Gene network module changes associated with the vertebrate fin to limb transition

Pasan C Fernando[1*], Paula M Mabee[2] and Erliang Zeng[3,4,5,6*]

[1] Department of Biology, University of South Dakota, Vermillion, SD, USA; Current affiliation: Department of Plant Sciences, University of Colombo, Sri Lanka

[2] Department of Biology, University of South Dakota, Vermillion, SD, USA; Current affiliation: National Ecological Observatory Network, 1625 38th St. #100, Boulder, CO 80301, USA

[3] Division of Biostatistics and Computational Biology, College of Dentistry, University of Iowa, Iowa City, IA, USA

[4] Department of Preventive & Community Dentistry, College of Dentistry, University of Iowa, Iowa City, IA, USA

[5] Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, IA, USA

[6] Department of Biomedical Engineering, College of Engineering, University of Iowa, Iowa City, IA, USA

[*] **Correspondence:** pasanfernando@pts.cmb.ac.lk and erliang-zeng@uiowa.edu

Keywords: gene networks, gene modules, evolution, fin to limb transition, network comparison

Abstract

Evolutionary phenotypic transitions, such as the fin to limb transition in vertebrate evolution, result from changes in associated genes and their interactions, often in response to changing environment. Identifying the associated changes in gene networks is vital to achieve a better understanding of these transitions. Previous experimental studies have been typically limited to manipulating a small number of genes. To expand the number of analyzed genes and hence, biological knowledge, we computationally isolated and compared the gene modules for paired fins (pectoral fin, pelvic fin) of fishes (zebrafish) to those of the paired limbs (forelimb, hindlimb) of mammals (mouse) using quality-enhanced gene networks from zebrafish and mouse. We ranked module genes according to their weighted-degrees and identified the highest-ranking hub genes, which were important for the module stability. Further, we identified genes conserved during the fin to limb transition and investigated the fates of zebrafish-specific and mouse-specific module genes in relation to their involvements in newly emerged or lost anatomical structures during the aquatic to terrestrial vertebrate transition. This paper presents the results of our investigations and demonstrates a general network-based computational workflow to study evolutionary phenotypic transitions involving diverse model organisms and anatomical entities.

1. Introduction

Phenotypes, such as fin development and limb development, are the result of multiple genes working together in complex biological pathways [1, 2]. Evolutionary modifications in phenotypes due to environmental or other changes involve rewiring gene interactions and their involvements in pathways [2, 3]. Most often, it is likely the network of multiple protein interactions rather than the contribution of a single protein that determines the resulting phenotype [1, 4]. Therefore, investigating the collection of genes and their interactions, i.e., modular gene structure [1], underlying phenotypes is important in evolutionary biology to understand the evolutionary mechanisms that drive phenotypic changes. Gene module analysis has become common in bioinformatics, and the concept of modular evolution has emerged to explain the changes in groups of genes rather than a single gene when studying the evolution of organisms [5-7]. However, most of these studies have focused on smaller protein complexes (typically containing less than 20 proteins) that determine molecular functions and biological pathways [8-10]. Phenotypes, such as fin and limb development, are resulted by a large number of proteins having diverse molecular functions and belonging to several biological pathways. Even the few protein network studies that focus on phenotypes have targeted human diseases [11, 12], and to our knowledge, there have been no evolutionary studies of modules to understand evolutionary phenotypic transitions. As there have been important anatomical changes associated with the vertebrate evolution, such as the fin to limb transition, module evolution studies for anatomical changes are essential to unravel new evolutionary information, which serves as the motivation for our work.

The fin to limb transition is an iconic anatomical change associated with the evolution of terrestrial vertebrates from aquatic fish-like ancestors [13, 14]. According to fossil record, the

transformation of fishes into land vertebrates began in the Devonian, 365-408 million years ago [13, 15]. This transformation is associated with many phenotypic changes in addition to the fin to limb transition, including changes in the cranial and axial skeleton [13]. The relationship between homologous anatomical structures of land and aquatic vertebrates is evident from several similar characteristics. For instance, the pectoral fin endoskeleton of panderichthyid fish fossils shows significant similarities with the limb skeletons of terrestrial vertebrates (tetrapods) such as the presence of a proximal humerus and two distal bones [14]. Such evidence indicates that forelimbs and hindlimbs of tetrapods are homologous to pectoral and pelvic fins of fishes, respectively.

Identifying the genetic changes associated with the fin to limb transition is a prominent topic in evolutionary biology [16, 17]. Many wet lab experiments have demonstrated the evolutionary importance of genes such as *shh* [14, 16]. Few computational studies, however, have been targeted on the fin to limb transition [17]. The recent availability of large PPI networks and the ability to perform module analysis through the advancement of network algorithms provide an opportunity for a new perspective on genetic changes associated with the fin to limb transition.

Graph theoretic methods are critical to the study of networks in biology. These methods are enabled by biological knowledge that is represented as computational graphs, such as protein-protein interaction (PPI) networks and biological ontologies. In graph theory, a module is defined as a set of nodes that are highly connected internally and sparsely connected with external nodes [1]. These network modules usually correspond to biological functions that contribute to phenotypes; hence, they are often referred to as ‘functional modules’ in biological vocabulary [6, 18].

There are a number of functional module detection algorithms that can be used to detect modules in a graph [5]. Some methods, such as graph partitioning [19], only consider the network structure and do not require any prior information. For modules that are known to involve a large number of genes in complex phenotypes, it is beneficial to perform module detection using prior knowledge as computational constraints [1, 4, 20]. These methods start from a set of known genes for a given phenotype and expand the module based on the 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 associated with the known phenotype are included in the module [1]. However, this method has proven to yield many false positives [1]. Therefore, network-based candidate gene prediction algorithms such as the Hishigaki method [21] and label propagation algorithm [22], which have been shown to be more accurate [1, 4, 21], are often used to predict new candidate genes for inclusion in a module.

One purpose of network analysis is the identification of hub genes, which are defined as important genes that are central to the stability of the module [23, 24]. Hub genes have a higher number of interactions (degrees) than other genes in the module. Their removal is most likely to disrupt the module organization, and thus the biological function(s) or phenotype(s) that is governed by the module. Using network analysis, a set of genes for a function or a phenotype can be transformed into a ranked list that is sorted based on their importance in the module. Usually, the number of interactions a gene forms within the module (degree) is used for the ranking [24].

The quality of PPI network data has been a problematic issue in previous network analyses because of the large portion of spurious PPIs generated by experimental methods, such as high-throughput yeast two-hybrid assay [4, 25]. Therefore, in our previous work [25], we

improved the quality of the PPI networks retrieved from the STRING database (STRING, RRID:SCR_005223) [26] by integrating existing experimental knowledge about gene-anatomy relationships available in literature using Uberon anatomy ontology [27]. First, semantic anatomy-based gene networks were generated by calculating the semantic similarity between anatomy terms annotated to different genes, and then, these semantic networks were integrated with the PPI networks for zebrafish and mouse, which improved the candidate gene prediction accuracy for anatomical entities [25]. In this study, we use these improved integrated networks to obtain the most accurate modules.

When considering the evolution of functional modules, most studies have focused on identifying the genes that are retained during evolution, i.e., conserved genes, and their organization in the respective modules [8, 10]. It has been hypothesized that gradual modular changes occur in evolution while maintaining the basic modular structure; this is because dramatic changes in gene interactions may destroy the proper function of an organism [7]. In support of this hypothesis, conserved genes are observed to play an important role in maintaining the stability of the gene modules during evolution [7, 8, 10]. 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 also identified as hub genes [7, 8]. While such conserved module genes may play a role in maintaining gene module structure, species-specific module genes that have been recruited or removed during the evolution may play important roles that contribute to evolutionary transitions [28].

In this work, our goal is to compare PPI network modules associated with fins and limbs to identify the genetic changes, such as the changes in involved genes and their importance,

which led to the anatomical changes that characterize the evolution of fins to limbs. From this analysis, we identify genes that are conserved between fins and limbs to understand their roles in the modular evolution, and we predict novel gene candidates with no previously known contributions to the development of paired fins or paired limbs. Further, we identify fin module-specific and limb module-specific genes and investigate their evolutionary roles. This work suggests some evolutionary hypotheses regarding the role of conserved genes versus fin or limb specific genes in the many evolutionary changes in these animals. Finally, this study demonstrates a general network-based computational model to perform gene module comparisons for evolutionary phenotypic transitions.

