 3	0.00176204
uberont 0003068	axial mesoderm	0.00176204
uberont 4000175	pectoral fin lepidotrichium	0.00176204
uberont 0001905	pineal body	0.00196393
uberont 0003079	floor plate	0.00206531
uberont 0003901	horizontal septum	0.00232987
uberont 0006334	posterior lateral line	0.00246051
uberont 0003056	pre-chordal neural plate	0.00257904
uberont 0002329	somite	0.00284081
uberont 0007812	post-anal tail	0.00318555
uberont 0002328	notochord	0.00334979

uberont 0009621	tail somite	0.00359526
uberont 0001908	optic tract	0.00359526
uberont 4000176	anal fin lepidotrichium	0.00359526
uberont 0000935	anterior commissure	0.00364652
uberont 0001264	pancreas	0.00371816
uberont 2001256	lateral floor plate	0.00482413
uberont 0003075	neural plate	0.00542672
uberont 0011778	motor nucleus of vagal nerve	0.00624516
uberont 2000356	gill raker	0.00624516
uberont 0010741	bone of pectoral complex	0.00624516
uberont 0006597	quadrate bone	0.00627714
uberont 0000926	mesoderm	0.00731927
uberont 0000925	endoderm	0.00731927
uberont 0003077	paraxial mesoderm	0.00796392
uberont 2001456	pectoral fin endoskeletal disc	0.00796392
uberont 0004880	chordamesoderm	0.00796392
uberont 0005362	vagus x ganglion	0.00796392
uberont 0003061	blood island	0.00963146
uberont 0011615	basihyal cartilage	0.00963146
uberont 2000073	somite 5	0.01018278
uberont 0010710	pectoral fin skeleton	0.01018278
uberont 0004375	bone of free limb or fin	0.01018278
uberont 2007008	ventral intermandibularis anterior	0.01018278
uberont 2007048	ventral intermandibularis posterior	0.01018278
uberont 0003117	pharyngeal arch 6	0.01018278
uberont 0002222	perichondrium	0.01018278
uberont 0000948	heart	0.01130549
uberont 0002196	adenohypophysis	0.0120689

Supplementary Table S2.11. The top 100 enriched Uberon terms for the pectoral fin conserved genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0000151	pectoral fin	2.85E-22
uberon 2001456	pectoral fin endoskeletal disc	5.23E-10
uberon 2001516	ceratobranchial cartilage	2.84E-09
uberon 0011242	ethmoid cartilage	2.84E-09
uberon 0005419	pectoral appendage bud	6.37E-09
uberon 0003107	meckel's cartilage	1.62E-08
uberon 2000250	opercle	1.75E-08
uberon 0003051	ear vesicle	2.68E-08
uberon 0004752	palatoquadrate cartilage	2.68E-08
uberon 0008896	post-hyoid pharyngeal arch	3.69E-08
uberon 0007215	trabecula cranii	4.46E-08
uberon 0001708	jaw skeleton	8.85E-08
uberon 0011607	hyomandibular cartilage	1.66E-07
uberon 0003079	floor plate	2.59E-07
uberon 0007812	post-anal tail	4.23E-07
uberon 0011610	ceratohyal cartilage	8.66E-07
uberon 0002329	somite	2.05E-06
uberon 2000558	posterior macula	2.43E-06
uberon 0005945	neurocranial trabecula	4.92E-06
uberon 0001049	neural tube	6.55E-06
uberon 0001264	pancreas	7.13E-06
uberon 0001976	epithelium of esophagus	1.02E-05
uberon 0002348	epicardium	1.02E-05
uberon 2005409	pars superior ear	1.02E-05
uberon 0007831	pectoral girdle skeleton	2.43E-05
uberon 0002531	paired fin bud	2.43E-05
uberon 0002328	notochord	3.42E-05
uberon 0003901	horizontal septum	3.43E-05
uberon 2001089	myoseptum	4.54E-05
uberon 0004742	dentary	4.74E-05
uberon 0007329	pancreatic duct	6.29E-05
uberon 2000168	anterior macula	8.13E-05
uberon 0000959	optic chiasma	8.13E-05

uberont 0003069	otic placode	8.47E-05
uberont 0005507	rhombomere 3	0.00010303
uberont 0005886	post-hyoid pharyngeal arch skeleton	0.00012482
uberont 0011615	basihyal cartilage	0.00013041
uberont 0001703	neurocranium	0.00013041
uberont 0000468	multicellular organism	0.000148
uberont 0003098	optic stalk	0.00015702
uberont 2001239	ceratobranchial 5 bone	0.00018976
uberont 0005515	rhombomere 5	0.00022664
uberont 0009635	parachordal cartilage	0.00022664
uberont 0002533	post-anal tail bud	0.00022939
uberont 0002241	chondrocranium	0.00026786
uberont 0000965	lens of camera-type eye	0.00033083
uberont 0004741	cleithrum	0.00036419
uberont 0004291	heart rudiment	0.00036419
uberont 2000422	retroarticular	0.00045298
uberont 2001425	basal plate cartilage	0.00045298
uberont 0002087	atrioventricular canal	0.00054642
uberont 0006860	swim bladder	0.00061179
uberont 0001135	smooth muscle tissue	0.00067652
uberont 0000948	heart	0.00071284
uberont 0002107	liver	0.00078239
uberont 0003278	skeleton of lower jaw	0.00083833
uberont 0001846	internal ear	0.00090271
uberont 2000657	entopterygoid	0.00094304
uberont 0001900	ventral thalamus	0.00094304
uberont 0001043	esophagus	0.00094304
uberont 4000175	pectoral fin lepidotrichium	0.00094304
uberont 2000694	ceratobranchial 5 tooth	0.00096343
uberont 0000935	anterior commissure	0.00106543
uberont 0002407	pericardium	0.00112019
uberont 0001277	intestinal epithelium	0.00117402
uberont 0005412	optic fissure	0.00117402
uberont 0002394	bile duct	0.00125195
uberont 0005387	olfactory glomerulus	0.00125195
uberont 0000058	duct	0.00125195
uberont 0003936	postoptic commissure	0.00141162
uberont 0003091	thyroid primordium	0.00160269
uberont 0000936	posterior commissure	0.00160269



uberont 0004753	scapulocoracoid	0.00160269
uberont 0007274	crista of ampulla of anterior semicircular duct of membranous labyrinth	0.00199471
uberont 2001256	lateral floor plate	0.00199471
uberont 0000931	proctodeum	0.00199471
uberont 0010170	region of neural crest	0.00242745
uberont 0002342	neural crest	0.00242745
uberont 2000040	median fin fold	0.00265588
uberont 0003077	paraxial mesoderm	0.00290037
uberont 2001076	intestinal bulb	0.00290037
uberont 0005305	thyroid follicle	0.00290037
uberont 0011611	ceratohyal bone	0.00290037
uberont 0003052	midbrain-hindbrain boundary	0.00393324
uberont 0002397	maxilla	0.00455478
uberont 0011606	hyomandibular bone	0.00455478
uberont 0003932	cartilage element of chondrocranium	0.00455478
uberont 2000476	branchiostegal ray	0.00455478
uberont 0003072	optic cup	0.00455478
uberont 2001069	ventral fin fold	0.00494971
uberont 0002518	otolith organ	0.00518303
uberont 0001213	intestinal villus	0.00518303
uberont 0001997	olfactory epithelium	0.0058488
uberont 0004117	pharyngeal pouch	0.0058488
uberont 0004740	basibranchial bone	0.00655156
uberont 0004745	parasphenoid	0.00655156
uberont 0001891	midbrain	0.00826705
uberont 0004739	pronephric glomerulus	0.0105616
uberont 0003081	lateral plate mesoderm	0.0106027
uberont 0001032	sensory system	0.01156592

Supplementary Table S2.12. The top 100 enriched Uberon terms for the forelimb conserved genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0002102	forelimb	4.43E-37
uberon 0001708	jaw skeleton	1.16E-28
uberon 0003128	cranium	4.11E-28
uberon 0011156	facial skeleton	1.81E-27
uberon 0002470	autopod region	3.24E-26
uberon 0005944	axial skeleton plus cranial skeleton	4.09E-26
uberon 0007811	craniocervical region	1.04E-25
uberon 0002091	appendicular skeleton	2.32E-25
uberon 0002544	digit	1.23E-24
uberon 0001434	skeletal system	1.90E-22
uberon 0000165	mouth	5.07E-22
uberon 0001684	mandible	2.12E-21
uberon 0002103	hindlimb	2.86E-21
uberon 0004716	conceptus	7.68E-21
uberon 0001456	face	2.78E-20
uberon 0000033	head	1.73E-19
uberon 0001690	ear	1.81E-19
uberon 0001716	secondary palate	2.24E-19
uberon 0001703	neurocranium	2.60E-19
uberon 0003822	forelimb stylopod	1.00E-18
uberon 0002105	vestibulo-auditory system	1.84E-18
uberon 0001756	middle ear	4.24E-17
uberon 0002397	maxilla	8.19E-17
uberon 0001007	digestive system	1.15E-16
uberon 0003216	hard palate	4.95E-16
uberon 0005619	secondary palatal shelf	5.08E-16
uberon 0002418	cartilage tissue	1.04E-15
uberon 0001049	neural tube	1.98E-15
uberon 0003823	hindlimb zeugopod	4.07E-15
uberon 0005871	palatine process of maxilla	9.49E-15
uberon 0006849	scapula	1.02E-14
uberon 0003252	thoracic rib cage	3.52E-14
uberon 0000975	sternum	5.39E-14

uberon 0001446	fibula	1.03E-13
uberon 0000976	humerus	1.44E-13
uberon 0001890	forebrain	4.82E-13
uberon 0001004	respiratory system	5.98E-13
uberon 0003107	meckel's cartilage	8.30E-13
uberon 0000209	tetrapod frontal bone	2.05E-12
uberon 0005417	forelimb bud	2.47E-12
uberon 0003450	upper jaw incisor	2.70E-12
uberon 0002228	rib	2.77E-12
uberon 0006428	basisphenoid bone	2.84E-12
uberon 0001677	sphenoid bone	2.91E-12
uberon 0000955	brain	7.90E-12
uberon 0005062	neural fold	1.18E-11
uberon 0002218	tympanic ring	1.38E-11
uberon 0001066	intervertebral disk	1.55E-11
uberon 0001689	malleus bone	2.21E-11
uberon 0001723	tongue	2.31E-11
uberon 0000922	embryo	2.75E-11
uberon 0002517	basicranium	3.02E-11
uberon 0004535	cardiovascular system	3.29E-11
uberon 0000979	tibia	3.38E-11
uberon 0001681	nasal bone	4.41E-11
uberon 0002104	visual system	6.79E-11
uberon 0000383	musculature of body	1.18E-10
uberon 0001424	ulna	1.39E-10
uberon 0003655	molar tooth	1.56E-10
uberon 0000012	somatic nervous system	1.56E-10
uberon 0002407	pericardium	1.62E-10
uberon 0006207	aortico-pulmonary spiral septum	2.18E-10
uberon 0000004	nose	2.59E-10
uberon 0002012	pulmonary artery	2.72E-10
uberon 0004356	apical ectodermal ridge	3.54E-10
uberon 0008828	presphenoid bone	4.91E-10
uberon 0000401	mandibular ramus	5.74E-10
uberon 0005620	primary palate	7.31E-10
uberon 0000080	mesonephros	8.46E-10
uberon 0003451	lower jaw incisor	1.03E-09
uberon 0000309	body wall	1.16E-09
uberon 0001737	larynx	1.36E-09

uberon 0000948	heart	1.47E-09
uberon 0002416	integumental system	1.49E-09
uberon 0006772	long bone epiphyseal plate hypertrophic zone	1.72E-09
uberon 0004347	limb bud	2.09E-09
uberon 0001435	carpal bone	2.26E-09
uberon 0004362	pharyngeal arch 1	2.55E-09
uberon 0002516	epiphyseal plate	3.17E-09
uberon 0001676	occipital bone	3.27E-09
uberon 0001423	radius bone	3.77E-09
uberon 0003697	abdominal wall	3.77E-09
uberon 0001738	thyroid cartilage	4.08E-09
uberon 0001894	diencephalon	4.44E-09
uberon 0004657	mandible condylar process	6.08E-09
uberon 0005175	chest organ	6.17E-09
uberon 0001682	palatine bone	6.18E-09
uberon 0002087	atrioventricular canal	6.48E-09
uberon 0001844	cochlea	8.06E-09
uberon 0004649	sphenoid bone pterygoid process	8.79E-09
uberon 0006721	alisphenoid bone	8.79E-09
uberon 0003221	phalanx	8.89E-09
uberon 0000019	camera-type eye	1.19E-08
uberon 0000210	tetrapod parietal bone	1.46E-08
uberon 0001008	renal system	1.66E-08
uberon 0010380	enteric nerve	1.70E-08
uberon 0003604	trachea cartilage	1.98E-08
uberon 0001688	incus bone	1.98E-08
uberon 0002328	notochord	2.23E-08
uberon 0003075	neural plate	2.30E-08

Supplementary Table S2.13. The top 100 enriched Uberon terms for the forelimb module-specific genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0002102	forelimb	1.10E-239
uberon 0002091	appendicular skeleton	2.35E-177
uberon 0002103	hindlimb	1.51E-122
uberon 0000976	humerus	8.36E-120
uberon 0003822	forelimb stylopod	4.90E-119
uberon 0001434	skeletal system	1.79E-113
uberon 0001424	ulna	1.34E-111
uberon 0001423	radius bone	1.77E-102
uberon 0003823	hindlimb zeugopod	2.32E-91
uberon 0005944	axial skeleton plus cranial skeleton	2.97E-76
uberon 0000979	tibia	3.92E-76
uberon 0002470	autopod region	3.30E-72
uberon 0000376	hindlimb stylopod	3.42E-72
uberon 0003252	thoracic rib cage	4.48E-72
uberon 0000981	femur	6.47E-70
uberon 0003128	cranium	6.45E-68
uberon 0007811	craniocervical region	1.29E-66
uberon 0002544	digit	4.86E-63
uberon 0011156	facial skeleton	6.22E-57
uberon 0002228	rib	2.29E-56
uberon 0003221	phalanx	1.11E-54
uberon 0001446	fibula	4.66E-53
uberon 0000165	mouth	8.61E-53
uberon 0001456	face	1.27E-52
uberon 0001708	jaw skeleton	4.02E-50
uberon 0000033	head	5.72E-50
uberon 0001703	neurocranium	6.78E-50
uberon 0002516	epiphyseal plate	3.16E-49
uberon 0002374	metacarpal bone	2.51E-47
uberon 0001684	mandible	1.92E-42
uberon 0000975	sternum	3.11E-42
uberon 0004716	conceptus	9.92E-42
uberon 0002471	zeugopod	6.47E-37

uberon 0006772	long bone epiphyseal plate hypertrophic zone	1.10E-36
uberon 0001007	digestive system	3.07E-34
uberon 0001435	carpal bone	4.61E-34
uberon 0001448	metatarsal bone	1.16E-33
uberon 0001716	secondary palate	1.57E-32
uberon 0001004	respiratory system	1.39E-29
uberon 0002483	trabecular bone tissue	8.58E-29
uberon 0002517	basicranium	2.16E-27
uberon 0002397	maxilla	2.54E-27
uberon 0006333	snout	2.88E-27
uberon 0006849	scapula	5.91E-26
uberon 0001049	neural tube	9.50E-25
uberon 0001130	vertebral column	9.92E-25
uberon 0002498	deltpectoral crest	1.29E-24
uberon 0001105	clavicle bone	2.09E-24
uberon 0001677	sphenoid bone	4.67E-24
uberon 0001676	occipital bone	2.17E-23
uberon 0002418	cartilage tissue	9.31E-23
uberon 0001690	ear	1.69E-22
uberon 0003216	hard palate	2.87E-22
uberon 0000922	embryo	5.19E-22
uberon 0002105	vestibulo-auditory system	7.47E-22
uberon 0004535	cardiovascular system	7.51E-22
uberon 0006771	long bone epiphyseal plate proliferative zone	1.65E-20
uberon 0001095	caudal vertebra	1.77E-20
uberon 0003861	neural arch	2.07E-20
uberon 0007812	post-anal tail	2.78E-20
uberon 0004347	limb bud	3.20E-20
uberon 0004356	apical ectodermal ridge	3.60E-20
uberon 0001008	renal system	4.07E-20
uberon 0000004	nose	4.88E-20
uberon 0001756	middle ear	1.32E-19
uberon 0000014	zone of skin	2.28E-19
uberon 0002229	interparietal bone	7.07E-19
uberon 0001681	nasal bone	7.20E-18
uberon 0001711	eyelid	1.21E-17
uberon 0000210	tetrapod parietal bone	2.18E-17
uberon 0000159	anal canal	2.67E-17
uberon 0002048	lung	2.80E-17

uberon 0002412	vertebra	3.27E-17
uberon 0003461	shoulder bone	3.27E-17
uberon 0000209	tetrapod frontal bone	4.02E-17
uberon 0001447	tarsal bone	1.06E-16
uberon 0002104	visual system	1.24E-16
uberon 0000982	skeletal joint	1.27E-16
uberon 0007830	pelvic girdle bone/zone	2.36E-16
uberon 0002208	sternebra	1.75E-15
uberon 0002446	patella	1.80E-15
uberon 0001066	intervertebral disk	3.19E-15
uberon 0000924	ectoderm	3.86E-15
uberon 0001723	tongue	5.35E-15
uberon 0000383	musculature of body	6.69E-15
uberon 0001439	compact bone tissue	7.86E-15
uberon 0000019	camera-type eye	7.88E-15
uberon 0001245	anus	8.40E-15
uberon 0000947	aorta	2.38E-14
uberon 0006430	xiphoid cartilage	2.45E-14
uberon 0002113	kidney	2.52E-14
uberon 0006428	basisphenoid bone	3.81E-14
uberon 0001678	temporal bone	4.02E-14
uberon 0000012	somatic nervous system	4.20E-14
uberon 0008867	trabecular network of bone	5.36E-14
uberon 0002347	thoracic vertebra	6.04E-14
uberon 0002413	cervical vertebra	1.90E-13
uberon 0001075	bony vertebral centrum	1.94E-13
uberon 0004747	supraoccipital bone	2.00E-13
uberon 0000323	late embryo	2.03E-13

Supplementary Table S2.14. The enriched Biological Process terms from the Gene Ontology for the mouse orthologs of the pectoral fin module-specific genes.

Term identifier	Term name	P-value
GO:0007275	multicellular organism development	5.68E-14
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	3.32E-10
GO:0009887	organ morphogenesis	6.91E-09
GO:0006355	regulation of transcription, DNA-templated	4.00E-08
GO:0006351	transcription, DNA-templated	3.72E-07
GO:0060070	canonical Wnt signaling pathway	4.94E-07
GO:0043588	skin development	6.59E-07
GO:0042475	odontogenesis of dentin-containing tooth	1.08E-06
GO:0006357	regulation of transcription from RNA polymerase II promoter	1.41E-06
GO:0045893	positive regulation of transcription, DNA-templated	1.22E-05
GO:0030154	cell differentiation	1.67E-05
GO:0016055	Wnt signaling pathway	1.91E-05
GO:0021766	hippocampus development	4.51E-05
GO:0010628	positive regulation of gene expression	5.57E-05
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	1.37E-04
GO:0090090	negative regulation of canonical Wnt signaling pathway	2.21E-04
GO:0035904	aorta development	3.56E-04
GO:0006024	glycosaminoglycan biosynthetic process	4.45E-04
GO:0021522	spinal cord motor neuron differentiation	4.94E-04
GO:0007507	heart development	5.13E-04
GO:0007224	smoothened signaling pathway	5.74E-04
GO:0030182	neuron differentiation	6.07E-04
GO:0043010	camera-type eye development	8.19E-04
GO:0060021	palate development	0.00113011
GO:0007281	germ cell development	0.00128033
GO:0001843	neural tube closure	0.00184155
GO:0002062	chondrocyte differentiation	0.00192196
GO:0001657	ureteric bud development	0.00217299
GO:0008284	positive regulation of cell proliferation	0.00252077
GO:0001756	somitogenesis	0.00337659
GO:0030111	regulation of Wnt signaling pathway	0.00412055
GO:0015012	heparan sulfate proteoglycan biosynthetic process	0.00412055
GO:0030324	lung development	0.00446791
GO:0071300	cellular response to retinoic acid	0.00513915



GO:0001701	in utero embryonic development	0.00546615
GO:0045669	positive regulation of osteoblast differentiation	0.00582864
GO:0001568	blood vessel development	0.00657098
GO:0045165	cell fate commitment	0.00683035
GO:0048844	artery morphogenesis	0.00788494
GO:0045665	negative regulation of neuron differentiation	0.00792833
GO:0030334	regulation of cell migration	0.00821812
GO:0051216	cartilage development	0.00976019
GO:0043066	negative regulation of apoptotic process	0.01109612
GO:0021983	pituitary gland development	0.0112363
GO:0001837	epithelial to mesenchymal transition	0.01196962
GO:0007417	central nervous system development	0.01218445
GO:0045892	negative regulation of transcription, DNA-templated	0.01259202
GO:0035116	embryonic hindlimb morphogenesis	0.01349739
GO:0010718	positive regulation of epithelial to mesenchymal transition	0.01429137
GO:0007399	nervous system development	0.01594534
GO:0061153	trachea gland development	0.01600806
GO:0060976	coronary vasculature development	0.01679091
GO:0001656	metanephros development	0.01766254
GO:0003281	ventricular septum development	0.01766254
GO:0071787	endoplasmic reticulum tubular network assembly	0.02128751
GO:0034653	retinoic acid catabolic process	0.02128751
GO:0008543	fibroblast growth factor receptor signaling pathway	0.02229888
GO:0001649	osteoblast differentiation	0.02292059
GO:0048706	embryonic skeletal system development	0.02328026
GO:0002051	osteoblast fate commitment	0.02653893
GO:0048755	branching morphogenesis of a nerve	0.02653893
GO:0001755	neural crest cell migration	0.02844596
GO:0001942	hair follicle development	0.03174493
GO:0072177	mesonephric duct development	0.03176245
GO:2000343	positive regulation of chemokine (C-X-C motif) ligand 2 production	0.03176245
GO:0060662	salivary gland cavitation	0.03176245
GO:0060789	hair follicle placode formation	0.03176245
GO:0042127	regulation of cell proliferation	0.0339043
GO:0060348	bone development	0.03636502
GO:0009913	epidermal cell differentiation	0.03695824
GO:1901522	positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus	0.03695824
GO:0021554	optic nerve development	0.03695824

GO:0071168	protein localization to chromatin	0.03695824
GO:0048752	semicircular canal morphogenesis	0.03695824
GO:0048793	pronephros development	0.03695824
GO:0048341	paraxial mesoderm formation	0.04212643
GO:0048702	embryonic neurocranium morphogenesis	0.04212643
GO:0030853	negative regulation of granulocyte differentiation	0.04212643
GO:0001947	heart looping	0.04374438
GO:0016567	protein ubiquitination	0.04533511

Supplementary Table S2.15. The top 100 enriched Uberon terms for the mouse orthologs of the pectoral fin module-specific genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0000922	embryo	6.92E-12
uberon 0004716	conceptus	2.01E-11
uberon 0007811	craniocervical region	8.10E-11
uberon 0001049	neural tube	3.87E-10
uberon 0005944	axial skeleton plus cranial skeleton	1.06E-09
uberon 0003128	cranium	1.35E-09
uberon 0000033	head	1.99E-09
uberon 0001456	face	5.43E-08
uberon 0001434	skeletal system	5.81E-07
uberon 0000165	mouth	5.91E-07
uberon 0001703	neurocranium	1.11E-06
uberon 0001007	digestive system	2.10E-06
uberon 0000926	mesoderm	2.55E-06
uberon 0001711	eyelid	2.75E-06
uberon 0004341	primitive streak	3.74E-06
uberon 0011156	facial skeleton	5.55E-06
uberon 0003457	head bone	7.51E-06
uberon 0005070	anterior neuropore	1.28E-05
uberon 0001819	palpebral fissure	2.93E-05
uberon 0001708	jaw skeleton	2.93E-05
uberon 0000478	extraembryonic structure	3.52E-05
uberon 0000947	aorta	4.31E-05
uberon 0003655	molar tooth	5.25E-05
uberon 0004022	germinal neuroepithelium	7.58E-05
uberon 0000948	heart	9.02E-05
uberon 0000019	camera-type eye	9.15E-05
uberon 0002104	visual system	9.22E-05
uberon 0005600	crus commune	9.42E-05
uberon 0001860	endolymphatic duct	0.0001457
uberon 0003451	lower jaw incisor	0.00017578
uberon 0002470	autopod region	0.00019574
uberon 0001508	arch of aorta	0.0002267
uberon 0002167	right lung	0.00023623

uberont 0001004	respiratory system	0.0002535
uberont 0001690	ear	0.00026285
uberont 0002105	vestibulo-auditory system	0.00028019
uberont 0003216	hard palate	0.00029064
uberont 0010190	pair of dorsal aortae	0.00029282
uberont 0002384	connective tissue	0.00030192
uberont 0004044	anterior visceral endoderm	0.00043905
uberont 0001818	tarsal gland	0.00049939
uberont 0035077	lateral nasal gland	0.00062633
uberont 0000014	zone of skin	0.0006311
uberont 0002544	digit	0.00066645
uberont 0002827	vestibulocochlear ganglion	0.00071149
uberont 0003955	molar crown	0.0007931
uberont 0001756	middle ear	0.00080729
uberont 0004535	cardiovascular system	0.00087226
uberont 0000423	eccrine sweat gland	0.00087311
uberont 0001601	extra-ocular muscle	0.00087311
uberont 0000004	nose	0.00090415
uberont 0002539	pharyngeal arch	0.00091495
uberont 0000084	ureteric bud	0.0009682
uberont 0001751	dentine	0.00097362
uberont 0003066	pharyngeal arch 2	0.00097362
uberont 0002168	left lung	0.00120309
uberont 0001688	incus bone	0.00128991
uberont 0000091	bilaminar disc	0.00140356
uberont 0011864	tendon collagen fibril	0.00148402
uberont 0001716	secondary palate	0.00151298
uberont 0003544	brain white matter	0.0016391
uberont 0000955	brain	0.00172361
uberont 0003051	ear vesicle	0.00180251
uberont 0003068	axial mesoderm	0.00184711
uberont 0007833	osseus semicircular canal	0.00187241
uberont 0000309	body wall	0.00193054
uberont 0002091	appendicular skeleton	0.00207267
uberont 0004573	systemic artery	0.00222527
uberont 0010513	strand of zigzag hair	0.00225969
uberont 0000945	stomach	0.0023366
uberont 0010409	ocular surface region	0.00233819
uberont 0004043	semicircular canal ampulla	0.00242659

uberon 0005356	rathke's pouch	0.00242659
uberon 0003950	inner ear canal	0.00254497
uberon 0001890	forebrain	0.00257106
uberon 0003252	thoracic rib cage	0.00267041
uberon 0008854	root of molar tooth	0.00268605
uberon 0008799	transverse palatine fold	0.00268605
uberon 0002418	cartilage tissue	0.00271496
uberon 0001199	mucosa of stomach	0.00287821
uberon 0005291	embryonic tissue	0.00300689
uberon 0001167	wall of stomach	0.00324016
uberon 0000007	pituitary gland	0.00332109
uberon 0001862	vestibular labyrinth	0.003632
uberon 0002413	cervical vertebra	0.00380992
uberon 0004362	pharyngeal arch 1	0.00403845
uberon 0001872	parietal lobe	0.00403845
uberon 0002329	somite	0.00405486
uberon 0001908	optic tract	0.00421892
uberon 0001274	ischium	0.00421892
uberon 0001853	utricle of membranous labyrinth	0.00477616
uberon 0001752	enamel	0.00477616
uberon 0002196	adenohypophysis	0.00516843
uberon 0005176	tooth enamel organ	0.00541814
uberon 0002487	tooth cavity	0.00541814
uberon 0010197	trunk of common carotid artery	0.00541814
uberon 0004090	periorbital region	0.00541814
uberon 0001894	diencephalon	0.00586178
uberon 0001854	sacculle of membranous labyrinth	0.00617638
uberon 0004212	glomerular capillary	0.00617638

Supplementary Table S2.16. The enriched Biological Process terms from the Gene Ontology for the zebrafish orthologs of the forelimb module-specific genes.

Term identifier	Term name	P-value
GO:0006355	regulation of transcription, DNA-templated	2.74E-11
GO:0007275	multicellular organism development	2.01E-09
GO:0009953	dorsal/ventral pattern formation	1.36E-08
GO:0016055	Wnt signaling pathway	6.90E-08
GO:0006351	transcription, DNA-templated	2.18E-07
GO:0030509	BMP signaling pathway	5.94E-07
GO:0030182	neuron differentiation	7.04E-07
GO:0035118	embryonic pectoral fin morphogenesis	1.60E-05
GO:0045165	cell fate commitment	5.38E-05
GO:0030510	regulation of BMP signaling pathway	1.57E-04
GO:0048701	embryonic cranial skeleton morphogenesis	4.43E-04
GO:0002072	optic cup morphogenesis involved in camera-type eye development	4.70E-04
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	0.00117724
GO:0040007	growth	0.00129193
GO:0060027	convergent extension involved in gastrulation	0.00135252
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	0.0015685
GO:0042074	cell migration involved in gastrulation	0.00165586
GO:0048468	cell development	0.00172114
GO:0060395	SMAD protein signal transduction	0.00188379
GO:0030903	notochord development	0.00205679
GO:0043408	regulation of MAPK cascade	0.00205679
GO:0007179	transforming growth factor beta receptor signaling pathway	0.00414164
GO:0021592	fourth ventricle development	0.00421122
GO:0045743	positive regulation of fibroblast growth factor receptor signaling pathway	0.00536967
GO:0048703	embryonic viscerocranium morphogenesis	0.00585852
GO:0060070	canonical Wnt signaling pathway	0.0061092
GO:0001944	vasculature development	0.00649015
GO:0015012	heparan sulfate proteoglycan biosynthetic process	0.00665665
GO:0002062	chondrocyte differentiation	0.00665665
GO:0030097	hemopoiesis	0.00670985
GO:0008543	fibroblast growth factor receptor signaling pathway	0.0095973
GO:0007417	central nervous system development	0.01064297
GO:0061386	closure of optic fissure	0.01302316
GO:0006024	glycosaminoglycan biosynthetic process	0.01302316

GO:0071910	determination of liver left/right asymmetry	0.01490327
GO:0007064	mitotic sister chromatid cohesion	0.01689251
GO:0009948	anterior/posterior axis specification	0.01898788
GO:0051216	cartilage development	0.01918023
GO:0007155	cell adhesion	0.01955582
GO:0061371	determination of heart left/right asymmetry	0.02012978
GO:0045595	regulation of cell differentiation	0.02118641
GO:0030198	extracellular matrix organization	0.0234852
GO:0060026	convergent extension	0.02377358
GO:0038108	negative regulation of appetite by leptin-mediated signaling pathway	0.02505648
GO:2000366	positive regulation of STAT protein import into nucleus	0.02505648
GO:0060912	cardiac cell fate specification	0.02505648
GO:1900745	positive regulation of p38MAPK cascade	0.02505648
GO:2000583	regulation of platelet-derived growth factor receptor-alpha signaling pathway	0.02505648
GO:0002076	osteoblast development	0.02505648
GO:0070587	regulation of cell-cell adhesion involved in gastrulation	0.02505648
GO:1990051	activation of protein kinase C activity	0.02505648
GO:0043010	camera-type eye development	0.02506223
GO:0060041	retina development in camera-type eye	0.02513516
GO:0060828	regulation of canonical Wnt signaling pathway	0.02588139
GO:0001503	ossification	0.02588139
GO:0060037	pharyngeal system development	0.02837219
GO:0050767	regulation of neurogenesis	0.02837219
GO:0002138	retinoic acid biosynthetic process	0.03734953
GO:0032481	positive regulation of type I interferon production	0.03734953
GO:0010332	response to gamma radiation	0.03734953
GO:0032965	regulation of collagen biosynthetic process	0.03734953
GO:0006978	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	0.03734953
GO:0010470	regulation of gastrulation	0.03734953
GO:0071733	transcriptional activation by promoter-enhancer looping	0.03734953
GO:0030917	midbrain-hindbrain boundary development	0.04515412
GO:0001568	blood vessel development	0.04877064
GO:0048730	epidermis morphogenesis	0.0494884
GO:0032868	response to insulin	0.0494884

Supplementary Table S2.17. The enriched Uberon terms for the zebrafish orthologs of the forelimb module-specific genes.

Uberon identifier	Term name	P-value
uberont0008823	neural tube derived brain	2.29E-07
uberont0000033	head	2.96E-07
uberont0000468	multicellular organism	3.10E-07
uberont0000019	camera-type eye	2.07E-05
uberont0001555	digestive tract	9.84E-05
uberont0001016	nervous system	0.00082865
uberont0002107	liver	0.00093463
uberont0001017	central nervous system	0.00141338
uberont2000084	yolk	0.00145089
uberont4000164	caudal fin	0.00325469
uberont0000948	heart	0.00876155
uberont0008895	splanchnocranium	0.01593721
uberont0002280	otolith	0.01730733
uberont0005281	ventricular system of central nervous system	0.02130098
uberont2000033	intermediate cell mass of mesoderm	0.02183289
uberont0001782	pigmented layer of retina	0.02256819
uberont0001231	nephron tubule	0.02976965
uberont0001155	colon	0.02976965
uberont0001708	jaw skeleton	0.0342243
uberont0005886	post-hyoid pharyngeal arch skeleton	0.03545473
uberont0001647	facial nerve	0.03865184
uberont0002407	pericardium	0.04825439



Supplementary Table S2.18. The top 100 enriched Biological Process terms from the Gene Ontology for the pelvic fin module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term identifier	Term name	P-value
GO:0016310	phosphorylation	1.07E-30
GO:0006468	protein phosphorylation	4.11E-19
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	1.03E-14
GO:0033333	fin development	1.66E-10
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	8.51E-09
GO:0007507	heart development	9.67E-09
GO:0006470	protein dephosphorylation	6.89E-08
GO:0007275	multicellular organism development	1.13E-07
GO:0030154	cell differentiation	2.15E-07
GO:0007264	small GTPase mediated signal transduction	2.32E-07
GO:0051216	cartilage development	4.89E-07
GO:0001947	heart looping	6.53E-07
GO:0038083	peptidyl-tyrosine autophosphorylation	1.22E-06
GO:0007188	adenylate cyclase-modulating G-protein coupled receptor signaling pathway	6.36E-06
GO:0033334	fin morphogenesis	7.58E-06
GO:0031290	retinal ganglion cell axon guidance	8.15E-06
GO:0006351	transcription, DNA-templated	1.15E-05
GO:0030509	BMP signaling pathway	1.51E-05
GO:0031101	fin regeneration	1.75E-05
GO:0006164	purine nucleotide biosynthetic process	2.05E-05
GO:0001649	osteoblast differentiation	3.11E-05
GO:0001889	liver development	3.85E-05
GO:0000028	ribosomal small subunit assembly	6.47E-05
GO:0043009	chordate embryonic development	6.89E-05
GO:0009953	dorsal/ventral pattern formation	7.59E-05
GO:0043010	camera-type eye development	8.17E-05
GO:0000027	ribosomal large subunit assembly	8.95E-05
GO:0006177	GMP biosynthetic process	9.26E-05
GO:0001568	blood vessel development	9.52E-05
GO:0030514	negative regulation of BMP signaling pathway	9.69E-05
GO:0006412	translation	1.15E-04

GO:0048066	developmental pigmentation	1.48E-04
GO:0030318	melanocyte differentiation	1.81E-04
GO:0003007	heart morphogenesis	2.04E-04
GO:0060536	cartilage morphogenesis	2.09E-04
GO:0045859	regulation of protein kinase activity	2.78E-04
GO:0007492	endoderm development	4.22E-04
GO:0009952	anterior/posterior pattern specification	5.33E-04
GO:0045087	innate immune response	5.44E-04
GO:0001501	skeletal system development	5.97E-04
GO:0086010	membrane depolarization during action potential	6.35E-04
GO:0048565	digestive tract development	7.98E-04
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	8.73E-04
GO:0008016	regulation of heart contraction	9.21E-04
GO:0016055	Wnt signaling pathway	0.001096677
GO:0060973	cell migration involved in heart development	0.00118875
GO:0021514	ventral spinal cord interneuron differentiation	0.00119353
GO:0060070	canonical Wnt signaling pathway	0.001316618
GO:0030902	hindbrain development	0.001530477
GO:0097324	melanocyte migration	0.001745185
GO:0035138	pectoral fin morphogenesis	0.001745185
GO:0043473	pigmentation	0.001763508
GO:0033339	pectoral fin development	0.002041423
GO:0045165	cell fate commitment	0.002349988
GO:0042127	regulation of cell proliferation	0.002429371
GO:0031284	positive regulation of guanylate cyclase activity	0.002430008
GO:0007263	nitric oxide mediated signal transduction	0.002430008
GO:0090131	mesenchyme migration	0.002430008
GO:0000398	mRNA splicing, via spliceosome	0.002440101
GO:0034097	response to cytokine	0.00244015
GO:0045893	positive regulation of transcription, DNA-templated	0.002507479
GO:0060026	convergent extension	0.002687487
GO:0060037	pharyngeal system development	0.002740327
GO:0048264	determination of ventral identity	0.002740327
GO:0060041	retina development in camera-type eye	0.003792248
GO:0048484	enteric nervous system development	0.003874613
GO:0043049	otic placode formation	0.003874613
GO:0035050	embryonic heart tube development	0.003928907
GO:0002574	thrombocyte differentiation	0.004285938

GO:0006355	regulation of transcription, DNA-templated	0.004473334
GO:0048703	embryonic viscerocranium morphogenesis	0.004655303
GO:0038061	NIK/NF-kappaB signaling	0.004767449
GO:0045776	negative regulation of blood pressure	0.004767449
GO:0031017	exocrine pancreas development	0.00495185
GO:0001525	angiogenesis	0.005226845
GO:0021984	adenohypophysis development	0.005453861
GO:0042476	odontogenesis	0.005453861
GO:0006383	transcription from RNA polymerase III promoter	0.005453861
GO:0030097	hemopoiesis	0.005661671
GO:0035118	embryonic pectoral fin morphogenesis	0.006112377
GO:0007179	transforming growth factor beta receptor signaling pathway	0.006112377
GO:0048010	vascular endothelial growth factor receptor signaling pathway	0.006794668
GO:0014003	oligodendrocyte development	0.006794668
GO:0006360	transcription from RNA polymerase I promoter	0.006794668
GO:0048709	oligodendrocyte differentiation	0.006794668
GO:0006457	protein folding	0.00701842
GO:0007368	determination of left/right symmetry	0.007696847
GO:0006809	nitric oxide biosynthetic process	0.00779477
GO:0045909	positive regulation of vasodilation	0.00779477
GO:0006207	'de novo' pyrimidine nucleobase biosynthetic process	0.00779477
GO:0060975	cardioblast migration to the midline involved in heart field formation	0.00779477
GO:0048935	peripheral nervous system neuron development	0.00779477
GO:0048699	generation of neurons	0.008314213
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	0.008666322
GO:0021522	spinal cord motor neuron differentiation	0.010017396
GO:0006260	DNA replication	0.010509899
GO:1902766	skeletal muscle satellite cell migration	0.011470527
GO:0044319	wound healing, spreading of cells	0.011470527
GO:0044211	CTP salvage	0.011470527
GO:0043097	pyrimidine nucleoside salvage	0.011470527

