Harnessing Single Cell RNA Sequencing to Uncover Developmental Genetic Underpinnings of Syngnathid Fishes' Derived Traits

by

Hope Meredith Healey

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Biology

Dissertation Committee:

Nadia Singh, Chair

William Cresko, Advisor

Matthew Barber, Core Member

John Postlethwait, Core Member

Kirstin Sterner, Institutional Representative

University of Oregon

Fall 2024

© 2024 Hope Meredith Healey
This work is openly licensed via CC BY.



DISSERTATION ABSTRACT

Hope Meredith Healey

Doctor of Philosophy in Biology

Fall 2024

Title: Harnessing Single Cell RNA Sequencing to Uncover Developmental Genetic Underpinnings of Syngnathid Fishes' Derived Traits

Syngnathid fishes are extraordinary creatures with highly derived features. Notably, these fishes have elongated, toothless snouts, male pregnancy, and exoskeletons. In addition to their highly altered morphologies, syngnathids have lost highly conserved signaling genes (*fgf3* and *4*), transcription factors (*eve1*), and tooth mineralization genes (*scpp* genes). Although these genomic changes are predicted to contribute to the development, or lack thereof, of syngnathid's derived traits, it is unknown whether this is the case. To investigate developmental origins of syngnathid's highly altered traits, I harnessed single cell RNA sequencing (scRNAseq) methods for my dissertation research. First, I identified limitations for completing scRNAseq analysis in non-traditional model organisms, then overcame these challenges using scISOseq to improve gene annotations. Then, I used these methods to study the unique craniofacial morphology (elongate head and loss of teeth) of syngnathids. I identified the cell types present and signaling genes expressed at the start of craniofacial elongation in Gulf pipefish. Additionally, I found unique changes in signaling gene expression patterns during early craniofacial development of Gulf pipefish. This thesis discovered local modifications (select Fgf, Wnt, and Bmp genes) in pipefish craniofacial development, but found no evidence for global gene network re-wiring. Overall, this study suggests that syngnathid's highly derived heads evolved from local network re-wiring.

This dissertation includes previously published and unpublished co-authored material.

ACKNOWLEDGMENTS

I am extraordinarily lucky to have worked with and been supported by amazing people during my dissertation. My advisor, Dr. Bill Cresko, has been instrumental for my scientific growth and achievement; he has always supported my scientific goals and pushed me to grow as a researcher. My advancement would not have been possible without my role models, Dr. Susie Bassham and Dr. Emily Beck, who showed me how to be a strong, successful woman in science, provided scientific guidance, and extended opportunities to me. I am also lucky to have worked with Dr. Clay Small, who has always provided key scientific input. Additionally, I thank Mark Currey and Tiffany Thornton (and many undergraduate assistants) for taking excellent care of all the fish used in this dissertation. I was fortunate to ford through graduate school with other Cresko Lab graduate students: Sophia Frantz, Shannon Snyder, Carla Campos, and Amberly Buer. Additionally, I am grateful for all the undergraduates who supported my work: Starla Chambrose, Thomas Skates, Micah Woods, Vithika Goyal, and Hayden Penn. Lastly, I want to thank rotation students who have inspired and assisted with the work from these chapters: LyAndra Lujan, Clara Rehmann, and Aubrey Mayer.

I am grateful for the support and guidance of my committee: Dr. Kirstin Sterner, Dr. Nadia Singh, Dr. Matt Barber, and Dr. John Postlethwait. Additionally, I am lucky to have had a second home in the Bioinformatics and Genomics Master's Program, and thank Dr. Stacey Wagner, Dr. Leslie Coonrod, Pete Batzel, and Jason Sydes. I am also grateful for the BGMP cohort from 2019-2020, especially Wyatt Eng, who grew with me as bioinformaticians. I also want to thank our wonderful collaborators from University of Idaho, Dr. Adam Jones and Dr. Balan Ramesh, for truly making this work possible. I am grateful for additional collaborators and scientific mentors including Dr. Thomas Desvignes, Dr. Angel Amores, Dr. Martin Stervander, Dr. Ingo Braasch, Dr. Andrew Thompson, and Dr. Emily Rose.

I would further like to thank my community here in the Institute of Ecology and Evolution and the Department of Biology. I am grateful for support from the Genetics Training Program, the GrEBES student group, and zebrafish groupie. But most importantly, I am grateful for my IE2 cohort: Victoria Caudill, Gabrielle Coffing, Aidan Short, and Caitlin Smith. I am also thankful for crucial administrative support from Sara Nash, Arlene Deyo, and Leah Frazier.

Finally, I am thankful for Women in Graduate Sciences, which brought together a supportive community ready to make positive impacts on our local community.

I am grateful for the funding that made this research possible. Specifically, funding from the Genetics Training Program (NIH T32GM007413), F31 funding from NIDCR (5F31DE032559-02), University of Oregon Research Excellence funds (Bill Cresko), and a National Science Foundation Grant (OPP-2015301; Bill Cresko, Susie Bassham, and Clay Small).

Finally, I want to thank all my loved ones for supporting and making this dream possible.

DEDICATION

For my grandfather, who gave me a love for learning

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 14

Introduction 17

References 17

II. SINGLE-CELL ISO-SEQUENCING ENABLES RAPID GENOME ANNOTATION FOR SCRNASEQ ANALYSIS 18

Introduction 18

Materials and Methods 19

Tissue Dissociation to Generate a Pool of Single Cells 19

Library Preparation 20

PacBio Data Processing 20

Annotation Analysis 22

ScRNAseq Analysis 22

Results and Discussion 23

ScISOr-Seq Captured Novel Isoforms and Improved 3' UTR 23

ScISOr-Seq Improved the Number of Reads Retained for ScRNAseq 25

ScISOr-Seq Enhances Biological Interpretation of Cell Clusters 28

Conclusion 30

References 32

Bridge 36

Chapter Page

III. SINGLE CELL RNA SEQUENCING PROVIDES CLUES FOR THE DEVELOPMENTAL GENETIC BASIS OF SYNGNATHID FISHES EVOLUTIONARY ADAPTATIONS 37

Introduction 37

Materials and Methods 40

Single Cell RNA Sequencing Libraries Preparation 40

Single Cell Atlas Construction 41

Single Cell Atlas Cluster Identification 42

Differentiation State Analysis 44

Single Cell KEGG Analysis 44

Single Cell Atlas Network Analysis 44

In situ Hybridization 45

Bone and Cartilage Staining 46

Gene Cluster Analysis 46

Results 47

Valuable cRNAseq Atlas for Studying Syngnathid Development 47

Discovery of Cell Cluster Function and State Using KEGG Analysis 49

Commonalities of Cell Clusters, Unique Networks, and Elusive Cell Types Identified in Network Analysis 51

Conserved Signaling Pathways Are Active during Syngnathid Craniofacial Development 53

Gulf Pipefish Retain Tooth Development Genes but Likely Lack Onset of Tooth Development 55
Tooth and Skeletal Genes Are Expressed during Dermal Armor Development 58
Epithelial Expression of Immune and Nutrient Processing Genes May Facilitate Embryo-Paternal Interactions in the Brood Pouch 59
Discussion 61
Our atlas Represents a Novel Resource for Evo-Devo Research 62
Conserved Pathways May Contribute to Derived Syngnathid Heads 62
No Evidence of Primordia Suggests Changes in Early, Not Late, Tooth Development Are Likely at the Root of Evolutionary Tooth Loss 64
Redeployment of the Bone Gene Network to Build Dermal Armor 65
Signatures of Embryonic Interactions within the Novel Pouch Environment 65
Conclusion 67
References 68
Bridge 81
IV. REDUNDACY AND PATHWAY EVOLUTION IN SYNGNATHID FISH CRANIOFACIAL GENE NETWORKS 82
Introduction 82
Materials and Methods 84
Production of ScRNAseq Sequencing Libraries 84
Production of ATACseq Sequencing Libraries 85
ScRNAseq Library Initial Processing 85
Ortholog Identification 85

ScRNAseq Atlas Construction 86
ScRNAseq Trajectory Analysis 87
ScRNAseq Differentially Expressed Gene Analysis 87
ScRNAseq Gene Network Analysis 87
ScRNAseq Cross-Species Correlation Analysis 87
ATACseq Read Processing and Analysis 88
Whole Genome Alignment and Identification of Conserved Elements 88
In situ Hybridization 89
Results 89
Successful Integration of Gulf Pipefish, Stickleback, and Zebrafish Single Cell Dataset… 89
Cranial Neural Crest Pharyngeal Arch Core Gene Regulatory Network is Largely Conserved in Gulf Pipefish 92
Gene Losses and Potential Novel Genes in Gulf Pipefish Pharyngeal Arch Cells Alter Gene Regulatory Networks 94
GENESPACE Analysis Identifies Candidates for Broader Syngnathid Evolution… 95
Numerous Candidate Genes Identified from Global Analyses, but few are Promising for Craniofacial Evolution 96
Gulf Pipefish has Novel Fgf Expression in Pharyngeal Arch Cells 96
Potential Changes in Gene Sequence and Regulation may Lead to Gulf Pipefish Expression Differences 98
Discussion 100

Extensive Teleost Genome Duplication Paralog Evolution but no Evidence for Global Re-wiring in Pipefish CNC 100

Local Gene Regulatory Network Re-wiring through Co-option of *Fgf23* in Pipefish… 101

Evolution of Redundancy in Fgf pathway may Enable Genome Evolution ... 102

References 104

V. CONCLUSION 112

APPENDICES 115

A. SUPPORTING INFORMATION FOR CHAPTER II 115

B. SUPPORTING INFORMATION FOR CHAPTER III 117

C. SUPPORTING INFORMATION FOR CHAPTER IV 198

LIST OF FIGURES

Figure Page

2.1 SQANTI3 Classification of ScISOr-Seq Data Revealed that the Majority of Isoforms were Previously Unannotated in the Stickleback Genome 24
2.2 ScISOr-Seq and Bulk Iso-Seq Captured a Limited Number of Annotated Transcripts, but Both Resulted in an Increase in Reads Mapping to the Transcriptome..... 25
2.3 ScISOr-Seq Improved Gene Models by Extending the 3'UTRs Illustrated Here by the Two Models for Nkx2.3..... 26
2.4 Annotation improvements using Iso-Seq (bulk Iso-Seq or ScISOr-Seq) data extended genes both 5' and 3', with 3' expansions leading to greater numbers of reads counted. 27
2.5 Using the Updated Gene Models from Our Merged and Pooled Dataset, We Identified Expected Cell Types in 70hpf Stickleback ScRNAseq Data and Illustrated How Annotations Improvements Are Relevant for Analysis 29
3.1 Gulf Pipefish Exemplify Syngnathid Derived Traits 40
3.2 Gulf Pipefish Single Cell Atlas Contains Cells from the Entire Embryo and Identifies Genetic Pathways Active in Different Cell Types 48
3.3 Weighted Gene Network Analysis (WGCNA) Identifies Gene Modules that Define and Unite Cell Clusters 50
3.4 Conserved Cell Types and Gene Pathways Build Unique Faces of Syngnathids. 52
3.5 Pipefish Do Not Possess Identifiable Tooth Primordium Cells, but Continue to Express Tooth Development Genes in Other Contexts 57
3.6 Tooth and Bone Development Genes Expressed during Exoskeleton Development 59
3.7 Gene Expression Signatures Suggest Embryonic Interactions within Brood Pouch Environment 60
4.1 Two Hypotheses for Syngnathid Craniofacial Development and Evolution 83

4.2 Integrated Atlas Encompasses Cells from the Entire Embryo 91

4.3 Number of Cells per Cluster Varied by Species.....92

4.4 Pharyngeal Arch Cells Express Highly Conserved Core Genetic Program 94

4.5 Novel Fgf and Wnt Pathway Expression Changes in Gulf Pipefish CNC Pharyngeal Arch Cells. 98

4.6 Potential Changes in Gene Sequence and Regulation may Lead to Gulf Pipefish Expression Differences 99

CHAPTER I
INTRODUCTION

The beauty of life is one of evolution's greatest mysteries. Solving life's riddles may be impossible, but uncovering the underlying mechanisms that create the incredible is worth it. In this thesis, I attempted to understand the extraordinary: the development and evolution of highly derived traits.

Highly derived traits appear across the diversity of life and often allow species to adapt to new niches. Although the ecological and evolutionary consequences of these traits is occasionally clear, their developmental genetic underpinnings are often challenging to discern. In practice, finding these needles in the haystack requires comparisons to close outgroups, hypothesis driven sampling, and different modalities of data. New genome assemblies have recently enabled these types of analyses in the charismatic clade of fishes known as syngnathids. These popular fishes include seahorses, pipefishes, and seadragons. They are known for their derived traits including male pregnancy, dermal armor, elongated snouts, and jaws without teeth. However, the developmental basis of these traits is unknown.

Syngnathids' genomes have been extensively studied to identify candidates for their trait evolution, and numerous candidates have been uncovered through these approaches. Notably, syngnathids have lost *fgf3* and *fgf4*, two deeply conserved signaling genes involved in craniofacial development (Small et al. 2022). The loss of *fgf3* is particularly intriguing given that the absence of *fgf3* in zebrafish is lethal (Herzog et al. 2004). Additionally, syngnathids have lost genes related to tooth development including *eve1* and most *scpp* genes (Small et al. 2016; Lin et al. 2016; Zhang et al. 2020). Beyond these protein coding losses, conserved non-coding elements (including in *dlx1a* and near *dlx2a*) and miRNAs (including *mir130* clusters and *mir196b*) are also missing in syngnathids (Small et al. 2015; Small et al. 2022). Connecting these genomic losses with syngnathid craniofacial evolution requires transcriptomic data to uncover whether these losses led to shifts in developmental networks.

To probe developmental transcription patterns in syngnathid fishes, I used a sequencing method called single cell RNA sequencing (scRNAseq). This technique determines the RNA expression profile of individual cells from a sample. Using these data, cell identities can be fleshed out post-hoc. Therefore, using this approach in non-traditional systems, including

syngnathids, is powerful because expression patterns of cell types can be deduced without dissection or genetic modifications. These datasets can be used to determine the cell types present at different stages of development, expression patterns of signaling pathway genes, and detect genetic networks.

In Chapter II, I identified a limitation of scRNAseq in non-traditional organisms and solved this problem using single cell ISO-sequencing (scISOseq). Although numerous developmental scRNAseq atlases exist for model systems, there are few atlases in non-model systems. Due to the 3' bias of scRNAseq reads, many sequencing reads land in the 3' UTR of gene models which are often poorly annotated in non-model organisms. I addressed this limitation using scISOseq, a sequencing method that used leftover cells from the scRNAseq library to produce full length RNA reads. Using this method, I lengthened 3' UTR annotations in the stickleback genome assembly leading to a 26.1% improvement in the number of scRNAseq reads counted. These findings suggest that scRNAseq is feasible in non-model systems but requires careful planning to ensure proper read quantification. This chapter has been published and is co-authored with Dr. Susan Bassham and Dr. William Cresko (Healey et al. 2022).

In Chapter III, I explored the developmental genetic basis of numerous syngnathid derived traits using scRNAseq. Although genomic candidates have been identified for their derived traits, there are few studies that examine how these changes relate to development. In this study, I developed a scRNAseq atlas that probed the elongate snout, loss of teeth, dermal armor, and male pregnancy (from the embryonic lens). Using Gulf pipefish embryos from early craniofacial elongation, I identified candidate genes (*sfrp1a, bmp4*, and *prdm16*) expressed in mesenchyme above the lengthening ethmoid plate in the face. Next, I searched for tooth primordial cells to determine whether tooth development initiates in syngnathids. Although I did not find these tooth primordial cells, expression of select tooth genes (*dlx3b* and *scpp1*) was observed in the developing dermal armor. Finally, I identified an enrichment of immune and metabolism genes in the embryo epidermis, potentially indicating nutrient uptake from the brood pouch. These findings suggest that syngnathid highly derived traits emerge from re-deployment of signaling pathways. This chapter has been published and is co-authored with Hayden Penn, Dr. Clay Small, Dr. Susan Bassham, Micah Woods, Vithika Goyal, and Dr. William Cresko (Healey et al. 2024).

In Chapter IV, I asked whether syngnathids have changes in the pharyngeal arch cells during early craniofacial development that could relate to their unique morphology. In laboratory and evolutionary models, changes in gene expression or numbers of cranial neural crest cells that build the pharyngeal arches can lead to changes in head shapes. Since syngnathids have an evolutionary knockout of *fgf3*, a gene essential during pharyngeal arch development, I probed the extent of genetic network changes in syngnathid pharyngeal arch cells. Using comparative scRNAseq, I identified that Gulf pipefish have largely conserved pharyngeal arch gene networks. Furthermore, I discovered that *fgf22* from the *fgf3* sub-family had evolved pharyngeal arch expression prior to the loss of *fgf3* in syngnathids. These findings suggest that redundancy permitted the loss of *fgf3* in syngnathids and, furthermore, there has not been extensive craniofacial gene regulatory network re-wiring in these fishes. This chapter is co-authored with Clara Rehmann, Dr. Susan Bassham, and Dr. William Cresko.

Together, these chapters form the foundation for successful developmental scRNAseq atlases in non-traditional systems. Chapters III and IV contribute significantly to the syngnathid literature by providing the first transcriptomic insights into their unique development.

References

Healey, Bassham, Cresko. 2022. Single-cell Iso-Sequencing enables rapid genome annotation for scRNAseq analysis. *Genetics* 220 (3): iyac017.

Healey, Penn, Small, Bassham, Goyal, Woods, Cresko. 2024. Single Cell RNA Sequencing Provides Clues for the Developmental Genetic Basis of Synganthidae's Evolutionary Adaptations. *eLIFE.*

Herzog, W., Sonntag, C., von der Hardt, S., Roehl, H. H., Varga, Z. M., & Hammerschmidt, M. (2004). Fgf3 signaling from the ventral diencephalon is required for early specification and subsequent survival of the zebrafish adenohypophysis. *Development*, *131*(15), 3681–3692.

Lin, Q., Fan, S., Zhang, Y., Xu, M., Zhang, H., Yang, Y., Lee, A. P., Woltering, J. M., Ravi, V., Gunter, H. M., Luo, W., Gao, Z., Lim, Z. W., Qin, G., Schneider, R. F., Wang, X., Xiong, P., Li, G., Wang, K., … Venkatesh, B. (2016a). The seahorse genome and the evolution of its specialized morphology. *Nature*, *540*(7633), 395–399.

Small, C. M., Bassham, S., Catchen, J., Amores, A., Fuiten, A. M., Brown, R. S., Jones, A. G., & Cresko, W. A. (2016). The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome Biology*, *17*(1), 258. https://doi.org/10.1186/s13059-016-1126-6

Small, Clayton M., Healey, H. M., Currey, M. C., Beck, E. A., Catchen, J., Lin, A. S. P., Cresko, W. A., & Bassham, S. (2022). Leafy and weedy seadragon genomes connect genic and repetitive DNA features to the extravagant biology of syngnathid fishes. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(26), e2119602119.

Zhang, Y., Ravi, V., Qin, G., Dai, H., Zhang, H., Han, F., Wang, X., & Liu, Y. (2020a). Comparative genomics reveal shared genomic changes in syngnathid fishes and signatures of genetic convergence with placental mammals ! Phylogenomics and evolutionary rate Transposable elements in syngnathids Expansion and contraction of gene families in t. *National Science Review*, *7*(6), 964–977.

CHAPTER II
SINGLE-CELL ISO-SEQUENCING ENABLES RAPID GENOME ANNOTATION FOR SCRNASEQ ANALYSIS

From Healey, Bassham, Cresko. 2022. Single-cell Iso-Sequencing enables rapid genome annotation for scRNAseq analysis. *Genetics* 220 (3): iyac017.

This chapter was published in Genetics. Susan Bassham and William Cresko are co-authors on this publication. The project originated from conversations between William Cresko and me. I prepared the sample for sequencing, performed the scRNAseq and scISOr-Seq analysis, and took the lead on writing the manuscript. Susan Bassham contributed to the analysis and made writing contributions. William Cresko oversaw the analysis, contributed to writing, and acquired funding.

The full supplemental material for this publication can be found at https://doi.org/10.25386/genetics.16811614.v1

Introduction

Single cell RNA sequencing (scRNAseq) is a revolutionary technique in biology that provides expression information from tissues and embryos (Shapiro et al. 2013). By barcoding RNA from individual cells directly from dissociated samples, scRNAseq allows for post hoc analysis of cell types and can be used to ascertain novel cell populations, explore developmental trajectories, and define gene regulatory networks (Luecken and Theis 2019).

To maximize the utility of scRNAseq datasets, however, 3' UTRs must be annotated for several reasons. ScRNAseq captures transcripts through poly(A) tails leading to a 3' bias in coverage (Hwang et al. 2018). Partially annotated genes may be represented in a dataset with lower read counts, leading to erroneous conclusions regarding their magnitude of expression across cell types. In addition, downstream scRNAseq analysis clusters cell types through the determination of a covariance structure across highly variable genes (Stuart et al. 2019). The systematic absence of such genes can lead to inferential errors in multivariate analyses and obscure biological reality.

The annotations of even intensively studied models, such as mice and zebrafish, continue to be improved (Gupta et al. 2018; Lawson et al. 2020). Most other organismal genome assemblies are even less well annotated. Facilitating a broader utility of scRNAseq requires more efficient methods for 3' UTR annotation. Recently, full length, single-molecule isoform sequencing (Iso-Seq) has been used to improve genome annotations (Kuo et al. 2017; Beiki et al. 2019; Kuo et al. 2020; Ali et al. 2021). PacBio's Iso-seq has been further adapted to use the 10x Genomics platform for scRNA barcoding (ScISOr-Seq) to track cell type specific isoform expression (Gupta et al. 2018; Zheng et al. 2020).

Here we show that ScISOr-Seq in the context of a scRNAseq experiment allows rapid 3'UTR annotation in threespine stickleback fish (*Gasterosteus aculeatus*). This fish has long been a focus of study in behavior, ecology and evolution (Bell and Foster 1994; Colosimo et al. 2004; Cresko et al. 2004a; Shapiro et al. 2004; Cresko et al. 2007; Hohenlohe et al. 2010; Reid et al. 2021), and is now a nascent system for biomedical research (Miller et al. 2007; Gardell et al. 2017; Small et al. 2017; Beck et al. 2020; Beck et al. 2021; Fuess et al. 2021). Although stickleback have a well assembled genome, its 3' UTR annotations are incomplete which limits scRNAseq's utility. We demonstrate that a single PacBio SMRT cell of ScISOr-Seq data is sufficient to significantly improve the stickleback annotations to an extent on par with zebrafish for the purpose of scRNAseq analysis at this stage. Our findings demonstrate that ScISOr-Seq will be a useful tool to efficiently improve genome annotations for scRNAseq in many organisms.

Materials and Methods

Tissue Dissociation to Generate a Pool of Single Cells

We crossed a laboratory line of stickleback originally isolated from Cushman Slough (Oregon) and raised embryos to 70 hours post fertilization (hpf) at 20°C using standard procedures from the Cresko Laboratory Stickleback Facility (Cresko et al. 2004b). 70hpf stickleback embryos are developmentally at an equivalent stage to 24 hpf zebrafish (Cresko et al. 2003). We euthanized 36 embryos in MS-222 following IACUC approved procedures then dechorionated and deyolked them at room temperature. We limited our embryo dissection to 20 minutes based on (Farrell et al. 2018). Following protocols from (Bresciani et al. 2018), we dissociated the cells for 6 minutes in 0.25% trypsin in PBS at 30°C, pipetting up and down every

30 seconds, then, stopped the dissociation using 10% FBS DMEM and spun down cells at 400xG for 3min at 4°C. We resuspended cells in 1mL of 0.1% BSA PBS, centrifuged at 400xG for 3min at 4°C, and resuspended in 100ul 0.04% BSA in PBS. At room temperature, we filtered cells through a 40uM cell strainer (Thomas scientific 1181X52) then washed the original tube twice with 100ul of 0.04% BSA in PBS and poured over the same cell strainer.

Library Preparation

The scRNAseq and ScISOr-Seq libraries were prepared and sequenced by the University of Oregon Genomics and Cell Characterization Core Facility (https://gc3f.uoregon.edu). The dissociated cells were diluted to 800 cells/ul using 0.04% BSA in PBS to target 10,000 cells for the 10X Genomics Single Cell 3' Genome Expression (GEX) mRNA-Seq prep with v3.1 NextGem chemistry. The single 10X preparation was split to create the scRNAseq and ScISOr-Seq libraries. The scRNAseq sample was sequenced on $1/7^{th}$ of a single S4 lane on a NovaSeq 6000 platform (Illumina). For the ScISOr-Seq library, 400ng of the amplified 10X cDNA was used as input for the SMRTbell Express Template Prep Kit 2.0 (P/N 100-938-900) without reamplification. Sample specific barcode (ATATAGCGCGCGTGTG) was added using the Barcoded Adapter Kit 8B—OVERHANG (P/N 101-628-500). The ScISOr-Seq library was sequenced on a single SMRT Cell 8M on the PacBio Sequel II platform using the v4 primer, v2.1 polymerase, 1 hour binding, 30-hour movie, and 2-hour pre-extension time at a loading concentration of 100pM.

PacBio Data Processing

The University of Oregon Genomics and Cell Characterization Core Facility (https://gc3f.uoregon.edu) generated “circular consensus” reads for our ScISOr-Seq dataset (ccs -j 39 --min-passes 3 --min-snr 2.5 --min-length 10 --max-length 50000 --min-rq 0.99; v6.6.0) and used lima (-j 39 –isoseq; v2.2.0) to remove a sample specific barcode (ATATAGCGCGCGTGTG) as a part of the PacBio SMRT Analysis software (Supplemental file 6). After barcode clipping, we used a custom ScISOr-Seq processing script (scISOr_Seq_processing.py) that removed the sample primers (3p: CTACACGACGCTCTTCCGATCT; 5p: CCCATGTACTCTGCGTTGATACCACTGCTT), removed and saved cell and UMI barcodes, removed poly(A) tails, then filtered out duplicated reads. This script outputs reads that had the expected presence and orientation of primers and

polyA tails for downstream analysis. The script additionally outputs reads without the expected primers in a separate file; however, these other reads were not used for further analysis.

We aligned reads to the stickleback genome (BROAD S1, 104.1 database version) using minimap2 (v2.7; parameters: -ax splice, -uf, -06,24, -B4; (Li 2018). We clustered for unique transcripts with collapse_isoforms_by_sam.py script (c=.99, i=.95; (Tseng 2021) from cDNA Cupcake (v27.0.0), classified transcripts with the sqanti3_qc.py script from SQANTI3 (v4.2), and refined transcripts with sqanti3_RulesFilter.py script (Tardaguila et al. 2018). We used the sqanti_classification.filtered_lite.gtf that resulted from the refining step (Supplemental file 2) to run Cell Ranger from 10x Genomics (v3.2.0) as described below.

We downloaded the raw data from (Naftaly et al. 2021) from NCBI. This adult Iso-Seq dataset contains gonad, pronephros, brain, and liver reads from both sexes, a total of 16 SMRT cells. We processed reads from this dataset with the following steps: we generated circular consensus reads with ccs (--min-rq=.9; v6.0.0), clipped barcodes listed in Naftaly, Pau, and White (2021) with lima (--isoseq –dump-clips; v2.2.0), removed polyA tails with isoseq3 refine (--require-polyA; v3.4.0-0), and clustered reads with isoseq3 cluster (v3.4.0-0) from the PacBio SMRT Analysis software (Supplemental file 6). For the analysis with solely these data, we aligned reads to the stickleback genome (BROAD S1, 104.1 database version), collapsed transcripts with cDNA Cupcake (v28.0.0), and filtered and classified with SQANTI3 (v4.2) as done with the ScISOr-Seq data. Again, we used the sqanti_classification.filtered_lite.gtf from SQANTI3 (Supplemental file 3) to run Cell Ranger as described below. Due to reduced novel and increased annotated genes relative to (Naftaly et al. 2021), we additionally tried using minimap2-2.15 with -ax splice -uf -C5 -secondary=no to match their parameters; however, we observed negligible changes. We propose that differences in our analyses arise from different earlier processing steps, different versions of stickleback gene models, and an updated version of SQANTI.

We pooled both sets of Iso-Seq reads and merged it with existing Ensembl annotations to create an improved version of the Ensembl annotations. Previously, we created a modified version of the Ensembl annotations where we extended the 3’ UTRs (Supplemental files 7,8) for several marker genes (*tbx16, sox10, sox32*, and *eya1*) and fgf/fgfr genes (*fgfr1a, fgfr1b, fgfrl1a, FGFRL1, fgfr2, fgfr3, fgfr4, fgf3, fgf4, fgf16, fgf17, fgf6, fgf6a, fgf8a,* and *fgf8b)*. Because we lacked Iso-Seq reads for *fgf4,* we used this modified gtf for the merging (Supplemental file 8).

We completed the processing of the ScISOr-Seq and bulk Iso-Seq data separately as described above then combined the files prior to alignment. We aligned the combined files with minimap2, collapsed transcripts with cDNA cupcake, and refined and classified isoforms with SQANTI3 as explained above.

We merged the sqanti_classification.filtered_lite.gtf resulting from SQANTI3 with the Ensembl annotation using TAMA (Kuo et al. 2020). To prepare the gtf files for TAMA Merge, we converted them to bed files using bedparse gtf2bed (--extraFields gene_id). Then, modified the output with awk to rearrange the columns of the bed file such that the gene id was separated from the transcript id by a semicolon (awk -v OFS='\t' '{print $1,$2,$3, $13 ";" $4, $5, $6,$7,$8,$9,$10,$11,$12}'). We merged with TAMA's script tama_merge.py (-s ensembl -cds ensembl -d merge_dup). We applied the ensembl gene names to genes with corresponding IDs using a custom script (tama_associating_ensembl_ids_with_genes.py). For our final gtf file, we converted the output bed file back to a gtf with TAMA's tama_convert_bed_gtf_ensembl_no_cds.py (Supplemental file 4). We also generated a gtf file specifically for scRNAseq analysis where mitochondrial genes started with "MT" by transforming the bed file with a custom script (tama_associating_ensembl_ids_with_genes_for_scRNAseq.py) and then converting the output back to a gtf in the same fashion as above (Supplemental file 5).

Annotation analysis

We assessed the Iso-Seq improvements to Ensembl gene models in two ways: 1) differences in starting and ending positions and 2) differences in reads captured. We compared the 5' start and 3' end of Ensembl and Iso-Seq generated gene models with a custom script written by H.M.H. (tama_observations_bed.py). Using the same custom script, we also compared the reads present in the first and last exons between Ensembl and Iso-Seq gene models.

scRNAseq analysis

We quantified gene counts using 10x Genomics Cell Ranger v3.0.2 (Supplemental file 1). We generated a reference using mkgtf and mkref, retaining protein coding and non-protein coding genes, and then aligned and counted reads using the stickleback genome (BROAD S1, 104.1). We analyzed the counts using the Seurat package (v3.2.3; (Stuart et al. 2019) on R (v4.0.2). We retained all cells for the analysis. We normalized the counts using SCTransform and

regressed out the mitochondrial genes using the glmGamPoi method (Hafemeister and Satija 2019). Based on the inflection point in the elbow plot of the PCA results, we chose 38 dimensions for generating the UMAP and identifying clusters. We identified cluster identities using three marker gene approaches.

First, we searched for marker genes identified in Farnsworth *et al*. (2020) for specific cell types found in a similar developmental stage from zebrafish (24hpf in zebrafish). Next, we identified markers with Seurat's FindAllMarkers searching for genes that were positively upregulated in each cluster compared to the other identities, in a minimum of 25% of cells, and a log fold change threshold of 0.25. Finally, we identified markers with Seurat's FindMarkers searching for genes that were positively upregulated in each cluster versus all the other cells. For the second and third approaches, we searched for the homologous gene in zebrafish and used expression data from zfin to predict which cell types expressed the gene. To quantify *frizzled* gene counts, we used FetchData to isolate raw counts for each gene and the number of cells expressing the gene. To determine the number of cell clusters each *frizzled* gene was expressed in, we saved a dotplot, subsetted for clusters where the percent of cells expressing the gene was greater than 10%, and then counted the number of clusters left for each gene.

Results and Discussion

ScISOr-Seq captured novel isoforms and improved 3' UTRs

A ScISOr-Seq library was produced using dissociated cells from 70 hpf stickleback embryos. The ScISOr-Seq reads were classified with SQANTI3 (Tardaguila et al. 2018) using Ensembl gene models to describe isoforms based on how well they match splice variants (Figure 2.1A). The most common structural category (45.15% of isoforms) was "novel not in catalog" containing at least one new splicing site relative to the existing annotation (Figure 2.1A and 2.1B). 17.42% of collapsed isoforms were "full splice matches" (FSM) that contained splicing sites and exons present in the Ensembl annotation but allowed differences in 5' or 3' ends. Notably, only 3% of the unique isoforms matched the existing gene annotations, while 97% improved existing or added gene models. Confirming that 3' UTRs were poorly annotated, 46.2% FSM isoforms had alternative 3' ends and 27.4% had alternative 3' and 5' ends relative to

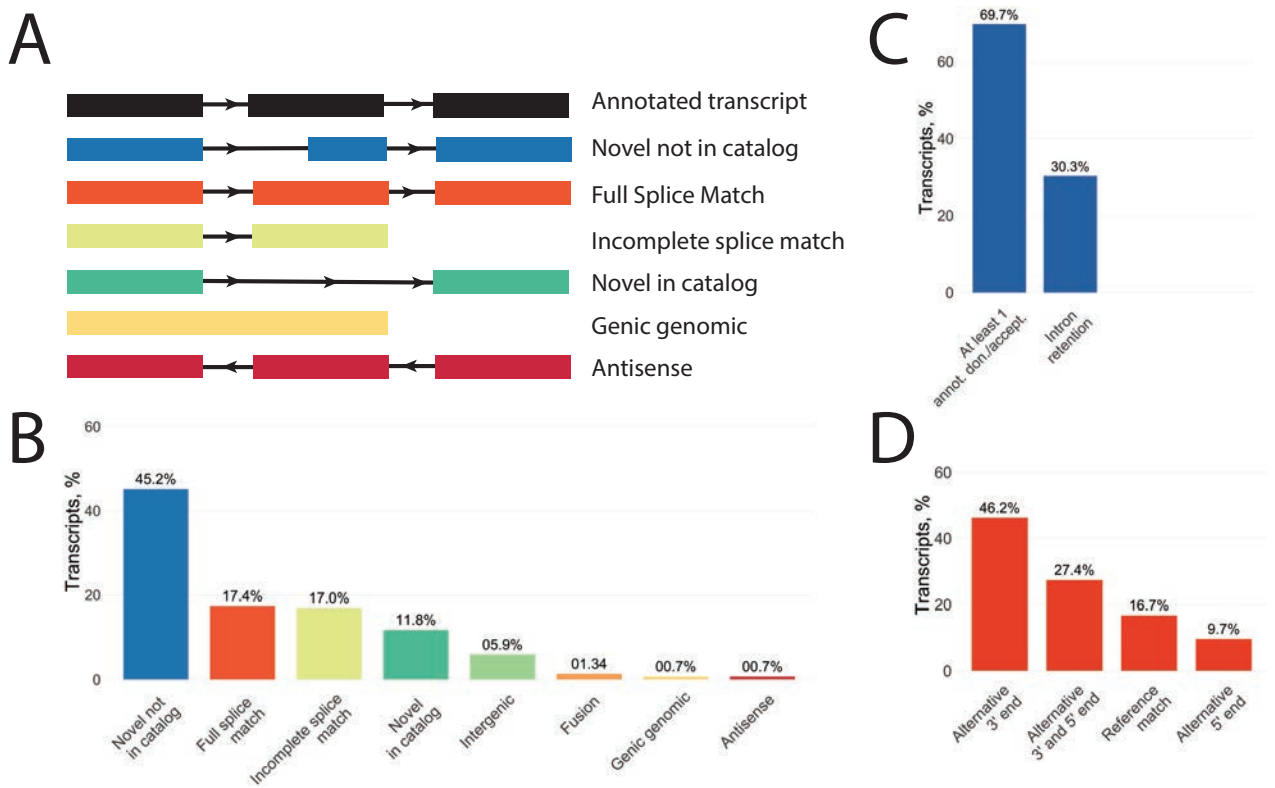


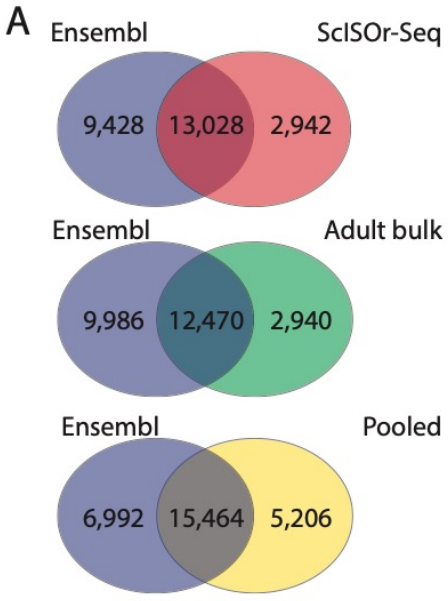
Figure 2.1: *SQANTI3 classification of ScISOr-Seq data revealed that the majority of isoforms were previously unannotated in the stickleback genome*. A) SQANTI3 categorizes isoforms based on their match to the reference, the major categories are shown here. B) the distribution of ScISOr-Seq isoforms in the major SQANTI3 classes illustrating that the bulk of isoforms are novel not in catalog. C) within the novel not in catalog category, isoforms are largely coming from cases where there is at least one new splicing donor or acceptor with some coming from cases of intron retention (for example, *nkx2.3* isoforms from Figure 2). D) the majority of isoforms that are full splice matches have at least an alternative 3' site while few are complete reference matches.

Ensembl gene models (Figure 2.1C). Therefore, the existing stickleback annotation - in addition to having incomplete 3' UTRs - is also missing many splice variants. Our work indicates that additional ScISOr-Seq or bulk Iso-Seq is necessary to capture these variants as well as prune erroneous transcript models from the Ensembl annotation.

ScISOr-Seq improved the number of reads retained for scRNAseq

We initially tested how well the existing stickleback gene models from Ensembl (BROAD S1, 104.1 database version) would capture scRNAseq reads. Strikingly, less than 50% of reads were retained for downstream scRNAseq analysis (Figure 2.2B; Supplemental file 1). Using such an incomplete annotation for scRNAseq would likely result in erroneous interpretation of gene expression patterns for genes lacking 3' UTR annotations.

Next, we used ScISOr-sequencing data to generate new gene models and tested how well these new models would capture 10X based Illumina scRNAseq reads (Supplemental file 2). Using the ScISOr-Seq dataset, 13,028 of 22,456 previously annotated genes and 2,942 novel genes were identified (Figure 2A). The ScISOr-Seq annotations lead to a notable 26.1% increase



B

	Number of cells	% of reads mapping to transcriptome	Median genes per cell
Ensembl	10,117	49.6%	1,197
ScISOr-Seq	11,343	75.7%	1,659
Adult bulk	10,902	64.7%	1,682
Pooled and merged	11,856	76.5%	1,597

Figure 2.2: *ScISOr-Seq and bulk Iso-Seq captured a limited number of annotated transcripts, but both resulted in an increase in reads mapping to the transcriptome*. Ultimately, pooling this data and merging with the ensembl annotation resulted in the largest gains in reads mapped. A) the number of genes in common and unique to each Iso-Seq read category is compared to the ensembl genome. B) results from cell ranger runs using the different genome annotation files illustrating the overall improvements by using Iso-Seq data particularly the ScISOr-Seq gene models.

in reads retained from scRNAseq compared to the Ensembl gene models alone (Figure 2B; Supplemental file 1). The alignment of scRNAseq reads with existing Ensembl and new ScISOr-Seq gene models illustrates that ScISOr-Seq models retain greater numbers of reads due to improved annotation of 3' UTRs (Figure 2.3; Figure S2.1). This result compares favorably with current research in the much better studied zebrafish model. Farnsworth, Saunders, and Miller (2020) completed their zebrafish atlas with 80% of reads retained in 5dpf (days post fertilization) fish. After (Lawson et al. 2020) improved the zebrafish transcriptome using RNAseq data, they noted a 4% increase in reads retained that lead to an increase of 2,257 cells and 8 clusters using the same 5 dpf atlas (Farnsworth et al. 2020). The similarity in number of retained reads in our dataset with those published in zebrafish improves the validity of stickleback scRNAseq investigations.

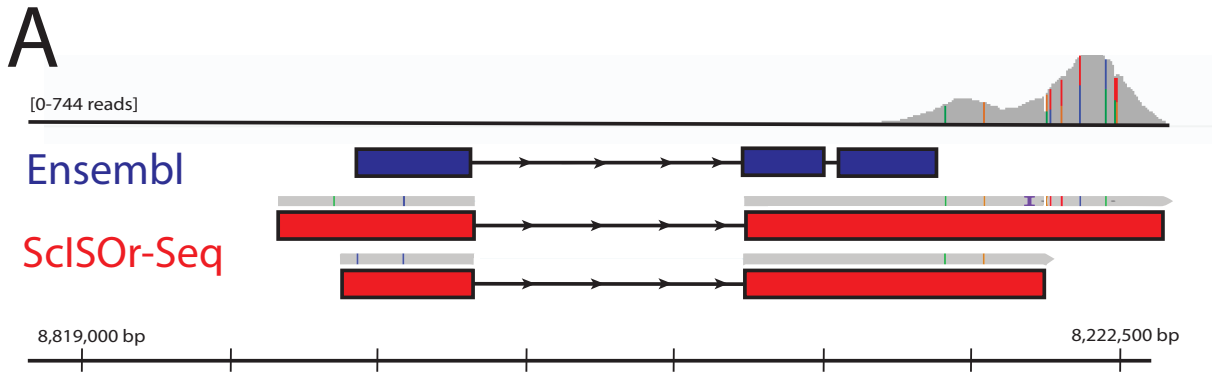


Figure 2.3: *ScISOr-Seq improved gene models by extending the 3' UTRs illustrated here by the two models for nkx2.3.* scRNAseq reads are shown at the top in grey. The existing ensembl model is shown in dark blue. Notably, this model does not overlap with the majority of the scRNAseq reads. The ScISOr-Seq reads (in grey) are shown above the respective gene models (in red) that they generated. Colored lines on the ScISOr-Seq and scRNAseq reads indicate a different base pair in the read than the stickleback reference genome (BROAD S1, 104.1 database version). These ScISOr-Seq gene models capture all of the scRNAseq reads.

Recently, (Naftaly et al. 2021) published a large bulk analysis of adult stickleback Iso-Seq data from 16 PacBio SMRT cells and 4 tissue types (gonad, brain, pronephros, and liver). We reanalyzed this dataset to test whether we would see comparable improvements in our scRNAseq analysis (Supplemental file 3). Although we identified similar overall annotated and novel genes, the bulk Iso-Seq annotation from adult stickleback captured 11% fewer reads than our embryonic ScISOr-Seq annotation (Figure 2.2A and 2.2B; Supplemental file 1). These results are likely due to differential expression of transcripts between the adult and embryonic

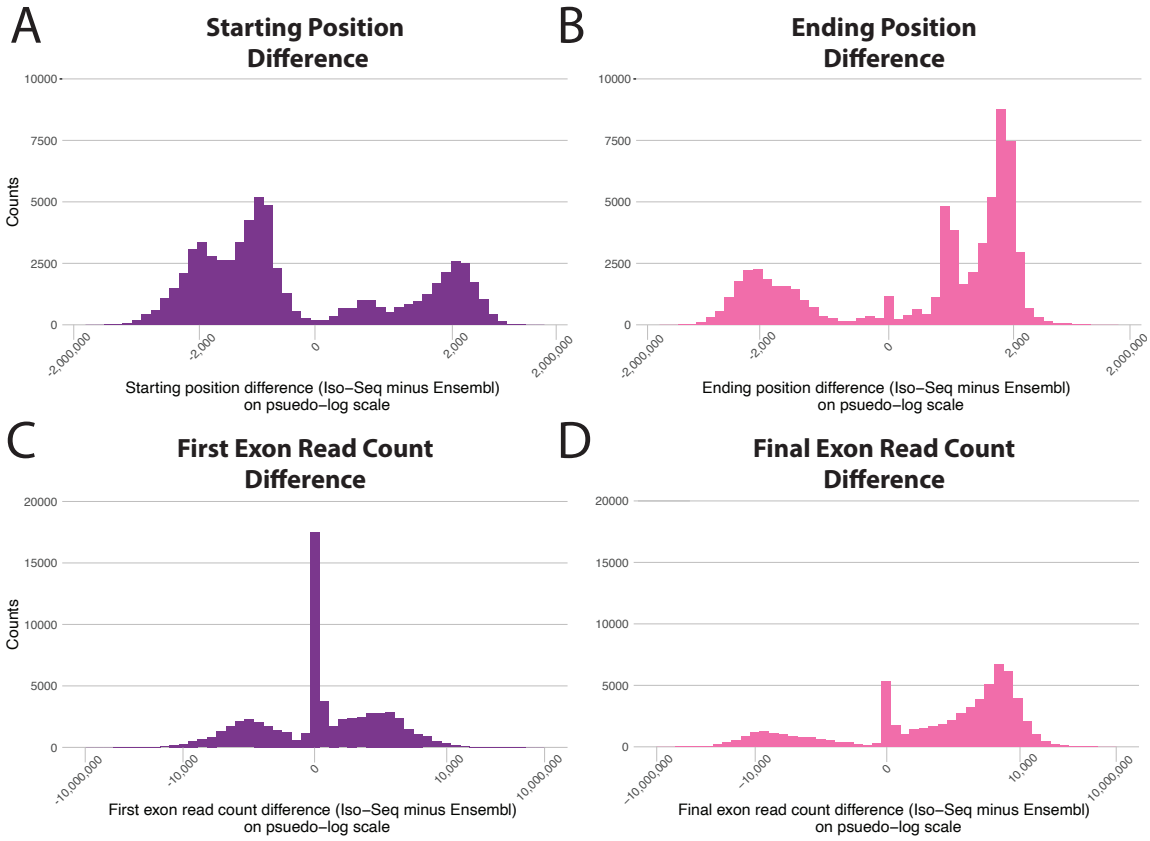


Figure 2.4: *Annotation improvements using Iso-Seq (bulk Iso-Seq or ScISOr-Seq) data extended genes both 5' and 3', with 3' expansions leading to greater numbers of reads counted*. Comparing the median difference in starting or ending position between the Ensembl annotations and pooled Iso-seq annotations shows that bulk Iso-Seq and ScISOr-Seq reads overall lead to earlier starting positions (median=26, a) and extended ending positions (median=94, b). These Iso-Seq differences increased the number of reads counted in the final exon of the gene (median=341, d); however, the alterations led to negligible changes of reads counted in the first exon (median=0, c), which highlights the 30 bias of scRNAseq reads. The x-axis is plotted on a pseudo log scale to account for the negative values.

samples and highlight that even for species with previous Iso-Seq libraries, additional sequencing may be needed for scRNAseq analysis in other developmental stages or tissue types.

Because both our ScISOr-Seq and the bulk Iso-Seq gtf files improved the number of scRNAseq reads retained, we pooled them to create a comprehensive annotation and also merged this pooled dataset with the Ensembl annotations. We retained all transcripts and united them under a single gene model for scRNAseq analysis (Supplemental file 4). The Iso-Seq generated gene models extended 3' ends of transcripts and increased reads counted in the final exon (Figure 2.4). This new annotation contained 6,992 genes unique to Ensembl, 15,464 genes containing transcripts from Ensembl and Iso-Seq, and 5,206 genes unique to the pooled Iso-Seq dataset (Figure 2.2A). Since reads are filtered during analysis, the 6,992 genes unique to Ensembl might still be expressed in 70hpf embryos or the adult tissues. For instance, 2,152 Ensembl annotated genes were removed from the pooled Iso-Seq dataset by SQANTI3's intra-priming and template switching filter. Although pooling and merging added 16 SMRT cells of Iso-Seq data as well as the full suite of Ensembl annotations, this effort only improved the proportion of reads retained by +0.8% beyond the annotation created by just our single ScISOr-Seq sample (Figure 2.2B; Supplemental file 1). Therefore, the gene models originating from a matching ScISOr-Seq library alone could be used for scRNAseq analysis if a system lacks a prior annotation.