2. Methods

(a) Selection of the integrated networks for module detection

Based on our previous work [25] of network-based candidate gene prediction using quality-enhanced PPI networks that were generated by four semantic similarity methods (Lin, Resnik, Schlicker, and Wang), the best performing gene networks for zebrafish and the mouse were selected for this project. These are referred to as ‘zebrafish integrated network’ and ‘mouse integrated network’ from herein.

(b) Detection of network modules

For module detection, genes with direct annotations to the pectoral fin (UBERON:0000151), forelimb (UBERON:0002102), pelvic fin (UBERON:0000152), and hindlimb (UBERON:0002103) were used as prior information and their anatomical profiles were extracted from the Monarch Initiative repository (https://monarchinitiative.org/; RRID:SCR_000824)

(06/20/2018)[29]. In addition, genes that were annotated to the parts (e.g., pectoral fin lepidotrichium and pectoral fin radial skeleton are parts of the pectoral fin) and the developmental precursors (pectoral fin bud, pelvic fin bud, forelimb bud, and hindlimb bud) of the above entities were extracted using the Uberon anatomy ontology relationships. The genes directly annotated to the anatomical entity of interest or annotated to a part or developmental precursor of the entity are collectively referred to as 'genes with original annotations'.

Beginning with the genes with original annotations, gene modules for the anatomical entities of interest were identified by predicting novel genes using the Hishigaki network-based candidate gene prediction method [21, 25]. First, the network-based candidate gene prediction performance for each anatomical entity of interest was evaluated using leave-one-out cross-validation [25], and ROC and precision-recall curves were generated. Then, a prediction precision threshold was used to predict new candidate genes. A trial and error method was used to select the best precision threshold for each gene module.

After predicting the candidate genes, the modules were extracted for the pectoral fin and the pelvic fin from the zebrafish integrated network and for the forelimb and the hindlimb from the mouse integrated network. The extracted modules were visualized using the Cytoscape software [30] (Cytoscape, RRID:SCR_003032).

(c) Validation of the predicted genes

The predicted candidate genes could be validated using either experimental methods, such as gene knockdown [31], or computational methods such as the one used in this work. First, 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 and *vice versa* to determine whether they were annotated to a homologous anatomical entity.

Second, enrichment analyses were performed to confirm for each module, whether the predicted genes shared similar Biological Process terms from Gene Ontology (GO-BP) as the genes with original annotations. Enrichment analyses were also performed to confirm for each module, whether the predicted genes shared similar Uberon anatomy annotations as the genes with original annotations.

Third, the weighted degree distributions of the predicted genes were compared with the weighted degree distributions of the genes with original annotations in each module. If the predicted genes have a higher weighted degree distribution, it indicates that the predicted genes have a similar or a higher importance as genes with original annotations.

(d) Comparison of the network modules

To study the fin to limb transition and identify the modular changes, the pectoral fin and pelvic fin modules of the zebrafish were compared with the forelimb and hindlimb modules of the mouse, respectively.

Teleost fishes, such as the zebrafish, have more genes than tetrapods, such as the mouse. A whole genome duplication event is proposed to have occurred at the origin of actinopterygian fishes, i.e., the teleost genome duplication [32]; hence, most of the mouse genes have duplicated copies in the zebrafish. To perform the module comparison, the gene ortholog mappings between mouse and zebrafish genes were retrieved from the Zebrafish Information Network [33] (ZFIN, 06/26/2018) (https://zfin.org/downloads) (Zebrafish Information Network, RRID:SCR_002560). During the comparison, if multiple zebrafish orthologs were present in a zebrafish module for a

single mouse gene, all zebrafish orthologs were retained. 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 network analysis, the degree of a gene (the number of interactions of the gene) is often used as an important metric [10, 24]. Genes with higher degrees in a module, i.e., hub genes, are considered 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 [23]. When analyzing networks with weights assigned for interactions (weighted networks), such as the integrated networks used here, weighted degree is preferred over the degree because it considers the different interaction weights rather (equation 1) than counting the number of interactions for a specific node [24].

$$Weighted\ degree = \sum_{v \in n(u)} sim(v,u) \qquad (1)$$

In equation 1, *n(u)* is the neighborhood of the gene of interest (*u*) and *v* iterates through all the neighbors of gene *u*. The gene similarity score for the interaction between genes *v* and *u*, which is represented by *sim(v,u)*, is used for the interaction weight. Weighted degree of gene *u* is the summation of all weights of interactions between gene *u* and all its neighbors.

The weighted degree for each gene in a module was calculated, and the genes were ranked accordingly. During the comparisons, the weighted degree of each zebrafish module gene was compared with the corresponding mouse ortholog. However, due to the size differences of the zebrafish and mouse modules, the weighted degree of each gene had to be normalized by the total number of genes in each module. Then, normalized weighted degree distributions for conserved genes, zebrafish module-specific genes, and mouse module-specific genes were

compared for pectoral fin *versus* forelimb and pelvic fin *versus* hindlimb to study the relative importance of genes in each group.

The fate of the zebrafish module-specific genes in mouse was investigated by extracting mouse orthologs for the pectoral and pelvic fin module-specific genes and performing enrichment analyses using Uberon and GO-BP terms. Similarly, the roles of the mouse module-specific genes in zebrafish were investigated using zebrafish orthologs for the forelimb and hindlimb module-specific genes. The DAVID (https://david.ncifcrf.gov/) (DAVID, RRID:SCR_001881) online functional enrichment analysis tool was used to perform gene set enrichment analysis using GO-BP terms. DAVID uses Fisher's exact test [34] to perform enrichment analyses. Although the GO is widely used for enrichment analysis, anatomy ontologies are rarely used. To perform enrichment analysis using the Uberon anatomy ontology and Fisher's exact test, a Python program (Uberon enrichment analysis program) was developed and used. Ontology terms with p-values less than 0.05 were considered as enriched terms.

3. Results and discussion

(a) Selection of the integrated networks for module detection

The integrated networks generated using the Lin and Schlicker methods were selected for module detection for zebrafish and mouse, respectively because they outperformed other integrated networks based on the results of our previous work [25]. The zebrafish Lin integrated network contained 17,394 genes and 730,855 interactions and the mouse Schlicker integrated network contained 18,002 genes and 613,671 interactions [25].

(b) Detection of network modules

The statistics showing the number of genes with original annotations to each anatomical entity are given in electronic supplementary material, table S1. The total number of genes for the pectoral fin (198) and the forelimb (267) were comparatively similar than the total number of genes for the pelvic fin (15) and the hindlimb (777). Detection of the pelvic fin module was challenging because of the low number of original gene annotations. 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 [35]. This may have been a potential reason for fewer annotations to the pelvic fin; the studied gene disruptions may have killed the larval zebrafish before the pelvic fin develops or the larvae may have been sacrificed at a pre-determined early stage.

The ROC and precision-recall curves generated for each anatomical entity during the network-based candidate gene prediction evaluations are given in electronic supplementary material, figures S1 and S2, respectively. According to the curves, all anatomical entities except the pelvic fin show high accuracies for network-based candidate gene predictions (the AUC values of ROC curves were higher than 0.85). This shows the high reliability of the network candidate gene predictions. The lower performance for the pelvic fin could be due to its low number of original gene annotations. It has been shown that the prediction accuracy improves with the size of the dataset/number of gene annotations, and anatomical entities with a low number of gene annotations can lead to lower AUC values [36].

The statistics for the extracted gene modules are given in electronic supplementary material, table S1. The genes with original annotations that were lost during the module extraction are listed in electronic supplementary material, table S2. 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 make the number of genes in the pelvic fin and the forelimb modules approximately similar.