Supplementary Table S2.19. The top 100 enriched Biological Process terms from the Gene Ontology for the pelvic fin conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term identifier	Term name	P-value
GO:0009953	dorsal/ventral pattern formation	7.42E-18
GO:0030509	BMP signaling pathway	1.10E-09
GO:0007275	multicellular organism development	1.69E-08
GO:0042664	negative regulation of endodermal cell fate specification	4.29E-08
GO:0010002	cardioblast differentiation	7.70E-08
GO:0001947	heart looping	9.05E-08
GO:0009952	anterior/posterior pattern specification	1.91E-07
GO:0007224	smoothened signaling pathway	2.23E-07
GO:0001756	somitogenesis	2.23E-07
GO:0001889	liver development	2.84E-07
GO:0048703	embryonic viscerocranium morphogenesis	3.87E-07
GO:0033339	pectoral fin development	5.44E-07
GO:0055002	striated muscle cell development	6.00E-07
GO:0008543	fibroblast growth factor receptor signaling pathway	8.67E-07
GO:0007517	muscle organ development	1.00E-06
GO:0033334	fin morphogenesis	1.81E-06
GO:0030903	notochord development	2.49E-06
GO:0031647	regulation of protein stability	2.52E-06
GO:0006355	regulation of transcription, DNA-templated	5.90E-06
GO:0007368	determination of left/right symmetry	1.10E-05
GO:0030182	neuron differentiation	1.22E-05
GO:0031101	fin regeneration	2.11E-05
GO:0006351	transcription, DNA-templated	2.11E-05
GO:0016310	phosphorylation	2.35E-05
GO:0040007	growth	4.03E-05
GO:0021508	floor plate formation	4.45E-05
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	4.95E-05
GO:0060395	SMAD protein signal transduction	6.02E-05
GO:0043408	regulation of MAPK cascade	6.61E-05
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	8.22E-05
GO:0060956	endocardial cell differentiation	1.53E-04
GO:0048795	swim bladder morphogenesis	1.53E-04
GO:0033333	fin development	1.83E-04

GO:0048557	embryonic digestive tract morphogenesis	2.54E-04
GO:0035143	caudal fin morphogenesis	2.54E-04
GO:0016539	intein-mediated protein splicing	2.54E-04
GO:0009880	embryonic pattern specification	2.70E-04
GO:0007267	cell-cell signaling	3.14E-04
GO:0048793	pronephros development	3.14E-04
GO:0042694	muscle cell fate specification	3.79E-04
GO:0021703	locus ceruleus development	3.79E-04
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	3.92E-04
GO:0021953	central nervous system neuron differentiation	7.04E-04
GO:0048663	neuron fate commitment	7.04E-04
GO:0031016	pancreas development	8.73E-04
GO:0031076	embryonic camera-type eye development	0.00112343
GO:0060325	face morphogenesis	0.00112343
GO:0048468	cell development	0.00117991
GO:0048263	determination of dorsal identity	0.00163684
GO:0061371	determination of heart left/right asymmetry	0.00164982
GO:0021984	adenohypophysis development	0.00192807
GO:0048709	oligodendrocyte differentiation	0.002242
GO:0048794	swim bladder development	0.002242
GO:0001649	osteoblast differentiation	0.002242
GO:0030878	thyroid gland development	0.00257841
GO:0048752	semicircular canal morphogenesis	0.00257841
GO:0060070	canonical Wnt signaling pathway	0.00321008
GO:0008015	blood circulation	0.00331769
GO:0006468	protein phosphorylation	0.00348935
GO:0048702	embryonic neurocranium morphogenesis	0.00414411
GO:0007417	central nervous system development	0.00502765
GO:0048264	determination of ventral identity	0.00505585
GO:0001945	lymph vessel development	0.00505585
GO:0050767	regulation of neurogenesis	0.00505585
GO:0016525	negative regulation of angiogenesis	0.00554316
GO:0036342	post-anal tail morphogenesis	0.00554316
GO:0043049	otic placode formation	0.00605114
GO:0071599	otic vesicle development	0.00657958
GO:0007179	transforming growth factor beta receptor signaling pathway	0.00769693
GO:0007411	axon guidance	0.00789129
GO:0030917	midbrain-hindbrain boundary development	0.00828542
GO:0038066	p38MAPK cascade	0.0101675

GO:0002275	myeloid cell activation involved in immune response	0.0101675
GO:0051895	negative regulation of focal adhesion assembly	0.0101675
GO:0003303	BMP signaling pathway involved in heart jogging	0.0101675
GO:0060912	cardiac cell fate specification	0.0101675
GO:0061131	pancreas field specification	0.0101675
GO:0033340	pelvic fin development	0.0101675
GO:0010831	positive regulation of myotube differentiation	0.0101675
GO:0031018	endocrine pancreas development	0.01016764
GO:0060041	retina development in camera-type eye	0.01023913
GO:0030901	midbrain development	0.01222062
GO:0030900	forebrain development	0.01222062
GO:0045165	cell fate commitment	0.01222062
GO:0030097	hemopoiesis	0.01223899
GO:0048839	inner ear development	0.01294194
GO:0003146	heart jogging	0.01294194
GO:0048618	post-embryonic foregut morphogenesis	0.01521292
GO:0031174	lifelong otolith mineralization	0.01521292
GO:0031290	retinal ganglion cell axon guidance	0.01521407
GO:0048565	digestive tract development	0.01681701
GO:0051146	striated muscle cell differentiation	0.02023297
GO:0001957	intramembranous ossification	0.02023297
GO:0014707	branchiomic skeletal muscle development	0.02023297
GO:0033336	caudal fin development	0.02023297
GO:0048855	adenohypophysis morphogenesis	0.02023297
GO:0030902	hindbrain development	0.02112173
GO:0008285	negative regulation of cell proliferation	0.02203214
GO:0007519	skeletal muscle tissue development	0.02485899
GO:0021523	somatic motor neuron differentiation	0.02522775

Supplementary Table S2.20. The top 100 enriched Biological Process terms from the Gene Ontology for the hindlimb conserved genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term identifier	Term name	P-value
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	1.33E-27
GO:0010628	positive regulation of gene expression	3.28E-25
GO:0030326	embryonic limb morphogenesis	3.43E-22
GO:0042475	odontogenesis of dentin-containing tooth	3.43E-22
GO:0001701	in utero embryonic development	1.22E-20
GO:0007507	heart development	2.62E-20
GO:0045893	positive regulation of transcription, DNA-templated	9.95E-19
GO:0008284	positive regulation of cell proliferation	4.76E-18
GO:0030324	lung development	8.41E-18
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	1.20E-17
GO:0008285	negative regulation of cell proliferation	3.66E-17
GO:0035116	embryonic hindlimb morphogenesis	1.74E-15
GO:0001501	skeletal system development	2.43E-15
GO:0048754	branching morphogenesis of an epithelial tube	4.20E-15
GO:0003007	heart morphogenesis	1.93E-14
GO:0045669	positive regulation of osteoblast differentiation	3.66E-14
GO:0043066	negative regulation of apoptotic process	4.55E-14
GO:0001649	osteoblast differentiation	1.95E-13
GO:0051216	cartilage development	2.59E-13
GO:0010629	negative regulation of gene expression	3.72E-13
GO:0060021	palate development	3.76E-13
GO:0007389	pattern specification process	4.51E-13
GO:0002053	positive regulation of mesenchymal cell proliferation	5.57E-13
GO:0001947	heart looping	8.50E-13
GO:0042733	embryonic digit morphogenesis	1.53E-12
GO:0007275	multicellular organism development	2.68E-12
GO:0002062	chondrocyte differentiation	3.67E-12
GO:0035108	limb morphogenesis	5.54E-12
GO:0009953	dorsal/ventral pattern formation	6.30E-12
GO:0030509	BMP signaling pathway	2.01E-11
GO:0001525	angiogenesis	3.86E-11
GO:0045892	negative regulation of transcription, DNA-templated	1.17E-10
GO:0060441	epithelial tube branching involved in lung morphogenesis	1.98E-10

GO:0009952	anterior/posterior pattern specification	2.01E-10
GO:0001657	ureteric bud development	3.26E-10
GO:0001658	branching involved in ureteric bud morphogenesis	3.26E-10
GO:0071542	dopaminergic neuron differentiation	6.89E-10
GO:0001822	kidney development	7.65E-10
GO:0021983	pituitary gland development	8.59E-10
GO:0070374	positive regulation of ERK1 and ERK2 cascade	1.08E-09
GO:0060445	branching involved in salivary gland morphogenesis	1.38E-09
GO:0031016	pancreas development	1.59E-09
GO:0031069	hair follicle morphogenesis	2.32E-09
GO:0048589	developmental growth	2.32E-09
GO:0030501	positive regulation of bone mineralization	3.30E-09
GO:0009887	organ morphogenesis	4.50E-09
GO:0021904	dorsal/ventral neural tube patterning	5.30E-09
GO:0030901	midbrain development	5.40E-09
GO:0048568	embryonic organ development	5.40E-09
GO:0045165	cell fate commitment	6.10E-09
GO:0045596	negative regulation of cell differentiation	6.42E-09
GO:0050679	positive regulation of epithelial cell proliferation	8.96E-09
GO:0045597	positive regulation of cell differentiation	1.30E-08
GO:0001934	positive regulation of protein phosphorylation	1.64E-08
GO:0009880	embryonic pattern specification	1.91E-08
GO:0030902	hindbrain development	1.91E-08
GO:0030900	forebrain development	2.14E-08
GO:0060070	canonical Wnt signaling pathway	2.72E-08
GO:0001944	vasculature development	2.94E-08
GO:0001889	liver development	4.30E-08
GO:0032355	response to estradiol	6.58E-08
GO:0008283	cell proliferation	7.65E-08
GO:0042127	regulation of cell proliferation	1.00E-07
GO:0010468	regulation of gene expression	1.19E-07
GO:0010718	positive regulation of epithelial to mesenchymal transition	1.22E-07
GO:0090263	positive regulation of canonical Wnt signaling pathway	1.30E-07
GO:0071773	cellular response to BMP stimulus	1.42E-07
GO:0034504	protein localization to nucleus	1.42E-07
GO:0007224	smoothened signaling pathway	1.85E-07
GO:0042472	inner ear morphogenesis	2.01E-07
GO:0001656	metanephros development	2.17E-07
GO:0045595	regulation of cell differentiation	2.49E-07



GO:0007179	transforming growth factor beta receptor signaling pathway	2.57E-07
GO:0043065	positive regulation of apoptotic process	2.60E-07
GO:0001569	patterning of blood vessels	2.83E-07
GO:0042493	response to drug	2.90E-07
GO:0043010	camera-type eye development	3.26E-07
GO:0048856	anatomical structure development	4.21E-07
GO:0010463	mesenchymal cell proliferation	4.21E-07
GO:0030154	cell differentiation	4.42E-07
GO:0001502	cartilage condensation	6.63E-07
GO:0048557	embryonic digestive tract morphogenesis	8.16E-07
GO:0006029	proteoglycan metabolic process	8.39E-07
GO:0007411	axon guidance	9.39E-07
GO:1902895	positive regulation of pri-miRNA transcription from RNA polymerase II promoter	9.95E-07
GO:0001503	ossification	1.12E-06
GO:0048646	anatomical structure formation involved in morphogenesis	1.44E-06
GO:0090090	negative regulation of canonical Wnt signaling pathway	1.60E-06
GO:0007267	cell-cell signaling	1.69E-06
GO:0048593	camera-type eye morphogenesis	1.71E-06
GO:0042542	response to hydrogen peroxide	1.88E-06
GO:0002052	positive regulation of neuroblast proliferation	2.01E-06
GO:0042487	regulation of odontogenesis of dentin-containing tooth	2.34E-06
GO:0060485	mesenchyme development	2.34E-06
GO:0042476	odontogenesis	2.36E-06
GO:0043392	negative regulation of DNA binding	3.17E-06
GO:0021522	spinal cord motor neuron differentiation	3.17E-06
GO:0048738	cardiac muscle tissue development	3.17E-06
GO:0035264	multicellular organism growth	4.13E-06
GO:0001837	epithelial to mesenchymal transition	4.18E-06

Supplementary Table S2.21. The top 100 enriched Biological Process terms from the Gene Ontology for the hindlimb module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term identifier	Term name	P-value
GO:0001501	skeletal system development	2.02E-34
GO:0001503	ossification	2.58E-28
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	4.76E-27
GO:0030326	embryonic limb morphogenesis	1.90E-26
GO:0045893	positive regulation of transcription, DNA-templated	3.11E-25
GO:0051216	cartilage development	4.68E-19
GO:0042733	embryonic digit morphogenesis	5.94E-19
GO:0008284	positive regulation of cell proliferation	9.08E-19
GO:0060021	palate development	1.79E-17
GO:0007275	multicellular organism development	3.26E-17
GO:0060349	bone morphogenesis	5.17E-13
GO:0009887	organ morphogenesis	7.17E-13
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	1.65E-12
GO:0035115	embryonic forelimb morphogenesis	3.13E-12
GO:0060173	limb development	4.71E-12
GO:0002062	chondrocyte differentiation	6.66E-12
GO:0008285	negative regulation of cell proliferation	9.13E-12
GO:0045669	positive regulation of osteoblast differentiation	1.62E-11
GO:0050680	negative regulation of epithelial cell proliferation	4.13E-11
GO:0009952	anterior/posterior pattern specification	9.27E-11
GO:0001958	endochondral ossification	1.25E-10
GO:0060348	bone development	1.27E-10
GO:0002063	chondrocyte development	2.14E-10
GO:0016055	Wnt signaling pathway	3.94E-10
GO:0009954	proximal/distal pattern formation	5.08E-10
GO:0001649	osteoblast differentiation	8.43E-10
GO:0048705	skeletal system morphogenesis	9.75E-10
GO:0007568	aging	9.77E-10
GO:0050679	positive regulation of epithelial cell proliferation	9.98E-10
GO:0030500	regulation of bone mineralization	1.20E-09
GO:0007507	heart development	1.34E-09
GO:0043066	negative regulation of apoptotic process	1.38E-09
GO:0002053	positive regulation of mesenchymal cell proliferation	1.65E-09

GO:0051091	positive regulation of sequence-specific DNA binding transcription factor activity	2.59E-09
GO:0048701	embryonic cranial skeleton morphogenesis	3.09E-09
GO:0009953	dorsal/ventral pattern formation	3.51E-09
GO:0030509	BMP signaling pathway	7.10E-09
GO:0046716	muscle cell cellular homeostasis	7.30E-09
GO:0035116	embryonic hindlimb morphogenesis	7.51E-09
GO:0090090	negative regulation of canonical Wnt signaling pathway	9.24E-09
GO:0043410	positive regulation of MAPK cascade	1.23E-08
GO:0030282	bone mineralization	1.42E-08
GO:0030178	negative regulation of Wnt signaling pathway	1.80E-08
GO:0035108	limb morphogenesis	2.40E-08
GO:0001843	neural tube closure	3.28E-08
GO:0031214	biomineral tissue development	3.40E-08
GO:0032332	positive regulation of chondrocyte differentiation	4.87E-08
GO:0045778	positive regulation of ossification	7.29E-08
GO:0010628	positive regulation of gene expression	7.55E-08
GO:0042493	response to drug	9.27E-08
GO:0001974	blood vessel remodeling	1.23E-07
GO:0003151	outflow tract morphogenesis	1.24E-07
GO:0090263	positive regulation of canonical Wnt signaling pathway	1.49E-07
GO:0048704	embryonic skeletal system morphogenesis	1.51E-07
GO:0036342	post-anal tail morphogenesis	1.54E-07
GO:0043065	positive regulation of apoptotic process	2.59E-07
GO:0060065	uterus development	3.12E-07
GO:0001701	in utero embryonic development	3.64E-07
GO:0001934	positive regulation of protein phosphorylation	3.92E-07
GO:0040014	regulation of multicellular organism growth	4.49E-07
GO:0001502	cartilage condensation	4.69E-07
GO:0060070	canonical Wnt signaling pathway	5.37E-07
GO:0030199	collagen fibril organization	5.71E-07
GO:0007517	muscle organ development	6.44E-07
GO:0040018	positive regulation of multicellular organism growth	9.00E-07
GO:0030217	T cell differentiation	9.00E-07
GO:0007628	adult walking behavior	1.12E-06
GO:0035136	forelimb morphogenesis	1.29E-06
GO:0008584	male gonad development	1.37E-06
GO:0006351	transcription, DNA-templated	1.43E-06
GO:0030198	extracellular matrix organization	1.70E-06

GO:0001568	blood vessel development	2.22E-06
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	2.34E-06
GO:0007389	pattern specification process	2.35E-06
GO:0006355	regulation of transcription, DNA-templated	2.38E-06
GO:0035019	somatic stem cell population maintenance	2.52E-06
GO:0045453	bone resorption	2.60E-06
GO:0016477	cell migration	2.98E-06
GO:0035137	hindlimb morphogenesis	3.04E-06
GO:0001666	response to hypoxia	3.21E-06
GO:0007417	central nervous system development	3.49E-06
GO:0030154	cell differentiation	3.69E-06
GO:0030501	positive regulation of bone mineralization	4.16E-06
GO:0007409	axonogenesis	4.51E-06
GO:0035987	endodermal cell differentiation	4.58E-06
GO:0042981	regulation of apoptotic process	5.33E-06
GO:0021915	neural tube development	6.20E-06
GO:0042060	wound healing	6.24E-06
GO:0048747	muscle fiber development	6.38E-06
GO:0014033	neural crest cell differentiation	6.71E-06
GO:0030335	positive regulation of cell migration	7.03E-06
GO:0010629	negative regulation of gene expression	7.25E-06
GO:0001942	hair follicle development	7.32E-06
GO:0030316	osteoclast differentiation	7.70E-06
GO:0060325	face morphogenesis	7.73E-06
GO:0042127	regulation of cell proliferation	9.04E-06
GO:0070374	positive regulation of ERK1 and ERK2 cascade	9.72E-06
GO:0045880	positive regulation of smoothed signaling pathway	9.82E-06
GO:0030279	negative regulation of ossification	1.22E-05
GO:0001756	somitogenesis	1.37E-05

Supplementary Table S2.22. The enriched Uberon terms for the pelvic fin module-specific genes.

Uberon identifier	Term name	P-value
uberon 0000151	pectoral fin	7.11E-07
uberon 4000163	anal fin	3.82E-06
uberon 2000040	median fin fold	4.79E-05
uberon 0005281	ventricular system of central nervous system	9.46E-05
uberon 0001017	central nervous system	0.00019108
uberon 2000694	ceratobranchial 5 tooth	0.0003055
uberon 0000152	pelvic fin	0.00031273
uberon 0002107	liver	0.00032183
uberon 2000106	extension	0.00067516
uberon 0003097	dorsal fin	0.00130209
uberon 0001555	digestive tract	0.00130304
uberon 0007812	post-anal tail	0.00164604
uberon 2001239	ceratobranchial 5 bone	0.00186514
uberon 4000172	lepidotrichium	0.00200288
uberon 2000084	yolk	0.00501523
uberon 0012438	blastema of regenerating fin/limb	0.00586938
uberon 2001280	branchiostegal ray 3	0.00590243
uberon 0000982	skeletal joint	0.00913717
uberon 0002120	pronephros	0.0119861
uberon 0015178	somite border	0.01298747
uberon 0001703	neurocranium	0.01579709
uberon 0000178	blood	0.01709498
uberon 0002513	endochondral bone	0.01987709
uberon 0002457	intersomitic artery	0.0218951
uberon 0007215	trabecula cranii	0.02246904
uberon 0000165	mouth	0.02399809
uberon 4000174	caudal fin lepidotrichium	0.02496839
uberon 0002422	fourth ventricle	0.02857919
uberon 0005362	vagus x ganglion	0.03250281
uberon 0000959	optic chiasma	0.03250281
uberon 0000468	multicellular organism	0.03847851
uberon 0002280	otolith	0.0390762
uberon 0003053	ventricular zone	0.0415756
uberon 2000411	posterior crista primordium	0.0415756

uberon 2000356	gill raker	0.04385078
uberon 0005419	pectoral appendage bud	0.04400473
uberon 0011944	subintestinal vein	0.04508881

Supplementary Table S2.23. The top 100 enriched Uberon terms for the pelvic fin conserved genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0003079	floor plate	1.08E-08
uberon 0000151	pectoral fin	8.38E-06
uberon 0007812	post-anal tail	1.15E-05
uberon 2000040	median fin fold	4.10E-05
uberon 2001089	myoseptum	8.16E-05
uberon 0001976	epithelium of esophagus	0.00010711
uberon 2001069	ventral fin fold	0.00013312
uberon 0001043	esophagus	0.00016953
uberon 2000250	opercle	0.00044826
uberon 0003097	dorsal fin	0.00053446
uberon 2000033	intermediate cell mass of mesoderm	0.00062034
uberon 2002200	hypobranchial muscle	0.00063788
uberon 0003901	horizontal septum	0.00071559
uberon 0003077	paraxial mesoderm	0.00082036
uberon 2001456	pectoral fin endoskeletal disc	0.00082036
uberon 4000172	lepidotrichium	0.00082036
uberon 0000160	intestine	0.00086843
uberon 0001264	pancreas	0.00110702
uberon 0003082	myotome	0.00113857
uberon 4000163	anal fin	0.00135149
uberon 0003069	otic placode	0.00169069
uberon 0002328	notochord	0.00234575
uberon 0005419	pectoral appendage bud	0.00303292
uberon 0002348	epicardium	0.00316901
uberon 0001135	smooth muscle tissue	0.00316901
uberon 0004291	heart rudiment	0.00350066
uberon 0002533	post-anal tail bud	0.00429825
uberon 0004752	palatoquadrate cartilage	0.00433197
uberon 0007269	pectoral appendage musculature	0.00439376
uberon 4000175	pectoral fin lepidotrichium	0.00439376
uberon 0002329	somite	0.0047405
uberon 0008896	post-hyoid pharyngeal arch	0.0049714
uberon 0003107	meckel's cartilage	0.0056048

uberont 0002394	bile duct	0.00580184
uberont 0007831	pectoral girdle skeleton	0.00580184
uberont 2001516	ceratobranchial cartilage	0.00595325
uberont 0003091	thyroid primordium	0.00738763
uberont 2000694	ceratobranchial 5 tooth	0.0088291
uberont 0006860	swim bladder	0.00883525
uberont 0002633	motor nucleus of trigeminal nerve	0.00914566
uberont 0007274	crista of ampulla of anterior semicircular duct of membranous labyrinth	0.00914566
uberont 2001256	lateral floor plate	0.00914566
uberont 0000931	proctodeum	0.00914566
uberont 0010170	region of neural crest	0.01107057
uberont 0002342	neural crest	0.01107057
uberont 4000174	caudal fin lepidotrichium	0.01107057
uberont 0007329	pancreatic duct	0.01107057
uberont 0011607	hyomandibular cartilage	0.01188946
uberont 4000164	caudal fin	0.01244072
uberont 0007298	pronephric proximal convoluted tubule	0.01263336
uberont 0000959	optic chiasma	0.0131571
uberont 2000558	posterior macula	0.01540015
uberont 0005507	rhombomere 3	0.01540015
uberont 0000965	lens of camera-type eye	0.01755751
uberont 0002196	adenohypophysis	0.01779469
uberont 0001049	neural tube	0.0181603
uberont 0014907	intersomitic vessel	0.01866956
uberont 0005805	dorsal aorta	0.01999588
uberont 0000152	pelvic fin	0.02033582
uberont 2000676	sagitta	0.02033582
uberont 0011944	subintestinal vein	0.02123921
uberont 0004117	pharyngeal pouch	0.02583882
uberont 0005515	rhombomere 5	0.02583882
uberont 0002241	chondrocranium	0.02879142
uberont 0001794	inner limiting layer of retina	0.02960362
uberont 0005499	rhombomere 1	0.02960362
uberont 2000544	pectoral fin actinotrichium	0.02960362
uberont 2000623	basipterygium bone	0.02960362
uberont 0003412	pelvic appendage bud mesenchyme	0.02960362
uberont 0006964	pars distalis of adenohypophysis	0.02960362
uberont 0007097	chordo neural hinge	0.02960362
uberont 2000674	interopercle	0.02960362



uberon 2007046	midbrain hindbrain boundary neural rod	0.02960362
uberon 2005346	extrapancreatic duct	0.02960362
uberon 2000284	subopercle	0.02960362
uberon 2001300	vagal placode 4	0.02960362
uberon 2001297	vagal placode 1	0.02960362
uberon 2001298	vagal placode 2	0.02960362
uberon 2001299	vagal placode 3	0.02960362
uberon 0002228	rib	0.02960362
uberon 2001277	anterior chamber swim bladder	0.02960362
uberon 0001638	vein	0.02960362
uberon 2007032	midbrain neural rod	0.02960362
uberon 0005422	pelvic appendage apical ectodermal ridge	0.02960362
uberon 0006859	swim bladder bud	0.02960362
uberon 0002514	intramembranous bone	0.02960362
uberon 0000948	heart	0.0313214
uberon 0004741	cleithrum	0.03507643
uberon 0003011	facial motor nucleus	0.0384002
uberon 0000926	mesoderm	0.0418392
uberon 0000925	endoderm	0.0418392
uberon 2001357	alar plate midbrain	0.04407914
uberon 0007124	pharyngeal pouch 3	0.04407914
uberon 0007125	pharyngeal pouch 4	0.04407914
uberon 0001093	vertebral bone 2	0.04407914
uberon 0002424	oral epithelium	0.04407914
uberon 2001430	pneumatic duct	0.04407914
uberon 0004376	fin bone	0.04407914
uberon 2005316	fin fold pectoral fin bud	0.04407914
uberon 0005283	tela choroidea	0.04407914

Supplementary Table S2.24. The top 100 enriched Uberon terms for the hindlimb conserved genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0002091	appendicular skeleton	3.19E-33
uberon 0002103	hindlimb	6.19E-33
uberon 0001708	jaw skeleton	2.38E-30
uberon 0011156	facial skeleton	1.01E-28
uberon 0005944	axial skeleton plus cranial skeleton	1.29E-27
uberon 0001456	face	1.66E-27
uberon 0000033	head	2.63E-27
uberon 0000165	mouth	1.31E-26
uberon 0003128	cranium	2.05E-26
uberon 0002102	forelimb	1.75E-24
uberon 0007811	craniocervical region	4.04E-24
uberon 0004716	conceptus	2.03E-23
uberon 0001434	skeletal system	2.95E-23
uberon 0001684	mandible	3.01E-21
uberon 0002470	autopod region	1.67E-20
uberon 0000383	musculature of body	1.57E-19
uberon 0001004	respiratory system	2.75E-19
uberon 0003823	hindlimb zeugopod	3.13E-19
uberon 0001049	neural tube	3.51E-19
uberon 0002544	digit	5.64E-19
uberon 0001716	secondary palate	6.56E-19
uberon 0001007	digestive system	1.14E-18
uberon 0001703	neurocranium	1.22E-18
uberon 0002397	maxilla	1.33E-18
uberon 0001690	ear	1.39E-17
uberon 0002105	vestibulo-auditory system	2.82E-17
uberon 0003252	thoracic rib cage	4.13E-17
uberon 0003216	hard palate	5.32E-17
uberon 0000014	zone of skin	6.27E-17
uberon 0001890	forebrain	9.98E-17
uberon 0002104	visual system	1.76E-16
uberon 0000955	brain	2.30E-16
uberon 0000948	heart	7.45E-16

uberon 0000019	camera-type eye	7.87E-16
uberon 0001008	renal system	8.45E-16
uberon 0004535	cardiovascular system	1.19E-15
uberon 0005619	secondary palatal shelf	1.80E-15
uberon 0003822	forelimb stylopod	3.01E-15
uberon 0001446	fibula	6.03E-15
uberon 0002516	epiphyseal plate	6.20E-15
uberon 0002416	integumental system	1.01E-14
uberon 0001285	nephron	1.07E-14
uberon 0001894	diencephalon	1.14E-14
uberon 0001229	renal corpuscle	1.15E-14
uberon 0000947	aorta	1.21E-14
uberon 0000975	sternum	2.24E-14
uberon 0000012	somatic nervous system	2.35E-14
uberon 0000004	nose	2.76E-14
uberon 0002012	pulmonary artery	3.00E-14
uberon 0000074	renal glomerulus	3.92E-14
uberon 0006772	long bone epiphyseal plate hypertrophic zone	4.88E-14
uberon 0001756	middle ear	4.90E-14
uberon 0000474	female reproductive system	5.73E-14
uberon 0001225	cortex of kidney	7.17E-14
uberon 0001681	nasal bone	9.46E-14
uberon 0000979	tibia	1.01E-13
uberon 0002483	trabecular bone tissue	1.01E-13
uberon 0000084	ureteric bud	1.50E-13
uberon 0000209	tetrapod frontal bone	2.85E-13
uberon 0001723	tongue	3.11E-13
uberon 0001677	sphenoid bone	4.49E-13
uberon 0006333	snout	1.19E-12
uberon 0000976	humerus	1.30E-12
uberon 0002228	rib	1.43E-12
uberon 0003655	molar tooth	1.96E-12
uberon 0002094	interventricular septum	1.99E-12
uberon 0004347	limb bud	2.89E-12
uberon 0002218	tympanic ring	3.02E-12
uberon 0003451	lower jaw incisor	3.02E-12
uberon 0002028	hindbrain	5.27E-12
uberon 0000924	ectoderm	6.31E-12
uberon 0000376	hindlimb stylopod	7.85E-12

uberon 0002517	basicranium	9.17E-12
uberon 0001130	vertebral column	9.40E-12
uberon 0002298	brainstem	9.72E-12
uberon 0001037	strand of hair	1.08E-11
uberon 0006849	scapula	1.84E-11
uberon 0003107	meckel's cartilage	1.84E-11
uberon 0010166	coat of hair	1.96E-11
uberon 0006207	aortico-pulmonary spiral septum	2.15E-11
uberon 0002080	heart right ventricle	2.45E-11
uberon 0001711	eyelid	2.98E-11
uberon 0004356	apical ectodermal ridge	2.98E-11
uberon 0003975	internal female genitalia	4.59E-11
uberon 0000922	embryo	6.10E-11
uberon 0000981	femur	7.75E-11
uberon 0002048	lung	7.78E-11
uberon 0001230	glomerular capsule	9.73E-11
uberon 0000059	large intestine	1.10E-10
uberon 0002084	heart left ventricle	1.13E-10
uberon 0001891	midbrain	2.01E-10
uberon 0005871	palatine process of maxilla	3.50E-10
uberon 0002407	pericardium	3.63E-10
uberon 0002418	cartilage tissue	4.47E-10
uberon 0001911	mammary gland	4.82E-10
uberon 0002113	kidney	4.82E-10
uberon 0001895	metencephalon	5.03E-10
uberon 0003221	phalanx	5.23E-10
uberon 0001439	compact bone tissue	5.32E-10
uberon 0008974	apocrine gland	5.42E-10

Supplementary Table S2.25. The top 100 enriched Uberon terms for the hindlimb module-specific genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0002103	hindlimb	2.05e-340
uberon 0002091	appendicular skeleton	3.0587604134e-320
uberon 0001434	skeletal system	1.62E-222
uberon 0003823	hindlimb zeugopod	6.11E-203
uberon 0000979	tibia	2.71E-190
uberon 0000376	hindlimb stylopod	8.00E-187
uberon 0000981	femur	4.30E-183
uberon 0002102	forelimb	4.60E-85
uberon 0005944	axial skeleton plus cranial skeleton	2.27E-80
uberon 0002483	trabecular bone tissue	2.04E-70
uberon 0002516	epiphyseal plate	1.09E-65
uberon 0003663	hindlimb muscle	3.02E-58
uberon 0008777	hypaxial musculature	6.71E-55
uberon 0001424	ulna	1.44E-53
uberon 0003128	cranium	1.28E-51
uberon 0003822	forelimb stylopod	4.18E-51
uberon 0007811	craniocervical region	4.56E-51
uberon 0000976	humerus	1.21E-50
uberon 0006772	long bone epiphyseal plate hypertrophic zone	2.64E-50
uberon 0001134	skeletal muscle tissue	4.82E-50
uberon 0000383	musculature of body	1.98E-49
uberon 0001423	radius bone	3.48E-48
uberon 0011156	facial skeleton	3.35E-46
uberon 0001439	compact bone tissue	6.40E-45
uberon 0002470	autopod region	5.17E-44
uberon 0001708	jaw skeleton	2.14E-43
uberon 0003252	thoracic rib cage	3.83E-43
uberon 0001446	fibula	2.16E-41
uberon 0001703	neurocranium	2.57E-41
uberon 0000165	mouth	5.13E-39
uberon 0002544	digit	1.45E-38
uberon 0001456	face	1.04E-37
uberon 0001684	mandible	1.32E-36

uberon 0000033	head	2.97E-36
uberon 0008867	trabecular network of bone	5.49E-36
uberon 0003221	phalanx	1.44E-35
uberon 0001130	vertebral column	6.99E-34
uberon 0002228	rib	9.84E-34
uberon 0001448	metatarsal bone	2.49E-30
uberon 0006771	long bone epiphyseal plate proliferative zone	9.20E-27
uberon 0000440	trabecula	1.06E-26
uberon 0001004	respiratory system	5.29E-26
uberon 0000475	organism subdivision	1.45E-25
uberon 0001389	soleus muscle	1.04E-24
uberon 0007812	post-anal tail	3.07E-24
uberon 0001007	digestive system	6.78E-24
uberon 0001388	gastrocnemius	9.31E-24
uberon 0002374	metacarpal bone	7.00E-23
uberon 0004535	cardiovascular system	3.17E-22
uberon 0002105	vestibulo-auditory system	6.14E-22
uberon 0001690	ear	8.89E-22
uberon 0000012	somatic nervous system	7.20E-21
uberon 0001780	spinal nerve	9.62E-21
uberon 0006333	snout	1.34E-20
uberon 0002397	maxilla	1.58E-20
uberon 0001095	caudal vertebra	1.64E-20
uberon 0002104	visual system	3.47E-19
uberon 0001385	tibialis anterior	1.62E-18
uberon 0002240	spinal cord	4.15E-18
uberon 0000468	multicellular organism	8.18E-18
uberon 0000019	camera-type eye	9.69E-18
uberon 0007830	pelvic girdle bone/zone	1.51E-17
uberon 0001756	middle ear	1.76E-17
uberon 0002412	vertebra	2.17E-17
uberon 0000975	sternum	2.99E-17
uberon 0001716	secondary palate	4.81E-17
uberon 0002048	lung	2.37E-16
uberon 0001438	metaphysis	2.93E-16
uberon 0001377	quadriceps femoris	7.63E-16
uberon 0002517	basicranium	1.87E-15
uberon 0001435	carpal bone	2.67E-15
uberon 0001008	renal system	2.77E-15

uberon 0001386	extensor digitorum longus	3.03E-15
uberon 0006849	scapula	1.06E-14
uberon 0001711	eyelid	1.08E-14
uberon 0000210	tetrapod parietal bone	1.21E-14
uberon 0001447	tarsal bone	1.69E-14
uberon 0000474	female reproductive system	2.49E-14
uberon 0004716	conceptus	1.41E-13
uberon 0002413	cervical vertebra	1.47E-13
uberon 0001049	neural tube	2.35E-13
uberon 0000014	zone of skin	2.41E-13
uberon 0000948	heart	4.92E-13
uberon 0001689	malleus bone	6.36E-13
uberon 0002347	thoracic vertebra	7.26E-13
uberon 0001723	tongue	7.97E-13
uberon 0001678	temporal bone	8.99E-13
uberon 0006861	diaphysis proper	1.22E-12
uberon 0001681	nasal bone	2.13E-12
uberon 0004347	limb bud	2.63E-12
uberon 0003861	neural arch	3.02E-12
uberon 0002446	patella	3.28E-12
uberon 0003461	shoulder bone	4.16E-12
uberon 0001013	adipose tissue	4.52E-12
uberon 0006068	bone of tail	4.91E-12
uberon 0002315	gray matter of spinal cord	7.28E-12
uberon 0001676	occipital bone	8.95E-12
uberon 0000982	skeletal joint	1.00E-11
uberon 0002416	integumental system	1.10E-11
uberon 0002418	cartilage tissue	1.24E-11

Supplementary Table S2.26. The top 100 enriched Biological Process terms from the Gene Ontology for the mouse orthologs of the pelvic fin module-specific genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Term identifier	Term name	P-value
GO:0006468	protein phosphorylation	5.19E-34
GO:0016310	phosphorylation	1.24E-30
GO:0046777	protein autophosphorylation	9.64E-22
GO:0018105	peptidyl-serine phosphorylation	1.63E-15
GO:0045893	positive regulation of transcription, DNA-templated	5.38E-15
GO:0018108	peptidyl-tyrosine phosphorylation	1.80E-14
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	3.24E-14
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	1.41E-13
GO:0010628	positive regulation of gene expression	1.67E-13
GO:0009887	organ morphogenesis	7.65E-12
GO:0030335	positive regulation of cell migration	1.14E-11
GO:0042493	response to drug	7.06E-11
GO:0018107	peptidyl-threonine phosphorylation	1.51E-10
GO:0043066	negative regulation of apoptotic process	4.82E-10
GO:0043627	response to estrogen	5.96E-10
GO:0001934	positive regulation of protein phosphorylation	6.88E-10
GO:0006351	transcription, DNA-templated	1.09E-09
GO:0038083	peptidyl-tyrosine autophosphorylation	1.75E-09
GO:0048565	digestive tract development	2.26E-09
GO:0035690	cellular response to drug	7.15E-09
GO:0008284	positive regulation of cell proliferation	9.34E-09
GO:0035556	intracellular signal transduction	1.19E-08
GO:0030154	cell differentiation	1.47E-08
GO:0035970	peptidyl-threonine dephosphorylation	2.43E-08
GO:0007275	multicellular organism development	4.69E-08
GO:0045665	negative regulation of neuron differentiation	8.95E-08
GO:0071773	cellular response to BMP stimulus	1.34E-07
GO:0007264	small GTPase mediated signal transduction	2.17E-07
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	2.52E-07
GO:0030182	neuron differentiation	3.94E-07
GO:0060070	canonical Wnt signaling pathway	4.66E-07
GO:0071407	cellular response to organic cyclic compound	7.62E-07
GO:0016477	cell migration	1.56E-06