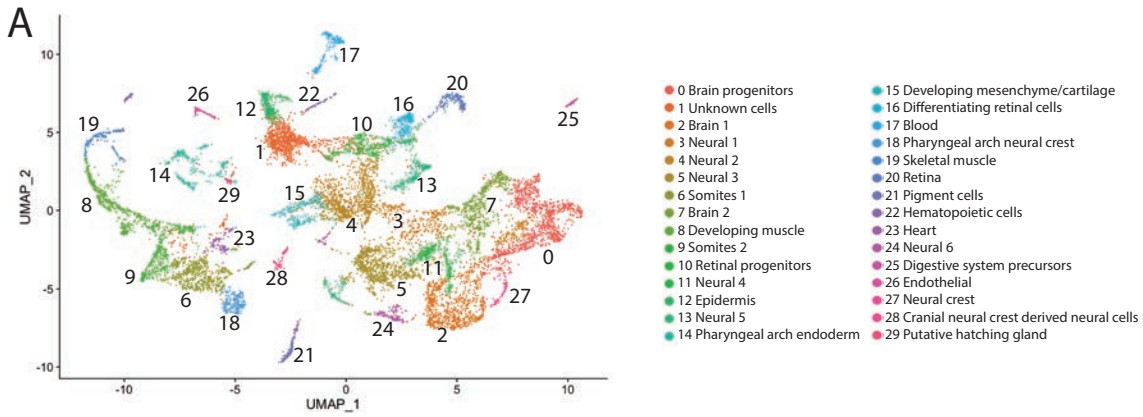
ScISOr-Seq enhances biological interpretation of cell clusters

Using the pooled (ScISOr-Seq and bulk Iso-Seq) annotation that was merged with Ensembl annotation, we proceeded with scRNAseq analysis (Supplemental file 5) to test whether biologically meaningful cell clusters would be created. We clustered cells by their transcriptional profiles into cell identities and analyzed the results with Seurat (Stuart et al. 2019). We identified 30 clusters of cells with 38 principal components (Figure 2.5A). We putatively annotated cell types via distinguishing sets of marker genes based on zebrafish literature (Howe et al. 2013; Wagner et al. 2018; Farnsworth et al. 2020).

Although zebrafish and stickleback diverged ~ 229.9 m.y.a. (Howe et al. 2021), we identified similar cell types as observed in existing zebrafish atlases for similar developmental stages as we analyzed in stickleback (Wagner et al. 2018; Farnsworth et al. 2020). Additionally, we compared the expression of *sox9a* and *sox9b* in our atlas to published descriptions of stickleback *in situ* expression at the same stage (Cresko et al. 2003). Supporting our cluster

annotations, we observed expression of both SOX genes in roughly the same cell types as defined by *in situ* hybridization (Figure S2.2A and S2.2B; Cresko et al. 2003). Despite our success using a single SMRT cell for ScISOr-Seq for isoform annotation, we would need to sequence additional SMRT cells to have enough data to correlate cell types and specific isoforms.

Since the pooled and merged dataset has higher numbers of cells, genes per cell, and total genes detected than the Ensembl scRNAseq dataset (Figure 2.2B), assessing the degree to which pooled and merged gene models improved the scRNAseq analysis is complex. For instance, a higher overall expression of a gene in a cluster might be the result of different clustering patterns, more cells retained, and/or more reads counted. We compared the raw counts of specific genes, number of cells expressing these genes, and number of clusters where at least



Genes	Total counts (ENS)	Total counts (P&M)	# cells expressing gene (ENS)	# cells expressing gene (P&M)	# clusters expressed (ENS)	# clusters expressed (P&M)
fzd1	**162**	**4076**	**160**	**2097**	**0**	**9**
fzd2	13	736	13	539	0	2
fzd3a	20	20	20	20	0	0
fzd3b	82	685	77	477	0	1
fzd4	103	3238	101	1625	0	9
fzd5	1352	1962	591	753	3	5
fzd6	132	2327	129	1594	0	8
fzd7	0	200	0	192	0	0
fzd8b	35	3164	35	1573	0	8
fzd9b	**3818**	**4263**	**1663**	**1807**	**12**	**10**
fzd10	689	693	420	424	2	1

Figure 2.5: *Using the updated gene models from our merged and pooled dataset, we identified expected cell types in 70hpf stickleback scRNAseq data and illustrated how annotations improvements are relevant for analysis*. A) the clustering and cell identities of the scRNAseq data is illustrated on the UMAP. B) frizzled gene family expression compared between the ensembl annotation scRNAseq analysis and the pooled and merged analysis using total raw counts, number of cells expressing each gene, and the number of clusters where over 10% of cells are expressing each gene. C) *fzd1* gene model is greatly extended in the pooled and merged annotation (purple) compared to the ensembl model (blue) which allows for much greater counting of reads (grey). D) *fzd9b* had minor changes in the pooled and merged model (purple), a longer 5' UTR, in comparison with the ensembl annotation (blue) which led to minimal increases in reads (grey) counted. The scRNAseq read pileups in C and D have colored lines on base pairs that are different from the stickleback reference genome (BROAD S1, 104.1 database version).

10% of cells expressed them (Figure 2.5B). Since Cell Ranger provided a subset of these values on a global scale (Figure 2.2B), we chose to examine changes in gene counts, cell number, and number of clusters with a case study: the *frizzled* genes, a family of Wnt-pathway signaling molecules expressed in a wide range of tissues (Huang and Klein 2004; Wang et al. 2016). For 10/11 *frizzled* genes, we observed increases in raw counts and number of cells expressing them in the pooled and merged annotation (Figure 5B). Some genes, however, exhibited dramatic increases (*fzd1, fzd2, fzd3b, fzd4, fzd6, fzd7,* and *fzd8b*) while others exhibited minor increases (*fzd5, fzd9b,* and *fzd10*). Comparing the gene models of *fzd1* and *fzd9b,* we determined which factors influenced expression detection. Importantly, the 5' and 3' UTRs expanded for *fzd1* in the merged and pooled annotation, however only the 5' end of *fzd9b* was extended (Figure 2.5C and 2.5D). Based on the read alignments (Figure 2.5C and 2.5D), the increase in counted reads for *fzd1* was due solely to 3' UTR changes, while a slight increase in *fzd9b* reads was due to 5' end extension. In addition to increased read counts, 7/11 genes were expressed in more cell clusters in the pooled and merged dataset. Two genes, *fzd9b* and *fzd10,* were found in one less cluster. We hypothesized that two clusters (containing *fzd9b* and *fzd10*) were combined in the merged and pooled dataset relative to Ensembl. Overall, this gene family comparison illustrated the problem of using an annotation with incomplete 3'UTRs. If the Ensembl dataset's *fzd* counts were accepted as their true qualitative and quantitative expression, flawed conclusions would have been made regarding the contributions of each gene in the family.

Conclusion

Incomplete 3'UTR annotations can hinder single cell transcriptional profiling studies. In addition to reducing the overall number of genes included in an analysis, systematic differences in preliminary UTR annotations could lead to significant inferential errors. We illustrated in stickleback fish that a minimal ScISOr-Seq dataset, generated concurrently with scRNAseq data, was capable of dramatic improvements in retained read counts (+26.1%). Additionally, we showed that pooling additional Iso-Seq reads from adult bulk samples and merging with existing Ensembl annotations improved scRNAseq reads retained negligibly beyond models solely from embryonic ScISOr-Seq (+0.8%). Using the improved annotation for scRNAseq permitted identification of cell types and increased the observed expression of numerous genes. Overall,

our work illustrates that ScISOr-Seq is a rapid and cost-effective method to annotate genomes of various organisms for scRNAseq and can improve efficacy of biological inferences.

References

Ali A, Thorgaard GH, Salem M. 2021. PacBio Iso-Seq improves the rainbow trout genome annotation and identifies alternative splicing associated with economically important phenotypes. *Front. Genet*. 12.

Beck EA, Currey MC, Small CM, Cresko WA. 2020. QTL mapping of intestinal neutrophil variation in threespine stickleback reveals possible gene targets connecting intestinal inflammation and systemic health. *G3*. 10(2):613–622.

Beck EA, Healey HM, Small CM, Currey MC, Desvignes T, Cresko WA, Postlethwait JH. 2021 Jul 29. Advancing human disease research with fish evolutionary mutant models. *Trends Genet*. 38(1):22-44.

Beiki H, Liu H, Huang J, Manchanda N, Nonneman D, Smith TPL, Reecy JM, Tuggle CK. 2019. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. *BMC Genomics*. 20(1):1–19.

Bell MA, Foster SA. 1994. The evolutionary biology of the threespine. 1st ed. New York: Oxford University Press.

Bresciani E, Broadbridge E, Liu PP. 2018. An efficient dissociation protocol for generation of single cell suspension from zebrafish embryos and larvae. *MethodsX*. 5:1287.

Colosimo P, Summers B, Kingsley D, Hosemann K. 2004. A simple and efficient microinjection protocol for making transgenic sticklebacks. *Behaviour*. 141(11–12):1345–1355.

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Phillips P, Bell MA, Kimmel CB, Postlethwait JH. 2004a. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc. Natl. Acad. Sci. U. S. A*. 101(16):6050–6055.

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Phillips P, Bell MA, Kimmel CB, Postlethwait JH. 2004b. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc. Natl. Acad. Sci*. 101(16):6050–6055.

Cresko WA, McGuigan KL, Phillips PC, Postlethwait JH. 2007. Studies of threespine stickleback developmental evolution: Progress and promise. *Genetica*. 129(1):105–126.

Cresko WA, Yan YL, Baltrus DA, Amores A, Singer A, Rodríguez-Marí A, Postlethwait JH. 2003. Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Dev. Dyn*. 228(3):480–489.

Farnsworth DR, Saunders LM, Miller AC. 2020. A single-cell transcriptome atlas for zebrafish development. *Dev. Biol*. 459(2):100–108.

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. 2018. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*. 360(6392).

Fuess LE, den Haan S, Ling F, Weber JN, Steinel NC, Bolnick DI. 2021. Immune gene expression covaries with gut microbiome composition in stickleback. *MBio*. 12(3).

Gardell AM, von Hippel FA, Adams EM, Dillon DM, Petersen AM, Postlethwait JH, Cresko WA, Buck CL. 2017. Exogenous iodide ameliorates perchlorate-induced thyroid phenotypes in threespine stickleback. *Gen. Comp. Endocrinol*. 243:60–69.

Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, Koopmans F, Barres B, Smit AB, Sloan SA, et al. 2018. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol*. 36(12):1197–1202.

Hafemeister C, Satija R. 2019. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. 20(1):296.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet*. 6(2).

Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SAT, et al. 2013. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. *Nucleic Acids Res*. 41(D1):D854–D860.

Howe KL, Achuthan P, Allen James, Allen Jamie, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al. 2021. Ensembl 2021. *Nucleic Acids Research*. 49(D1):D884–D891.

Huang H-C, Klein PS. 2004. The Frizzled family: receptors for multiple signal transduction pathways. *Genome Biol*. 5(7):1–7.

Hwang B, Lee JH, Bang D. 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. and Mol. Med*. 50(8).

Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. 2020. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*. 21(1):1–22.

Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. 2017. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics*. 18(1):1–19.

Lawson ND, Li R, Shin M, Grosse A, Onur YMS, Stone OA, Kucukural A, Zhu LJ. 2020. An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes. *eLife*. 9:1–76.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34(18):3094–3100.

Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol*. 15(6).

Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, Kingsley DM. 2007. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell*. 131(6):1179–1189.

Naftaly AS, Pau S, White MA. 2021. Long-read RNA sequencing reveals widespread sex-specific alternative splicing in threespine stickleback fish. *Genome Res*. 31(8):1486–1497.

Reid K, Bell MA, Veeramah KR. 2021. Threespine stickleback: a model system for evolutionary genomics. *Annu Rev Genomics Hum Genet*. 22:357–383.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Gen*. 14(9):618–630.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*. 428(6984):717–723.

Small CM, Milligan-Myhre K, Bassham S, Guillemin K, Cresko WA. 2017. Host genotype and microbiota contribute asymmetrically to transcriptional variation in the threespine stickleback gut. *Genome Biol. Evol*. 9(3):504–520.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive integration of single-cell data. *Cell*. 177(7):1888-1902.e21.

Tardaguila M, Fuente L de la, Marti C, Pereira C, Pardo-Palacios FJ, Risco H del, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. 2018. SQANTI: extensive

characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res*. 28(3):396.

Tseng E. 2021. cDNA cupcake.

Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. 2018. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*. 360(6392):981–987.

Wang Y, Chang H, Rattner A, Nathans J. 2016. Frizzled receptors in development and disease. *Curr. Top. Dev. Biol*. 117:113.

Zheng Y-F, Chen Z-C, Shi Z-X, Hu K-H, Zhong J-Y, Wang C-X, Shi W, Chen Y, Xie S-Q, Luo F, et al. 2020 Jul 28. HIT-scISOseq: high-throughput and high-accuracy single-cell full-length isoform sequencing for corneal epithelium. *bioRxiv*.:2020.07.27.222349.

BRIDGE

Chapter II detailed challenges to using single cell RNA sequencing (scRNAseq) in less traditional model organisms. This technique captures reads at the 3' end of mRNA transcripts, therefore it requires high quality 3' UTR annotations that are often incomplete in non-model systems. I harnessed single cell ISO-sequencing data to improve 3' UTR annotations in threespine stickleback. Through lengthening 3' UTR sequences, I increased the percentage of scRNAseq reads retained by 26.5%.

In Chapter III, I demonstrated the utility of these techniques for analyzing the development of highly derived traits in Gulf pipefish. I created a single cell atlas for the Gulf pipefish, using single cell ISO-sequencing to improve gene models and single cell RNA-sequencing to quantify gene expression. Using this atlas, I studied the development of pipefish's elongate snout and exoskeleton, explored developmental origins of their loss of teeth, and asked whether epidermal cells express genes related to development within the brood pouch. This study provided preliminary evidence that pipefish's derived traits may originate from changes within developmental gene networks.

CHAPTER III

SINGLE CELL RNA SEQUENCING PROVIDES CLUES FOR THE DEVELOPMENTAL GENETIC BASIS OF SYNGNATHID FISHES EVOLUTIONARY ADAPTATIONS

From Healey, Penn, Small, Bassham, Goyal, Woods, Cresko. 2024. Single Cell RNA Sequencing Provides Clues for the Developmental Genetic Basis of Synganthidae's Evolutionary Adaptations. *eLIFE.*

The full supplemental material for thesis chapter can be found at https://doi.org/10.7554/eLife.97764.1

This project was conceptualized by William Cresko and I. Sequencing data was generated by Clay Small and me. I completed the scRNAseq analysis. The *in situ* hybridizations were largely completed by Hayden Penn and myself with assistance from Susan Bassham, Vithika Goyal, and Micah Woods. I took the lead on writing the manuscript. The paper was reviewed and revised by all authors.

Introduction

Seahorses, pipefishes, and seadragons are extraordinary fishes in the family Syngnathidae with diverse body plans, coloration, and elaborate structures for paternal brooding. The syngnathid clade comprises over 300 diverse species that vary in conservation status, distribution, ecology, and morphology (Leysen et al., 2011; Manning et al., 2019; Schneider, Woltering, et al., 2023; Stiller et al., 2022). Syngnathids have numerous highly altered traits, trait losses, and evolutionary novelties. They have elongated snouts bearing small, toothless jaws (Leysen et al., 2011) specialized for capture of small zooplankton (Van Wassenbergh et al., 2009). Additionally, syngnathids are distinctly protected by a bony dermal armor rather than scales (Jungerson, 1910). Other skeletal differences include a lack of ribs and pelvic fins, and an expansion in the number of vertebrae (Schneider, Woltering, et al., 2023). Finally, syngnathids exhibit male pregnancy, which has involved the evolution of specialized brooding tissues and structures (Whittington & Friesen, 2020). Paternal investment varies among lineages; for example, seadragons tether embryos externally to their tails while seahorses and some pipefishes

have enclosed brood pouches proposed to support embryos through nutrient transfer and osmotic regulation (Carcupino, 2002; Melamed et al., 2005; Ripley & Foran, 2006).

Despite advances in understanding the ecology and evolution of syngnathid novelties, the developmental genetic basis for these traits is largely unknown. The recent production of high quality syngnathid genome assemblies (Qu et al., 2021; Ramesh et al., 2023; Small et al., 2022; Wolf et al., 2024) provides initial clues for the developmental genetic basis of some evolutionary changes. Studies have found that syngnathids lack several genes with deeply conserved roles in vertebrate development, including pharyngeal arch development (*fgf3*), tooth development (*fgf3*, *fgf4*, and *eve1*, and most *scpp* genes), fin development (*tbx4*), and immune function (MHC pathway components) (Lin et al., 2016; Qu et al., 2021; Small et al., 2016; Small et al., 2022; Zhang et al., 2020). Though these gene losses are highly suggestive of leading to unique changes, exploration of the actual developmental consequences of their losses is needed.

To fill this gap in knowledge, we used single cell RNA sequencing (scRNAseq) to investigate how these striking genomic changes have affected the developmental genetic and cellular basis of syngnathids' evolutionary derived traits. The Gulf pipefish (*Syngnathus scovelli*) is an attractive model for this study (Figure 3.1A). This species has a high-quality reference genome assembly annotated by NCBI and is amenable to laboratory culture (Anderson & Jones, 2019; Ramesh et al., 2023). Furthermore, species from the *Syngnathus* genus are used worldwide to address questions about syngnathid evolution in microbial, developmental, histological, transcriptomic, ecotoxicological, and genomic studies (Berglund et al., 1986; Carcupino, 2002; Fuiten & Cresko, 2021; Harada et al., 2022; Harlin-Cognato et al., 2006; Partridge et al., 2007; Ripley & Foran, 2006a; Rose et al., 2023; Roth et al., 2012; Small et al., 2016; Small et al., 2013). In this paper, we focus on a subset of unique traits, including their elongated head, toothlessness, dermal armor, and development of embryos inside the brood pouch. These traits represent the diversity of evolutionary changes observed in the syngnathid clade (highly altered, lost, and novel traits) and hypotheses from studies of model organisms suggest developmental pathways involved in their evolution (Lin et al., 2016b; Roth et al., 2020; Small et al., 2016; Small et al., 2022).

scRNAseq atlases are a powerful complement to genomic analyses (Shema et al., 2019; Ton et al., 2020). They can provide crucial insights into types of cells present, genes that distinguish cell types (marker genes), active gene networks, and a means to identify expression of

genes of interest within predicted cell types (Farnsworth et al., 2020; Farrell et al., 2018; Williams et al., 2019). Specifically, scRNAseq captures RNA expression profiles from individual cells allowing cell types to be inferred *post-hoc*. scRNAseq has successfully been applied to syngnathid adult kidneys (Parker et al., 2022), but there are no published syngnathid developmental atlases.

Here we report the first developmental scRNAseq atlas for syngnathids from late embryogenesis staged Gulf pipefish. We delineate the overall structure of this atlas, which describes 38 cell clusters composed of 35,785 cells, and use these data to make inferences about the morphological evolution of several syngnathid innovations. In addition to inferring present cell types and their underlying genetic networks, we detail *in situ* spatial expression patterns of select marker and other candidate genes in pipefish embryos and juveniles. We found conserved signaling pathways expressed during craniofacial development but did not detect evidence of tooth primordia. The embryonic dermis and epidermis, respectively, expressed genes for the dermal armor development (bone development pathways) and genes potentially involved in interaction with the male brood pouch (e.g., nutrient acquisition genes). Overall, this atlas provides a deeper understanding of the development of Gulf pipefish and identifies gene candidates for understanding the development of syngnathid evolutionary innovations. In

addition to these discoveries, this atlas provides a significant resource for researchers studying syngnathid evolution and development.

Materials and Methods

Single Cell RNA Sequencing Libraries Preparation

We created scRNAseq atlases from embryos of wild caught Gulf pipefish (*Syngnathus scovelli,* acquired from our collaborator Emily Rose using collection permit SAL-21-0182-E),

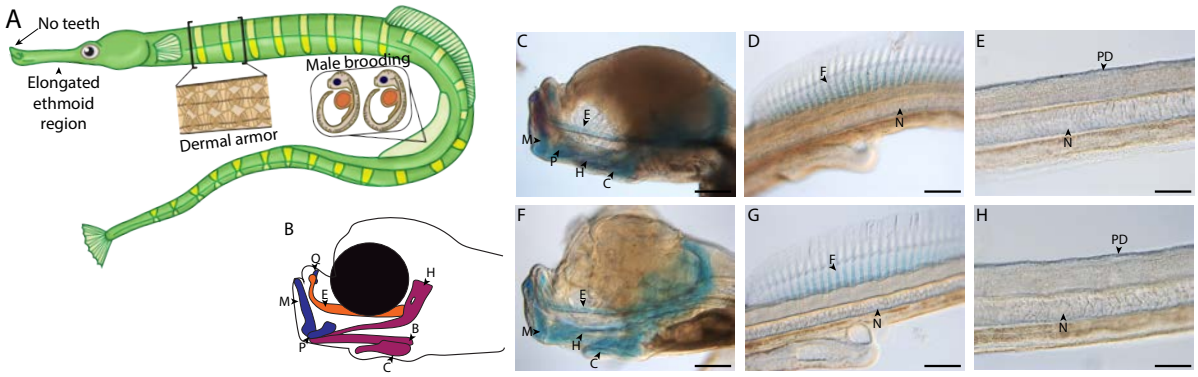


Figure 3.1: *Gulf pipefish exemplify syngnathid derived traits.* Gulf pipefish have elongate snouts, have lost teeth on their oral and pharyngeal jaws, possess dermal armor, and have brood pouches in males (A-B). Cartilage (alcian) and bone (alizarin) stained clutch siblings of embryos from the two scRNAseq samples are shown in panels C-H. Embryos have cartilaginous craniofacial skeletons (panels B, C, F; E marks the (Mes)ethmoid cartilage, C indicates the Ceratohyal, H shows the Hyosymplectic cartilage, M marks the Meckel's cartilage, Q indicates the quadrate, B shows the Basihyal, and P marks the palatoquadrate) with the onset of ossification in the jaw. They have cartilaginous fin radials in the dorsal fin (D and G; F indicates fin radials and N denotes the notochord). The embryos do not have signs of ossification in the trunk where the exoskeleton will form later (panels E and H; PD shows the region where dermal armor primordia will arise). Panels C, D, F, and G, scale bar 200µm; E, H scale bar 100 µm.

and all work was performed under an approved IACUC protocol. We harvested 20 embryos per pouch from two wild caught male pipefish. The embryos were at a stage before the tubular face was fully elongated, and while the head skeleton is cartilaginous with minimal signs of mineralization of superficial intramembranous bones. This corresponds to a stage termed “frontal jaws” in a recent description of pipefish development (Sommer et al., 2012).

We dissociated the embryos using 460ul of .25% trypsin in water and 40ul 100 mg/mL collagenase I (Sigma C0130-200mg) for 16 minutes. We filtered cells using a 40uM cell strainer

(Thomas Scientific #1181X52). We quantified cell concentrations using the TC20 Automated Cell Counter (Biorad) then diluted the samples to 800 cells/ul in 4% BSA in PBS. The University of Oregon Genomics and Cell Characterization Core (GC3F; https://gc3f.uoregon.edu) prepared single cell libraries for each sample using 10X Genomics Single Cell 3' Genome Expression mRNAseq kit with NextGEM v3.1 chemistry. We sequenced these libraries on an S4 lane on the NovaSeq 6000 at the GC3F. To improve the 3' UTR genome annotations, we also prepared scISOrSeq libraries from the first embryonic sample and from dissociated pouch cells from pregnant and nonpregnant males. These libraries were produced in accordance with (Healey et al., 2022). Embryonic, pregnant pouch and non-pregnant pouch libraries were sequenced separately on PacBio Sequel II - SMRT Cells 8M.

To turn the scISOrSeq reads into gene models, we followed the pipeline from Healey et al. (2022). We ran our custom script (scISOr_Seq_processing.py) to remove barcodes, identify cell barcodes, and demultiplex with the single cell flag and appropriate barcodes (5' CCCATGTACTCTGCGTTGATACCACTGCT and 3' CTACACGACGCTCTTCCGATCT). We aligned the reads to the 2022 Gulf pipefish genome GenBank: GCA_024217435.2) using minimap v2.9 (cite: Li 2018). We filtered the reads using cDNA cupcake to remove duplicate transcripts (Tseng, 2021). We used SQANTI3 to identify gene models and filter them (Tardaguila et al., 2018). We merged the SQANTI3 annotations with the Gulf pipefish genome assembly (NCBI: ROL_Ssco_1.1) using TAMA merge (Kuo et al., 2017). Since the Gulf pipefish genome assembly does not contain mitochondrial genes, we appended the genome files with the Gulf pipefish mitochondrial genome (NCBI RefSeq: NC_065499.1).

Single Cell Atlas Construction

We ran Cell Ranger (10X Genomics 3.0.2) using our scRNAseq reads, the Gulf pipefish genome assembly with mitochondrial genome, and the modified gene annotations. Cell Ranger estimated 20,733 cells for sample one, 23,682 genes expressed, and 21,039 mean reads per cell. For sample two, Cell Ranger predicted 17,626 cells, 23,740 genes expressed, and 29,804 mean reads per cell. We analyzed Cell Ranger's output using Seurat (v4.1.0) on R (v4.0.2; Butler et al., 2018; Hafemeister & Satija, 2019).

To remove extraneous RNA counts from the dataset, we used SoupX (v1.5.2; Young & Behjati, 2020). We identified doublet scores for our dataset using scrublet (v0.2.3). The doublet removal step reduced the first sample by 114 cells (from 20,733 cells to 20,619 cells) and second sample by 167 cells (from 17,626 cells to 17,459 cells). We finally removed cells with less than 500 features, greater than 9000 features, greater than 1E5 RNA counts, with a scrublet score greater than the detected threshold (.76 for sample 2 and .21 for sample 2), or greater than 10% mitochondrial reads. The second filtering step removed 727 cells from sample one (20,619 cells to 19,892 cells) and 1,566 cells from sample two (from 17,459 cells to 15,893 cells).

We normalized the datasets with SCTransform (v0.3.3). We used Seurat's integration tools, SelectIntegrationFeatures using 3,000 feature genes, FindIntegrationAnchors using SCT normalization, and IntegrateData using SCT normalization, to integrate the two datasets. After integration, our combined atlas had 35,785 cells (Supplementary File 1,2). We then used the integrated dataset to complete PCA analysis. We tested using a variety of principle components for further analysis and chose 30 PCs for our analysis based on the clear delineation of major cell types. We next clustered the cells using 30 PCs and plotted the data on a UMAP with Seurat.

Single Cell Atlas Cluster Identification

To identify cluster identities, we used the RNA assay of the scRNAseq data to find cluster markers with Seurat's FindAllMarkers command with the parameters only.pos = TRUE and logfc.threshold = 0.25, requiring markers to be upregulated in the cluster and have a log fold change of at least .25. We found a second set of cluster markers through our custom function which searched through all genes and identified genes uniquely expressed in greater than 60% of cells in the cluster and in less than 10% of cells in every other cluster using Seurat's DotPlots. We searched for our identified markers in available zebrafish datasets (Fabian et al., 2022; Farnsworth et al., 2020; Lange et al., 2023), ZFIN (Howe et al., 2013), NCBI, medlineplus.gov, and genecards to give the clusters initial annotations. For each cluster, we examined multiple genes using DotPlots and FeaturePlots to propose the cluster identity.

Next, we identified one gene for each cluster which marked the cluster best (expressed in the most cells in the focal cluster and expressed in as few of the other clusters as possible) through consulting the two marker gene lists and examining markers with Dot Plots (Supplementary File

3). Using these markers, we completed a set of *in situ* hybridizations to hone our cluster annotations. Due to challenges culturing Gulf pipefish, we used both embryos and larvae from *Syngnathus leptorhynchus*, a local pipefish from the same genus, and from cultured Gulf pipefish for the cluster annotation *in situ* experiments. We caught pregnant male *Syngnathus leptorhynchus* using a beach seine near Coos Bay, Oregon under Oregon Department of Fisheries and Wildlife permit number 26987.

Syngnathus scovelli used for *in situ* experiments were purchased from Alyssa's Seahorse Saavy and Gulf Specimens Marine Lab then reared in our facility at 25 degrees C water and 25-28 PPT Salinity. Since *Syngnathus leptorhynchus* does not have a published genome assembly, we designed probes using NCBI Primer Blast with the Gulf pipefish genome and produced these probes using Gulf pipefish embryonic cDNA pools. To create the probes, we completed two rounds of PCR. The first round used a gene specific forward primer with a reverse gene specific primer that had 10 nucleotides of T7 promotor sequence attached. PCR products were cleaned with Zymo clean and concentrator DNA kit and eluted in 15 ul of elution buffer. Round two of PCR used the same gene specific forward primer with a modified T7 promoter sequence (TGGACTAATACGACTCACTATAGGG) as the reverse primer, and finally the product was cleaned again with Zymo clean and concentrator DNA kit and eluted in 15 ul of elution buffer.

Round one PCR conditions are as followed: 95 degrees Celsius for 3:00 minutes, 40 cycles of denaturation (95 degrees for 30 seconds), annealing (30 seconds, annealing temperature varied by probe), and extension (72 degrees for 1:00 minute), and a final extension step of 3:00 minutes. Round two PCR conditions are as followed: 98 degrees Celsius for 30 seconds, 30 degrees for 10 seconds, 72 degrees for 50 seconds, then 35 rounds of denaturation (98 degrees for 10 seconds), annealing (50 degrees for 10 seconds), and extension (72 degrees for 50 seconds), and a final extension step of 10 minutes at 72 degrees. For round two PCRs with multiple bands (specifically, *ifitm5*), the band of the expected size was excised with a razorblade and DNA was extracted with Zymoclean gel DNA recovery kit.

The round two PCR product was Sanger sequenced to confirm identity. All the marker genes primers and successful PCR conditions are in Supplementary File 4. The probes were transcribed with T7 polymerase for 2-6 hours then cleaned with Zymo RNA clean and concentrator and eluted into 30 ul of water. For the *in situ* hybridizations, we selected embryos and newly spawned larvae close to the developmental stage used in the atlas. We completed *in situ*

hybridizations in keeping with Thisse & Thisse, 2007, leaving the embryos in stain until background was observed. For spawned larvae, we completed a bleaching step (1% H2O2 and .5% KOH for 8 minutes) prior to the proteinase K digest. After imaging, we used the levels tool in Adobe Photoshop (v23.4.2) to white balance the pictures.

Differentiation State Analysis

To assess whether proposed primordial cell clusters were composed of undifferentiated cells relative to other clusters from their lineages, we completed a differentiation analysis using CytoTRACE (v0.3.3). Clusters from similar lineages (neural: 0, 3, 7, 8, 12, 22, 25, 33, and 35; muscle: 2, 10, 17, and 37; connective: 4, 5, 6, 9, 15, 16, 18, 24, 27, 28, and 29) were isolated using subset. Cell counts were gathered using as.Matrix(GetAssayData), then CytoTRACE was run on these counts. Data was plotted on a VlnPlot to show the variability.

Single Cell KEGG Analysis

To identify pathways upregulated in cell clusters, we completed a KEGG analysis. We downloaded gulf pipefish KEGG pathways (https://www.kegg.jp). For the KEGG analysis, we used the marker genes identified from our FindAllMarkers list as the input genes. We converted these gene ids using keggConv to KEGG ids. We used a Wilcoxon enrichment test to ask whether cluster marker genes were enriched for each KEGG pathway.

Single Cell Atlas Network Analysis

To identify genetic networks present in our atlas, we completed a weighted gene network correlation analysis using WGCNA (v1.72-1,Langfelder & Horvath, 2008). We selected 3,000 variable features from the integrated assay of the single cell dataset for the WGCNA. We created an adjacency matrix from the data using bicor with a maxPOutliers of .05. To decide on a beta value or the soft threshold power, we created an adjacency matrix plot using pickSoftThreshold and picked the threshold where the scale free topology model fit leveled off. We selected the value of two. Then, we raised the adjacency matrix to the power of two.

We calculated the dissimilarity matrix by calculating TOM similarity of the adjacency matrix and subtracting it from one. We then created a gene tree through hclust of the dissimilarity matrix with the average method. From the gene tree, we created modules using

cutreeDynamic with deepSplit of two and the minClusterSize of 15 (setting the smallest cluster size to 15). We calculated module eigengenes using moduleEigengenes of the adjacency matrix then calculated the dissimilarity of the module eigengenes with cor of the module eigengenes subtracted from one and clustered the module eigengenes with hclust. We then chose a dissimilarity of 35% as the cut off for merging modules; however, no modules had dissimilarity scores with each other below 40%. Since the module eigengenes are calculated for each cell in the dataset, we used these values to calculate the t-statistic for each cell cluster (considered each sample) in the module (all of the module eigengene scores were used as the population).

We next tested the hypothesis that certain cell clusters were strongly associated with specific gene modules through a two-way permutation test (1000 permutations) and corrected p-values to control the False Discovery Rate (FDR). Additionally, we tested whether specific cell clusters drove underlying gene network structure by measuring network connectivity. The network connectivity was first calculated using the entire dataset, then each cell cluster was progressively dropped from the dataset and the connectivity was remeasured. This resulted in changes in connectivity scores for each module-cluster pair. To assess whether these changes were significant, we completed 1000 permutations whereby cells were randomly dropped (the number of cells dropped was equal to the cell cluster size of the focal cluster), connectivity measured, and the change in connectivity recorded. The p-value is the number of instances where the change in connectivity is greater in the permutations than in the focal cell cluster run. P-values were corrected to control the FDR.

Using the module gene lists, we identified the number of genes from each module that were found in each KEGG pathway. Since the KEGG modules do not have p-values associated with the genes, we could not complete a Wilcoxon enrichment test. We instead removed any pathways where there were fewer than three genes present for any pathway and note that there is no statistical test run on these KEGG results. We visualized the networks using Cytoscape (v3.10.0).

In situ Hybridization

For genes chosen for further follow up analysis, we completed *in situ* hybridizations in Gulf pipefish (*Syngnathus scovelli*) or bay pipefish (*Syngnathus leptorhynchus*). Primer sequences were designed using NCBI Primer Blast with the Gulf pipefish genome assembly and

synthesized using Gulf pipefish embryonic cDNA. The *sfrp1a* probe was synthesized using the PCR based probe preparation protocol described in the Single Cell Identification section. All other probes (*bmp4*, *dlx2a*, *dlx3b*, *fgf22*, *lhx6a*, *pitx2*, and *scpp1*) were prepared using TOPO cloning. Probe primer sequences as well as the species used for the *in situ* experiments are described in supplementary file In_Situ_Probe_and_Samples_Used.xlsx. Pregnant male bay pipefish were caught as described above. For Gulf pipefish embryos, we allowed the fish to mate in our facility then harvested embryos once they reached the appropriate stages. Fish were reared in 25 degrees C water with 25-28 ppt salinity. Embryos and larvae were fixed in 4% PFA, dehydrated through a series of PBT/MeOH washes, and stored in MeOH at -20C. 5-12 embryos were used for each probe. We completed *in situ* hybridizations in keeping with Thisse & Thisse, 2007, leaving the embryos in stain until background was observed. After imaging, we used the levels tool in Adobe Photoshop (v23.4.2) to white balance the pictures.

Bone and Cartilage Staining

We used alcian and alizarin stains to mark cartilage and bones. Specifically, we assayed cartilage and bone development in siblings of the scRNAseq samples. These embryos were fixed in 4% PFA then stored at -20C in MeOH. We followed the protocol from Walker and Kimmel (2007) with minor alterations. We stored samples in 50% glycerol/0.1% KOH at 4C and imaged them in 100% glycerol. After imaging, we white balanced the photographs using Photoshop (v23.4.2) levels tool.

Gene Cluster Analysis

To examine close syngnathid outgroups, we downloaded the 2023 mandarin dragonet genome assembly from NCBI (GenBank assembly accession: GCA_027744825.1) and used an unpublished highly contiguous cornetfish genome assembly. Since these genome assemblies were unannotated, we manually identified *scpp* genes. We searched for *scpp* genes using BLASTN with medaka and additional fish sequences as the query. We also searched for these genes with mVISTA plots across conserved gene synteny regions (LAGAN alignment using translated anchoring) with medaka as the focal species. To identify *scpp* genes in additional species, we gathered cluster information from NCBI and ensembl. Additionally, we used mVISTA plots to further search for unannotated *scpp* genes.

Results

Valuable scRNAseq atlas for studying syngnathid development

We produced the first developmental scRNAseq atlas for a syngnathid from two samples comprising 20 similarly staged embryos from pregnant, wild-caught Gulf pipefish (*Syngnathus scovelli*) males. The samples represent a late organogenesis developmental stage (Figure 1B-H). These embryos had a primarily cartilaginous skeleton with minimal mineralization, including jaw cartilages that were at the onset of mineralization and ethmoid elongation. The embryos also possessed cartilaginous dorsal fin pterygiophores but had no signs of dermal armor mineralization. This stage is referred to as “frontal jaws” in syngnathid literature (Sommer et al., 2012).

The atlas included 35,785 cells (19,892 and 15,893 cells from each sample; S3.1:S3.2), which formed 38 cell clusters (Figure 3.2A, Supplementary File 2). We classified cells into 4 different broad tissue types – epithelial, connective, neural, and muscle – using Seurat identified marker genes and published model organism resources (S3.3:S3.57). We next used Seurat identified marker genes to pinpoint single marker genes that were most unique to each cluster (Figure S3.58, Supplementary File 3). We completed *in situ* hybridization using Gulf and bay pipefish embryos for cell clusters for which examining gene expression would help hone and

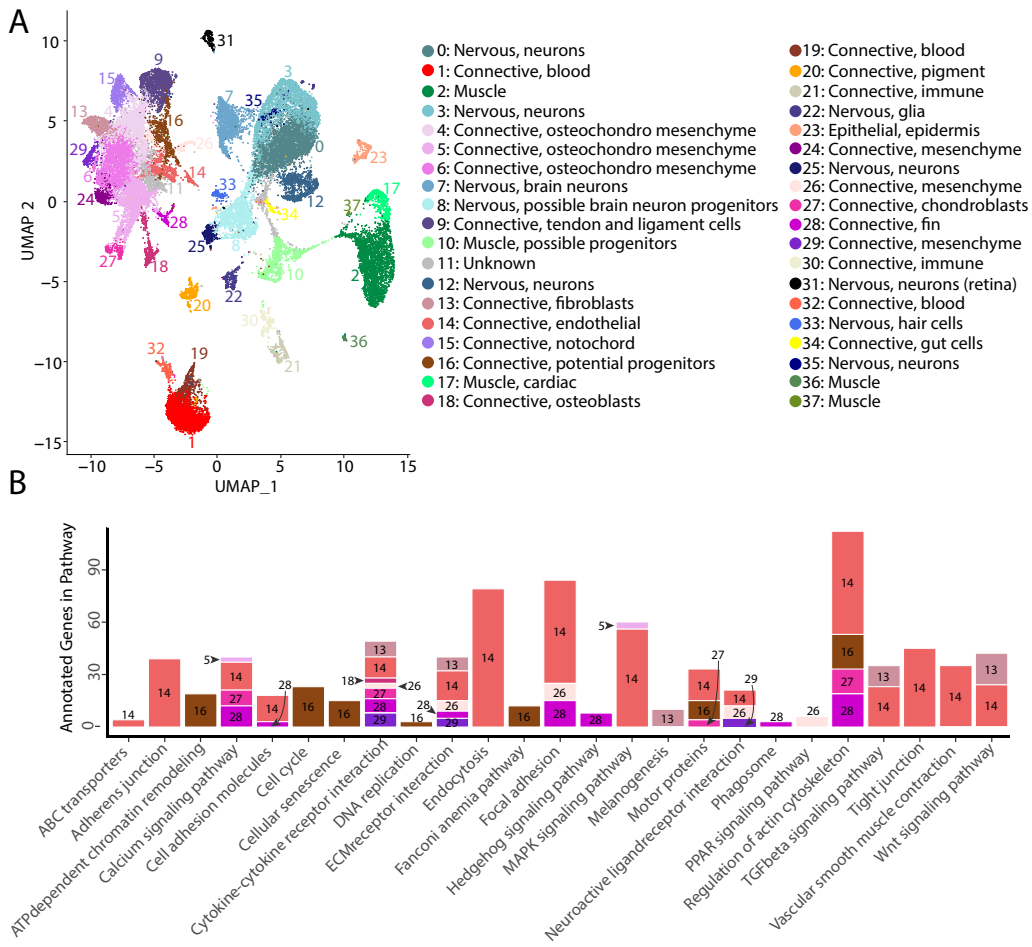


Figure 3.2: *Gulf pipefish single cell atlas contains cells from the entire embryo and identifies genetic pathways active in different cell types.* The UMAP plot (panel A) shows all of the cell clusters and their identities reduced to the first two UMAP dimensions. The graph in panel B displays results of the KEGG pathway analysis in cell clusters identified as connective tissue (excluding blood, pigment, digestive, and immune cells). The number of Seurat identified marker genes for each cluster that was a part of each pathway is displayed on the y axis. Bars are colored and labeled by cell cluster.

validate cluster annotations (Figure S3.2, S3.59:S3.71).

In total, our atlas contained 13,027 connective tissue cells (excluding cells from blood, immune, and the digestive system) from 14 clusters, 10,112 nervous system cells from ten clusters, 4,363 muscle cells from five clusters, 4,133 blood cells from three clusters, 650 immune cells from two clusters, 432 pigment cells from one cluster, 370 epidermal cells from one cluster, and 137 gut cells from one cluster. Within the connective tissue cell types, we also identified cartilage (302 cells), developing bone (442 cells), fins (253 cells), and notochord (693 cells). The number of recovered cells per identity may not necessarily represent organismal cellular proportions because of potential variability in dissociation success for different cell types (Denisenko et al., 2020; Venema et al., 2022).

Discovery of cell cluster function and state using KEGG analysis

To affirm identities and discover potential properties of each cluster, we completed a KEGG pathway analysis for each cluster using Seurat's marker genes (Figure S3.4, Figure 3.2B). For eight of the clusters (1, 4, 6, 9, 11, 15, 19, and 24), we did not find any significantly enriched pathways, possibly due to similar gene expression profiles across cell types that reduced the number of identified markers. However, we found one or more significantly enriched pathways for the other 29 cell clusters. We observed enriched pathway terms that supported cluster annotations. For example, phototransduction in the retina cluster, melanogenesis in the pigment cluster, cardiac muscle contraction in muscle clusters, and neuroactive ligand receptor interaction in neuronal clusters.

The inferred KEGG pathways demonstrated some commonalities across the different tissue types, including in signaling pathways and cell states. Notably, our identified KEGG terms delineated progenitor and differentiated cell clusters. Based on their KEGG terms, we classified clusters 8, 10, and 16 as possible neural, muscle, and connective tissue progenitor cells, respectively. We also detected expression of *pax3a* and *pax3b*, muscle primordia markers, in cluster 10, supporting this annotation. These clusters had enriched KEGG terms associated with cell division ('cell cycle', 'DNA replication', 'nucleotide excision repair', and 'homologous recombination'), and lacked enrichment for KEGG pathways present with differentiated cell types of their lineage. Specifically, cluster 8 lacked the neural KEGG term 'neuroactive ligand receptor

interaction', cluster 10 lacked muscle KEGG terms 'adrenergic signaling in cardiomyocytes', 'calcium signaling pathways', and 'cardiac muscle contraction', and cluster 16 lacked connective

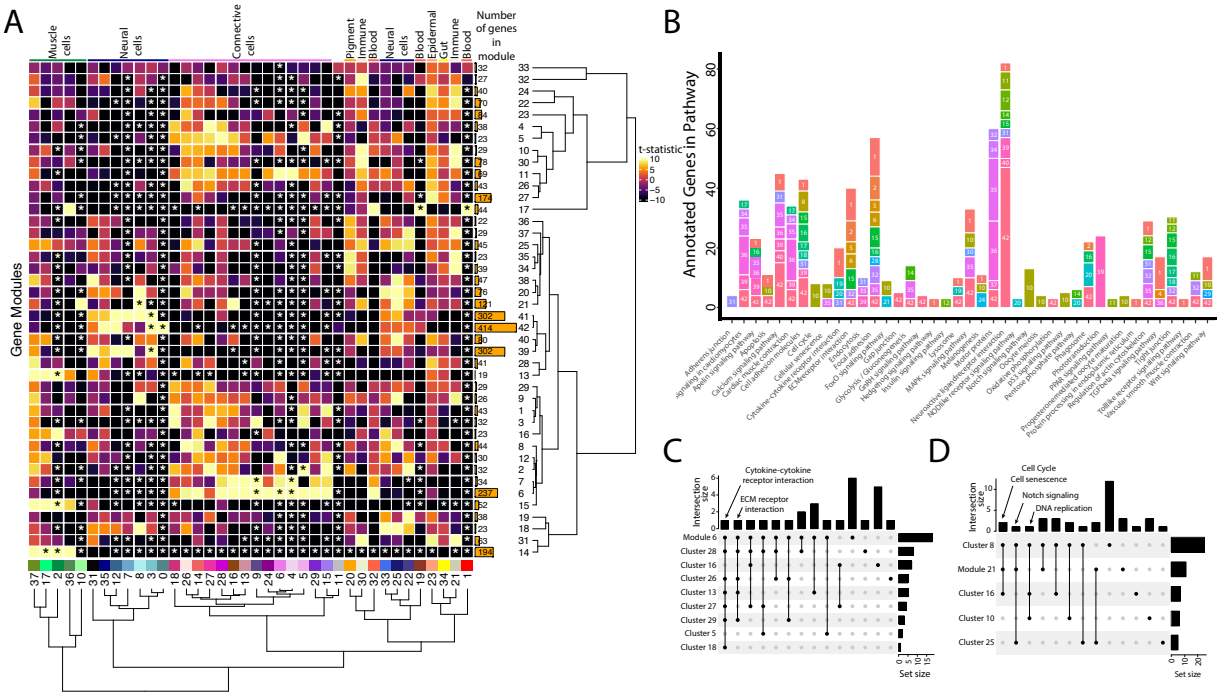


Figure 2.3: *Weighted Gene Network Analysis (WGCNA) identifies gene modules that define and unite cell clusters.* A) The strength of association between the gene modules and cell clusters is shown in panel A with dendrogram clustering illustrating the distance between modules and cell clusters. Gene modules are represented by rows and cell clusters by columns. The modules and clusters are clustered using the Pearson distance method. The number of genes in each gene module are shown in the right hand bar plots. Cell clusters are colored based on their identity. The asterisks indicate the module-cluster relationships that have a p-value less than .05 from a two-sided permutation test after correction for multiple tests (FDR). The heatmap is colored by t-statistics in a range of -10 to 10, with highly positive values in yellow and highly negative values in black. B) The identified gene modules possess genes from KEGG pathways. The bars are labeled with the gene module and the size of each bar corresponds to the number of genes from the KEGG pathway in the module. Since WGCNA modules do not have p-values, only KEGG pathways with more than two genes included in the gene module are shown on the plot. C) Identified gene modules contain similar KEGG pathways as the cell clusters that correlated with them. These relationships are shown in Upset plots where each row is a cell cluster or gene module, each column represents KEGG pathways shared by the modules and clusters (shared condition is shown filled in black dots connected by lines), the interaction size is the number of pathways in common between the set of modules and clusters, and the set size is the number of pathways that are enriched in each cluster and module. Panel C1 highlights that 'cytokine-cytokine receptor interaction' and 'ECM receptor interaction' are present in module 6 as well as 6 and 4 connective cell clusters respectively. Panel C2 shows that 'cell cycle' and 'senescence' are present in module 21 as well as clusters 8 and 16, 'Notch signaling' genes are present in clusters 8 and 25 as well as module 21, and 'DNA replication' is present in clusters 8, 10, and 16.

tissue term ‘ECM receptor interaction’. To complement these findings, we completed a cell differentiation analysis using CytoTRACE for neural, muscle, and connective clusters (Figure S3.73). Clusters 8, 10, and 16 had the lowest scores in each respective comparison, which indicated undifferentiated cell states. Thus, it is likely that clusters 8, 10, and 16 represented undifferentiated cells within the major lineages of neural, muscle, and connective cells.