The visualizations of the resulting modules for the pectoral fin, pelvic fin, forelimb, and hindlimb are given in electronic supplementary material, figures S3, S4, S5, and S6, respectively. The companion Cytoscape network files for these modules are available in electronic supplementary material, files S1, S2, S3, and S4. The genes in the pectoral fin, pelvic fin, forelimb, and hindlimb modules ranked based on the weighted degree are listed in electronic supplementary material, files S5, S6, S7, and S8, respectively.

(c) Validation of the predicted genes

The list of predicted genes for pectoral fin, pelvic fin, forelimb, and hindlimb modules are given in electronic supplementary material, tables S3, S4, S5, and S6, respectively. Of the 45 predicted genes for the pectoral fin, 14 genes had mouse orthologs that were associated with the forelimb (9 direct annotations, 2 annotations only to the parts or the developmental precursors, and 3 predicted genes). Of the 605 predicted genes for the pelvic fin, 78 genes had mouse orthologs that were associated with the hindlimb (46 direct annotations, 20 annotations only to the parts or the developmental precursors, and 12 predicted genes). Of the 18 predicted genes for the forelimb, 6 genes had mouse orthologs that were associated with the pectoral fin (2 direct annotations, 1 annotation only to the parts or the developmental precursors, and 3 predicted genes). Of the 32 predicted genes for the hindlimb, 12 genes had mouse orthologs that were associated with the pelvic fin (all 12 were predicted genes). These results indicate that the orthologs of the predicted genes are annotated to homologous anatomical entities, providing a certain level of validation for the predicted genes.

The enriched GO-BP terms that are common to the predicted genes and genes with original annotations to pectoral fin, pelvic fin, forelimb, and hindlimb are listed in electronic supplementary material, tables S7, S8, S9, and S10, 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 electronic supplementary material, tables S11, S12, S13, and S14, respectively. There were several common enriched GO-BP terms for all the modules, some of which were related to paired fins and limbs, such as pectoral fin development, fin development, embryonic limb morphogenesis, embryonic digit morphogenesis. Some of the common enriched Uberon terms, such as median fin fold, ventral fin fold, caudal fin, appendicular skeleton and limb, were related with fin or limb development.

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 figure 1. In all the modules, the weighted degree distributions of the predicted genes were higher than the genes with original annotations. This indicates that predicted genes as a group are important in the modules and central to the function of the modules, which supports the biological significance of the predicted genes.

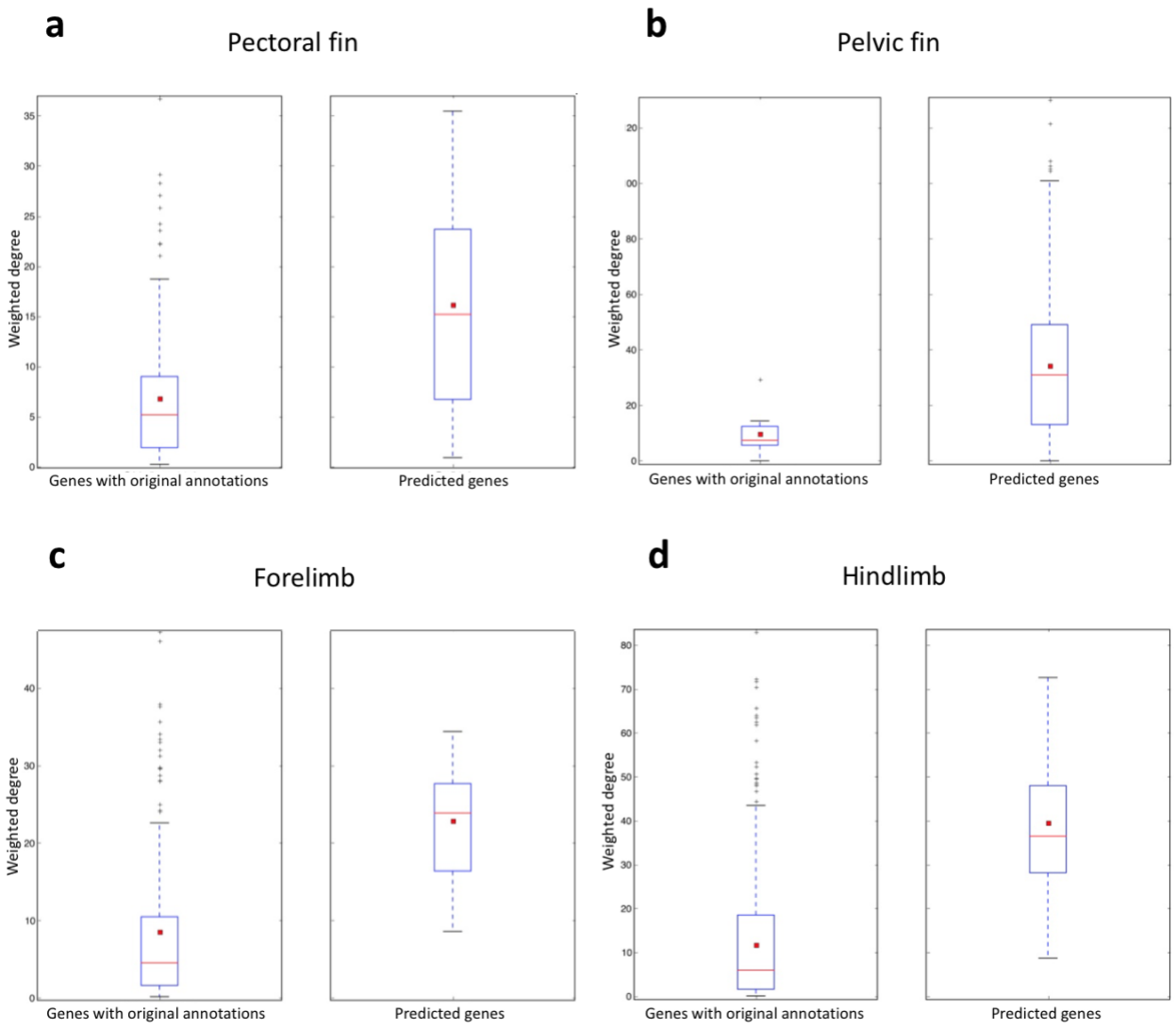


Figure 1. The boxplot comparisons of the weighted degree distributions for the predicted genes *versus* genes with original annotations for each module. In the boxplots, the red line and the square represent the median and mean, respectively.

(d) Comparison of the network modules

(i) Pectoral fin and forelimb comparison

317 According to the comparison, 183 genes were specific to the pectoral fin module, 207 genes
318 were specific to the forelimb module. 37 genes were shared (conserved genes) between the
319 pectoral fin and forelimb (electronic supplementary material, table S15, figure 2).

320 In the pectoral fin module, the top-ranked hub gene based on the weighted degree was
321 *shha* (sonic hedgehog a) (electronic supplementary material, file S5), whose role has been well-
322 documented in pectoral fin development [16]. Its ortholog, *Shh*, is important in the development
323 and morphogenesis of limbs in tetrapods including humans [37], and it was also highly ranked in
324 the forelimb module (4^{th}, see electronic supplementary material, table S15). The loss or gain of
325 activity in the sonic hedgehog signaling pathway in tetrapods results in lost, gained, or
326 malformed limbs [37]. The *shh* gene has long been considered an important gene associated with
327 fin to limb transition because it is important in the morphological patterning of paired fins and
328 limbs [14].

329 The highest-ranking gene in the forelimb module was *bmp4* (bone morphogenetic protein
330 4), another gene closely associated with limb formation and morphogenesis in tetrapods [38].
331 Mutations in *bmp4* affect the *bmp4* signaling pathway to cause abnormalities in limb and digit
332 formation in tetrapods [38]. *Bmp4* was ranked 2^{nd} in the pectoral fin module (electronic
333 supplementary material, table S15) and was predicted during module detection.