GO:0006355	regulation of transcription, DNA-templated	1.98E-06
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	2.25E-06
GO:0001942	hair follicle development	3.29E-06
GO:0043525	positive regulation of neuron apoptotic process	3.31E-06
GO:0007049	cell cycle	4.64E-06
GO:0043524	negative regulation of neuron apoptotic process	5.40E-06
GO:0007179	transforming growth factor beta receptor signaling pathway	6.68E-06
GO:0006357	regulation of transcription from RNA polymerase II promoter	8.40E-06
GO:0048485	sympathetic nervous system development	8.61E-06
GO:0045740	positive regulation of DNA replication	9.30E-06
GO:0071363	cellular response to growth factor stimulus	1.61E-05
GO:0007507	heart development	1.70E-05
GO:0048663	neuron fate commitment	1.77E-05
GO:0030509	BMP signaling pathway	2.27E-05
GO:0001701	in utero embryonic development	2.45E-05
GO:0008584	male gonad development	2.93E-05
GO:0006470	protein dephosphorylation	3.41E-05
GO:0071333	cellular response to glucose stimulus	3.48E-05
GO:0030513	positive regulation of BMP signaling pathway	3.59E-05
GO:0031175	neuron projection development	3.65E-05
GO:0045165	cell fate commitment	3.86E-05
GO:0003151	outflow tract morphogenesis	4.33E-05
GO:0043406	positive regulation of MAP kinase activity	4.33E-05
GO:0042127	regulation of cell proliferation	5.34E-05
GO:0010629	negative regulation of gene expression	7.71E-05
GO:0016569	covalent chromatin modification	8.05E-05
GO:0051496	positive regulation of stress fiber assembly	1.15E-04
GO:0060041	retina development in camera-type eye	1.17E-04
GO:0010468	regulation of gene expression	1.23E-04
GO:0003007	heart morphogenesis	1.31E-04
GO:0045429	positive regulation of nitric oxide biosynthetic process	1.31E-04
GO:0007399	nervous system development	1.38E-04
GO:0045597	positive regulation of cell differentiation	1.49E-04
GO:0006366	transcription from RNA polymerase II promoter	1.71E-04
GO:0008285	negative regulation of cell proliferation	1.73E-04
GO:0007165	signal transduction	2.05E-04
GO:0045931	positive regulation of mitotic cell cycle	2.26E-04
GO:0032212	positive regulation of telomere maintenance via telomerase	2.26E-04
GO:0045471	response to ethanol	2.30E-04

GO:0007224	smoothened signaling pathway	2.52E-04
GO:0007167	enzyme linked receptor protein signaling pathway	2.85E-04
GO:0007204	positive regulation of cytosolic calcium ion concentration	3.02E-04
GO:0048709	oligodendrocyte differentiation	3.03E-04
GO:0001525	angiogenesis	3.37E-04
GO:0030324	lung development	3.55E-04
GO:0001764	neuron migration	3.55E-04
GO:0048015	phosphatidylinositol-mediated signaling	3.99E-04
GO:0048538	thymus development	4.03E-04
GO:0016055	Wnt signaling pathway	4.15E-04
GO:0090090	negative regulation of canonical Wnt signaling pathway	4.48E-04
GO:0010001	glial cell differentiation	4.49E-04
GO:0071260	cellular response to mechanical stimulus	4.50E-04
GO:0030501	positive regulation of bone mineralization	4.54E-04
GO:0009790	embryo development	4.87E-04
GO:0070301	cellular response to hydrogen peroxide	4.90E-04
GO:0045909	positive regulation of vasodilation	5.16E-04
GO:0060412	ventricular septum morphogenesis	5.16E-04
GO:0070374	positive regulation of ERK1 and ERK2 cascade	5.21E-04
GO:0048661	positive regulation of smooth muscle cell proliferation	5.26E-04
GO:0008283	cell proliferation	5.53E-04
GO:0051091	positive regulation of sequence-specific DNA binding transcription factor activity	6.17E-04
GO:0001666	response to hypoxia	6.21E-04
GO:0060045	positive regulation of cardiac muscle cell proliferation	6.46E-04
GO:0007623	circadian rhythm	6.56E-04
GO:0048646	anatomical structure formation involved in morphogenesis	7.64E-04
GO:0048854	brain morphogenesis	7.64E-04
GO:0000082	G1/S transition of mitotic cell cycle	7.71E-04

Supplementary Table S2.27. The top 100 Uberon terms for the mouse orthologs of the pelvic fin module-specific genes. The full enriched term list is available at

[https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon identifier	Term name	P-value
uberon 0000922	embryo	6.50E-16
uberon 0004716	conceptus	1.53E-13
uberon 0004365	vitelline blood vessel	2.69E-09
uberon 0000478	extraembryonic structure	3.51E-09
uberon 0001007	digestive system	8.53E-09
uberon 0000014	zone of skin	2.47E-08
uberon 0010190	pair of dorsal aortae	4.77E-08
uberon 0001049	neural tube	1.02E-07
uberon 0008852	visceral yolk sac	1.23E-07
uberon 0000948	heart	1.60E-07
uberon 0004535	cardiovascular system	4.34E-07
uberon 0000468	multicellular organism	5.93E-07
uberon 0000016	endocrine pancreas	8.87E-07
uberon 0002067	dermis	1.91E-06
uberon 0001004	respiratory system	3.10E-06
uberon 0000358	blastocyst	3.19E-06
uberon 0000033	head	3.57E-06
uberon 0002012	pulmonary artery	4.49E-06
uberon 0002407	pericardium	1.34E-05
uberon 0002048	lung	1.66E-05
uberon 0004374	vitelline vasculature	2.27E-05
uberon 0001987	placenta	2.69E-05
uberon 0003087	anterior cardinal vein	3.00E-05
uberon 0000006	islet of langerhans	3.49E-05
uberon 0001264	pancreas	4.05E-05
uberon 0000947	aorta	4.50E-05
uberon 0001456	face	6.01E-05
uberon 0003512	lung blood vessel	6.31E-05
uberon 0002240	spinal cord	7.35E-05
uberon 0003946	placenta labyrinth	0.00010402
uberon 0002368	endocrine gland	0.00011026
uberon 0001792	ganglionic layer of retina	0.00012935
uberon 0002062	endocardial cushion	0.00014493

uberon 0001508	arch of aorta	0.00015353
uberon 0001083	myocardium of ventricle	0.0001601
uberon 0007811	cranio cervical region	0.00016979
uberon 0001809	enteric ganglion	0.0001807
uberon 0000084	ureteric bud	0.00020974
uberon 0002087	atrioventricular canal	0.00022946
uberon 0002315	gray matter of spinal cord	0.00024038
uberon 0004647	liver lobule	0.00026764
uberon 0001818	tarsal gland	0.00026935
uberon 0000165	mouth	0.00027721
uberon 0001675	trigeminal ganglion	0.00027999
uberon 0006207	aortico-pulmonary spiral septum	0.00028524
uberon 0002416	integumental system	0.00028781
uberon 0010172	bulb of aorta	0.00030071
uberon 0002005	enteric nervous system	0.00030515
uberon 0006524	alveolar system	0.00031454
uberon 0008870	pulmonary alveolar parenchyma	0.00034866
uberon 0000088	trophoblast	0.00039593
uberon 0002094	interventricular septum	0.000402
uberon 0000383	musculature of body	0.00042772
uberon 0001711	eyelid	0.00043278
uberon 0003216	hard palate	0.00044958
uberon 0001280	liver parenchyma	0.00046467
uberon 0000011	parasympathetic nervous system	0.0004649
uberon 0002073	hair follicle	0.00052472
uberon 0002370	thymus	0.00057905
uberon 0008856	stomach muscularis externa	0.0006099
uberon 0000019	camera-type eye	0.00061682
uberon 0002408	parietal serous pericardium	0.00062547
uberon 0001496	ascending aorta	0.00063348
uberon 0000117	respiratory tube	0.00063848
uberon 0002165	endocardium	0.00073687
uberon 0003066	pharyngeal arch 2	0.00073687
uberon 0002511	trabecula carnea	0.00077776
uberon 0000013	sympathetic nervous system	0.0008281
uberon 0000012	somatic nervous system	0.00090485
uberon 0002342	neural crest	0.00093289
uberon 0003618	aorta tunica media	0.00093289
uberon 0002080	heart right ventricle	0.00100621

uberon 0001637	artery	0.0010228
uberon 0001708	jaw skeleton	0.00113097
uberon 0002104	visual system	0.00120939
uberon 0002069	stratum granulosum of epidermis	0.00121228
uberon 0001074	pericardial cavity	0.00142689
uberon 0001041	foregut	0.00145277
uberon 0008874	pulmonary acinus	0.00148651
uberon 0004493	cardiac muscle tissue of myocardium	0.00151061
uberon 0002410	autonomic nervous system	0.00154176
uberon 0001135	smooth muscle tissue	0.00168825
uberon 0001716	secondary palate	0.00180283
uberon 0003501	retina blood vessel	0.00195196
uberon 0002025	stratum basale of epidermis	0.00203946
uberon 0011156	facial skeleton	0.0021246
uberon 0003823	hindlimb zeugopod	0.00219172
uberon 0001534	right subclavian artery	0.00222747
uberon 0005970	brain commissure	0.00223824
uberon 0003073	lens placode	0.00223824
uberon 0003128	cranium	0.00240653
uberon 0001003	skin epidermis	0.0024443
uberon 0010513	strand of zigzag hair	0.00258297
uberon 0001213	intestinal villus	0.00269833
uberon 0001806	sympathetic ganglion	0.00286072
uberon 0005870	olfactory pit	0.00286072
uberon 0005343	cortical plate	0.00286898
uberon 0005062	neural fold	0.00286898
uberon 0004663	aorta wall	0.0028703
uberon 0010512	strand of guard hair	0.00311381

Supplementary Table S2.28. The enriched Biological Process terms from the Gene Ontology for the zebrafish orthologs of the hindlimb module-specific genes.

Term identifier	Term name	P-value
GO:0006355	regulation of transcription, DNA-templated	1.34E-09
GO:0007275	multicellular organism development	5.81E-09
GO:0040007	growth	7.61E-07
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	1.22E-06
GO:0060395	SMAD protein signal transduction	1.89E-06
GO:0043408	regulation of MAPK cascade	2.33E-06
GO:0048701	embryonic cranial skeleton morphogenesis	1.03E-05
GO:0048468	cell development	1.49E-05
GO:0051216	cartilage development	1.50E-05
GO:0030182	neuron differentiation	3.49E-05
GO:0016055	Wnt signaling pathway	5.61E-05
GO:0007417	central nervous system development	1.10E-04
GO:0060070	canonical Wnt signaling pathway	1.97E-04
GO:0006351	transcription, DNA-templated	2.70E-04
GO:0008543	fibroblast growth factor receptor signaling pathway	3.76E-04
GO:0001501	skeletal system development	5.14E-04
GO:0007155	cell adhesion	6.49E-04
GO:0048703	embryonic viscerocranium morphogenesis	9.92E-04
GO:0043401	steroid hormone mediated signaling pathway	0.00114721
GO:0030903	notochord development	0.00117063
GO:0005975	carbohydrate metabolic process	0.00129702
GO:0009612	response to mechanical stimulus	0.00135076
GO:0030510	regulation of BMP signaling pathway	0.00160837
GO:0030513	positive regulation of BMP signaling pathway	0.00204619
GO:0001503	ossification	0.00204619
GO:0045165	cell fate commitment	0.00207662
GO:0000165	MAPK cascade	0.00225021
GO:0002062	chondrocyte differentiation	0.00225021
GO:0007626	locomotory behavior	0.0023328
GO:0031101	fin regeneration	0.00239174
GO:0043010	camera-type eye development	0.0025896
GO:0007517	muscle organ development	0.00308006
GO:0009953	dorsal/ventral pattern formation	0.00369912
GO:0005978	glycogen biosynthetic process	0.0039571

GO:0030916	otic vesicle formation	0.0039571
GO:0035118	embryonic pectoral fin morphogenesis	0.00553237
GO:0007179	transforming growth factor beta receptor signaling pathway	0.00553237
GO:0030282	bone mineralization	0.00628086
GO:0060349	bone morphogenesis	0.00737634
GO:0030902	hindbrain development	0.00739625
GO:0070654	sensory epithelium regeneration	0.00769009
GO:0031099	regeneration	0.00927093
GO:0007519	skeletal muscle tissue development	0.0106517
GO:0043627	response to estrogen	0.01086062
GO:0061024	membrane organization	0.01086062
GO:0021587	cerebellum morphogenesis	0.01086062
GO:0043697	cell dedifferentiation	0.01086062
GO:0009948	anterior/posterior axis specification	0.01102739
GO:0042074	cell migration involved in gastrulation	0.01173397
GO:0045595	regulation of cell differentiation	0.01296272
GO:0048839	inner ear development	0.01440797
GO:0008045	motor neuron axon guidance	0.01440797
GO:0061300	cerebellum vasculature development	0.01492536
GO:0030198	extracellular matrix organization	0.01507944
GO:0060536	cartilage morphogenesis	0.01737942
GO:0060021	palate development	0.01953552
GO:0046716	muscle cell cellular homeostasis	0.01953552
GO:0060538	skeletal muscle organ development	0.01953552
GO:0048884	neuromast development	0.01986392
GO:0060037	pharyngeal system development	0.01986392
GO:0035567	non-canonical Wnt signaling pathway	0.02253359
GO:0042981	regulation of apoptotic process	0.02458938
GO:0072661	protein targeting to plasma membrane	0.02465756
GO:0006874	cellular calcium ion homeostasis	0.02703075
GO:0055113	epiboly involved in gastrulation with mouth forming second	0.02923793
GO:0016203	muscle attachment	0.0302594
GO:0030500	regulation of bone mineralization	0.0302594
GO:0048840	otolith development	0.03165263
GO:0007267	cell-cell signaling	0.03311103
GO:0001889	liver development	0.03497684
GO:0051482	positive regulation of cytosolic calcium ion concentration involved in phospholipase C-activating G-protein coupled signaling pathway	0.03505962
GO:0060027	convergent extension involved in gastrulation	0.03507928

GO:0001568	blood vessel development	0.03507928
GO:0055001	muscle cell development	0.03631032
GO:0030199	collagen fibril organization	0.03631032
GO:0060041	retina development in camera-type eye	0.03668877
GO:0006357	regulation of transcription from RNA polymerase II promoter	0.03680213
GO:0060059	embryonic retina morphogenesis in camera-type eye	0.03864788
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	0.0397757
GO:0006508	proteolysis	0.04168701
GO:0007631	feeding behavior	0.04278097



Supplementary Table S2.29. The enriched Uberon terms for the zebrafish orthologs of the hindlimb module-specific genes.

Uberon identifier	Term name	P-value
uberont0000019	camera-type eye	7.22E-18
uberont0000468	multicellular organism	5.69E-17
uberont0000033	head	2.90E-15
uberont0008823	neural tube derived brain	1.67E-12
uberont0001016	nervous system	5.42E-10
uberont0001017	central nervous system	1.81E-08
uberont0002107	liver	4.19E-07
uberont0000948	heart	6.07E-07
uberont0001555	digestive tract	1.70E-06
uberont0001032	sensory system	1.93E-06
uberont0007812	post-anal tail	2.28E-06
uberont0002407	pericardium	1.47E-05
uberont0005886	post-hyoid pharyngeal arch skeleton	3.02E-05
uberont0000084	yolk	3.05E-05
uberont0005281	ventricular system of central nervous system	6.07E-05
uberont0010314	structure with developmental contribution from neural crest	0.00074386
uberont0001846	internal ear	0.0008888
uberont0002329	somite	0.00089648
uberont0001003	skin epidermis	0.00170376
uberont0003102	surface structure	0.00190061
uberont0002100	trunk	0.00252997
uberont0001708	jaw skeleton	0.00287797
uberont0004141	heart tube	0.00345967
uberont0001782	pigmented layer of retina	0.00380629
uberont0000017	exocrine pancreas	0.00384968
uberont0000164	caudal fin	0.0046263
uberont0001945	superior colliculus	0.00488419
uberont0012438	blastema of regenerating fin/limb	0.00591299
uberont0000965	lens of camera-type eye	0.00652003
uberont0000966	retina	0.0085354
uberont0002280	otolith	0.01008715
uberont0008897	fin	0.01130278
uberont0002028	hindbrain	0.01447067
uberont0001890	forebrain	0.01678628

uberont 0014907	intersomitic vessel	0.01866955
uberont 0005884	hyoid arch skeleton	0.01978595
uberont 2000106	extension	0.02390583
uberont 0005310	pronephric nephron tubule	0.02958931
uberont 2001456	pectoral fin endoskeletal disc	0.0328624
uberont 0011611	ceratohyal bone	0.0328624
uberont 0003052	midbrain-hindbrain boundary	0.03515936
uberont 0002328	notochord	0.03668385
uberont 0002082	cardiac ventricle	0.03785239
uberont 0006283	future cardiac ventricle	0.04193983
uberont 0005087	tooth placode	0.04412975
uberont 0002422	fourth ventricle	0.04604822

Supplementary Table S2.30. The enriched Biological Process terms from the Gene Ontology that are common to the predicted genes and genes with original annotations for the pectoral fin. The enriched terms are sorted based on the p-value of those terms for the predicted genes.

Term identifier	Term name	P-value for predicted genes	P-value for original genes
GO:0007275	multicellular organism development	2.41E-12	1.85E-09
GO:0009953	dorsal/ventral pattern formation	6.32E-12	2.23E-06
GO:0051216	cartilage development	2.28E-11	4.70E-17
GO:0009880	embryonic pattern specification	4.19E-09	0.0160149
GO:0006355	regulation of transcription, DNA-templated	4.82E-09	1.70E-07
GO:0014032	neural crest cell development	2.91E-08	0.00016897
GO:0001756	somitogenesis	6.37E-08	2.80E-07
GO:0007368	determination of left/right symmetry	1.71E-07	4.78E-09
GO:0030182	neuron differentiation	3.70E-07	9.35E-05
GO:0007224	smoothened signaling pathway	6.03E-07	0.00117078
GO:0048703	embryonic viscerocranium morphogenesis	3.24E-06	2.95E-16
GO:2000223	regulation of BMP signaling pathway involved in heart jogging	3.31E-06	0.0037629
GO:0042476	odontogenesis	4.29E-06	2.35E-06
GO:0021984	adenohypophysis development	4.29E-06	0.00442453
GO:0030902	hindbrain development	4.52E-06	4.34E-08
GO:0009952	anterior/posterior pattern specification	7.59E-06	0.00060755
GO:0001947	heart looping	8.61E-06	6.61E-15
GO:0016055	Wnt signaling pathway	1.66E-05	6.53E-06
GO:0048793	pronephros development	2.04E-05	1.01E-05
GO:0003143	embryonic heart tube morphogenesis	3.81E-05	3.94E-08
GO:0031018	endocrine pancreas development	5.91E-05	0.02264448
GO:0042694	muscle cell fate specification	9.42E-05	0.04578302
GO:0031290	retinal ganglion cell axon guidance	0.000111737	9.33E-06
GO:0048701	embryonic cranial skeleton morphogenesis	0.000292195	5.06E-09
GO:0060070	canonical Wnt signaling pathway	0.00042765	8.61E-05
GO:0021508	floor plate formation	0.000564411	0.00513581
GO:0048264	determination of ventral identity	0.001287915	0.00055075
GO:0060037	pharyngeal system development	0.001287915	0.00055075
GO:0060041	retina development in camera-type eye	0.001442926	5.42E-06
GO:0071599	otic vesicle development	0.001684552	0.0148082
GO:0030917	midbrain-hindbrain boundary development	0.002132019	0.00117078

GO:0060059	embryonic retina morphogenesis in camera-type eye	0.002132019	0.01854958
GO:0045893	positive regulation of transcription, DNA-templated	0.002141963	0.04413722
GO:0008543	fibroblast growth factor receptor signaling pathway	0.003568028	0.03020605
GO:0031016	pancreas development	0.003981011	9.33E-06
GO:0001501	skeletal system development	0.003981011	0.00293934
GO:0043010	camera-type eye development	0.007171621	0.00060755
GO:0007507	heart development	0.010543118	1.44E-08
GO:0003342	proepicardium development	0.012678336	0.00059083
GO:0010002	cardioblast differentiation	0.022708189	3.72E-05
GO:0045892	negative regulation of transcription, DNA-templated	0.024331089	0.00642797
GO:0030166	proteoglycan biosynthetic process	0.025200109	0.00259184
GO:0021986	habenula development	0.032638807	0.00442453
GO:0048384	retinoic acid receptor signaling pathway	0.035106069	0.00015684
GO:0001649	osteoblast differentiation	0.035106069	0.00015684
GO:0048709	oligodendrocyte differentiation	0.035106069	3.28E-06
GO:0030198	extracellular matrix organization	0.047350757	0.00940935

Supplementary Table S2.31. The enriched Biological Process terms from the Gene Ontology that are common to the predicted genes and genes with original annotations for the pelvic fin. The enriched terms are sorted based on the p-value of those terms for the predicted genes.

Term identifier	Term name	P-value for the predicted genes	P-value for the original genes
GO:0033333	fin development	7.11E-10	9.72E-08

Supplementary Table S2.32. The enriched Biological Process terms from the Gene Ontology that are common to the predicted genes and genes with original annotations for the forelimb. The enriched terms are sorted based on the p-value of those terms for the predicted genes.

Term identifier	Term name	P-value for the predicted genes	P-value for the original genes
GO:0030509	BMP signaling pathway	2.12E-17	2.19E-10
GO:0007275	multicellular organism development	5.15E-12	5.81E-25
GO:0007389	pattern specification process	1.71E-09	1.90E-15
GO:0030326	embryonic limb morphogenesis	2.62E-09	1.44E-40
GO:0045669	positive regulation of osteoblast differentiation	3.87E-09	2.15E-10
GO:0045893	positive regulation of transcription, DNA-templated	1.91E-08	8.27E-22
GO:0001701	in utero embryonic development	2.06E-07	3.67E-14
GO:0042475	odontogenesis of dentin-containing tooth	3.08E-07	2.14E-16
GO:0042733	embryonic digit morphogenesis	3.72E-07	6.50E-38
GO:0051216	cartilage development	8.94E-07	1.05E-21
GO:0060021	palate development	1.03E-06	1.88E-18
GO:0060070	canonical Wnt signaling pathway	1.24E-06	1.80E-11
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	1.27E-06	5.94E-32
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	2.30E-06	3.42E-23
GO:0001501	skeletal system development	2.60E-06	9.14E-36
GO:0009952	anterior/posterior pattern specification	3.12E-06	1.71E-17
GO:0001649	osteoblast differentiation	3.23E-06	3.89E-10
GO:0001707	mesoderm formation	4.43E-06	4.00E-05
GO:0071773	cellular response to BMP stimulus	4.43E-06	4.00E-05
GO:0007411	axon guidance	9.70E-06	1.08E-07
GO:0001658	branching involved in ureteric bud morphogenesis	1.16E-05	3.84E-12
GO:0009953	dorsal/ventral pattern formation	1.24E-05	1.42E-21
GO:0008285	negative regulation of cell proliferation	2.11E-05	1.63E-13
GO:0003007	heart morphogenesis	2.78E-05	2.07E-09
GO:0090263	positive regulation of canonical Wnt signaling pathway	3.19E-05	8.99E-12
GO:0016055	Wnt signaling pathway	3.95E-05	2.85E-11
GO:0001503	ossification	9.34E-05	1.17E-21
GO:0090090	negative regulation of canonical Wnt signaling pathway	0.000111887	1.05E-10
GO:0009887	organ morphogenesis	0.000136246	1.77E-11

GO:0045892	negative regulation of transcription, DNA-templated	0.00014868	3.08E-08
GO:0045596	negative regulation of cell differentiation	0.000155607	0.00172739
GO:0048646	anatomical structure formation involved in morphogenesis	0.000226843	5.78E-06
GO:0042474	middle ear morphogenesis	0.000226843	5.78E-06
GO:0003148	outflow tract septum morphogenesis	0.000226843	0.000132268
GO:0001822	kidney development	0.000229367	1.74E-17
GO:0021983	pituitary gland development	0.00035634	1.84E-05
GO:0001837	epithelial to mesenchymal transition	0.000380704	0.00036855
GO:0060349	bone morphogenesis	0.00040586	4.87E-11
GO:0035115	embryonic forelimb morphogenesis	0.000514369	6.35E-37
GO:0030501	positive regulation of bone mineralization	0.000543461	1.62E-13
GO:0060325	face morphogenesis	0.000635432	7.74E-05
GO:0048568	embryonic organ development	0.000635432	4.91E-06
GO:0007492	endoderm development	0.000635432	0.000990141
GO:0030901	midbrain development	0.000635432	1.15E-08
GO:0030514	negative regulation of BMP signaling pathway	0.000840378	0.000152854
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	0.000914879	1.46E-05
GO:0001657	ureteric bud development	0.000914879	9.46E-07
GO:0045599	negative regulation of fat cell differentiation	0.000914879	0.001967343
GO:0003151	outflow tract morphogenesis	0.001156789	7.48E-09
GO:0045668	negative regulation of osteoblast differentiation	0.001243528	0.00346626
GO:0001756	somitogenesis	0.001243528	1.82E-07
GO:0045165	cell fate commitment	0.002045776	2.19E-11
GO:0042472	inner ear morphogenesis	0.002045776	1.01E-07
GO:0050679	positive regulation of epithelial cell proliferation	0.002276056	1.05E-13
GO:0006355	regulation of transcription, DNA-templated	0.0030741	1.70E-09
GO:0007267	cell-cell signaling	0.004133083	1.62E-09
GO:0030336	negative regulation of cell migration	0.004371326	0.030620272
GO:0006351	transcription, DNA-templated	0.005714263	2.85E-08
GO:0030324	lung development	0.005931081	6.63E-18
GO:0048762	mesenchymal cell differentiation	0.010296133	0.00020907
GO:0060272	embryonic skeletal joint morphogenesis	0.01122718	1.20E-07
GO:0043616	keratinocyte proliferation	0.012157404	0.008847589
GO:0061053	somite development	0.013086804	0.00045
GO:0008284	positive regulation of cell proliferation	0.013313841	3.23E-18
GO:0014032	neural crest cell development	0.015870067	8.96E-07
GO:0003203	endocardial cushion morphogenesis	0.015870067	3.15E-05

GO:0035137	hindlimb morphogenesis	0.015870067	1.96E-08
GO:0032331	negative regulation of chondrocyte differentiation	0.016796179	2.91E-08
GO:0051145	smooth muscle cell differentiation	0.016796179	4.01E-05
GO:0060037	pharyngeal system development	0.016796179	0.016734602
GO:0048856	anatomical structure development	0.016796179	4.01E-05
GO:0021904	dorsal/ventral neural tube patterning	0.017721471	4.21E-08
GO:0008283	cell proliferation	0.017771413	3.67E-05
GO:0045879	negative regulation of smoothened signaling pathway	0.019569595	7.65E-05
GO:0023019	signal transduction involved in regulation of gene expression	0.021414447	0.002033958
GO:0060425	lung morphogenesis	0.022335647	0.002305628
GO:0007498	mesoderm development	0.0241756	0.002913859
GO:0007507	heart development	0.024473032	1.18E-18
GO:0042476	odontogenesis	0.025094353	4.27E-07
GO:0035108	limb morphogenesis	0.026012293	6.57E-17
GO:0001702	gastrulation with mouth forming second	0.026012293	0.038461084
GO:0071542	dopaminergic neuron differentiation	0.026929419	7.12E-10
GO:0048663	neuron fate commitment	0.029675922	2.55E-05
GO:0035116	embryonic hindlimb morphogenesis	0.030589801	1.87E-19
GO:0048589	developmental growth	0.032415128	1.19E-10
GO:0034504	protein localization to nucleus	0.032415128	0.000592382
GO:0048754	branching morphogenesis of an epithelial tube	0.033326579	4.55E-12
GO:0002053	positive regulation of mesenchymal cell proliferation	0.034237221	1.34E-18
GO:0030154	cell differentiation	0.034609309	1.49E-12
GO:0060412	ventricular septum morphogenesis	0.035147055	0.008576143
GO:0030879	mammary gland development	0.035147055	6.02E-05
GO:0001656	metanephros development	0.035147055	7.45E-09
GO:0033077	T cell differentiation in thymus	0.036056083	0.000898871
GO:0048706	embryonic skeletal system development	0.040589145	4.11E-11
GO:0045597	positive regulation of cell differentiation	0.042396746	1.13E-05
GO:0048705	skeletal system morphogenesis	0.049595145	1.36E-07



Supplementary Table S2.33. The enriched Biological Process terms from the Gene Ontology that are common to the predicted genes and genes with original annotations for the hindlimb. The enriched terms are sorted based on the p-value of those terms for the predicted genes.

Term identifier	Term name	P-value for the predicted genes	P-value for the original genes
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	4.34E-14	7.62E-37
GO:0045893	positive regulation of transcription, DNA-templated	3.71E-12	4.70E-31
GO:0001649	osteoblast differentiation	1.27E-11	5.39E-12
GO:0050679	positive regulation of epithelial cell proliferation	4.24E-11	2.50E-09
GO:0001837	epithelial to mesenchymal transition	1.74E-09	0.00157941
GO:0060325	face morphogenesis	6.66E-09	0.00497598
GO:0008285	negative regulation of cell proliferation	1.07E-08	1.18E-17
GO:0010629	negative regulation of gene expression	1.13E-08	7.93E-09
GO:0008284	positive regulation of cell proliferation	1.36E-08	8.36E-26
GO:0010628	positive regulation of gene expression	1.49E-08	5.42E-17
GO:0001658	branching involved in ureteric bud morphogenesis	1.72E-08	6.90E-07
GO:0009887	organ morphogenesis	2.72E-08	2.99E-14
GO:0001701	in utero embryonic development	2.89E-08	3.27E-14
GO:0030335	positive regulation of cell migration	4.24E-08	1.74E-05
GO:0030326	embryonic limb morphogenesis	6.92E-08	9.15E-40
GO:0045669	positive regulation of osteoblast differentiation	1.02E-07	8.74E-17
GO:0007507	heart development	2.36E-07	8.32E-18
GO:0060021	palate development	3.14E-07	8.09E-23
GO:0010718	positive regulation of epithelial to mesenchymal transition	3.16E-07	0.01455875
GO:0030509	BMP signaling pathway	3.53E-07	2.76E-12
GO:0045892	negative regulation of transcription, DNA-templated	3.57E-07	4.21E-07
GO:0042060	wound healing	5.21E-07	1.23E-05
GO:0048701	embryonic cranial skeleton morphogenesis	5.57E-07	8.42E-08
GO:0007275	multicellular organism development	5.67E-07	1.89E-21
GO:0001503	ossification	5.78E-07	1.14E-28
GO:0001934	positive regulation of protein phosphorylation	6.10E-07	1.08E-08
GO:0001501	skeletal system development	9.93E-07	4.60E-43
GO:0001657	ureteric bud development	1.30E-06	0.00213229
GO:0030324	lung development	2.06E-06	3.43E-11
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	2.43E-06	7.84E-20

GO:0007389	pattern specification process	2.81E-06	4.16E-11
GO:0042475	odontogenesis of dentin-containing tooth	3.92E-06	6.87E-14
GO:0042733	embryonic digit morphogenesis	4.73E-06	5.30E-25
GO:0030500	regulation of bone mineralization	5.94E-06	1.79E-05
GO:0045165	cell fate commitment	6.70E-06	6.90E-07
GO:0060395	SMAD protein signal transduction	8.76E-06	9.49E-06
GO:0042474	middle ear morphogenesis	9.01E-06	0.00415561
GO:0051216	cartilage development	1.13E-05	6.37E-26
GO:0060070	canonical Wnt signaling pathway	1.56E-05	3.09E-09
GO:0001843	neural tube closure	2.19E-05	3.09E-06
GO:0090090	negative regulation of canonical Wnt signaling pathway	2.67E-05	4.37E-10
GO:0071773	cellular response to BMP stimulus	2.88E-05	0.00038339
GO:0001707	mesoderm formation	2.88E-05	4.37E-06
GO:0048754	branching morphogenesis of an epithelial tube	3.14E-05	5.48E-06
GO:0030501	positive regulation of bone mineralization	3.41E-05	3.34E-09
GO:0030879	mammary gland development	3.70E-05	7.76E-05
GO:0043066	negative regulation of apoptotic process	3.88E-05	3.04E-15
GO:0045596	negative regulation of cell differentiation	4.13E-05	0.01495266
GO:0030514	negative regulation of BMP signaling pathway	6.59E-05	0.03947855
GO:0048705	skeletal system morphogenesis	0.00010666	2.20E-09
GO:0043406	positive regulation of MAP kinase activity	0.00010666	0.00073993
GO:0045668	negative regulation of osteoblast differentiation	0.00011893	0.00015728
GO:0048762	mesenchymal cell differentiation	0.00015495	0.00017231
GO:0009612	response to mechanical stimulus	0.00016899	6.14E-05
GO:0061053	somite development	0.00025555	0.00048879
GO:0007179	transforming growth factor beta receptor signaling pathway	0.00028339	1.47E-07
GO:0070374	positive regulation of ERK1 and ERK2 cascade	0.00028588	3.19E-09
GO:0016477	cell migration	0.00030363	1.90E-06
GO:2000679	positive regulation of transcription regulatory region DNA binding	0.00033627	0.00934741
GO:0003203	endocardial cushion morphogenesis	0.0003807	0.00108701
GO:0016055	Wnt signaling pathway	0.00045893	1.58E-09
GO:0008283	cell proliferation	0.00051842	4.12E-06
GO:1902895	positive regulation of pri-miRNA transcription from RNA polymerase II promoter	0.00064319	0.02270901
GO:0032355	response to estradiol	0.00067863	4.88E-06
GO:0023019	signal transduction involved in regulation of gene expression	0.0007037	0.0035404
GO:0007267	cell-cell signaling	0.00071851	1.79E-07
GO:0031214	biomineral tissue development	0.00112204	0.00012938

GO:0001822	kidney development	0.00140877	2.39E-08
GO:0031016	pancreas development	0.00145299	2.71E-06
GO:0040007	growth	0.00172815	5.48E-06
GO:0002053	positive regulation of mesenchymal cell proliferation	0.00182495	2.88E-17
GO:0030154	cell differentiation	0.00184529	1.77E-08
GO:0007492	endoderm development	0.0021305	0.00497598
GO:0030901	midbrain development	0.0021305	1.26E-06
GO:0010595	positive regulation of endothelial cell migration	0.00223737	0.02729736
GO:0043065	positive regulation of apoptotic process	0.00245285	4.03E-10
GO:0045740	positive regulation of DNA replication	0.00245859	0.00679759
GO:0048706	embryonic skeletal system development	0.00257293	0.00020375
GO:0045597	positive regulation of cell differentiation	0.00280899	4.36E-06
GO:0010862	positive regulation of pathway-restricted SMAD protein phosphorylation	0.00305488	6.90E-07
GO:0045599	negative regulation of fat cell differentiation	0.00305488	6.90E-07
GO:0007568	aging	0.00317746	8.86E-08
GO:0048468	cell development	0.00318148	0.00040482
GO:0009953	dorsal/ventral pattern formation	0.00318148	4.20E-16
GO:0021915	neural tube development	0.00344197	1.25E-06
GO:0006355	regulation of transcription, DNA-templated	0.0035698	3.72E-08
GO:0001756	somitogenesis	0.00413529	3.60E-07
GO:0001666	response to hypoxia	0.00425666	8.11E-06
GO:0001541	ovarian follicle development	0.0042811	0.00102775
GO:0043408	regulation of MAPK cascade	0.00457982	0.00022813
GO:0071560	cellular response to transforming growth factor beta stimulus	0.00488793	4.73E-05
GO:0048812	neuron projection morphogenesis	0.00520538	0.00186241
GO:0003007	heart morphogenesis	0.00536758	1.28E-12
GO:0071363	cellular response to growth factor stimulus	0.00536758	6.98E-05
GO:0048839	inner ear development	0.0056989	8.92E-05
GO:0090263	positive regulation of canonical Wnt signaling pathway	0.00586801	3.47E-11
GO:0001938	positive regulation of endothelial cell proliferation	0.00586801	0.03967443
GO:0050680	negative regulation of epithelial cell proliferation	0.00674768	1.02E-11
GO:0000187	activation of MAPK activity	0.00749191	0.00128144
GO:0060548	negative regulation of cell death	0.00827159	0.02477654
GO:0048661	positive regulation of smooth muscle cell proliferation	0.00827159	1.38E-05
GO:0007050	cell cycle arrest	0.008472	1.56E-05
GO:0016485	protein processing	0.00950655	0.01045742

GO:0007417	central nervous system development	0.01015302	2.40E-08
GO:0061312	BMP signaling pathway involved in heart development	0.0102439	0.00041145
GO:0051897	positive regulation of protein kinase B signaling	0.01196962	0.00160387
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	0.01292758	1.37E-07
GO:1904948	midbrain dopaminergic neuron differentiation	0.0136359	0.01964074
GO:0043410	positive regulation of MAPK cascade	0.01366739	7.62E-11
GO:0051092	positive regulation of NF-kappaB transcription factor activity	0.0149405	0.01163548
GO:0061384	heart trabecula morphogenesis	0.01532769	0.0247888
GO:0008584	male gonad development	0.0157281	7.20E-05
GO:0009952	anterior/posterior pattern specification	0.0157281	4.58E-17
GO:0006468	protein phosphorylation	0.01621133	3.88E-05
GO:0060039	pericardium development	0.01701666	0.03041854
GO:0007165	signal transduction	0.01815958	0.01179635
GO:0007435	salivary gland morphogenesis	0.01870283	0.00305644
GO:0007219	Notch signaling pathway	0.01934164	0.00078917
GO:0030308	negative regulation of cell growth	0.01934164	4.31E-07
GO:0042493	response to drug	0.01991886	3.17E-11
GO:0060272	embryonic skeletal joint morphogenesis	0.02038621	0.00025275
GO:0060389	pathway-restricted SMAD protein phosphorylation	0.02206678	0.0050812
GO:0006366	transcription from RNA polymerase II promoter	0.02326995	0.04072805
GO:0007411	axon guidance	0.02682883	9.88E-07
GO:0061036	positive regulation of cartilage development	0.02709179	0.00934741
GO:0035567	non-canonical Wnt signaling pathway	0.02876123	0.01111749
GO:0060038	cardiac muscle cell proliferation	0.02876123	0.01111749
GO:0035137	hindlimb morphogenesis	0.02876123	1.99E-07
GO:0060037	pharyngeal system development	0.0304279	0.01306762
GO:0017015	regulation of transforming growth factor beta receptor signaling pathway	0.03209181	0.00169332
GO:0046579	positive regulation of Ras protein signal transduction	0.03706695	0.02270901
GO:0006351	transcription, DNA-templated	0.03711583	4.28E-09
GO:0090190	positive regulation of branching involved in ureteric bud morphogenesis	0.03871983	3.17E-05
GO:0045778	positive regulation of ossification	0.03871983	5.75E-09
GO:0035050	embryonic heart tube development	0.03871983	0.02558431
GO:0009880	embryonic pattern specification	0.04036995	4.13E-05
GO:0048646	anatomical structure formation involved in morphogenesis	0.04036995	2.99E-06
GO:0045216	cell-cell junction organization	0.04036995	0.02864555

GO:0032967	positive regulation of collagen biosynthetic process	0.04366199	0.03532201
GO:0035987	endodermal cell differentiation	0.04530391	8.45E-05
GO:0042476	odontogenesis	0.04530391	5.02E-07
GO:0042981	regulation of apoptotic process	0.04548052	8.66E-07
GO:0035108	limb morphogenesis	0.0469431	1.11E-16
GO:0055010	ventricular cardiac muscle tissue morphogenesis	0.0469431	0.00097892
GO:0043392	negative regulation of DNA binding	0.04857957	0.00012938
GO:0071542	dopaminergic neuron differentiation	0.04857957	6.06E-08

Supplementary Table S2.34. The enriched Uberon terms that are common to the predicted genes and genes with original annotations for the pectoral fin. The enriched terms are sorted based on the p-value of those terms for the predicted genes.