Commonalities of cell clusters, unique networks, and elusive cell types identified in network analysis

We built gene networks/modules from 3,000 variable genes using Weighted Gene Network Correlation Analysis (WGCNA; Langfelder & Horvath, 2008). This produced 43 gene modules in total (Supplementary File 5,6), assessed for each cluster-module pair for their strength of association (Figure 3.3A, Supplementary File 7,8) and every module’s dependence on each cluster for their network connectivity (Figure S3.5, Supplementary File 9,10). Using the genes from each network, we completed a KEGG pathway analysis to identify whether gene modules indicated specific cellular pathways or states (Figure 3.3B). We initially explored whether these network-cluster associations could reveal commonalities between cell clusters or identify whether particular clusters contained multiple cell identities.

First, we asked whether gene modules that associate with three or more cell clusters signify commonalities between clusters that have similar cell types. We identified seven gene modules (6, 7, 21, 14, 17, 41, and 42) that each associate with three or more cell clusters. These gene modules do connect clusters of similar identities or cell states. For example, Modules 6 and 7 associate with connective tissue cells. Module 6, the larger of the two modules, contains numerous other KEGG pathways found in most connective cell clusters, such as ‘cytokine-cytokine receptor interaction’ (Figure 3.3C). Interestingly, Module 21, associated with clusters 8, 10, 16, and 25, contains genes from KEGG pathways related to the ‘cell cycle’ and ‘cellular senescence’, supporting the results found in our cluster based KEGG pathway analysis (Figure 3.3D). Module 21 also contained ‘Notch signaling’ genes, possibly due to similar correlations with cell cycle genes; however, these were only expressed in the neural cell clusters (8 and 25).

Where one cell cluster is associated with multiple gene networks, we wondered if multiple cell identities existed within the cell cluster. We explored this possibility by examining the pigment

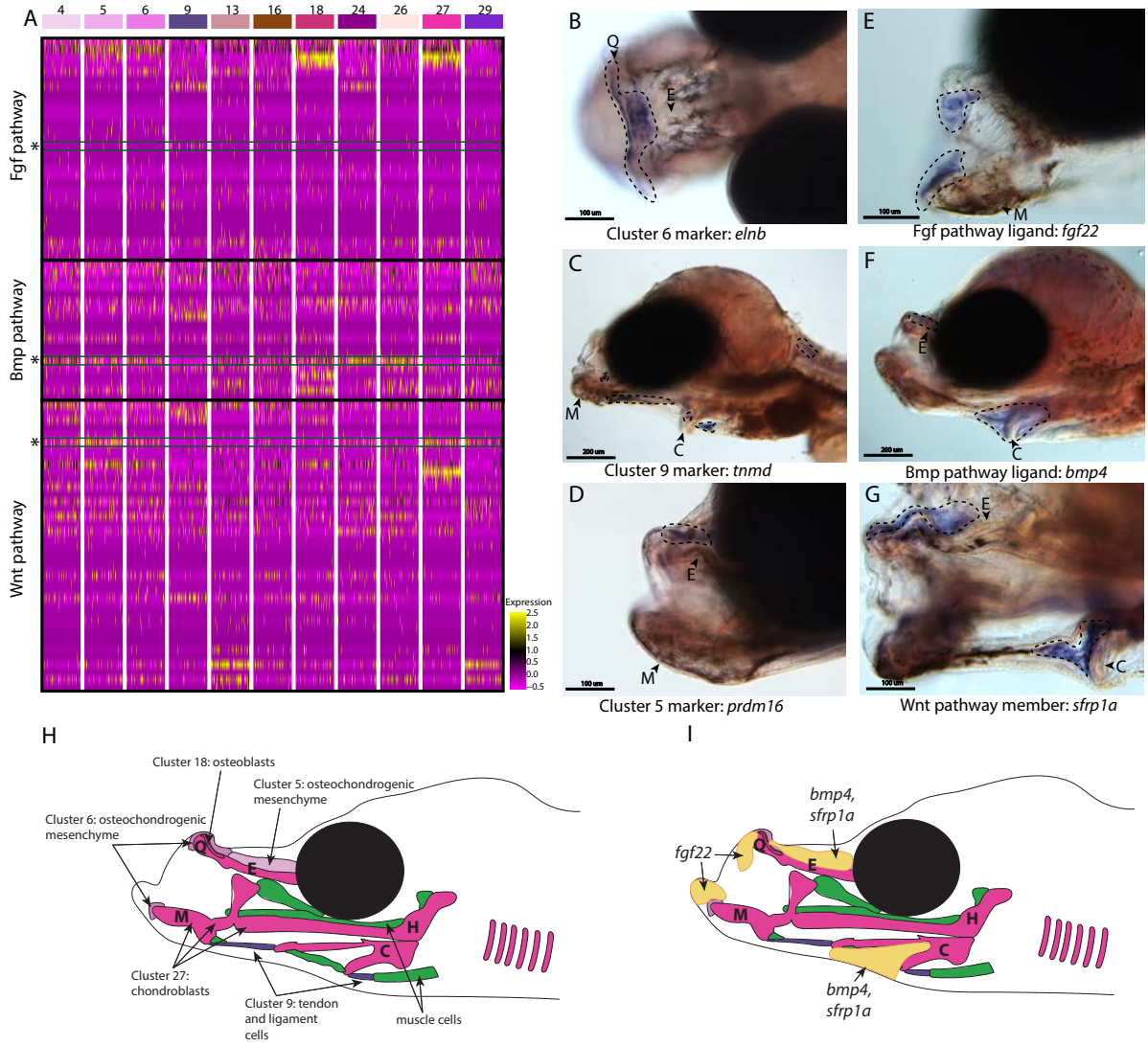


Figure 3.4: *Conserved cell types and gene pathways build unique faces of syngnathids.* Three main conserved signaling pathways are enriched in connective cell types, MAPK signaling (including Fgf signaling), TGF-beta signaling (including Bmp signaling), and Wnt signaling. Receptors and ligands expression patterns are shown in panel A heatmap from all cell types with cells present in the head. This heatmap features 100 cells down sampled from each cluster and illustrates that many genes from these families are expressed in these cells. Yellow lines indicate high expression of a gene, while hot pink lines indicate no expression. The pathways are boxed in black. Rows representing *fgf22, bmp4*, and *sfrp1a* expression are marked with an asterisk and green box for each respective section of signaling (Fgf, Bmp, and Wnt). Panels B, C, and D are *in situ* hybridizations of three marker genes, *elnb*, *tnmd,* and *prdm16*. The genes *prdm16* and *elnb* mark osteochondrogenic mesenchyme and *tnmd* marks tendons and ligaments. Panels E, F, and G, show expression patterns of three pathway representatives (*fgf22, bmp4*, and *sfrp1a*). All three genes are expressed in the face: *fgf22* at the tip of the mandible and *bmp4* and *sfrp1a* above the ethmoid and near the ceratohyal. Staining is circled with dashed lines. The Meckel's cartilage (M), (mes)Ethmoid cartilage (E), Quadrate (Q), and Ceratohyal (C) are labeled. Panel C is a dorsal view. Panels B, C, E, F, G, H, I are in lateral view. *In situs* of *fgf22* were completed using 10dpf Gulf pipefish. *In situs* of *bmp4*, *sfrp1a*, *tnmd*, *elnb*, and *prdm16* were completed using wild caught bay pipefish at the onset craniofacial elongation. Panels H and I are summary illustrations of our findings, panel H shows where cells from various clusters were present in the developing head and panel I illustrates where *bmp4*, *sfrp1a*, and *fgf22* were expressed.

and 38) are expressed in distinct subgroups of cells within the cluster (S3.75). We found cases of more than one genetic network (modules 35 and 38) expressed in the same cells. Modules 35 and 38 contain conserved pigment genes, *pmela*, *mlana*, and *dct* in module 35 and *tryp1b* and *pmel* in module 38, that mark melanocytes (Figure S3.6; Du et al., 2003; Johnson et al., 2011; Lamason et al., 2005; B. Thisse & Thisse, 2004). However, we also found non-overlapping expression of networks, notably modules 36 and 37. We inferred that module 36 is associated with xanthophores and xanthoblasts due to the inclusion of *plin6* and *scarb1*, genes involved with lipid binding and activity in xanthophores (Ahi et al., 2020). On the other hand, module 37 likely represents iridophores and iridoblasts because it contains the gene *pnp4a,* which is involved in purine-nucleoside phosphorylase activity in iridoblasts (Kimura et al., 2017).

Conserved signaling pathways are active during syngnathid craniofacial development

The specialized pipefish feeding apparatus is composed of an elongate, tubular snout, toothless mandible and pharyngeal jaws, large tendons, and associated muscles. Therefore, numerous cell types contribute to their distinct faces: cartilage, bone, tendon, muscle, and connective tissues as well as their progenitors. We sought to identify markers of these cell types and signaling pathways active in them.

We found marker genes uniquely expressed in the face, genes that mark cell types important to craniofacial development, and markers with potentially relevant functions for craniofacial development using *in situ* hybridizations of cell cluster marker genes. For instance, we observed the marker for osteochondro-mesenchymal cells (cluster #6), *elnb*, specifically expressed at the intersection between the ethmoid plate and palatoquadrate as well as on the Meckel's cartilage (Figure 4B). Although *elnb* expression is observed in the zebrafish cranial skeleton, it is primarily studied for its proposed role in teleost heart evolution (Miao et al., 2007; Moriyama et al., 2016).

Other genes identified here as cell markers were not uniquely craniofacial but provide insights into the cell types that build the face. For example, *tnmd* expression marked tendons and ligaments (cluster #9) throughout the face and body (Figure 3.4C; Figure S3.65). Our finding is consistent with *tnmd's* role in tenocyte development in model systems, namely zebrafish and mouse (J. W. Chen & Galloway, 2014; Docheva et al., 2005). Identifying tenocyte cells is particularly relevant in syngnathid fishes where tendons are enlarged and store elastic energy

necessary for their specialized feeding (Van Wassenbergh et al., 2007).

The last category of markers contained genes with regulatory roles important for craniofacial development such as *prdm16,* the marker for osteochondro-mesenchymal cells (cluster #5). This gene mediates the methylation of histones and regulates gene expression, which is key for promoting craniofacial chondrocyte differentiation (Ding et al., 2013; Kaneda-Nakashima et al., 2022; Shull et al., 2020), We found *prdm16* expressed in mesenchymal cells directly above the ethmoid plate and in fins (Figure 3.4D; Figure S3.61). Comparison of pipefish with a related, short snouted fish (stickleback) identified conserved fin expression, but discrepancy in craniofacial staining (S3.76). Stickleback expressed *prdm16* in the hindbrain, gill arches, and lower jaw, consistent with published zebrafish data, but staining was restricted to the ethmoid region in pipefish (Ding et al., 2013). Our observations are consistent with the literature, except *prdm16* expression has not been observed above the ethmoid in zebrafish.

We next examined signaling pathways active in craniofacial development. Our KEGG pathway analysis revealed that MAPK, Wnt, and TGF-beta signaling pathways were significantly enriched in one or more craniofacial contributing cell clusters (module #6; Figure 3.2B, Figure 3.3C, Figure 3.4A). In addition to these KEGG findings, expression of ligands and receptors encompassed by these pathways (Fgf, a MAPK pathway, Wnt, and Bmp, a TGF-beta pathway) was observed in differentiated and undifferentiated clusters (Figure 3.4A). We chose three genes, one from each major pathway, for *in situ* hybridizations: *fgf22* (a fgf ligand present in the actively dividing cells module), *bmp4* (a bmp ligand present in the largest connective tissue module), and *sfrp1a* (a Wnt pathway enabler present in a cartilage gene module). Although mouse and zebrafish studies only identified *fgf22* expression in the nervous system (Miyake & Itoh, 2013; Umemori et al., 2004), we found *fgf22* is expressed at the tip of the palatoquadrate and Meckel's cartilage, the gill arches, fins, and brain in pipefish (figure 3.4E; S3.77). These same patterns were observed in stickleback, suggesting a surprising co-option of fin and craniofacial expression in percomorph fish.

We observed *bmp4* and *sfrp1a* expression above the ethmoid plate and along the ceratohyal in pipefish (figure 3.4F-G). The gene *bmp4* has a conserved role in craniofacial development, particularly important at later stages for driving chondrocyte differentiation (Wang et al., 2024; Zhou et al., 2013). However, the specificity of *bmp4* expression to mesenchyme around the ceratohyal and ethmoid was not observed in stickleback fishes which had broad craniofacial

expression (including the jaws, tooth germs, and gill arches; S3.78). Interestingly, *sfrp1a* expression has not been observed in the palate of mice or ethmoid region of zebrafish, but *sfrp1a* craniofacial expression was identified in stickleback (lower jaw and gill arches; S3.79) and in other fishes (Ahi et al., 2014; Schneider, Gunter, et al., 2023).

Gulf pipefish retain tooth development genes but likely lack onset of tooth development

Previous papers have identified possible candidate genes for the loss of teeth in syngnathid fishes including those from genes that initiate tooth bud formation (*fgf4, eve1*), regulate tooth morphogenesis (*fgf3, fgf4*), and synthesize tooth minerals (*scpp4, scpp7, scpp9, odam*, and *scpp5*; Lin et al., 2016; Qu et al., 2021; Small et al., 2016; Small et al., 2022; Zhang et al., 2020). However, it is unknown whether syngnathid tooth development initiates then halts or whether it never begins. We searched for signs of early tooth primordia within our atlas to ask whether tooth development might initiate in syngnathids. Additionally, we examined whether genes present in mature teeth are still expressed in syngnathids and what types of cells express them.

No tooth primordium cell cluster was identified during our thorough cluster annotation process. We therefore searched for tooth primordia by examining expression of specific odontogenesis marker genes (*aldh1a2*, *bmp4*, *dlx2a*, *dlx3b*, *lef1*, *lhx6a*, *lhx8*, *msx1a*, *msx2*, and *wnt10a*; Figure 3.5A). We observed several primordium genes expressed in our atlas. However, there was no cluster with every marker gene expressed in over 10% of cells. Several markers (*dlx3b*, *lef1*, *msx1a*, *msx2*, and *wnt10a*) were expressed in cluster #28, a fin cluster distinguished by *hoxa13a* and *hoxa13b* expression. Since we previously noted that cluster #16 seems to be a primordial connective tissue cluster, we wondered if it could contain tooth primordial cells. In this cluster, we found the following percentage of expression of each gene in the cells: *bmp4* in 46%, *aldh1a2* in 29%, *lef1* in 24%, *dlx2a* in 18%, *dlx3b* in 16.4%, *eda* in 12.7%, *msx1a* in 10.6%, *pitx2* in 10%, *lhx6a* in 8%, *wnt10a* in 4.6%, *msx2* in 2.65%, *lhx8* in 1.8%, and *shha* in 1%.

We suggest that, given the low expression of most tooth marker genes, cluster #16 is unlikely to contain tooth primordial cells. To test this assertion, we examined spatial gene expression using *in situ* hybridization of *bmp4*, *pitx2*, *lhx6a*, *dlx2a*, and *dlx3b* in pipefish to ask whether the two definitive cell types are present, namely the dental epithelium (marked by *pitx2*

and *bmp4*) and dental mesenchyme (distinguished by *dlx2a*, *dlx3b*, and *lhx6a*; Figure 5B,C,D,E,F; Gibert et al., 2019; A. Tucker & Sharpe, 2004). Tooth-specific expression of *dlx2a* is observed solely in the dental mesenchyme in zebrafish and mice, however, it is also expressed in the dental epithelium in medaka (Stock et al., 2006). For this study, we labelled it as a mesenchyme marker, though it could be expressed in both dental tissues in a syngnathid outgroup. We found expression of all genes except *dlx2a* in the developing jaws. However, *bmp4*, *pitx2*, *lhx6a*, and *dlx3b* were expressed throughout the face rather than in the punctate pattern observed in tooth primordia development.

We next investigated *scpp* genes (*enam*, *scpp1*, and *spp1*), which are expressed in late tooth development and tooth maintenance in other vertebrates (Figure 3.5A). These genes also have

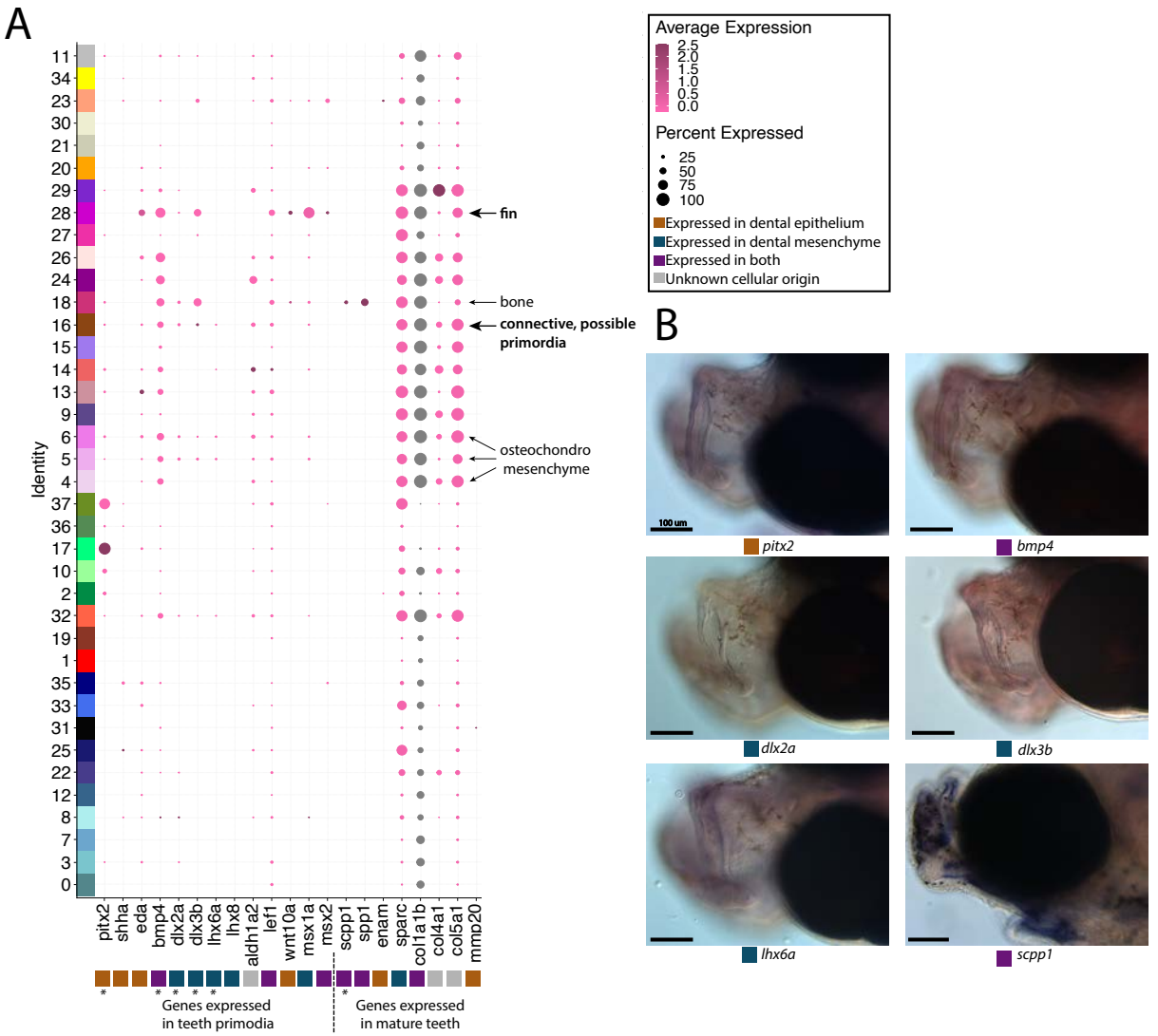


Figure 3.5: *Pipefish do not possess identifiable tooth primordium cells, but continue to express tooth development genes in other contexts*. Panel A presents a dotplot of genes classified by the tissue layer in which they are reported to be expressed during tooth development in other vertebrate (Tucker and Sharpe 2004; Gilbert et al. 2019; Kawasaki 2009). The x-axis contains the assayed genes, with asterisks under the genes that were also examined with *in situ* hybridizations. The y-axis contains all of the cell clusters. The size of the dot represents of the percentage of cells from the cluster that express the gene. The color of the dots is average expression of the gene in the cluster (darker purples represent higher expression). Panel B includes *in situ* hybridizations of selected tooth primordia markers (*bmp4, pitx2, lhx6a, dlx2a*, and *dlx3b*) and mature tooth marker (*scpp1*). *In situs* of *bmp4, pitx2, lhx6a, dlx2a,* and *dlx3b* were completed with wild caught bay pipefish that had begun craniofacial elongation. *In situs* of *scpp1* were completed using 9dpf Gulf pipefish. The scale bars for all images represents 100 µm.

some expression outside of teeth such as in dental bone (*scpp1*, *spp1;* Kawasaki, 2009) and fins (*enam;* Jain et al., 2007). We identified *spp1* and *scpp1* expression in 54.5% and 24% of bone cells, respectfully, and sparse *spp1* expression (<3%) in other connective tissue cell types. We found minimal expression of *enam* (in 11.4%) of epidermal cells and in less than 8% of muscle cells. In whole mount *in situ* hybridization, we found that *scpp1* is expressed in all developing pipefish bones, both endochondral and dermal (Figure 3.5G). Since *scpp* gene losses observed in syngnathids have been hypothesized to be responsible for their tooth loss (Lin et al., 2016a; Qu et al., 2021; Zhang et al., 2020a), we explored *scpp* gene family content in the close toothed relative to syngnathids, the blue spotted cornetfish, and we found several gene losses (*scpp4*, *scpp7*, and *scpp9*) and a likely pseudogene (*scpp5*; Figure S80; Hughes et al., 2018; Stiller et al., 2022).

Exploration of additional tooth maturation genes (*col1a1b, col4a1, col5a1, sparc*, and *mmp20b*) similarly found that these genes were expressed in non-tooth derivatives, including connective tissue, smooth muscle, and neural cells.

Tooth and skeletal genes are expressed during dermal armor development

Syngnathid dermal armor is mineralized dermal bone underneath the skin (Figure 3.6D, G, H). It is unknown when dermal armor primordia initiate and how they are patterned. Spatial expression analysis of *pitx2* and *dlx3b* in search of tooth primordia instead revealed expression of these genes in possible dermal armor primordia (Figure 3.6A, B, E, F). We found *pitx2* staining localized dorsally to striated muscle underneath developing dermal armor. The gene *dlx3b* is expressed in a repeating pattern along the body in the epidermal and dermal tissues. Both staining patterns were in discrete regions of the muscle and epidermal layers rather than being continuously expressed across the tissues. We did not observe the expression of other tooth primordium genes (*bmp4, lhx6a*, and *dlx2a*) in this region.

Because the embryos from our atlas had not begun dermal armor mineralization, the atlas cannot be directly used for discovery of genes active in dermal armor. However, the atlas contains osteoblasts from craniofacial bones which we used to create osteoblast specific gene networks. We therefore asked whether these osteoblast genes were expressed in mineralizing dermal armor at later stages. *In situ* hybridization expression analysis revealed that *scpp1*, an

osteoblast and tooth mineralization gene, and *ifitm5*, an osteoblast gene, were expressed in the dermal armor at the onset of mineralization (Figure 3.6C; Figure S68).

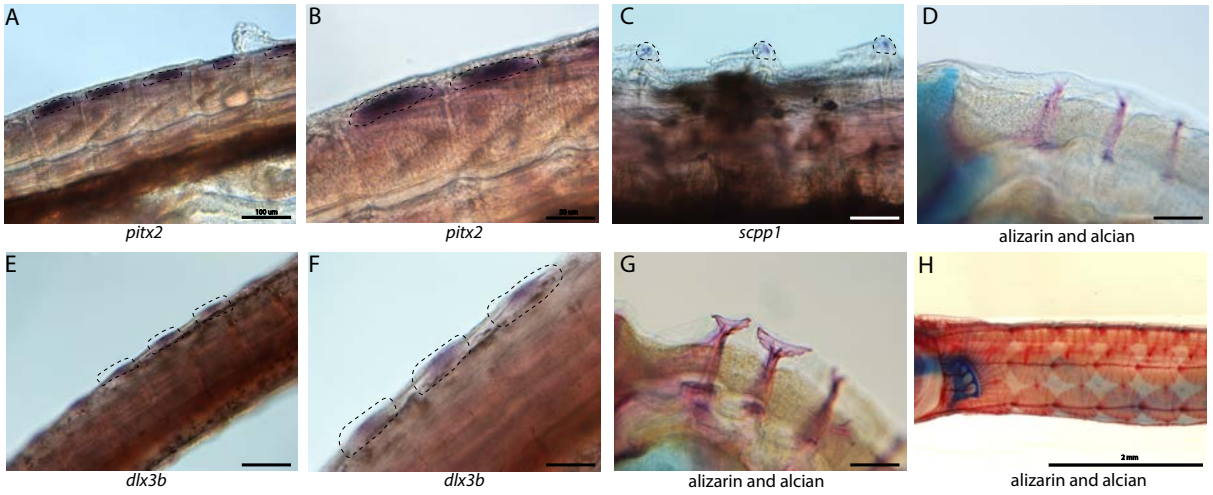


Figure 3.6: *Tooth and bone development genes expressed during exoskeleton development.* We discovered *pitx2* (A and B) and *dlx3b* (E and F) expression during possible emergence of exoskeletal primordium in wild caught bay pipefish. The embryos used for these *in situs* were the same stage as those from Figure 2.4, at the beginning of craniofacial elongation. Figures A and E have 100 µm scale bars, figures B and F are taken from the same fish and have 50 µm scale bars. We further found *scpp1* is expressed at the mineralization front of the exoskeleton in 12dpf Gulf pipefish (C) and has a 50 µm scale bar. Alizarin and alcian stained pipefish are shown in panels D (12dpf), G (1dpf), and H (adult Gulf pipefish) to illustrate how the exoskeleton forms. Panels D and G have 100 µm scale bars.

Epithelial expression of immune and nutrient processing genes may facilitate embryo-paternal interactions in the brood pouch

Within the brood pouch, embryos could interact with male placenta-like tissues, the male brood pouch epithelium, and/or the pouch microbiome. Once the thin chorion is shed, the embryos' epidermis is directly exposed to the pouch environment. We therefore asked if the embryonic epidermal cells expressed nutrient acquisition and/or immune genes that would indicate active transfer of nutrients and immune response.

Within our larger KEGG analysis, we asked whether nutrient absorption and immune KEGG terms were among the enriched pathways for the epidermal cells. We identified 106 enriched genes in the endocytosis pathway (p-value=.036; Figure 3.7A). Four metabolism pathways (galactose, glutathione, sphingolipid, and starch and sucrose) are also enriched. No

immune related KEGG terms were enriched in the epidermis. For comparison, we investigated whether these KEGG terms are also enriched in epidermal cells of non-brooding fishes, we completed a KEGG pathway analysis on a comparably staged zebrafish single cell RNA sequencing atlas (3dpf; Lange et al., 2023). The 3dpf zebrafish epidermal cells did not have a significant enrichment of the endocytosis pathway (23 up-regulated genes, p-value=.99) or any metabolism pathway. However, there are 11 endocytosis genes up regulated in both zebrafish and

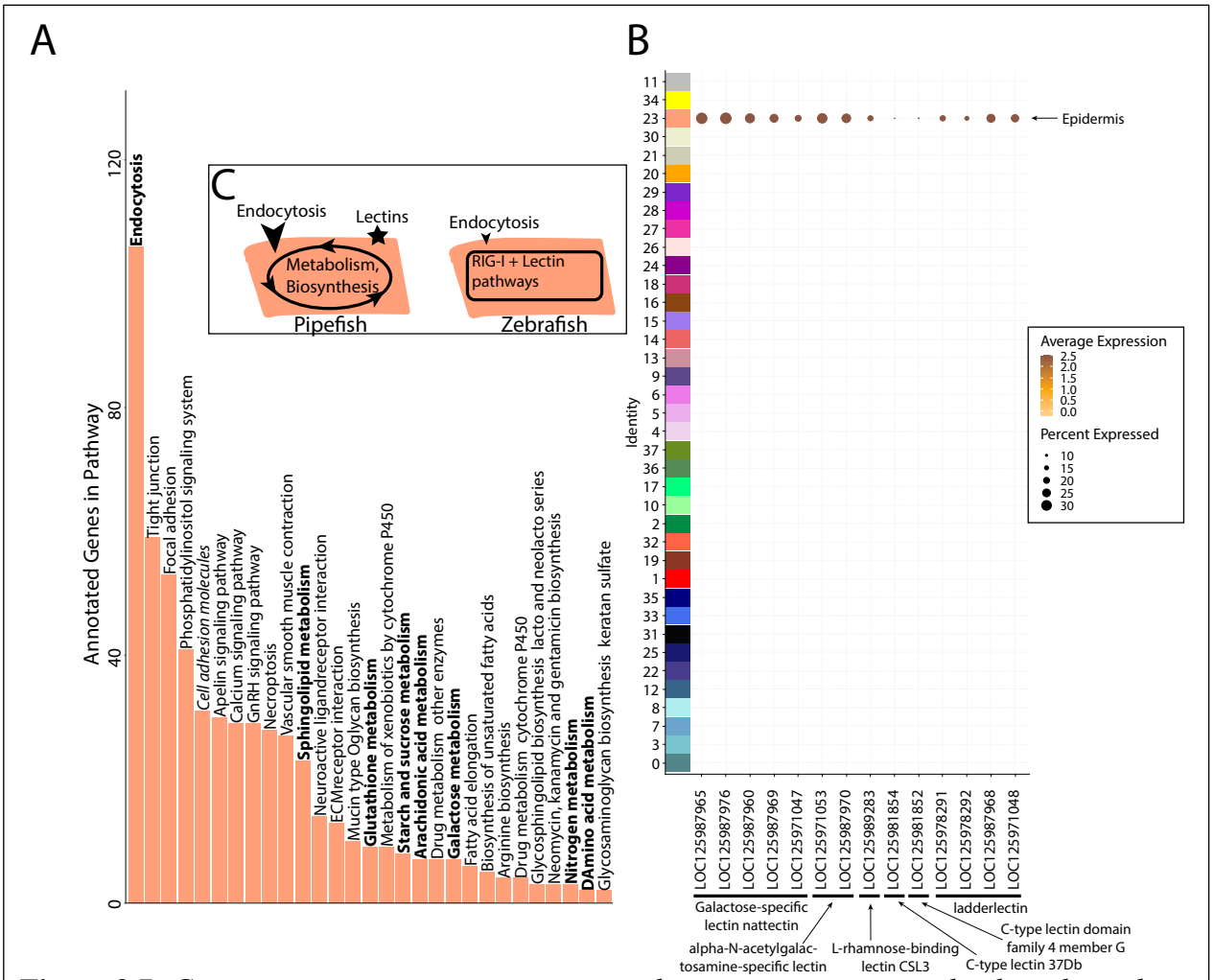


Figure 3.7: *Gene expression signatures suggest embryonic interactions within brood pouch environment.* Epidermal cells (panel A), with pathways that suggest increased endocytosis and metabolism pathways are in bold text. Pathways upregulated in 3dpf zebrafish epidermal cells are in italics. Pipefish epidermal cells also express 15 lectin genes not found in other cell types (panel B). We suggest an epidermal cell model (panel C), in which we predict pipefish have an enrichment of nutrient processing genes and lectins in comparison with zebrafish cells.

pipefish epidermal cells, suggesting conserved expression of these genes in pipefish.

We next examined epidermal gene networks for the presence of nutrient absorption and immune genes. We found that the largest epidermal gene network (#16) contained a striking enrichment of C type lectin genes, carbohydrate binding proteins that possess antimicrobial properties (Figure 3.7B). This network contained 14 total lectin genes expressed in the epidermal cells: five galactose-specific lectin nattectin, 2 alpha-N-acetylgalactosamine-specific lectins, one L-rhamnose-binding lectins, four ladderlectin, one C-type lectin 37Dd-like, and one C-type lectin domain family 4 member G-like. Through examining lectin gene expression in the entire dataset, we found that these genes were specific to epidermal cells. Interestingly, previous literature has identified an upregulation of C-type lectins in brood pouch tissues throughout different stages of syngnathid pregnancy (Roth et al., 2020; Small et al., 2013; Whittington et al., 2015).

Unlike pipefish, we did not find any expression of C-type lectin genes in zebrafish epidermal cells (Figure S3.81). However, it is possible there are unannotated C-type lectin genes in zebrafish which would not have been present in the atlas. Interestingly, the 'C-type lectin receptor signaling pathway' was significantly enriched in zebrafish epidermal cells (13 genes, p-value=.04) but not in pipefish epidermal cells (34 genes, p-value=.4). Although these results seem paradoxical, the KEGG term 'C-type lectin signaling pathway' does not include any of the C-type lectin genes themselves. Overall, these data suggest that the expression of C-type lectin genes in the pipefish embryonic epidermis is potentially unique and warrants further investigation.

Discussion

Our study examines the development of syngnathids, with a particular focus on novel and adaptive characters, using single cell RNA sequencing of Gulf pipefish embryos coupled with *in situ* experiments of gene expression. Our single cell atlas represents early craniofacial skeleton development in Gulf pipefish, at a stage when the cartilages of the head skeleton were formed but the face has not elongated. We used the atlas to explore craniofacial and dermal armor development and to investigate potential interactions between the embryos and the brood pouch environment. Our dataset is both an opportunity to explore the developmental genetic underpinnings of syngnathid innovations, and a resource for teleost researchers for future studies in this fascinating lineage.

Our atlas represents a novel resource for Evo-Devo research

Developmental single cell atlases have elucidated cell identities and genetic pathways active in zebrafish, mice, and chick (Farnsworth et al., 2020; Farrell et al., 2018; Feregrino et al., 2019; Morrison et al., 2017; Soldatov et al., 2019; D. E. Wagner et al., 2018; Williams et al., 2019). In less traditional models, the majority of scRNAseq atlases are produced from adult tissues, allowing investigations into cell types, population differences, and genetic networks (e.g. Chari et al., 2021; Fuess & Bolnick, 2023; Hain et al., 2022; Hong et al., 2023; Koiwai et al., 2023; Parker et al., 2022; Potts et al., 2022; Royan et al., 2021; Songco-Casey et al., 2022; Vonk et al., 2023; Woych et al., 2022). Creation of developmental scRNAseq atlases in non-model organisms is just beginning to accelerate, but many emerging models still lack such a resource, limiting our understanding of their development (e.g. Healey et al., 2022; Salamanca-Díaz et al., 2022; Steger et al., 2022; Ton et al., 2023). Our single cell atlas is one of the first created to understand the development of derived traits in a non-model organism.

For syngnathids specifically, this atlas represents an important step towards understanding the genetic nature of unique syngnathid traits. Numerous developmental genetic changes can lead to evolutionary innovations, including the evolution of novel genes, gene duplications, gene losses, gene family expansions or contractions, evolution of regulatory elements, co-option of gene regulatory networks, re-wiring of gene networks, assembly of novel gene networks, and/or the emergence of novel cells (Arendt et al., 2016; Cãestro et al., 2007; Teichmann & Babu, 2004; A. Wagner, 2011; G. P. Wagner & Lynch, 2010). Considering these possibilities, we examined select syngnathid traits and speculated on developmental genetic mechanisms influencing their evolution. Continuing to investigate these proposed mechanisms through expanded scRNAseq atlases and other studies will be critical for understanding syngnathid evolution.

Conserved pathways may contribute to derived syngnathid heads

Syngnathids have highly derived heads including an elongated ethmoid region, uniquely shaped hyoid, and altered muscles and tendons to support specialized “pivot feeding”. The developmental underpinnings of these derived traits have remained underexplored. In our atlas, we identified cell types present in the developing pipefish head and genetic pathways active in those cell types. We identified numerous cells that were present in the developing face: cartilage,

bone, tendons, ligaments, osteochondrogenic mesenchyme, fibroblasts, and unclassified connective tissue cells. Overall, we did not find any unrecognizable cell types, suggesting that genetic modifications within conserved cell types may drive craniofacial modifications.

We next investigated signaling pathways expressed in these cells to determine whether and to what extent developmental genetic reorganization might have occurred. Specifically, we examined the expression of one gene each from three different highly conserved signaling pathways: Wnt (*sfrp1a*), TGF-beta (*bmp4*), and MAPK (*fgf22*). Using *in situ* hybridizations, we found *sfrp1a* and *bmp4* expressed dorsal of the elongating ethmoid plate and surrounding the ceratohyal, suggesting that Wnt and BMP signaling may be active in the lengthening structures. The genes *sfrp1a* and *bmp4* are proposed to influence the development of elongated and broadened craniofacial morphologies in other species (Ahi et al., 2014; Schneider, Gunter, et al., 2023; A. S. Tucker et al., 2000). Pipefish *prdm16* is similarly expressed dorsal to the elongating ethmoid plate. Since *prdm16* regulates Wnt and TGF-beta signaling and these genes regulate chondrocyte differentiation (Bjork et al., 2010; Kaneda-Nakashima et al., 2022; Shull et al., 2022; Wang et al., 2024), their co-expression might suggest a prolonged period of chondrocyte differentiation along the pipefish ethmoid region. Future studies investigating this testable hypothesis would clarify whether prolonged chondrocyte differentiation broadly underlies craniofacial diversity.

We found *fgf22* expressed in the mandible, gill arches, and fins using *in situ* hybridizations, but not in the elongating regions of the head. Interestingly, *fgf22* has not been reported in craniofacial development of any other species, but we found similar craniofacial and fin expression in stickleback fishes (Miyake & Itoh, 2013). Fgf signaling, however, is a conserved and essential pathway for craniofacial development (Crump et al., 2004; Leerberg et al., 2019; McCarthy et al., 2016; Szabo-Rogers et al., 2008; Walshe & Mason, 2003; Woronowicz & Schneider, 2019), raising the possibility that *fgf22* has been co-opted into a role played by different *fgf* genes in other species. Future work should consider a relationship between *fgf22*'s novel craniofacial expression and the loss of the genes that code for Fgf ligands (*fgf3* and *fgf4)* in syngnathids. If *fgf22* is active in existing gene networks, particularly those where *fgf3* or *fgf4* is active in other species, then its novel expression may indicate evolved genetic compensation.

Our analysis suggests ways in which unique syngnathid craniofacial structure could have evolved through genetic network evolution. Unusual expression location (e.g., *fgf22*) and specificity (e.g., *sfrp1a*, *bmp4*, and *prdm16*) in pipefish compared to zebrafish and stickleback suggests that changes in signaling gene deployment and/or content within craniofacial gene networks, particularly genes from Wnt, Fgf, or TGF-beta signaling pathways, could underly the exceptionally elongated syngnathid face.

No evidence of primordia suggests changes in early, not late, tooth development is likely at the root of evolutionary tooth loss

Tooth loss has occurred independently in numerous lineages and has often been studied to understand the developmental basis of character loss. For instance, research in birds and turtles found that tooth programs initiate but are subsequently truncated, explaining toothlessness in mature animals (Y. P. Chen et al., 2000; Tokita et al., 2013). Additional studies in birds have found losses in tooth maturation genes (specifically *scpp* genes, Sire et al., 2008). Since numerous primordium and maturation genes are lost in syngnathids (Lin et al., 2016; Qu et al., 2021; Small et al., 2016; Small et al., 2022; Zhang et al., 2020a), we asked if syngnathids begin tooth development at all.

We found that early tooth development genes were still expressed in pipefish, which is unsurprising given their pleiotropic roles, but found no convincing evidence either in the atlas or in spatial gene expression analysis tissues for tooth primordia. However, these assays surveyed for a transient cell population during a singular developmental timepoint, creating the possibility that this population was present at an earlier or later stage. Tooth primordia may have been undetectable due to syngnathid *fgf3* and *fgf4* losses which could have resulted in insufficient Fgf signaling from the oral epithelium to the dental mesenchyme for tooth initiation (Small et al., 2022; Stock et al., 2006).

If our finding for the loss of earliest stages of tooth development is consistent across developmental stages and the clade, why then have syngnathids retained some members of the Scpp tooth maturation gene cluster? Since studies in birds propose *scpp* gene losses can occur from relaxed selection (Sire et al., 2008), we speculated syngnathids lost *scpp* genes with expression limited to teeth and retained genes with ancestrally pleiotropic expression patterns. In Gulf pipefish, we found the retained genes *scpp1*, *spp1*, and *enam,* are expressed in structures outside of tooth development, suggesting developmental pleiotropic constraint. Specifically, we found *spp1*

and *scpp1* expressed in osteoblasts which is consistent with zebrafish (Bergen et al., 2022; Kawasaki, 2009; Liu et al., 2016) and *enam* expressed in the epidermis which has not been reported in zebrafish (Goldsmith et al., 2003; Jain et al., 2007; Liu et al., 2016). Through examining the conservation of the *scpp* genes in close syngnathid relatives, we found that most *scpp* gene losses (*scpp4*, *scpp7*, and *scpp9* and a functional *scpp5*) are shared with a tooth-bearing outgroup to the family, and likely occurred prior to the loss of teeth in syngnathids. Overall, our analysis favors the hypothesis that pleiotropic *scpp* genes were retained in syngnathids while other *scpp* genes were lost due to relaxed selection.

Redeployment of the bone gene network to build dermal armor

The syngnathid dermal armor is a type of evolutionary novelty, which can arise through either differentiation of serially repeated elements or *de novo* origination, derived from either the redeployment of existing gene networks, rewiring of existing gene networks, or the assemblage of new gene networks (Wagner & Lynch, 2010). Currently, there is no understanding of the developmental genetic underpinnings of the syngnathid dermal armor.

We identified *dlx3b* and *pitx2* expression in tissues where the dermal plates mineralize later in development. Using *in situ* hybridizations, we showed that epithelial and dermal layers expressed *dlx3b* and underlying muscle cells expressed *pitx2*. In some species, dermal bone, plate, or denticle development occurs from the re-deployment of tooth gene regulatory networks (Mori & Nakamura, 2022). However, this does not appear to be the case in syngnathids because the dermal armor lacks the characteristic epithelial–mesenchyme interactions distinguished by *pitx2* expression in the epithelia.

Instead, dermal armor might originate from the co-option of existing bone development gene regulatory networks. Expression of *dlx3b* has been observed in epithelia and mesenchyme during dermal and perichondral bone development in zebrafish (Verreijdt et al., 2006). In addition to *dlx3b*, we observed bone development genes *scpp1* and *ifitm5* expressed in the ossifying dermal armor. Future studies could test our hypothesis that the dermal armor evolved through re-deployed osteoblast networks by examining osteoblast gene network expression over time.

Signatures of embryonic interactions within the novel pouch environment

Syngnathid embryos are reared within the brood pouch, a novel structure and environment composed of male derived tissues (epithelium and placental-like tissues that include specialized cell types) that harbors a pouch microbiome (Stölting & Wilson, 2007). During pregnancy, the

male brood pouch undergoes numerous changes including increased vascularization and altered expression of immune genes (Harada et al., 2022; Ripley et al., 2010; Roth et al., 2020; Small et al., 2013; Whittington et al., 2015). Researchers predict that these changes relate to nutrient and waste transfer and prevention of embryonic rejection and bacterial infection (Dudley et al., 2021; Whittington & Friesen, 2020). However, there are few studies that examine whether and how embryos interact with the brood pouch environment (Kvarnemo et al., 2011; Ripley & Foran, 2006). To consider whether embryos have specializations for life in this brood pouch environment, we asked about cell type-specific expression of nutrient acquisition and/or immune genes.

Pipefish embryos uptake paternally derived carbohydrates, proteins, and lipids (Kvarnemo et al., 2011; Ripley & Foran, 2006). Our data suggests that this uptake could occur through the embryonic epidermis. Specifically, we noticed an enrichment of endocytosis and metabolism genes in epidermal cells. Epidermal absorption of maternally derived nutrients has been suggested in viviparous fishes (Tengfei et al., 2021; Wourms, 1981). Interestingly, microvilli, a type of cellular projection, have been observed on the anal fin of developing seahorses (Wetzel & Wourms, 2004). Possibly, these microvilli may be functionally equivalent to those on the small intestine, maximizing nutrient absorption from the environment.

During pregnancy, the male brood pouch increases expression of C-type lectin genes (Roth et al., 2020; Small et al., 2013; Whittington et al., 2015). These genes are transmembrane or secreted receptors that sense self or non-self and are primarily studied for their role in innate and adaptive immunity (Brown et al., 2018). We identified 14 C-type lectin genes expressed in the embryonic epidermis. Our work suggests that lectin genes are produced by both the father and the embryos, but their function is still unclear. Syngnathid research has primarily suggested that lectin genes are produced to prevent bacterial infection (Melamed et al., 2005), though they could be important for male-embryo recognition.

Overall, our findings suggest that pipefish embryos have evolved to be specialized for development within the brood pouch by expressing genes related to nutrient acquisition and immunity. Future studies could provide insights into when nutrient acquisition and lectin genes are expressed in development, their functional role, and how their expression varies across syngnathid lineages that have exposed versus enclosed embryos, for example, to examine how embryonic development has been impacted by the brood pouch.

Conclusions

Our study represents the first scRNAseq developmental atlas in syngnathids, and one of the first non-model developmental scRNAseq atlases, providing a major step forward for evo-devo research. We used our atlas to begin addressing questions on the evolution and development of syngnathid innovations including their unique craniofacial structure, loss of teeth, dermal armor, and development within the male brood pouch. By combining scRNAseq analysis with spatial expression data from *in situ* hybridization, we made important discoveries in cell type identity and distribution as well as spatial expression of marker and signaling genes. We found that syngnathids express conserved signaling pathways during craniofacial development, suggesting that alterations within these pathways may be important for the evolution of their craniofacial skeletons. We did not find evidence for tooth primordia within syngnathids and propose that genetic changes early in tooth development could have led to their loss of teeth. We propose that the re-deployment of bone gene networks, but probably not tooth networks, could play a role in the dermal armor development. Finally, we observed an enrichment of endocytosis genes and many C-type lectin genes in epidermal cells, which suggests ways these cells might interact with the brood pouch environment. Our atlas pushes forward our understanding of syngnathid development and evolution and provides resources for developmental genetic analysis in nascent evo-devo model species.

References

Ahi, E. P., Kapralova, K. H., Pálsson, A., Maier, V. H., Gudbrandsson, J., Snorrason, S. S., Jónsson, Z. O., & Franzdóttir, S. R. (2014). Transcriptional dynamics of a conserved gene expression network associated with craniofacial divergence in Arctic charr. *EvoDevo*, *5*(1), 1–19.

Ahi, E. P., Lecaudey, L. A., Ziegelbecker, A., Steiner, O., Glabonjat, R., Goessler, W., Hois, V., Wagner, C., Lass, A., & Sefc, K. M. (2020). Comparative transcriptomics reveals candidate carotenoid color genes in an East African cichlid fish. *BMC Genomics 2020 21:1*, *21*(1), 1–15.

Anderson, A. P., & Jones, A. G. (2019). Choosy Gulf pipefish males ignore age but prefer active females with deeply keeled bodies. *Animal Behaviour*, *155*, 37–44.

Arendt, D., Musser, J. M., H Baker, C. V, Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D., & Wagner, G. P. (2016). *The origin and evolution of cell types*.

Bergen, D. J. M., Tong, Q., Shukla, A., Newham, E., Zethof, J., Lundberg, M., Ryan, R., Youlten, S. E., Frysz, M., Croucher, P. I., Flik, G., Richardson, R. J., Kemp, J. P., Hammond, C. L., & Metz, J. R. (2022). Regenerating zebrafish scales express a subset of evolutionary conserved genes involved in human skeletal disease. *BMC Biology 2021 20:1*, *20*(1), 1–25.