334 When considering the conserved genes (figure 2), some of the important hub genes in the
335 pectoral fin module, such as *shha*, *bmp4*, *bmp2b*, and *bmp7a*, had retained their importance
336 demonstrated by their higher ranks based on the weighted degree in the forelimb module
337 (electronic supplementary material, table S15). Other genes such as *sox9* were elevated in rank
338 during the transition from pectoral fin to forelimb. In the pectoral fin module, *sox9a* and *sox9b*
339 genes were ranked 83^{rd} and 104^{th}, respectively, while in the mouse, the ortholog *sox9* was

elevated to 15th (electronic supplementary material, table S15). *Sox9* is well known to be involved with digit patterning in the limbs of tetrapods due to its participation in the a bmp-sox9-wnt Turing network [17, 39]. Because digits emerged after the transition from fins to limbs [13, 14], the involvement of *sox9* in a digit patterning pathway could have increased the number of interactions with other genes in the forelimb module, and hence, the increased importance.

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 are given in figure 3. The conserved genes in both modules have higher normalized weighted degree distributions compared to the respective module-specific genes. This indicates that the conserved genes share more interactions within the module as a group and are more central to modular stability. From an evolutionary point of view, during the transition from pectoral fin to the forelimb, it appears that genes with higher degrees in the pectoral fin module, such as *shha*, *bmp4*, were conserved in the forelimb and new forelimb module-specific genes were recruited surrounding those conserved genes.

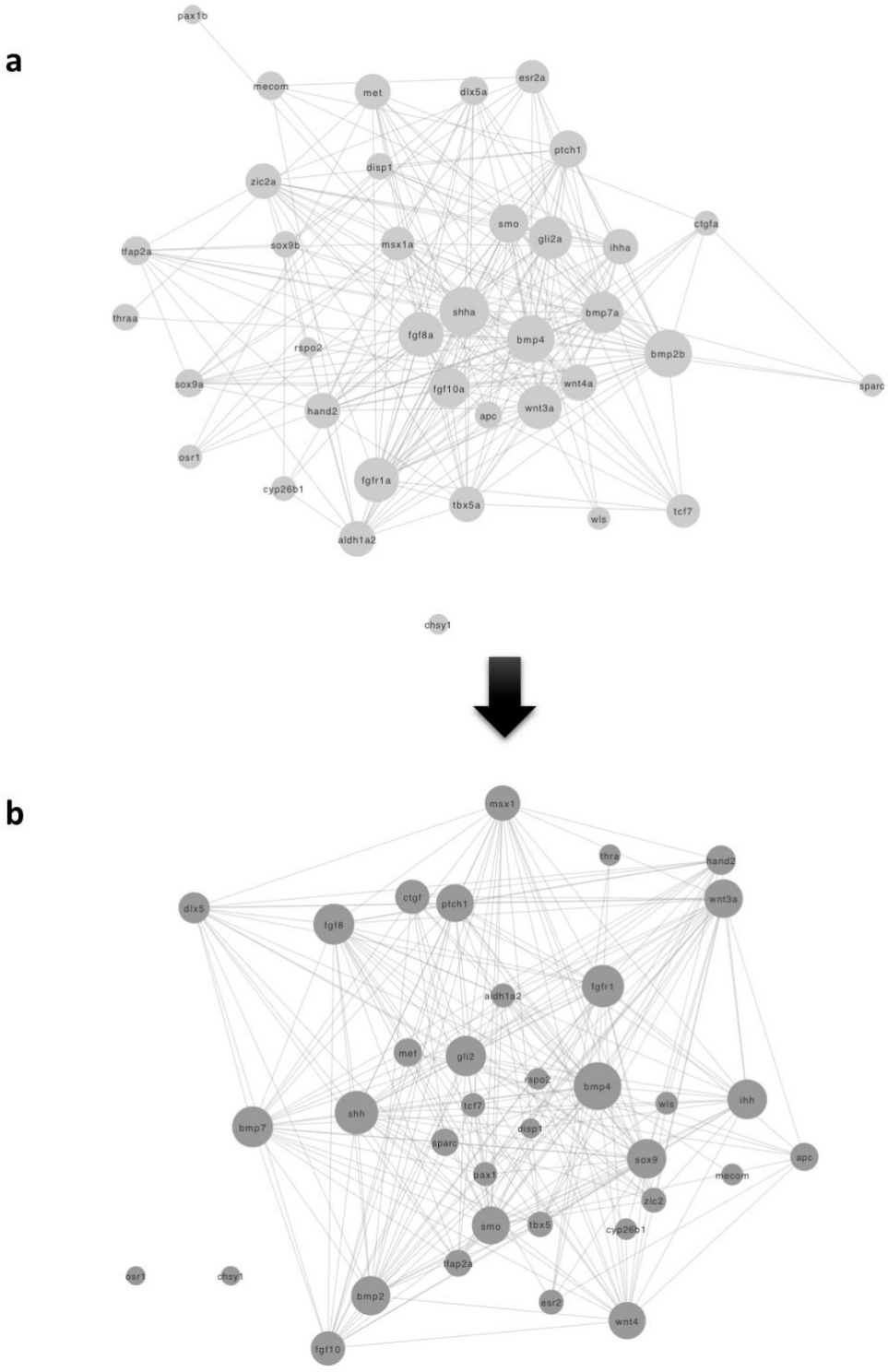


Figure 2. Networks of the 37 conserved genes that are common to and extracted from (a) the pectoral fin module and (b) the forelimb module. Node size is proportional to the degree (number

of interactions) of the gene. Hub genes, such as *bmp4*, *shh*, *smo*, *bmp7, sox9*, and *gli2*, are shown in larger node sizes. The arrow represents the direction of modular evolution.

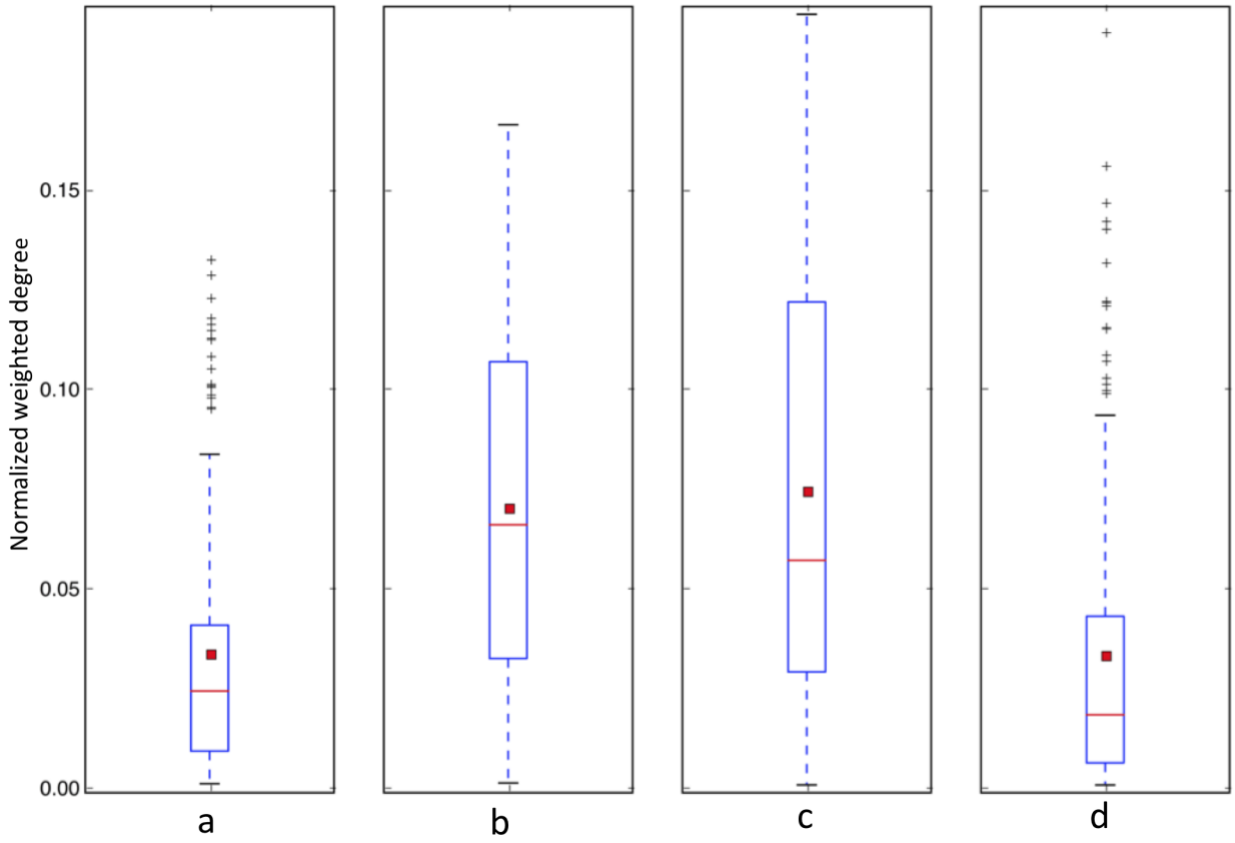


Figure 3. Boxplot comparison of normalized weighted degree distributions for (a) pectoral fin module-specific genes, (b) pectoral fin conserved genes, (c) forelimb conserved genes, and (d) forelimb module-specific genes. In the boxplots, the red line and the square represent the median and mean, respectively.