Uberon term identifier	Uberon term name	P-value for the predicted genes	P-value for the original genes
uberon_0011610	ceratohyal cartilage	3.81E-13	1.90E-14
uberon_0005886	post-hyoid pharyngeal arch skeleton	1.88E-12	2.56E-19
uberon_2001516	ceratobranchial cartilage	9.63E-11	0.00102618
uberon_0011242	ethmoid cartilage	9.63E-11	0.00029496
uberon_0003107	meckel's cartilage	1.84E-10	3.21E-12
uberon_0003079	floor plate	1.42E-09	0.03374939
uberon_0002533	post-anal tail bud	9.84E-08	0.00338073
uberon_0007215	trabecula cranii	1.50E-07	0.00035249
uberon_0008896	post-hyoid pharyngeal arch	1.85E-07	2.03E-11
uberon_0007812	post-anal tail	6.22E-07	0.00521186
uberon_0002329	somite	7.57E-07	0.00590808
uberon_0001016	nervous system	2.59E-06	0.03822373
uberon_0002328	notochord	6.81E-06	0.00808243
uberon_0011607	hyomandibular cartilage	1.03E-05	6.45E-06
uberon_0001049	neural tube	2.10E-05	0.0227542
uberon_0004752	palatoquadrate cartilage	2.14E-05	2.87E-09
uberon_2000250	opercle	2.74E-05	1.51E-05
uberon_0000165	mouth	3.80E-05	0.00014836
uberon_0000468	multicellular organism	4.20E-05	1.09E-09
uberon_0001894	diencephalon	5.15E-05	0.0201601
uberon_0003901	horizontal septum	7.47E-05	0.00192572
uberon_2001256	lateral floor plate	8.54E-05	0.03796748
uberon_0003051	ear vesicle	0.00019633	2.14E-06
uberon_0005945	neurocranial trabecula	0.0002281	0.00646992
uberon_2001239	ceratobranchial 5 bone	0.00033971	0.00024377
uberon_0003099	cranial neural crest	0.00040528	0.017947
uberon_0002028	hindbrain	0.0005451	0.00387881
uberon_0000044	dorsal root ganglion	0.00085431	0.03471745
uberon_0000965	lens of camera-type eye	0.00096984	0.02462029
uberon_0001890	forebrain	0.00100664	0.00629125
uberon_0001708	jaw skeleton	0.00115871	7.84E-20
uberon_2000694	ceratobranchial 5 tooth	0.00170211	7.74E-06

uberont_2001089	myoseptum	0.00170211	0.01399613
uberont_0002240	spinal cord	0.0017408	0.00254371
uberont_0001264	pancreas	0.00181488	0.00023088
uberont_0003936	postoptic commissure	0.00248307	0.00081772
uberont_0007329	pancreatic duct	0.003566	0.00556203
uberont_0011615	basihyal cartilage	0.00372936	0.0016559
uberont_0003077	paraxial mesoderm	0.00425656	0.00706401
uberont_2000040	median fin fold	0.0046312	1.80E-17
uberont_2000558	posterior macula	0.00500389	0.0087837
uberont_0000033	head	0.0061048	6.69E-08
uberont_0001032	sensory system	0.0062966	4.43E-05
uberont_0003072	optic cup	0.00666504	0.01290014
uberont_0003098	optic stalk	0.00666504	0.00167394
uberont_4000164	caudal fin	0.00707562	9.31E-11
uberont_2001069	ventral fin fold	0.00853811	2.60E-07
uberont_0009635	parachordal cartilage	0.00854196	3.23E-06
uberont_0004741	cleithrum	0.01174598	8.41E-06
uberont_0003278	skeleton of lower jaw	0.0124274	1.01E-06
uberont_0001898	hypothalamus	0.01291407	0.0053699
uberont_0003011	facial motor nucleus	0.01291407	0.0053699
uberont_0000926	mesoderm	0.01413075	0.03471745
uberont_0000948	heart	0.01508289	0.00333015
uberont_0001891	midbrain	0.01623867	0.01553909
uberont_0000019	camera-type eye	0.01756108	0.00298894
uberont_0005884	hyoid arch skeleton	0.0224617	0.00079602
uberont_0000935	anterior commissure	0.02393768	0.00050308
uberont_0011085	palatoquadrate arch	0.02551321	0.01749
uberont_0004375	bone of free limb or fin	0.03295212	0.0093632
uberont_0001703	neurocranium	0.03765523	1.19E-07
uberont_0001976	epithelium of esophagus	0.04102281	0.0137538
uberont_0002348	epicardium	0.04102281	0.00099481
uberont_2002193	dorsolateral septum	0.04102281	0.00099481
uberont_0011004	pharyngeal arch cartilage	0.04146017	2.83E-10
uberont_2001257	medial floor plate	0.04902763	0.0188574
uberont_0005598	trunk somite	0.04902763	0.0188574
uberont_0003114	pharyngeal arch 3	0.04902763	0.0188574

Supplementary Table S2.35. The enriched Uberon terms that are common to the predicted genes and genes with original annotations for the pelvic fin. The enriched terms are sorted based on the p-value of those terms for the predicted genes.

Uberon term identifier	Uberon term name	P-value for the predicted genes	P-value for the original genes
uberon_0000151	pectoral fin	3.26E-07	1.42E-09
uberon_2000040	median fin fold	5.40E-06	8.70E-05
uberon_4000163	anal fin	0.00016979	4.43E-11
uberon_4000172	lepidotrichium	0.00351758	2.42E-06
uberon_0003097	dorsal fin	0.02306165	1.03E-11



Supplementary Table S2.36. The top 100 enriched Uberon terms that are common to the predicted genes and genes with original annotations for the forelimb. The enriched terms are sorted based on the p-value of those terms for the predicted genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon term identifier	Uberon term name	P-value for the predicted genes	P-value for the original genes
uberon_0001703	neurocranium	9.49E-19	3.67E-52
uberon_0011156	facial skeleton	2.21E-18	1.36E-65
uberon_0003128	cranium	5.88E-16	9.57E-79
uberon_0001676	occipital bone	7.46E-16	4.99E-20
uberon_0001708	jaw skeleton	2.83E-15	2.61E-61
uberon_0004716	conceptus	7.34E-14	2.32E-48
uberon_0007811	craniocervical region	7.49E-14	4.41E-77
uberon_0000209	tetrapod frontal bone	1.20E-13	1.61E-17
uberon_0002091	appendicular skeleton	1.27E-13	1.39E-187
uberon_0001434	skeletal system	2.87E-13	2.37E-122
uberon_0005944	axial skeleton plus cranial skeleton	3.25E-13	1.13E-87
uberon_0000165	mouth	1.11E-12	3.64E-61
uberon_0002517	basicranium	1.36E-12	2.14E-27
uberon_0001692	basioccipital bone	2.48E-12	0.00016976
uberon_0003252	thoracic rib cage	2.49E-12	4.72E-74
uberon_0001756	middle ear	3.11E-11	1.26E-24
uberon_0001456	face	5.09E-11	6.78E-61
uberon_0006428	basisphenoid bone	7.87E-11	5.22E-16
uberon_0001049	neural tube	1.06E-10	3.88E-29
uberon_0001684	mandible	1.12E-10	1.56E-51
uberon_0000210	tetrapod parietal bone	1.25E-10	1.03E-16
uberon_0003216	hard palate	1.35E-10	4.89E-27
uberon_0004747	supraoccipital bone	4.05E-10	6.38E-08
uberon_0000033	head	6.18E-10	2.24E-58
uberon_0001690	ear	6.79E-10	9.92E-30
uberon_0001716	secondary palate	1.22E-09	1.29E-40
uberon_0001685	hyoid bone	1.35E-09	1.21E-10
uberon_0001677	sphenoid bone	1.46E-09	8.95E-27
uberon_0001689	malleus bone	1.90E-09	8.78E-13
uberon_0002105	vestibulo-auditory system	2.07E-09	1.06E-28
uberon_0001007	digestive system	3.28E-09	1.27E-40

uberon_0002229	interparietal bone	4.28E-09	3.38E-18
uberon_0002228	rib	4.82E-09	2.08E-59
uberon_0001004	respiratory system	1.82E-08	9.08E-34
uberon_0000955	brain	2.00E-08	1.48E-14
uberon_0001678	temporal bone	2.39E-08	1.11E-14
uberon_0006721	alisphenoid bone	3.22E-08	5.08E-06
uberon_0003450	upper jaw incisor	6.27E-08	9.49E-07
uberon_0003051	ear vesicle	9.95E-08	1.97E-06
uberon_0008828	presphenoid bone	9.95E-08	1.88E-12
uberon_0000922	embryo	1.55E-07	1.90E-25
uberon_0003451	lower jaw incisor	1.66E-07	2.76E-08
uberon_0002218	tympanic ring	1.66E-07	1.04E-15
uberon_0002418	cartilage tissue	5.94E-07	8.02E-30
uberon_0003655	molar tooth	6.91E-07	1.66E-12
uberon_0003966	gonial bone	9.41E-07	0.00011778
uberon_0010389	pterygoid bone	9.41E-07	6.42E-06
uberon_0002510	anterior fontanel	1.34E-06	0.00254357
uberon_0002328	notochord	1.35E-06	2.01E-07
uberon_0001066	intervertebral disk	1.77E-06	2.31E-19
uberon_0011933	vibrissa unit	2.02E-06	3.00E-05
uberon_0001738	thyroid cartilage	2.79E-06	1.22E-07
uberon_0001681	nasal bone	3.03E-06	6.39E-22
uberon_0001890	forebrain	3.25E-06	3.76E-13
uberon_0004649	sphenoid bone pterygoid process	4.50E-06	1.93E-08
uberon_0002329	somite	4.85E-06	6.28E-09
uberon_0004660	mandible coronoid process	5.01E-06	4.14E-07
uberon_0005871	palatine process of maxilla	5.01E-06	1.37E-09
uberon_0001695	squamous part of temporal bone	6.79E-06	7.79E-07
uberon_0005942	hair outer root sheath	6.79E-06	0.01132058
uberon_0003075	neural plate	8.18E-06	0.00165765
uberon_0000924	ectoderm	8.52E-06	6.64E-17
uberon_0018242	palatine bone horizontal plate	9.74E-06	1.20E-07
uberon_0001694	petrous part of temporal bone	9.74E-06	2.04E-05
uberon_0002073	hair follicle	1.01E-05	7.57E-10
uberon_0002224	thoracic cavity	1.06E-05	9.88E-09
uberon_0000401	mandibular ramus	1.15E-05	4.55E-11
uberon_0001737	larynx	1.80E-05	1.91E-10
uberon_0006772	long bone epiphyseal plate hypertrophic zone	1.97E-05	2.91E-40
uberon_0001687	stapes bone	2.19E-05	6.84E-08

uberon_0003861	neural arch	2.26E-05	2.25E-18
uberon_0002470	autopod region	2.28E-05	5.54E-91
uberon_0002103	hindlimb	2.43E-05	2.57E-138
uberon_0001752	enamel	2.98E-05	0.00014902
uberon_0003107	meckel's cartilage	2.98E-05	4.98E-17
uberon_0002413	cervical vertebra	3.01E-05	1.16E-14
uberon_0001894	diencephalon	3.06E-05	5.38E-11
uberon_0001706	nasal septum	3.53E-05	1.69E-09
uberon_0001092	vertebral bone 1	3.73E-05	1.52E-10
uberon_0001682	palatine bone	3.93E-05	1.83E-10
uberon_0002416	integumental system	3.96E-05	2.80E-16
uberon_0000004	nose	4.01E-05	2.04E-24
uberon_0002028	hindbrain	4.59E-05	3.77E-08
uberon_0005354	malleus processus brevis	4.75E-05	0.00721838
uberon_0005619	secondary palatal shelf	4.82E-05	3.00E-17
uberon_0001130	vertebral column	5.03E-05	5.26E-27
uberon_0001691	external ear	5.83E-05	3.87E-15
uberon_0001232	collecting duct of renal tubule	6.68E-05	8.78E-05
uberon_0003982	mature ovarian follicle	6.68E-05	0.00364138
uberon_0001075	bony vertebral centrum	6.98E-05	6.89E-13
uberon_0001675	trigeminal ganglion	6.98E-05	0.000653
uberon_0001998	sternocostal joint	7.60E-05	1.87E-09
uberon_0003461	shoulder bone	8.04E-05	3.33E-19
uberon_0002129	cerebellar cortex	8.27E-05	0.00314678
uberon_0002414	lumbar vertebra	8.67E-05	6.55E-10
uberon_0005867	mandibular prominence	8.70E-05	0.00080115
uberon_0001037	strand of hair	0.00010094	3.10E-07
uberon_0003941	cerebellum anterior vermis	0.00010272	0.01504756
uberon_0002217	synovial joint	0.00010272	0.00100864
uberon_0001043	esophagus	0.00011262	6.33E-07

Supplementary Table S2.37. The top 100 enriched Uberon terms that are common to the predicted genes and genes with original annotations for the hindlimb. The enriched terms are sorted based on the p-value of those terms for the predicted genes. The full enriched term list is available at [https://github.com/pasanfernando/Chapter2\\_datafiles](https://github.com/pasanfernando/Chapter2_datafiles) repository.

Uberon term identifier	Uberon term name	P-value for the predicted genes	P-value for the original genes
uberont_0003128	cranium	4.01E-24	5.82E-57
uberont_0011156	facial skeleton	1.11E-22	4.74E-53
uberont_0001434	skeletal system	1.42E-21	1.19E-223
uberont_0005944	axial skeleton plus cranial skeleton	4.39E-21	3.58E-88
uberont_0007811	craniocervical region	3.88E-19	2.99E-57
uberont_0001456	face	4.19E-19	2.62E-45
uberont_0000033	head	1.62E-18	5.98E-44
uberont_0001708	jaw skeleton	2.11E-17	1.39E-53
uberont_0001703	neurocranium	3.94E-17	2.49E-44
uberont_0000165	mouth	3.90E-16	1.59E-47
uberont_0002091	appendicular skeleton	1.44E-15	0
uberont_0002105	vestibulo-auditory system	1.25E-14	1.85E-25
uberont_0001690	ear	3.73E-14	1.27E-25
uberont_0001684	mandible	4.59E-14	8.20E-43
uberont_0001007	digestive system	1.41E-13	1.04E-28
uberont_0006333	snout	2.45E-13	8.50E-22
uberont_0002483	trabecular bone tissue	1.84E-12	4.59E-72
uberont_0000210	tetrapod parietal bone	1.91E-12	6.03E-14
uberont_0001756	middle ear	1.96E-12	9.35E-20
uberont_0001004	respiratory system	4.07E-12	8.92E-32
uberont_0000209	tetrapod frontal bone	3.47E-11	3.65E-13
uberont_0000383	musculature of body	4.96E-11	1.02E-56
uberont_0001630	muscle organ	1.56E-10	0.00011844
uberont_0002517	basicranium	3.84E-10	1.66E-17
uberont_0001049	neural tube	7.14E-10	2.88E-19
uberont_0001676	occipital bone	1.39E-09	1.36E-12
uberont_0002229	interparietal bone	1.95E-09	1.81E-10
uberont_0001677	sphenoid bone	2.11E-09	1.50E-13
uberont_0000955	brain	3.27E-09	2.55E-14
uberont_0006428	basisphenoid bone	4.44E-09	2.35E-07

uberon_0002228	rib	4.70E-09	4.14E-37
uberon_0003252	thoracic rib cage	9.23E-09	5.47E-50
uberon_0004716	conceptus	1.03E-08	3.81E-21
uberon_0003450	upper jaw incisor	1.09E-08	2.32E-05
uberon_0006772	long bone epiphyseal plate hypertrophic zone	2.06E-08	6.26E-55
uberon_0002416	integumental system	3.26E-08	9.53E-16
uberon_0003451	lower jaw incisor	3.59E-08	2.90E-09
uberon_0000014	zone of skin	4.54E-08	9.99E-19
uberon_0001689	malleus bone	5.05E-08	2.31E-12
uberon_0001130	vertebral column	5.84E-08	2.46E-37
uberon_0003107	meckel's cartilage	1.03E-07	2.75E-12
uberon_0004535	cardiovascular system	2.17E-07	5.30E-28
uberon_0002113	kidney	2.72E-07	3.61E-13
uberon_0006721	alisphenoid bone	4.21E-07	6.91E-06
uberon_0000474	female reproductive system	4.37E-07	7.29E-19
uberon_0003216	hard palate	4.55E-07	3.86E-17
uberon_0000019	camera-type eye	4.68E-07	1.67E-23
uberon_0002516	epiphyseal plate	4.92E-07	4.29E-72
uberon_0001678	temporal bone	6.21E-07	4.05E-12
uberon_0002418	cartilage tissue	6.60E-07	3.22E-14
uberon_0001890	forebrain	6.60E-07	2.57E-12
uberon_0005070	anterior neuropore	6.63E-07	0.04722767
uberon_0001692	basioccipital bone	9.19E-07	0.00016117
uberon_0002101	limb	9.91E-07	4.55E-08
uberon_0001008	renal system	1.01E-06	5.55E-21
uberon_0002104	visual system	1.18E-06	1.15E-25
uberon_0002397	maxilla	1.25E-06	8.65E-29
uberon_0001681	nasal bone	1.28E-06	2.59E-17
uberon_0002370	thymus	1.86E-06	5.49E-13
uberon_0002405	immune system	1.88E-06	1.05E-06
uberon_0003697	abdominal wall	2.07E-06	1.51E-07
uberon_0002218	tympanic ring	2.14E-06	5.31E-11
uberon_0001716	secondary palate	2.86E-06	2.15E-25
uberon_0000948	heart	2.92E-06	1.31E-18
uberon_0002390	hematopoietic system	3.26E-06	3.56E-07
uberon_0002544	digit	3.57E-06	2.75E-48
uberon_0003461	shoulder bone	4.45E-06	2.81E-14
uberon_0001752	enamel	5.00E-06	1.17E-06
uberon_0006849	scapula	5.00E-06	1.09E-18

uberon_0002491	lambdoid suture	5.12E-06	0.00065007
uberon_0003975	internal female genitalia	5.36E-06	5.15E-14
uberon_0000323	late embryo	8.00E-06	4.11E-10
uberon_0002428	limb bone	8.82E-06	0.00139297
uberon_0000309	body wall	9.97E-06	7.79E-08
uberon_0002048	lung	1.23E-05	3.26E-20
uberon_0002470	autopod region	1.37E-05	1.95E-55
uberon_0001229	renal corpuscle	1.41E-05	4.22E-05
uberon_0001075	bony vertebral centrum	1.53E-05	1.30E-09
uberon_0002328	notochord	1.71E-05	2.82E-07
uberon_0001285	nephron	1.79E-05	1.62E-06
uberon_0002073	hair follicle	1.82E-05	3.40E-07
uberon_0001225	cortex of kidney	2.68E-05	1.80E-05
uberon_0001987	placenta	2.68E-05	1.80E-05
uberon_0006861	diaphysis proper	3.28E-05	3.71E-12
uberon_0006771	long bone epiphyseal plate proliferative zone	3.45E-05	2.52E-29
uberon_0001016	nervous system	3.73E-05	5.57E-07
uberon_0005956	outflow part of left ventricle	3.78E-05	0.00017418
uberon_0001037	strand of hair	4.12E-05	3.95E-06
uberon_0002137	aortic valve	4.12E-05	0.00021436
uberon_0001695	squamous part of temporal bone	4.43E-05	0.00058499
uberon_0010166	coat of hair	5.19E-05	1.65E-06
uberon_0002329	somite	5.99E-05	0.00015746
uberon_0001758	periodontium	6.34E-05	3.57E-09
uberon_0003051	ear vesicle	6.89E-05	0.00027719
uberon_0008828	presphenoid bone	6.89E-05	3.51E-08
uberon_0001683	jugal bone	6.89E-05	8.05E-11
uberon_0002224	thoracic cavity	6.89E-05	5.32E-05
uberon_0002412	vertebra	7.08E-05	2.93E-18
uberon_0000945	stomach	7.42E-05	0.00021632
uberon_0000012	somatic nervous system	7.46E-05	1.69E-27

### **Chapter 3: Integrate large-scale trait data with large phylogenies by computationally solving the challenges associated with big data integration.**

#### **Abstract**

When studying the evolution of anatomical characters, such as the pectoral fin and the pelvic fin, it is important to infer their evolutionary history using ancestral state reconstructions. Performing large-scale ancestral state reconstructions using large trait data matrices and large phylogenetic trees is important when conducting macroevolutionary studies, but previous ancestral state reconstructions were limited in scale due to computational constraints. Using a large synthetic morphological supermatrix for paired fins retrieved from the Phenoscape Knowledgebase (KB) and a large species-level tree for teleost fishes retrieved from the Open Tree of Life (Open Tree), the integration challenges were solved by developing new computational methods that mostly focused on extending the original data volume and minimizing the data loss. For example, the data propagation algorithm developed during this work extended the original data significantly and reduced the missing data percentage from 85.9% to 38.4%. This was critical for conducting efficient ancestral state reconstructions. These methods were arranged into a bioinformatics pipeline (PhenTree pipeline) that can be used to integrate any anatomical character from the Phenoscape KB with a large phylogenetic tree retrieved from the Open Tree, which then can be used to perform large-scale ancestral state reconstructions to study the evolution of anatomical characters.

### 3.1 Background

Understanding the evolutionary history of anatomical traits is extremely important in evolutionary biology. For instance, the absence of the pectoral fin in fishes such as eels has fascinated the researchers since at least the time of Aristotle (Leunissen, 2010; Ogle, 1882) and is investigated continuously until now. To understand the evolution of anatomical traits, it is important not only to observe the character states of extant species but also to understand the characteristics of ancestral taxa, which are extinct or available as fossils. The process of inferring unknown ancestral states based on observed states of existing species is identified as the ancestral state reconstruction (Cunningham, 1999; Didier, 2017). This requires the integration of a trait matrix with a phylogenetic tree. Ancestral state reconstructions enable tracing the evolution of a particular trait through the evolutionary history and identify important events, such as changes from the presence to the absence or potential regains of anatomical entities.

Currently, biological data are accumulated at a rapid pace; next-generation experimental methods generate large volumes of data, which has introduced the concept of big data analysis (Fan, et al., 2014). However, ancestral state reconstructions are mostly conducted at a smaller scale due to the limited availability of data sources and integration challenges (Jackson, et al., 2018). The availability of large-scale trait matrices and large phylogenies were limited until recently when resources such as the Phenoscape Knowledgebase (Phenoscape KB; <https://kb.phenoscape.org>) (Dececchi, et al., 2015; Jackson, et al., 2018) and the Open Tree of Life (Open Tree; <https://tree.opentreeoflife.org>) (Jackson, et al., 2018; Rees and Cranston, 2017) emerged.

The Phenoscape KB contains ontology-annotated phenotypic data for vertebrates, which are primarily based on published character matrices (Dahdul, et al., 2015; Dececchi, et al., 2015;



Edmunds, et al., 2016). The use of ontology terms to annotate free text phenotypic descriptions from literature enables the Phenoscape KB data to be readily extractable and computable, which is ideal for large-scale studies across multiple species. The phenotypic data from the Phenoscape KB are associated with the taxonomic names from the Vertebrate Taxonomy Ontology (VTO) (Midford, et al., 2013), which primarily uses the Catalog of Fishes (Eschmeyer, 2013) for extant fishes, the National Center for Biotechnology Information (NCBI) taxonomy for extant tetrapods, and the PaleoDb (Melott, 2008) to supplement the extinct taxa. The phenotypic characters in the Phenoscape KB are annotated using the Entity-Quality (EQ) formalism (Dececchi, et al., 2015; Mungall, et al., 2010; Mungall, et al., 2007), where the anatomical entities such as pectoral fin and their relationships are based on the Uberon anatomy ontology (Haendel, et al., 2014; Mungall, et al., 2012), and the quality terms that represent the variation in the anatomical entities, such as size, presence/absence, and shape are drawn from the Phenotype and Trait Ontology (PATO) (Gkoutos, et al., 2005).

The Open Tree is a comprehensive data source for the retrieval of large phylogenetic trees (Hinchliff, et al., 2015). It uses the ‘propinquity’ supertree pipeline to dynamically synthesize trees by using multiple input phylogenies along with a reference taxonomy (Redelings and Holder, 2017). The input phylogenies for the Open Tree are published trees that are manually curated to align tips with the Open Tree reference taxonomy (Rees and Cranston, 2017). The single rooted supertrees constructed by the pipeline can be customized according to the user preference. If the species relationships of the tree are unresolved, they can be manually curated and adjusted to retrieve a better-resolved tree, which is important when performing ancestral state reconstructions (Jackson, et al., 2018). Even without user modifications, the Open

Tree interface (<https://tree.opentreeoflife.org>) allows the user to easily retrieve large phylogenetic trees for any taxonomic group of interest.

Despite the availability of these resources, integration of large trait matrices with large phylogenetic trees has remained a challenge (Jackson, et al., 2018). The computational algorithms required for the integration and the visualization of large phylogenetic trees require improvements (Harmon, et al., 2013; Harris and Arbuckle, 2016; Hunt and Slater, 2016). The integration of trait data with phylogenetic trees requires a successful transfer of trait data to the phylogenetic tree, which involves minimizing the amount of data lost during the integration. Especially when the two data sources use different taxonomic naming systems, matching the correct names at a large scale is challenging. Therefore, different taxonomic reconciliation solutions, such as the Taxonomic Name Resolution Service (TNRS), which aggregates taxonomic data from different data sources (Boyle, et al., 2013), have appeared. As the size of the trait data and the phylogenetic trees increases, these solutions must be improved to facilitate the extra volume of data and new solutions must be developed; this is the primary objective of this Chapter.

Integration of trait data retrieved from the Phenoscope KB and phylogenetic trees retrieved from the Open Tree is required for performing large-scale ancestral state reconstructions. However, this integration cannot be achieved using existing computational solutions, thus requires manual processing, which is not applicable for large-scale data. Therefore, new computational solutions were developed to solve the challenges associated with the integration. The developed algorithms were arranged into a bioinformatics pipeline (PhenTree pipeline), which can be used to easily integrate any trait data matrix that contains a single anatomical character retrieved from the Phenoscope KB with any phylogenetic tree

downloaded from the Open Tree. The functionality of the pipeline was demonstrated using the evolution of paired fins in teleost fishes as the use case, which involved a phylogeny that contained over 38,000 taxa (Jackson, et al., 2018). This pipeline contains computational algorithms to reduce the number of mismatches between the two data sources and also to extend the original data amount. This Chapter discusses the challenges that are associated with integrating large trait matrices with large phylogenies and the computational methods developed, which were used to solve those challenges.

## **3.2 Methods**

### *3.2.1 Retrieval of a synthetic morphological supermatrix from the Phenoscape KB*

The Phenoscape KB generates synthetic morphological supermatrices, which are large-scale presence/absence trait matrices that can be retrieved using the OntoTrace tool (Dececchi, et al., 2015). A synthetic morphological supermatrix for the pectoral fin and the pelvic fin for Teleostei used in Jackson, et al. (2018), which was downloaded from the OntoTrace was used for this work. This matrix was in NeXML format (Vos, et al., 2012) and contained data provenance in metadata. In the matrix, the presence is denoted by ‘1’ and the absence is denoted by ‘0’. At the time of the retrieval, the Phenoscape KB contained 21,569 character states annotated with 526,221 phenotypes for 5,208 extant and fossil vertebrates from 171 comparative studies (Jackson, et al., 2018).

### *3.2.2 Retrieval of a species-level tree/phylogeny from the Open Tree*

The Open Tree (Hinchliff, et al., 2015) can be used to retrieve a species-level tree for any taxonomic group using the public interface (<https://tree.opentreeoflife.org>). The trees retrieved

from the interface are based on the publicly available supertree (Open Tree 2.10) generated by the propinquity pipeline (Redelings and Holder, 2017). It also enables the user to customize the tree if the taxonomic groups in the publicly available supertree are not fully resolved. The phylogenies retrieved from the Open Tree are in Newick format (Cardona, et al., 2008). For this work, a customized Teleostei species-level tree was obtained from Jackson, et al. (2018), which was originally retrieved from the Open Tree.

### 3.2.3 Pre-processing the synthetic morphological supermatrix and the species-level tree

The first challenge when integrating the synthetic morphological supermatrix with the Teleostei species-level tree is pre-processing them to remove resource specific anomalies. For instance, the synthetic morphological supermatrix is in NeXML format, which is not ideal for downstream analyses. Therefore, it was converted into the tab-delimited format, which is easy to manipulate and read. Furthermore, the synthetic morphological supermatrices retrieved from the Phenoscope KB can contain missing character states denoted by '?', which were removed during the pre-processing.

The phylogenetic trees downloaded from the Open Tree also require pre-processing. The Newick format tree file, retrieved from the Open Tree, contained Open Tree identification numbers at the end of each taxon name (e.g., *Scleropages\_jardinii\_ott335719*, *Danio\_rerio\_ott1005914*), which needed to be removed before further processing. Moreover, some of the large phylogenetic trees from the Open Tree may contain unifurcating nodes, i.e., internal nodes on the tree with exactly one child node (Sukumaran and Holder, 2018). For instance, in the Teleostei species-level tree, there are some genera with exactly one species for each genus (e.g., genus *Heterophotus* only contains *Heterophotus ophistoma* species). The

presence of unifurcating nodes typically causes problems when manipulating the tree with phylogenetic software (Sukumaran and Holder, 2018). Particularly, Mesquite (Maddison and Maddison, 2016), which was used to perform ancestral state reconstructions in this work, generates errors when a tree contains unifurcating nodes. Therefore, the unifurcating nodes on the tree were suppressed using the DendroPy Python library (Sukumaran and Holder, 2010; Sukumaran and Holder, 2018).

#### *3.2.4 Removal of apparent polymorphisms and conflicts*

The supermatrices retrieved from the Phenoscope KB contain taxa with both presence and absence for a particular character denoted by '0&1'. Using the provenance reports available in the metadata, the reasons for observing '0&1' can be identified; these are due to either actual polymorphisms, apparent polymorphisms, or conflicts (Jackson, et al., 2018). Actual polymorphisms are described at the species level and both presence and absence are declared in the same reference. For instance, in fishes in which the anatomical entity is present in one gender (e.g., female) and absent in the other gender (e.g., male), both presence and absence could be represented in the matrix. Apparent polymorphisms are described at higher taxonomic levels (e.g., genus, family), and the author does not provide the information about the species in which the anatomical entity is present or missing. In such instances, where '0&1' is assigned to higher-level taxa, it is difficult to trace which species lack or retain the anatomical entity. Conflicts occur when two different authors make conflicting statements about presence and absence for the same anatomical entity for the same taxon. These conflicts could occur at any taxonomic level.

It is important to identify the reason for these '0&1' states. If it is due to conflicts, the conflicting states can be investigated and corrected if one author has incorrectly described the

state of the anatomical entity. Using the metadata for the morphological supermatrix, the taxa with ‘0&1’ states were identified for further investigation. If the ‘0&1’ states were present in higher-level taxa due to apparent polymorphisms or conflicts, those states were removed from the pre-processed matrix to facilitate data propagation from higher-level taxa to species at a later stage (section 3.2.6).

### *3.2.5 Distinguishing inference versus assertion*

The phenotypic data available in the Phenoscape KB are tagged with ontology terms, which gives the advantage of inferring the phenotypic state of an anatomical entity using indirect descriptions of the entity and its parts (Jackson, et al., 2018). The ontology-based inference has been shown to greatly expand the available data (Dececchi, et al., 2015), a desirable feature given the insufficiency of direct statements/assertions by authors regarding the anatomical phenotypes.

An example that shows the use of ontology relationships to infer phenotypic states is given in Fig. 3.1 (Jackson, et al., 2018). In the example, the phenotype “pectoral fin rays are unbranched” was annotated as Entity: “pectoral fin ray,” Quality: “branched”, and based on the ontology relationships, the pectoral fin was inferred as present because pectoral fin ray is a part of the pectoral fin (Fig. 3.1). However, the converse is not true; the presence of pectoral fin does not mean that all the parts of the pectoral fin are present.

Although ontology-based inference expands the amount of data available in supermatrices from the Phenoscape KB, they are based on computational predictions. Therefore, the reliability of inferred data is lower than direct author assertions about phenotypes, which are based on specimens of the species. When using the supermatrices from the Phenoscape KB, it is

important to distinguish between the inferred *versus* asserted data to understand the proportion of data based on inference. This was done using a Python script that uses the metadata associated with the morphological supermatrix. This script adds a new column (inferred status column) to the matrix (the matrix with the ‘0&1’ states removed from higher-level taxa), which indicates the inferred status of the anatomical character for each taxon in the matrix. In the case of an inferred presence for a particular taxon, the character state of ‘2’ is indicated in the inferred status column for the taxon; the character state ‘3’ is indicated in the inferred status column for a taxon with inferred absence. When presence and absence is asserted, the original states of ‘1’ and ‘0’ is copied to the inferred status column, respectively. This inferred status column can be used to differentially visualize inferred *versus* asserted character states in the phylogenetic tree after performing the ancestral state reconstruction.

### 3.2.6 Data propagation

The morphological supermatrices from Phenoscope KB contain character states annotated to higher-level taxa (e.g., family, genus, etc.). In such cases, the authors wanted to convey that the character states such as presence and absence for higher-level taxa would apply to all their descendants at the species level. However, character states at higher-level taxa cannot be utilized by current tools for ancestral state reconstruction without manual editing, which is not feasible for large-scale data (Jackson, et al., 2018). For example, the PhyTools R package (Revell, 2012) has not implemented the functionality to perform reconstructions using data at internal nodes, but it has developed a workaround method, which could not be applied to large-scale phylogenetic trees (Jackson, et al., 2018). It is important to use this data at higher-level taxa when performing large-scale ancestral state reconstructions because of the paucity of data in trait matrices.

Therefore, an algorithm was developed to propagate character data annotated to families and genera to their corresponding species.

The use of VTO to annotate taxa in the Phenoscape KB helps in implementation of the propagation algorithm. Ontologies are usually stored as directed acyclic graphs (Pesquita, et al., 2009), which simplifies the manipulations and analyses that use the ontology data. The propagation algorithm stores the VTO relationships in a graph data structure, which is ideal for retrieving the relationships between higher-level taxa (families and genera) and their species. This makes it easier to propagate data from higher-level taxa to species, reducing the running time of the algorithm. The propagation algorithm works in two iterations (Fig. 3.2). During the first iteration, character states from genera are propagated to the corresponding species. During the second iteration, character states from families are propagated to the remaining species that did not receive propagated data from the first iteration. During the propagation, species with existing character states are not replaced but considered as propagation conflicts (Fig. 3.2). This applies to the species with directly asserted or inferred character states and species with data propagated from genera during the first iteration. Taxonomic levels above family are not considered for the propagation. After the propagation, species that were not included in the original morphological supermatrix were added to the matrix with the character states propagated from their higher-level taxa. To distinguish the propagated character states from the non-propagated character states, a new column was added to the matrix to indicate the propagated status. This column can also be used during the visualizations of the matrix data in the species-level tree to visualize the distribution of the propagated data in the tree.

The propagation algorithm propagates the data from higher-level taxa (families and genera) as intended by the authors. For example, when an author states that pectoral fin is present



for a certain genus, it is meant that the pectoral fin is present for all the species of that genus. However, during the propagation process, there are occasions where the propagated character state is conflicting with an already existing state of a species. For instance, when the author states that pectoral fin is present for a certain genus, and if the pectoral fin is absent in one of the species in that genus, it can be considered as a propagation conflict. Furthermore, during the second propagation iteration, when character states are propagated from a family to corresponding species, these states can conflict with an opposite state that was propagated during the first iteration from a genus that is also a descendent of that family. Propagation algorithm retains the original character state of the species in these situations, but it is important to estimate the number of conflicts to indicate the reliability of the propagation algorithm. Therefore, the number of conflicts occurred during the propagation for the Teleostei morphological supermatrix with two characters (pectoral fin and pelvic fin) were counted. To further estimate the reliability of the propagation algorithm in a much larger matrix, a new matrix was downloaded from the Phenoscope KB using the OntoTrace to include all the parts of the paired fin and include all characters with values for Teleostei (obtained 9/1/2018). The propagation algorithm was implemented on the larger matrix, one character at a time, and the number of propagation conflicts was recorded.

### *3.2.7 Reconciliation of taxon names*

Another challenge when integrating the Phenoscope KB supermatrices with the Open Tree phylogenetic trees is reconciling taxon names between the two data files. The Phenoscope KB uses VTO for the taxon names and the Open Tree taxon names are based on the NCBI taxonomy system, the Interim Register for Marine and Non-marine Genera (IRMNG), the Global

Biodiversity Information Facility (GBIF), and the World Register of Marine Species (WoRMS) (Jackson, et al., 2018). Therefore, the taxon names must be reconciled between the two data files to reduce the number of mismatches. A combined algorithm was developed to reconcile the taxon names by first matching the taxon names using the NCBI taxonomy IDs (Sayers, et al., 2009) as the common identifier and then matching the remaining taxa using exact taxon names. The mismatched taxa were printed as a list to identify the reason for the mismatches. This algorithm converts the VTO taxon names of the matrix with propagated data into Open Tree taxon names to easily integrate with the Open Tree species-level tree.

### *3.2.8 Ancestral state reconstruction*

The final output matrix generated in the previous section (3.2.7) was used by Jackson, et al. (2018) to perform ancestral state reconstructions to study the evolution of the pectoral and the pelvic fin.