Berglund, A., Rosenqvist, G., & Svensson, I. (1986). *Reversed sex roles and parental energy investment in zygotes of two pipefish (Syngnathidae) species*. *29*, 209–215.

Bjork, B. C., Turbe-Doan, A., Prysak, M., Herron, B. J., & Beier, D. R. (2010). Prdm16 is required for normal palatogenesis in mice. *Human Molecular Genetics*, *19*(5), 774–789.

Brown, G. D., Willment, J. A., & Whitehead, L. (2018). C-type lectins in immunity and homeostasis. *Nature Reviews Immunology*, *18*(6), 374–380.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420.

Cãestro, C., Yokoi, H., & Postlethwait, J. H. (2007). Evolutionary developmental biology and genomics. In *Nature Reviews Genetics* (Vol. 8, Issue 12, pp. 932–942). Nature Publishing

Group.

Carcupino, M. (2002). Functional significance of the male brood pouch in the reproductive strategies of pipefishes and seahorses: a morphological and ultrastructural comparative study on three anatomically different pouches. *Journal of Fish Biology*, *61*(6), 1465–1480.

Chari, T., Weissbourd, B., Gehring, J., Ferraioli, A., Leclère, L., Herl, M., Gao, F., Chevalier, S., Copley, R. R., Houliston, E., Anderson, D. J., & Pachter, L. (2021). Whole-animal multiplexed single-cell RNA-seq reveals transcriptional shifts across Clytia medusa cell types. *Science Advances*, *7*(48), 1683.

Chen, J. W., & Galloway, J. L. (2014). The development of zebrafish tendon and ligament progenitors. *Development*, *141*(10), 2035–2045. https://doi.org/10.1242/DEV.104067

Chen, Y. P., Zhang, Y., Jiang, T. X., Barlow, A. J., St. Amand, T. R., Hu, Y., Heaney, S., Francis-West, P., Chuong, C. M., & Maas, R. (2000). Conservation of early odontogenic signaling pathways in Aves. *Proceedings of the National Academy of Sciences*, *97*(18), 10044–10049.

Crump, J. G., Maves, L., Lawson, N. D., Weinstein, B. M., & Kimmel, C. B. (2004). An essential role for Fgfs in endodermal pouch formation influences later craniofacial skeletal patterning. *Development*, *131*(22), 5703–5716.

Denisenko, E., Guo, B. B., Jones, M., Hou, R., De Kock, L., Lassmann, T., Poppe, D., Poppe, D., Clément, O., Simmons, R. K., Simmons, R. K., Lister, R., & Forrest, A. R. R. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biology*, *21*(1), 1–25.

Ding, H. L., Clouthier, D. E., & Artinger, K. B. (2013). Redundant roles of PRDM family members in zebrafish craniofacial development. *Developmental Dynamics*, *242*(1), 67–79.

Docheva, D., Hunziker, E. B., Fässler, R., & Brandau, O. (2005). Tenomodulin Is Necessary for Tenocyte Proliferation and Tendon Maturation. *Molecular and Cellular Biology*, *25*(2), 699–705.

Du, J., Miller, A. J., Widlund, H. R., Horstmann, M. A., Ramaswamy, S., & Fisher, D. E. (2003). MLANA/MART1 and SILV/PMEL17/GP100 Are Transcriptionally Regulated by MITF in Melanocytes and Melanoma. *The American Journal of Pathology*, *163*(1), 333–343.

Dudley, J. S., Hannaford, P., Dowland, S. N., Lindsay, L. A., Thompson, M. B., Murphy, C. R., Van Dyke, J. U., & Whittington, C. M. (2021). Structural changes to the brood pouch of

male pregnant seahorses (Hippocampus abdominalis) facilitate exchange between father and embryos. *Placenta*, *114*, 115–123.

Fabian, P., Tseng, K.-C., Thiruppathy, M., Arata, C., Chen, H.-J., Smeeton, J., Nelson, N., & Crump, J. G. (2022). Lifelong single-cell profiling of cranial neural crest diversification in zebrafish. *Nature Communications 2022 13:1*, *13*(1), 1–13.

Farnsworth, D. R., Saunders, L. M., & Miller, A. C. (2020). A single-cell transcriptome atlas for zebrafish development. *Developmental Biology*, *459*(2), 100–108.

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., & Schier, A. F. (2018a). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, *360*(6392).

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., & Schier, A. F. (2018b). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, *360*(6392).

Feregrino, C., Sacher, F., Parnas, O., & Tschopp, P. (2019). A single-cell transcriptomic atlas of the developing chicken limb. *BMC Genomics*, *20*(1), 1–15.

Fuess, L. E., & Bolnick, D. I. (2023). GBE Single-Cell RNA Sequencing Reveals Microevolution of the Stickleback Immune System. *Genome Biology and Evolution*, *15*.

Fuiten, A. M., & Cresko, W. A. (2021). Evolutionary divergence of a Hoxa2b hindbrain enhancer in syngnathids mimics results of functional assays. *Development Genes and Evolution*, *231*(3–4), 57–71.

Gibert, Y., Samarut, E., Ellis, M. K., Jackman, W. R., & Laudet, V. (2019). The first formed tooth serves as a signalling centre to induce the formation of the dental row in zebrafish. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1904).

Goldsmith, M. I., Fisher, S., Waterman, R., & Johnson, S. L. (2003). Saltatory control of isometric growth in the zebrafish caudal fin is disrupted in long fin and rapunzel mutants. *Developmental Biology*, *259*(2), 303–317.

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, *20*(1), 296.

Hain, D., Gallego-Flores, T., Klinkmann, M., Macias, A., Ciirdaeva, E., Arends, A., Thum, C., Tushev, G., Kretschmer, F., Tosches, M. A., & Laurent, G. (2022). Molecular diversity and evolution of neuron types in the amniote brain. *Science*, *377*(6610).

Harada, A., Shiota, R., Okubo, R., Yorifuji, M., Sogabe, A., Motomura, H., Hiroi, J., Yasumasu, S., & Kawaguchi, M. (2022). Brood pouch evolution in pipefish and seahorse based on histological observation. *Placenta*, *120*, 88–96.

Harlin-Cognato, A., Hoffman, E. A., & Jones, A. G. (2006). Gene cooption without duplication during the evolution of a male-pregnancy gene in pipefish. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(51), 19407–19412.

Healey, H. M., Bassham, S., & Cresko, W. A. (2022). Single-cell Iso-Sequencing enables rapid genome annotation for scRNAseq analysis. *Genetics*, *220*(3).

Hong, X., Wang, Y., Wang, K., Wei, C., Li, W., Yu, L., Xu, H., Zhu, J., Zhu, X., & Liu, X. (2023). Single-Cell Atlas Reveals the Hemocyte Subpopulations and Stress Responses in Asian Giant Softshell Turtle during Hibernation. *Biology*, *12*(7), 994.

Howe, D. G., Bradford, Y. M., Conlin, T., Eagle, A. E., Fashena, D., Frazer, K., Knight, J., Mani, P., Martin, R., Moxon, S. A. T., Paddock, H., Pich, C., Ramachandran, S., Ruef, B. J., Ruzicka, L., Schaper, K., Shao, X., Singer, A., Sprunger, B., … Westerfield, M. (2013). ZFIN, the Zebrafish Model Organism Database: Increased support for mutants and transgenics. *Nucleic Acids Research*, *41*(D1), D854–D860.

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., Arcila, D., Betancur, R., Li, C., Becker, L., Bellora, N., Zhao, X., Li, X., Wang, M., Fang, C., Xie, B., Zhoui, Z., Huang, H., Chen, S., … Shi, Q. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6249–6254.

Jain, I., Stroka, C., Yan, J., Huang, W. M., & Iovine, M. K. (2007). Bone growth in zebrafish fins occurs via multiple pulses of cell proliferation. *Developmental Dynamics*, *236*(9),

Johnson, S. L., Nguyen, A. T. N., & Lister, J. A. (2011). mitfa is required at multiple stages of melanocyte differentiation but not to establish the melanocyte stem cell. *Developmental Biology*, *350*(2), 405–413.

Jungerson, H. F. E. (1910). *Ichthyotomical Contributions: The structure of the Aulostomidae, Syngnathidae, and Solenostomidae*.

Kaneda-Nakashima, K., Igawa, K., Suwanruengsri, M., Naoyuki, F., Ichikawa, T., Funamoto, T., Kurogi, S., Sekimoto, T., Yamashita, Y., Chosa, E., Yamaguchi, R., & Morishita, K. (2022). Role of Mel1/Prdm16 in bone differentiation and morphology. *Experimental Cell*

Research, *410*(2), 112969.

Kawasaki, K. (2009). The SCPP gene repertoire in bony vertebrates and graded differences in mineralized tissues. *Development Genes and Evolution*, *219*(3), 147–157.

Kimura, T., Takehana, Y., & Naruse, K. (2017). Pnp4a Is the causal gene of the medaka iridophore Mutant guanineless. *G3: Genes, Genomes, Genetics*, *7*(4), 1357–1363.

Koiwai, K., Kondo, H., & Hirono, I. (2023). scRNA-seq Analysis of Hemocytes of Penaeid Shrimp Under Virus Infection. *Marine Biotechnology*, *25*(3), 488–502.

Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L., & Burt, D. W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics*, *18*(1), 1–19.

Kvarnemo, C., Mobley, K. B., Partridge, C., Jones, A. G., & Ahnesjö, I. (2011). Evidence of paternal nutrient provisioning to embryos in broad-nosed pipefish Syngnathus typhle. *Journal of Fish Biology*, *78*(6), 1725–1737.

Lamason, R. L., Mohideen, M. A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., Sinha, S., Moore, J. L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P. M., O'Donnell, D., Kittles, R., … Cheng, K. C. (2005). Genetics: SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, *310*(5755), 1782–1786.

Lange, M., Granados, A., VijayKumar, S., Bragantini, J., Ancheta, S., Santhosh, S., Borja, M., Kobayashi, H., McGeever, E., Solak, A. C., Yang, B., Zhao, X., Liu, Y., Detweiler, A. M., Paul, S., Mekonen, H., Lao, T., Banks, R., Kim, Y.-J., … Royer, L. A. (2023). Zebrahub – Multimodal Zebrafish Developmental Atlas Reveals the State-Transition Dynamics of Late-Vertebrate Pluripotent Axial Progenitors. *BioRxiv*, 2023.03.06.531398.

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 1–13.

Leerberg, D. M., Hopton, R. E., & Draper, B. W. (2019). Fibroblast growth factor receptors function redundantly during zebrafish embryonic development. *Genetics*, *212*(4), 1301–1319.

Leysen, H., Roos, G., & Adriaens, D. (2011). Morphological variation in head shape of pipefishes and seahorses in relation to snout length and developmental growth. *Journal of Morphology*, *272*(10), 1259–1270.

Lin, Q., Fan, S., Zhang, Y., Xu, M., Zhang, H., Yang, Y., Lee, A. P., Woltering, J. M., Ravi, V., Gunter, H. M., Luo, W., Gao, Z., Lim, Z. W., Qin, G., Schneider, R. F., Wang, X., Xiong, P., Li, G., Wang, K., … Venkatesh, B. (2016a). The seahorse genome and the evolution of its specialized morphology. *Nature*, *540*(7633), 395–399.

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., Jiang, C., Sun, L., Wang, R., Zhang, Y., Zhou, T., Zeng, Q., Fu, Q., Gao, S., Li, N., Koren, S., Jiang, Y., Zimin, A., Xu, P., … Waldbieser, G. C. (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nature Communications*, *7*.

Manning, C. G., Foster, S. J., & Vincent, A. C. J. (2019). A review of the diets and feeding behaviours of a family of biologically diverse marine fishes (Family Syngnathidae). *Reviews in Fish Biology and Fisheries 2019 29:2*, *29*(2), 197–221.

McCarthy, N., Sidik, A., Bertrand, J. Y., & Eberhart, J. K. (2016). An Fgf-Shh signaling hierarchy regulates early specification of the zebrafish skull. *Developmental Biology*, *415*(2), 261–277.

Melamed, P., Xue, Y., Poon, J. F. D., Wu, Q., Xie, H., Yeo, J., Foo, T. W. J., & Chua, H. K. (2005). The male seahorse synthesizes and secretes a novel C-type lectin into the brood pouch during early pregnancy. *The FEBS Journal*, *272*(5), 1221–1235.

Miao, M., Bruce, A. E. E., Bhanji, T., Davis, E. C., & Keeley, F. W. (2007). Differential expression of two tropoelastin genes in zebrafish. *Matrix Biology*, *26*(2), 115–124.

Miyake, A., & Itoh, N. (2013). Fgf22 regulated by Fgf3/Fgf8 signaling is required for zebrafish midbrain development. *Biology Open*, *2*(5), 515–524.

Mori, S., & Nakamura, T. (2022). Redeployment of odontode gene regulatory network underlies dermal denticle formation and evolution in suckermouth armored catfish. *Scientific Reports 2022 12:1*, *12*(1), 1–16.

Moriyama, Y., Ito, F., Takeda, H., Yano, T., Okabe, M., Kuraku, S., Keeley, F. W., & Koshiba-Takeuchi, K. (2016). Evolution of the fish heart by sub/neofunctionalization of an elastin gene. *Nature Communications 2016 7:1*, *7*(1), 1–10.

Morrison, J. A., McLennan, R., Wolfe, L. A., Gogol, M. M., Meier, S., McKinney, M. C., Teddy, J. M., Holmes, L., Semerad, C. L., Box, A. C., Li, H., Hall, K. E., Perera, A. G., & Kulesa, P. M. (2017). Single-cell transcriptome analysis of avian neural crest migration reveals signatures of invasion and molecular transitions. *ELife*, *6*.

Parker, J., Guslund, N. C., Jentoft, S., & Roth, O. (2022). Characterization of Pipefish Immune Cell Populations Through Single-Cell Transcriptomics. *Frontiers in Immunology*, *13*, 820152.

Partridge, C., Shardo, J., & Boettcher, A. (2007). Osmoregulatory role of the brood pouch in the euryhaline Gulf pipefish, Syngnathus scovelli. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology*, *147*(2 SPEC. ISS.), 556–561.

Potts, H. G., Lemieux, M. E., Rice, E. S., Warren, W., Choudhury, R. P., & Mommersteeg, M. T. M. (2022). Discordant Genome Assemblies Drastically Alter the Interpretation of Single-Cell RNA Sequencing Data Which Can Be Mitigated by a Novel Integration Method. *Cells*, *11*(4), 608.

Qu, M., Liu, Y., Zhang, Y., Wan, S., Ravi, V., Qin, G., Jiang, H., Wang, X., Zhang, H., Zhang, B., Gao, Z., Huysseune, A., Zhang, Z., Zhang, H., Chen, Z., Yu, H., Wu, Y., Tang, L., Li, C., … Lin, Q. (2021). Seadragon genome analysis provides insights into its phenotype and sex determination locus. *Science Advances*, *7*(34), eabg5196.

Ramesh, B., Small, C. M., Healey, H., Johnson, B., Barker, E., Currey, M., Bassham, S., Myers, M., Cresko, W. A., & Jones, A. G. (2023). Improvements to the Gulf pipefish Syngnathus scovelli genome. *GigaByte*, *2023*.

Ripley, J.L., Williams, P. S., & Foran, C. M. (2010). Morphological and quantitative changes in paternal brood-pouch vasculature during embryonic development in two syngnathus pipefishes. *Journal of Fish Biology*, *77*, 67–79.

Ripley, Jennifer L., & Foran, C. M. (2006a). Differential parental nutrient allocation in two congeneric pipefish species (Syngnathidae: Syngnathus spp.). *Journal of Experimental Biology*, *209*(6), 1112–1121.

Ripley, Jennifer L., & Foran, C. M. (2006b). Differential parental nutrient allocation in two congeneric pipefish species (Syngnathidae: Syngnathus spp.). *Journal of Experimental Biology*, *209*(6), 1112–1121.

Rose, E., Johnson, B. D., Armitage, V., & Jones, A. G. (2023). The response of the brood pouch transcriptome to synthetic estrogen exposure in the Gulf pipefish (Syngnathus scovelli). *Frontiers in Marine Science*, *10*(June), 1–12.

Roth, O., Klein, V., Beemelmanns, A., Scharsack, J. P., & Reusch, T. B. H. (2012). Male pregnancy and biparental immune priming. *American Naturalist*, *180*(6), 802–814.

Roth, O., Solbakken, M. H., Tørresen, O. K., Bayer, T., Matschiner, M., Baalsrud, H. T., Hoff, S. N. K., Brieuc, M. S. O., Haase, D., Hanel, R., Reusch, T. B. H., & Jentoft, S. (2020a). Evolution of male pregnancy associated with remodeling of canonical vertebrate immunity in seahorses and pipefishes. *Proceedings of the National Academy of Sciences*, 201916251.

Roth, O., Solbakken, M. H., Tørresen, O. K., Bayer, T., Matschiner, M., Baalsrud, H. T., Hoff, S. N. K., Brieuc, M. S. O., Haase, D., Hanel, R., Reusch, T. B. H., & Jentoft, S. (2020b). Evolution of male pregnancy associated with remodeling of canonical vertebrate immunity in seahorses and pipefishes. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9431–9439.

Royan, M. R., Siddique, K., Csucs, G., Puchades, M. A., Nourizadeh-Lillabadi, R., Bjaalie, J. G., Henkel, C. V., Weltzien, F. A., & Fontaine, R. (2021). 3D Atlas of the Pituitary Gland of the Model Fish Medaka (Oryzias latipes). *Frontiers in Endocrinology*, *12*, 719843.

Salamanca-Díaz, D. A., Schulreich, S. M., Cole, A. G., & Wanninger, A. (2022). Single-Cell RNA Sequencing Atlas From a Bivalve Larva Enhances Classical Cell Lineage Studies. *Frontiers in Ecology and Evolution*, *9*, 783984.

Schneider, R. F., Gunter, H. M., Salewski, I., Woltering, J. M., & Meyer, A. (2023). Growth dynamics and molecular bases of evolutionary novel jaw extensions in halfbeaks and needlefishes (Beloniformes). *Molecular Ecology*, *32*(21), 5798–5811.

Schneider, R. F., Woltering, J. M., Adriaens, D., & Roth, O. (2023). A comparative analysis of the ontogeny of syngnathids (pipefishes and seahorses) reveals how heterochrony contributed to their diversification. *Developmental Dynamics*, *252*(5), 553–588.

Shema, E., Bernstein, B. E., & Buenrostro, J. D. (2019). Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nature Genetics*, *51*(1), 19–25.

Shull, Lomeli C., Lencer, E. S., Kim, H. M., Goyama, S., Kurokawa, M., Costello, J. C., Jones, K., & Artinger, K. B. (2022). PRDM paralogs antagonistically balance Wnt/β-catenin activity during craniofacial chondrocyte differentiation. *Development (Cambridge, England)*, *149*(4).

Shull, Lomeli Carpio, Sen, R., Menzel, J., Goyama, S., Kurokawa, M., & Artinger, K. B. (2020). The conserved and divergent roles of Prdm3 and Prdm16 in zebrafish and mouse craniofacial development. *Developmental Biology*, *461*(2), 132–144.

Sire, J. Y., Delgado, S. C., & Girondot, M. (2008). Hen's teeth with enamel cap: From dream to impossibility. *BMC Evolutionary Biology*, *8*(1), 1–11.

Small, C. M., Bassham, S., Catchen, J., Amores, A., Fuiten, A. M., Brown, R. S., Jones, A. G., & Cresko, W. A. (2016). The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome Biology*, *17*(1), 258. https://doi.org/10.1186/s13059-016-1126-6

Small, Clayton M., Harlin-Cognato, A. D., & Jones, A. G. (2013). Functional similarity and molecular divergence of a novel reproductive transcriptome in two male-pregnant Syngnathus pipefish species. *Ecology and Evolution*, *3*(12), 4092–4108.

Small, Clayton M., Healey, H. M., Currey, M. C., Beck, E. A., Catchen, J., Lin, A. S. P., Cresko, W. A., & Bassham, S. (2022). Leafy and weedy seadragon genomes connect genic and repetitive DNA features to the extravagant biology of syngnathid fishes. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(26), e2119602119.

Soldatov, R., Kaucka, M., Kastriti, M. E., Petersen, J., Chontorotzea, T., Englmaier, L., Akkuratova, N., Yang, Y., Häring, M., Dyachuk, V., Bock, C., Farlik, M., Piacentino, M. L., Boismoreau, F., Hilscher, M. M., Yokota, C., Qian, X., Nilsson, M., Bronner, M. E., … Adameyko, I. (2019). Spatiotemporal structure of cell fate decisions in murine neural crest. *Science*, *364*(6444).

Sommer, S., Whittington, C. M., & Wilson, A. B. (2012). Standardised classification of pre-release development in male-brooding pipefish, seahorses, and seadragons (Family Syngnathidae). *BMC Developmental Biology*, *12*, 12–15.

Songco-Casey, J. O., Coffing, G. C., Piscopo, D. M., Pungor, J. R., Kern, A. D., Miller, A. C., & Niell, C. M. (2022). Cell types and molecular architecture of the Octopus bimaculoides visual system. *Current Biology : CB*, *32*(23), 5031.

Steger, J., Cole, A. G., Denner, A., Lebedeva, T., Genikhovich, G., Ries, A., Reischl, R., Taudes, E., Lassnig, M., & Technau, U. (2022). Single-cell transcriptomics identifies conserved regulators of neuroglandular lineages. *Cell Reports*, *40*(12).

Stiller, J., Short, G., Hamilton, H., Saarman, N., Longo, S., Wainwright, P., Rouse, G. W., & Simison, W. B. (2022). Phylogenomic analysis of Syngnathidae reveals novel relationships, origins of endemic diversity and variable diversification rates. *BMC Biology 2022 20:1*, *20*(1), 1–21.

Stock, D. W., Jackman, W. R., & Trapani, J. (2006). Developmental genetic mechanisms of evolutionary tooth loss in cypriniform fishes. *Development*, *133*(16), 3127–3137.

Stölting, K. N., & Wilson, A. B. (2007). Male pregnancy in seahorses and pipefish: Beyond the mammalian model. *BioEssays*, *29*(9), 884–896.

Szabo-Rogers, H. L., Geetha-Loganathan, P., Nimmagadda, S., Fu, K. K., & Richman, J. M. (2008). FGF signals from the nasal pit are necessary for normal facial morphogenesis. *Developmental Biology*, *318*(2), 289–302.

Tardaguila, M., Fuente, L. de la, Marti, C., Pereira, C., Pardo-Palacios, F. J., Risco, H. del, Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V., & Conesa, A. (2018). SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research*, *28*(3), 396.

Teichmann, S. A., & Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nature Genetics*, *36*(5), 492–496.

Tengfei, D., Yongshuang, X., Haixia, Z., Li, Z., Qinghua, L., Xueying, W., Jun, L., Shihong, X., Yanfeng, W., Jiachen, Y., Lele, W., Yunong, W., & Guang, G. (2021). Multiple Fetal Nutritional Patterns Before Parturition in Viviparous Fish Sebastes schlegelii (Hilgendorf, 1880). *Frontiers in Marine Science*, *7*, 571946.

Thisse, B., & Thisse, C. (2004). *Fast Release Clones: A High Throughput Expression Analysis*. ZFIN Direct Data Submission.

Thisse, C., & Thisse, B. (2007). High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nature Protocols 2008 3:1*, *3*(1), 59–69.

Tokita, M., Chaeychomsri, W., & Siruntawineti, J. (2013). Developmental basis of toothlessness in turtles: Insight into convergent evolution of vertebrate morphology. *Evolution*, *67*(1), 260–273.

Ton, Mai Linh N., Guibentif, C., & Göttgens, B. (2020). Single cell genomics and developmental biology: moving beyond the generation of cell type catalogues. In *Current Opinion in Genetics and Development* (Vol. 64, pp. 66–71). Elsevier Ltd.

Ton, Mai Linh Nu, Keitley, D., Theeuwes, B., Guibentif, C., Ahnfelt-Rønne, J., Andreassen, T. K., Calero-Nieto, F. J., Imaz-Rosshandler, I., Pijuan-Sala, B., Nichols, J., Benito-Gutiérrez,

È., Marioni, J. C., & Göttgens, B. (2023). An atlas of rabbit development as a model for single-cell comparative genomics. *Nature Cell Biology 2023 25:7*, *25*(7), 1061–1072.

Tseng, E. (2021). *cDNA cupcake* (v27.0.0).

Tucker, A. S., Headon, D. J., Schneider, P., Ferguson, B. M., Overbeek, P., Tschopp, J., & Sharpe, P. T. (2000). Edar/Eda interactions regulate enamel knot formation in tooth morphogenesis. *Development*, *127*(21), 4691–4700.

Tucker, A., & Sharpe, P. (2004). The cutting-edge of mammalian development; how the embryo makes teeth. *Nature Reviews Genetics 2004 5:7*, *5*(7), 499–508.

Umemori, H., Linhoff, M. W., Ornitz, D. M., & Sanes, J. R. (2004). FGF22 and its close relatives are presynaptic organizing molecules in the mammalian brain. *Cell*, *118*(2), 257–270.

Van Wassenbergh, S., Roos, G., Genbrugge, A., Leysen, H., Aerts, P., Adriaens, D., & Herrel, A. (2009). Suction is kid's play: extremely fast suction in newborn seahorses. *Biology Letters*, *5*(2), 200–203.

Van Wassenbergh, S., Strother, J. A., Flammang, B. E., Ferry-Graham, L. A., & Aerts, P. (2007). Extremely fast prey capture in pipefish is powered by elastic recoil. *Journal of The Royal Society Interface*, *5*(20), 285–296.

Venema, W. T. C., Ramírez-Sánchez, A. D., Bigaeva, E., Withoff, S., Jonkers, I., McIntyre, R. E., Ghouraba, M., Raine, T., Weersma, R. K., Franke, L., Festen, E. A. M., & van der Wijst, M. G. P. (2022). Gut mucosa dissociation protocols influence cell type proportions and single-cell gene expression levels. *Scientific Reports 2022 12:1*, *12*(1), 1–14.

Verreijdt, L., Debiais-Thibaud, M., Borday-Birraux, V., Van Der Heyden, C., Sire, J. Y., & Huysseune, A. (2006). Expression of the dlx gene family during formation of the cranial bones in the zebrafish (Danio rerio): Differential involvement in the visceral skeleton and braincase. *Developmental Dynamics*, *235*(5), 1371–1389.

Vonk, A. C., Zhao, X., Pan, Z., Hudnall, M. L., Oakes, C. G., Lopez, G. A., Hasel-Kolossa, S. C., Kuncz, A. W. C., Sengelmann, S. B., Gamble, D. J., & Lozito, T. P. (2023). Single-cell analysis of lizard blastema fibroblasts reveals phagocyte-dependent activation of Hedgehog-responsive chondrogenesis. *Nature Communications 2023 14:1*, *14*(1), 1–20.

Wagner, A. (2011). The molecular origins of evolutionary innovations. *Trends in Genetics*, *27*(10), 397–410.

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G., & Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, *360*(6392), 981–987.

Wagner, G. P., & Lynch, V. J. (2010). Evolutionary novelties. In *Current Biology* (Vol. 20, Issue 2, pp. R48–R52). Cell Press.

Walshe, J., & Mason, I. (2003). Fgf signalling is required for formation of cartilage in the head. *Developmental Biology*, *264*(2), 522–536.

Wang, W., Yang, N., Wang, L., Zhu, Y., Chu, X., Xu, W., Li, Y., Xu, Y., Gao, L., Zhang, B., Zhang, G., Sun, Q., Wang, W., Wang, Q., Zhang, W., & Chen, D. (2024). The TET-Sall4-BMP regulatory axis controls craniofacial cartilage development. *Cell Reports*, *43*(3), 113873.

Wetzel, J. T., & Wourms, J. P. (2004). Embryogenesis in the Dwarf Seahorse, Hippocampus zosterae (Syngnathidae). *Gulf and Caribbean Research*, *16*(1), 27–35.

Whittington, C. M., & Friesen, C. R. (2020). The evolution and physiology of male pregnancy in syngnathid fishes. *Biological Reviews*, *95*(5), 1252–1272.

Whittington, C. M., Griffith, O. W., Qi, W., Thompson, M. B., & Wilson, A. B. (2015a). Seahorse Brood Pouch Transcriptome Reveals Common Genes Associated with Vertebrate Pregnancy. *Molecular Biology and Evolution*, *32*(12), msv177.

Williams, R. M., Candido-Ferreira, I., Repapi, E., Gavriouchkina, D., Senanayake, U., Ling, I. T. C., Telenius, J., Taylor, S., Hughes, J., & Sauka-Spengler, T. (2019). Reconstruction of the Global Neural Crest Gene Regulatory Network In Vivo. *Developmental Cell*, *51*(2), 255-276.e7.

Wolf, M., Ferrette, B. L. da S., Coimbra, R. T. F., Jong, M. de, Nebenführ, M., Prochotta, D., Schöneberg, Y., Zapf, K., Rosenbaum, J., Intyre, H. A. M., Maier, J., Souza, C. C. S. de, Gehlhaar, L. M., Werner, M. J., Oechler, H., Wittekind, M., Sonnewald, M., Nilsson, M. A., Janke, A., & Winter, S. (2024). Near chromosome-level and highly repetitive genome assembly of the snake pipefish Entelurus aequoreus (Syngnathiformes: Syngnathidae). *GigaByte*, *2024*, 1–13.

Woronowicz, K. C., & Schneider, R. A. (2019). Molecular and cellular mechanisms underlying the evolution of form and function in the amniote jaw. In *EvoDevo* (Vol. 10, Issue 1, pp. 1–21). BioMed Central Ltd.

Wourms, J. P. (1981). *Viviparity: The Maternal-Fetal Relationship in Fishes*. *21*, 473–515.

Woych, J., Gurrola, A. O., Deryckere, A., Jaeger, E. C. B., Gumnit, E., Merello, G., Gu, J., Araus, A. J., Leigh, N. D., Yun, M., Simon, A., & Tosches, M. A. (2022). Cell-type profiling in salamanders identifies innovations in vertebrate forebrain evolution. *Science*, *377*(6610).

Young, M. D., & Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, *9*(12), 1–10.

Zhang, Y., Ravi, V., Qin, G., Dai, H., Zhang, H., Han, F., Wang, X., & Liu, Y. (2020a). Comparative genomics reveal shared genomic changes in syngnathid fishes and signatures of genetic convergence with placental mammals ! Phylogenomics and evolutionary rate Transposable elements in syngnathids Expansion and contraction of gene families in t. *National Science Review*, *7*(6), 964–977.

Zhou, J., Gao, Y., Lan, Y., Jia, S., & Jiang, R. (2013). Pax9 regulates a molecular network involving Bmp4, Fgf10, Shh signaling and the Osr2 transcription factor to control palate morphogenesis. *Development*, *140*(23), 4709–4718.

BRIDGE

In the previous chapter, I documented Gulf pipefish development through building a single cell atlas at the onset of craniofacial elongation. I identified genes expressed in craniofacial cell types, specifically *prdm16*, *bmp4*, and *sfrp1a*. I observed the expression of these three genes in the mesenchymal cells above the ethmoid plate, where they may be involved with craniofacial elongation.

In this chapter, I examined early craniofacial development in syngnathids using single cell atlases. I used comparative single cell analysis to identify gene expression changes in early development of Gulf pipefish embryos. Although global craniofacial networks were conserved in pipefish, I found novel expression of *fgf22* in stickleback and pipefish pharyngeal arch cells. Additionally, I detected unique endocrine Fgf (*fgf19* and *fgf23*) expression in these cells. These findings suggest that select changes in syngnathid's craniofacial gene networks led to redundancy with *fgf3* and may relate to their elongate snouts.

CHAPTER IV
CO-OPTION FACIILITATES LOCAL CHANGES IN CRANIOFACIAL GENE REGULATORY NETWORKS IN SYNGNATHID FISHES

This chapter is unpublished and was co-authored by Clara Rehmann, Susan Bassham, and William Cresko. Susan Bassham, William Cresko, and I conceptualized the project. I generated the scRNAseq data and completed the analyses. Clara Rehmann developed software for the data analysis. I took the lead on writing the manuscript, and all authors revised the manuscript.

Introduction

Cranial neural crest (CNC) cells are a vertebrate innovation responsible for forming most of the head and jaw. Despite high levels of conservation in CNC gene regulatory networks (GRNs), these cells produce a great array of craniofacial morphologies (Fish, 2019). Morphological changes can originate from GRN re-wiring in CNC or interacting cells that impact cell survival, proliferation, and/or migration into pharyngeal arches (PAs) or the frontonasal stream (Fish, 2016). In a case of local re-wiring, SNPs in the gene *lbh* are suggested to impact CNC migration contributing to divergence in cichlid craniofacial morphology (Powder et al., 2014). However, in comparisons with larger evolutionary distance and similar craniofacial structures, there are numerous sequence changes in genes and regulatory elements that influence CNC expression patterns (ex: ~6% species biased enhancers in Human vs Chimpanzee; Prescott et al., 2015). The scope of CNC network changes in organisms with both high evolutionary divergence and derived heads is largely unknown.

Syngnathid fishes (seahorses, pipefishes, and seadragons) are an exciting system to investigate GRNs underlying derived craniofacial structures. Enabling rapid pivot feeding, their modified heads include a tube-like snout created by elongated ethmoid bones, specialized hyoids that allow the head to pivot, and a lack of teeth (Leysen et al., 2011; Van Wassenbergh et al., 2009). These fishes have lost deeply conserved signaling genes (*fgf3, fgf4*), transcription factors (*eve1, tbx4*), and regulatory elements (*hoxa2b;* Fuiten & Cresko, 2021; Small et al., 2016; Small et al., 2021). The loss of *fgf3* is particularly shocking given that zebrafish *fgf3* mutants lack gill arches (PA 3-7 derivatives) and prematurely die (Crump et al., 2004; Herzog et al., 2004). Despite this loss (*fgf3*), syngnathids possess gill arches. Therefore, CNC networks in zebrafish

and syngnathids must have differences in structure and/or composition beyond the described gene losses.

Based on their elongated snouts and surprising loss of *fgf3*, we wondered what the extent (local vs global) of craniofacial gene network re-wiring in syngnathids (Figure 4.1). Gene networks can be modified through evolution of gene regulatory elements, gene co-option, gene loss, evolution of novel genes, and gene or genome duplication (one such event occurred at the base of the teleost lineage; Amores et al., 1998; Peter & Davidson, 2011). Gene duplicates can have differential fates, such as gene loss, subfunction partitioning, and neofunctionalization which can lead to phenotypic evolution (Arnegard et al., 2010; Cresko et al., 2003; Force et al., 1999). Under the local re-wiring hypothesis, we propose that one or two genetic changes led to redundancy with *fgf3* prior to its loss and these same or few other alterations promoted craniofacial elongation. In the global re-wiring hypothesis, we consider that there could be extensive, global re-wiring in the CNC network that reduces the importance of Fgf pathway in CNC cells and creates a derived head. The two mentioned hypotheses exist on opposite ends of a continuum of GRN modifications, with possibility for the true scenario to be within the spectrum rather than encompassed by the opposing local and global hypotheses.

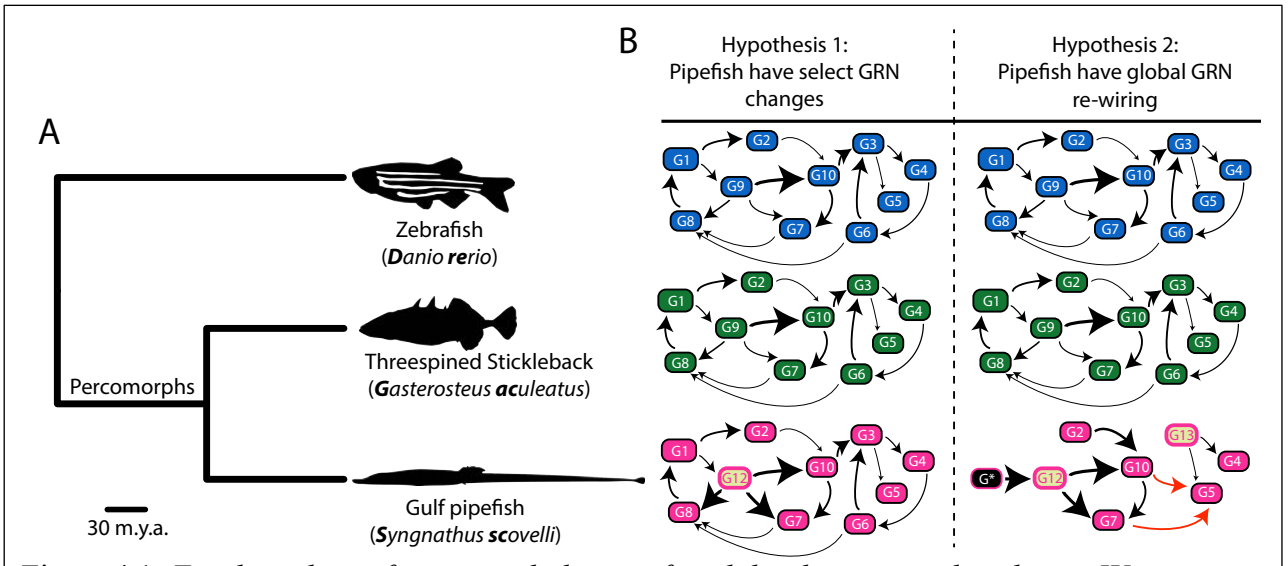


Figure 4.1: *Two hypotheses for syngnathid craniofacial development and evolution*. We examined datasets from three species, zebrafish, threespine stickleback, and Gulf pipefish (panel A). We proposed two hypotheses to understand genomic and morphological evolution of syngnathids (panel B). After the loss of a gene (G9), our local GRN evolution hypothesis (#1) suggests that another gene (G12) could evolve CNC expression and replace G9 in the network; however, in doing so, it may have different relationships with other network members and drive syngnathid morphological evolution. Our second hypothesis is that there are extensive changes to the CNC network that lead to syngnathid craniofacial changes: numerous gene losses (G1, 3, 6, and 9), cases of co-option (G12 and G13), novel gene evolution (G*), and changes in pathway interactions (G7 and G10 to G5).

To detect changes in syngnathid craniofacial GRNs and determine whether these changes are more extensive than in other cell lineages, we harnessed single cell RNA sequencing (scRNAseq). This technique permits the detection and comparison of cell types, gene expression patterns, and GRNs without necessitating dissection (Boroviak et al., 2018; Wang et al., 2021). Although CNC cells can be isolated in model systems through genetic manipulation and cell sorting, these approaches are unfeasible in syngnathids whereby embryos develop within the brood pouch. Furthermore, scRNAseq libraries from syngnathids successfully identified insights into craniofacial elongation as well as other highly derived traits (Healey et al., 2024; Parker et al., 2022). Therefore, scRNAseq permits the direct comparison of CNC cells across species because cell types are distinguished post-hoc.

We created scRNAseq libraries from a syngnathid (*Syngnathus scovelli*, Gulf pipefish) and another percomorph fish (*Gasterosteus aculeatus*, threespine stickleback). Then, compared these atlases to published zebrafish data (*Danio rerio*). We selected these teleosts to include a syngnathid (Gulf pipefish) and two fish with conventional snouts (threespine stickleback and zebrafish) while also accounting for phylogenetic relationships (zebrafish is a cypriniform and stickleback and pipefish are percomorphs. Although we detected some changes in CNC expression, we found overall conservation of the core PA CNC gene network in syngnathids. Strikingly, we identified undescribed expression of *fgf22* in stickleback and pipefish CNC derived pharyngeal arch cells. Additionally, we detected novel expression of endocrine Fgf members, *fgf23*, in PA, mesenchymal, somite, and retinal cells in pipefish. These results suggest syngnathid craniofacial gene network and morphological evolution may have occurred from local changes in GRNs rather than global re-wiring.

Materials and Methods

Production of scRNAseq sequencing libraries

We created single cell RNA sequencing atlases from two replicates each of threespine stickleback and Gulf pipefish. Samples were sequenced at similar developmental timepoints, the onset of pigment in the retina (70hpf for stickleback and around 5-6 dpf for pipefish). For stickleback 70hpf sample one, we used the scRNAseq data from Healey et al., 2022. For stickleback 70hpf sample two, 27 70hpf stickleback embryos were dissociated in .25% trypsin as per Healey et al. (2022) and filtered with a 40uM cell strainer (MTC Bio C4040). Pipefish

samples 1and 2 came from a 5dpf and 6dpf clutch, respectively. Fathers were euthanized using MS-222, embryos were removed from the pouch then dissociated. Pipefish embryos were dissociated using 460 uL trpLE Express (Gibco #12605-010) and 40 uL collagenase I (sigma C0120-100MG) for 8.5 minutes, filtered through 40um cell strainer, quantified with a Bio-Rad cell counter, diluted, and submitted at a concentration 800 cells/uL. All samples were prepared with the 10X Genomics Single Cell 3' v3.1 Dual Index Genome Expression (GEX) – mRNA-Seq kit at the Genomics and Cell Characterization Core Facility (GC3F) at the University of Oregon. Samples were sequenced on an Illumina S4 lane with Paired End 150nt.

Production of ATACseq sequencing libraries

ATACseq libraries were produced from threespine stickleback and Gulf pipefish at the onset of retinal pigmentation. For Gulf pipefish, dissociated cells from the scRNAseq libraries were used as the input to the ATACseq library prep kit. For threespine stickleback, we dissociated cells from 15-20 70 hpf embryos using typLE Express. 340,000 cells were used as the input to the Active Motif kit (part # 105320) to prepare ATACseq libraries. Using the Active Motif kit, we attached sequencing barcodes: pipefish sample 1 (i7 TCGCCTTA, i5 TAGATCGC), pipefish sample 2 (i7 CTAGTACG, i5 CTCTCTAT), stickleback sample 1 (i7 TCGCCTTA , i5 TAGATCGC), and stickleback sample 2 (i7CTAGTACG, i5 CTCTCTAT). Samples were sequenced on an Illumina S4 lane using Paired End 150nt.

scRNAseq library initial processing

Two zebrafish 24hpf (Farnsworth et al., 2020) scRNAseq libraries were downloaded using SRA toolkit (3.0.7) . ScRNAseq reads for all species were processed using CellRanger (v7.2). For zebrafish, we aligned reads to the Lawson genome file (GRCz11_4.3.2/danRer11) and annotation (v4.3.2; Lawson et al., 2020). For stickleback, we aligned reads to the GCF_016920845.1 assembly with its respective annotations extended with scISOseq reads from Healey et al. (2022). We aligned Gulf pipefish scRNAseq reads to the GCF_024217435.1 genome and annotation file from Healey et al. (2024). Using the CellRanger outputs, we identified doublets using scrublet.

Ortholog identification

We identified orthologs between zebrafish, pipefish, and stickleback using GENESPACE (Lovell et al., 2022). We tested variations in the number of species used to identify orthologs, from solely the three species to including 27 additional, high-quality genome assemblies, to find

the highest number of 1:1:1 orthologs (Supplementary file). We also assembled a test set of genes from stickleback and zebrafish to test for accuracy in ortholog identification. After finding the highest number of 1:1:1 orthologs and accuracy of ortholog identification, we chose to use our 3 species and 25 additional species for the GENESPACE (v1.2.3) analysis. From the GENESPACE output, we filtered the pangene output using parse_output.py for stickleback, zebrafish, and pipefish. Then, integrated the data sets and analyzed outputs using custom scripts. Finally, we renamed genes from the cell ranger filtered_feature_bc_matrix output files according to their 1:1:1 ortholog using rename_seq_hopesversion.py. For zebrafish, genes were not renamed. For stickleback, gene names were changed to zebrafish names if the gene had a 1:1 ortholog. For pipefish, gene names were changed to zebrafish names if the gene had a 1:1 ortholog. Additionally, if genes did not have a 1:1 ortholog with zebrafish but it did have one for stickleback, gene names were changed to the stickleback name. For 1: many ortholog cases, genes were not renamed.

scRNAseq atlas construction

Using Seurat (v5.0.3), we pre-processed all the datasets separately. We removed ambient reads with SoupX (v1.6.2) and doublets with scrublet. We then filtered the datasets further to remove cells with high numbers of reads (required less than 100,000 counts in stickleback and pipefish, and less than 40,000 counts in zebrafish), low numbers of detected genes (required greater than 500 genes), high numbers of detected genes (required less than 10,000 detected genes in stickleback and pipefish and 6,000 genes in zebrafish)s, and high percentage of mitochondrial genes (required less than 5% of reads to be mitochondrial). Finally, samples were randomly down sampled to the size of the smallest scRNAseq atlas (pipefish 5dpf). The six datasets were integrated using Seurat CCA method. The integrated data was clustered with 40 PCs at a 1.9 clustering resolution. The success of integration was evaluated using LISI (v1.0). Marker genes were identified using FindAllMarkers and FindConservedMarkers from Seurat. To annotate cluster identity, we examined marker gene expression from zfin and Daniocell then proposed identity based on several markers for each cluster (Farrell et al., 2018; Howe et al., 2013; Sur et al., 2023; Thisse & Thisse, 2004). We further examined marker gene expression in different species using DotPlots. We visualized gene expression patterns, cell abundances, and cell distribution in multidimensional space using Seurat, ggplot2 (v3.5.1), and scCustomize (v2.1.2).

scRNAseq trajectory analysis

To identify whether species' samples were at similar developmental stages, we predicted differentiation status by species using CytoTRACE (v0.3.3; Gulati et al., 2020). For each cluster, we randomly sampled 125 cells from each species, extracted the counts layer, and used iCytotrace to identify differentiation status.

scRNAseq differentially expressed gene analysis

To identify genes that were differentially expressed by species, we examined each cluster individually. Since the number of cells per cluster varied by species, we randomly down sampled each species' cells to 100 cells. For each cluster, we ran FindMarkers in three pairwise comparisons (stickleback vs pipefish, pipefish vs zebrafish, and zebrafish vs stickleback). Next, we examined these data in different subsets: genes expressed in pipefish but not in the other fish, genes expressed in stickleback but not the other fish, genes expressed in zebrafish but not the other fish, genes expressed in both pipefish and stickleback but not zebrafish, genes expressed in both zebrafish and pipefish but not stickleback, and genes expressed in both stickleback and zebrafish but not pipefish. Genes were then filtered to remove genes with p-values above .05 and logFCs below 1. Finally, genes were removed from the analysis if they were not 1:1:1 orthologs (i.e. no lost or potentially novel genes were retained).

scRNAseq gene network analysis

To determine whether Gulf pipefish has derived genetic networks, we built genetic networks using WGCNA (v1.72-5; Langfelder & Horvath, 2008). First, we built gene networks from the entire atlas and species-specific derivations of the atlas. For each approach, we normalized the dataset, ran FindVariableFeatures to identify 2,000 variable features, and scaled the data. Next, we created gene expression matrices using each dataset's variable genes, picked an appropriate soft threshold using pickSoftThreshold, transformed the dataset with the soft threshold, and created a dissimilarity matrix. To build gene networks, we then clustered genes using hclust, created modules with cutreeDynamic (minimum cluster size of 15, guide hang of .05, and hang of .03), and calculated module eigengenes and merged modules with a module eigenegene correlation of 70% or higher. We determined the correlation of each module with cell clusters by calculating the mean correlation of module eigengenes of each cell in the cluster.

scRNAseq cross-species correlation analysis

To assess whether Gulf pipefish have derived core expression profiles, we completed a correlation analysis comparing species in a pairwise manner. Using code derived from Ton et al.(2023), compare_nhoods.R, we built neighbors of cells for each species using miloR (1.8.1; k = 40, d=40, prop=.3, CCA integrated dimension). Then, we calculated neighborhood correlations between every neighborhood in the two species using 8,000 highly variable genes which were filtered to only include 1:1:1 genes. To compare clusters across species, correlation scores were filtered to ensure that the index cell of compared neighborhoods belonged to the same cluster. We completed this cluster level comparison with binned (combining clusters with similar annotations) and raw cell clusters. For the binned cell cluster approach, we created 11 categories: pharyngeal arch CNC cells (cluster 16), mesenchymal, somites, and notochord cells (1, 2,15,and 35), unknown cells (10), neural cells (0,3,4,5,6,7,9,12,13,14,19,29,31,32,33,34,36,37,41,43, and 46), red blood cells (8,11,17,21,22, and 39), endothelial cells (24), epithelial cells (18, 23, 26, and 28), muscle cells (25, 27, 44, and 45), immune cells (30), pigment cells (38 and 40), and endodermal cells (20 and 42).