(ii) Pelvic fin and hindlimb comparison

According to the comparison, 536 genes were specific to the pelvic fin module, 601 genes were specific to the hindlimb module, and 81 genes were conserved between pectoral fin and forelimb modules. (electronic supplementary material, table S16 and figure 4).

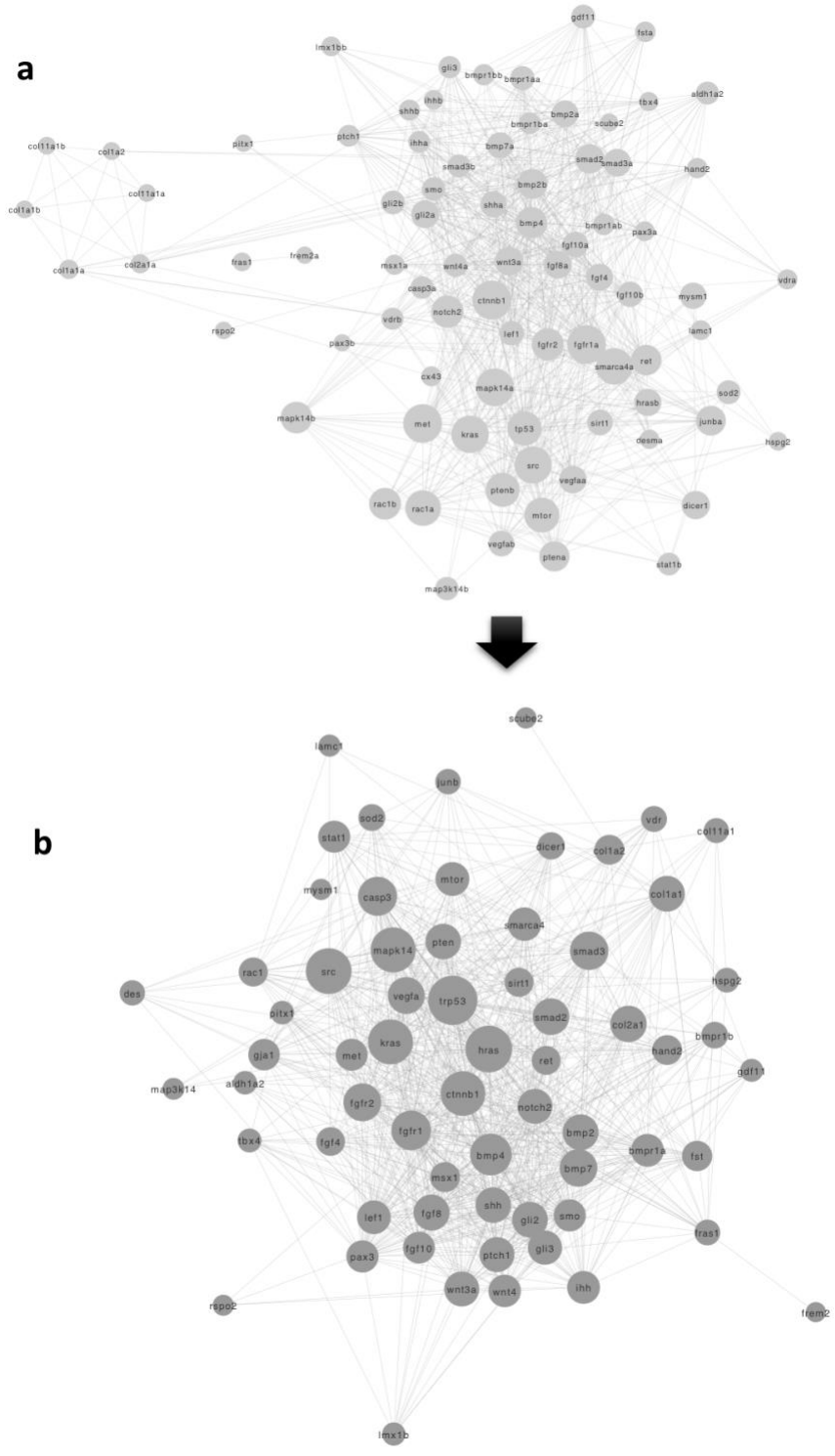
In the pelvic fin module, the highest-ranking gene was *hsp90ab* (predicted) (electronic supplementary material, file S6). Although it is a heat shock protein and does not have known effects on the pelvic fin, studies have shown that the inhibition of its expression causes defects in zebrafish, especially in eye development [40]. Furthermore, the disruption of *hsp90ab* expression has been associated with caudal fin fold defects in the zebrafish [40], which is not recorded in the ZFIN or the Monarch Initiative repository. Our computational results, together with a noted effect on a fin, indicate that *hsp90ab* is a prime new candidate gene for pelvic fin development that may have a key role in the module stability.

The top ranked hub gene in the hindlimb module based on weighted degree was *trp53* (electronic supplementary material, file S8), which has been associated with embryonic hindlimb development in mouse [41]. When *trp53* is disrupted, mouse limbs are deformed [42]. *Trp53* was also found in the pelvic fin module (predicted gene) but it had a lower rank (24th) based on the weighted degree (figure 4 and electronic supplementary material, table S16).

When comparing the conserved genes between the pelvic fin and the hindlimb modules (electronic supplementary material, table S16 and figure 4), several that are central to the modular stability were identified. For example, the *ctnnb1* gene, predicted and ranked 4th in the pelvic fin module, was also highly ranked (3rd) in the forelimb module. *Ctnnb1* is essential for the β-catenin pathway, which is necessary for the hindlimb initiation in the mouse [43]. Although it does not have known association to either of the paired fins in the zebrafish, it is known to be essential in fish development [44].

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 is given in figure 5. The conserved genes in both modules show higher normalized

394 weighted degree distributions compared to their respective module-specific genes. As observed
395 for the pectoral fin, this indicates the higher importance of the conserved genes for the stability
396 of the modules.
397
398 (iii) The fate of zebrafish paired fin module-specific genes in the mouse
399 A large number of zebrafish fin module genes (183 for pectoral fin and 536 for pelvic fin) were
400 not included in the mouse limb modules (electronic supplementary material, files S5 and S6),
401 implying these genes had not been maintained in limb development. To understand the roles of
402 those zebrafish pectoral and pelvic fin module-specific genes in the mouse, the enriched GO-BP
403 and Uberon terms for the mouse orthologs for these fin module-specific genes are given in
404 electronic supplementary material, tables S17, S18, S19 and S20. They were enriched for a
405 number of novel anatomical entities and related biological processes unique to tetrapods [13]
406 (electronic supplementary material, tables S21 and S22).
407