There were three types of data in the final output matrix: asserted data based on direct author statements, inferred only data generated using ontology-based inference (excluding data with both inferred and asserted states; they are considered as asserted data), and data propagated from higher-level taxa. Different colors were assigned to asserted, inferred only, and propagated data using the additional columns in the final output matrix for pectoral and pelvic fins, which indicate the inferred only and propagated status for a particular character state. These data were visualized in Mesquite (Maddison and Maddison, 2016) using ancestral state reconstructions.

### 3.2.9 The PhenTree bioinformatics pipeline

Although this work demonstrates the use of the developed computational algorithms to integrate a large species-level tree with a synthetic morphological supermatrix to study the evolution of paired fins in teleost fishes, they can be applied to study the evolution of any anatomical character. To make the computational algorithms reusable, they were assembled into a bioinformatics pipeline called the ‘PhenTree’. The PhenTree pipeline includes the discussed steps from the pre-processing the data files (section 3.2.3) to the reconciliation of taxon names (section 3.2.7). Although OntoTrace can generate synthetic morphological supermatrices with multiple characters, the pipeline works for matrices with one single character because ancestral state reconstructions are performed on one character at a time. The output of the pipeline is a matrix that can be integrated with any species-level tree from the Open Tree. The pipeline was developed in Python programming language (Van Rossum and Drake, 2011) and contains six steps (matrix conversion, pre-processing the input matrix, removal of apparent polymorphisms and conflicts, distinguishing inference *versus* assertion, data propagation, and reconciliation of taxon names) (Fig. 3.3). This pipeline is available as a command-line tool ([https://github.com/pasanfernando/generic\\_pipeline\\_for\\_trait\\_integration](https://github.com/pasanfernando/generic_pipeline_for_trait_integration)) to be downloaded and run on a computer with any operating system (Windows, Mac, Linux) with Python. More details of the command-line version can be found in the Fernando, et al. (2017). Furthermore, the PhenTree pipeline is also available as a web tool (<http://phentree.biocomps.org/>), which is easier to use. A detailed tutorial about the PhenTree web tool is also included. A snapshot of the PhenTree web tool is shown in Fig. 3.4.

### 3.3 Results

#### 3.3.1 *The synthetic morphological supermatrix and the species-level tree*

The synthetic morphological supermatrix from the OntoTrace contained two characters (pectoral fin and pelvic fin) associated with 3047 taxa (2,663 species, 132 genera, 223 families, and 29 supra-familial taxa) from 87 studies (Jackson, et al., 2018). There were 6,094 total cells in the synthetic morphological matrix of which 4,853 were populated. Out of the 4,853 populated cells, 616 contained only directly asserted data, 3,953 contained only inferred data, and 284 contained both asserted and inferred data. For the pectoral fin, 246 taxa have only asserted data, 2,020 taxa have only inferred data, and 42 taxa have both asserted and inferred data. For the pelvic fin, 370 taxa have only asserted data, 1,933 taxa have only inferred data, and 242 taxa have both asserted and inferred data. The species-level tree for Teleostei obtained from the Open Tree contained 38,419 species-level tips and 560 families.

#### 3.3.2 *Removal of apparent polymorphisms and conflicts*

Conflicts and apparent polymorphic character states were identified for 74 higher-level taxa (50 families and 24 genera for pelvic fin and 4 families for pectoral fin), which were removed from the matrix (Jackson, et al., 2018). Actual polymorphic character states (within species variation identified by a single author) were found only for the pelvic fin in five species (a catfish, *Glanapteryx anguilla*; two hatchet herrings, *Pristigaster cayana* and *Pristigaster sp.*; and two priapumfishes, *Phallostethus lehi* and *Phallostethus dunkeri*). Conflicts that were found at the species-level as a result of data aggregation and inference were all between asserted and inferred states. These were found in the pelvic fin for five species (the eel catfish, *Channallabes apus*, two air-breathing catfishes, *Dolichallabes microphthalmus*, and *Gymnallabes typus*, the

cobia, *Rachycentron canadum*, and the three-spined stickleback *Gasterosteus aculeatus*), and in the pectoral fin for one species (the bobtail snipe eel, *Neocyema erythrosoma*) (Jackson, et al., 2018). Conflicts and the actual polymorphisms at the species level were retained in the matrix because they do not affect the propagation.

### 3.3.3 Data propagation

The propagation algorithm transferred asserted and inferred data from a total of 182 families and 119 genera to the corresponding species that otherwise lacked data for the pectoral and pelvic fins (Jackson, et al., 2018). This resulted in an addition of 11,293 new species to the pre-processed data matrix that contained 2,663 species, which increased the total number of species to 13,956 in the resulted propagated matrix. When considering the two fins separately, the number of species with data for the pectoral fin increased from 2,023 to 11,594 (9,571 species added) and species for the pelvic fin increased from 2,478 to 6,878 (4,400 species added) as a result of the propagation. Of the propagated species, there were 12 instances of propagation conflicts (6 conflicts for the pectoral fin and 6 conflicts for the pelvic fin) when the propagated character state does not agree with the existing species-level character state out of 4,501 instances of existing species level states for both fins (2,023 for pectoral fin and 2,478 for pelvic fin). This percentage (0.27%) of propagation conflicts is quite negligible compared to the number of propagation attempts made on existing species data.

To further analyze the number of propagation conflicts, a larger matrix for Teleostei that consists all the available parts for the paired fin (pectoral fin and pelvic fin) was retrieved. This matrix contained 115 anatomical characters (Supplementary Table S3.1) for 3,109 taxa (2,725 species, 132 genera, 223 families, and 29 supra-familial taxa). The propagation algorithm was

implemented on each character separately and the propagation statistics for each character is shown in Supplementary Table S3.1. In total, the propagation added species-level data to 420,254 cells that otherwise did not have any data for all the characters. Furthermore, out of 42,611 cells with existing species-level data for all the characters, propagated states only conflicted with 255 (0.60%) cells with existing data, which is relatively low compared to the amount of propagated states.

### *3.3.4 Reconciliation of taxon names*

The combined algorithm for the taxon name reconciliation, which initially uses the NCBI IDs for matching and then uses exact taxon name matches, managed to successfully match 12,582 of the 13,956 species in the matrix with the 38,419 species in the Teleostei species-level tree (Jackson, et al., 2018). This number is higher than using either method alone (NCBI taxonomic IDs: 4,423 matches; exact taxon name matching: 12,500 matches). Of the mismatched 1,374 species (of the 13,956), 72 are fossil species which are not included in the Open Tree taxonomic sources, 362 are species with unconventional names that were added to the VTO because they are referenced in publications (e.g., “*Notropis* sp. sawfin shiner (Coburn and Cavender 1992)”), and 940 are mismatched for multiple other reasons (e.g., taxonomic name changes between the two data sources, extinct species that are not marked as such in the VTO). The mismatched species are listed in the Supplementary Table S3.2.

Before the propagation, the pre-processed data matrix contained 2663 species for two characters (pectoral fin and pelvic fin), with 3538 populated cells for species (85.9% missing data) (Table 3.1). After the propagation, the final output matrix contained 12,582 species with 16,408 populated cells (34.8% missing data) (Table 3.1).

Of the 16,408 populated cells in the final output matrix, 494 (150 pectoral, 344 pelvic) contained only directly asserted data (Fig. 3.5) where the presence of the pectoral fin was asserted in 123 species, and absence asserted in 30. The presence of the pelvic fin was directly asserted in 150 species and absence asserted in 194. In the remaining cells, 3,044 (1,511 pectoral, 1,533 pelvic) contained only inferred data, and 12,870 cells (8,798 pectoral, 4,072 pelvic) contained propagated data (Fig. 3.5). Of the 8,798 species for which pectoral fin data were propagated, 5,077 of these were propagated from asserted family and genus-level data. Of the 4,072 species for which pelvic fin data were propagated, 2,906 of these were propagated from asserted family and genus-level data. These are numbers after the reconciliation of taxon names; hence, only the species that matched with the species-level tree file are considered.

### *3.3.5 Ancestral state reconstruction*

The results of the ancestral state reconstructions are available in Jackson, et al. (2018). Data propagation and ontology-based inference play a major role in extending the original data volume to perform large-scale ancestral state reconstructions. Fig. 3.6 visually represents the proportion of propagated and inferred only data after performing the ancestral state reconstructions for the pectoral fin and the pelvic fin. It is clear that asserted data (shown in dark blue) that are based on direct author statements represent a smaller portion in the visualizations for both the fins. Without inferred (light blue) and propagated (green) data these large-scale ancestral state reconstructions become inefficient due to the low data volume.

### 3.4 Discussion

Integrating large-scale trait matrices with large phylogenetic trees introduces several computational challenges, which were investigated during this work. Computational solutions were developed using a large morphological supermatrix for paired fins retrieved from the Phenoscape KB and a Teleostei species-level tree from the OpenTree. By implementing these solutions, data for the pectoral fin and the pelvic fin were successfully mapped to the Teleostei species-level tree, which constituted 38,419 taxa. The computational solutions were assembled into a generic bioinformatics pipeline, 'PhenTree', which can be used to integrate any trait matrix that contains a single anatomical character retrieved from the Phenoscape KB with any phylogenetic tree obtained from the Open Tree. Using the pipeline, ancestral state reconstructions that were usually limited to a smaller number of taxa can now be performed at a larger scale for large taxonomic groups, such as Teleostei, which helps to identify useful evolutionary patterns. In an era where large quantities of biological data are generated using next-generation methods and 'big data analysis' is essential for identifying biological patterns, ancestral state reconstructions must move into the realm of the big data analysis. This is enabled by the computational methods in the pipeline that automate the integration process and remove any manual modifications required, which is essential to perform large-scale ancestral state reconstructions.

The first challenge when performing large-scale ancestral state reconstructions is finding reliable data sources to retrieve large trait data matrices and phylogenetic trees. These were lacking until recently when data sources such as the Phenoscape KB and the Open Tree emerged. The ability of the Phenoscape KB to generate large synthetic morphological supermatrices by combining smaller published matrices is critical for macroevolutionary studies.



When performing large-scale ancestral state reconstructions, it is important to integrate the maximum possible amount of trait data with the phylogenetic tree. This can be achieved through two main approaches: (1) extending the original data volume by computational methods, (2) minimizing the data loss during the integration. The anatomical data in the Phenoscape KB are tagged with Uberon anatomy ontology terms, which can be used to extend the original data volume by ontology-based inference (Dececchi, et al., 2015; Jackson, et al., 2018). In cases where authors do not directly state the presence and absence of an anatomical entity such as the pectoral fin but make indirect statements about the anatomical parts, ontology relationships are used to infer the presence or the absence of the original anatomical entity (Fig. 3.1). It appears that the contribution of inference is significant compared to direct assertions in terms of the data volume. For instance, in the morphological supermatrix used for this work, 3,044 cells (1,511 for pectoral fin and 1,533 for pelvic fin) with species-level data are inferred only compared to the 494 (150 for pectoral fin and 344 for pelvic fin) cells with direct author assertions (Fig. 3.5). The impact of data extension through inference on the ancestral state reconstructions for the pectoral fin and the pelvic fin is illustrated in Fig. 3.6. It is evident how the inference (shown in light blue) extends mapped data in the Teleostei species-level tree, leading to a more complete analysis of the evolution of paired fins.

The data propagation algorithm, which propagates data from families and genera to corresponding species without data, is another computational approach that extended the volume of the data. The supermatrices retrieved from Phenoscape KB contain data that are annotated to higher-level taxa. For example, the supermatrix obtained for this work contained pectoral or pelvic fin data for 182 families and 119 genera (after removing higher-level taxa with ‘0&1’ states), which would have been lost if they were not propagated to the corresponding species.

Although there were data annotated to taxa above the family level (e.g., orders), they were not propagated due to the increasing expectation of evolutionary changes in character states with increasing divergence time. Furthermore, the apparent polymorphisms and conflict states ('0&1') were removed from higher-level taxa before the propagation (section 3.2.4). In apparent polymorphisms, the '0&1' state labeled for a higher-level taxon such as a genus indicates that some of the species of the genus contain the fin and some lack the fin. Without knowing the exact species which lack or contain the fin, this data cannot be propagated to the species level.

During the propagation, existing species-level data were never replaced by data coming from the families and genera. In some instances, the data at species level conflicted with the data from the corresponding families and genera. For instance, the species, *Moringua edwardsi* is annotated with pectoral fin presence ('1') but the corresponding family, *Moringuidae* is annotated with pectoral fin absence ('0'). These are considered as propagation conflicts and they affect the reliability of the propagation algorithm. However, the percentage of propagation conflicts is very low compared to the number of existing species-level data. In the matrix for the paired fins (2 characters), this was 0.27%, and in the larger matrix that included parts of the paired fins (115 characters), the propagation conflict percentage was 0.60% (Supplementary Table S3.1). Therefore, the reliability of the propagation algorithm can be considered as high. Rather than not using the propagation due to potential propagation conflicts, it can be used to avoid losing a significant proportion of data annotated to families and genera. Not only the propagation can be used to avoid missing data but also it extends the data volume by adding data to species without original annotations. As depicted in Fig. 3.5, the propagation added data to 12,870 cells (8,798 for pectoral fin and 4,072 for pelvic fin) corresponding to species without original data in the morphological supermatrix for the paired fins. There are two main reasons for

the pectoral fin to have more propagated data than the pelvic fin: (1) data were propagated from more families and genera for the pectoral fin (151 families and 97 genera) compared to the pelvic fin (90 families and 71 genera), (2) most families in the VTO that were involved with pectoral fin data propagation are more speciose (e.g., Loricariidae, 899 species; Callichthyidae, 211 species) (Jackson, et al., 2018). The importance of propagated data for ancestral state reconstruction is evident in Fig. 3.6 as propagated data (shown in green) completes a significant portion of the Teleostei species-level tree.

The amount of data extended by the propagation is significantly higher compared to the amount of data extended by the inference (Fig. 3.5). However, when considered together, the amount of data extended by both inference and propagation is substantial as shown in Fig. 3.5 and Fig. 3.6, and essential when integrating large trait data with large phylogenetic trees. In the final output matrix that was integrated with the Teleostei species-level tree, the missing data percentage was reduced from 98% to 85.9% using the inference, and further reduced from 85.9% to 34.8% using the propagation. The PhenTree pipeline implements the propagation algorithm on any trait matrix for any single anatomical entity retrieved from the Phenoscope KB and enables the user to differentially visualize the distribution of propagated and inferred data in the species-level tree from the Open Tree. This is important because despite the propagation and inference extending the original data volume, they are still extended by computational methods and the reliability may not be up to par with asserted data that are based on species specimens. Therefore, the user should have the option to distinguish between asserted, inferred and propagated data types after mapping them to the phylogenetic tree.

Reconciliation of taxon names is another important challenge associated with integrating trait matrices from the Phenoscope KB with phylogenetic trees from the Open Tree as they use

different taxon naming systems (Phenoscape KB: VTO; Open Tree: NCBI taxonomy system). This is a common challenge when integrating large data from different sources that use different naming systems. Taxon name reconciliation is an active research area, and current methods frequently use the taxon name as the integrative unit (Patterson, 2003), which introduces several challenges, such as resolving synonyms, abbreviations, misspellings, and handling improper naming syntax (Cranston, et al., 2014; Patterson, et al., 2016). Furthermore, a single taxon name can belong to multiple tips within the same taxonomy, which is identified as homonyms (Rees and Cranston, 2017). There are several solutions developed for the reconciliation of taxon names. There are online servers, such as the Taxonomic Name Resolution Service (TNRS), which act as scientific name repositories that aggregate data from different sources (Boyle, et al., 2013). Furthermore, there are software, such as the toolkit distributed by the Global Names Architecture (GNA), which can perform taxon name reconciliations (Patterson, et al., 2016). Furthermore, when the Open Tree integrates multiple source taxonomies to build the Open Tree taxonomy, the taxon names are matched from individual taxonomic sources to identify the correct taxonomic name (Redelings and Holder, 2017; Rees and Cranston, 2017). Unfortunately, the VTO, which the taxon names from the Phenoscape KB are based on, does not support the aforementioned taxon name reconciliation solutions, thus required the development of an efficient taxon name reconciliation method (section 3.2.7). The taxonomy ID can be used as an alternative for the taxon name (Thomson and Shaffer, 2010). However, depending solely on NCBI taxonomy IDs for reconciliation was inefficient because a large number of VTO taxa (9,522) in the final output matrix did not have references to NCBI taxonomy IDs. Therefore, an algorithm was developed that first uses NCBI taxonomy IDs and then uses taxon names for the reconciliation. This improved the number of matches compared to those that independently use NCBI taxonomy IDs

and taxon names. However, there were still some mismatches due to various reasons. For example, some of the taxa in the VTO are extinct and not found in the Open Tree and some taxa in the VTO contain unconventional names (Supplementary Table S3.2). To reduce these mismatches, VTO must either use the aforementioned taxon name reconciliation solutions or pre-process the taxa names to keep only the standard names. Reconciliation of taxon names still remains a challenge. More efficient solutions must be developed, and existing solutions must be improved in the future to accommodate emerging data recourses that can be used for macroevolutionary analyses.

When the Phenoscope KB aggregates multiple matrices to generate morphological supermatrices, data conflicts can arise due to different statements made by different authors regarding the same anatomical phenotype. The ability to isolate the conflicts is critical to further investigate the reason for the conflicts, which gives a better understanding of the anatomical entities in question and may lead to useful discoveries. They may be due to errors in author statements, observing different specimens of the same species, or due to different interpretations of the observations. Because ontology-based inference is used in the Phenoscope KB supermatrices, a conflict can arise between an inferred state and an asserted state for the same entity, which is the common conflict type reported for the supermatrices (Dececchi, et al., 2015). In the supermatrix generated for this work, only 0.04% of the species-level data (6 of 16,408 populated cells) were conflicted (excluding actual polymorphisms), and all of them were between asserted and inferred states. For example, one author (Nelson, 2006) asserted the absence of the pelvic fin in the eel catfish, *Channallabes apus* and another author (De Pinna, 1993) described the thickness of the first pelvic fin ray of the same species. The first pelvic fin ray is a part of the pelvic fin; hence, the presence of the pelvic fin ray leads to the inference of

the presence of the pelvic fin, which conflicts the asserted absence by Nelson (2006). Similar to this conflict, identification of other conflicts is important for further investigations. The PhenTree pipeline isolates these conflicts for any single character matrix given the metadata, which is a useful feature to broaden the knowledge about the anatomical entities and correct some potential errors made by authors.

The goal of performing ancestral state reconstructions is to study the evolution of characters by investigating their ancestral states. For example, questions such as how often, and in which taxa, the paired fins were lost are of utmost importance in ichthyology (Larouche, et al., 2017; Nelson, 1990; Yamanoue, et al., 2010). The automation achieved through the computational solutions developed in this work enabled to perform a macroevolutionary analysis using a large-scale supermatrix and a large species-level tree to study the evolution of paired fins in teleosts and identify taxa where the fins were regained after a loss (see Jackson, et al. (2018) for details).

As demonstrated, performing ancestral state reconstructions at a larger scale enables to identify important evolutionary patterns and points to certain phylogenetic clades that need more investigation. Although the computational solutions developed in this work were used to integrate a supermatrix for paired fins with a Teleostei species-level tree to study the evolution of paired fins, these methods can be applied to study the evolution of any anatomical entity available in the Phenoscape KB using the PhenTree pipeline (Fernando, et al., 2017). Reusability of computational solutions is important in bioinformatics that is at times taken for granted. The PhenTree pipeline ensures that the computational algorithms developed to solve the challenges associated with the integration of large trait matrices with large phylogenetic trees are reusable to study more evolutionary hypotheses. However, still, some challenges remain for the future,

especially relating to the visualization of large phylogenetic trees. Visualization and manipulation of large phylogenetic trees are challenging using current phylogenetic software (Harmon, et al., 2013). Although software such as Mesquite enables manipulation of large phylogenetic trees, some functions require manual modifications, which is cumbersome at a large scale (Jackson, et al., 2018). The next-generation tools must facilitate easier navigation and support more complex evolutionary analyses (Gruenstaeudl, 2016). With projects such as Arbor (Harmon, et al., 2013) developing for visualizations and comparative analyses of large phylogenetic trees, the evolutionary biology field seems to be moving in the right direction. Therefore, the PhenTree pipeline is a timely addition to the bioinformatics tools that perform macroevolutionary analyses and the computational solutions included in the pipeline will benefit future macroevolutionary studies.

### **3.5 Conclusion**

This work focused on understanding and solving computational challenges associated with the integration of large trait matrices with large phylogenetic trees using a morphological supermatrix for paired fins retrieved from the Phenoscope KB and a Teleostei species-level tree retrieved from the Open Tree. This work showed the importance of data sources such as the Phenoscope KB and the Open Tree for performing macroevolutionary studies. Importantly, it was clear that computational solutions focused on extending the original data volume and minimizing the data loss were required for the integration. For example, the data propagation algorithm, developed during this work, was able to further extend the data in the morphological supermatrix that was originally extended through ontology-based inference, which was critical for the analysis. The combined method that uses both the NCBI ID and the taxon name for taxon

name reconciliation was able to minimize the data loss during the integration. This work showcased the importance of such computational techniques for macroevolutionary studies. Most importantly, the computational methods developed in this work are now available as a reusable bioinformatics pipeline (PhenTree pipeline), which can be used to integrate large-scale data for any anatomical character obtained from the Phenoscape KB with a large phylogenetic tree from the Open Tree. The evolution of a single anatomical character, such as the pectoral fin, can be treated as a single evolutionary hypothesis. Using the PhenTree pipeline, the evolution of each anatomical character with data in the Phenoscape KB can be studied, broadening the number of evolutionary hypothesis that can be made, which is important for performing macroevolutionary studies in the era of the big data analysis.



## Tables

Table 3.1. Percentage of missing data before and after data propagation. The table contains the change in the percentage of missing data before propagation in the pre-processed matrix compared to after propagation in the final output matrix. Missing percentages are relative to the total number of species in the final output matrix (12,582 species; 25,164 cells).

	Cells with data for pectoral fin	Cells with data for pelvic fin	Total populated cells	Percentage of missing data in the final output matrix
Before propagation (Pre-processed matrix)	1,661	1,877	3,538	85.9%
After propagation (Final output matrix)	10,459	5,949	16,408	34.8%



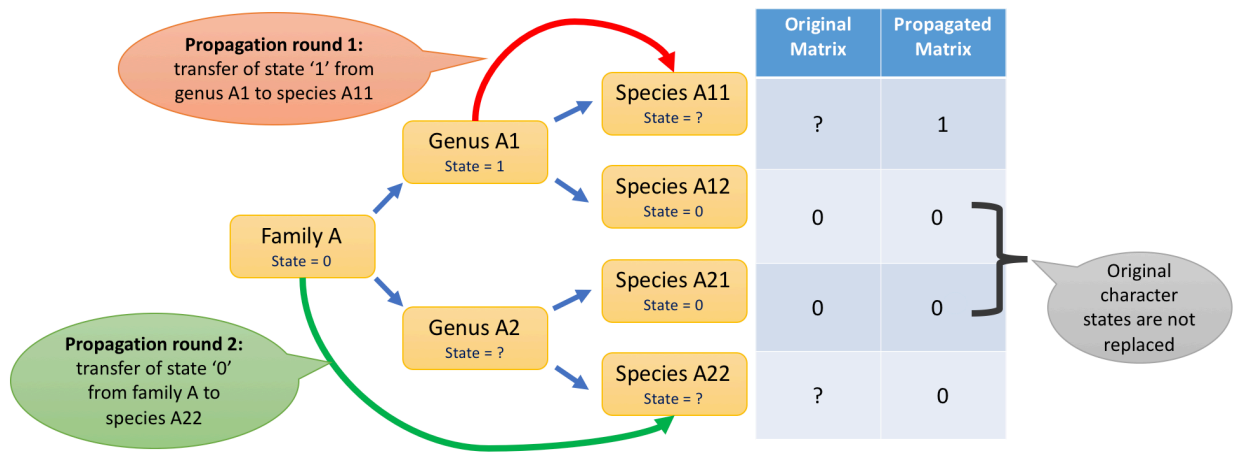


Figure 3.2. A schematic representation of the propagation algorithm. During the first iteration (red arrow), data are propagated from genera to corresponding species. For instance, state 1 from Genus A1 is propagated to Species A11, which initially lacked data. During the second iteration (green arrow), data are propagated from families to the remaining species with missing data (Species A22). The character states of the species with existing data are not modified by the propagation during each iteration. For example, the character states of species that had original data (Species A12 and Species A21) are not replaced during first iteration and character states propagated from genera during the first iteration (Species A11) are not replaced during the second iteration.

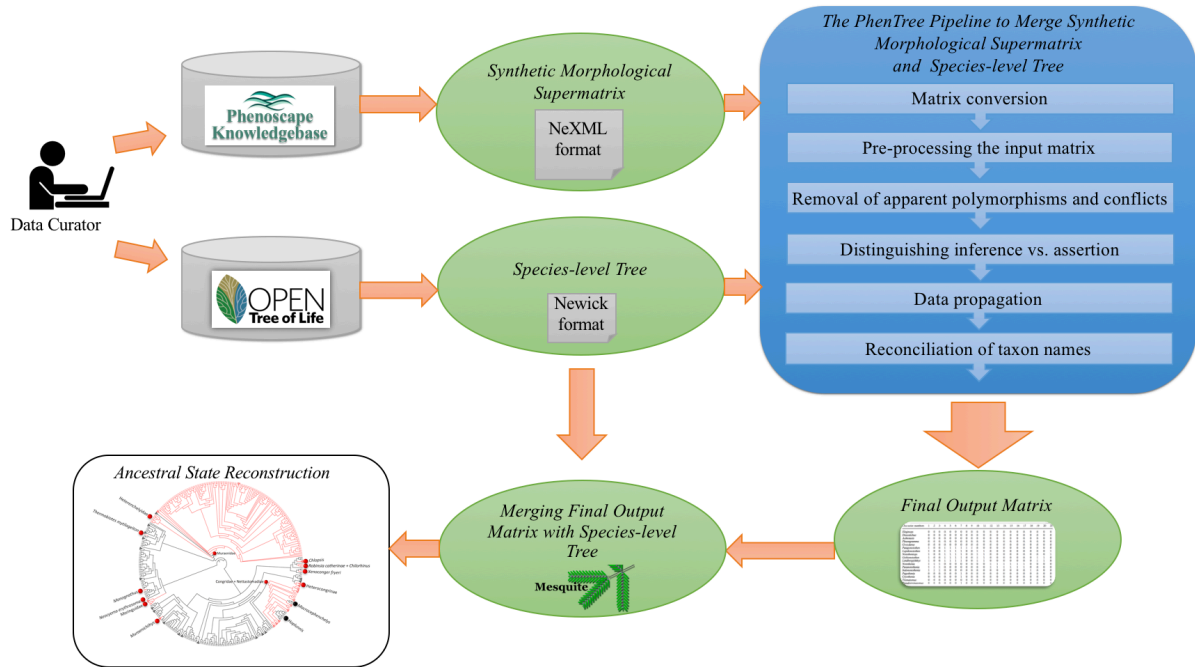


Figure 3.3. The general workflow for integrating a synthetic morphological supermatrix retrieved from the Phenoscape Knowledgebase with a species-level tree obtained from the Open Tree of Life to be used for ancestral state reconstruction. The PhenTree pipeline (shown in blue) converts the supermatrix step by step to a version that can be merged with the species-level tree. This figure was adapted from Jackson, et al. (2018).



# PhenTree

[Home](#)

[Tutorial](#)

[About](#)

[Biocombs.org](#)

This pipeline integrates a phenotypic character matrix generated from Phenoscape Knowledgebase (KB) with a phylogeny from the Open Tree of Life.

User Input:

Two user inputs are required for the pipeline: a synthetic character matrix downloaded from Phenoscape KB and a synthetic phylogeny downloaded from the Open Tree of Life.

Please retrieve your matrix from Phenoscape using the link below:



Please upload the OntoTrace matrix (.xml):



No file chosen

Please upload the OntoTrace metadata file (.xml) (optional):



No file chosen

Please retrieve your phylogeny from the Open Tree of Life using the link below:



Please input the phylogeny (.tre):



No file chosen

Fig. 3.4. A snapshot of the user interface of the PhenTree web tool. A detailed tutorial is available in the tutorial tab. The tool can be accessed using the following link:

<http://phentree.biocombs.org/>.

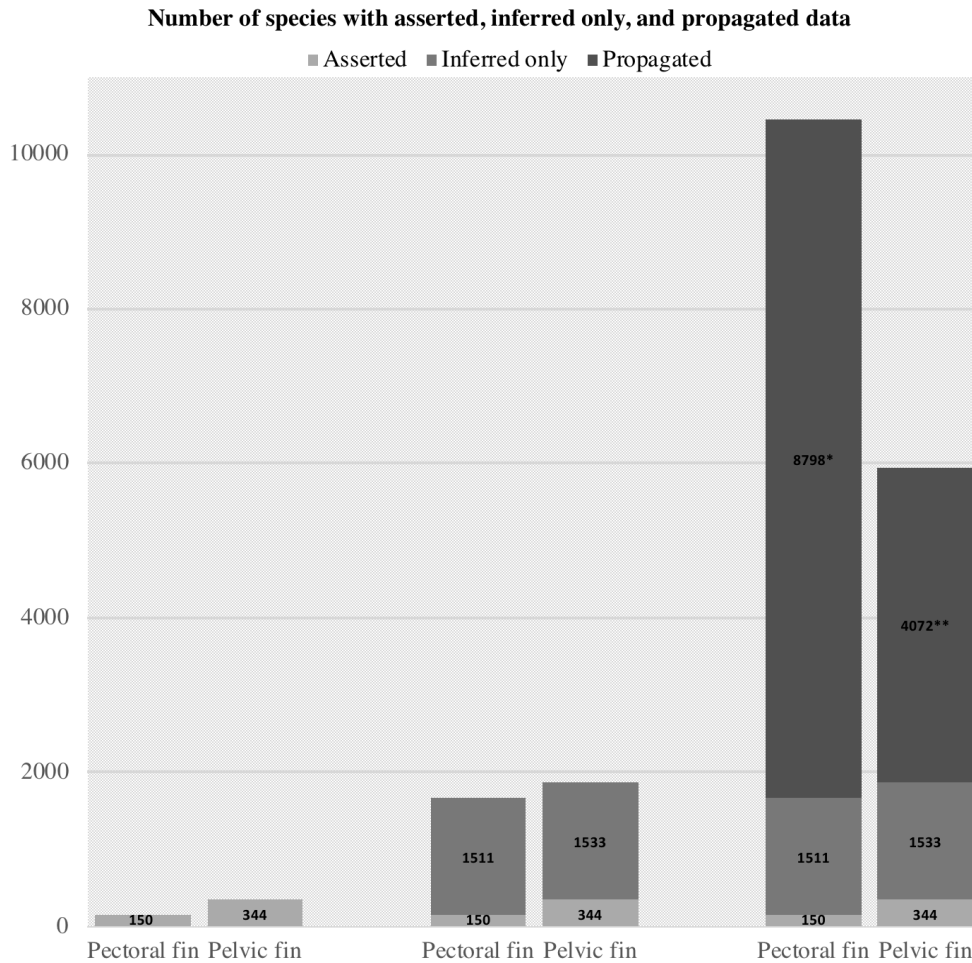
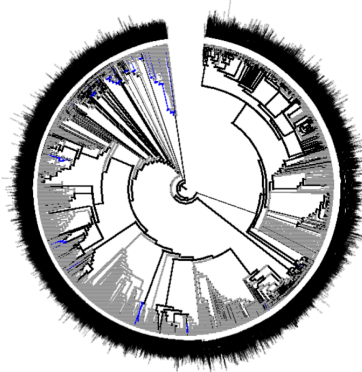


Fig. 3.5. Combined usage of inference and propagation extends morphological data. The bar charts show the number of species with asserted (light gray), inferred only (medium gray), and propagated (dark grey) data for the pectoral fin and pelvic fin. Increase in the number of species with data after inference and then propagation demonstrate the importance of these steps in reducing missing data. \*Of the 8,798 species for which pectoral fin data are propagated from family and genus-level data, 5,077 are propagated from asserted data, and 3,721 are propagated from inferred data. \*\*Of the 4,072 species for which pelvic fin data are propagated from family and genus-level data, 2,906 are propagated from asserted data, and 1,166 are propagated from inferred data.

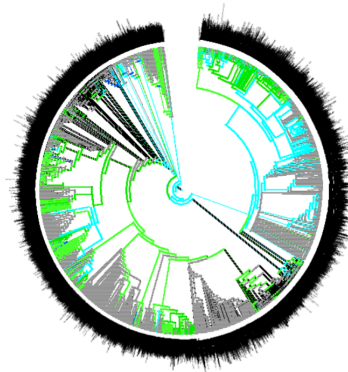
## Pectoral Fin



Asserted data

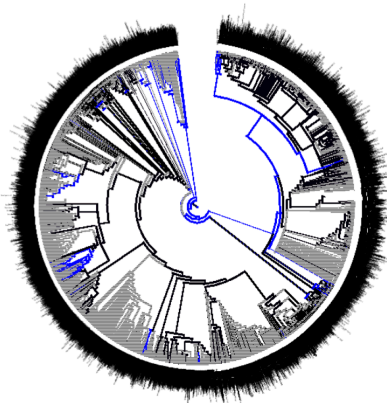


Asserted + Inferred only  
data

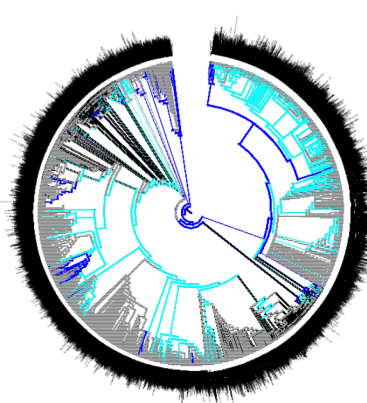


Asserted + Inferred only +  
Propagated data

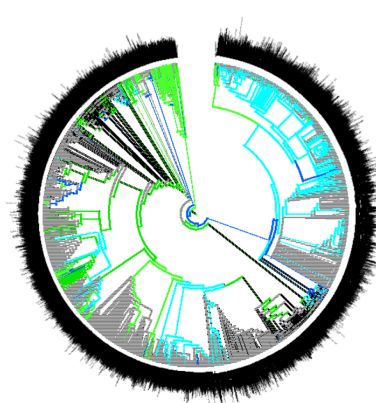
## Pelvic Fin



Asserted data



Asserted + Inferred only  
data



Asserted + Inferred only +  
Propagated data

Figure 3.6. The distribution of asserted (dark blue), inferred only (light blue), and propagated data (green) after performing ancestral state reconstructions for the pectoral and pelvic fins.

## Supplementary Tables

Supplementary Table S3.1. Propagation statistics for the matrix that contained parts of the paired fins (115 characters) for Teleostei. The propagation algorithm was implemented on one character at a time; therefore, the statistics are recorded separately for each character.