ATACseq read processing and analysis

We downloaded two samples of 24hpf zebrafish data from (Franke et al., 2021)using SRA toolkit (3.0.7). Reads from stickleback, pipefish, and zebrafish were aligned to their respective genome assemblies using Bowtie2 (v2.3.4). Mitochondrial reads were then removed from the threespine stickleback and zebrafish genome assemblies using grep (the Gulf pipefish genome does not have mitochondrial sequences). Reads were filtered out with low quality scores (q10) and sorted the files using samtools (v1.19). Next, we marked and removed duplicate reads with picard (v2.6.0). To prepare reads for peak identification, we sorted reads by name using samtools (v1.19). Finally, identified peaks using genrich (v0.6.1) with both sample replicates.

Whole genome alignment and Identification of conserved elements

To determine whether genes with detected expression changes in Gulf pipefish may have alterations in regulatory elements, we paired out ATAC-seq analysis with a conserved element approach. First, we selected 12 species (5 syngnathids: Gulf pipefish, dwarf seahorse, leafy seadragon, scribbled pipefish, and blue Striped Pipefish; 1 long-snouted outgroup: bluespotted cornetfish; 1 outgroup with a “short” snout: mandarin dragonet; 5 additional species: zebrafish, northern pike, japanese medaka, threespine stickleback, and southern bluefin tuna). Identifier repeat elements with RepeatModeler (v 2.0.5) and masked them with RepeatMasker (v 4.1.5).

For RepeatMasker, we used concatenated repeat files generated from all species RepeatModeler results as the repeat library input. Aligned masked genome assemblies using cactus (v2.8.1; Armstrong et al., 2020). Then, for each chromosome of each species, generated maf files using cactus-hal2maf and identified conserved elements using phastcons (v1.5; target-coverage 0.125, expected-length 12, most-conserved). Conserved element bed files were concatenated to include all chromosomes for each species separately. Then, chain files which created pairwise species maps were generated using cactus-hal2chains. Finally, conserved elements and ATAC-seq peaks were lifted over using these chain files to respective genome assemblies using UCSC liftover (minMatch=.75).

In situ Hybridization

For *fgf22*, we completed *in situ* hybridizations in Gulf pipefish (*Syngnathus scovelli*). Primer sequences were designed using NCBI Primer Blast and synthesized embryonic cDNA libraries. Probes were prepared using TOPO cloning. For Gulf pipefish embryos, we allowed the fish to mate in our facility then harvested embryos once they reached the appropriate stages. Fish were reared in 25 degrees C water with 25-28 ppt salinity. For stickleback embryos, we completed in vitro fertilization with Cushman Slough individuals. Embryos were reared at 20 degrees C using standard stickleback rearing procedures. Zebrafish were obtained from the Aquatic Animal Care Services. Embryos and larvae were fixed in 4% PFA, dehydrated through a series of PBT/MeOH washes, and stored in MeOH at -20C. 5-12 embryos were used for each experiment. We completed *in situ* hybridizations in keeping with Thisse & Thisse, 2007, leaving the embryos in stain until background was observed. After imaging, we used the levels tool in Adobe Photoshop (v23.4.2) to white balance the pictures.

Results

Successful integration of Gulf pipefish, stickleback and zebrafish single cell data sets

To compare syngnathid craniofacial gene expression to species with non-derived faces that possess *fgf3*, we completed a cross species single cell analysis using Gulf pipefish, stickleback, and zebrafish at the pharyngeal arch stage, the timepoint where the loss of *fgf3* is detrimental to zebrafish head development. The integrated atlas contained 32,814 cells (5,469

cells per sample) that contributed to 47 cell clusters (Figure 4.2). From cluster annotation with *de novo* marker genes analyzed using published zebrafish resources (S4.1:S4.20; Farrell et al., 2018; Howe et al., 2013; Sur et al., 2023; Thisse & Thisse, 2004), we identified cells from all major lineages. These included neural (such as CNC derived neurons), somites, mesenchyme, epidermis, periderm, pigment, immune, erythrocytes, pharyngeal pouch, and pharyngeal arch cells. Cell clusters are well distinguished from one another which enables detailed analyses (S4.21). Furthermore, apart from erythrocyte clusters, species' cells mixed well in UMAP space, suggesting a successfully integrated atlas (S2; Tran et al., 2020). Poor erythrocyte integration is likely due to a high imbalance in cell proportions (Figure 4.22), these cells contribute to roughly 20% of the pipefish cells but around 2% of stickleback and zebrafish samples.

Although erythrocyte clusters represent an extreme case, all cell clusters except for the somitic mesoderm had variation in abundance across species. We suggest that this variance may originate from dissociation or biological differences rather than indicate major developmental mismatch. We found that species largely had congruent cellular differentiation states suggesting that these datasets are comparable (S4.23; Gulati et al., 2020). Additionally, cell number disparities spanned cell lineages (such as increased zebrafish cells from the epidermis/periderm) that are not explainable by cells being in different developmental stages. Thus, biological (differences in cell proliferation/death) or dissociation (methodological differences that alter ease of cell-cell separation) explanations may illuminate these cell number differences.

We focused our analyses around three main cell clusters, pharyngeal arch cells derived from CNC (#16), pharyngeal pouch cells (#28), and CNC derived neural cells (#32), to identify developmental genetic changes relevant for syngnathid craniofacial evolution. Pharyngeal arch and pouch cells contribute to building the vertebrate head while CNC neural cells can be used as a CNC comparator. Although pipefish had fewer pharyngeal pouch cells than stickleback and zebrafish, they had higher numbers of CNC PA cells (3.35% of pipefish cells) compared to

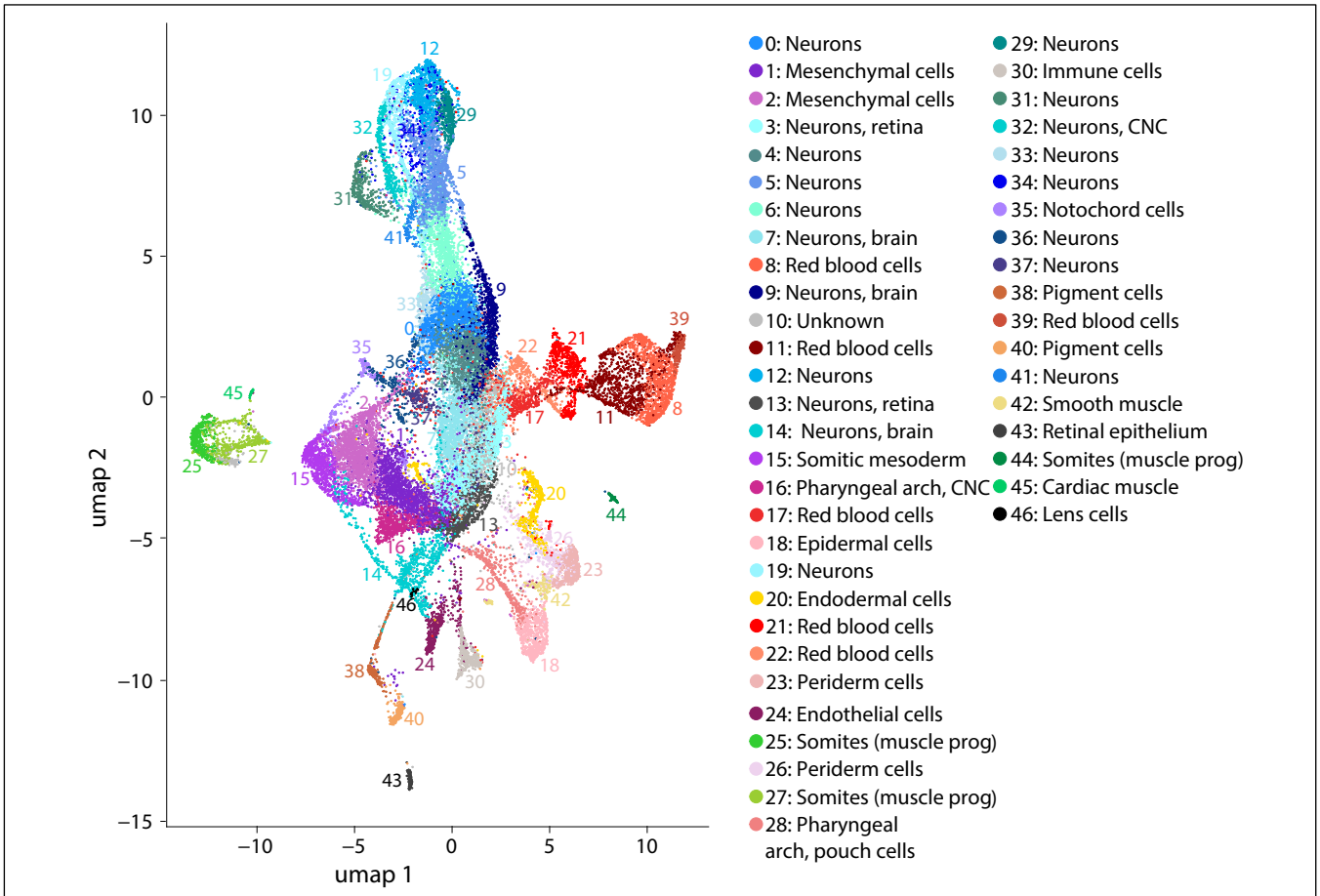


Figure 4.2: *Integrated atlas encompasses cells from the entire embryo*. Cells are plotted on the first two CCA integrated dimensions and colored by cell cluster. Cell cluster annotations are to the right.

stickleback (1.85%) and zebrafish (1.35%). To probe whether this change in PA cells could relate to a higher overall number of CNC cells or a greater number of CNC cells differentiating into skeletogenic fates, we compared PA and neural cells derived from the CNC. We found that stickleback and pipefish had roughly the same number of total cells (stickleback = 403; pipefish = 409), but that pipefish had a higher number of PA cells than neural cells (PA: stickleback = 199, pipefish = 365; neural: stickleback = 204, pipefish = 44). Zebrafish, like stickleback, had relatively equal numbers of PA and neural derived CNC cells (141 neural cells; 151 PA cells). This finding provides preliminary evidence for differential cell abundance of PA cells in Gulf pipefish, and further motivates investigations on CNC PA cells gene networks.

Cranial neural crest pharyngeal arch core gene regulatory network is largely conserved in Gulf pipefish

To determine whether there are changes in gene regulatory networks in Gulf pipefish, we built networks using WGCNA (Langfelder & Horvath, 2008). Despite using similar parameters across each species' analysis, we only recovered a PA network in the pipefish atlas. We compared expression of the pipefish network genes across species (Figure 4.4A) and observed a high level of similarities in presence/absence across species (*dlx* genes, *barx1, lhx6, mab21l2, sox9a*, and *emx2*). However, *G9886,* from the pipefish network, was not present in stickleback and zebrafish gene lists. This gene has an ortholog in stickleback, *LOC120826634*, but it was filtered out of the dataset because it is a ncRNA. This analysis suggests that the PA network detected in pipefish is largely conserved.

Since the WGCNA analysis did not detect networks in stickleback or zebrafish, we completed an additional exploration to determine similarities across cell types. Specifically, to discern whether strength of pharyngeal arch cell associations between species is reduced in comparisons with pipefish, we measured the correlation of cell neighborhoods across species and associated these with cell identities (Ton et al., 2023). We identified cell neighborhoods from each species' data then completed a pairwise neighborhood correlation analysis with 3,320

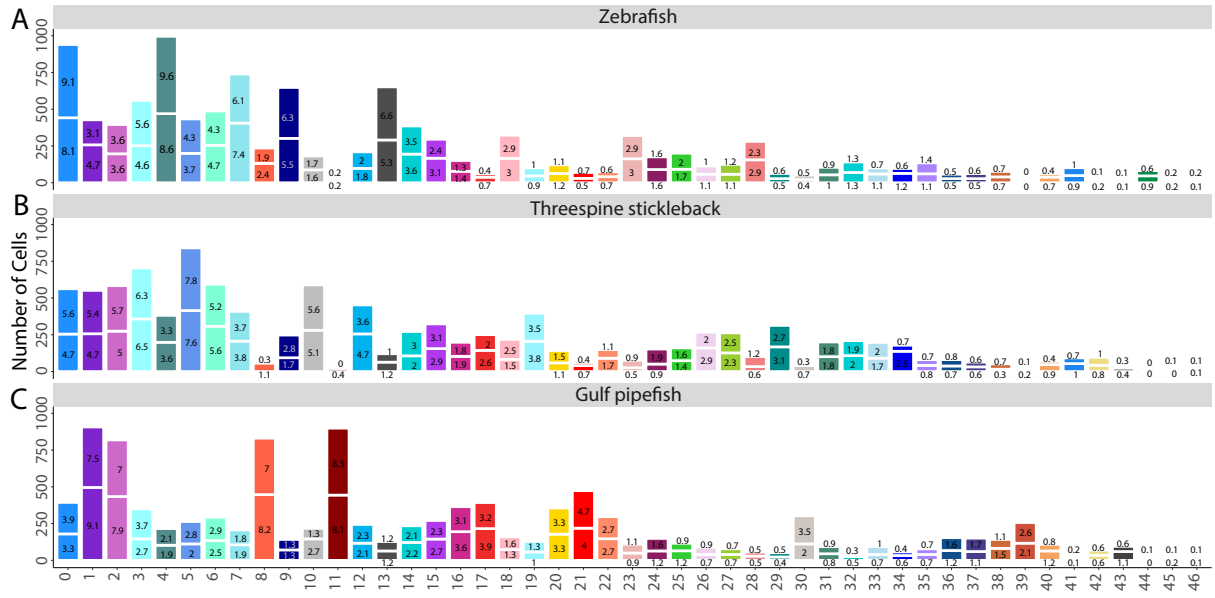


Figure 4.3: *Number of cells per cluster varied by species*. Species data is separated into three panels with the two samples distinguished by a white line. The percentage that each cluster contributes to the sample's library is labeled on the bars.

highly variable 1:1:1 orthologs. Examining species' correlations across the 47 cell clusters revealed that the PA CNC cluster is highly conserved (within the top six highly correlative clusters; S4.24; S4.25). To reduce bias from species differences in cell abundances, we binned clusters into cell types. This approach again identified pharyngeal arch cells as one of the highest correlative cell types across all comparisons (Figure 4.4C-E). Despite preservation of the high PA rank, median correlation scores varied among comparisons with PA cells: pipefish-stickleback (.87), stickleback-zebrafish (.85), and pipefish-zebrafish (.80; S4.26). Overall, these data suggest a maintenance of core pharyngeal arch gene networks in Gulf pipefish despite select GRN changes.

We next asked how many genes have differing expression profiles in PA cells in Gulf pipefish. For each cluster we compared 1:1:1 ortholog expression using pairwise species contrasts, applied stringent filtering, and compiled data to analyze specific groupings (ex: genes expressed in pipefish PA cells but not present in stickleback or zebrafish). We observed a strong influence of shared evolutionary history on gene expression (Figure 4.4B; S4.27). While the "short" snouted fish and both morphologically and phylogenetically distant (pipefish and zebrafish) comparisons had only 3 unique genes each, the percomorphs *vs.* zebrafish contrast had 296 genes for PA cells. This same pattern is observed in other cell types as well, which may be the result of phylogenetic distance or technical differences (staging could be closer in pipefish and stickleback than zebrafish, zebrafish samples were prepared with 10X genomics V2 chemistry while the percomorph samples were isolated with V3 scRNAseq chemistry). We next identified 1:1:1 orthologs that were only expressed in one species. In CNC PA cells, pipefish had the highest number of biased expressed genes (28). Stickleback and Zebrafish had 5 and 2 genes, respectively. All pipefish PA candidates were expressed across numerous cell types (including somites, mesenchyme, and notochord). This suggests pipefish instances of gene co-option is not specific to PA cells.

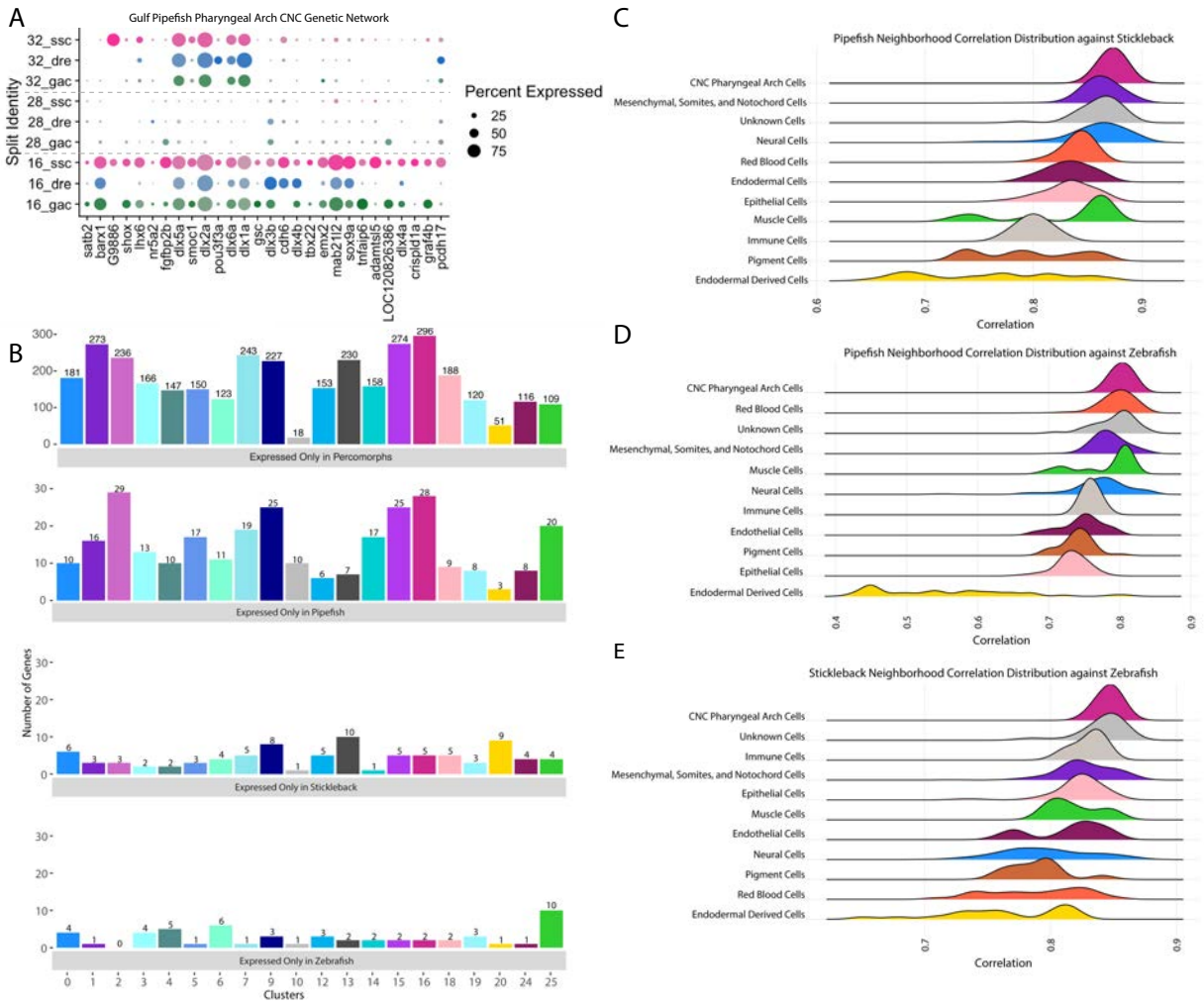


Figure 4.4: *Pharyngeal arch cells express highly conserved core genetic program.* In panel A, the genes from the Gulf pipefish genetic network which correlated with pharyngeal arch cells are plotted on a DotPlot. In panel B, cluster gene expression is compared between species to identify 1:1:1 orthologs that have divergent expression profiles. Cells for each species per cluster were down sampled to 100. Cell types are separated by species. The size of the dot is proportional to the percentage of cells that express the gene. In panels C-E, pairwise correlation analyses demonstrate that pharyngeal arch cells have maintained core gene expression profile in all species. Cell types are binned into categories based on common identities.

Gene losses and potential novel genes in Gulf pipefish pharyngeal arch cells alter gene regulatory networks

After examining 1:1:1 ortholog expression, we considered the role of gene loss on Gulf pipefish craniofacial networks. We identified genes potentially missing in each species using our ortholog datasets. This data was derived using GENESPACE by supplying 28 species' genome assemblies to identify orthologs between species. We took advantage of this phylogenetic sampling to identify genes that were not detected in 5 species of syngnathids (including Gulf pipefish), 3 species of cypriniforms (including zebrafish), or 3 species of perciforms (including

stickleback). By also removing genes undetected in their closest outgroup, we derived a list of potential unique losses for each clade including 182 genes in syngnathids, 1 gene in cypriniforms, and 12 genes in perciforms (S4.28: S4.29). The syngnathid unidentified gene list includes known losses: *odam, fgf3, fgf4,* fgf9, *eve1*, and *tbx4*. Again, none of the potentially missing genes were specific for PA cells. While there is no evidence of extensive craniofacial gene loss, syngnathid craniofacial evolution may have been influenced by loss of select genes.

After showing that Syngnathids do not have widespread PA-specific gene losses, we wondered whether they could have gained novel genes with key roles in craniofacial development. We, again, took advantage of our GENESPACE analysis to distinguish genes present in pipefish, zebrafish, and stickleback that were found in one or more other members of their lineage but not annotated in other species. We found 293 pipefish genes only detected in 1 or more other syngnathids, 566 zebrafish genes in cypriniforms, and 16 stickleback genes in perciforms. Since the number of unique genes is dependent on the evolutionary divergence of species included in the GENESPACE run, we focused interpretation on gene expression patterns rather than gene quantity comparisons. 47 of the 293 syngnathid only genes have the highest percent expression in the pharyngeal arch cells; however, all these genes are broadly expressed (S4.30: S4.31). Overall, these data did not find evidence for the evolution of novel genes specific to syngnathid craniofacial gene networks.

GENESPACE analysis identifies candidates for broader syngnathid evolution

Although the primary focus of this study was craniofacial development, we found interesting gene candidates for other syngnathid derived traits. For instance, the evolution of the syngnathid male brood pouch is suspected to be related to syngnathid immune gene losses (Roth et al., 2020). The gene *marco*, a macrophage receptor, was not detected in any syngnathid via GENESPACE; however, a *marco*-like gene is present in Gulf pipefish (*loc125973236*). Therefore, this gene may be derived in syngnathids. Next, we observed interesting changes in respect to DNA repair and TE expansion. The gene *lrrc31*, a repressor of DNA repair, was not detected in any syngnathids. Additionally, three *tigger* TE derived genes (*loc125967667, loc125980098, loc125967667*) were only identified in syngnathids. This finding complements previous studies that noted an overrepresentation of *tigger* repeats in syngnathids (Small et al., 2022).

Numerous candidate genes identified from global analyses, but few are promising for craniofacial evolution

Although global analyses (gene network, expression changes, and genome content) identified numerous changes in pipefish, few candidate genes emerged with relevance for craniofacial development. The gene *efcc1*, associated with the disease osteogenesis imperfecta, was not detected in syngnathid genomes though the functional consequence of this loss is unclear (S4.28; Smith et al., 2020). We additionally found three genes that were present in syngnathid genomes but did not have pipefish PA expression (in contrast to stickleback and zebrafish): *cul3a*, involved in protein ubiquitination; *akr1a1b*, involved in cellular detoxification of aldehyde; and *dynlt3,*a member of dynein light chains (Aktary et al., 2021; Qi et al., 2021; Sakaue et al., 2017). However, none of these genes appeared relevant for syngnathid craniofacial evolution.

Despite having 293 genes uniquely identified in syngnathids, none of the genes were unique to the pharyngeal arch. Candidates present in all species but with pipefish biased expression in PA cells include *podxl2, arhgap4a, ecrg4a, otos*, and *crispld1a*. The genes *podxl2* and *arhgap4a* are regulators of cellular movements, suggesting potential alterations in pipefish PA cell migration (Sauteur et al., 2017; Vogt et al., 2007). Interestingly, the three other genes, *ecrg4a, otos*, and *crispld1a*, are expressed in PA derived cell types in zebrafish but not until 42-48 hpf when cartilage cells begin to differentiate (Baanannou et al., 2020; Melvin et al., 2013; Sur et al., 2023). These limited data suggest potential premature activation of cartilage differentiation in Gulf pipefish.

Of relevance to craniofacial development, we found changes to conserved genes between percomorphs and zebrafish. Differences between percomorphs and zebrafish include teleost gene duplicates: *ednrab* (an endothelin pathway ligand, expressed in zebrafish), *fgfbp2b* (an Fgf pathway binding protein, expressed in percomorphs), and *wnt5a* (a Wnt pathway ligand, exclusive to percomorphs). One deviation from this phylogenetic pattern was in *snai1b* which is expressed in zebrafish and pipefish but not stickleback. Considering syngnathid evolution, these results indicate that few genes have changed expression in Gulf pipefish PA cells, but that paralog evolution post-duplication is on-going.

Gulf pipefish has novel Fgf expression in pharyngeal arch cells

In addition to detecting the above-described genes in our global analyses, we also found unique changes in conserved signaling pathway genes (such as *fgf23* uniquely expressed in pipefish PA cells). These observations as well as the loss of *fgf3* and *4* in syngnathids motivated a thorough analysis of key signaling pathways for craniofacial development. We, specifically, compared Fgf, Bmp, Notch, Endothelian, and Wnt ligands, receptors, and interactors expression between zebrafish, stickleback, and pipefish (S4.32: S4.37). Since samples varied in sequencing depth and preparation, we focused on changes in gene presence or absence rather than abundance.

While the presence/absence of signaling gene expression in CNC derived PA cells was largely conserved among all three species, we found select cases of unique loss and gain of signaling gene expression patterns in Gulf pipefish. Genes previously identified as missing, *fgf3* and *fgf4*, in syngnathid genomes had PA cell expression in zebrafish and stickleback CNC derived and pouch cells (Figure 4.5). In pipefish, we detected two additional gene losses, *wnt7aa* (however, this is retained in *Corythoichthys intestinalis*) and *fzd1.* The genes *wnt7aa* and *fzd1* were expressed in PA pouch cells in stickleback and zebrafish; additionally, *fzd1* was expressed in CNC derived PA cells. From genes present in all three species, we observed four intriguing genes: *fgf22, fgf23, fgf19*, and *bmp7b.* All three *fgf* genes were absent in zebrafish CNC derived PA cells, but present in pipefish; however, *fgf22* was also minimally expressed in stickleback PA cells. Percomorph specific PA expression of *fgf22* was confirmed using *in situ* hybridizations, and an additional novel domain was observed in percomorph fin buds (S4.38). Interestingly, *fgf19* is adjacent to *fgf3/4* in the genome, perhaps suggesting co-option of regulatory elements. Like *fgf19*, *fgf23* was weakly expressed in PA cells and belongs to the endocrine category of Fgf ligands. Outside of the Fgf family, *bmp7b* was expressed in CNC derived PA cells of all species, but uniquely absent in the PA pouch of Gulf pipefish. Overall, this analysis identified minimal alterations to signaling gene expression presence/absence in Gulf pipefish.

In addition to syngnathid specific findings, we found deviations in percomorphs vs zebrafish expression of signaling genes. Primarily, these changes were seen in paralogs derived from the teleost whole genome duplication event. As discussed earlier, we found differential gene expression of *wnt5* paralogs, *wnt5a* and *wnt5b* (Figure 4.5). Pipefish and stickleback expressed *wnt5a* and *wnt5b* in CNC derived PA cells, however, zebrafish only expressed *wnt5b* in these cells. We also found evidence for differential gene retention of paralogs, whereby

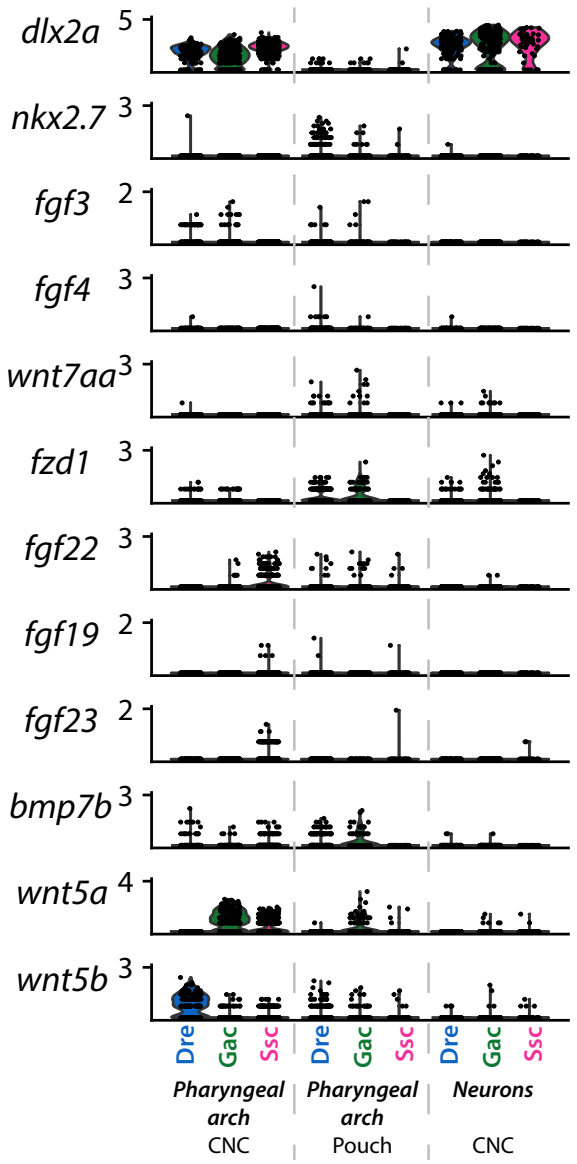


Figure 4.5: *Novel Fgf and Wnt pathway expression changes in Gulf pipefish CNC pharyngeal arch cells*. Expression pattern of signaling genes is displayed in a Violin plot, whereby each dot is a cell and species are plotted individually. The gene *dlx2a* marks CNC derived cells (cluster 16 and 32) and *nkx2.7* marks pharyngeal arch epidermal cells (cluster 18). Select signaling genes are highlighted for differences in expression between species.

CEs identified across species.

zebrafish retained *fgf18b*, *jag2a*, *fzd8b, dkk1b, bmp2a, bmpr1ab, bmpr2b, sfrp1b, nkd2b,* and *fgfbp2a* but these genes have not been identified in stickleback or pipefish genomes (S4.32: S4.37). This gene duplicate analysis indicates that stickleback data is essential for interpreting developmental changes relevant for Gulf pipefish craniofacial evolution.

Potential changes in gene sequence and regulation may lead to Gulf pipefish expression differences

To hypothesize whether changes in expression may relate to changes in regulatory regions, we produced a whole genome alignment of 12 fishes to identify conserved elements (CEs) and analyzed ATACseq data of zebrafish, stickleback, and pipefish to find open chromatin regions. We first examined genes lost in syngnathids including *fgf3* and *4, fzd1*, and *wnt7aa* for sequence evolution. For instance, although *fgf3* and *4* are still present in the syngnathid outgroup (the bluespotted cornetfish), *fgf4* did not align with sticklebacks and two small *fgf3* CEs from exons 2 and 3 were not found (S4.39). Potentially, this indicates sequence degradation prior to gene loss in syngnathids. In contrast to *fgf3's* case, we did not detect losses in cornetfish *fzd1* CEs (Figure 6A). The gene *wnt7aa,* which also was not detected in most syngnathids, had sparse

Next, we examined genes with alterations in expression including *fgf19, fgf22, fgf23,* and *bmp7b.* For *fgf22,* there was a high degree of CE correspondence between species (apart from zebrafish; figure 4.6B). These conserved elements aligned with ATACseq peaks from stickleback and pipefish, suggesting a functional relevance. In the case of *bmp7b,* whereby Gulf pipefish lacked expression in the PA pouch, there was a CE missing in the first intron of most syngnathids except *doryrhamphus excisus.* However, this CE did not overlap with ATACseq peaks. Despite presence in all species except for *D. excisus*, syngnathid *fgf19* and *23* sequences did not align with non-syngnathid species (S4.40; Figure 4.6C). These misalignments are indicative of genomic rearrangements, *fgf19* neighbors the *fgf3/4* locus and *fgf23* neighbors *tigarb* which is lost in most syngnathids, and these rearrangements may be relevant for gene regulation.

Since different fgf ligands have preferential

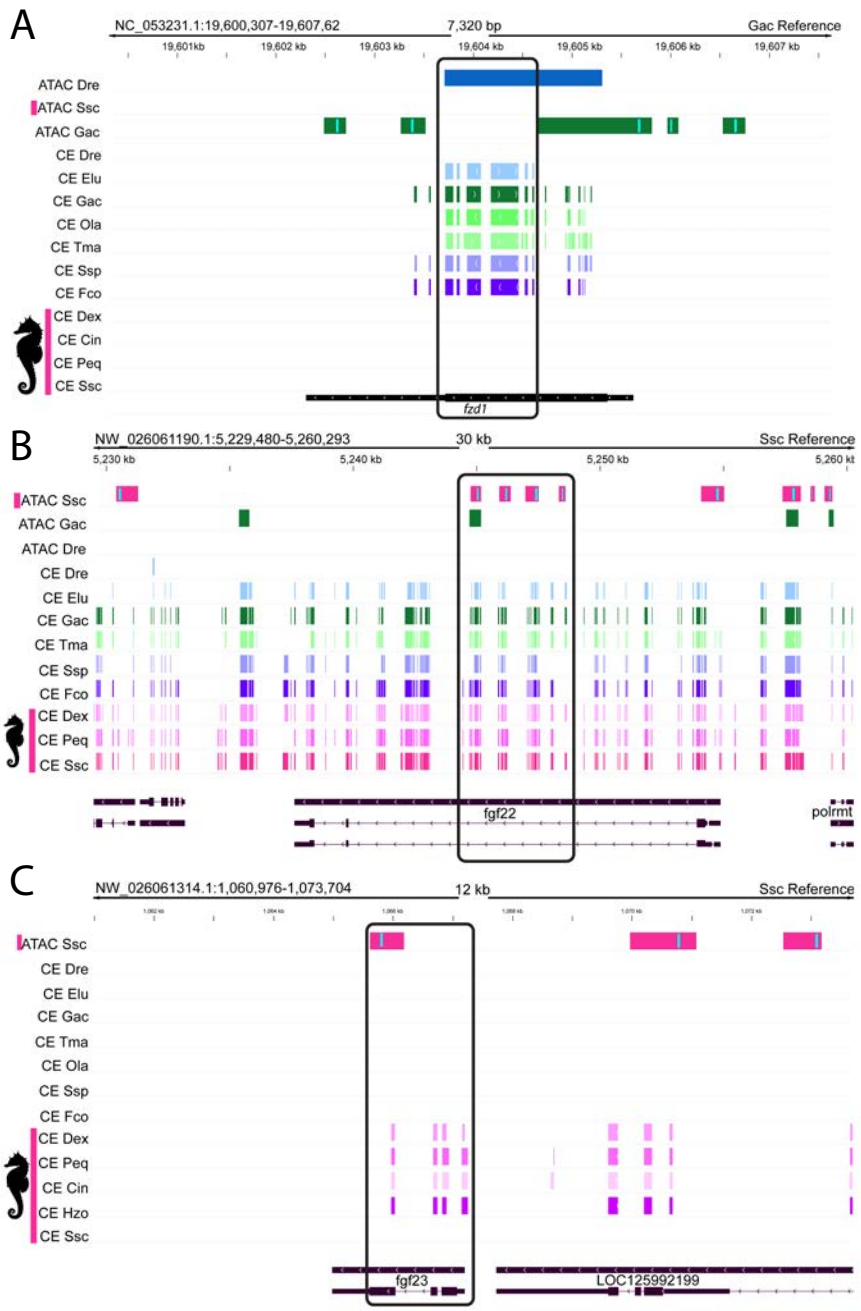


Figure 4.6: *Potential changes in gene sequence and regulation may lead to Gulf pipefish expression differences*. Panels A-C illustrate identified conserved elements (CE) and ATACseq peaks. Pink bars and seahorses to the left of the species names indicate syngnathids. Gene models are shown at the base of the graph, whereby thicker regions show exons and thinner regions show introns. Plots illustrate changes in conserved elements in genes that have altered expression or content in Gulf pipefish.

binding to fgf receptors, we, lastly, examined the CE landscape around fgf receptors. Although there were no syngnathid CE losses around *fgfr1a* and *fgfr2, fgfr3* and *fgfr4* have several syngnathid CE losses (S4.41). The gene *fgfr1b,* in contrast, did not have many annotated CEs and is not present in the syngnathus genus. However, it interestingly gene models in stickleback indicate a loss of splice isoform b which was confirmed using VISTA analysis (S4.42).

Discussion

In this study, we created the first comparative, developmental scRNAseq atlas to determine the scale of GRN re-wiring in syngnathid craniofacial development. We showed that Gulf pipefish CNC cells have not experienced global GRN re-wiring despite having highly derived head and jaws and numerous gene losses. Although we found gene turnover in syngnathids (293 genes unidentified in other species and 182 suspected losses), most candidates were not expressed in CNC cells or were likely auxiliary to the core network based on their broad expression. However, local GRN changes, including co-option of *fgf22, fgf19,* and *fgf23* in CNC*,* cells may explain syngnathid's resiliency to *fgf3* loss and could relate to their lengthened snouts. We propose that FGF pathway evolution in percomorphs led to redundancy of *fgf22* and *fgf3* in CNC cells, permitting *fgf3* loss in syngnathids, and that *fgf22* may be upregulated in syngnathids which could result in increased CNC proliferation. Overall, our data supports the hypothesis that pathway redundancy enabled gene loss and phenotype evolution occurred without widespread network changes in syngnathids. This case study of GRN evolution showcased that it is possible to find the "needles in the haystack" that drive derived trait development due to the conservation of core gene regulatory networks.

<u>Extensive teleost genome duplication paralog evolution but no evidence for global re-wiring in pipefish CNC</u>

Examinations of developmental pathways, gene losses, potential novel genes, and gene networks found few candidate genes for the evolution of syngnathid derived faces. In addition to Fgf pathway changes, we identified gene losses in the Wnt (*fzd1* in all syngnathids and *wnt7aa* in syngnathids apart from *D. excisus*) and an expression loss in the Bmp (*bmp7b*) pathway. These genes are all expressed in the pharyngeal arch, but there is no evidence that their loss would be detrimental (Baldera et al., 2023; Beretta et al., 2011; Duncan et al., 2015; Shawi & Serluca, 2008). Extending from known signaling pathways to *de novo* candidates, we found evidence for

gene turnover in syngnathids but did not identify suggestive candidates for craniofacial evolution. Consistent with examinations from other species (birds, cichlids), these analyses suggest local rather than global re-wiring drive syngnathid craniofacial evolution (Bhullar et al., 2015; Powder et al., 2014).

In comparison to the minimal GRN changes between syngnathids and "short" snouted fishes, we observed a strong phylogenetic signal in the expression data particularly with TGD paralogs. We identified 11 signaling pathway gene losses and 2 expression differences in percomorphs vs zebrafish TGD paralogs. One expression discrepancy was observed in highly conserved craniofacial development genes, *ednraa* and *ednrab*. These genes are partially redundant in zebrafish, though knockdowns of *ednrab* knockdowns have shortened lower jaws and joint fusions (Nair et al., 2007). However, our analysis found that percomorph members did not express *ednrab* in pharyngeal arch cells. This discrepancy suggests differential subfunction partitioning between paralogs in the two lineages (Force et al., 2005). Beyond the biological finding, this observation highlights that the loss of *ednrab* expression in Gulf pipefish PA cells could have been mis-interpreted as developmentally relevant if threespine stickleback had not been included in the analysis.

Although our analysis did not find evidence for global GRN re-wiring in syngnathids, the observed phylogenetic signal is indicative of developmental systems drift (Haag & True, 2021). Under this principle, changes in development can lead to similar phenotypic outcomes. Therefore, teasing out genetic changes unrelated to morphological evolution is critical for removing false positives from evo-devo studies. It, further, suggests that cross species scRNAseq is challenging to interpret without phylogenetic sampling.

Local gene regulatory network re-wiring through co-option of *fgf23* in pipefish

The Fgf signaling pathway is a highly conserved, pleiotropic pathway active in craniofacial, somite, limb, and neural development in vertebrates (Ornitz & Itoh, 2015). Our dataset revealed a surprising deviation from this pattern in Gulf pipefish endocrine Fgf ligands (primarily *fgf23* with minimal *fgf19* dosage). Although *fgf23* in model systems is involved with the regulation of calcium and phosphate metabolism by kidneys and bone (Fukumoto & Yamashita, 2007; Huitema et al., 2012; Lin et al., 2017), we detected it in pipefish PA CNC, mesenchymal, somite, and retinal cells. The sequence of *fgf23* is conserved in syngnathids, suggesting maintenance of binding affinities, but loss of *tigarb* around the locus may have led to

these novel expression domains. Based on that hypothesis, *fgf23* from the basal syngnathid, *doryrhampus excisus,* may lack PA expression because it has retained *tigarb* (Stiller et al., 2022).

These novel *fgf23* expression domains may represent a shift from endocrine to paracrine signaling. This gene is suspected to engage in paracrine or autocrine signaling in mouse osteoblasts due to expression of *fgf23* and its receptor complex (*alpha-klotho* and *fgfr1;* Rhee et al., 2011). However, our dataset found minimal klotho expression (<10% of cells within clusters), suggesting pipefish arch *fgf23* signal transduction does not occur via canonical the MAPK pathway. An alternative signaling mechanism for *fgf23* has been identified cardiac myocytes where phospholipase Cy acts as a co-factor and the calcineurin/NAFT pathway is activated. (Gottfried et al., 2018). Since phospholipase C genes are expressed in arch and mesenchymal cells, with *plcd3a* and *plcd4b* unique to pipefish arches, it is possible that *fgf23* is signaling through this alternative pathway in developing pipefish. Although the role of *fgf23* on pipefish development is unknown, it is a case of gene co-option in GRNs of the PA, somites, and mesenchyme.

Evolution of redundancy in Fgf pathway may enable genome evolution

Our study proposed loss and gains in Fgf pathway redundancy in cypriniform and percomorph fishes are relevant for syngnathid evolution. Considering syngnathid's loss of *fgf3*, we examined the other members of its subfamily and detected PA expression of fgf*10a, 10b*, and *22* in stickleback, fgf*10a, 10b*, and *22* in pipefish, and *fgf10a* and *10b* in zebrafish. Based on ancestral expression of *fgf3* (CNC among other domains) and *fgf22* (neural), it is parsimonious to hypothesize that *fgf22* gained expression in percomorph PA cells (and fin, identified with *in situs*). In this case of local GRN re-wiring, *fgf3* may have become redundant in percomorph PA CNC cells with *fgf22* which could have enabled the non-lethal loss of *fgf3* in syngnathids.

In the zebrafish, cell proliferation in the midbrain and defective roof plate formation and tectum specification is decreased in *fgf22* knockdowns (Miyake & Itoh, 2013). We propose *fgf22* may have co-opted similar functions in stickleback and pipefish pharyngeal arches and fins. Indeed, *fgf22* is expressed in the proliferative mesenchyme medial to the apical ectodermal ridge in stickleback and pipefish pectoral fins. This expression mirrors *fgf10a's* fin domains, suggesting co-option and a similar role regulating limb outgrowth (Fischer et al., 2003; Ng et al., 2002).

Although *fgf22* was weakly detected in stickleback pharyngeal arches, we observed strong expression in the gill arches and around the periphery of arches one and two in Gulf pipefish. Our data, further, found a higher abundance and average expression of *fgf22* in Gulf pipefish CNC PA cells than *fgf22* and *fgf3* in stickleback. Higher *fgf22* expression may lead to increased proliferation in these arches in Gulf pipefish. In concordance with this hypothesis, we found that Gulf pipefish had a greater number of CNC derived PA cells than stickleback and zebrafish. Although it is possible that this may be a false positive given differences in scRNAseq dissociations, these findings warrant follow-up studies. The second arch, in particular, forms highly derived cartilages in syngnathids, hyoid elements. Therefore, teasing out the role of *fgf22* on CNC proliferation in arch 2 may provide an important insight into the development of their derived skeleton (Kimmel et al., 2001; Leysen et al., 2010).

Although *fgf22* is an example of where redundancy was gained in the Fgf pathway in the percomorph lineage, there are other cases where redundancy is lost. We suggest that *fgfr1* paralogs have decreased redundancy in percomorphs. The gene *fgfr1* has two paralogs in teleost fishes, which each have two splice variants (1b and 1c) that vary in fgf affinity. We noticed that percomorphs are missing the exon for the 1b splice variant in *fgfr1b*, and syngnathus fishes are missing the *fgfr1b* gene completely (Small et al., 2022). Although *fgfr1a* and *fgfr1b* are redundant in zebrafish, the loss of *fgfr1a* in medaka is lethal (Leerberg et al., 2019; Yokoi et al., 2007). The loss of the *1b* splice variant in *fgfr1b* likely leads to heavily reduced binding of fgf3, 7, and 8 by fgfr1b, possibly explaining the detrimental effects of *fgfr1a* gene loss in medaka (Zhang et al., 2006).

Combined, these evolutionary (syngnathids) and lab (zebrafish and medaka) knockouts suggest varying redundancy in the Fgf signaling pathway across teleosts. Redundancy in GRNs increases robustness against perturbations, illustrated by syngnathids' PA 3-7 maintenance despite *fgf3* loss (Cooke et al., 1997). However, there is an interplay between developmental robustness and evolvability (Chen & Lin, 2011). For instance, *fgfr1b* in cichlids, which may have also lost the 1b splice variant, is a top candidate for scale shape evolution (Albertson et al., 2018). Observed variability in Fgf pathway robustness suggests on a broad scale that speculation of phenotypic outcome from gene loss events requires developmental context.

References

Aktary, Z., Conde-Perez, A., Rambow, F., Di Marco, M., Amblard, F., Hurbain, I., Raposo, G., Delevoye, C., Coscoy, S., & Larue, L. (2021). A role for Dynlt3 in melanosome movement, distribution, acidity and transfer. *Communications Biology 2021 4:1*, *4*(1), 1–15.

Albertson, R. C., Kawasaki, K. C., Tetrault, E. R., & Powder, K. E. (2018). Genetic analyses in Lake Malawi cichlids identify new roles for Fgf signaling in scale shape variation. *Communications Biology*, *1*(1), 1–11.

Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M., & Postlethwait, J. H. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science*, *282*(5394), 1711–1714.

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., … Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature 2020 587:7833*, *587*(7833), 246–251.

Arnegard, M. E., Zwickl, D. J., Lu, Y., & Zakon, H. H. (2010). Old gene duplication facilitates origin and diversification of an innovative communication system - Twice. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(51), 22172–22177.

Baanannou, A., Rastegar, S., Bouzid, A., Takamiya, M., Gerber, V., Souissi, A., Beil, T., Jrad, O., Strähle, U., & Masmoudi, S. (2020). Gene duplication and functional divergence of the zebrafish otospiralin genes. *Development Genes and Evolution*, *230*(1), 27–36.

Baldera, D., Baxendale, S., van Hateren, N. J., Marzo, M., Glendenning, E., Geng, F. S., Yokoya, K., Knight, R. D., & Whitfield, T. T. (2023). Enhancer trap lines with GFP driven by smad6b and frizzled1 regulatory sequences for the study of epithelial morphogenesis in the developing zebrafish inner ear. *Journal of Anatomy*, *243*(1), 78–89.