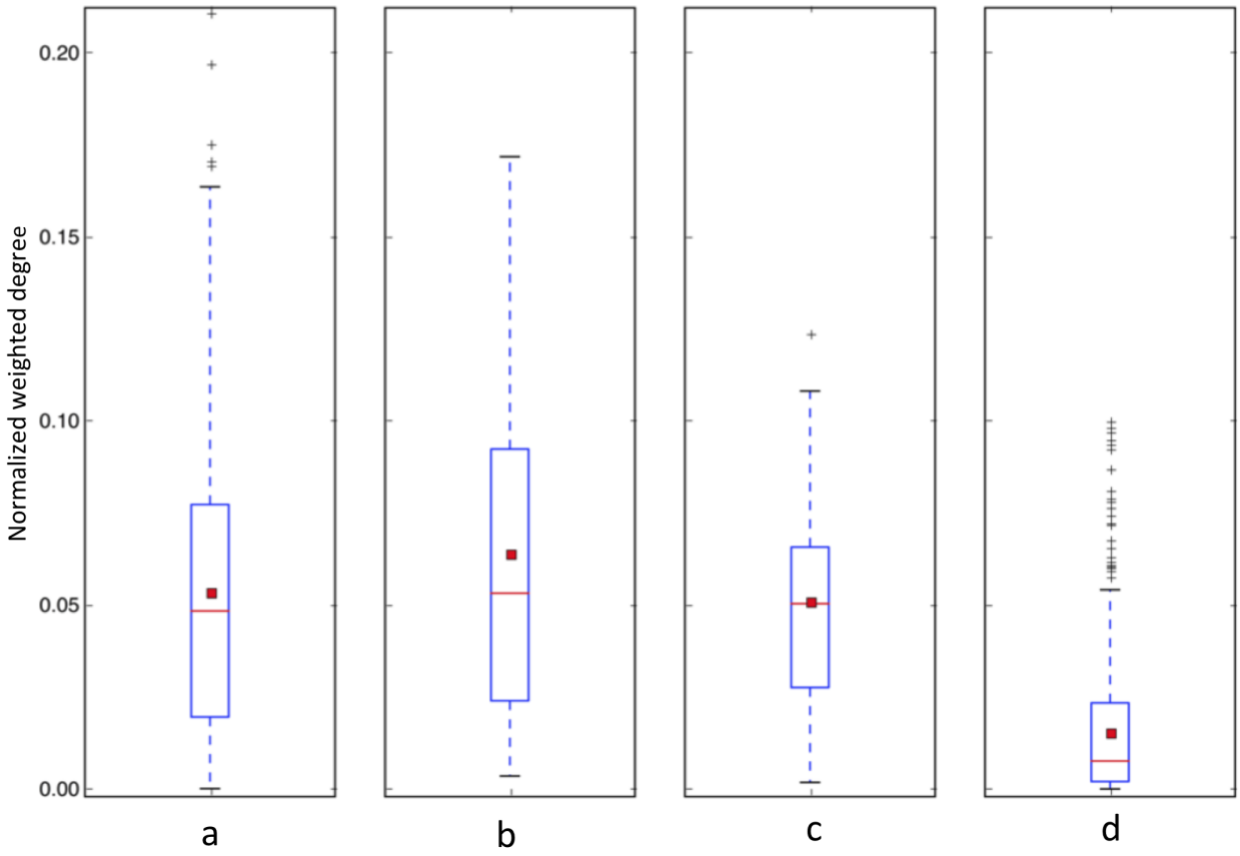


408

409 Figure 4. Networks of the 81 conserved genes common to and extracted from (a) the pelvic fin

410 module and (b) the hindlimb module. Node size is proportional to the degree (number of

411 interactions) of the gene. Hub genes, such as *bmp4*, *shh*, *ctnb1*, *bmp7, trp53*, and *hras*, are shown
412 in larger node sizes. The arrow represents the direction of modular evolution.



413

414 Figure 5. Boxplot comparison of normalized weighted degree distributions for (a) pelvic fin
415 module-specific genes, (b) pelvic fin conserved genes, (c) hindlimb conserved genes, and (d)
416 hindlimb module-specific genes. In the boxplots, the red line and the square represent the median
417 and mean, respectively.

418

419 For instance, the pectoral fin module-specific gene *lef1*, an important (ranked 7th)
420 member in the pectoral fin module, is involved with palate development, trachea gland
421 development, and associated with neck-related phenotypes [45, 46]. The neck evolved in
422 tetrapods and allowed them to support the head, which was crucial for their success in land [47,

48]. In the pelvic fin module-specific genes, *mapk1* (ranked 12th) is also associated with neck-related phenotypes, such as thymus development and trachea formation [49, 50]. It is additionally involved with the lung phenotypes and the development of the lung [50], another structure which progressively evolved in tetrapods that enabled them to breath and thrive in terrestrial environments [51]. *Lama5*, a gene found in both the pectoral fin and pelvic fin modules, is an example of another module-specific gene that is involved with lung development in the mouse [52]. Furthermore, it is also involved with hair follicle development and hair-related phenotypes [53], which are other anatomical entities specific for mammals [54]. These examples point to the possibility that many of genes used in fin development were recruited in the development of novel anatomical entities that enabled tetrapods to thrive in a terrestrial environment.

(iv) The role of mouse module-specific limb genes in the zebrafish

A large number of module-specific genes for the forelimb and hindlimb (207 for forelimb and 601 for hindlimb) did not appear in pectoral fin or pelvic fin modules (electronic supplementary material, files S7 and S8), and the question of their developmental function in the zebrafish occurred. To understand the function of the limb module-specific genes in zebrafish, the enriched GO-BP and Uberon terms for the limb module-specific genes are given in electronic supplementary material, tables S23, S24, S25 and S26. According to the enrichment analyses, these mouse limb module-specific genes were enriched to the head of the zebrafish, specifically, the jaw skeleton and post-hyoid pharyngeal arch skeleton (electronic supplementary material, tables S27 and S28). The latter region includes the gill chamber and contains parts such as gill rakers [55] that have been lost in tetrapods. For instance, *fst* is a crucial forelimb module-specific

gene, which has a zebrafish ortholog (*fsta*) with phenotypes related to splanchnocranium [56] and post-hyoid pharyngeal arch skeleton [57] that supports the gill chamber. Furthermore, *twist1* is module-specific for both forelimb and hindlimb, and it has two zebrafish orthologs (*twist1a* and *twist1b*) that are involved with pharyngeal system development [58].

There are some mouse module-specific genes, e.g., *tgfbr3*, which are involved in the development of both the forelimb and the hindlimb, that is associated with the development of the caudal fin in zebrafish [59]. Another example, *lep*, which is module-specific for both forelimb and hindlimb, is associated with otolith development in zebrafish [60]. Otoliths are located in the inner ear cavity of all teleost fishes where they aid in hearing and serve as balance organs [61]. The enrichment analyses point to the possibility that genes associated with various fish-specific structures such as gill arches and the caudal fin, were recruited for limb development as they were lost during the transition to tetrapods.

4. Conclusion

The goal of this work was to study the modular changes associated with the fin to limb transition using gene networks. This computational study expanded the number of genes that could be analyzed compared to wet lab methods and enabled the study of gene network structure rather than individual genes. Employing the quality-enhanced integrated networks ensured that the module detections, gene predictions, and identification of important genes in the modules were accurate, as evidenced from the results. To our knowledge, this is the first work that uses PPI networks to study the fin to limb transition. We discovered important information such as the hub genes responsible for the stability of paired fin and limb modules and changes in the importance of module genes associated with the transition. Some of the module genes were

469 predicted during the module detection, with evidence confirming their involvement with the
470 respective fins or limbs. This paper tabulates rankings of module genes based on their
471 importance, predicted candidates, and comparisons of the importance of the module genes,
472 which will be useful for future studies on fin to limb transition.

473 Furthermore, we discovered that the conserved genes were more likely to be hub genes
474 than the module-specific genes. Thus, it appeared that during the fin to limb transition, most of
475 the crucial hub genes of fin modules were conserved in the limb, and limb-specific genes were
476 recruited to surround this conserved 'appendage' core network. Moreover, our data suggested
477 that zebrafish fin module-specific genes were additionally employed in anatomical structures that
478 emerged after the aquatic to terrestrial vertebrate transition, such as lung and neck. Furthermore,
479 the evidence implied that mouse limb module-specific genes were involved with anatomical
480 structures, such as the gill rakers in the zebrafish that were lost during the transition. These
481 results provide the groundwork for evolutionary developmental biologists to experimentally
482 investigate aforementioned hypotheses. Most importantly, this work demonstrates how gene
483 networks can be used to study evolutionary phenotypic transitions and this computational
484 workflow can be used to perform large-scale network analyses to study evolutionary transitions
485 involving any model organism and anatomical entity with sufficient data, which is a valuable
486 addition to evolutionary biology.