Character name	Uberon ID	Number of families contributed to the propagation	Number of genera contributed to the propagation	Number of existing species with data	Number of newly added species	Total number of species after the propagation	Number of propagation conflicts
Pectoral fin radial cartilage	Uberon_2 201586	52	29	974	2294	3268	0
Pelvic fin skeleton	Uberon_0 010711	132	78	2230	8480	10710	8
Mesopterygium element	Uberon_4 300081	11	24	55	1206	1261	0
Pectoral fin distal radial element 2	Uberon_2 102279	11	7	104	1145	1249	0
Pectoral fin spine	Uberon_2 001787	45	13	1045	3916	4961	4
Pectoral fin radial bone	Uberon_2 001586	51	29	960	2292	3252	0
Paired fin skeleton	Uberon_0 010713	181	112	2580	10880	13460	0
Pectoral fin radial skeleton	Uberon_4 440009	52	29	987	2294	3281	0
Paired fin radial skeleton	Uberon_4 300013	76	38	986	1919	2905	0
Branched pectoral fin ray	Uberon_2 001993	11	11	315	1153	1468	0
Pectoral fin proximal radial element	Uberon_2 101587	15	30	475	1236	1711	0
Paired fin spine	Uberon_4 500009	57	14	1085	3341	4426	0
Paired fin radial element	Uberon_1 600006	76	38	986	1919	2905	0
Pelvic fin ray	Uberon_4 300117	93	26	795	6854	7649	36
Pelvic fin lepidotrichium	Uberon_4 000173	109	46	1449	7131	8580	65



Paired fin radial bone	Uberon_1 500006	60	37	959	1760	2719	0
Pectoral fin radial element	Uberon_2 101586	52	29	987	2294	3281	0
Pectoral fin lepidotrichium	Uberon_4 000175	53	36	1595	4216	5811	4
Pelvic fin middle radial bone	Uberon_1 500010	107	26	519	7359	7878	1
Paired fin lepidotrichium	Uberon_4 440011	62	45	1875	3497	5372	0
Paired fin radial cartilage	Uberon_1 700006	41	22	946	1149	2095	0
Lateropterygium	Uberon_2 002077	91	26	732	6829	7561	36
Pectoral fin distal radial element 3	Uberon_2 102280	11	7	104	1145	1249	0
Pelvic fin ray 1	Uberon_2 001776	91	26	777	6829	7606	36
Pectoral fin skeleton	Uberon_0 010710	156	98	2053	10220	12273	15
Pelvic splint	Uberon_2 001788	91	27	596	6846	7442	16
Anterior dentation of pectoral fin spine	Uberon_2 002001	12	10	310	1183	1493	0
Anterior distal serration of pectoral fin spine	Uberon_2 002002	12	7	50	1179	1229	0
Pectoral fin proximal radial bone 4	Uberon_2 002029	11	7	76	1145	1221	0
Pectoral fin proximal radial cartilage	Uberon_2 201587	15	30	462	1236	1698	0
Pectoral fin proximal radial bone 3	Uberon_2 002028	11	10	139	1149	1288	0
Posterior dentation of pectoral fin spine	Uberon_2 002000	12	9	306	1186	1492	0

Pectoral fin distal radial bone 3	Uberon_2 002280	11	7	104	1145	1249	0
Pectoral fin proximal radial bone	Uberon_2 001587	14	30	448	1234	1682	0
Pelvic radial cartilage	Uberon_2 001538	106	42	295	7460	7755	3
Pectoral fin ray	Uberon_4 500007	12	12	422	1156	1578	0
Pectoral fin distal radial cartilage 2	Uberon_2 202279	11	7	104	1145	1249	0
Pectoral fin proximal radial bone 2	Uberon_2 002027	11	7	105	1145	1250	0
Pectoral fin proximal radial bone 1	Uberon_2 002026	14	7	105	1166	1271	0
Metapterygium cartilage	Uberon_4 400000	11	7	28	1145	1173	0
Mesopterygium bone	Uberon_4 300087	11	24	55	1206	1261	0
Pelvic axillary process	Uberon_2 002086	91	26	260	6865	7125	0
Pelvic fin radial bone	Uberon_2 000508	106	42	575	7441	8016	1
Pectoral axillary process	Uberon_2 002087	11	7	78	1145	1223	0
Mesopterygium cartilage	Uberon_1 500007	11	24	55	1206	1261	0
Pectoral fin proximal radial element 1	Uberon_2 102026	14	7	118	1166	1284	0
Pectoral fin distal radial cartilage 3	Uberon_2 202280	11	7	104	1145	1249	0
Pectoral fin proximal radial element 2	Uberon_2 102027	11	7	105	1145	1250	0
Pelvic fin radial skeleton	Uberon_4 440010	121	43	295	7617	7912	3
Pectoral fin proximal radial cartilage 2	Uberon_2 202027	11	7	105	1145	1250	0

Pelvic fin radial element	Uberon_2 100508	121	43	295	7617	7912	3
Pectoral fin proximal radial cartilage 1	Uberon_2 202026	14	7	105	1166	1271	0
Pectoral fin proximal radial cartilage 3	Uberon_2 202028	11	10	139	1149	1288	0
Pelvic fin spine	Uberon_2 002270	111	40	347	7623	7970	17
Pectoral fin ray 1	Uberon_2 001761	12	8	137	1148	1285	0
Pectoral fin ray 2	Uberon_2 001762	11	7	92	1145	1237	0
Pectoral fin proximal radial element 3	Uberon_2 102028	11	10	139	1149	1288	0
Pectoral fin proximal radial element 4	Uberon_2 102029	11	7	76	1145	1221	0
Propterygium element	Uberon_4 300083	12	8	73	1148	1221	0
Propterygium cartilage	Uberon_2 001589	12	8	73	1148	1221	0
Pectoral fin distal radial bone 2	Uberon_2 002279	11	7	104	1145	1249	0
Pectoral fin distal radial element	Uberon_2 101588	11	7	105	1145	1250	0
Pelvic fin ray 6	Uberon_2 001781	91	26	294	6865	7159	0
Pectoral fin distal radial cartilage	Uberon_2 201588	11	7	105	1145	1250	0
Pectoral fin proximal radial cartilage 4	Uberon_2 202029	11	7	76	1145	1221	0
Pectoral fin distal radial bone	Uberon_2 001588	11	7	105	1145	1250	0
Pelvic cartilage	Uberon_4 300016	94	27	285	6928	7213	7
Pectoral fin base	Uberon_4 300147	25	7	31	1709	1740	0
Pelvic fin clasper	Uberon_0 010518	91	26	548	6829	7377	0
Pectoral splint	Uberon_4 300155	12	9	41	1188	1229	0

Metapterygi um element	Uberon_4 300082	11	7	28	1145	1173	0
Rudimentar y pectoral fin ray	Uberon_4 300103	11	7	55	1145	1200	0
Pelvic fin ray 2	Uberon_2 001777	91	26	227	6865	7092	0
Pelvic fin ray 3	Uberon_2 001778	92	26	225	6883	7108	0
Pelvic fin distal radial bone	Uberon_1 500008	91	26	518	6844	7362	0
Pelvic fin distal radial element 1	Uberon_2 101417	91	26	218	6865	7083	0
Pelvic fin distal radial bone 3	Uberon_2 001416	91	26	518	6844	7362	0
Pelvic radial 2 cartilage	Uberon_2 001541	91	26	218	6865	7083	0
Pelvic fin distal radial bone 2	Uberon_2 001415	91	26	518	6844	7362	0
Pelvic fin distal radial bone 1	Uberon_2 001417	91	26	518	6844	7362	0
Ventral marginal cartilage	Uberon_4 300018	91	26	548	6829	7377	0
Pelvic fin distal radial cartilage 2	Uberon_2 201415	91	26	218	6865	7083	0
Propterygiu m bone	Uberon_4 300089	11	7	28	1145	1173	0
Pelvic fin distal radial element 2	Uberon_2 101415	91	26	218	6865	7083	0
Unbranched pelvic fin ray	Uberon_4 500011	91	26	218	6865	7083	0
Pelvic fin basipterygial radial	Uberon_1 500009	91	26	518	6844	7362	0
Clasper plate	Uberon_0 018315	91	26	548	6829	7377	0
Pelvic radial 1 cartilage	Uberon_2 001542	91	26	218	6865	7083	0
Mesenchym e pelvic fin	Uberon_0 003935	91	26	218	6865	7083	0
Pelvic fin distal radial element	Uberon_1 600008	91	26	218	6865	7083	0
Pectoral fin distal radial element 1	Uberon_2 102277	11	7	28	1145	1173	0

Pelvic fin distal radial cartilage 3	Uberon_2 201416	91	26	218	6865	7083	0
Metapterygium bone	Uberon_4 300088	11	7	28	1145	1173	0
Pelvic fin distal radial element 3	Uberon_2 101416	91	26	218	6865	7083	0
Pelvic fin distal radial cartilage 1	Uberon_2 201417	91	26	218	6865	7083	0
Pelvic fin ray 5	Uberon_2 001780	91	26	218	6865	7083	0
Pelvic fin middle radial element	Uberon_1 600010	91	26	218	6865	7083	0
Pectoral fin distal radial cartilage 1	Uberon_2 202277	11	7	28	1145	1173	0
Pectoral fin intermediate radial bone	Uberon_4 200208	11	7	28	1145	1173	0
Pelvic fin ray 7	Uberon_2 001782	91	26	218	6865	7083	0
Pectoral fin fold	Uberon_2 005317	11	7	28	1145	1173	0
Apical ectodermal ridge pelvic fin	Uberon_2 001450	91	26	218	6865	7083	0
Pelvic fin actinotrichium	Uberon_2 000596	91	26	218	6865	7083	0
Pelvic fin ray 4	Uberon_2 001779	91	26	218	6865	7083	0
Pelvic radial 3 cartilage	Uberon_2 001540	91	26	218	6865	7083	0
Unbranched pectoral fin ray	Uberon_4 500010	11	7	28	1145	1173	0
Pectoral fin actinotrichium	Uberon_2 000544	11	7	28	1145	1173	0
Basal scute	Uberon_4 200165	19	1	3	952	955	0
Pectoral fin distal radial bone 1	Uberon_2 002277	11	7	28	1145	1173	0
Mesenchyme pectoral fin	Uberon_0 003934	11	7	28	1145	1173	0
Pectoral fin ray 7	Uberon_2 001767	11	7	28	1145	1173	0

Pectoral fin ray 6	Uberon_2001766	11	7	28	1145	1173	0
Pectoral fin ray 5	Uberon_2001765	11	7	28	1145	1173	0
Pectoral fin ray 4	Uberon_2001764	11	7	28	1145	1173	0
Pectoral fin ray 3	Uberon_2001763	11	7	28	1145	1173	0

Supplementary Table S3.2. The species that were mismatched between the final output matrix that contains Vertebrate Taxonomy Ontology (VTO) taxon names and the Teleostei species-level tree from the Open Tree of Life that contains taxon names based on the NCBI taxonomy system.

The number of mismatched taxa that is extinct: 72	The number of mismatched taxa that has sp. in name (improper naming): 362	The remaining number of mismatches taxa: 940
<p><i>Saurorhamphus freyeri</i>  <i>Batrachoides nidificans</i>  <i>Joffrichthys triangulpterus</i>  <i>Enchodus brevis</i>  <i>Bolcyrus formosissimus</i>  <i>Phractocephalus nassi</i>  <i>Idiacanthus trispinosus</i>  <i>Astroscoptes countermani</i>  <i>Cretophareodus alberticus</i>  <i>Carangidarum americanus</i>  <i>Corydoras revelatus</i>  <i>Batrachoides antiquior</i>  <i>Idiacanthus bellistriatus</i>  <i>Idiacanthus cameratus</i>  <i>Diodon circumflexus</i>  <i>Socnopaea grandis</i>  <i>Eosolea formosa</i>  <i>Lampanyctus latesulcatus</i>  <i>Austromola angerhoferi</i>  <i>Enchodus gracilis</i>  <i>Hoplopteryx antiquus</i>  <i>Cimolichthys nepaholica</i>  <i>Conger vetustus</i>  <i>Myripristis homopterygius</i>  <i>Tautoga (Protautoga) conidens</i>  <i>Holocentrum pygaeum</i>  <i>Beryx radians</i>  <i>Eosolea bartonensis</i>  <i>Nematonotus bottae</i>  <i>Phareodus encaustus</i>  <i>Conger meridies</i>  <i>Beryx ornatus</i>  <i>Eosolea subglabra</i>  <i>Holocentrum pygmaeum</i>  <i>Palaeolycus dreginensis</i>  <i>Congeris brevior</i>  <i>Voltaconger latispinus</i>  <i>Beryx germanus</i>  <i>Eosolea texana</i>  <i>Phractocephalus acreornatus</i>  <i>Albula dunklei</i>  <i>Brychaetus muelleri</i></p>	<p><i>Glyptothorax sp. (de Pinna 1993)</i>  <i>Genus 2 (Bockmann 1998)</i>  <i>Genus 7 sp. (Bockmann 1998)</i>  <i>Parotocinclus sp. (Britto 2003)</i>  <i>Stegophilus sp. (Schaefer 1990)</i>  <i>Phractura sp. (Mo 1991)</i>  <i>Diplomystes sp. (de Pinna 1993)</i>  <i>Pterodoras sp. (Kailola 2004)</i>  <i>Akysis sp. (de Pinna 1993)</i>  <i>Coregonus sp. (Fink and Fink 1981)</i>  <i>Salmo sp. (Fink and Fink 1981)</i>  <i>Poecilocharax sp. (Fink and Fink 1981)</i>  <i>Hypostomus sp. round snout 1 (Armbruster 2004)</i>  <i>Pareiorhina sp. (Britto 2003)</i>  <i>Clarias sp. (Fink and Fink 1981)</i>  <i>Otothyris sp. (Schaefer 1991)</i>  <i>Orinocodoras sp. (Royero 1999)</i>  <i>Chiloglanis sp. E (Vigliotta 2008)</i>  <i>Astroblepus sp. (Britto 2003)</i>  <i>Farlowella sp. (de Pinna 1993)</i>  <i>Leptodoras sp. (Vigliotta 2008)</i>  <i>Pareiorhina sp. (Armbruster 2004)</i>  <i>Dianema sp. (Kailola 2004)</i>  <i>Oxyropsis sp. (de Pinna 1993)</i>  <i>Imparfinis sp. B (Bockmann 1998)</i>  <i>Acrochordonichthys sp. (de Pinna 1993)</i>  <i>Scopaeocharax sp. (Burns et al 1995)</i>  <i>Pterygoplichthys sp. (Schaefer 1987)</i>  <i>Neosilurus sp. 1 (de Pinna 1993)</i>  <i>Exastilithoxus sp. (Armbruster 2004)</i>  <i>Loricariidae sp. (Mo 1991)</i>  <i>Sturisoma sp. (Schaefer 1987)</i>  <i>Corydoras sp. (Mo 1991)</i>  <i>Hoplomyzon sp. (Friel 1994)</i>  <i>Micralestes cf. elongatus (Zanata and Vari 2005)</i></p>	<p><i>Hippocampus europaeus</i>  <i>Patagonotothen shagensis</i>  <i>Leiocassis similis</i>  <i>Salarias basilisca</i>  <i>Rivulus tocantinensis</i>  <i>Kryptopterus eugeneiatus</i>  <i>Aspidoras aff. poecilus (Britto 2003)</i>  <i>Megalebias wolterstorffi</i>  <i>Rivulus peruanus</i>  <i>Homaloptera smithi</i>  <i>Rivulus siegfriedi</i>  <i>Silurus palavanensis</i>  <i>Hippocampus planifrons</i>  <i>Rivulus uatuman</i>  <i>Dinotopterus jacksoni</i>  <i>Canthigaster rostratus</i>  <i>Liparis quasimodo</i>  <i>Trachycorystes obscurus</i>  <i>Leptocephalus holti</i>  <i>Megalodoras granulatus</i>  <i>Hypsoblennius hentzi</i>  <i>Ctenopoma fasciolatum</i>  <i>Hippocampus chinensis</i>  <i>Platybelone platura</i>  <i>Rivulus sape</i>  <i>Duboisialestes tumbensis</i>  <i>Rivulus amanan</i>  <i>Rivulus jurubatibensis</i>  <i>Odax acroptilus</i>  <i>Peckoltia snethlageae</i>  <i>Farlowella platyrynus</i>  <i>Cichlasoma ornatum</i>  <i>Heterandria litoperas</i>  <i>Isorineloricaria spinosissimus</i>  <i>Synodontis fascipinna</i>  <i>Pseudotrematomus lepidorhinus</i>  <i>Hime purpurissata</i>  <i>Lycodes pacificus</i>  <i>Rivulus elongatus</i>  <i>Hypostomus commersonii</i>  <i>Rivulus montium</i></p>

<p><i>Joffrichthys symmetropterus</i>  <i>Albula eppi</i>  <i>Enchodus venator</i>  <i>Taubateia paraiba</i>  <i>Myctophum americanum</i>  <i>Genartina texana</i>  <i>Albula lapidosa</i>  <i>Brachyplatystoma promagdalenae</i>  <i>Beryx zippei</i>  <i>Conger fornicatus</i>  <i>Rharbichthys ferox</i>  <i>Myripristis leptacanthus</i>  <i>Teratichthys antiquitatus</i>  <i>Ranzania grahami</i>  <i>Ranzania tenneyorum</i>  <i>Lampadena jacksoni</i>  <i>Eosolea claibornensis</i>  <i>Conger sanctus</i>  <i>Channa elliptica</i>  <i>Parenchodus longipterygius</i>  <i>Phareodus testis</i>  <i>Eurypholis pulchellus</i>  <i>Eurypholis boissieri</i>  <i>Beryx microcephalus</i>  <i>Enchodus petrosus</i>  <i>Eosolea aquitanica</i>  <i>Fajumia schweinfurthi</i>  <i>Claibornichthys troelli</i>  <i>Genartina hampshiresis</i>  <i>Enchodus marchesettii</i></p>	<p><i>Garra</i> sp. '<i>Discognathichthys</i>'  (Coburn and Cavender 1992)  <i>Creagrutus</i> sp. (Fink and Fink 1981)  <i>Synodontis</i> sp. (de Pinna 1993)  <i>Roeboides</i> sp. B (Lucena and Menezes 1998)  <i>Ixinandria</i> sp. (Britto 2003)  <i>Glyptothorax</i> sp. (de Pinna 1996)  Genus 12 sp. (Bockmann 1998)  <i>Bunocephalus</i> sp. (de Pinna 1993)  <i>Myoglanis</i> sp. (Bockmann 1998)  <i>Xyliphius</i> cf. <i>melanopterus</i> (Friel 1994)  <i>Zaireichthys</i> sp. B (Vigliotta 2008)  <i>Hypostomus</i> sp. (Schaefer 1987)  <i>Euchilichthys</i> sp. A (Vigliotta 2008)  <i>Garra</i> sp. (Coburn and Cavender 1992)  <i>Astroblepus festae</i> sp. 2 (Mo 1991)  <i>Gyrinocheilus</i> sp. (Fink and Fink 1981)  <i>Neoplecostomus</i> sp. (Schaefer 1990)  <i>Farlowella</i> sp. (Schaefer 1987)  <i>Hypostomus</i> sp. round snout 2 (Armbruster 2004)  <i>Erethistes</i> sp. (de Pinna 1993)  Genus 5 sp. B (Bockmann 1998)  <i>Arius</i> sp. (Royero 1999)  <i>Chiloglanis</i> sp. A (Vigliotta 2008)  <i>Ageneiosus</i> sp. 3 (Mo 1991)  <i>Hoplomyzon</i> n. sp. (Friel 1994)  <i>Tytocharax</i> sp. (Burns et al 1995)  <i>Charax</i> sp. (Buckup 1998)  <i>Hypoptopoma</i> sp. (Schaefer 1991)  <i>Planaltina</i> sp. (Burns et al 1995)  <i>Ernstichthys</i> sp. (de Pinna 1993)  <i>Astroblepus</i> sp. (Fink and Fink 1981)  <i>Loricariichthys</i> sp. (Britto 2003)  <i>Tanganikallabes</i> sp. (de Pinna 1993)  <i>Micralestes</i> cf. <i>acutidens</i> (Zanata and Vari 2005)  <i>Leptodoras</i> sp. (Friel 1994)  <i>Sternarchogiton</i> sp. B (Albert 2001)  <i>Platystacus</i> sp. (Mo 1991)  <i>Argopleura</i> sp. (Weitzman and Fink 1985)  <i>Corydoras barbatus</i> sp. II (Britto 2003)  <i>Pangasius</i> sp. (de Pinna 1993)  <i>Leporinus</i> cf. <i>ecuadorensis</i> (Sidlauskas and Vari 2008)  <i>Carpiodes</i> sp. (Fink and Fink 1981)  <i>Imparfinis</i> sp. A (Bockmann 1998)</p>	<p><i>Nothobranchius mkuziensis</i>  <i>Rivulus rossoi</i>  <i>Hypsidoris farsonensis</i>  <i>Rivulus lanceolatus</i>  <i>Monotrete brevirostris</i>  <i>Doryrhamphus multiannulatus</i>  <i>Tetraodon leiurus</i>  <i>Pseudotrematomus loenbergi</i>  <i>Nannoplecostomus eleonorae</i>  <i>Rivulus micropus</i>  <i>Lipophrys adriaticus</i>  <i>Plicofollis angyropleuron</i>  <i>Leiocassis longispinalis</i>  <i>Procatopus kabae</i>  <i>Procatopus loemensis</i>  <i>Bathophilus metallicus</i>  <i>Gnathagnus egregius</i>  <i>Cyprinus buggenhagii</i>  <i>Xyrichtys woodi</i>  <i>Cottus altaicus</i>  <i>Microphis biocellatus</i>  <i>Astephus resimus</i>  <i>Syngnathus argentatus</i>  <i>Pseudotrematomus scotti</i>  <i>Bassanago albescens</i>  <i>Hemiancistrus itacua</i>  <i>Centromochlus marthae</i>  <i>Microphis ocellatus</i>  <i>Cathorops higuchi</i>  <i>Limnothrissa stappersii</i>  <i>Oligancistrus punctatissimus</i>  <i>Alutera monoceros</i>  <i>Sciades sona</i>  <i>Solea nasuta</i>  <i>Rivulus kayabi</i>  <i>Aseraggodes pavoninus</i>  <i>Malapterurus baarbatus</i>  <i>Neoarius coatesi</i>  <i>Platybelone lovii</i>  <i>Rivulus torrenticola</i>  <i>Isorineloricaria spinosissima</i>  <i>Procatopus cabindae</i>  <i>Simpsonichthys flavicaudatus</i>  <i>Lebias anatoliae</i>  <i>Cryptoheros septemfasciatus</i>  <i>Austrolebias salviai</i>  <i>Lipophrys dalmatinus</i>  <i>Xenobalistes caeruleolineatus</i>  <i>Brachirus salinarum</i>  <i>Simpsonichthys janaubensis</i>  <i>Moringua raitaborus</i>  <i>Andamia expansa</i>  <i>Gila alvordensis</i>  <i>Ariopsis seemanni</i>  <i>Parapteronotus bonapartii</i>  <i>Simpsonichthys constanciae</i></p>
---	--	---



<p><i>Ichthyoborus</i> sp. (Fink and Fink 1981)  <i>Chrysichthys</i> sp. 1 (de Pinna 1993)  <i>Rita</i> sp. (Kailola 2004)  <i>Trichomycteridae</i> sp. (de Pinna 1993)  <i>Cetopsorhamdia</i> sp. n. B (Bockmann 1998)  <i>Akysis</i> sp. 2 (Chen 1994)  <i>Engraulis</i> sp. (Arratia 1999)  <i>Chiloglanis</i> sp. B (Vigliotta 2008)  <i>Erethistes</i> sp. (Chen 1994)  <i>Gymnocorymbus</i> sp. (Fink and Fink 1981)  <i>Brycon</i> sp. (Lucena and Menezes 1998)  <i>Albula</i> sp. (Fink and Fink 1981)  <i>Auchenipterichthys</i> sp. (de Pinna 1993)  <i>Sturisoma</i> sp. (Friel 1994)  <i>Tytocharax</i> sp. A (Weitzman and Fink 1985)  <i>Chaetostoma</i> sp. (Friel 1994)  <i>Astroblepus</i> sp. (Armbruster 2004)  <i>Cetopsidae</i> sp. (de Pinna 1993)  <i>Scleromystax</i> sp. (Britto 2003)  <i>Allothrissops</i> sp. (Arratia 1999)  <i>Genus C species C</i> (Malabarba 1998)  <i>Bunocephalus</i> sp. (de Pinna 1996)  <i>Catostomus</i> sp. (Fink and Fink 1981)  <i>Megalancistrus</i> sp. (Schaefer 1987)  <i>Neoplecostomus</i> sp. (Schaefer 1991)  <i>Ageneiosus</i> sp. 2 (Mo 1991)  <i>Gagata</i> sp. (de Pinna 1993)  <i>Xyliphius</i> cf. <i>lepturus</i> (Friel 1994)  <i>Osteoglossum</i> sp. (de Pinna and Grande 2003)  <i>Rineloricaria</i> sp. (Britto 2003)  <i>Synodontis</i> sp. (Royero 1999)  <i>Ancistrus</i> sp. (Britto 2003)  <i>Rhamdia</i> sp. (Kailola 2004)  <i>Erethistes</i> sp. (de Pinna 1996)  <i>Microlepidogaster</i> sp. (Armbruster 2004)  <i>Bagrichthys</i> sp. 1 (de Pinna 1993)  <i>Akysis</i> sp. 1 (Chen 1994)  <i>Hysteronotus</i> sp. (Weitzman and Menezes 1998)  <i>Xyliphius</i> sp. (Chen 1994)  <i>Auchenipterus</i> sp. (de Pinna 1993)  <i>Tetranematichthys</i> n. sp. (Royero 1999)  <i>Loricaria</i> sp. (Armbruster 2004)  <i>Amphilius</i> sp. A (Vigliotta 2008)</p>	<p><i>Trachelyichthys</i> n. sp. 1 (Royero 1999)  <i>Rivulus glaucus</i>  <i>Hippocampus tuberculatus</i>  <i>Scoloplax baileyi</i>  <i>Selar crumenophthalmus</i>  <i>Chiasmodon lavenbergi</i>  <i>Rivulus ornatus</i>  <i>Synodontis marmorata</i>  <i>Simpsonichthys flagellatus</i>  <i>Chlorophthalmus agassiz</i>  <i>Microphis platyrhynchus</i>  <i>Halidesmus waltairiensis</i>  <i>Rivulus nicoi</i>  <i>Phalloceros malabarai</i>  <i>Rivulus pacificus</i>  <i>Gymnothorax prionodon</i>  <i>Orestias frontosus</i>  <i>Pseudoscopus vityazi</i>  <i>Thalassenchelys coheni</i>  <i>Rivulus caurae</i>  <i>Eomola bimaxillaria</i>  <i>Brachysynodontis batensoda</i>  <i>Soleichthys multifasciatus</i>  <i>Acanthostracion bucephalus</i>  <i>Archaeotetraodon jamestyleri</i>  <i>Tatia creutzbergi</i>  <i>Dinotopterus gigas</i>  <i>Pseudoscopus albeolus</i>  <i>Bothus thompsoni</i>  <i>Rivulus salmonicaudus</i>  <i>Exoglossum maxillingua</i>  <i>Symptericthys politus</i>  <i>Uncisudis longirostris</i>  <i>Epiplatys berkenkampii</i>  <i>Protoclupea chilensis</i>  <i>Knightia bohaiensis</i>  <i>Simpsonichthys magnificus</i>  <i>Scartichthys xiphiodon</i>  <i>Simpsonichthys hellneri</i>  <i>Rivulus intermittens</i>  <i>Rivulus illuminatus</i>  <i>Channa pleurophthalmus</i>  <i>Astyanax incaucus</i>  <i>Paraperca mimaseana</i>  <i>Parasilurus asotus</i>  <i>Salarias reticulatus</i>  <i>Artedidraco loennbergi</i>  <i>Lycodes fasciatus</i>  <i>Cryptobalistes brevis</i>  <i>Protriacanthus gortanii</i>  <i>Odax cyanomelas</i>  <i>Himantolophus rostratus</i>  <i>Rondeletia bicolor</i>  <i>Paranotothenia trigramma</i>  <i>Plotosus anguillaris</i></p>
---	---

<p><i>Bunocephalus n. sp. 4</i> (Friel 1994)  <i>Ageneiosus cf. guianensis</i> (Mo 1991)  <i>Otothyris sp.</i> (Britto 2003)  <i>Phenacorhamdia sp. B</i> (Bockmann 1998)  <i>Exostoma sp.</i> (de Pinna 1996)  <i>Liposarcus sp.</i> (Friel 1994)  <i>Characidium sp.</i> (Fink and Fink 1981)  <i>Phenacorhamdia sp. D</i> (Bockmann 1998)  <i>Bunocephalus n. sp. 5</i> (Friel 1994)  <i>Gladioglanis sp. A</i> (Bockmann 1998)  <i>Leptoglanis sp.</i> (de Pinna 1993)  <i>Roeboides sp. A</i> (Lucena and Menezes 1998)  <i>Humbertia sp.</i> (Arratia 1999)  <i>Zaireichthys sp. A</i> (Vigliotta 2008)  <i>Arius sp.</i> (Kailola 2004)  <i>Pseudobunocephalus cf. bifidus</i> (Friel 1994)  <i>Trichomycterus sp.</i> (Schaefer 1990)  <i>Loricariichthys sp.</i> (Schaefer 1987)  <i>Imparfinis sp. n.</i> (Bockman 1998)  <i>Genus 7</i> (Bockmann 1998)  <i>Anaethalion cf. A. subovatus</i> (Arratia 1999)  <i>Distichodus sp.</i> (Sidlauskas and Vari 2008)  <i>Ageneiosus sp.</i> (Soares-Porto 1998)  <i>Hemiancistrus sp.</i> (Armbruster 2004)  <i>Ageneiosus sp.</i> (de Pinna 1993)  <i>Hemipsilichthys sp.</i> (Armbruster 2004)  <i>Pimelodella sp. n.</i> (Bockmann 1998)  <i>Euchilichthys sp. B</i> (Vigliotta 2008)  <i>Pachythrissops sp.</i> (Arratia 1999)  <i>Centromochlus sp.</i> (Rio Negro) (Royero 1999)  <i>Phenacorhamdia sp. A</i> (Bockmann 1998)  <i>Pterygoplichthys sp.</i> (de Pinna 1993)  <i>Peckoltia sp. big spot</i> (Armbruster 2004)  <i>Chiloglanis sp. C</i> (Vigliotta 2008)  <i>Neosilurus sp. 2</i> (de Pinna 1993)  <i>Chiloglanis sp.</i> (de Pinna 1993)  <i>Chaetostomus sp.</i> (Britto 2003)  <i>Loricariinae sp.</i> (Friel 1994)  <i>Ancistrus sp.</i> (Britz and Hoffman 2006)  <i>Peckoltia sp. 2</i> (Armbruster 2004)</p>	<p><i>Pseudobagrus vachellii</i>  <i>Rivulus litteratus</i>  <i>Lipophrys canevae</i>  <i>Antennarius analis</i>  <i>Gymnelus barsukovi</i>  <i>Rivulus kayapo</i>  <i>Synodus cressseyi</i>  <i>Liosaccus pachygaster</i>  <i>Clarotes macrocephalus</i>  <i>Synodontis melanoptera</i>  <i>Pseudobagrus aurantiacus</i>  <i>Amblydoras bolivarensis</i>  <i>Hemiaris kessleri</i>  <i>Cynopanchax bukobanus</i>  <i>Barbus callensis</i>  <i>Tharsis dubius</i>  <i>Halichoeres chrysotaenia</i>  <i>Rhinecanthus echarpe</i>  <i>Scobinichthys granulatus</i>  <i>Synodontis ornatissima</i>  <i>Hypostomus cordovae</i>  <i>Hypostomus butantanis</i>  <i>Canthidermis maculatus</i>  <i>Chatrabus damaranus</i>  <i>Pagothenia phocae</i>  <i>Protoclupea atacamensis</i>  <i>Rivulus bororo</i>  <i>Sphoeroides hyperostosus</i>  <i>Hexanematichthys mastersi</i>  <i>Archaeotetraodon winterbottomi</i>  <i>Parotocinclus halbothi</i>  <i>Platybelone pierura</i>  <i>Heterandria obliqua</i>  <i>Hypostomus emarginatus 1</i> (Armbruster 2004)  <i>Acentronura breviperula</i>  <i>Pelteobagrus nudiceps</i>  <i>Lophiodes abdituspinus</i>  <i>Tetraodon cochinchinensis</i>  <i>Arius laticutatus</i>  <i>Omobranchus lineolatus</i>  <i>Simpsonichthys chacoensis</i>  <i>Conger dissimilis</i>  <i>Poecilothrissa congica</i>  <i>Bembrops philippinus</i>  <i>Erethistes serratus</i>  <i>Microglanis aff. iberingi</i> (Shibatta 1998)  <i>Rivulus megaroni</i>  <i>Trachinotus bailloni</i>  <i>Synodontis punctulata</i>  <i>Notarius troschelii</i>  <i>Rivulus bahianus</i>  <i>Rivulus monticola</i>  <i>Neotropius khavalchor</i>  <i>Rivulus scalaris</i></p>
--	---

<p> <i>Trachydoras</i> sp. (Vigliotta 2008)  <i>Astroblepus</i> sp. (Schaefer 1990)  <i>Hypoptopoma</i> sp. (Britto 2003)  <i>Hypostomus</i> sp. (de Pinna 1993)  <i>Amblygaster</i> sp. (DiDario 2004)  <i>Thrissops</i> cf. <i>T. subovatus</i> (Arratia 1999)  <i>Hypoptopoma</i> sp. (Armbruster 2004)  <i>Chrysichthys</i> sp. 2 (de Pinna 1993)  <i>Ameiurus</i> sp. (Vigliotta 2008)  <i>Sternopygus</i> sp. (Fink and Fink 1981)  <i>Corydoras</i> n. sp. (Britto 2003)  <i>Thrissops</i> cf. <i>T. formosus</i> (Arratia 1999)  <i>Hemiodus</i> sp. (Fink and Fink 1981)  <i>Phractura</i> sp. (de Pinna 1993)  <i>Scoloplax</i> sp. (de Pinna 1993)  <i>Phractura</i> sp. (Vigliotta 2008)  <i>Astroblepus</i> sp. (de Pinna 1993)  <i>Amphilius</i> sp. B (Vigliotta 2008)  <i>Otocinclus</i> sp. (Schaefer 1987)  <i>Neosilurus</i> sp. (Kailola 2004)  <i>Limatulichthys</i> sp. (Britto 2003)  <i>Kronichthys</i> sp. (Schaefer 1987)  <i>Glyptothorax</i> sp. (Friel 1994)  <i>Trachelyichthys</i> n. sp. 2 (Royero 1999)  <i>Pseudobunocephalus</i> sp. (Friel 1994)  Genus 2 sp. B (Bockmann 1998)  <i>Cetopsidium</i> sp. (de Pinna, Ferraris and Vari 2007)  <i>Loricaria</i> sp. (Britto 2003)  <i>Centrodoras</i> sp. (Vigliotta 2008)  <i>Centromochlus</i> n. sp. (Royero 1999)  <i>Trachycorystes</i> sp. (de Pinna 1993)  <i>Leporinus</i> cf. <i>moralesi</i> (Sidlauskas and Vari 2008)  <i>Chiloglanis</i> sp. D (Vigliotta (2008)  <i>Ameiurus</i> sp. (de Pinna 1993)  <i>Notropis</i> sp. ""sawfin shiner"" (Coburn and Cavender 1992)  <i>Bunocephalus</i> n. sp. 3 (Friel 1994)  <i>Microlepidogaster</i> sp. (Britto 2003)  <i>Leptorhamdia</i> sp. (Bockmann 1998)  <i>Pseudancistrus</i> sp. (Armbruster 2004)  <i>Pseudopimelodus</i> sp. (Friel 1994)  <i>Ageneiosus</i> sp. 1 (Mo 1991)  <i>Synodontis</i> sp. (Vigliotta 2008)  <i>Megalonema</i> sp. (de Pinna 1993)  <i>Loricaria</i> sp. (Schaefer 1987)  <i>Bunocephalus</i> sp. (Mo 1991) </p>	<p> <i>Antennarius senegalensis</i>  <i>Symphurus arabicus</i>  <i>Cirrhilabrus ryukyuensis</i>  <i>Laimosemion ubim</i>  <i>Rivulus dibaphus</i>  <i>Jenynsia pygogramma</i>  <i>Syngnathus hymenolomus</i>  <i>Pseudotrematomus pennellii</i>  <i>Synodontis caudovittata</i>  <i>Tarletonbeania tenua</i>  <i>Simpsonichthys izecksohni</i>  <i>Sciades platypogon</i>  <i>Symphurus sayademalensis</i>  <i>Rivulus romeri</i>  <i>Halicampus crinitus</i>  <i>Hippocampus taeniopterus</i>  <i>Sardinella janeiro</i>  <i>Poeciliopsis letoni</i>  <i>Orthrias tigris</i>  <i>Cichlasoma salvini</i>  <i>Stemonosudis elongatus</i>  <i>Melletes papilio</i>  <i>Hisonotus candombe</i>  <i>Navodon xanthopterus</i>  <i>Anguilla borneensis</i>  <i>Pseudobagrus tokiensis</i>  <i>Pseudobagrus pictus</i>  <i>Pshekhadiodon parini</i>  <i>Synaphobranchus capensis</i>  <i>Cichlasoma beani</i>  <i>Rivulus monikae</i>  <i>Leiuranus cyclorhinus</i>  <i>Chrysichthys furcatus</i>  <i>Tylosurus imperialis</i>  <i>Solea lascaris</i>  <i>Leptolepis coryphaenoides</i>  <i>Parauchenoglanis loennbergi</i>  <i>Poecilia parae</i>  <i>Astyanax novae</i>  <i>Corythoichthys isigakius</i>  <i>Hippocampus villosus</i>  <i>Mastacembelus flavomarginatus</i>  <i>Spectrolebias reticulatus</i>  <i>Epiplatys azureus</i>  <i>Electrona subasper</i>  <i>Cottus haemusi</i>  <i>Zenopsis nebulosa</i>  <i>Phaenomonas forsteri</i>  <i>Hippocampus dahli</i>  <i>Platybelone trachura</i>  <i>Hippichthys cyanospilus</i>  <i>Lipophrys velifer</i>  <i>Pseudobagrus wittenburgii</i>  <i>Cretatricanthus guidottii</i>  <i>Mystriophis porphyreus</i>  <i>Rivulus speciosus</i> </p>
---	--

<p><i>Tytocharax</i> sp. B (Weitzman and Fink 1985)  cf. <i>Arius macrorhynchus</i> (Kailola 2004)  <i>Kronichthys</i> sp. 1 (Armbruster 2004)  <i>Platydoras</i> sp. (Royero 1999)  <i>Bunocephalus</i> sp. (Fink and Fink 1981)  <i>Hemisorubim</i> sp. (de Pinna 1993)  Genus 5 (Bockmann 1998)  <i>Pterodoras</i> sp. (Royero 1999)  <i>Pseudopimelodus</i> sp. (de Pinna 1993)  <i>Bunocephalus</i> sp. 3 (Chen 1994)  <i>Farlowella</i> sp. (Britto 2003)  <i>Astroblepus festae</i> sp. 3 (Mo 1991)  <i>Microsynodontis</i> sp. (Vigliotta 2008)  <i>Acentronichthys</i> sp. n. B (Bockmann 1998)  <i>Bunocephalus</i> n. sp. 2 (Friel 1994)  <i>Pseudoloricaria</i> sp. (Britto 2003)  <i>Ictalurus</i> sp. (Vigliotta 2008)  <i>Clarias</i> sp. (de Pinna 1993)  <i>Bunocephalus</i> sp. 2 (Chen 1994)  <i>Phenacorhamdia</i> sp. C (Bockmann 1998)  <i>Neoplecostomus</i> sp. (de Pinna 1993)  <i>Microlepidogaster</i> sp. (Schaefer 1991)  <i>Entomocorus</i> sp. (de Pinna 1993)  Genus 1 (Royero 1999)  <i>Leiocassis</i> sp. (Mo 1991)  <i>Megalops</i> sp. (Fink and Fink 1981)  <i>Austroglanis</i> sp. (de Pinna 1993)  <i>Citharinus</i> sp. (Sidlauskas and Vari 2008)  <i>Neoplecostomus</i> sp. (Britto 2003)  <i>Liobagrus</i> cf. <i>marginatus</i> (Mo 1991)  <i>Decapogon</i> sp. (Schaefer 1987)  Genus 5 sp. A (Bockmann 1998)  <i>Trachelyopterus</i> sp. (Vigliotta 2008)  <i>Schizolecis</i> sp. (Britto 2003)  <i>Corydoras barbatus</i> sp. I (Britto 2003)  <i>Ictalurus</i> sp. (Fink and Fink 1981)  <i>Hydrocynus</i> cf. <i>brevis</i> (Zanata and Vari 2005)  <i>Esox</i> sp. (Fink and Fink 1981)  <i>Clarias</i> sp. (Britz and Hoffman 2006)  <i>Osteogeneiosus</i> sp. (Mo 1991)  <i>Arius</i> sp. (Schaefer 1990)</p>	<p><i>Microphis leiaspis</i>  <i>Ostracion clippertonense</i>  <i>Mola chelonopsis</i>  <i>Ostichthys sufensis</i>  <i>Rivulus corpulentus</i>  <i>Synodus houlti</i>  <i>Pseudotrematomus hansonii</i>  <i>Lycodes colletti</i>  <i>Ascalabos voithii</i>  <i>Brachypetersius gabonensis</i>  <i>Lacustricola lualabaensis</i>  <i>Creedia bilineatus</i>  <i>Ariopsis assimilis</i>  <i>Scartichthys fernandezensis</i>  <i>Rivulus zygonectes</i>  <i>Dinotopterus atribranchus</i>  <i>Sciades felis</i>  <i>Patagonotothen occidentalis</i>  <i>Doras fimbriatus</i>  <i>Leptocephalus thorianus</i>  <i>Pseudobagrus ransonnettii</i>  <i>Alticus aldabraensis</i>  <i>Zungaro luetkeni</i>  <i>Compsaraia compsus</i>  <i>Gordichthys conquensis</i>  <i>Neoarius velutinus</i>  <i>Astephus antiquus</i>  <i>Dexillus muelleri</i>  <i>Hime formosanus</i>  <i>Hisonotus leptochilus</i>  <i>Bothrocara tanakae</i>  <i>Brycinus</i> aff. <i>nurse</i> (Zanata and Vari 2005)  <i>Selene setipinnis</i>  <i>Hemipsilichthys?</i> (Armbruster 2004)  <i>Synodontis leoparda</i>  <i>Procatopus lamberti</i>  <i>Regalecus pacificus</i>  <i>Rivulus rutilicaudus</i>  <i>Halieutopsis nasuta</i>  <i>Lycoptera davidi</i>  <i>Zebrias japonica</i>  <i>Halichoeres kneri</i>  <i>Peckoltia arenaria</i>  <i>Brachirus selheimi</i>  <i>Phoxocampus kampeni</i>  <i>Cottus kuznetzovi</i>  <i>Plotosus brevibarbus</i>  <i>Rhambdella gilli</i>  <i>Vladichthys gloverensis</i>  <i>Zebrias cochinchensis</i>  <i>Plancterus zebrinus</i>  <i>Austrolebias vasferreirai</i>  <i>Channichthys normani</i>  <i>Rivulus xanthonotus</i></p>
---	--