Beretta, C. A., Brinkmann, I., & Carl, M. (2011). All four zebrafish Wnt7 genes are expressed during early brain development. *Gene Expression Patterns*, *11*(3–4), 277–284. https://doi.org/10.1016/J.GEP.2011.01.004

Bhullar, B. A. S., Morris, Z. S., Sefton, E. M., Tok, A., Tokita, M., Namkoong, B., Camacho, J., Burnham, D. A., & Abzhanov, A. (2015). A molecular mechanism for the origin of a key evolutionary innovation, the bird beak and palate, revealed by an integrative approach to

major transitions in vertebrate history. *Evolution*, *69*(7), 1665–1677.

Boroviak, T., Stirparo, G. G., Dietmann, S., Hernando-Herraez, I., Mohammed, H., Reik, W., Smith, A., Sasaki, E., Nichols, J., & Bertone, P. (2018). Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development (Cambridge)*, *145*(21).

Chen, B.-S., & Lin, Y.-P. (2011). On the Interplay between the Evolvability and Network Robustness in an Evolutionary Biological Network: A Systems Biology Approach. *Evolutionary Bioinformatics*, *7*(7), EBO.S8123.

Cooke, J., Nowak, M. A., Boerlijst, M., & Maynard-Smith, J. (1997). Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends in Genetics*, *13*(9), 360–364.

Cresko, W. A., Yan, Y. L., Baltrus, D. A., Amores, A., Singer, A., Rodríguez-Marí, A., & Postlethwait, J. H. (2003). Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Developmental Dynamics*, *228*(3), 480–489.

Crump, J. G., Maves, L., Lawson, N. D., Weinstein, B. M., & Kimmel, C. B. (2004). An essential role for Fgfs in endodermal pouch formation influences later craniofacial skeletal patterning. *Development*, *131*(22), 5703–5716.

Duncan, R. N., Panahi, S., Piotrowski, T., & Dorsky, R. I. (2015). Identification of Wnt Genes Expressed in Neural Progenitor Zones during Zebrafish Brain Development. *PLOS ONE*, *10*(12), e0145810.

Farnsworth, D. R., Saunders, L. M., & Miller, A. C. (2020). A single-cell transcriptome atlas for zebrafish development. *Developmental Biology*, *459*(2).

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., & Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, *360*(6392).

Fischer, S., Draper, B. W., & Neumann, C. J. (2003). The zebrafish fgf24 mutant identifies and additional level of Fgf signaling involved in vertebrate forelimb initiation. *Development*, *130*(15), 3515–3524.

Fish, J. L. (2016). Developmental mechanisms underlying variation in craniofacial disease and evolution. In *Developmental Biology* (Vol. 415, Issue 2, pp. 188–197). Academic Press Inc.

Fish, J. L. (2019). Evolvability of the vertebrate craniofacial skeleton. *Seminars in Cell &*

Developmental Biology, *91*, 13–22.

Force, A., Cresko, W. A., Pickett, F. B., Proulx, S. R., Amemiya, C., & Lynch, M. (2005). The origin of subfunctions and modular gene regulation. *Genetics*, *170*(1), 433–446.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y., & Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics*, *151*, 1531–1545.

Franke, M., De la Calle-Mustienes, E., Neto, A., Almuedo-Castillo, M., Irastorza-Azcarate, I., Acemel, R. D., Tena, J. J., Santos-Pereira, J. M., & Gómez-Skarmeta, J. L. (2021). CTCF knockout in zebrafish induces alterations in regulatory landscapes and developmental gene expression. *Nature Communications*, *12*.

Fuiten, A. M., & Cresko, W. A. (2021). Evolutionary divergence of a Hoxa2b hindbrain enhancer in syngnathids mimics results of functional assays. *Development Genes and Evolution*, *231*(3–4), 57–71.

Fukumoto, S., & Yamashita, T. (2007). FGF23 is a hormone-regulating phosphate metabolism-Unique biological characteristics of FGF23. *Bone*, *40*(5), 1190–1195.

Gottfried, R., De Borst, M., Leifheit-Nestler, M., & Haffner, D. (2018). *Paracrine effects of FGF23 on the Heart*. *9*, 1.

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., & Newman, A. M. (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, *367*(6476), 405–411.

Haag, E. S., & True, J. R. (2021). *Developmental System Drift*.

Healey, H. M., Bassham, S., & Cresko, W. A. (2022). Single-cell Iso-Sequencing enables rapid genome annotation for scRNAseq analysis. *Genetics*, *220*(3).

Healey, H. M., Penn, H. B., Small, C. M., Bassham, S., Goyal, V., Woods, M. A., & Cresko, W. A. (2024). Single Cell RNA Sequencing Provides Clues for the Developmental Genetic Basis of Syngnathidae's Evolutionary Adaptations. *ELife*, *13*.

Herzog, W., Sonntag, C., von der Hardt, S., Roehl, H. H., Varga, Z. M., & Hammerschmidt, M. (2004). Fgf3 signaling from the ventral diencephalon is required for early specification and subsequent survival of the zebrafish adenohypophysis. *Development*, *131*(15), 3681–3692.

Howe, D. G., Bradford, Y. M., Conlin, T., Eagle, A. E., Fashena, D., Frazer, K., Knight, J., Mani, P., Martin, R., Moxon, S. A. T., Paddock, H., Pich, C., Ramachandran, S., Ruef, B. J., Ruzicka, L., Schaper, K., Shao, X., Singer, A., Sprunger, B., … Westerfield, M. (2013). ZFIN, the Zebrafish Model Organism Database: Increased support for mutants and transgenics. *Nucleic Acids Research*, *41*(D1), D854–D860.

Huitema, L. F. A., Apschner, A., Logister, I., Spoorendonk, K. M., Bussmann, J., Hammond, C. L., & Schulte-Merker, S. (2012). Entpd5 is essential for skeletal mineralization and regulates phosphate homeostasis in zebrafish. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(52), 21372–21377.

Kimmel, C. B., Miller, C. T., & Moens, C. B. (2001). Specification and morphogenesis of the zebrafish larval head skeleton. In *Developmental Biology* (Vol. 233, Issue 2, pp. 239–257). Academic Press Inc.

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 1–13.

Lawson, N. D., Li, R., Shin, M., Grosse, A., Onur, Y. M. S., Stone, O. A., Kucukural, A., & Zhu, L. J. (2020). An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes. *ELife*, *9*, 1–76.

Leerberg, D. M., Hopton, R. E., & Draper, B. W. (2019). *Fibroblast Growth Factor Receptors Function Redundantly During Zebrafish Embryonic Development*.

Leysen, H., Jouk, P., Brunain, M., Christiaens, J., & Adriaens, D. (2010). Cranial architecture of tube-snouted gasterosteiformes (Syngnathus rostellatus and Hippocampus capensis). *Journal of Morphology*, *271*(3), 255–270.

Leysen, H., Roos, G., & Adriaens, D. (2011). Morphological variation in head shape of pipefishes and seahorses in relation to snout length and developmental growth. *Journal of Morphology*, *272*(10), 1259–1270.

Lin, C. H., Hu, H. J., & Hwang, P. P. (2017). Molecular Physiology of the Hypocalcemic Action of Fibroblast Growth Factor 23 in Zebrafish (Danio rerio). *Endocrinology*, *158*(5), 1347–

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *ELife*, *11*.

Melvin, V. S., Feng, W., Hernandez-Lagunas, L., Artinger, K. B., & Williams, T. (2013). A

morpholino-based screen to identify novel genes involved in craniofacial morphogenesis. *Developmental Dynamics*, *242*(7), 817–831.

Miyake, A., & Itoh, N. (2013). Fgf22 regulated by Fgf3/Fgf8 signaling is required for zebrafish midbrain development. *Biology Open*, *2*(5), 515–524.

Nair, S., Li, W., Cornell, R., & Schilling, T. F. (2007). Requirements for Endothelin type-A receptors and Endothelin-1 signaling in the facial ectoderm for the patterning of skeletogenic neural crest cells in zebrafish. *Development*, *134*(2), 335–345.

Ng, J. K., Kawakami, Y., Büscher, D., Raya, Á., Itoh, T., Koth, C. M., Esteban, C. R., Rodríguez-León, J., Garrity, D. M., Fishman, M. C., & Belmonte, J. C. I. (2002). The limb identity gene Tbx5 promotes limb initiation by interacting with Wnt2b and Fgf10. *Development*, *129*(22), 5161–5170.

Ornitz, D. M., & Itoh, N. (2015). The fibroblast growth factor signaling pathway. *Wiley Interdisciplinary Reviews: Developmental Biology*, *4*(3), 215–266.

Parker, J., Guslund, N. C., Jentoft, S., & Roth, O. (2022). Characterization of Pipefish Immune Cell Populations Through Single-Cell Transcriptomics. *Frontiers in Immunology*, *13*, 820152.

Peter, I., & Davidson, E. (2011). Evolution of Gene Regulatory Networks that Control Embryonic Development of the Body Plan. *Cell*, *144*(6), 970–985.

Powder, K. E., Cousin, H., McLinden, G. P., & Craig Albertson, R. (2014). A Nonsynonymous Mutation in the Transcriptional Regulator lbh Is Associated with Cichlid Craniofacial Adaptation and Neural Crest Cell Development. *Molecular Biology and Evolution*, *31*(12), 3113–3124.

Prescott, S. L., Srinivasan, R., Marchetto, M. C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F. H., Swigut, T., & Wysocka, J. (2015). Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell*, *163*(1), 68–83.

Qi, H., Schmöhl, F., Li, X., Qian, X., Tabler, C. T., Bennewitz, K., Sticht, C., Morgenstern, J., Fleming, T., Volk, N., Hausser, I., Heidenreich, E., Hell, R., Nawroth, P. P., & Kroll, J. (2021). Reduced Acrolein Detoxification in akr1a1a Zebrafish Mutants Causes Impaired Insulin Receptor Signaling and Microvascular Alterations. *Advanced Science*, *8*(18), 2101281.

Rhee, Y., Bivi, N., Farrow, E., Lezcano, V., Plotkin, L. I., White, K. E., & Bellido, T. (2011).

Parathyroid hormone receptor signaling in osteocytes increases the expression of fibroblast growth factor-23 in vitro and in vivo. *Bone*, *49*(4), 636–643.

Roth, O., Solbakken, M. H., Tørresen, O. K., Bayer, T., Matschiner, M., Baalsrud, H. T., Hoff, S. N. K., Brieuc, M. S. O., Haase, D., Hanel, R., Reusch, T. B. H., & Jentoft, S. (2020). Evolution of male pregnancy associated with remodeling of canonical vertebrate immunity in seahorses and pipefishes. *Proceedings of the National Academy of Sciences*, 201916251.

Sakaue, T., Maekawa, M., Nakayama, H., & Higashiyama, S. (2017). Prospect of divergent roles for the CUL3 system in vascular endothelial cell function and angiogenesis. *The Journal of Biochemistry*, *162*(4), 237–245.

Sauteur, L., Affolter, M., & Belting, H. G. (2017). Distinct and redundant functions of Esama and VE-cadherin during vascular morphogenesis. *Development (Cambridge)*, *144*(8), 1554–1565.

Shawi, M., & Serluca, F. C. (2008). Identification of a BMP7 homolog in zebrafish expressed in developing organ systems. *Gene Expression Patterns*, *8*(6), 369–375.

Small, C. M., Bassham, S., Catchen, J., Amores, A., Fuiten, A. M., Brown, R. S., Jones, A. G., & Cresko, W. A. (2016). The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome Biology*, *17*(1), 258. https://doi.org/10.1186/s13059-016-1126-6

Small, Clayton M., Healey, H. M., Currey, M. C., Beck, E. A., Catchen, J., Lin, A. S. P., Cresko, W. A., & Bassham, S. (2022). Leafy and weedy seadragon genomes connect genic and repetitive DNA features to the extravagant biology of syngnathid fishes. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(26), e2119602119.

Small, Clayton M, Healey, H. M., Currey, M. C., Beck, E. A., Catchen, J., P Lin, A. S., Cresko, W. A., Bassham, S., & Bassham, S. (2021). Leafy and Weedy Seadragon Genomes Connect Genic and Repetitive DNA Features to the Extravagant Biology of Syngnathid Fishes. *BioRxiv*, 2021.09.24.461757.

Smith, C. E. L., Whitehouse, L. L. E., Poulter, J. A., Wilkinson Hewitt, L., Nadat, F., Jackson, B. R., Manfield, I. W., Edwards, T. A., Rodd, H. D., Inglehearn, C. F., & Mighell, A. J. (2020). A missense variant in specificity protein 6 (SP6) is associated with amelogenesis imperfecta. *Human Molecular Genetics*, *29*(9), 1417–1425.

Stiller, J., Short, G., Hamilton, H., Saarman, N., Longo, S., Wainwright, P., Rouse, G. W., &

Simison, W. B. (2022). Phylogenomic analysis of Syngnathidae reveals novel relationships, origins of endemic diversity and variable diversification rates. *BMC Biology 2022 20:1*, *20*(1), 1–21.

Sur, A., Wang, Y., Capar, P., Margolin, G., & Farrell, J. A. (2023). Single-cell analysis of shared signatures and transcriptional diversity during zebrafish development. *BioRxiv : The Preprint Server for Biology*.

Thisse, B., & Thisse, C. (2004). *Fast Release Clones: A High Throughput Expression Analysis*. ZFIN Direct Data Submission.

Thisse, C., & Thisse, B. (2007). High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nature Protocols 2008 3:1*, *3*(1), 59–69.

Ton, M. L. N., Keitley, D., Theeuwes, B., Guibentif, C., Ahnfelt-Rønne, J., Andreassen, T. K., Calero-Nieto, F. J., Imaz-Rosshandler, I., Pijuan-Sala, B., Nichols, J., Benito-Gutiérrez, È., Marioni, J. C., & Göttgens, B. (2023). An atlas of rabbit development as a model for single-cell comparative genomics. *Nature Cell Biology 2023 25:7*, *25*(7), 1061–1072.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1), 1–32.

Van Wassenbergh, S., Roos, G., Genbrugge, A., Leysen, H., Aerts, P., Adriaens, D., & Herrel, A. (2009). Suction is kid's play: extremely fast suction in newborn seahorses. *Biology Letters*, *5*(2), 200–203.

Vogt, D. L., Gray, C. D., Young, W. S., Orellana, S. A., & Malouf, A. T. (2007). ARHGAP4 is a novel RhoGAP that mediates inhibition of cell motility and axon outgrowth. *Molecular and Cellular Neuroscience*, *36*(3), 332–342.

Wang, J., Sun, H., Jiang, M., Li, J., Zhang, P., Chen, H., Mei, Y., Fei, L., Lai, S., Han, X., Song, X., Xu, S., Chen, M., Ouyang, H., Zhang, D., Yuan, G. C., & Guo, G. (2021). Tracing cell-type evolution by cross-species comparison of cell atlases. *Cell Reports*, *34*(9), 108803.

Yokoi, H., Shimada, A., Carl, M., Takashima, S., Kobayashi, D., Narita, T., Jindo, T., Kimura, T., Kitagawa, T., Kage, T., Sawada, A., Naruse, K., Asakawa, S., Shimizu, N., Mitani, H., Shima, A., Tsutsumi, M., Hori, H., Wittbrodt, J., … Takeda, H. (2007). Mutant analyses reveal different functions of fgfr1 in medaka and zebrafish despite conserved ligand-receptor relationships. *Developmental Biology*, *304*(1), 326–337.

Zhang, X., Ibrahimi, O. A., Olsen, S. K., Umemori, H., Mohammadi, M., & Ornitz, D. M. (2006). Receptor Specificity of the Fibroblast Growth Factor Family: THE COMPLETE MAMMALIAN FGF FAMILY*. *The Journal of Biological Chemistry*, *281*(23), 15694.

CHAPTER V
CONCLUSION

Evolutionary novelties and derived traits form the basis of clade evolution. From the evolution of jaws to placentas to electric organs, these novel traits fundamentally altered organisms' niches. Although uncovering the developmental genetic bases of these traits is relevant for understanding human disease etiology as well as evolutionary trajectories, these studies are challenging. Teasing apart macroevolutionary changes requires careful comparative studies, examining overall morphology, cellular properties, genomes, transcriptomes, and epigenomes. In this thesis, I added a new tool to the evo-devo toolkit (single cell RNA sequencing) and illustrated its value in understanding developmental origins of derived traits.

In chapter II, I identified limitations of scRNAseq in non-traditional organisms and overcame these challenges. Using threespine stickleback, I found that incomplete gene model annotations impaired scRNAseq analysis, a particular problem in non-model systems, but these annotations could be lengthened using scISOseq. After these gene models were extended to include the 3' UTR, I increased the number of reads retained in the analysis by 26.5%. This manuscript provides a guide for researchers studying non-model systems to plan successful scRNAseq experiments and includes the first application of scISOseq in gene model annotation. Findings from this chapter enabled my following chapters by identifying effective sequencing methods and successful bioinformatic pipelines.

In chapter III, I produced the first syngnathid developmental scRNAseq atlas and analyzed the transcriptomic basis of derived traits. I created a scRNAseq atlas from Gulf pipefish at the onset of craniofacial elongation to examine craniofacial and dermal armor development as well as consider embryonic interactions with the brood pouch. During craniofacial elongation, my study found signaling genes (*bmp4* and *sfrpr1a*) and a transcription regulator (*prdm16*) expressed in mesenchyme above the lengthening ethmoid region. I probed for tooth germs using tooth development genes and found no signs of tooth bud initiation. However, two genes of these genes (*dlx3b* in early stages and *scpp1* as mineralization began) were expressed in developing dermal armor. Finally, I found evidence for nutrient exchange between embryos and the brood pouch environment. The examined pipefish evolutionary innovations were composed of recognizable cell types, suggesting derived features originate from changes within existing gene networks. Specifically for tooth development, our study gives support for the hypothesis that of

early tooth development gene losses were crucial for syngnathid's teeth loss. Beyond providing new hypotheses and avenues for syngnathid research, this atlas provides an amazing resource for syngnathid researchers.

In chapter IV, I presented a comparative scRNAseq analysis that identified unique changes in Gulf pipefish GRNs. This analysis identified overall gene network conservation between Gulf pipefish, threespine stickleback, and zebrafish during pharyngeal arch development. Gulf pipefish had several changes including novel *fgf23* expression in PA cells and loss of *fzd1.* Furthermore, pipefish and stickleback expressed *fgf22* in PA cells while zebrafish restricted *fgf22* expression to neural cells. We also found a higher number of PA cells in pipefish than the "short" snouted fish. These findings suggest syngnathid derived snouts evolved from select changes in gene networks which may have impacted cell proliferation or differentiation. Additionally, the *fgf22* finding indicates redundancy may have enabled syngnathid's loss of *fgf3*. These findings add to the growing evo-devo literature that has identified select GRN changes associated with derived traits (bird beaks, cichlid jaws, and more). My study provides a framework for future comparative scRNAseq analyses to identify the basis of derived traits.

Chapters III and IV identified signaling genes (*fgf23, fgf22, bmp4,* and *sfrp1a*) and transcription regulators (*prmd16*) expressed during syngnathid craniofacial development. However, it is still unclear which developmental pathways, Fgf, Wnt, and/or BMP, regulate craniofacial elongation. Future studies could tease apart the role of each pathway using chemical inhibitors to block pathway activity. Although the brood pouch presents a unique challenge to exposure, dosage could either occur during pregnancy, in embryos removed from the pouch, or with hatched larvae. These tests would determine whether blocking the Fgf, Wnt, and/or BMP pathways leads to a shortened snout, providing a fundamental understanding to genetic mechanisms involved with producing syngnathid's elongate snout.

In addition to functionally validating genomic candidates, future studies could expand my preliminary finding of increased pharyngeal arch cellular abundance in Gulf pipefish. Although cellular numbers can be predicted from scRNAseq data, the variability in species dissociations could lead to false differences. Prospective studies could quantify syngnathid CNC cell numbers, size, and distribution then compare findings with non-syngnathids. These analyses would reveal whether syngnathids have early changes in cellular attributes that influence their elongate skeleton. In addition to potential CNC changes, there may also be important alterations in cell

proliferation or morphology during craniofacial elongation which could be fleshed out through similar studies. Uncovering the cellular basis of elongation would illustrate whether there may be extended periods of cellular proliferation and/or differentiation in Syngnathid craniofacial development.

Beyond the evolution of syngnathid's derived snouts, this study presented a hypothesis (redundancy through *fgf22*) to understand the loss of *fgf3* in syngnathids. Prospective research could tease this idea apart through completing *fgf22* and *fgf3* knockouts in zebrafish and stickleback. These assays would determine whether the loss of *fgf3* would have the same developmental impacts in stickleback as in zebrafish (lethal) or whether it would have minimal impacts due to evolved redundancy with *fgf22*. Through testing this hypothesis, this work would suggest whether the loss of *fgf3* in syngnathids had any developmental impacts.

In this thesis, I demonstrated the challenges and power of scRNAseq in non-traditional model systems. Through addressing scRNAseq limitations in Chapter II, I provided a toolkit for researchers to use in their own model systems. Then, I used these tools to study the evolution and development of syngnathid's derived traits in Chapters III and IV. These projects identified novel candidates for their elongate snout, loss of teeth, male brooding, and dermal armor. This body of work found that syngnathids have conserved gene regulatory networks with slight modifications that may lead to derived features. Additionally, detection of pathway evolution within multiple teleosts suggests that applying findings from model systems to non-models requires taking context into account. Combined, these findings illustrate the importance of assessing genomic and phenotypic evolution through the lens of gene regulatory networks. Beyond providing novel insights for syngnathid research, I aim for my thesis to inspire future evo-devo studies uncovering the developmental basis of evolutionary innovations.

APPENDIX A

SUPPORTING INFORMATION FOR CHAPTER II

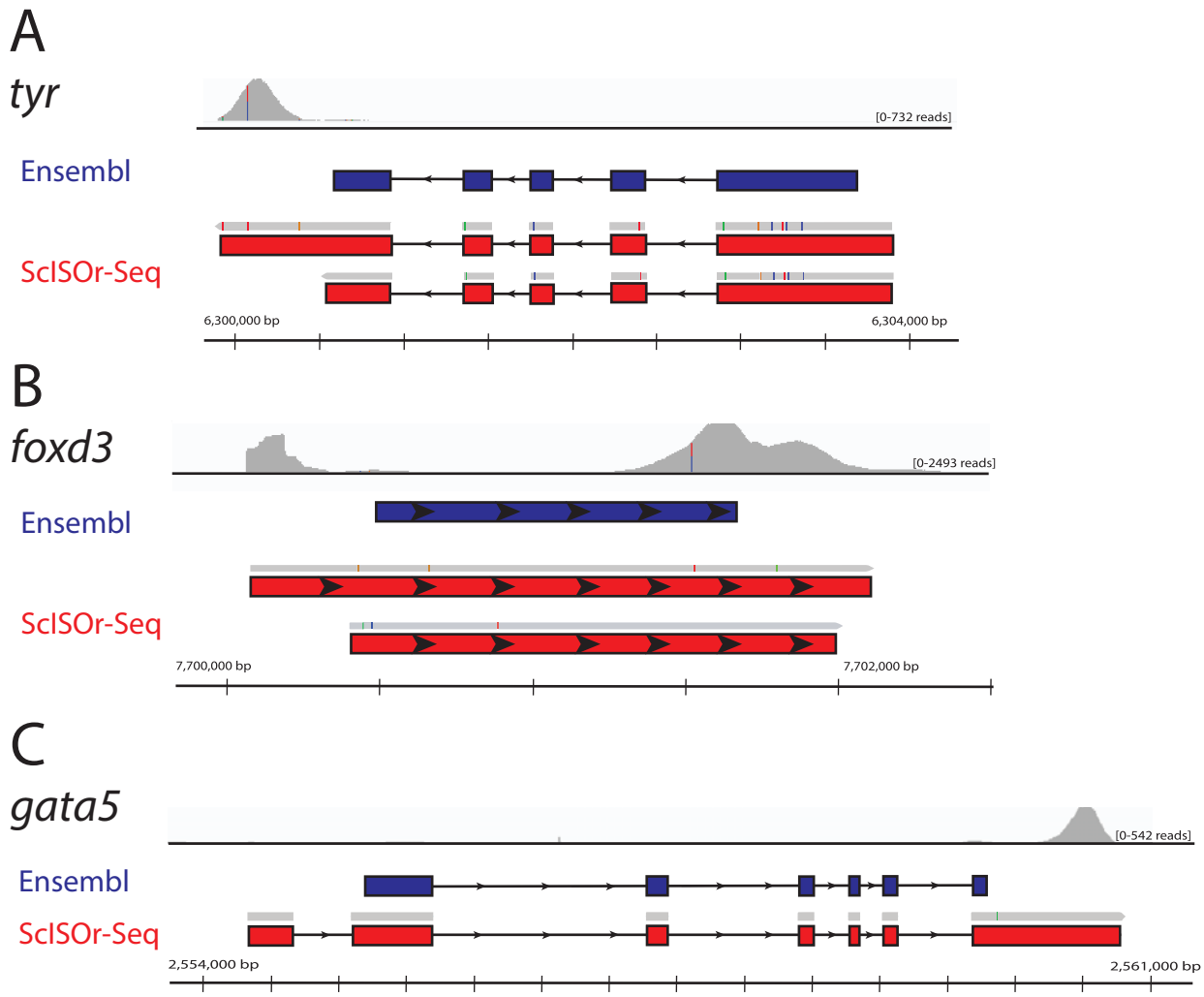


Figure S2.1: *ScISOr-Seq improved gene models by extending the 3' UTRs illustrated by tyr (ENSGACG00000007960), foxd3 (ENSGACG00000006933), and gata5 (ENSGACG00000003649).* scRNAseq reads are shown at the top for each gene in grey. The label to the right of the scRNAseq reads indicate the scale of the reads. The existing Ensembl models are shown in dark blue. The ScISOr-Seq reads (in grey) are shown above the respective gene models (in red) that they generated. Colored lines on the ScISOr-Seq and scRNAseq reads indicate a different base pair in the read than the stickleback reference genome (BROAD S1, 104.1 database version). The scale bar at the bottom indicates the position of each gene model. The ScISOr-Seq generated gene models capture the majority of the scRNAseq reads.

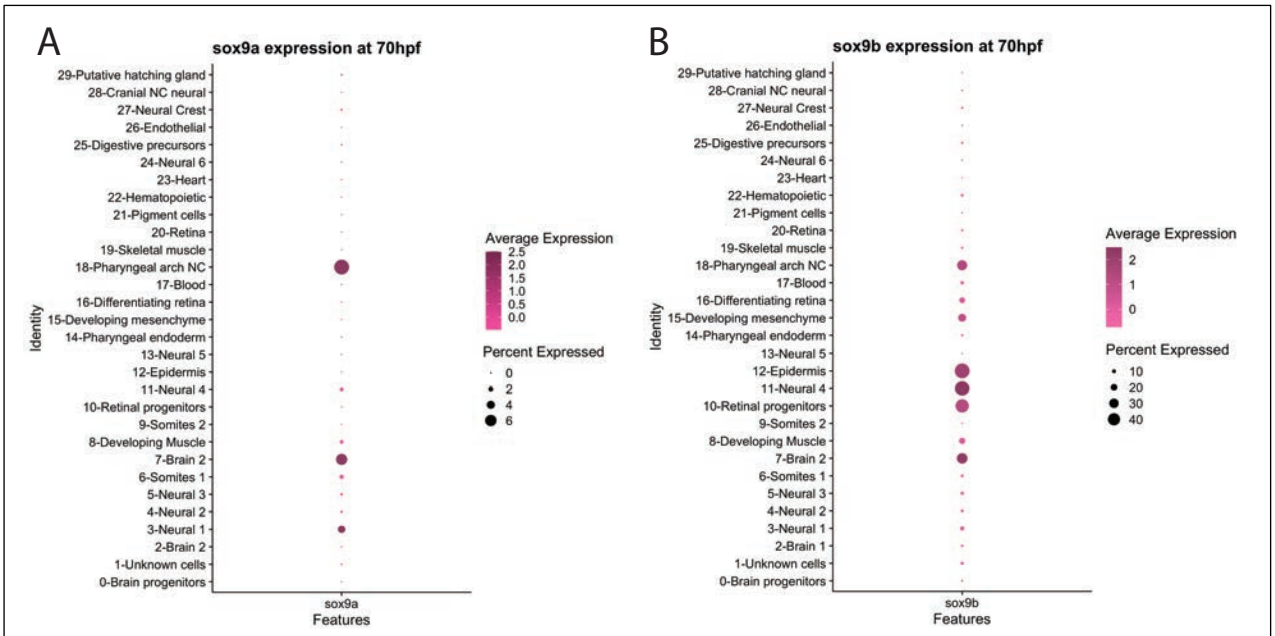


Figure S2.2: *Expression patterns of sox9a and sox9b recapitulate expression seen in Cresko et al. 2003*. The two paralogs are plotted on separate Dot plots due to an order of magnitude difference in expression. A) as observed in Cresko *et al.* 2003, *sox9a* is expressed in the pharyngeal arches, brain, and somites. B) We found *sox9b* expression in the pharyngeal arches, retina, brain, and weakly in somites, similar to Cresko *et al.* 2003.

APPENDIX B

SUPPORTING INFORMATION FOR CHAPTER III

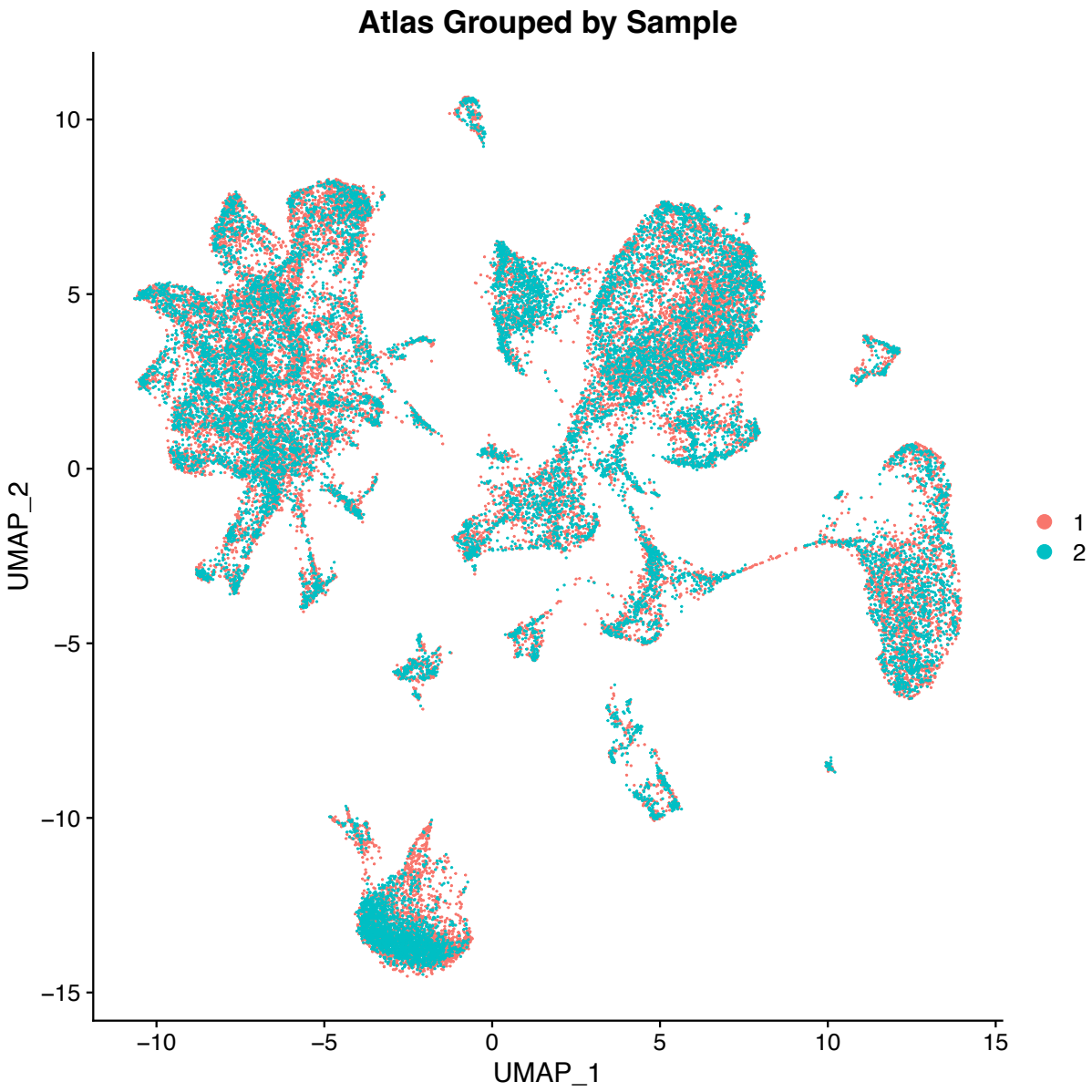


Figure S3.1: *Two Gulf pipefish samples have similar contributions to cell clusters.* This UMAP features cells plotted by sample (sample one in orange and sample two in cyan) in the first two UMAP dimensions.

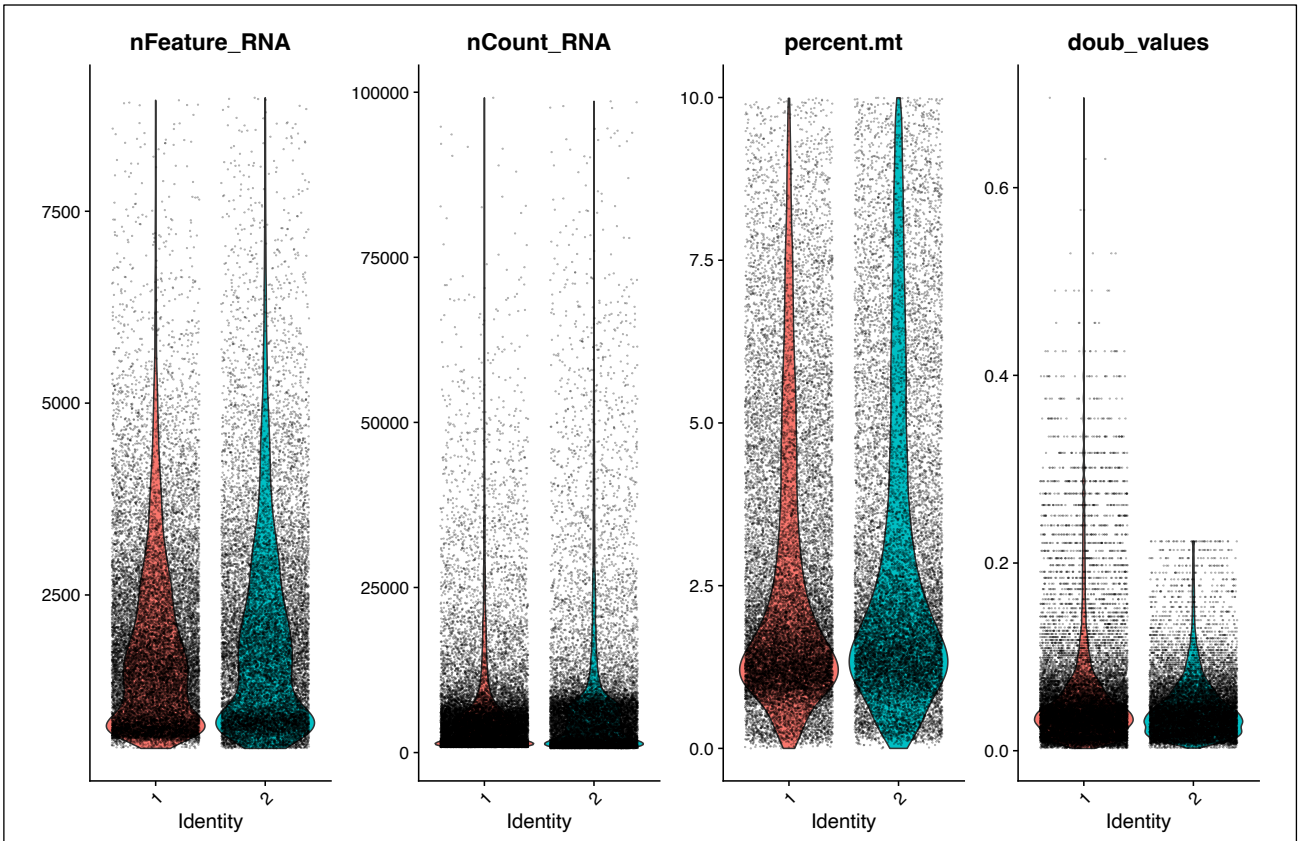


Figure S3.2: *QC metrics for both Gulf pipefish samples show similar sample quality after initial filtering.* Sample quality was assessed using the number of genes identified (nFeatureRNA), the total RNA counts (nCountRNA), the percentage of mitochondrial reads (percentmt), and the likelihood of being a doublet (doubvalues). Gulf pipefish sample data post-filtering is displayed for each of these metrics.

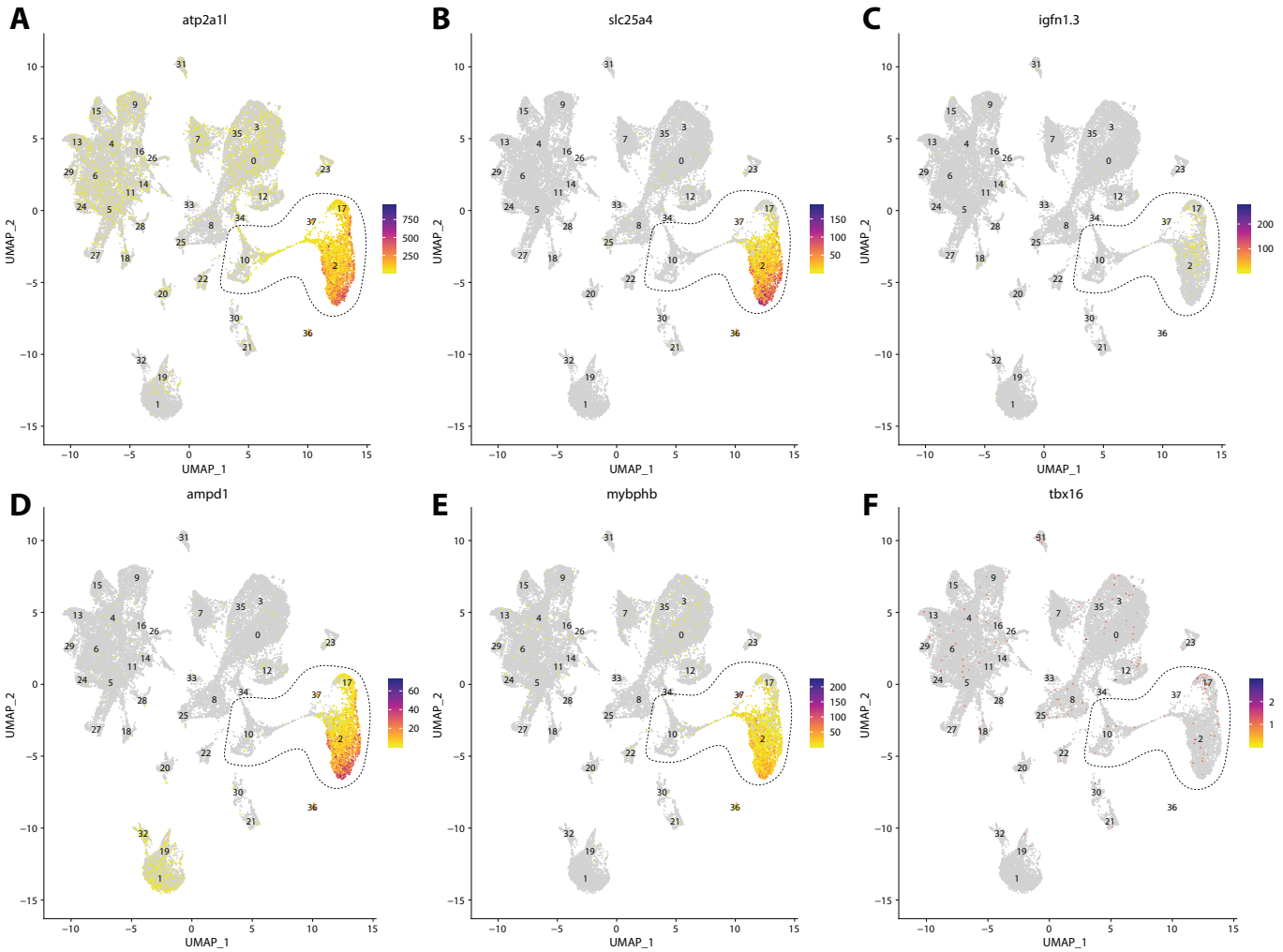


Figure S3.3: *Zebrafish muscle cell marker genes expression patterns correspond with annotated muscle cells in Gulf pipefish atlas*. Feature plots of zebrafish muscle cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated muscle cell clusters in pipefish are distinguished by a dashed outline.

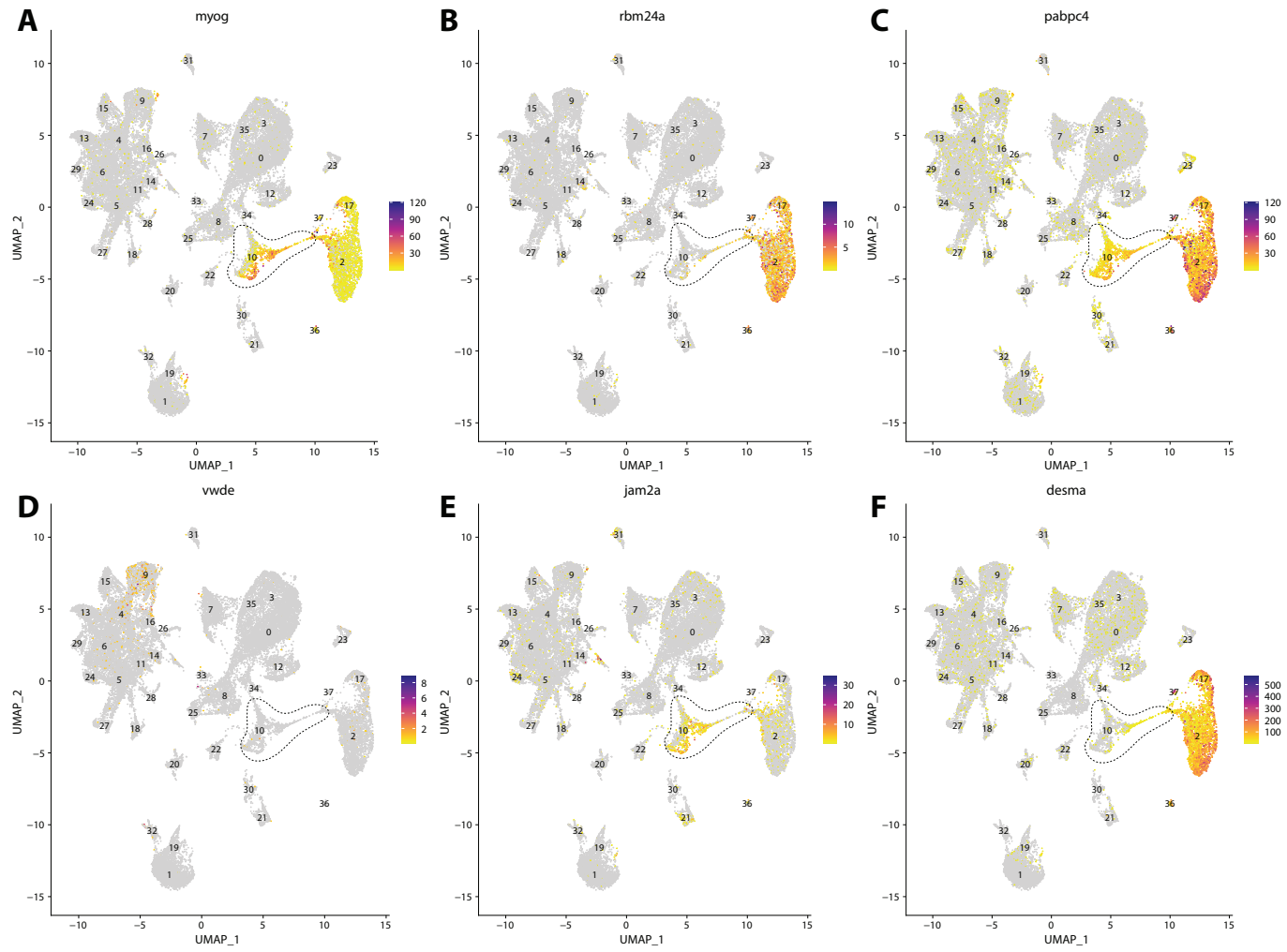


Figure S3.4: *Zebrafish muscle satellite cell marker genes expression patterns correspond with annotated muscle progenitor cells in Gulf pipefish atlas*. Feature plots of zebrafish muscle satellite cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated muscle progenitor cell cluster in pipefish is distinguished by a dashed outline.

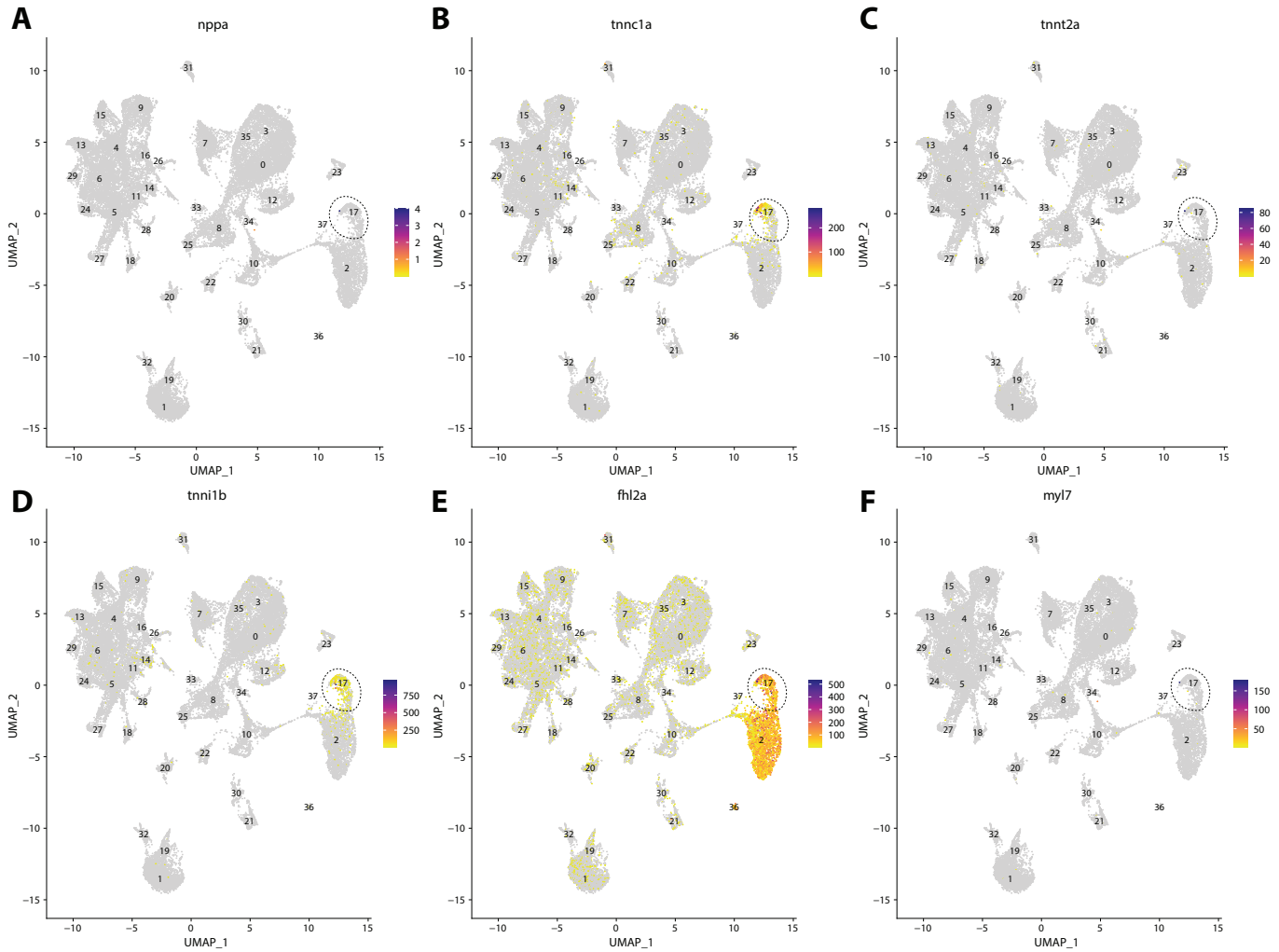


Figure S3.5: *Zebrafish cardiac muscle cell marker genes expression patterns correspond with annotated cardiac muscle cells in Gulf pipefish atlas*. Feature plots of zebrafish cardiac muscle cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated cardiac muscle cell cluster in pipefish is distinguished by a dashed outline.