487

488 **Data accessibility**

489 The network files and the anatomy profiles used for the candidate gene predictions are available
490 at https://doi.org/10.6084/m9.figshare.13589579.v1 and the Python scripts used for this analysis
491 are available at https://doi.org/10.5281/zenodo.4445583.

Authors' contributions

All authors planned and designed the experiments. PCF wrote the Python scripts for the analysis and performed the experiments under the supervision of PMM and EZ. All authors analyzed the results, read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests. The founding sponsors played no role in the design of this study; the collection, analyses or interpretation of the data; the writing of the manuscript; or the decision to publish the results.

Funding

This work was supported by the National Science Foundation (NSF) collaborative grant DBI-1062542, the Phenotype Research Coordination Network (NSF 0956049) and partially by NSF EPSCoR grant IIA–1355423. This work was also supported by the University of Iowa faculty startup fund to EZ. The High-Performance Computing clusters used in this work were supported by an NSF grant OAC-1626516. The views expressed in this paper do not necessarily reflect those of the NSF. Furthermore, PCF received support from the University of South Dakota (USD) Graduate Academic and Creative Research Grant and the Nelson Fellowship from Biology Department.

Acknowledgements

The authors thank J. P. Balhoff for assisting with the retrieval of gene-anatomical entity relationships from the Monarch Initiative repository. The authors thank T. J. Vision, D. Goodman, B. Wone, W. M. Dahdul, and L. M. Jackson for their assistance and helpful comments to improve this research.

References

[1] Cowen, L., Ideker, T., Raphael, B.J. & Sharan, R. 2017 Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*.
[2] Yamada, T. & Bork, P. 2009 Evolution of biomolecular networks [mdash] lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol* **10**, 791-803. (doi:http://www.nature.com/nrm/journal/v10/n11/suppinfo/nrm2787_S1.html).
[3] Defoort, J., Van de Peer, Y. & Vermeirssen, V. 2018 Function, dynamics and evolution of network motif modules in integrated gene regulatory networks of worm and plant. *Nucleic Acids Res.*
[4] Sharan, R., Ulitsky, I. & Shamir, R. 2007 Network-based prediction of protein function. *Molecular systems biology* **3**, 88.
[5] Tripathi, S., Moutari, S., Dehmer, M. & Emmert-Streib, F. 2016 Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC bioinformatics* **17**, 129. (doi:10.1186/s12859-016-0979-8).
[6] Tang, X., Wang, J., Liu, B., Li, M., Chen, G. & Pan, Y. 2011 A comparison of the functional modules identified from time course and static PPI network data. *BMC bioinformatics* **12**, 339.
[7] Vespignani, A. 2003 Evolution thinks modular. *Nat. Genet.* **35**. (doi:10.1038/ng1003-118).
[8] Wuchty, S., Oltvai, Z.N. & Barabási, A.L. 2003 Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.* **35**. (doi:10.1038/ng1242).
[9] Alhindi, T., Zhang, Z., Ruelens, P., Coenen, H., Degroote, H., Iraci, N. & Geuten, K. 2017 Protein interaction evolution from promiscuity to specificity with reduced flexibility in an increasingly complex network. *Scientific reports* **7**, 44948.
[10] Shui, Y. & Cho, Y.-R. 2016 Alignment of PPI Networks Using Semantic Similarity for Conserved Protein Complex Prediction. *IEEE transactions on nanobioscience* **15**, 380-389.
[11] García del Valle, E.P., Lagunes García, G., Prieto Santamaría, L., Zanin, M., Menasalvas Ruiz, E. & Rodríguez-González, A. 2019 Disease networks and their contribution to disease understanding: A review of their evolution, techniques and data sources. *J. Biomed. Inf.* **94**, 103206. (doi:https://doi.org/10.1016/j.jbi.2019.103206).
[12] Cheng, F., Jia, P., Wang, Q., Lin, C.-C., Li, W.-H. & Zhao, Z. 2014 Studying Tumorigenesis through Network Evolution and Somatic Mutational Perturbations in the Cancer Interactome. *Molecular Biology and Evolution* **31**, 2156-2169. (doi:10.1093/molbev/msu167).
[13] Shubin, N. 2008 *Your inner fish: a journey into the 3.5-billion-year history of the human body*, Vintage.
[14] Amaral, D.B. & Schneider, I. 2018 Fins into limbs: Recent insights from sarcopterygian fish. *genesis* **56**, e23052.

[15] Vorobyeva, E.I. & Schultze, H.-P. 1991 Description and systematics of panderichthyid fishes with comments on their relationship to tetrapods. *Origins of the higher groups of tetrapods: controversy and consensus*, 68-109.
[16] 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., et al. 2018 A conserved Shh cis-regulatory module highlights a common developmental origin of unpaired and paired fins. *Nat. Genet.* **50**, 504-509. (doi:10.1038/s41588-018-0080-5).
[17] Onimaru, K., Marcon, L., Musy, M., Tanaka, M. & Sharpe, J. 2016 The fin-to-limb transition as the re-organization of a Turing pattern. *Nature communications* **7**, 11582.
[18] Zeng, E., Ding, C., Mathee, K., Schneper, L. & Narasimhan, G. 2012 Gene Function Prediction and Functional Network: The Role of Gene Ontology. In *Data Mining: Foundations and Intelligent Paradigms* (eds. D.E. Holmes & L.C. Jain), pp. 123-162, Springer Berlin Heidelberg.
[19] Kernighan, B.W. & Lin, S. 1970 An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal* **49**, 291-307.
[20] Zeng, E., Yang, C., Li, T. & Narasimhan, G. Year On the effectiveness of constraints sets in clustering genes. IEEE.
[21] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. & Takagi, T. 2001 Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast* **18**, 523-531. (doi:10.1002/yea.706).
[22] Liu, X. & Murata, T. 2010 Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A: Statistical Mechanics and its Applications* **389**, 1493-1500.
[23] Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q. & Wrana, J.L. 2009 Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotech* **27**, 199-204. (doi:http://www.nature.com/nbt/journal/v27/n2/suppinfo/nbt.1522_S1.html).
[24] Tang, X., Wang, J., Zhong, J. & Pan, Y. 2014 Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **11**, 407-418.
[25] Fernando, P.C., Mabee, P.M. & Zeng, E. 2020 Integration of anatomy ontology data with protein–protein interaction networks improves the candidate gene prediction accuracy for anatomical entities. *BMC bioinformatics* **21**, 442. (doi:10.1186/s12859-020-03773-2).
[26] Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. 2017 The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362-D368. (doi:10.1093/nar/gkw937).
[27] Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E. & Haendel, M.A. 2012 Uberon, an integrative multi-species anatomy ontology. *Genome biology* **13**, R5. (doi:10.1186/gb-2012-13-1-r5).
[28] Liang, Z., Xu, M., Teng, M. & Niu, L. 2006 Comparison of protein interaction networks reveals species conservation and divergence. *BMC bioinformatics* **7**, 457.
[29] Mungall, C.J., McMurry, J.A., Köhler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. 2017 The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **45**, D712-D722. (doi:10.1093/nar/gkw1128).