<p> <i>Cnidoglanis</i> sp. (Mo 1991)  <i>Leporinus</i> cf. <i>niceforoi</i> (Sidlauskas and Vari 2008)  <i>Hybognathus</i> sp. (Coburn and Cavender 1992)  <i>Amblydoras</i> sp. (Friel 1994)  <i>Akysis</i> sp. 1 (de Pinna 1996)  <i>Tridentopsis</i> sp. (Schaefer 1990)  <i>Microglanis</i> sp. (de Pinna 1993)  <i>Heteropneustes</i> sp. (de Pinna 1993)  <i>Brycon</i> sp. (Fink and Fink 1981)  <i>Lipopterichthys</i> sp. (Britto 2003)  <i>Cetopsorhamdia</i> sp. n. C (Bockmann 1998)  <i>Trachyglanis</i> sp. (de Pinna 1993)  <i>Opsariichthys</i> sp. (Cavender and Coburn 1992)  <i>Imparfinis</i> sp. cf. <i>minutus</i> (Bockman 1998)  <i>Kronichthys</i> sp. 2 (Armbruster 2004)  <i>Ernstichthys</i> sp. (de Pinna 1996)  <i>Pseudopimelodus</i> sp. 2 (Shibatta 1991)  <i>Distichodus</i> sp. (Fink and Fink 1981)  <i>Pimelodella</i> sp. B (Bockmann 1998)  <i>Hypostomus</i> sp. (Britto 2003)  <i>Panaque</i> sp. (de Pinna 1993)  <i>Mesoborus</i> sp. (Fink and Fink 1981)  <i>Bunocephalus</i> cf. <i>knerii</i> (Friel 1994)  <i>Rhamdiopsis</i> sp. n. A (Bockmann 1998)  <i>Leporinus</i> cf. <i>fasciatus</i> (Sidlauskas and Vari 2008)  <i>Eutropius</i> sp. (de Pinna 1993)  <i>Hoplomyzon</i> cf. <i>papillatus</i> (Friel 1994)  <i>Ctenolucius</i> sp. (Fink and Fink 1981)  <i>Bunocephalus</i> sp. 1 (Chen 1994)  <i>Ernstichthys</i> sp. (Friel 1994)  <i>Pseudancistrus</i> sp. gold spot (Armbruster 2004)  <i>Hemidoras</i> sp. (Vigliotta 2008)  <i>Henonemus</i> sp. (Schaefer 1990)  <i>Hypoptopoma</i> sp. (Schaefer 1987)  <i>Malapterurus</i> sp. (Vigliotta 2008)  <i>Adontosternarchus</i> sp. (Fink and Fink 1981)  <i>Helogenes</i> sp. (Mo 1991)  <i>Abramites</i> sp. (Fink and Fink 1981)  <i>Lycoptera</i> <i>davidi</i> cf. <i>L. tokunagai</i> (Arratia 1999) </p>	<p> <i>Glyptothorax</i> <i>sausii</i>  <i>Gila</i> <i>boraxobius</i>  <i>Paralipophrys</i> <i>trigloides</i>  <i>Cichlasoma</i> <i>istlanum</i>  <i>Syngnathus</i> <i>pellegrini</i>  <i>Gagata</i> <i>gasawyuh</i>  <i>Domeykos</i> <i>profetaesis</i>  <i>Leptolepides</i> <i>haerteisi</i>  <i>Zebrasoma</i> <i>veliferum</i>  <i>Notropis</i> <i>dorsalis</i>  <i>Halichoeres</i> <i>annularis</i>  <i>Lipophrys</i> <i>bauchotae</i>  <i>Clupeonella</i> <i>muhlsi</i>  <i>Nemacheilus</i> <i>insignis</i>  <i>Antennarius</i> <i>ocellatus</i>  <i>Sinopangasius</i> <i>semicultratus</i>  <i>Tachysurus</i> <i>mica</i>  <i>Parapercis</i> <i>tetracanthus</i>  <i>Zaireichthys</i> <i>rhodesiensis</i>  <i>Solea</i> <i>orientalis</i>  <i>Lasiancistrus</i> <i>nationi</i>  <i>Rivulus</i> <i>uakti</i>  <i>Simpsonichthys</i> <i>igneus</i>  <i>Spinacanthus</i> <i>cuneiformis</i>  <i>Trichomycterus</i> <i>anhanga</i>  <i>Tautoga</i> (<i>Protautoga</i>)  <i>Antennarius</i> <i>scriptissimus</i>  <i>Synodontis</i> <i>guttata</i>  <i>Gelanoglanis</i> <i>tracieso</i>  <i>Paruroconger</i> <i>drachi</i>  <i>Chlorophthalmus</i> <i>productus</i>  <i>Ophisoma</i> <i>prorigerum</i>  <i>Caulophryne</i> <i>bacescui</i>  <i>Epiplatys</i> <i>infracasciatus</i>  <i>Mystus</i> <i>elongatus</i>  <i>Aphyosemion</i> <i>jeanpoli</i>  <i>Antennarius</i> <i>drombus</i>  <i>Neoarius</i> <i>latirostris</i>  <i>Rineloricaria</i> <i>rupestris</i>  <i>Xyrichtys</i> <i>pentadactylus</i>  <i>Ophichthus</i> <i>unicolor</i>  <i>Xyrichtys</i> <i>bimaculatus</i>  <i>Comephorus</i> <i>baikalensis</i>  <i>Ancistrus</i> <i>albihoai</i>  <i>Rivulus</i> <i>planaltinus</i>  <i>Kali</i> <i>caribbaea</i>  <i>Notarius</i> <i>phrygiatus</i>  <i>Rivulus</i> <i>lazzarotoi</i>  <i>Rivulus</i> <i>vittatus</i>  <i>Synaptura</i> <i>nigra</i>  <i>Platybelone</i> <i>argalus</i>  <i>Dinotopterus</i> <i>filicibarbis</i>  <i>Pseudobagrus</i> <i>nitidus</i>  <i>Rivulus</i> <i>cladophorus</i>  <i>Hoplostethus</i> <i>racurictus</i>  <i>Danio</i> <i>erythromicron</i> </p>
--	--

<p> <i>Genus sp. 1</i> (Royero 1999)  <i>Synodontis ? sp.</i> (Royero 1999)  <i>Compasaria sp. 1</i> (Albert 2001)  <i>Propimelodus sp.</i> (Mo 1991)  <i>Auchenipterus sp. (cuyuni)</i> (Royero 1999)  <i>Parotocinclus sp.</i> (Schaefer 1991)  <i>Lycoptera cf. L. sinensis</i> (Arratia 1999)  <i>Pimelodella sp. A</i> (Bockmann 1998)  <i>Peckoltia sp. 1</i> (Armbruster 2004)  <i>Harttia sp.</i> (Armbruster 2004)  <i>Wallago sp.</i> (Mo 1991)  <i>Planiloricaria sp.</i> (Britto 2003)  <i>Ompok sp.</i> (de Pinna 1993)  <i>Akysis sp.</i> (Friel 1994)  <i>Apareiodon sp.</i> (Fink and Fink 1981)  <i>Bunocephalus sp.</i> (Friel 1994)  <i>Euchilichthys sp. C</i> (Vigliotta 2008)  <i>Synodontis sp.</i> (Mo 1991)  <i>Pseudotothyris sp.</i> (Schaefer 1991)  <i>Chasmocranus sp.</i> (de Pinna 1993)  <i>Corydoras sp.</i> (de Pinna 1993)  <i>Trachelyopterus sp.</i> (de Pinna 1993)  <i>Pseudopimelodus sp. 1</i> (Shibatta 1991)  <i>Rhinodoras sp.</i> (Vigliotta 2008)  <i>Pristigaster sp.</i> (DiDario 1999)  <i>Bagrichthys sp. 2</i> (de Pinna 1993)  <i>Akysis sp. 3</i> (de Pinna 1996)  <i>Phreatobius sp. 2</i> (de Pinna 1993)  <i>Diplomystes sp.</i> (Vigliotta 2008)  <i>Gephyrocharax sp.</i> (Weitzman and Fink 1985)  <i>Silurichthys sp.</i> (Bornbusch 1991)  <i>Lasiancistrus sp.</i> (Armbruster 2004)  <i>Acrobrycon sp.</i> (Burns et al 1995)  <i>Amaralia n. sp.</i> (Friel 1994)  <i>Astephus sp.</i> (Lundberg 1992)  <i>Ancistrus sp.</i> (Armbruster 2004)  <i>Micralestes sp.</i> (Zanata and Vari 2005)  <i>Vandellia sp.</i> (Fink and Fink 1981)  <i>Crossoloricaria sp.</i> (Armbruster 2004)  <i>Tatia sp.</i> (de Pinna 1993)  <i>Hypostomus sp.</i> (Schaefer 1990)  <i>Astroblepus festae sp. 1</i> (Mo 1991)  <i>Panaque sp.</i> (Friel 1994)  <i>Chiloglanis sp. F</i> (Vigliotta 2008)  <i>Entomocorus n. sp.</i> (Royero 1999)  <i>Chiloglanis sp. G</i> (Vigliotta 2008)  <i>Pterobunocephalus sp.</i> (Friel 1994)  <i>Callichthys sp.</i> (Mo 1991) </p>	<p> <i>Formosania lacustre</i>  <i>Stomias colubrinus</i>  <i>Rivulus atratus</i>  <i>Proaracana dubia</i>  <i>Alosa pontica</i>  <i>Belone acus</i>  <i>Antennarius radiusus</i>  <i>Barbus versluysii</i>  <i>Caranx otrynter</i>  <i>Ophichthus fowleri</i>  <i>Paratrachichthys pulsator</i>  <i>Gymnodraco victori</i>  <i>Trichomycterus florense</i>  <i>Muraenesox yamaguchiensis</i>  <i>Ipnops pristibrachium</i>  <i>Scomberoides commersonianus</i>  <i>Poropanchax myersi</i>  <i>Nemapteryx caelatus</i>  <i>Brachaluteres fahaqa</i>  <i>Rivulus crixas</i>  <i>Doryrhamphus dactyliophorus</i>  <i>Chelonodon dapsilis</i>  <i>Rhinecanthus rectangularis</i>  <i>Bathycongrus baranesi</i>  <i>Synaptura marginata</i>  <i>Antennarius duescus</i>  <i>Tetraodon palembangensis</i>  <i>Pseudocaranx cheilio</i>  <i>Echiophis mordax</i>  <i>Zignoichthys oblongus</i>  <i>Procatopus schioetzi</i>  <i>Microphis aculeatus</i>  <i>Halichoeres hyrtlil</i>  <i>Platybelone annobonensis</i>  <i>Nemacheilus savona</i>  <i>Pellona mayrinki</i>  <i>Batrachthys felinus</i>  <i>Aulopus nanae</i>  <i>Rhombus minimus</i>  <i>Bolcabalistes varii</i>  <i>Aplocheilichthys luluae</i>  <i>Liauchenoglanis maculatus</i>  <i>Pseudoscopelus microps</i>  <i>Electrona rissoi</i>  <i>Ophisurus rotundus</i>  <i>Cyprinella zanema</i>  <i>Brachirus fitzroiensis</i>  <i>Syngnathus acicularis</i>  <i>Pachythrissops laevis</i>  <i>Pachythrissops vectensis</i>  <i>Gymnothorax miliais</i>  <i>Rivulus derhami</i>  <i>Pseudolabrus crassilabris</i>  <i>Batasio niger</i>  <i>Arothron perspicillaris</i>  <i>Antennablennius sexfasciatus</i> </p>
---	---

	<p> <i>Pterygoplichthys sp.</i> (Vigliotta 2008)  <i>Anaethalion sp.</i> (Arratia 1999)  Genus 12 (Bockmann 1998)  <i>Phreatobius sp. 1</i> (de Pinna 1993)  Genus 6 sp. (Bockmann 1998)  <i>Corydoras sp. A</i> (Britto 2003)  <i>Aspidoras cf. pauciradiatus</i> (Britto 2003)  <i>Akysis sp. 2</i> (de Pinna 1996)  <i>Exostoma sp.</i> (Chen 1994)  <i>Mystus sp.</i> (Kailola 2004)  <i>Glyptothorax sp.</i> (Chen 1994)  Genus 2 sp. A (Bockmann 1998)  <i>Roestes sp.</i> (Lucena and Menezes 1998)  <i>Ictalurus sp.</i> (Mo 1991)  <i>Exostoma sp.</i> (de Pinna 1993)  Genus 6 (Bockmann 1998)  <i>Characidium cf. zebra</i> (Buckup 1998)  <i>Homaloptera sp.</i> (Fink and Fink 1981)  Genus 11 sp. (Bockmann 1998)  <i>Eigenmannia sp.</i> (Fink and Fink 1981)  <i>Cetopsorhamdia sp. n. A</i> (Bockmann 1998)  <i>Reganella sp.</i> (Britto 2003)  <i>Hemiancistrus sp. Brazil</i> (Armbruster 2004)  <i>Bunocephalus n. sp. 1</i> (Friel 1994)  <i>Loricaria sp.</i> (Friel 1994)  <i>Imparfinis sp.</i> (de Pinna 1993)  <i>Oxyropsis sp.</i> (Schaefer 1991)  Genus 11 (Bockmann 1998) </p>	<p> <i>Gymnelus platycephalus</i>  <i>Rivulus modestus</i>  <i>Parbatmya brazosensis</i>  <i>Gymnoscopelus aphyia</i>  <i>Gila bicolor</i>  <i>Bathophilus cwyanorum</i>  <i>Fundulopanchax sjoestedti</i>  <i>Anaethalion angustissimus</i>  <i>Arius parkii</i>  <i>Lampanyctus cupriarius</i>  <i>Ammodytes idai</i>  <i>Rivulus depressus</i>  <i>Albula goreensis</i>  <i>Hypostomus plecostomus 1</i> (Armbruster 2004)  <i>Eospinus daniltshenkoi</i>  <i>Stomias pacificus</i>  <i>Rhynchocypris oxycephalus</i>  <i>Simpsonichthys mediopapillatus</i>  <i>Aplocheilichthys atripinnis</i>  <i>Rivulus tessellatus</i>  <i>Rivulus leucurus</i>  <i>Pseudolabrus mortonii</i>  <i>Uropterygius goslinei</i>  <i>Orthrias panthera</i>  <i>Fundulopanchax kribianus</i>  <i>Hippocampus moluccensis</i>  <i>Bothus ypsigrammus</i>  <i>Microphis millepunctatus</i>  <i>Clarias tijsmanni</i>  <i>Rhinecanthus rectangulus</i>  <i>Labrus cyaedus</i>  <i>Hippocampus natalensis</i>  <i>Belone euxini</i>  <i>Heterandria dirempta</i>  <i>Echiodon anchiperus</i>  <i>Synodontis resupinata</i>  <i>Solea bleekeri</i>  <i>Acanthostracion polygonia</i>  <i>Varasichthys ariasi</i>  <i>Hemibagrus johorensis</i>  <i>Bovichtus elongatus</i>  <i>Notarius bonillai</i>  <i>Dinotopterus loweae</i>  <i>Barchatus cirrhosa</i>  <i>Trichiurus nanhaiensis</i>  <i>Chascanopsetta danae</i>  <i>Careproctus cryptacanthoides</i>  <i>Paramphilius goodi</i>  <i>Protacanthodes nimesensis</i>  <i>Lycodes gracilis</i>  <i>Paranotothenia angustata</i>  <i>Astephus calvus</i>  <i>Osmerus dentex</i>  <i>Alticus orientalis</i>  <i>Laimosemion jauaperi</i> </p>
--	---	---

		<p> <i>Rivulus faucireticulatus</i>  <i>Ompok krattensis</i>  <i>Tetraodon turgidus</i>  <i>Balistomorphus spinosus</i>  <i>Tylosurus fodiator</i>  <i>Leptolebias cruzi</i>  <i>Hippocampus suezensis</i>  <i>Lactoria fornasini</i>  <i>Tetraodon suvattii</i>  <i>Scomberoides moadetta</i>  <i>Pseudotrematomus eulepidotus</i>  <i>Rivulus limoncochae</i>  <i>Gymnothorax kikako</i>  <i>Plectocretacicus clarae</i>  <i>Gymnothorax argus</i>  <i>Astyanax vermilion</i>  <i>Phalacronotus micronema</i>  <i>Rivulus mazaruni</i>  <i>Clupea pallasii</i>  <i>Hippocampus bicuspis</i>  <i>Scartichthys rubropunctatus</i>  <i>Balistes villosus</i>  <i>Coloconger saldanhai</i>  <i>Harpagifer marionensis</i>  <i>Istiblennius enosimae</i>  <i>Epiplatys steindachneri</i>  <i>Lasiancistrus castelnaui</i>  <i>Rivulus riograndensis</i>  <i>Gymnothorax ruepellii</i>  <i>Anaethalion knorri</i>  <i>Triphoturus microchir</i>  <i>Nematogenys cuivi</i>  <i>Pseudojuloides inornatus</i>  <i>Anaethalion angustus</i>  <i>Neostethus geminus</i>  <i>Hypostomus emarginatus 2</i>  <i>(Armbruster 2004)</i>  <i>Pseudotrematomus nicolai</i>  <i>Coreobagrus okadai</i>  <i>Chelidonichthys gurnardus</i>  <i>Rivulus pinima</i>  <i>Dinotopterus foveolatus</i>  <i>Acanthopleurus collettei</i>  <i>Stlegicottus scutigera</i>  <i>Pterolebias hoignei</i>  <i>Cichlasoma festae</i>  <i>Bryx heraldi</i>  <i>Erethistes maesotensis</i>  <i>Nimbapanchax melanopterygius</i>  <i>Liparis takashimensis</i>  <i>Symphodus ocellaris</i>  <i>Synaptura commersonnii</i>  <i>Saurenhelys elongatum</i>  <i>Zignodon fornasieroae</i>  <i>Antennarius multiocellatus</i>  <i>Aspistor insculptus</i> </p>
--	--	--



		<p> <i>Halichoeres lamarii</i>  <i>Gephyroglanis congicus</i>  <i>Protacanthodes ombonii</i>  <i>Aplocheilichthys nimbaensis</i>  <i>Prohollardia avita</i>  <i>Carangoides ciliaris</i>  <i>Rivulus kuelpmanni</i>  <i>Hippocampus polytaenia</i>  <i>Peckoltia ucayalensis</i>  <i>Xenoclaris holobranchus</i>  <i>Chilomycterus reticulatus</i>  <i>Rivulus giarettai</i>  <i>Pseudobagrus hoi</i>  <i>Megalebias elongatus</i>  <i>Lycodes schmidti</i>  <i>Ostracion meretrix</i>  <i>Mastacembelus goro</i>  <i>Microphis vaillantii</i>  <i>Nototheniops larseni</i>  <i>Simpsonichthys semiocellatus</i>  <i>Corydoras bertonii</i>  <i>Holocentrus adscensionsis</i>  <i>Lycoptera sinensis?</i>  <i>Thalassenchelys foliaceus</i>  <i>Acentronura mossambica</i>  <i>Halicampus vittatus</i>  <i>Iniistius javanicus</i>  <i>Crystalloides enderburyensis</i>  <i>Lampanyctus bensoni</i>  <i>Rivulus parnaibensis</i>  <i>Rhamdia quelin</i>  <i>Bodianus albostrigatus</i>  <i>Orthogonikleithrus leichi</i>  <i>Hemiancistrus pankimpuju</i>  <i>Syngnathus heptagonus</i>  <i>Rivulus altivelis</i>  <i>Rivulus lungi</i>  <i>Lacustricola moeruensis</i>  <i>Rivulus gaucheri</i>  <i>Eoplectus bloti</i>  <i>Rivulus wassmanni</i>  <i>Arius danicus</i>  <i>Pseudocaranx georgianus</i>  <i>Laciris pelagicus</i>  <i>Microphis lineatus</i>  <i>Niwaella multifasciata</i>  <i>Pterygoplichthys ambrosettii</i>  <i>Rivulus elegans</i>  <i>Leiocassis macropterus</i>  <i>Arius harmandi</i>  <i>Rivulus javahe</i>  <i>Omobranchus ferox</i>  <i>Amarginops hildae</i>  <i>Uropterygius makatei</i>  <i>Myxocephalus quadricornis</i>  <i>Triglopsis quadricornis</i> </p>
--	--	--

		<p> <i>Halicampus ensenadae</i>  <i>Pseudeutropius atherinoides</i>  <i>Micronema cheveyi</i>  <i>Hypostomus plecostomus</i> 2  (Armbruster 2004)  <i>Synodontis cuangoana</i>  <i>Pardachirus rautheri</i>  <i>Pseudotocinclus jaquiae</i>  <i>Chirolophis snyderi</i>  <i>Rivulus holmiae</i>  <i>Thrissops regleyi</i>  <i>Leptolepides sprattiformis</i>  <i>Melanostomias spilorhynchus</i>  <i>Ichthyapus acutirostris</i>  <i>Sphyraena asotus</i>  <i>Archiaphyosemion petersi</i>  <i>Synaptura lusitanica</i>  <i>Aspistor luniscutis</i>  <i>Oligolactoria bubiki</i>  <i>Elassoma evergladi</i>  <i>Protobalistum imperialis</i>  <i>Amphichthys hildebrandi</i>  <i>Sardinella dayi</i>  <i>Doryrhamphus malus</i>  <i>Gnathagnus elongatus</i>  <i>Thrissops cephalus</i>  <i>Nelusetta ayraud</i>  <i>Rhynchocypris poljakowi</i>  <i>Aphyosemion viride</i>  <i>Glyptoperichthys lituratus</i>  <i>Phenacorhamdia tenuis</i>  <i>Megalocottus taeniopterus</i>  <i>Photostomias tantillux</i>  <i>Synodontis notata</i>  <i>Synodus amaranthus</i>  <i>Micropanchax pumilus</i>  <i>Encheliophis homei</i>  <i>Micropanchax rudolfianus</i>  <i>Notogoneus osculus</i>  <i>Rivulus rubromarginatus</i>  <i>Luisichthys vinalesensis</i>  <i>Monocentris japonica</i>  <i>Corythoichthys conspicillatus</i>  <i>Soleonassus finis</i>  <i>Aphyosemion obscurum</i>  <i>Mystus amemiyai</i>  <i>Harengula jaaguana</i>  <i>Megacheiroduon unicus</i>  <i>Simpsonichthys costai</i>  <i>Fundulus notti</i>  <i>Rivulus taeniatus</i>  <i>Leptocephalus proboscideus</i>  <i>Heterandria cataractae</i>  <i>Caranx ruber</i>  <i>Ariosoma sanzoi</i>  <i>Allothrissops mesogaster</i> </p>
--	--	--

		<p> <i>Rivulus cyanopterus</i>  <i>Aseraggodes smithi</i>  <i>Poecilia amazonica</i>  <i>Petrocephalus guttatus</i>  <i>Pangasius micronema</i>  <i>Trichomycterus cachiaensis</i>  <i>Rivulus kirovskyi</i>  <i>Glyptothorax horai</i>  <i>Monochirus atlanticus</i>  <i>Pseudotrematomus tokarevi</i>  <i>Chiasmodon bolangeri</i>  <i>Rivulus decoratus</i>  <i>Doumea alula</i>  <i>Rhynchocypris oxycephala</i>  <i>Rivulus kirovskyi</i>  <i>Pisodonophis semicinctus</i>  <i>Lampanyctus gemmifer</i>  <i>Trichomycterus quechuorum</i>  <i>Phenacogrammus altus</i>  <i>Clarias camerunensus</i>  <i>Microdevario kubotai</i>  <i>Opsanus tao</i>  <i>Strabozebrias cancellatus</i>  <i>Pseudotrematomus vicarius</i>  <i>Triodon antiquus</i>  <i>Ageneiosus rondoni</i>  <i>Lagocephalus cheesemanii</i>  <i>Halichoeres dimidiatus</i>  <i>Halichoeres penrosei</i>  <i>Platybelone platyura</i>  <i>Conger marginatus</i>  <i>Chrysichthys stappersii</i>  <i>Acanthopleurus trispinosus</i>  <i>Malthopsis retifera</i>  <i>Aethotaxis mitopteryx</i>  <i>Micrognathus pygmaeus</i>  <i>Rivulus egens</i>  <i>New genus (Schaefer 1991)</i>  <i>Rivulus simplicis</i>  <i>Encheliophis boraborensis</i>  <i>Paralichthys coeruleosticta</i>  <i>Rhabdoblennius rhadotrachelus</i>  <i>Thamnaconus garretti</i>  <i>Acanthopsoides graciroides</i>  <i>(Sawada 1982)</i>  <i>Phoxinus perenurus</i>  <i>Rivulus sucubti</i>  <i>Solea triophthalma</i>  <i>Bathycongrus macrocercus</i>  <i>Cyprinodon hubbsi</i>  <i>Synaptura annularis</i>  <i>Rivulus beniensis</i>  <i>Pholidophorus bechei</i>  <i>Hara filamentosus</i>  <i>Phrynorhombus regius</i>  <i>Moclaybalistes danekrus</i> </p>
--	--	--

		<p> <i>Aphyosemion microphthalmum</i>  <i>Rhynchostracion nasus</i>  <i>Aplocheilichthys loati</i>  <i>Muraenesox ferox</i>  <i>Alticus magnusi</i>  <i>Festucalex townsendi</i>  <i>Liparis niger</i>  <i>Rivulus boehlkei</i>  <i>Limia nicholsi</i>  <i>Lagocephalus guntheri</i>  <i>Xyrichtys cyanifrons</i>  <i>Tetraodon baileyi</i>  <i>Liparis meridionalis</i>  <i>Parika scaber</i>  <i>Moema ortegai</i>  <i>Orthogonikleithrus hoelli</i>  <i>Ctenopoma machadoi</i>  <i>Solea hexophthalma</i>  <i>Rivulus amanapira</i>  <i>Netuma proxima</i>  <i>Cyprinus niloticus</i>  <i>Anguilla huangi</i>  <i>Trachinotus glaucus</i>  <i>Diplomystus dentatus</i>  <i>Acanthurus sandvicensis</i>  <i>Micropoecilia parae</i>  <i>Aphyosemion bochtleri</i>  <i>Sympterychthys verrucosus</i>  <i>Rivulus villwocki</i>  <i>Nipponocypris temminckii</i>  <i>Syngnathus acusimilis</i>  <i>Oxycheilinus digrammus</i>  <i>Simpsonichthys bokermanni</i>  <i>Simpsonichthys harmonicus</i>  <i>Schilbe djemeri</i>  <i>Rivulus xinguensis</i>  <i>Rivulus karaja</i>  <i>Pseudobagrus longirostris</i>  <i>Notoscopelus kroyeri</i>  <i>Hippocampus tristis</i>  <i>Ichthyscopus inermis</i>  <i>Astyanax armandoi</i>  <i>Poeciliopsis sonoriensis</i>  <i>Thalassoma septemfasciatum</i>  <i>Aplocheilichthys pumilis</i>  <i>Careproctus entargyreus</i>  <i>Albula argentea</i>  <i>Channa striatus</i>  <i>Synodontis melanosticta</i>  <i>Chelonodon fluviatilis</i>  <i>Notropis boucardi</i>  <i>Carapus smithvillensis</i>  <i>Merodoras nheco</i>  <i>Doryrhamphus paulus</i>  <i>Lindbergichthys nudifrons</i>  <i>Doryrhamphus melanopleura</i> </p>
--	--	---

		<p> <i>Parapteronotus bonaparti</i>  <i>Pachythrissops furcatus</i>  <i>Pseudobagrus virgatus</i>  <i>Choerodon caninus</i>  <i>Tylosurus melanotus</i>  <i>Clupea bentincki</i>  <i>Symphurus novemfasciatus</i>  <i>Cosmocampus coccineus</i>  <i>Monotrete ocellaris</i>  <i>Pseudocoris philippina</i>  <i>Hippocampus takakurae</i>  <i>Stemonosudis similis</i>  <i>Heptadiodon echinus</i>  <i>Synodontis multimaculata</i>  <i>Acanthaluteres brwonii</i>  <i>Hemiarius maculatus</i>  <i>Liparis barbatus</i>  <i>Orestias agassii</i>  <i>Selar crumnophthalmus</i>  <i>Notarius quadriscutis</i>  <i>Rivulus rubripunctatus</i>  <i>Soleichthys nigrostriolatus</i>  <i>Siderea pictus</i>  <i>Corydoras esperanze</i>  <i>Trichiurus nitens</i>  <i>Eotetraodon pygmaeus</i>  <i>Ostracion camurun</i>  <i>Rivulus erberi</i>  <i>Hippocampus arnei</i>  <i>Poecilia zonata</i>  <i>Muraenichthys malabonensis</i>  <i>Barbourichthys zanzibaricus</i>  <i>Acanthopleurus serratus</i>  <i>Astronesthes barbatus</i>  <i>Rivulus parsi</i>  <i>Pelteobagrus sinensis</i>  <i>Rivulus paracatuensis</i>  <i>Rivulus igneus</i>  <i>Solegnathus naso</i>  <i>Careproctus entomelas</i>  <i>Ancistrus sp (Hoffmann and Britz 2006)</i>  <i>Hemisynodontis membranaceus</i>  <i>Corythoichthys waitei</i>  <i>Antennarius dorehensis</i>  <i>Nemacheilus rupecula</i>  <i>Pterocryptis afghana</i>  <i>Ageneiosus valenciennesi</i>  <i>Aulotrachichthys fernandezianus</i>  <i>Ompok anostomus</i>  <i>Synodontis ilebrebis</i>  <i>Pseudobagrus taphraphilus</i>  <i>Synodontis aterrima</i>  <i>Oreoglanis pumatensis</i>  <i>Pachythrissops propterus</i>  <i>Leuciscus cephalus</i> </p>
--	--	---

		<p> <i>Synodontis macropunctata</i>  <i>Pseudobagarius kyphus</i>  <i>Hippocampus bleekeri</i>  <i>Hippocampus kampylotrachelos</i>  <i>Eolactoria sorbinii</i>  <i>Malacosteus indicus</i>  <i>Hippocampus manadensis</i>  <i>Pictichromis caitinae</i>  <i>Festucalex amakusensis</i>  <i>Chionodraco kathleenae</i>  <i>Chilomycterus atringa</i>  <i>Mystus fortis</i>  <i>Antennarius bermudensis</i>  <i>Ophichthus altipennis</i>  <i>Balistomorphus ovalis</i>  <i>Trachyrhamphus bicoractatus</i>  <i>Antennarius sanguineus</i>  <i>Clarias dialonensis</i>  <i>Synodontis zanzibarica</i>  <i>Antennarius rosaceus</i>  <i>Balistomorphus orbiculatus</i>  <i>Tetraodon cutcutia</i>  <i>Holtbyrnia melanocephala</i>  <i>Ctenopoma oxyrhynchum</i>  <i>Pterolebias zonatus</i>  <i>Rivulus birkhahni</i>  <i>Poecilia dominicensis</i>  <i>Eubalichthys quadrispinus</i>  <i>Paratrachichthys novaezelandicus</i>  <i>Hypophthalmus longifilis</i>  <i>Notropis sallaei</i>  <i>Pelteobagrus argentivittatus</i>  <i>Peckoltia kuhlmanni</i>  <i>Ophichthus rotidoderma</i>  <i>Mystus guilio</i>  <i>Alepes melanoptra</i>  <i>Synodontis macrophthalma</i>  <i>Microphis yoshi</i>  <i>Peckoltia filicaudata</i>  <i>Hypsopanchax modestus</i>  <i>Sciades emphysetus</i>  <i>Cichlasoma facetum</i>  <i>Ariopsis guatemalensis</i>  <i>Xyrichtys twistii</i>  <i>Cottus japonicus</i>  <i>Tetraodon cambodgiensis</i>  <i>Oligobalistes robustus</i>  <i>Synodontis vermiculata</i>  <i>Encheliophis dubius</i>  <i>Phaenomonas foresti</i>  <i>Paralichthys oblongus</i>  <i>Leptobotia curta</i>  <i>Pterygoplichthys anisitsi</i>  <i>Socnopaea horai</i>  <i>Melanocetus polyactis</i>  <i>Amphilophus robertsoni</i> </p>
--	--	---

		<p> <i>Kryptopterus bleekeri</i>  <i>Duopalatinus olallae</i>  <i>Blennius maoricus</i>  <i>Helicolenus hilgendorffi</i>  <i>Rivulus christinae</i>  <i>Ariosoma gnanadossi</i>  <i>Rasbora dusonensis</i> 2 (Cavender  and Coburn 1992)  <i>Rivulus dapazi</i>  <i>Carpathospinosus propheticus</i>  <i>Ompok siluroides</i>  <i>Mastacembelus batesii</i>  <i>Cichlasoma atromaculatum</i>  <i>Chilomycterus schoepfi</i>  <i>Choeroichthys valencienni</i>  <i>Rita itchkeea</i>  <i>Astyanax rutilus</i>  <i>Ompok mysoricus</i>  <i>Simpsonichthys flammeus</i>  <i>Rhamdia parvus</i>  <i>Antennablennius velifer</i>  <i>Sprattus phalerica</i>  <i>Pseudoscopelus stellatus</i>  <i>Coreobagrus ichikawai</i>  <i>Synodontis dorsomaculata</i>  <i>Synodontis albolineata</i>  <i>Achiropsis nattereri</i>  <i>Prodiodon tenuispinus</i>  <i>Synodontis kogoensis</i>  <i>Cathorops festae</i>  <i>Tischlingerichthys vlohli</i>  <i>Catathyridium lorentzi</i>  <i>Clupeonella tscharchalensis</i>  <i>Bathophilus novicki</i>  <i>Parapterygotrigla multiocellata</i>  <i>Thrissops subovatus</i>  <i>Brachirus breviceps</i>  <i>Liparis lindbergi</i>  <i>Thalassoma quinquevittatus</i>  <i>Notarius rugispinis</i>  <i>Heterandria jonesii</i>  <i>Poecilothrissa moeruensis</i>  <i>Nothobranchius kiyawensis</i>  <i>Aphyosemion melinoeides</i>  <i>Rivulus nudiventris</i>  <i>Tylosurus rafale</i>  <i>Nannoptopoma sternoptychum</i>  <i>Centromochlus musaica</i>  <i>Lebias fasciatus</i>  <i>Prodiodon erinaceus</i>  <i>Tetraodon abei</i>  <i>Simpsonichthys antenori</i>  <i>Parapercis naevosa</i>  <i>Lindbergichthys mizops</i>  <i>Aphyosemion margaretae</i>  <i>Pangasius tubbi</i> </p>
--	--	--

		<i>Nototheniops loesha</i> <i>Ophichthus lithinus</i> <i>Dinotopterus nyasensis</i> <i>Doryrhamphus extensus</i> <i>Charitosomous lineolatus</i> <i>Heterandria anzueto</i> <i>Paraliparis wildi</i> <i>Gambusia rachovii</i> <i>Rivulus uraenis</i> <i>Liparis ingens</i> <i>Hypoptopoma joberti</i> <i>Channomuraena bennettii</i> <i>Chauliodus slaoni</i> <i>Rivulus cearensis</i>
--	--	---



## REFERENCES

- Abbasi, A.A. Evolution of vertebrate appendicular structures: Insight from genetic and palaeontological data. *Developmental Dynamics* 2011;240(5):1005-1016.
- Aittokallio, T. and Schwikowski, B. Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics* 2006;7(3):243-255.
- Akimenko, M.-A. and Ekker, M. Anterior Duplication of the Sonic hedgehog Expression Pattern in the Pectoral Fin Buds of Zebrafish Treated with Retinoic Acid. *Developmental Biology* 1995;170(1):243-247.
- Albalat, R., Baquero, M. and Minguillón, C. Identification and characterisation of the developmental expression pattern of *tbx5b*, a novel *tbx5* gene in zebrafish. *Gene Expression Patterns* 2010;10(1):24-30.
- Alexandre, D., Clarke, J.D., Oxtoby, E., Yan, Y.L., Jowett, T. and Holder, N. Ectopic expression of *Hoxa-1* in the zebrafish alters the fate of the mandibular arch neural crest and phenocopies a retinoic acid-induced phenotype. *Development* 1996;122(3):735-746.
- Amaral, D.B. and Schneider, I. Fins into limbs: Recent insights from sarcopterygian fish. *genesis* 2018;56(1):e23052.
- Amatruda, J.F., Gattermeir, D.J., Karpova, T.S. and Cooper, J.A. Effects of null mutations and overexpression of capping protein on morphogenesis, actin distribution and polarized secretion in yeast. *The Journal of cell biology* 1992;119(5):1151-1162.
- Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R. and Magrane, M. UniProt: the universal protein knowledgebase. *Nucleic acids research* 2004;32(suppl\_1):D115-D119.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 2000;25(1):25-29.
- Austin, C.J. Evo-devo: a science of dispositions. *European Journal for Philosophy of Science* 2017;7(2):373-389.
- Bader, G. and Hogue, C. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics* 2003;4(1):2.
- Bakrania, P., Efthymiou, M., Klein, J.C., Salt, A., Bunyan, D.J., Wyatt, A., Ponting, C.P., Martin, A., Williams, S., Lindley, V., Gilmore, J., Restori, M., Robson, A.G., Neveu, M.M., Holder, G.E., Collin, J.R.O., Robinson, D.O., Farndon, P., Johansen-Berg, H., Gerrelli, D. and Rague, N.K. Mutations in *BMP4* Cause Eye, Brain, and Digit Developmental Anomalies: Overlap between the *BMP4* and Hedgehog Signaling Pathways. *The American Journal of Human Genetics* 2008;82(2):304-319.
- Bandyopadhyay, A., Tsuji, K., Cox, K., Harfe, B.D., Rosen, V. and Tabin, C.J. Genetic Analysis of the Roles of *BMP2*, *BMP4*, and *BMP7* in Limb Patterning and Skeletogenesis. *PLOS Genetics* 2006;2(12):e216.
- Baumgartner Jr, W.A., Cohen, K.B., Fox, L.M., Acquah-Mensah, G. and Hunter, L. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 2007;23(13):i41-i48.
- Beck, M. and Baumeister, W. Cryo-electron tomography: can it reveal the molecular sociology of cells in atomic detail? *Trends in cell biology* 2016;26(11):825-837.

- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. GenBank. *Nucleic acids research* 2008;36(Database issue):D25.
- Boucherat, O., Nadeau, V., Bérubé-Simard, F.-A., Charron, J. and Jeannotte, L. Crucial requirement of ERK/MAPK signaling in respiratory tract development. *Development* 2014;141(16):3197.
- Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J.A., Mozzherin, D., Rees, T., Matasci, N., Narro, M.L., Piel, W.H., McKay, S.J., Lowry, S., Freeland, C., Peet, R.K. and Enquist, B.J. The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC bioinformatics* 2013;14(1):16.
- Braasch, I., Peterson, S.M., Desvignes, T., McCluskey, B.M., Batzel, P. and Postlethwait, J.H. A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. *Journal of experimental zoology. Part B, Molecular and developmental evolution* 2014.
- Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D.G., Knight, J., Mani, P., Martin, R., Moxon, S.A.T., Paddock, H., Pich, C., Ramachandran, S., Ruef, B.J., Ruzicka, L., Bauer Schaper, H., Schaper, K., Shao, X., Singer, A., Sprague, J., Sprunger, B., Van Slyke, C. and Westerfield, M. ZFIN: enhancements and updates to the zebrafish model organism database. *Nucleic acids research* 2011;39(suppl\_1):D822-D829.
- Browne, M.L., Carter, T.C., Kay, D.M., Kuehn, D., Brody, L.C., Romitti, P.A., Liu, A., Caggana, M., Druschel, C.M. and Mills, J.L. Evaluation of genes involved in limb development, angiogenesis, and coagulation as risk factors for congenital limb deficiencies. *American Journal of Medical Genetics Part A* 2012;158(10):2463-2472.
- Cardona, G., Rosselló, F. and Valiente, G. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC bioinformatics* 2008;9(1):532.
- Chua, H.N., Sung, W.-K. and Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 2006;22(13):1623-1630.
- Clack, J.A. Gaining ground: the origin and evolution of tetrapods. Indiana University Press; 2012.
- Coates, M.I. The origin of vertebrate limbs. *Development (Cambridge, England). Supplement* 1994:169-180.
- Coates, M.I. and Cohn, M.J. Fins, limbs, and tails: outgrowths and axial patterning in vertebrate evolution. *BioEssays* 1998;20(5):371-381.
- Consortium, G.O. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* 2004;32(suppl 1):D258-D261.
- Cormen, T.H. Introduction to algorithms. MIT press; 2009.
- Cowen, L., Ideker, T., Raphael, B.J. and Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* 2017.
- Cranston, K.A., Harmon, L.J., O'Leary, M.A. and Lisle, C. Best practices for data sharing in phylogenetic research. *PLoS Currents* 2014;6:ecurrents.tol.bf01eff04a06b60ca4825c69293dc59645.
- Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J. and Sander, C. Pathway and network analysis of cancer genomes. *Nature methods* 2015;12(7):615.
- Cunningham, C.W. Some limitations of ancestral character-state reconstruction when testing evolutionary hypotheses. *Systematic Biology* 1999;48(3):665-674.

- Dahdul, W., Dececchi, T.A., Ibrahim, N., Lapp, H. and Mabee, P. Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy. *Database* 2015;2015.
- Dahdul, W.M., Balhoff, J.P., Blackburn, D.C., Diehl, A.D., Haendel, M.A., Hall, B.K., Lapp, H., Lundberg, J.G., Mungall, C.J., Ringwald, M., Segerdell, E., Van Slyke, C.E., Vickaryous, M.K., Westerfield, M. and Mabee, P.M. A unified anatomy ontology of the vertebrate skeletal system. *PLoS One* 2012;7(12):e51070.
- Dahdul, W.M., Lundberg, J.G., Midford, P.E., Balhoff, J.P., Lapp, H., Vision, T.J., Haendel, M.A., Westerfield, M. and Mabee, P.M. The teleost anatomy ontology: anatomical representation for the genomics age. *Systematic Biology* 2010;59(4):369-383.
- Dal-Pra, S., Fürthauer, M., Van-Celst, J., Thisse, B. and Thisse, C. Noggin1 and Follistatin-like2 function redundantly to Chordin to antagonize BMP activity. *Developmental Biology* 2006;298(2):514-526.
- Dalcq, J., Pasque, V., Ghaye, A., Larbuisson, A., Motte, P., Martial, J.A. and Muller, M. RUNX3, EGR1 and SOX9B Form a Regulatory Cascade Required to Modulate BMP-Signaling during Cranial Cartilage Development in Zebrafish. *PLOS ONE* 2012;7(11):e50140.
- Das, A. and Crump, J.G. Bmps and Id2a Act Upstream of Twist1 To Restrict Ectomesenchyme Potential of the Cranial Neural Crest. *PLOS Genetics* 2012;8(5):e1002710.
- De Pinna, M.C. PhD Dissertation. New York: City University of New York; 1993. Higher-level Phylogeny of Siluriformes, with a New Classification of the Order (Teleostei: Ostariophysi).
- Dececchi, T.A., Balhoff, J.P., Lapp, H. and Mabee, P.M. Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Systematic Biology* 2015;64(6):936-952.
- Dececchi, T.A., Mabee, P.M. and Blackburn, D.C. Data sources for trait databases: comparing the phenomic content of monographs and evolutionary matrices. *PLoS ONE* 2016;11(5):e0155680.
- Defoort, J., Van de Peer, Y. and Vermeirssen, V. Function, dynamics and evolution of network motif modules in integrated gene regulatory networks of worm and plant. *Nucleic acids research* 2018.
- Didier, G. Time-Dependent-Asymmetric-Linear-Parsimonious Ancestral State Reconstruction. *Bulletin of Mathematical Biology* 2017;79(10):2334-2355.
- Don, E.K., Currie, P.D. and Cole, N.J. The evolutionary history of the development of the pelvic fin/hindlimb. *Journal of anatomy* 2013;222(1):114-133.
- Duan, D., Yue, Y., Zhou, W., Labe, B., Ritchie, T.C., Grosschedl, R. and Engelhardt, J.F. Submucosal gland development in the airway is controlled by lymphoid enhancer binding factor 1 (LEF1). *Development* 1999;126(20):4441-4453.
- Duan, G. and Walther, D. The roles of post-translational modifications in the context of protein interaction networks. *PLoS computational biology* 2015;11(2):e1004049.
- Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* 2009;10:48-48.
- Edmunds, R.C., Su, B., Balhoff, J.P., Eames, B.F., Dahdul, W.M., Lapp, H., Lundberg, J.G., Vision, T.J., Dunham, R.A., Mabee, P.M. and Westerfield, M. Phenoscope: Identifying

- Candidate Genes for Evolutionary Phenotypes. *Molecular Biology and Evolution* 2016;33(1):13-24.
- Elhanan, G., Ochs, C., Mejino Jr, J.L.V., Liu, H., Mungall, C.J. and Perl, Y. From SNOMED CT to Uberon: transferability of evaluation methodology between similarly structured ontologies. *Artificial intelligence in medicine* 2017;79:9-14.
- Erard, N., Knott, S.R.V. and Hannon, G.J. A CRISPR Resource for Individual, Combinatorial, or Multiplexed Gene Knockout. *Molecular cell* 2017;67(2):348-354.
- Erdin, S., Lisewski, A.M. and Lichtarge, O. Protein function prediction: towards integration of similarity metrics. *Current Opinion in Structural Biology* 2011;21(2):180-188.
- Eschmeyer, W. Catalog of Fishes: Genera, Species, References. In: Eschmeyer, W., Fricke, R. and Van der Laan, R., editors.: California Academy of Sciences; 2013.
- Fan, J., Han, F. and Liu, H. Challenges of big data analysis. *National science review* 2014;1(2):293-314.
- Fernando, P.C., Jackson, L.M., Zeng, E., Mabee, P.M. and Balhoff, J.P. A generic bioinformatics pipeline to integrate large-scale trait data with large phylogenies. In: IEEE; 2017. p. 2235-2237.
- Ficklin, S.P., Dunwoodie, L.J., Poehlman, W.L., Watson, C., Roche, K.E. and Feltus, F.A. Discovering Condition-Specific Gene Co-Expression Patterns Using Gaussian Mixture Models: A Cancer Case Study. *Scientific reports* 2017;7(1):8617.
- Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P. and von Mering, C. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research* 2013;41(D1):D808-D815.
- Fraser, A.G. and Marcotte, E.M. A probabilistic view of gene function. *Nature genetics* 2004;36(6):559.
- Freitas, R., Gómez-Marín, C., Wilson, J.M., Casares, F. and Gómez-Skarmeta, J.L. Hoxd13 contribution to the evolution of vertebrate appendages. *Developmental cell* 2012;23(6):1219-1229.
- Gagneur, J., Krause, R., Bouwmeester, T. and Casari, G. Modular decomposition of protein-protein interaction networks. *Genome biology* 2004;5(8):R57.
- Gao, J., DeRouen, M.C., Chen, C.-H., Nguyen, M., Nguyen, N.T., Ido, H., Harada, K., Sekiguchi, K., Morgan, B.A. and Miner, J.H. Laminin-511 is an epithelial message promoting dermal papilla development and function during early hair morphogenesis. *Genes & development* 2008;22(15):2111-2124.
- Gene Ontology, C. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research* 2016;45(D1):D331-D338.
- Georgii, E., Dietmann, S., Uno, T., Pagel, P. and Tsuda, K. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics* 2009;25(7):933-940.
- Gillis, J.A., Modrell, M.S. and Baker, C.V.H. Developmental evidence for serial homology of the vertebrate jaw and gill arch skeleton. *Nature communications* 2013;4:1436.
- Gkoutos, G.V., Green, E.C., Mallon, A.M., Hancock, J.M. and Davidson, D. Using ontologies to describe mouse phenotypes. *Genome biology* 2005;6(1):R8.
- Gkoutos, G.V., Mungall, C., Dölken, S., Ashburner, M., Lewis, S., Hancock, J., Schofield, P., Köhler, S. and Robinson, P.N. Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO. *Conference Proceedings* 2009;2009:7069-7072.

- Grandel, H., Draper, B.W. and Schulte-Merker, S. *dackel* acts in the ectoderm of the zebrafish pectoral fin bud to maintain AER signaling. *Development* 2000;127(19):4169.
- Grandel, H. and Schulte-Merker, S. The development of the paired fins in the Zebrafish (*Danio rerio*). *Mechanisms of Development* 1998;79(1):99-120.
- Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E. and Sealfon, S.C. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* 2015;47(6):569.
- Gregory, S. Finding overlapping communities in networks by label propagation. *New Journal of Physics* 2010;12(10):103018.
- Gruenstaedl, M. WARACS: wrappers to automate the reconstruction of ancestral character states. *Applications in Plant Sciences* 2016;4(2):1500120.
- Gu, L., Frommel, S.C., Oakes, C.C., Simon, R., Grupp, K., Gerig, C.Y., Bär, D., Robinson, M.D., Baer, C., Weiss, M., Gu, Z., Schapira, M., Kuner, R., Sültmann, H., Provenzano, M., Cancer, I.P.o.E.O.P., Yaspo, M.-L., Brors, B., Korb, J., Schlomm, T., Sauter, G., Eils, R., Plass, C. and Santoro, R. BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence. *Nature genetics* 2014;47:22.
- Haendel, M.A., Balhoff, J.P., Bastian, F.B., Blackburn, D.C., Blake, J.A., Bradford, Y., Comte, A., Dahdul, W.M., Dececchi, T.A. and Druzinsky, R.E. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics* 2014;5:21.
- Haendel, M.A., Balhoff, J.P., Bastian, F.B., Blackburn, D.C., Blake, J.A., Bradford, Y., Comte, A., Dahdul, W.M., Dececchi, T.A., Druzinsky, R.E., Hayamizu, T.F., Ibrahim, N., Lewis, S.E., Mabee, P.M., Niknejad, A., Robinson-Rechavi, M., Sereno, P.C. and Mungall, C.J. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics* 2014;5(1):21.
- Hall, B., Limaye, A. and Kulkarni, A.B. Overview: generation of gene knockout mice. *Current protocols in cell biology* 2009:19.12. 11-19.12. 17.
- Han, J., Pei, J. and Kamber, M. Data mining: concepts and techniques. Elsevier; 2011.
- Harmon, L.J., Baumes, J., Hughes, C., Soberon, J., Specht, C.D., Turner, W., Lisle, C. and Thacker, R.W. Arbor: comparative analysis workflows for the Tree of Life. *PLoS Currents* 2013;5:ecurrents.tol.099161de099165eabdee099073fd099163d099121a044518dc.
- Harris, J.R. and Arbuckle, K. Tempo and mode of the evolution of venom and poison in tetrapods. *Toxins* 2016;8(7).
- Hart, G.T., Ramani, A.K. and Marcotte, E.M. How complete are current yeast and human protein-interaction networks? *Genome biology* 2006;7(11):120.
- Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., Gude, K., Hibbett, D.S., Katz, L.A., Laughinghouse, H.D., I.V., McTavish, E.J., Midford, P.E., Owen, C.L., Ree, R.H., Rees, J.A., Soltis, D.E., Williams, T. and Cranston, K.A. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences of the United States of America* 2015;112(41):12764-12769.

- Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast* 2001;18(6):523-531.
- Hoffman, B.G., Williams, K.L., Tien, A.H., Lu, V., de Algara, T.R., Ting, J.-y. and Helgason, C.D. Identification of novel genes and transcription factors involved in spleen, thymus and immunological development and function. *Genes and immunity* 2006;7(2):101.
- Horridge, M. and Bechhofer, S. The owl api: A java api for owl ontologies. *Semantic Web* 2011;2(1):11-21.
- Huang, D.W., Sherman, B.T. and Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 2009;37(1):1-13.
- Huang, L., Jin, J., Deighan, P., Kiner, E., McReynolds, L. and Lieberman, J. Efficient and specific gene knockdown by small interfering RNAs produced in bacteria. *Nature biotechnology* 2013;31(4):350-356.
- Hunt, G. and Slater, G. Integrating paleontological and phylogenetic approaches to macroevolution. *Annual Review of Ecology, Evolution, and Systematics* 2016;47(1):189-213.
- Im, C.H., Chau, K., Lopez, M., Sun, X.Z., Tran, J., Cuya, S. and Sharp, S.B. Hindlimb Muscles of 17.5–18.5 dpc Mice Double Null for MyoD and Trp53 Appear Indistinguishable from Muscles of Mice Null for Either Gene. *The FASEB Journal* 2016;30(1\_supplement):1035-1032.
- Jackson, L.M., Fernando, P.C., Hanscom, J.S., Balhoff, J.P. and Mabee, P.M. Automated Integration of Trees and Traits: A Case Study Using Paired Fin Loss Across Teleost Fishes. *Systematic Biology* 2018:syx098-syx098.
- Jackson, L.M., Fernando, P.C., Hanscom, J.S., Balhoff, J.P. and Mabee, P.M. Automated integration of trees and traits: a case study using paired fin loss across teleost fishes. *Systematic biology* 2018;67(4):559-575.
- Jeong, J.C. and Chen, X.w. Evaluating topology-based metrics for GO term similarity measures. In, *2013 IEEE International Conference on Bioinformatics and Biomedicine*. 2013. p. 43-48.
- Jiang, R., Gan, M. and He, P. Constructing a gene semantic similarity network for the inference of disease genes. *BMC Systems Biology* 2011;5(2):S2.
- Jones, D.T. and Swindells, M.B. Getting the most from PSI-BLAST. In.: Elsevier; 2002.
- Joos, J.P., Saadatmand, A.R., Schnabel, C., Viktorinová, I., Brand, T., Kramer, M., Nattel, S., Dobrev, D., Tomancak, P. and Backs, J. Ectopic expression of S28A-mutated Histone H3 modulates longevity, stress resistance and cardiac function in *Drosophila*. *Scientific reports* 2018;8(1):2940.
- Kamaid, A., Molina-Villa, T., Mendoza, V., Pujades, C., Maldonado, E., Ispizua Belmonte, J.C. and López-Casillas, F. Betaglycan knock-down causes embryonic angiogenesis defects in zebrafish. *genesis* 2015;53(9):583-603.
- Kawakami, Y., Marti, M., Kawakami, H., Itou, J., Quach, T., Johnson, A., Sahara, S., O’Leary, D.D.M., Nakagawa, Y. and Lewandoski, M. Islet1-mediated activation of the  $\beta$ -catenin pathway is necessary for hindlimb initiation in mice. *Development* 2011;138(20):4465-4473.

- Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R. and Ideker, T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS* 2003;100.
- Kernighan, B.W. and Lin, S. An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal* 1970;49(2):291-307.
- Kibbe, W.A., Arze, C., Felix, V., Mitra, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., Parkinson, H. and Schriml, L.M. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research* 2015;43(D1):D1071-D1078.
- Kim, J.D., Ohta, T., Tateisi, Y. and Tsujii, J.i. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* 2003;19(suppl\_1):i180-i182.
- Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F. and Boycott, K.M. The human phenotype ontology in 2017. *Nucleic acids research* 2016;45(D1):D865-D876.
- Kourmpetis, Y.A.I., van Dijk, A.D.J., Bink, M.C.A.M., van Ham, R.C.H.J. and ter Braak, C.J.F. Bayesian Markov Random Field Analysis for Protein Function Prediction Based on Network Data. *PLOS ONE* 2010;5(2):e9293.
- Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W. and Ma'ayan, A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* 2016;44(W1):W90-W97.
- Laisney, J.A.G.C., Braasch, I., Walter, R.B., Meierjohann, S. and Schartl, M. Lineage-specific co-evolution of the Egf receptor/ligand signaling system. *BMC Evolutionary Biology* 2010;10(1):27.
- Larouche, O., Zelditch, M.L. and Cloutier, R. Fin modules: an evolutionary perspective on appendage disparity in basal vertebrates. *BMC Biol* 2017;15(1):32.
- Le, D.-H. and Dang, V.-T. Ontology-based disease similarity network for disease gene prediction. *Vietnam Journal of Computer Science* 2016;3(3):197-205.
- Lee, T.-C. and Threadgill, D.W. Generation and validation of mice carrying a conditional allele of the epidermal growth factor receptor. *genesis* 2008;47(2):85-92.
- Letelier, J., de la Calle-Mustienes, E., Pieretti, J., Naranjo, S., Maeso, I., Nakamura, T., Pascual-Anaya, J., Shubin, N.H., Schneider, I., Martinez-Morales, J.R. and Gómez-Skarmeta, J.L. A conserved Shh cis-regulatory module highlights a common developmental origin of unpaired and paired fins. *Nature genetics* 2018;50(4):504-509.
- Leunissen, M. Explanation and Teleology in Aristotle's Science of Nature. New York: Cambridge University Press; 2010.
- Li, M., Hu, X., Zhu, J., Zhu, C., Zhu, S., Liu, X., Xu, J., Han, S. and Yu, Z. Overexpression of miR-19b impairs cardiac development in zebrafish by targeting *ctnbl1*. *Cellular Physiology and Biochemistry* 2014;33(6):1988-2002.
- Liang, Z., Xu, M., Teng, M. and Niu, L. Comparison of protein interaction networks reveals species conservation and divergence. *BMC bioinformatics* 2006;7(1):457.
- Lin, D. An information-theoretic definition of similarity. In, *ICML*. 1998. p. 296-304.
- Liu, J., Jing, L. and Tu, X. Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. *BMC Cardiovascular Disorders* 2016;16(1):54.

- Liu, Q., Dalman, M., Chen, Y., Akhter, M., Brahmandam, S., Patel, Y., Lowe, J., Thakkar, M., Gregory, A.-V., Phelps, D., Riley, C. and Londraville, R.L. Knockdown of leptin A expression dramatically alters zebrafish development. *General and Comparative Endocrinology* 2012;178(3):562-572.
- Liu, X. and Murata, T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A: Statistical Mechanics and its Applications* 2010;389(7):1493-1500.
- Liu, Z., Liu, Q., Sun, H., Hou, L., Guo, H., Zhu, Y., Li, D. and He, F. Evidence for the additions of clustered interacting nodes during the evolution of protein interaction networks from network motifs. *BMC Evolutionary Biology* 2011;11(1):133.
- Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory* 1982;28(2):129-137.
- Long, J.A. and Gordon, M.S. The Greatest Step in Vertebrate History: A Paleobiological Review of the Fish-Tetrapod Transition\*. *Physiological and Biochemical Zoology* 2004;77(5):700-719.
- Lopez-Rios, J. The many lives of SHH in limb development and evolution. *Seminars in Cell & Developmental Biology* 2016;49:116-124.
- Maddison, W. and Maddison, D. 2016. Mesquite: a modular system for evolutionary analysis. Release 3.10. <http://mesquiteproject.org>
- Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* 2005;33(suppl\_1):D54-D58.
- Manda, P., Balhoff, J.P. and Vision, T.J. Measuring the importance of annotation granularity to the detection of semantic similarity between phenotype profiles. *bioRxiv* 2016:086306.
- Martínez-López, B., Perez, A.M. and Sánchez-Vizcaíno, J.M. Social network analysis. Review of general concepts and use in preventive veterinary medicine. *Transboundary and emerging diseases* 2009;56(4):109-120.
- Melott, A.L. Long-term cycles in the history of life: periodic biodiversity in the paleobiology database. *PLoS One* 2008;3(12):e4044.
- Mercader, N. Early steps of paired fin development in zebrafish compared with tetrapod limb development. *Development, growth & differentiation* 2007;49(6):421-437.
- Meyer, A. and Schartl, M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Current opinion in cell biology* 1999;11(6):699-704.
- Midford, P.E., Dececchi, T.A., Balhoff, J.P., Dahdul, W.M., Ibrahim, N., Lapp, H., Lundberg, J.G., Mabee, P.M., Sereno, P.C. and Westerfield, M. The vertebrate taxonomy ontology: a framework for reasoning across model organism and species phenotypes. *Journal of biomedical semantics* 2013;4(1):34.
- Minguez, P., Letunic, I., Parca, L., Garcia-Alonso, L., Dopazo, J., Huerta-Cepas, J. and Bork, P. PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic acids research* 2014;43(D1):D494-D502.
- Müller, G.B. Evo-devo: extending the evolutionary synthesis. *Nature Reviews Genetics* 2007;8(12):943-949.
- Mullins, M.C., Hammerschmidt, M., Kane, D.A., Odenthal, J., Brand, M., Van Eeden, F.J., Furutani-Seiki, M., Granato, M., Haffter, P. and Heisenberg, C.-P. Genes establishing dorsoventral pattern formation in the zebrafish embryo: the ventral specifying genes. *Development* 1996;123(1):81-93.



- Mungall, C.J., Gkoutos, G.V., Smith, C.L., Haendel, M.A., Lewis, S.E. and Ashburner, M. Integrating phenotype ontologies across multiple species. *Genome biology* 2010;11:R2.
- Mungall, C.J., Gkoutos, G.V., Washington, N. and Lewis, S.E. Representing phenotypes in OWL. In, *OWLED proceedings*. 2007.
- Mungall, C.J., McMurtry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J.P., Jacobsen, J.O.B., Keith, D., Laraway, B., Lewis, S.E., NguyenXuan, J., Shefchek, K., Vasilevsky, N., Yuan, Z., Washington, N., Hochheiser, H., Groza, T., Smedley, D., Robinson, P.N. and Haendel, M.A. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research* 2017;45(D1):D712-D722.
- Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E. and Haendel, M.A. Uberon, an integrative multi-species anatomy ontology. *Genome biology* 2012;13(1):R5.
- Nawshad, A. and Hay, E.D. TGF $\beta$ 3 signaling activates transcription of the LEF1 gene to induce epithelial mesenchymal transformation during mouse palate development. *The Journal of cell biology* 2003;163(6):1291-1301.
- Nelson, J.S. Analysis of the multiple occurrence of pelvic fin absence in extant fishes. *Matsya* 1990;15/16:21-38.
- Nelson, J.S. *Fishes of the World*. New York: John Wiley & Sons, Inc.; 2006.
- Nguyen, N.M., Miner, J.H., Pierce, R.A. and Senior, R.M. Laminin  $\alpha$ 5 is required for lobar septation and visceral pleural basement membrane formation in the developing mouse lung. *Developmental biology* 2002;246(2):231-244.
- Odersky, M. The Scala programming language. URL <http://www.scala-lang.org> 2008.
- Ogle, W. Aristotle: On the Parts of Animals. K. Paul, French & Company; 1882.
- Onimaru, K., Marcon, L., Musy, M., Tanaka, M. and Sharpe, J. The fin-to-limb transition as the re-organization of a Turing pattern. *Nature communications* 2016;7:11582.
- Osborne, J.D., Flatow, J., Holko, M., Lin, S.M., Kibbe, W.A., Zhu, L., Danila, M.I., Feng, G. and Chisholm, R.L. Annotating the human genome with Disease Ontology. *BMC Genomics* 2009;10(1):S6.
- Patterson, D.J. Progressing towards a biological names register. *Nature* 2003;422(6933):661.
- Patterson, D.J., Mozzherin, D., Shorthouse, D. and Thessen, A. Challenges with using names to link digital biodiversity information. *Biodiversity Data Journal* 2016;4:e8080.
- Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences* 1999;96(8):4285.
- Peng, J., Bai, K., Shang, X., Wang, G., Xue, H., Jin, S., Cheng, L., Wang, Y. and Chen, J. Predicting disease-related genes using integrated biomedical networks. *BMC Genomics* 2017;18(1):1043.
- Pereira-Leal, J.B., Enright, A.J. and Ouzounis, C.A. Detection of functional modules from protein interaction networks. *PROTEINS: Structure, Function, and Bioinformatics* 2004;54(1):49-57.
- Pesquita, C., Faria, D., Falcão, A.O., Lord, P. and Couto, F.M. Semantic Similarity in Biomedical Ontologies. *PLoS Comput Biol* 2009;5(7):e1000443.
- Pizzuti, C. and Rombo, S.E. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics* 2014;30(10):1343-1352.

- Raff, R.A. Evo-devo: the evolution of a new discipline. *Nature Reviews Genetics* 2000;1(1):74-79.
- Rao, V.S., Srinivas, K., Sujini, G.N. and Kumar, G.N. Protein-protein interaction detection: methods and analysis. *International journal of proteomics* 2014;2014.
- Raspopovic, J., Marcon, L., Russo, L. and Sharpe, J. Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science* 2014;345(6196):566-570.
- Redelings, B.D. and Holder, M.T. A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. *PeerJ* 2017;5:e3058.
- Rees, J. and Cranston, K. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal* 2017;5:e12581.
- Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007* 1995.
- Revell, L.J. Phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 2012;3(2):217-223.
- Robertson, D.L. and Lovell, S.C. Evolution in protein interaction networks: co-evolution, rewiring and the role of duplication. In.: Portland Press Limited; 2009.
- Rodríguez Mendoza, R.P. Otoliths and their applications in fishery science. *Croatian Journal of Fisheries: Ribarstvo* 2006;64(3):89-102.
- Routledge, R. Fisher's exact test. *Encyclopedia of biostatistics* 2005.
- Rubinov, M. and Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 2010;52(3):1059-1069.
- Salter-Townshend, M., White, A., Gollini, I. and Murphy, T.B. Review of statistical network analysis: models, algorithms, and software. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 2012;5(4):243-264.
- Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvermin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E. and Ye, J. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 2009;37(Database issue):D5-15.
- Schlicker, A., Domingues, F.S., Rahnenführer, J. and Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics* 2006;7(1):302.
- Schwikowski, B., Uetz, P. and Fields, S. A network of protein-protein interactions in yeast. *Nat Biotech* 2000;18(12):1257-1261.
- Shameer, K., Naika, M.B., Mathew, O.K. and Sowdhamini, R. POEAS: Automated Plant Phenomic Analysis Using Plant Ontology. *Bioinformatics and biology insights* 2014;8:209.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 2003;13(11):2498-2504.
- Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M. and Ideker, T. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America* 2005;102(6):1974-1979.

- Sharan, R., Ulitsky, I. and Shamir, R. Network-based prediction of protein function. *Molecular systems biology* 2007;3(1):88.
- Sheth, R., Marcon, L., Bastida, M.F., Junco, M., Quintana, L., Dahn, R., Kmita, M., Sharpe, J. and Ros, M.A. Hox genes regulate digit patterning by controlling the wavelength of a Turing-type mechanism. *Science* 2012;338(6113):1476-1480.
- Shoemaker, B.A. and Panchenko, A.R. Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLoS computational biology* 2007;3(3):e42.
- Shoemaker, B.A. and Panchenko, A.R. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS computational biology* 2007;3(4):e43.
- Shubin, N. Your inner fish: a journey into the 3.5-billion-year history of the human body. Vintage; 2008.
- Shui, Y. and Cho, Y.-R. Alignment of PPI Networks Using Semantic Similarity for Conserved Protein Complex Prediction. *IEEE transactions on nanobioscience* 2016;15(4):380-389.
- Smith, C.L. and Eppig, J.T. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mammalian Genome* 2012;23(9-10):653-668.
- Smith, K.A., Chocron, S., von der Hardt, S., de Pater, E., Soufan, A., Bussmann, J., Schulte-Merker, S., Hammerschmidt, M. and Bakkers, J. Rotation and asymmetric development of the zebrafish heart requires directed migration of cardiac progenitor cells. *Developmental cell* 2008;14(2):287-297.
- Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L. and Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011;27(3):431-432.
- Sordino, P. and Duboule, D. A molecular approach to the evolution of vertebrate paired appendages. *Trends in Ecology & Evolution* 1996;11(3):114-119.
- Stoick-Cooper, C.L., Weidinger, G., Riehle, K.J., Hubbert, C., Major, M.B., Fausto, N. and Moon, R.T. Distinct Wnt signaling pathways have opposing roles in appendage regeneration. *Development* 2007;134(3):479.
- Sukumaran, J. and Holder, M.T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 2010;26(12):1569-1571.
- Sukumaran, J. and Holder, M.T. The DendroPy Phylogenetic Computing Library Documentation. In.; 2018.
- Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE* 2011;6(7):e21800.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J. and von Mering, C. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research* 2015;43(D1):D447-D452.
- Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J. and von Mering, C. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research* 2017;45(D1):D362-D368.
- Tang, X., Wang, J., Liu, B., Li, M., Chen, G. and Pan, Y. A comparison of the functional modules identified from time course and static PPI network data. *BMC bioinformatics* 2011;12(1):339.

- Tang, X., Wang, J., Zhong, J. and Pan, Y. Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2014;11(2):407-418.
- Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q. and Wrana, J.L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotech* 2009;27(2):199-204.
- Thomson, R.C. and Shaffer, H.B. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Systematic Biology* 2010;59(1):42-58.
- Threadgill, D.W., Dlugosz, A.A., Hansen, L.A., Tennenbaum, T., Lichti, U., Yee, D., LaMantia, C., Mourton, T., Herrup, K. and Harris, R.C. Targeted disruption of mouse EGF receptor: effect of genetic background on mutant phenotype. *Science* 1995;269(5221):230-234.
- To, M.D., Wong, C.E., Karnezis, A.N., Del Rosario, R., Di Lauro, R. and Balmain, A. Kras regulatory elements and exon 4A determine mutation specificity in lung cancer. *Nature genetics* 2008;40(10):1240.
- Tripathi, S., Moutari, S., Dehmer, M. and Emmert-Streib, F. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC bioinformatics* 2016;17:129.
- Tripathi, S., Moutari, S., Dehmer, M. and Emmert-Streib, F. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC bioinformatics* 2016;17(1):129.
- Ulitsky, I. and Shamir, R. Identification of functional modules using network topology and high-throughput data. *BMC systems biology* 2007;1(1):8.
- Van Rossum, G. and Drake, F.L. The python language reference manual. Network Theory Ltd.; 2011.
- Vares, G., Wang, B., Tanaka, K., Shang, Y., Fujita, K., Hayata, I. and Neno, M. Trp53 activity is repressed in radio-adapted cultured murine limb bud cells. *Journal of radiation research* 2011;52(6):727-734.
- Vespignani, A. Evolution thinks modular. *Nature genetics* 2003;35.
- von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. and Bork, P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research* 2005;33(Database issue):D433-437.
- von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;417(6887):399-403.
- Vorobyeva, E.I. and Schultze, H.-P. Description and systematics of panderichthyid fishes with comments on their relationship to tetrapods. *Origins of the higher groups of tetrapods: controversy and consensus* 1991:68-109.
- Vos, R.A., Balhoff, J.P., Caravas, J.A., Holder, M.T., Lapp, H., Maddison, W.P., Midford, P.E., Priyam, A., Sukumaran, J., Xia, X. and Stoltzfus, A. NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Systematic Biology* 2012;61(4):675-689.
- Wang, B., Ohyama, H., Haginoya, K., Odaka, T., Itsukaichi, H., Yukawa, O., Yamada, T. and Hayata, I. Adaptive response in embryogenesis. III. Relationship to radiation-induced apoptosis and Trp53 gene status. *Radiation research* 2000;154(3):277-282.

- Wang, J., Ma, Z., Carr, S.A., Mertins, P., Zhang, H., Zhang, Z., Chan, D.W., Ellis, M.J.C., Townsend, R.R., Smith, R.D., McDermott, J.E., Chen, X., Paulovich, A.G., Boja, E.S., Mesri, M., Kinsinger, C.R., Rodriguez, H., Rodland, K.D., Liebler, D.C. and Zhang, B. Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. *Molecular & Cellular Proteomics* 2017;16(1):121-134.
- Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. and Chen, C.F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;23(10):1274-1281.
- Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S. and Sabatini, D.M. Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. *Cell* 2017;168(5):890-903.
- Wang, X., Gulbahce, N. and Yu, H. Network-based methods for human disease gene prediction. *Briefings in Functional Genomics* 2011;10(5):280-293.
- Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M. and Lewis, S.E. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology* 2009;7(11):2662.
- Wehner, D., Cizelsky, W., Vasudevaro, Mohankrishna D., Özhan, G., Haase, C., Kagermeier-Schenk, B., Röder, A., Dorsky, Richard I., Moro, E., Argenton, F., Kühl, M. and Weidinger, G. Wnt/ $\beta$ -Catenin Signaling Defines Organizing Centers that Orchestrate Growth and Differentiation of the Regenerating Zebrafish Caudal Fin. *Cell Reports* 2014;6(3):467-481.
- Westerfield, M., Doerry, E., Kirkpatrick, A.E. and Douglas, S.A. Zebrafish informatics and the ZFIN database. In, *Methods in cell biology*. Elsevier; 1998. p. 339-355.
- Willer, G.B., Lee, V.M., Gregg, R.G. and Link, B.A. Analysis of the zebrafish perplexed mutation reveals tissue-specific roles for de novo pyrimidine synthesis during development. *Genetics* 2005;170(4):1827-1837.
- Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016.
- Wuchty, S., Oltvai, Z.N. and Barabási, A.L. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature genetics* 2003;35.
- Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M. and Eisenberg, D. DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res.* 2001;29(1):239-241.
- Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Hönigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., Richter, L., Ashkenazy, H., Punta, M., Schlessinger, A., Bromberg, Y., Schneider, R., Vriend, G., Sander, C., Ben-Tal, N. and Rost, B. PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic acids research* 2014;42(W1):W337-W343.
- Yamada, T. and Bork, P. Evolution of biomolecular networks [mdash] lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol* 2009;10(11):791-803.
- Yamanoue, Y., Setiamarga, D.H. and Matsuura, K. Pelvic fins in teleosts: structure, function and evolution. *Journal of fish biology* 2010;77(6):1173-1208.
- Yan, Y.-L., Willoughby, J., Liu, D., Crump, J.G., Wilson, C., Miller, C.T., Singer, A., Kimmel, C., Westerfield, M. and Postlethwait, J.H. A pair of Sox: distinct and overlapping functions of zebrafish sox9 co-orthologs in craniofacial and pectoral fin development. *Development* 2005;132(5):1069.

- Ye, P., Peyser, B.D., Pan, X., Boeke, J.D., Spencer, F.A. and Bader, J.S. Gene function prediction from congruent synthetic lethal interactions in yeast. *Molecular systems biology* 2005;1(1).
- Yeyati, P.L., Bancewicz, R.M., Maule, J. and van Heyningen, V. Hsp90 Selectively Modulates Phenotype in Vertebrate Development. *PLoS Genetics* 2007;3(3):e43.
- Yokoi, H., Nishimatsu, A., Ozato, K. and Yoda, K. Cloning and embryonic expression of six wnt genes in the medaka (*Oryzias latipes*) with special reference to expression of wnt5a in the pectoral fin buds. *Development, Growth & Differentiation* 2003;45(1):51-61.
- Yu, D., Kim, M., Xiao, G. and Hwang, T.H. Review of biological network data and its applications. *Genomics & informatics* 2013;11(4):200-210.
- Yu, G. GO-terms Semantic Similarity Measures. *Bioinformatics* 2010;26(7):976-978.
- Zeng, E., Ding, C., Mathee, K., Schnepfer, L. and Narasimhan, G. Gene Function Prediction and Functional Network: The Role of Gene Ontology. In: Holmes, D.E. and Jain, L.C., editors, *Data Mining: Foundations and Intelligent Paradigms*. Springer Berlin Heidelberg; 2012. p. 123-162.
- Zeng, E., Ding, C., Narasimhan, G. and Holbrook, S.R. Estimating support for protein-protein interaction data with applications to function prediction. In, *Computational Systems Bioinformatics: (Volume 7)*. World Scientific; 2008. p. 73-84.
- Zhang, C., Freddolino, P.L. and Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic acids research* 2017;45(W1):W291-W299.
- Zhang, J., Wagh, P., Guay, D., Sanchez-Pulido, L., Padhi, B.K., Korzh, V., Andrade-Navarro, M.A. and Akimenko, M.A. Loss of fish actinotrichia proteins and the fin-to-limb transition. *Nature* 2010;466(7303):234-237.
- Zhang, M., Zhang, J., Lin, S.-C. and Meng, A.  $\beta$ -Catenin 1 and  $\beta$ -catenin 2 play similar and distinct roles in left-right asymmetric development of zebrafish embryos. *Development* 2012;dev-074435.
- Zhang, S., Ning, X. and Zhang, X.-S. Identification of functional modules in a PPI network by clique percolation clustering. *Computational biology and chemistry* 2006;30(6):445-451.
- Zickenrott, S., Angarica, V.E., Upadhyaya, B.B. and Del Sol, A. Prediction of disease-gene-drug relationships following a differential network analysis. *Cell death & disease* 2017;7(1):e2040.
- Zuzarte-Luis, V., Montero, J.A., Rodriguez-Leon, J., Merino, R., Rodriguez-Rey, J.C. and Hurle, J.M. A new role for BMP5 during limb development acting through the synergic activation of Smad and MAPK pathways. *Developmental biology* 2004;272(1):39-52.