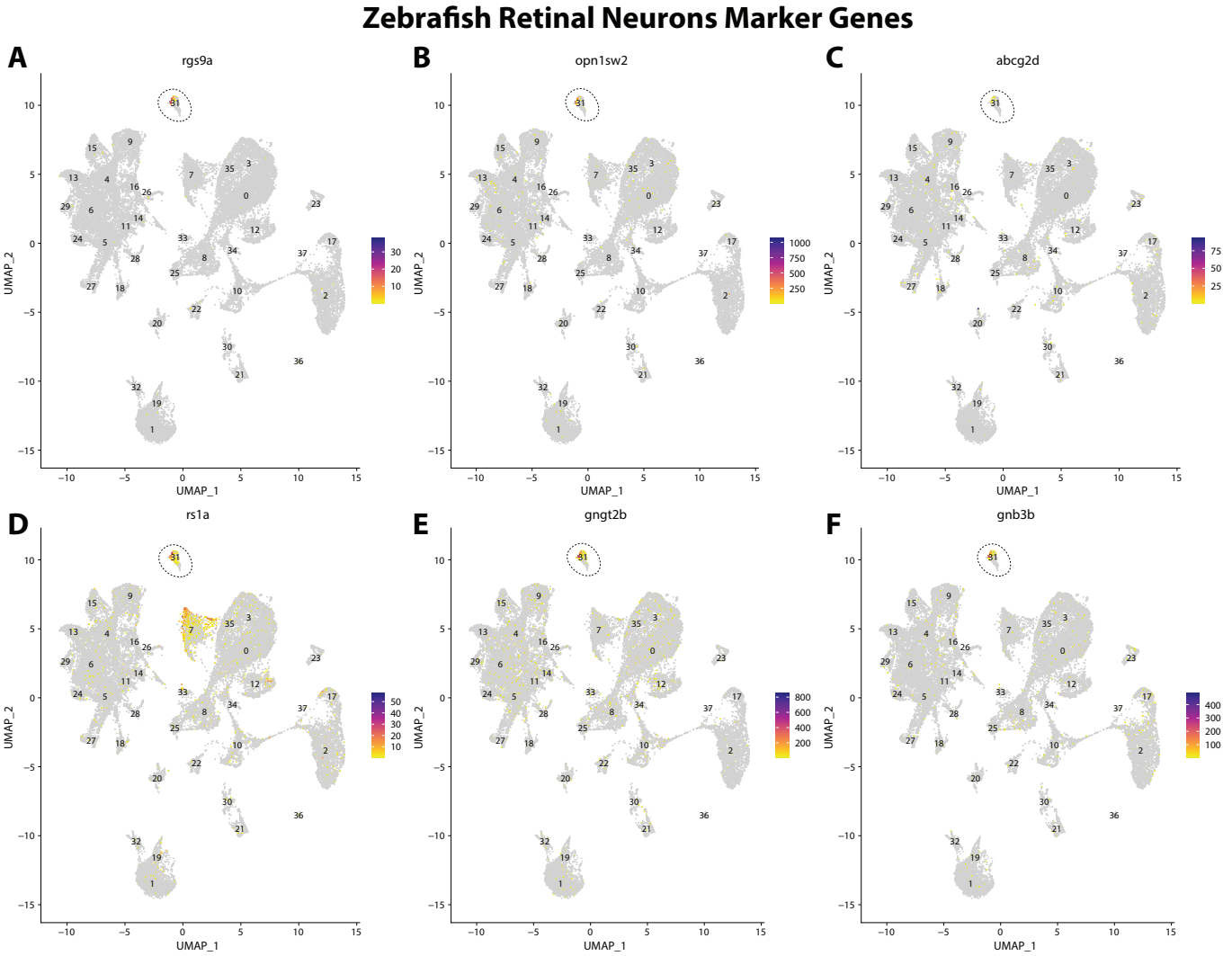


Figure S3.6: *Zebrafish retinal cell marker genes expression patterns correspond with annotated retinal cells in Gulf pipefish atlas.* Feature plots of zebrafish retinal cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated retinal cell cluster in pipefish is distinguished by a dashed outline.

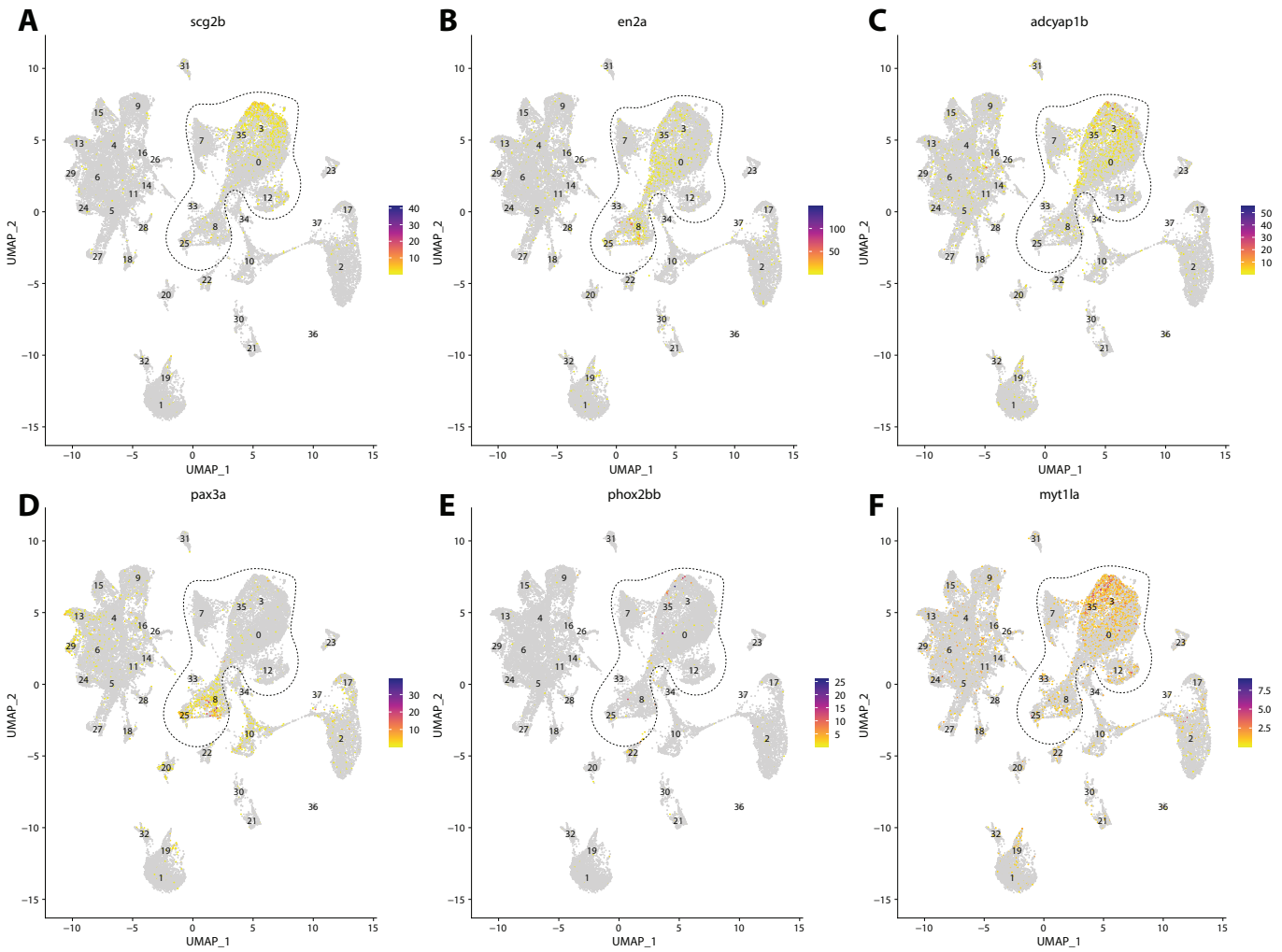


Figure S3.7: *Zebrafish neuronal cell marker genes expression patterns correspond with annotated neuronal cells in Gulf pipefish atlas.* Feature plots of zebrafish neuronal cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated neuronal cell clusters in pipefish are distinguished by a dashed outline.

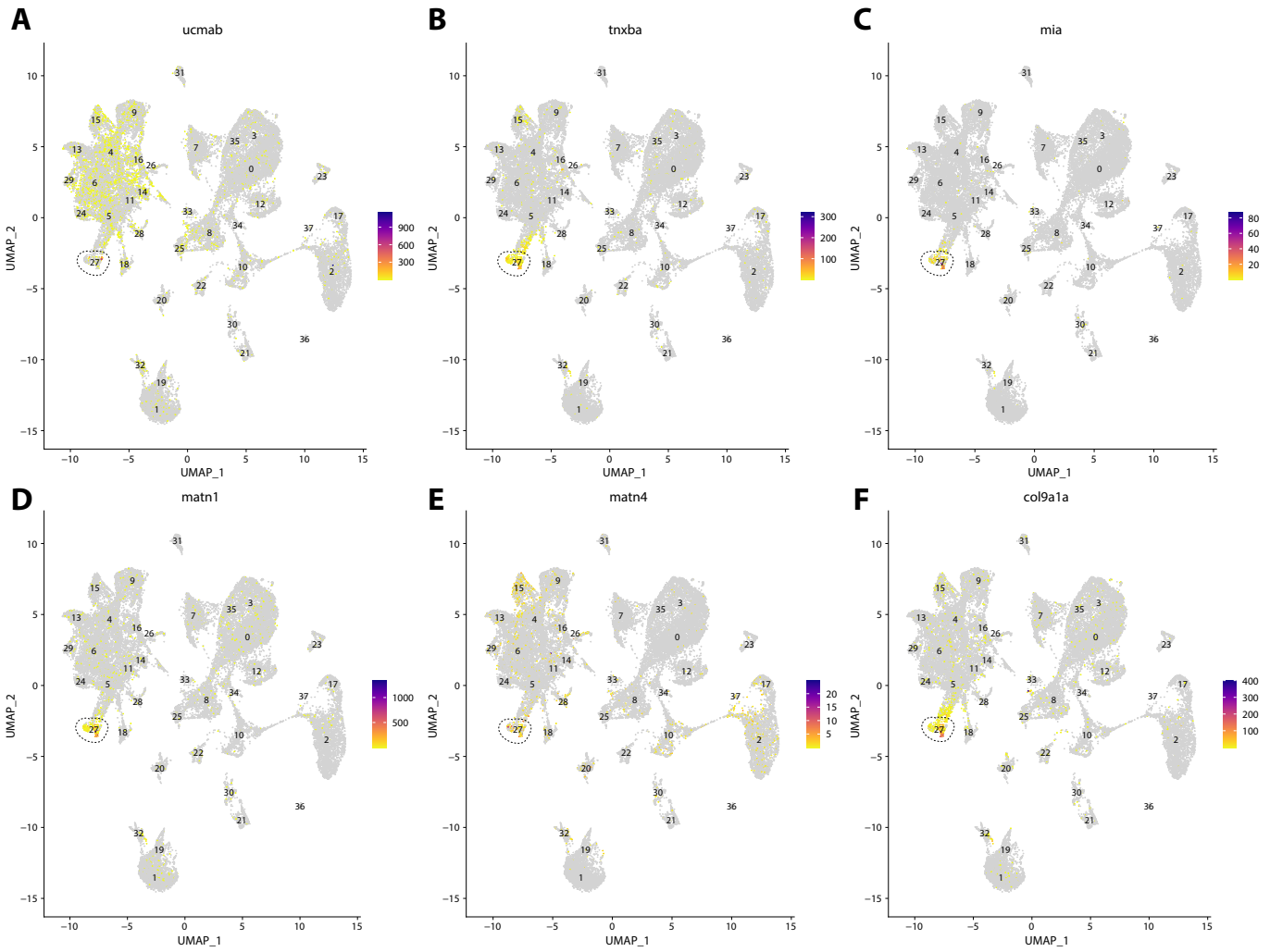


Figure S3.8: *Zebrafish chondroblast cell marker genes expression patterns correspond with annotated chondroblast cells in Gulf pipefish atlas.* Feature plots of zebrafish chondroblast cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated chondroblast cell cluster in pipefish is distinguished by a dashed outline.

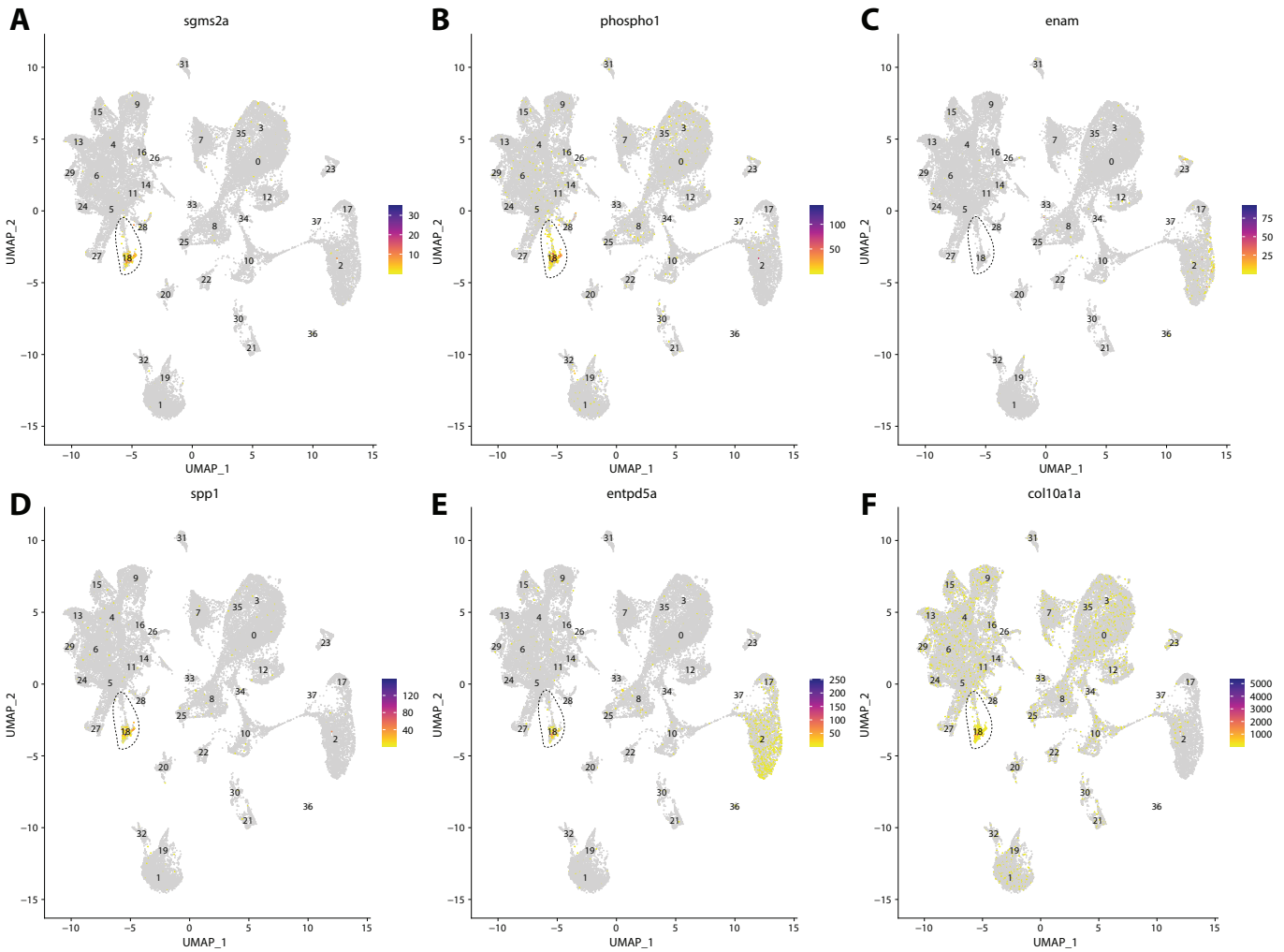


Figure S3.9: *Zebrafish osteoblast cell marker genes expression patterns correspond with annotated osteoblast cells in Gulf pipefish atlas.* Feature plots of zebrafish osteoblast cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated osteoblast cell cluster in pipefish is distinguished by a dashed outline.

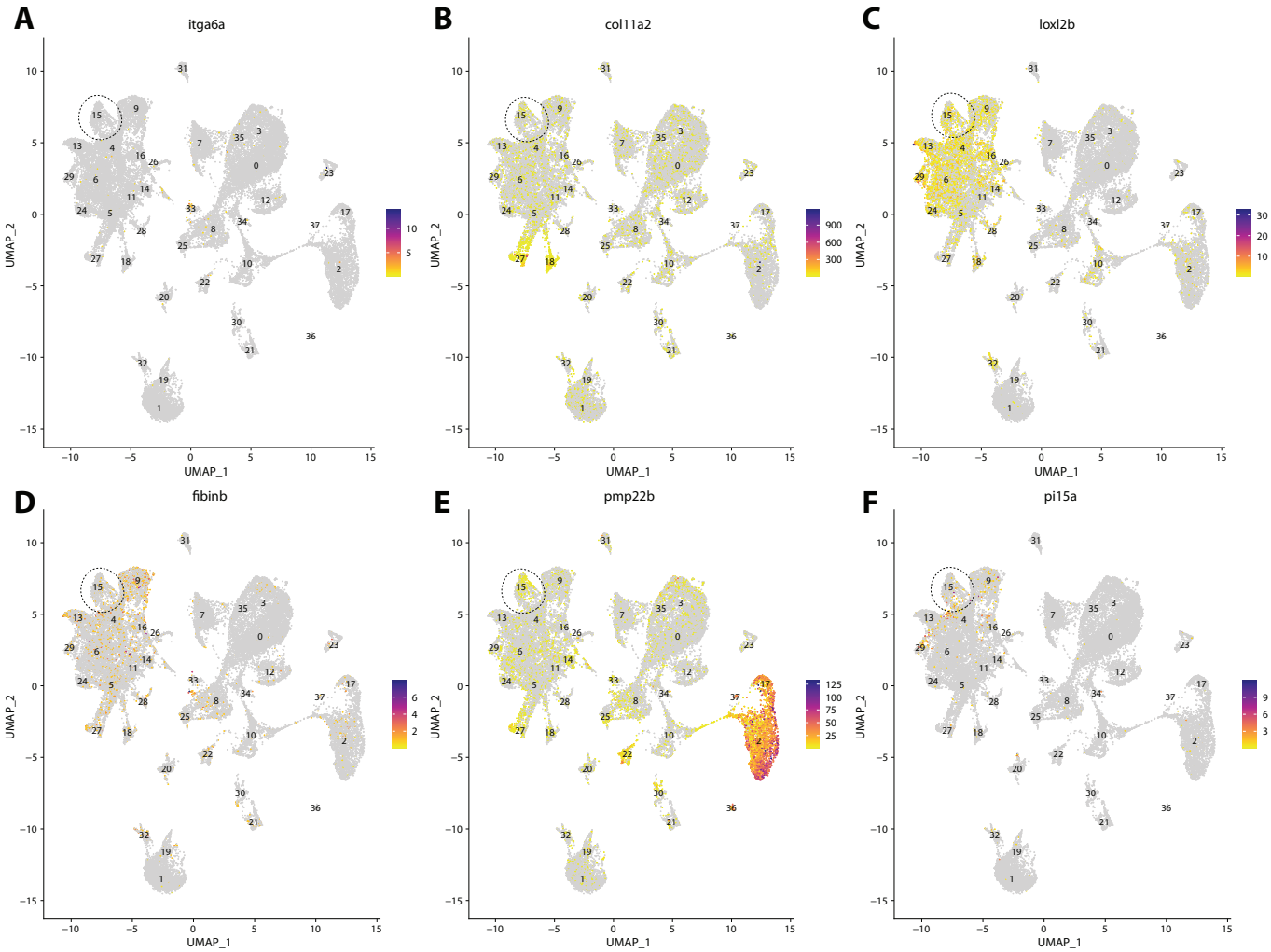


Figure S3.10: *Zebrafish notochord cell marker genes expression patterns correspond with annotated notochord cells in Gulf pipefish atlas.* Feature plots of zebrafish notochord cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated notochord cell cluster in pipefish is distinguished by a dashed outline.

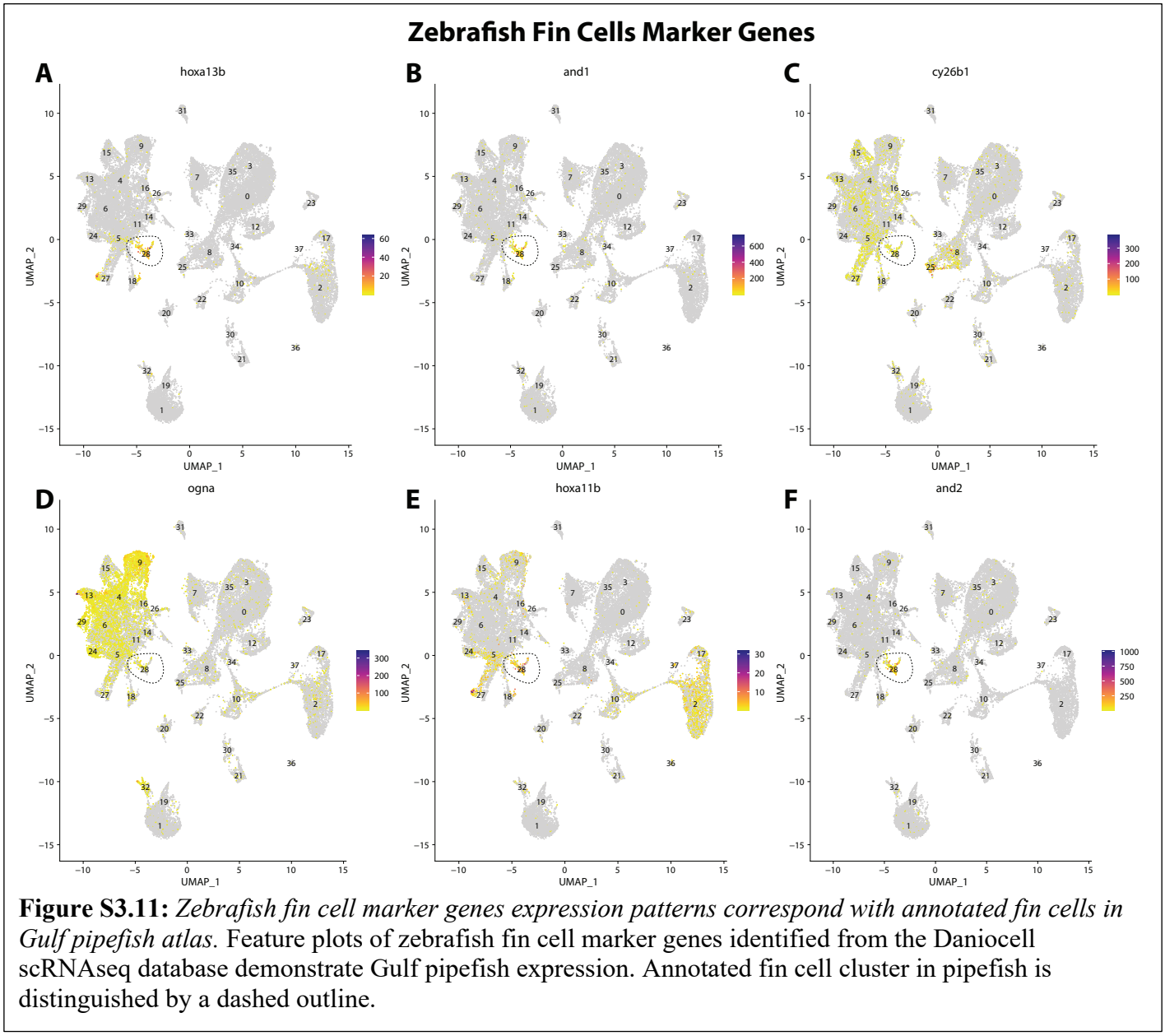


Figure S3.11: *Zebrafish fin cell marker genes expression patterns correspond with annotated fin cells in Gulf pipefish atlas.* Feature plots of zebrafish fin cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated fin cell cluster in pipefish is distinguished by a dashed outline.

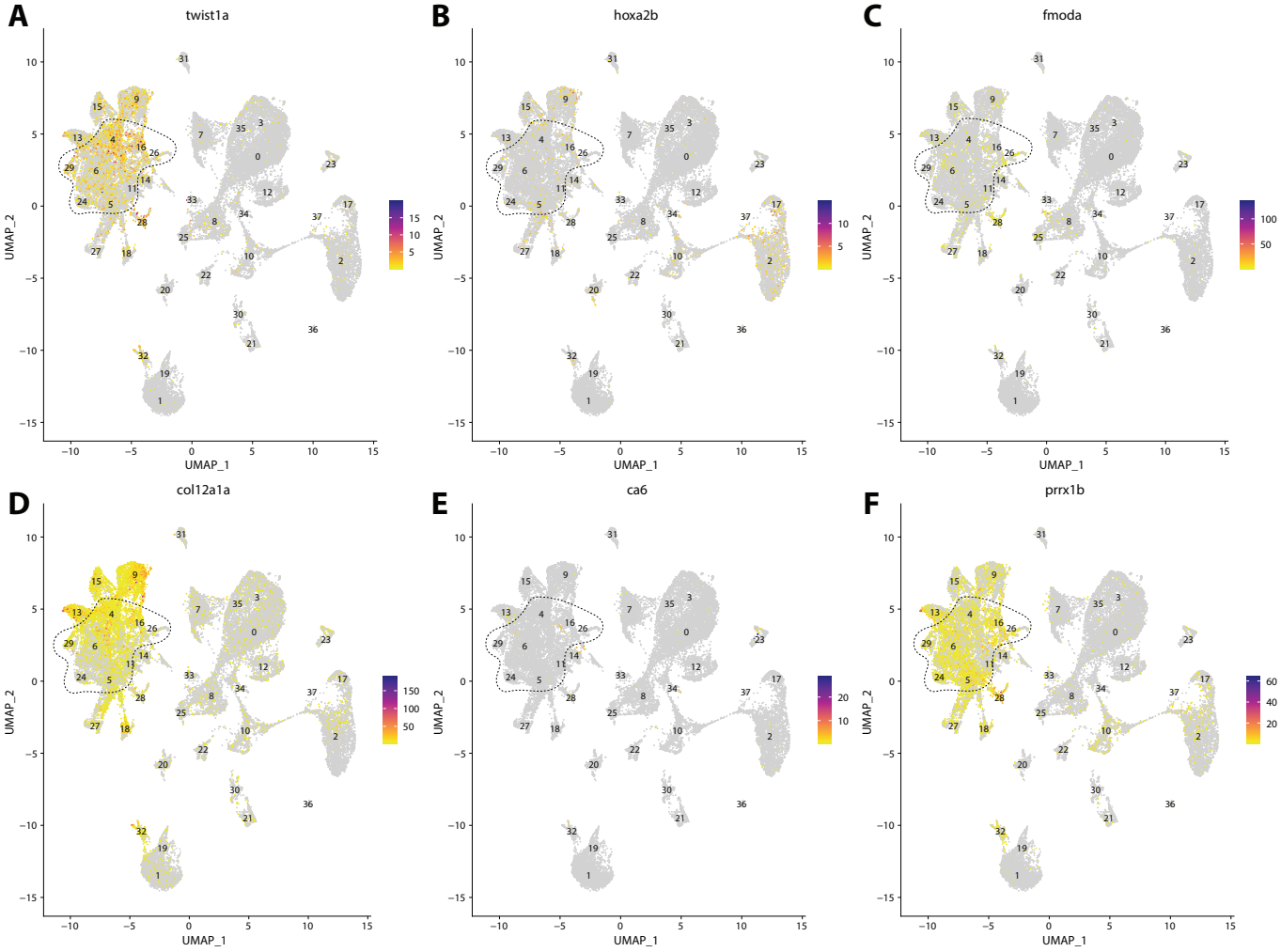


Figure S3.12: *Zebrafish mesenchymal cell marker genes expression patterns correspond with annotated mesenchymal cells in Gulf pipefish atlas.* Feature plots of zebrafish mesenchymal cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated mesenchymal cell clusters in pipefish are distinguished by a dashed outline.

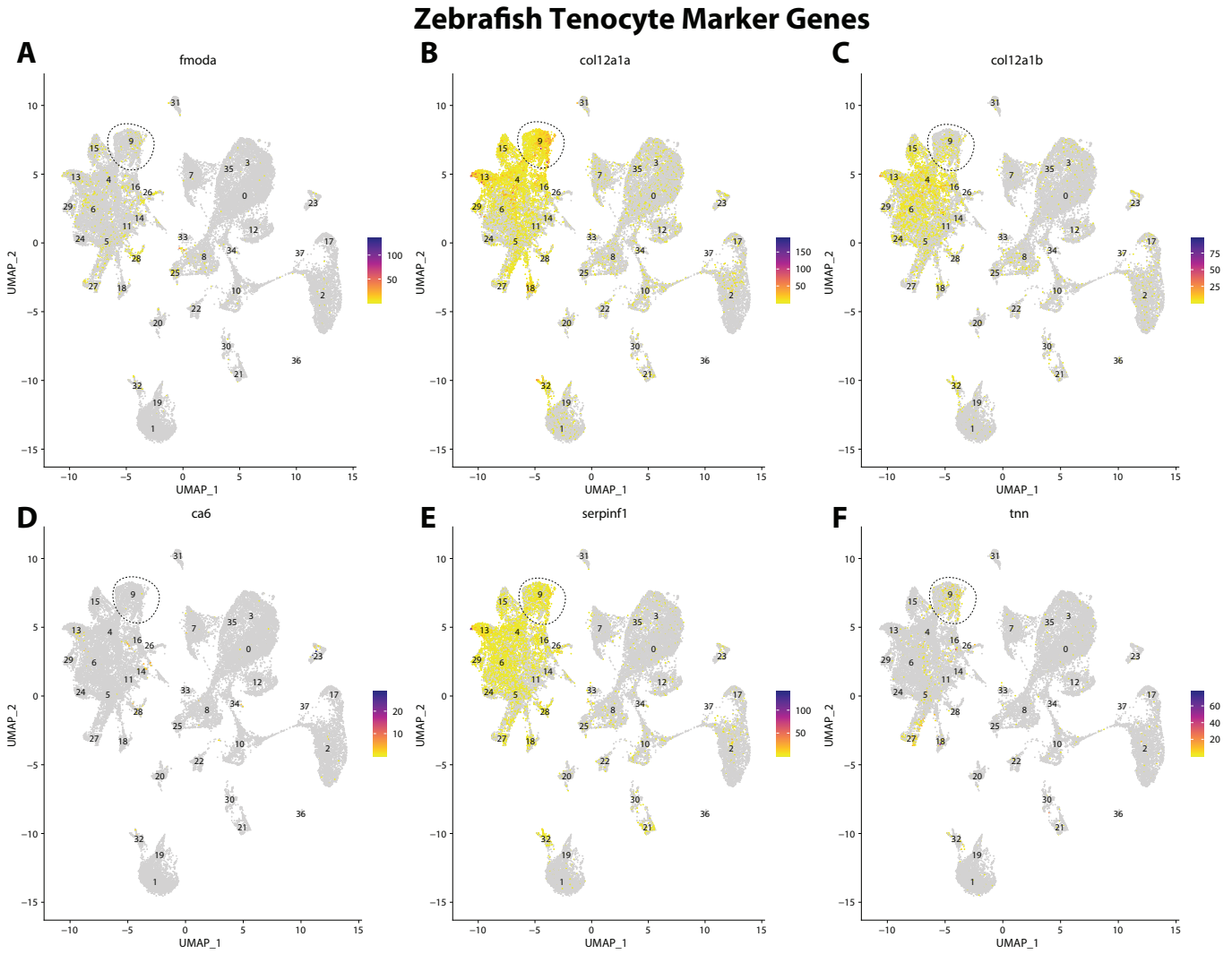


Figure S3.13: *Zebrafish tenocyte cell marker genes expression patterns correspond with annotated tenocyte cells in Gulf pipefish atlas.* Feature plots of zebrafish tenocyte cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated tenocyte cell cluster in pipefish is distinguished by a dashed outline.

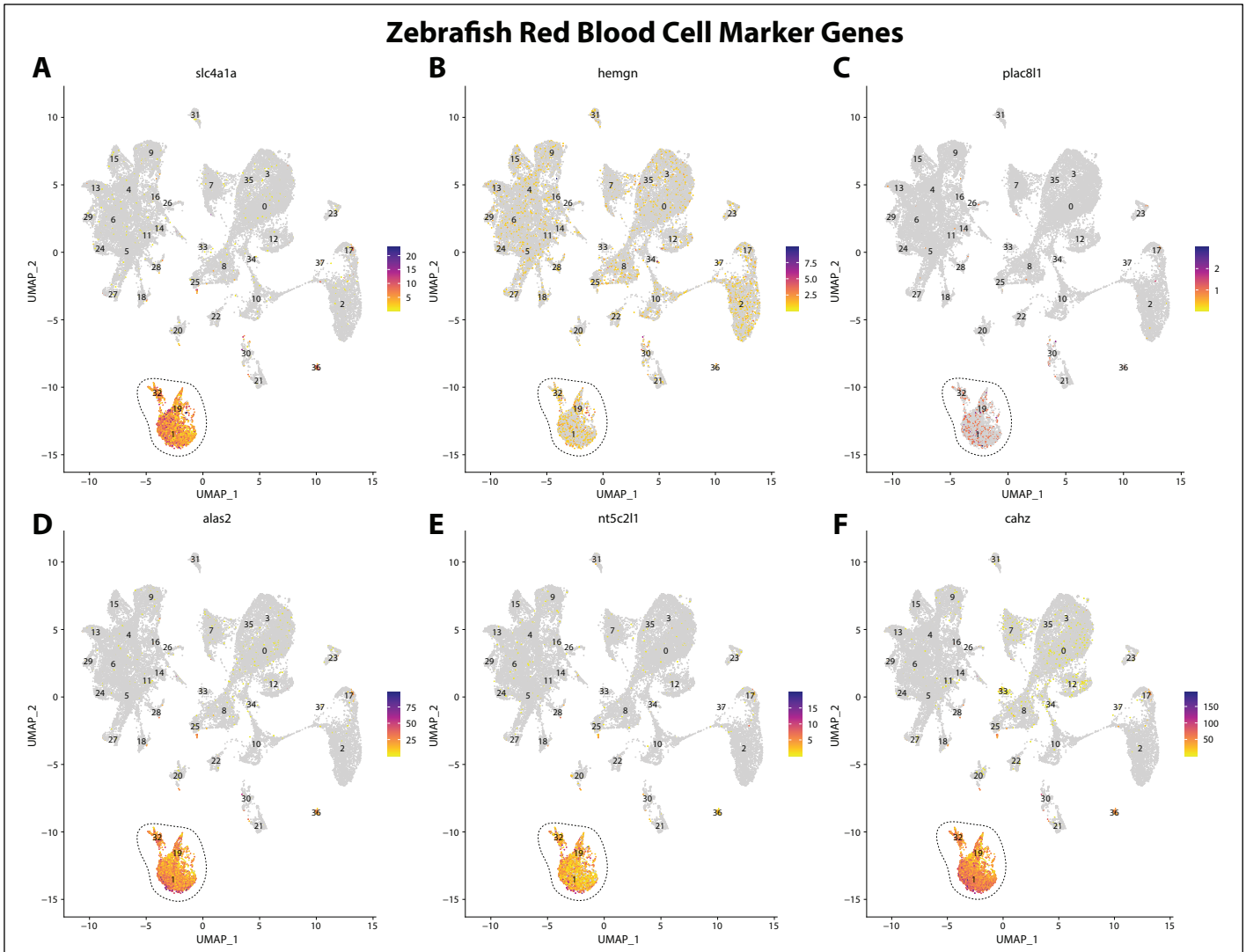


Figure S3.14: *Zebrafish erythrocyte cell marker genes expression patterns correspond with annotated erythrocyte cells in Gulf pipefish atlas.* Feature plots of zebrafish erythrocyte cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated erythrocyte cell clusters in pipefish are distinguished by a dashed outline.

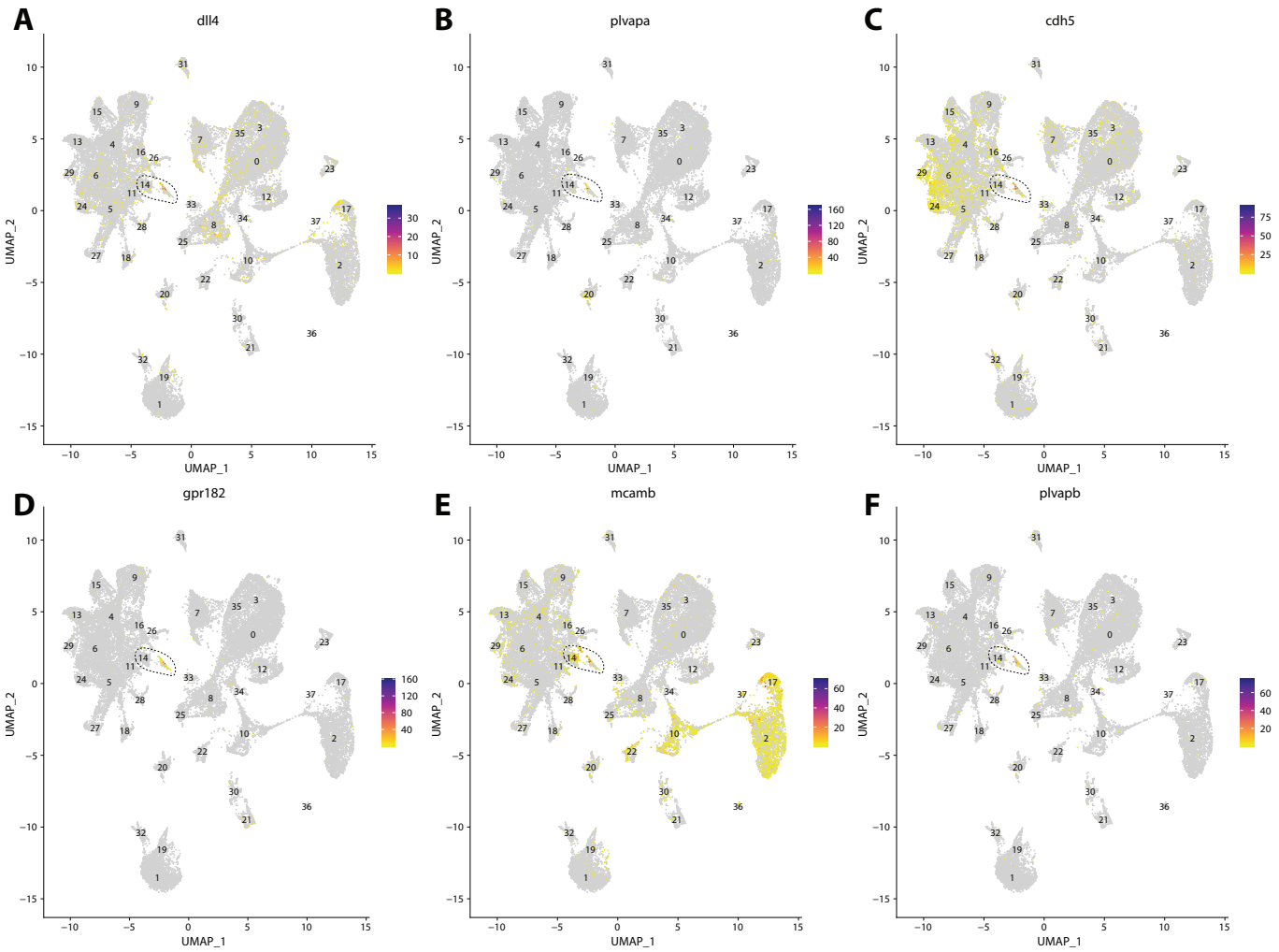


Figure S3.15: *Zebrafish endothelial cell marker genes expression patterns correspond with annotated endothelial cells in Gulf pipefish atlas.* Feature plots of zebrafish endothelial cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated endothelial cell cluster in pipefish is distinguished by a dashed outline.

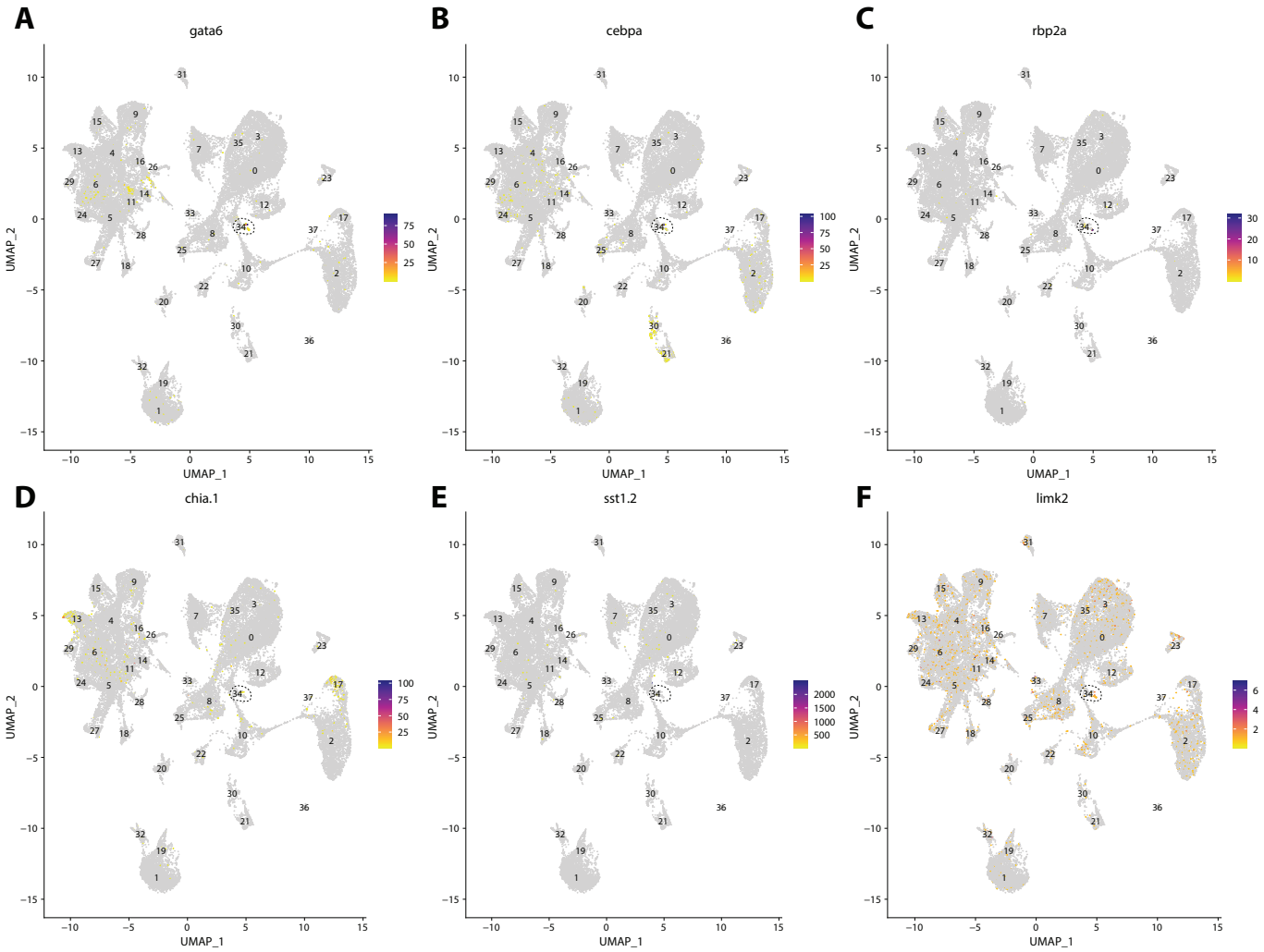


Figure S3.16: *Zebrafish gut and liver cell marker genes expression patterns correspond with annotated gut and liver cells in Gulf pipefish atlas.* Feature plots of zebrafish gut and liver cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated liver and gut cell cluster in pipefish is distinguished by a dashed outline.

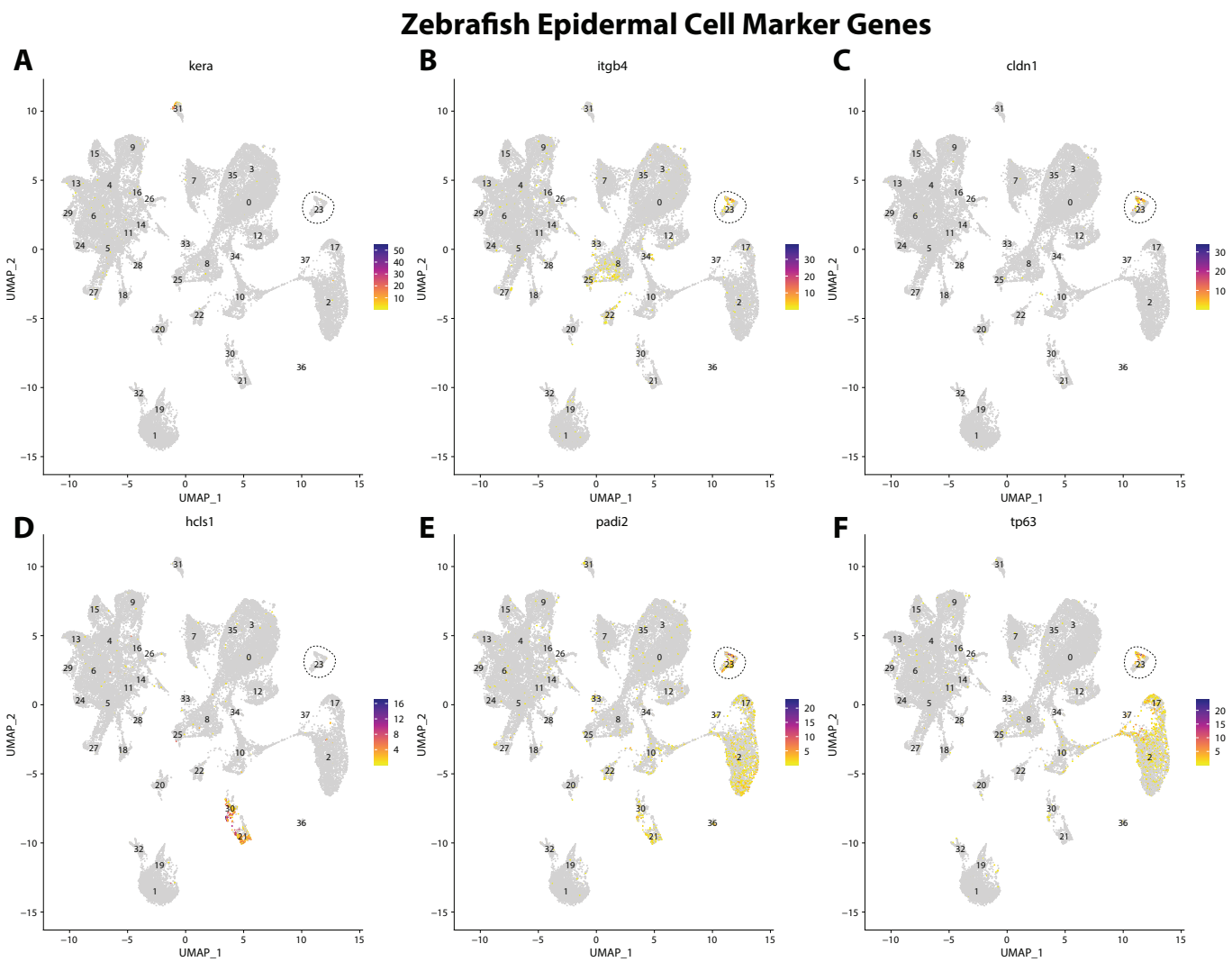


Figure S3.17: *Zebrafish epidermal cell marker genes expression patterns correspond with annotated epidermal cells in Gulf pipefish atlas.* Feature plots of zebrafish epidermal cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated epidermal cell cluster in pipefish is distinguished by a dashed outline.