601 [30] Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. 2011 Cytoscape 2.8: new
602 features for data integration and network visualization. *Bioinformatics* **27**, 431-432.
603 [31] Huang, L., Jin, J., Deighan, P., Kiner, E., McReynolds, L. & Lieberman, J. 2013 Efficient
604 and specific gene knockdown by small interfering RNAs produced in bacteria. *Nature*
605 *biotechnology* **31**, 350-356.
606 [32] Meyer, A. & Schartl, M. 1999 Gene and genome duplications in vertebrates: the one-to-four
607 (-to-eight in fish) rule and the evolution of novel gene functions. *Current opinion in cell biology*
608 **11**, 699-704.
609 [33] Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D.G., Knight, J., Mani,
610 P., Martin, R., Moxon, S.A.T., et al. 2011 ZFIN: enhancements and updates to the zebrafish
611 model organism database. *Nucleic Acids Res.* **39**, D822-D829. (doi:10.1093/nar/gkq1077).
612 [34] Routledge, R. 2005 Fisher's exact test. *Encyclopedia of biostatistics*.
613 [35] Grandel, H. & Schulte-Merker, S. 1998 The development of the paired fins in the Zebrafish
614 (Danio rerio). *Mechanisms of Development* **79**, 99-120. (doi:https://doi.org/10.1016/S0925-
615 4773(98)00176-2).
616 [36] Varoquaux, G. 2018 Cross-validation failure: small sample sizes lead to large error bars.
617 *Neuroimage* **180**, 68-77.
618 [37] Lopez-Rios, J. 2016 The many lives of SHH in limb development and evolution. *Seminars*
619 *in Cell & Developmental Biology* **49**, 116-124.
620 (doi:https://doi.org/10.1016/j.semcdb.2015.12.018).
621 [38] Bakrania, P., Efthymiou, M., Klein, J.C., Salt, A., Bunyan, D.J., Wyatt, A., Ponting, C.P.,
622 Martin, A., Williams, S., Lindley, V., et al. 2008 Mutations in BMP4 Cause Eye, Brain, and
623 Digit Developmental Anomalies: Overlap between the BMP4 and Hedgehog Signaling
624 Pathways. *The American Journal of Human Genetics* **82**, 304-319.
625 (doi:https://doi.org/10.1016/j.ajhg.2007.09.023).
626 [39] Raspopovic, J., Marcon, L., Russo, L. & Sharpe, J. 2014 Digit patterning is controlled by a
627 Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science* **345**, 566-570.
628 [40] Yeyati, P.L., Bancewicz, R.M., Maule, J. & van Heyningen, V. 2007 Hsp90 Selectively
629 Modulates Phenotype in Vertebrate Development. *PLOS Genetics* **3**, e43.
630 (doi:10.1371/journal.pgen.0030043).
631 [41] Im, C.H., Chau, K., Lopez, M., Sun, X.Z., Tran, J., Cuya, S. & Sharp, S.B. 2016 Hindlimb
632 Muscles of 17.5–18.5 dpc Mice Double Null for MyoD and Trp53 Appear Indistinguishable
633 from Muscles of Mice Null for Either Gene. *The FASEB Journal* **30**, 1035-1032.
634 [42] Vares, G., Wang, B., Tanaka, K., Shang, Y., Fujita, K., Hayata, I. & Nenoi, M. 2011 Trp53
635 activity is repressed in radio-adapted cultured murine limb bud cells. *Journal of radiation*
636 *research* **52**, 727-734.
637 [43] Kawakami, Y., Marti, M., Kawakami, H., Itou, J., Quach, T., Johnson, A., Sahara, S.,
638 O'Leary, D.D.M., Nakagawa, Y. & Lewandoski, M. 2011 Islet1-mediated activation of the β-
639 catenin pathway is necessary for hindlimb initiation in mice. *Development* **138**, 4465-4473.
640 [44] Zhang, M., Zhang, J., Lin, S.-C. & Meng, A. 2012 β-Catenin 1 and β-catenin 2 play similar
641 and distinct roles in left-right asymmetric development of zebrafish embryos. *Development*, dev-
642 074435.
643 [45] Duan, D., Yue, Y., Zhou, W., Labed, B., Ritchie, T.C., Grosschedl, R. & Engelhardt, J.F.
644 1999 Submucosal gland development in the airway is controlled by lymphoid enhancer binding
645 factor 1 (LEF1). *Development* **126**, 4441-4453.

[46] Nawshad, A. & Hay, E.D. 2003 TGFβ3 signaling activates transcription of the LEF1 gene to induce epithelial mesenchymal transformation during mouse palate development. *The Journal of cell biology* **163**, 1291-1301.

[47] Sallan, L.C. 2012 Tetrapod-like axial regionalization in an early ray-finned fish. *Proceedings of the Royal Society B: Biological Sciences* **279**, 3264-3271. (doi:10.1098/rspb.2012.0784).

[48] Ericsson, R., Knight, R. & Johanson, Z. 2013 Evolution and development of the vertebrate neck. *J. Anat.* **222**, 67-78. (doi:10.1111/j.1469-7580.2012.01530.x).

[49] Hoffman, B.G., Williams, K.L., Tien, A.H., Lu, V., de Algara, T.R., Ting, J.-y. & Helgason, C.D. 2006 Identification of novel genes and transcription factors involved in spleen, thymus and immunological development and function. *Genes and immunity* **7**, 101.

[50] Boucherat, O., Nadeau, V., Bérubé-Simard, F.-A., Charron, J. & Jeannotte, L. 2014 Crucial requirement of ERK/MAPK signaling in respiratory tract development. *Development* **141**, 3197.

[51] Tatsumi, N., Kobayashi, R., Yano, T., Noda, M., Fujimura, K., Okada, N. & Okabe, M. 2016 Molecular developmental mechanism in polypterid fish provides insight into the origin of vertebrate lungs. *Scientific Reports* **6**, 30580. (doi:10.1038/srep30580).

[52] Nguyen, N.M., Miner, J.H., Pierce, R.A. & Senior, R.M. 2002 Laminin α5 is required for lobar septation and visceral pleural basement membrane formation in the developing mouse lung. *Developmental biology* **246**, 231-244.

[53] Gao, J., DeRouen, M.C., Chen, C.-H., Nguyen, M., Nguyen, N.T., Ido, H., Harada, K., Sekiguchi, K., Morgan, B.A. & Miner, J.H. 2008 Laminin-511 is an epithelial message promoting dermal papilla development and function during early hair morphogenesis. *Genes & development* **22**, 2111-2124.

[54] Wu, D.-D., Irwin, D.M. & Zhang, Y.-P. 2008 Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. *BMC Evolutionary Biology* **8**, 241. (doi:10.1186/1471-2148-8-241).

[55] Gillis, J.A., Modrell, M.S. & Baker, C.V.H. 2013 Developmental evidence for serial homology of the vertebrate jaw and gill arch skeleton. *Nature communications* **4**, 1436.

[56] Dalcq, J., Pasque, V., Ghaye, A., Larbuisson, A., Motte, P., Martial, J.A. & Muller, M. 2012 RUNX3, EGR1 and SOX9B Form a Regulatory Cascade Required to Modulate BMP-Signaling during Cranial Cartilage Development in Zebrafish. *PLOS ONE* **7**, e50140. (doi:10.1371/journal.pone.0050140).

[57] Dal-Pra, S., Fürthauer, M., Van-Celst, J., Thisse, B. & Thisse, C. 2006 Noggin1 and Follistatin-like2 function redundantly to Chordin to antagonize BMP activity. *Developmental Biology* **298**, 514-526. (doi:https://doi.org/10.1016/j.ydbio.2006.07.002).

[58] Das, A. & Crump, J.G. 2012 Bmps and Id2a Act Upstream of Twist1 To Restrict Ectomesenchyme Potential of the Cranial Neural Crest. *PLOS Genetics* **8**, e1002710. (doi:10.1371/journal.pgen.1002710).

[59] Kamaid, A., Molina-Villa, T., Mendoza, V., Pujades, C., Maldonado, E., Ispizua Belmonte, J.C. & López-Casillas, F. 2015 Betaglycan knock-down causes embryonic angiogenesis defects in zebrafish. *genesis* **53**, 583-603. (doi:10.1002/dvg.22876).

[60] Liu, Q., Dalman, M., Chen, Y., Akhter, M., Brahmandam, S., Patel, Y., Lowe, J., Thakkar, M., Gregory, A.-V., Phelps, D., et al. 2012 Knockdown of leptin A expression dramatically alters zebrafish development. *General and Comparative Endocrinology* **178**, 562-572. (doi:https://doi.org/10.1016/j.ygcen.2012.07.011).

691 [61] Rodríguez Mendoza, R.P. 2006 Otoliths and their applications in fishery science. *Croatian*
692 *Journal of Fisheries: Ribarstvo* **64**, 89-102.
693