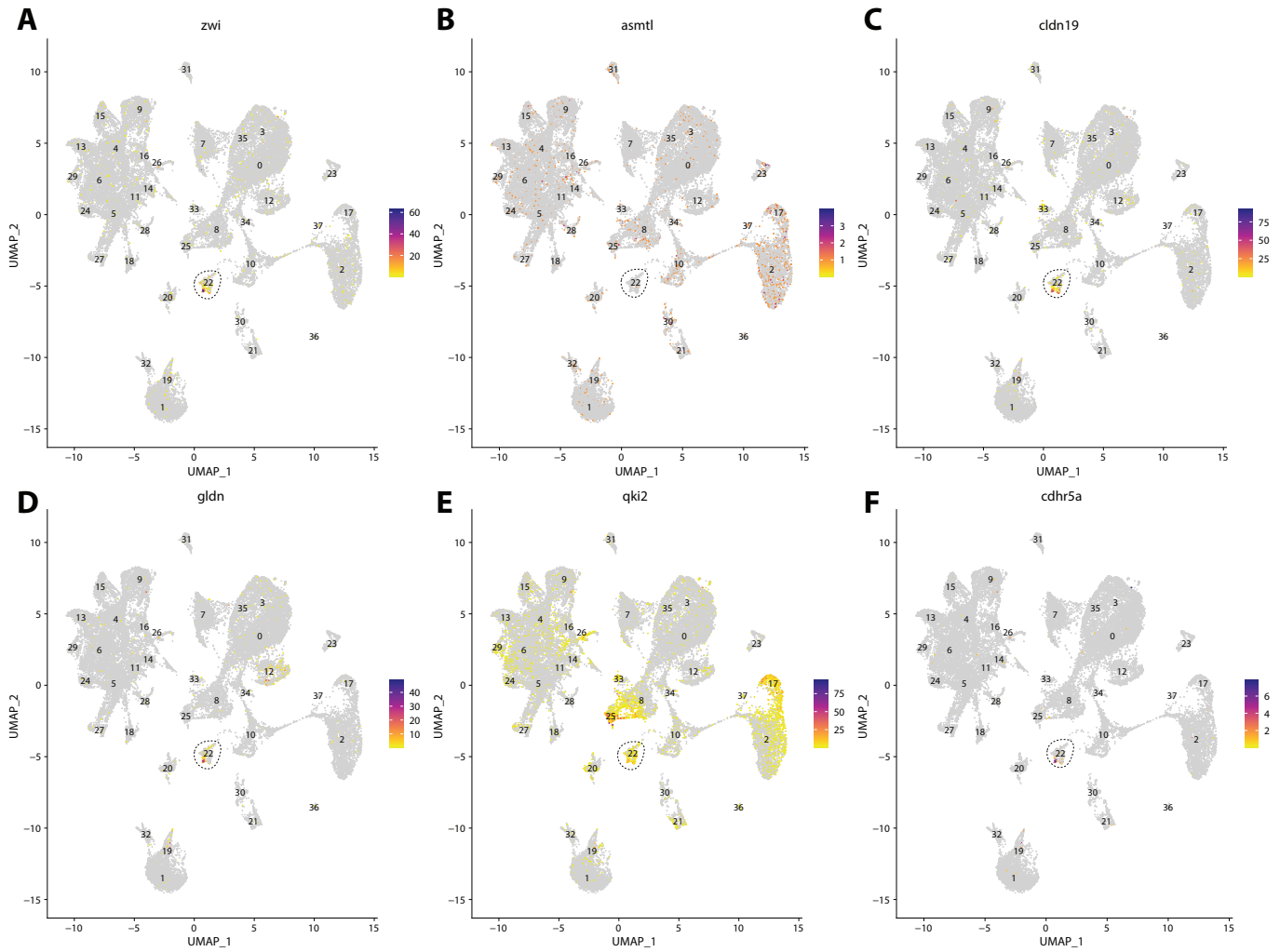


Figure S3.18: *Zebrafish glial cell marker genes expression patterns correspond with annotated glial cells in Gulf pipefish atlas.* Feature plots of zebrafish glial cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated glial cell cluster in pipefish is distinguished by a dashed outline.

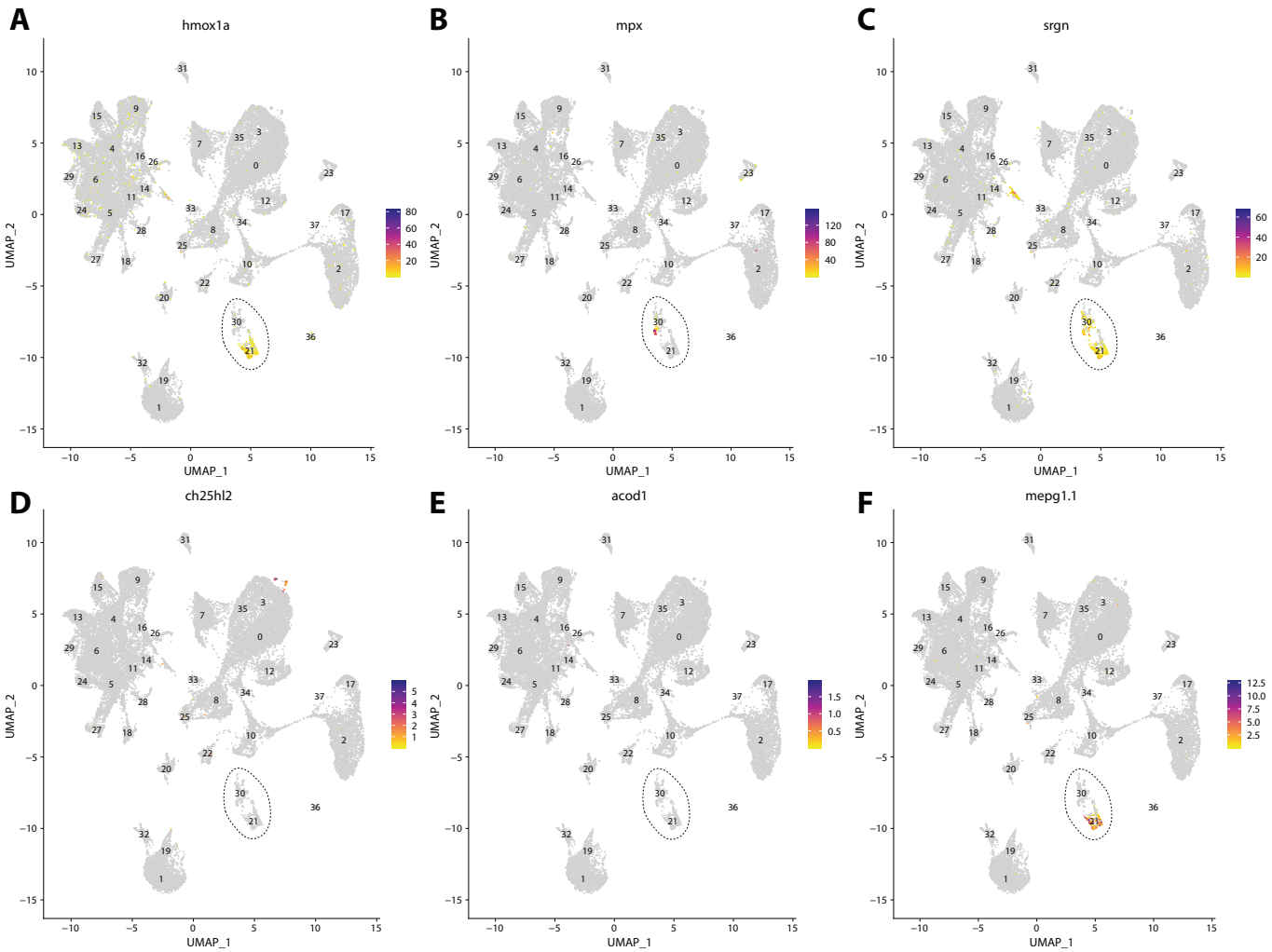


Figure S3.19: *Zebrafish immune cell marker genes expression patterns correspond with annotated immune cells in Gulf pipefish atlas.* Feature plots of zebrafish immune cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated immune cell clusters in pipefish are distinguished by a dashed outline.

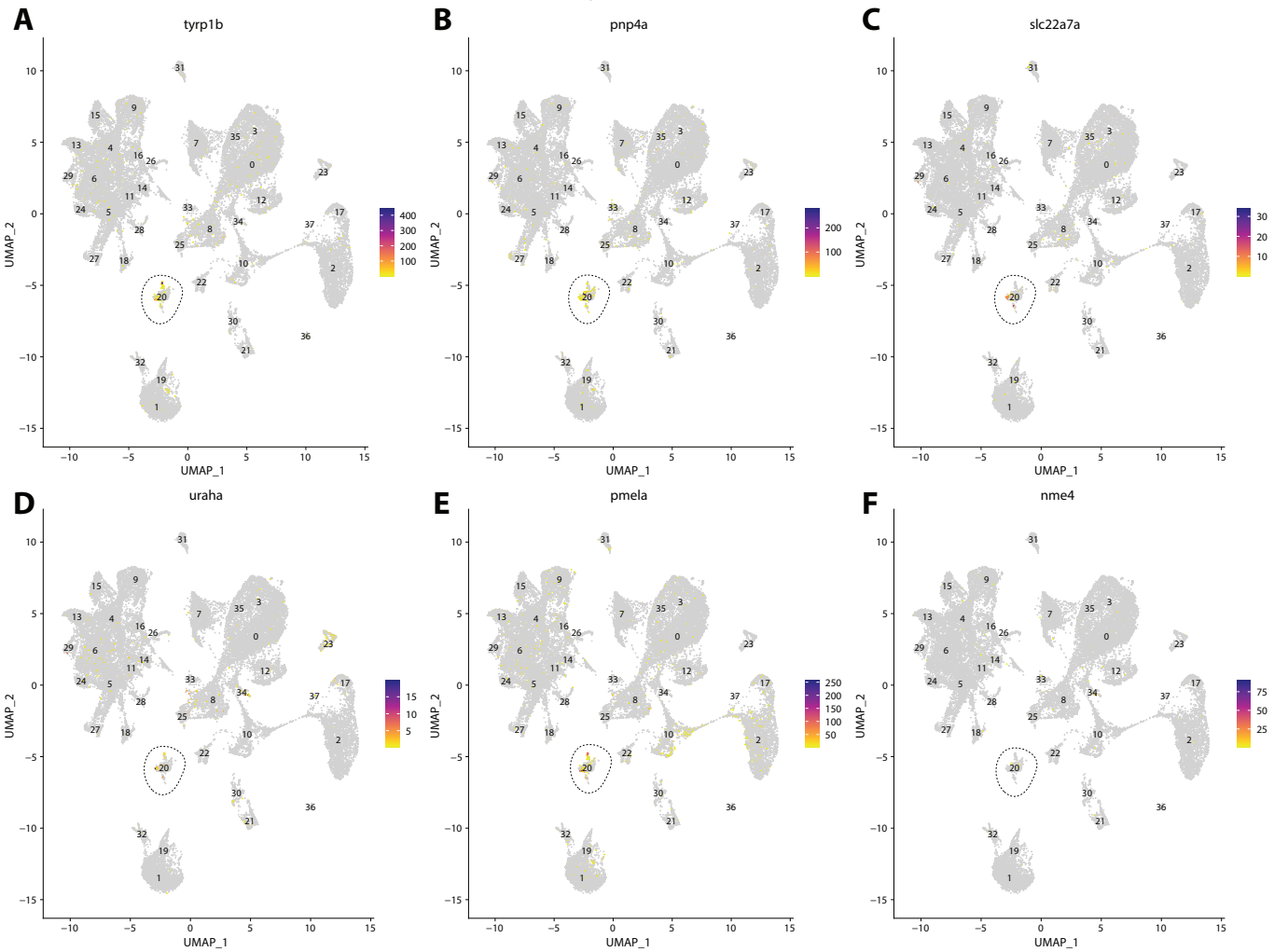


Figure S3.20: *Zebrafish pigment cell marker genes expression patterns correspond with annotated pigment cells in Gulf pipefish atlas.* Feature plots of zebrafish pigment cell marker genes identified from the Daniocell scRNAseq database demonstrate Gulf pipefish expression. Annotated pigment cell cluster in pipefish is distinguished by a dashed outline.

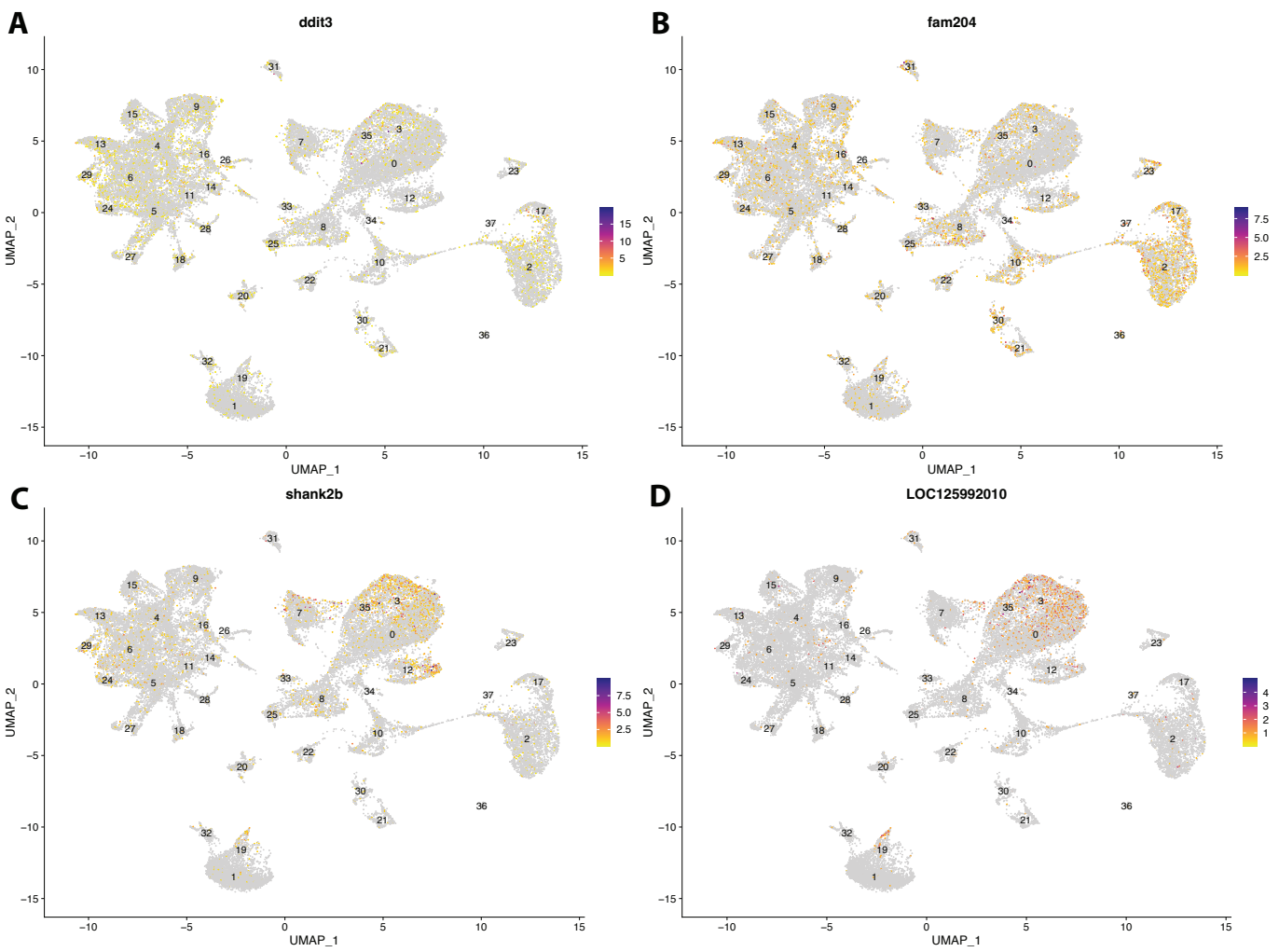


Figure S3.21: *Gulf pipefish cluster 0 markers.* FeaturePlots of top markers for cluster 0 identified from marker gene sets.

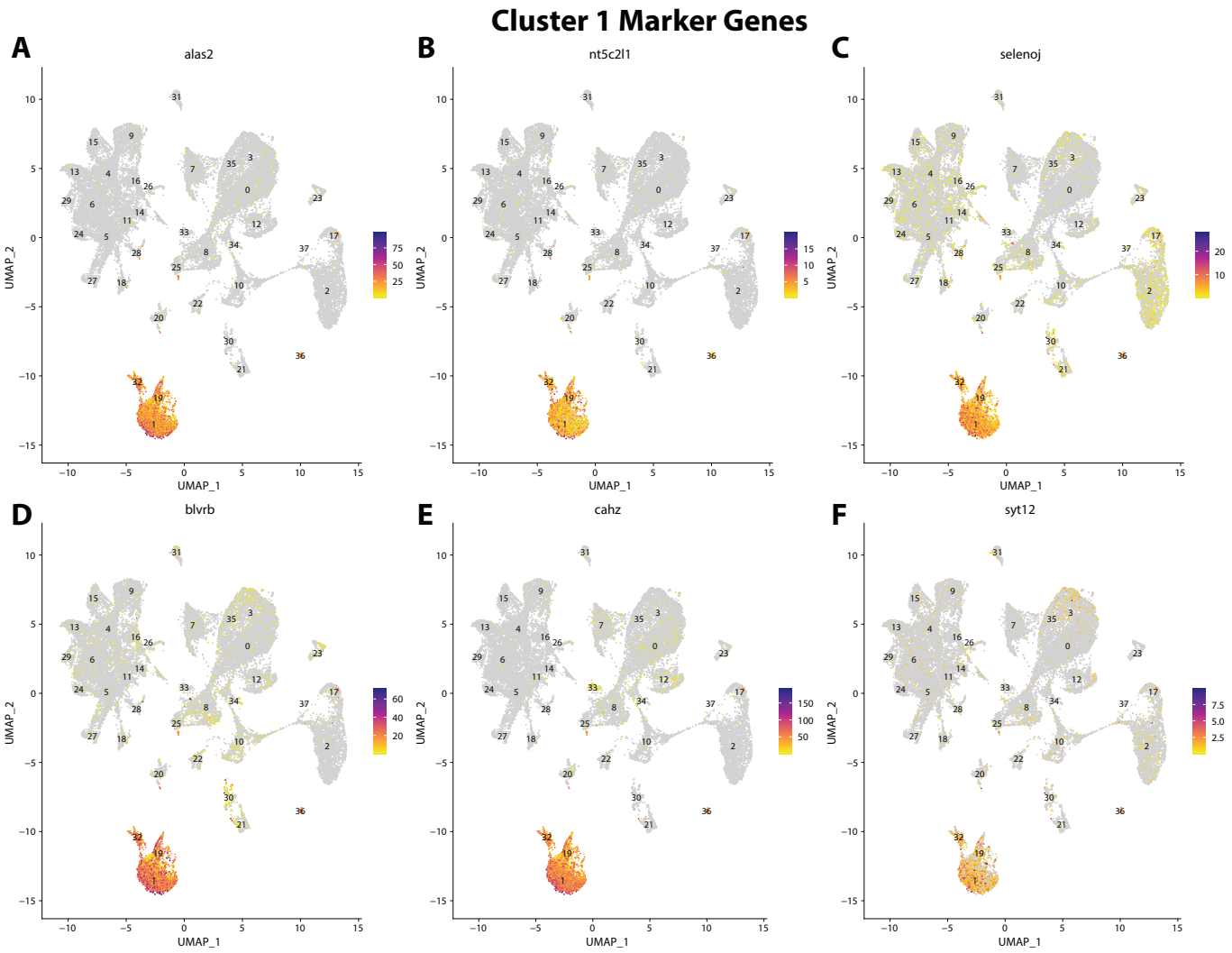


Figure S3.22: *Gulf pipefish cluster 1 markers.* FeaturePlots of top markers for cluster 1 identified from marker gene sets.

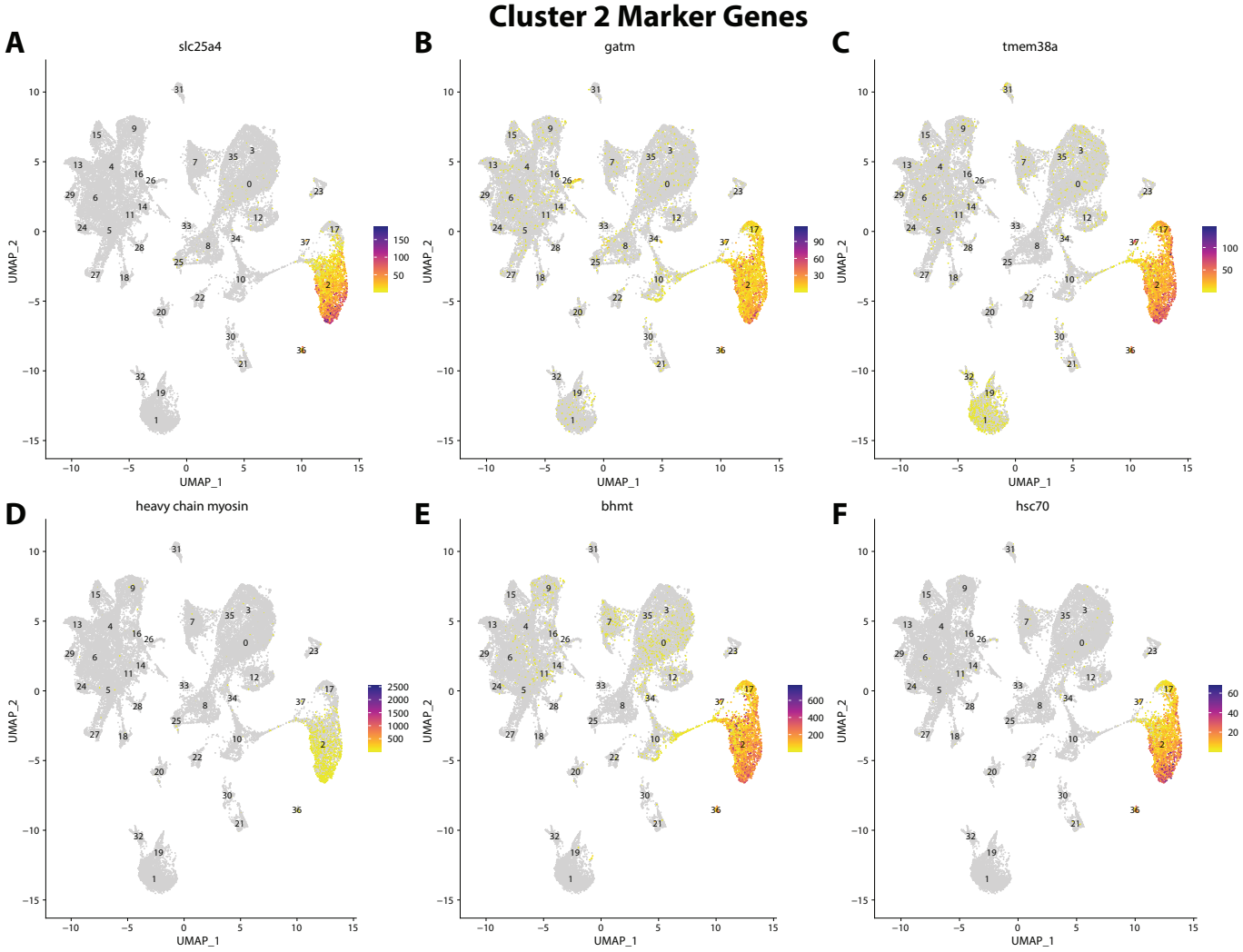


Figure S3.22: *Gulf pipefish cluster 2 markers.* FeaturePlots of top markers for cluster 2 identified from marker gene sets.

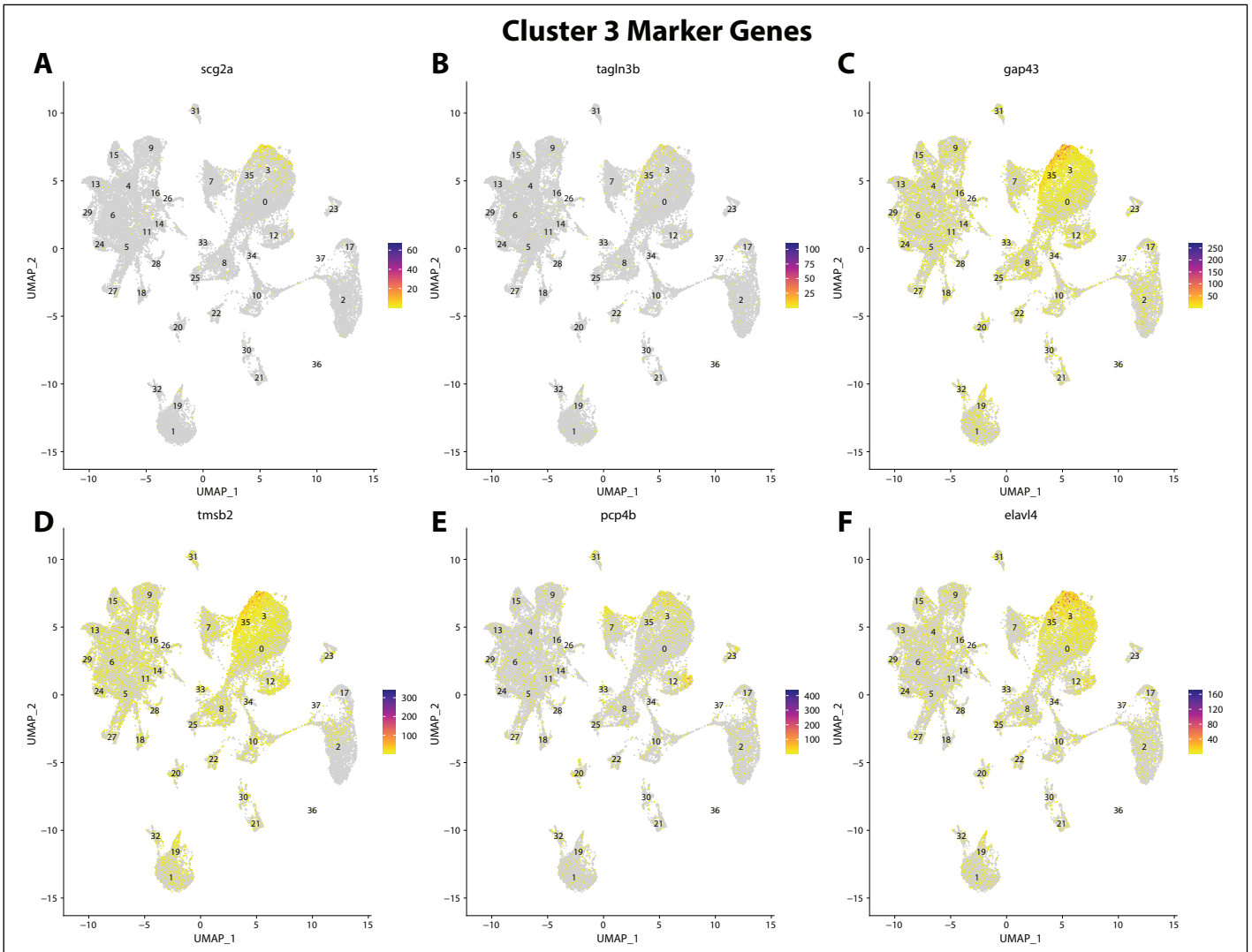


Figure S3.24: *Gulf pipefish cluster 3 markers.* FeaturePlots of top markers for cluster 3 identified from marker gene sets.

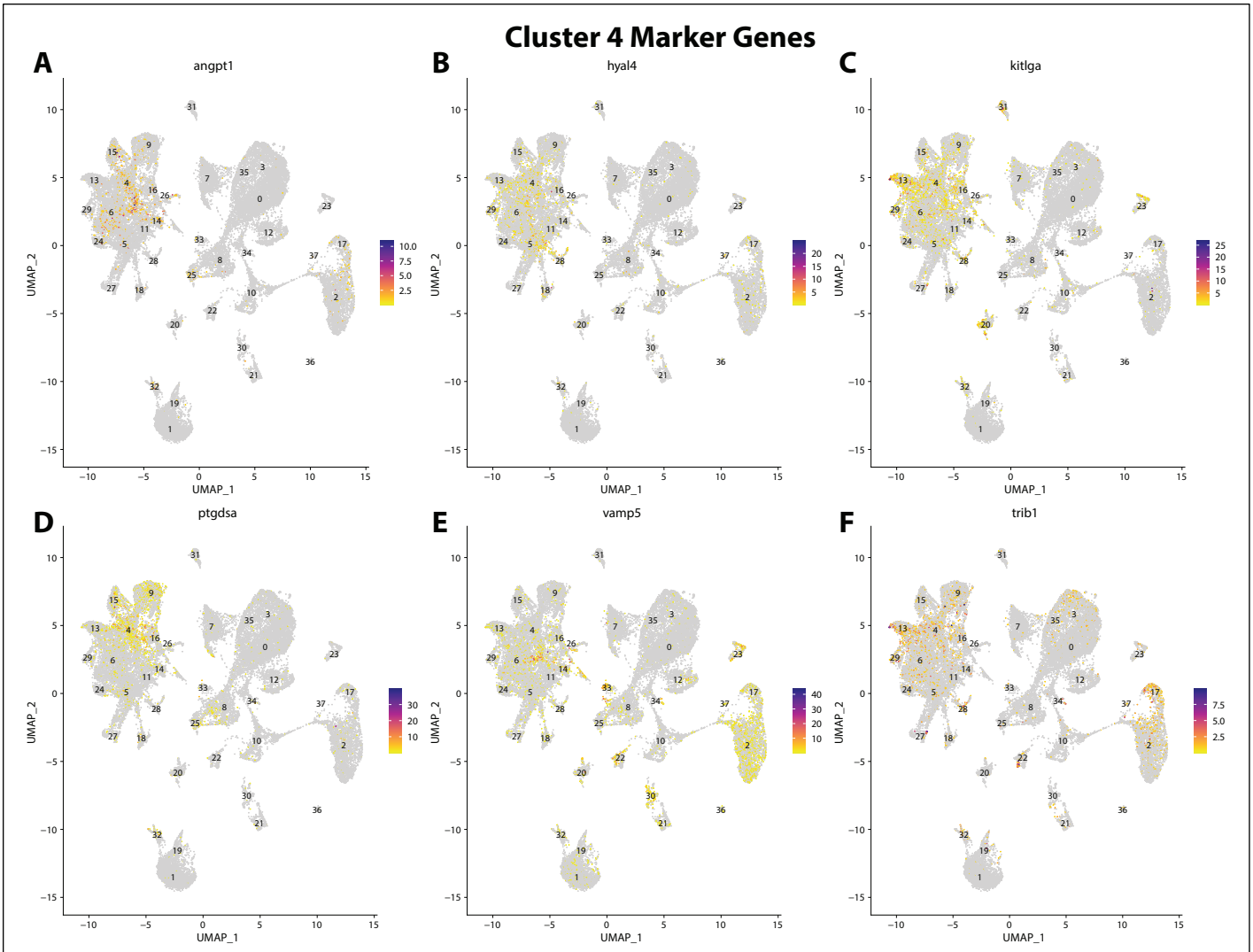


Figure S3.25: *Gulf pipefish cluster 4 markers.* FeaturePlots of top markers for cluster 4 identified from marker gene sets.

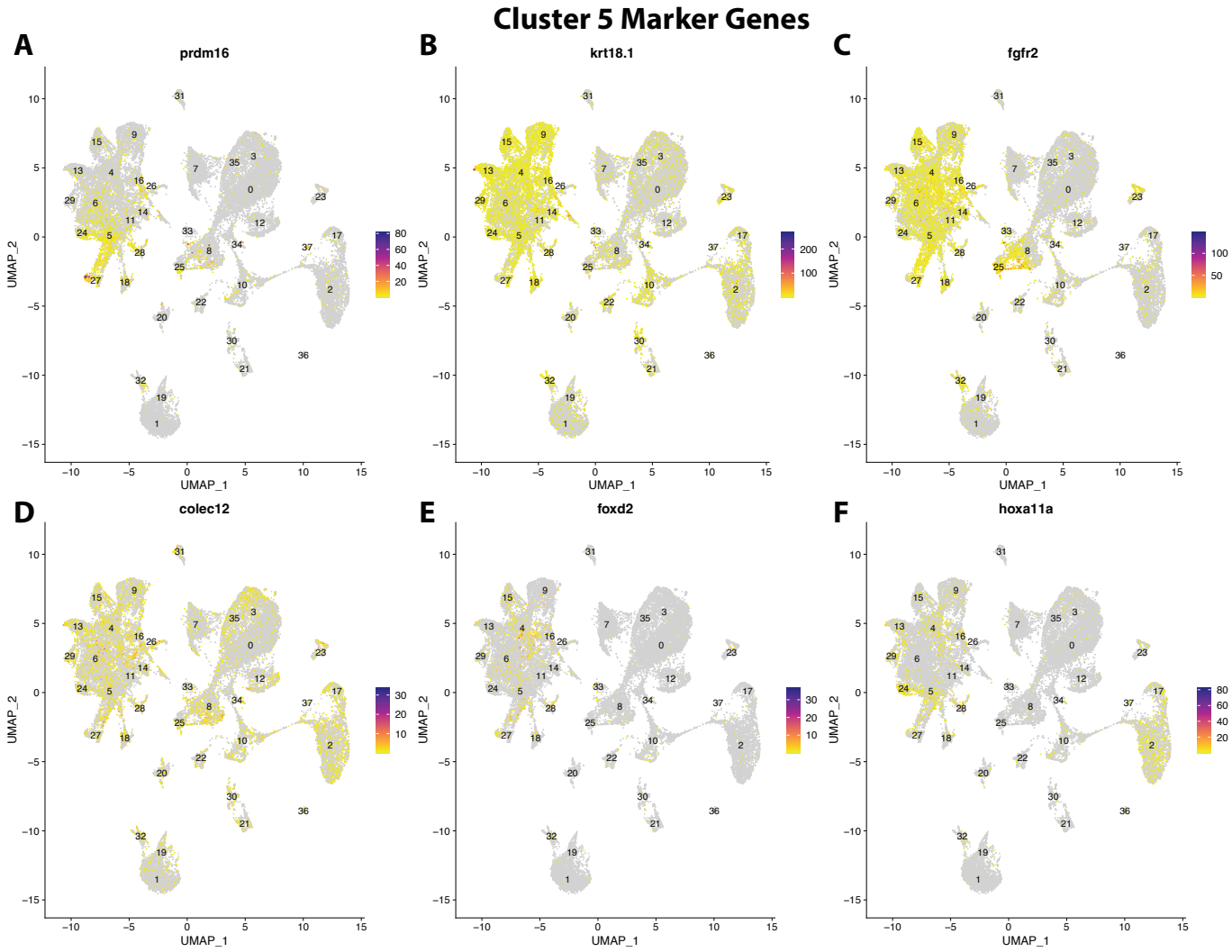


Figure S3.26: *Gulf pipefish cluster 5 markers.* FeaturePlots of top markers for cluster 5 identified from marker gene sets.

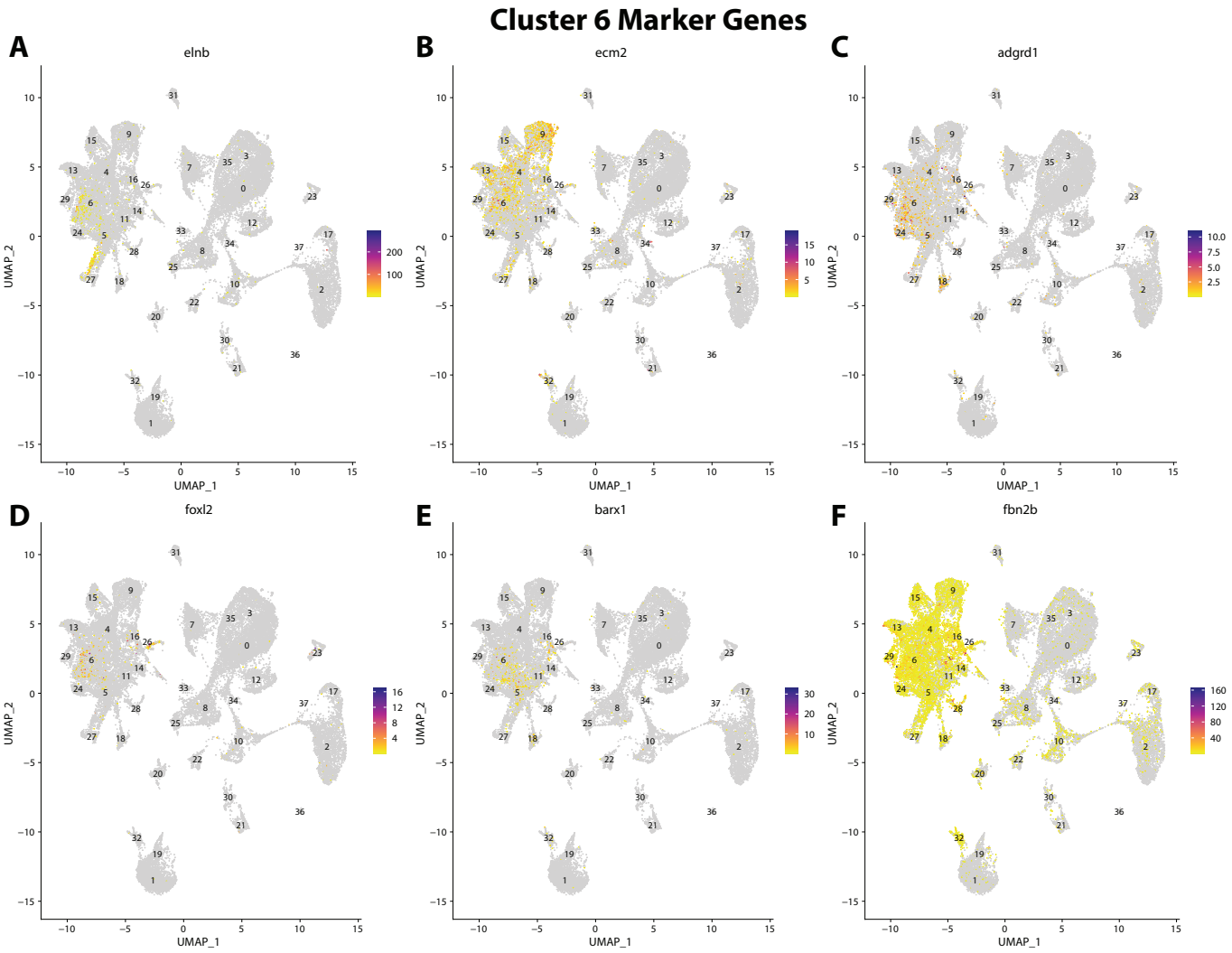


Figure S3.27: *Gulf pipefish cluster 6 markers.* FeaturePlots of top markers for cluster 6 identified from marker gene sets.

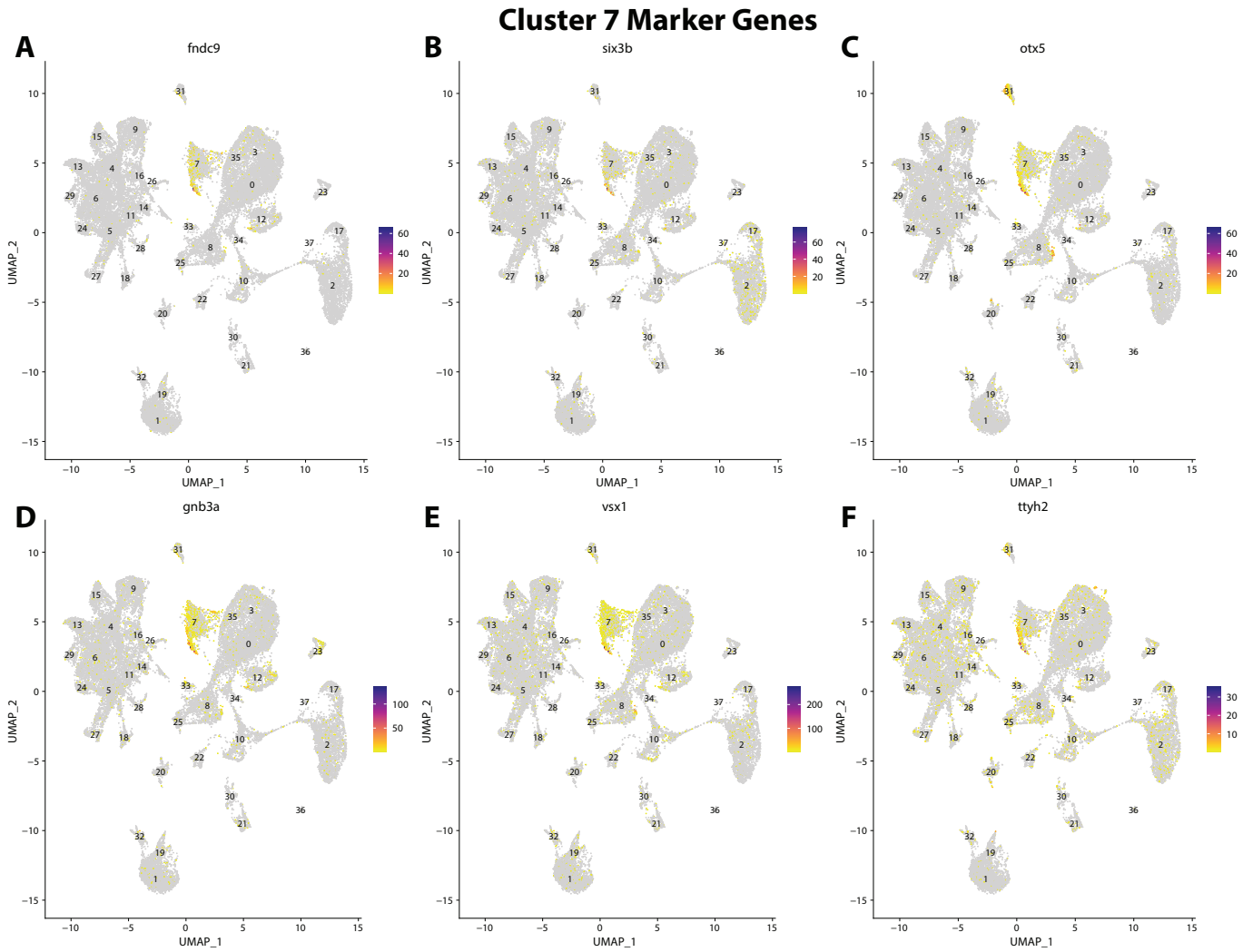


Figure S3.28: *Gulf pipefish cluster 7 markers.* FeaturePlots of top markers for cluster 7 identified from marker gene sets.

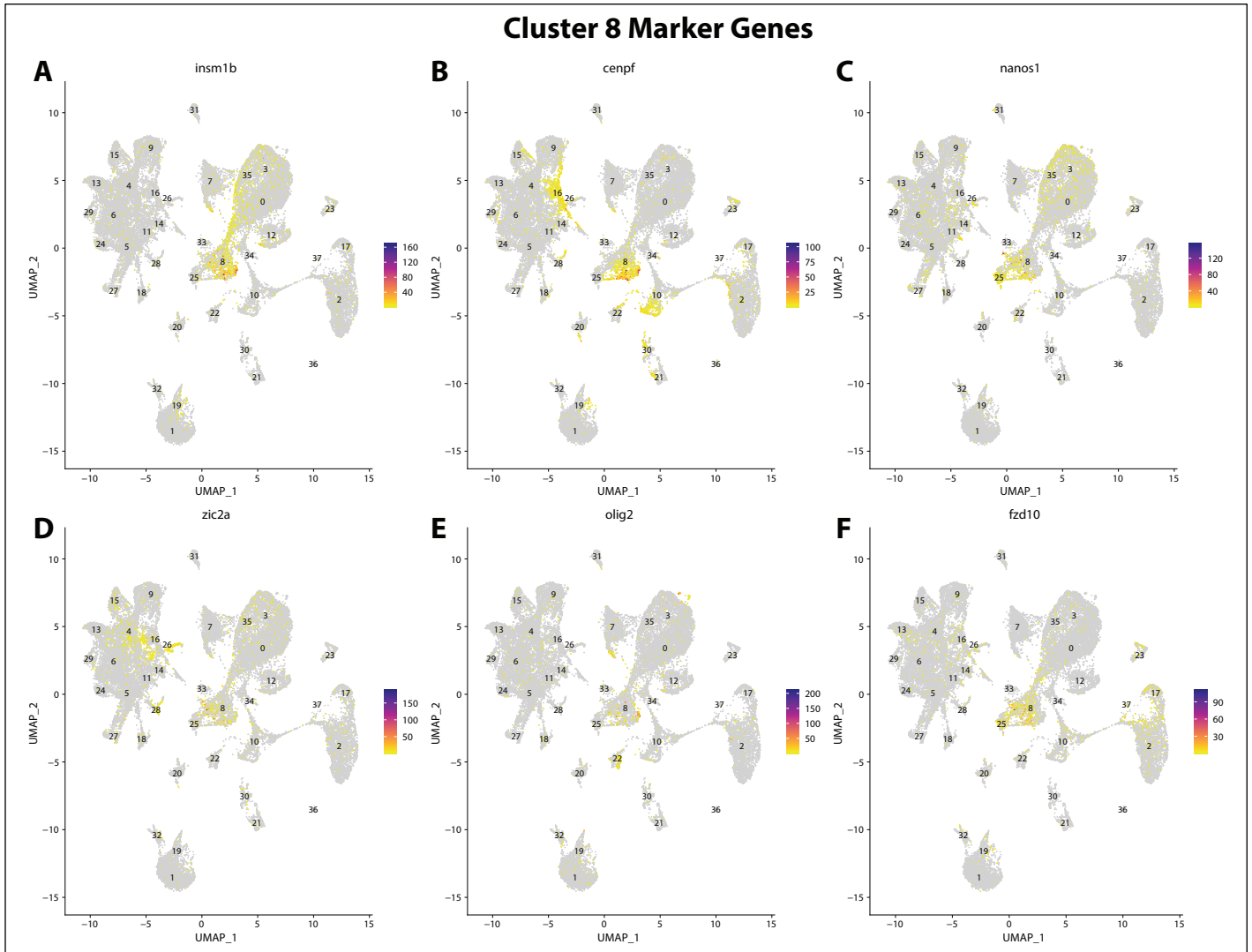


Figure S3.29: *Gulf pipefish cluster 8 markers.* FeaturePlots of top markers for cluster 8 identified from marker gene sets.

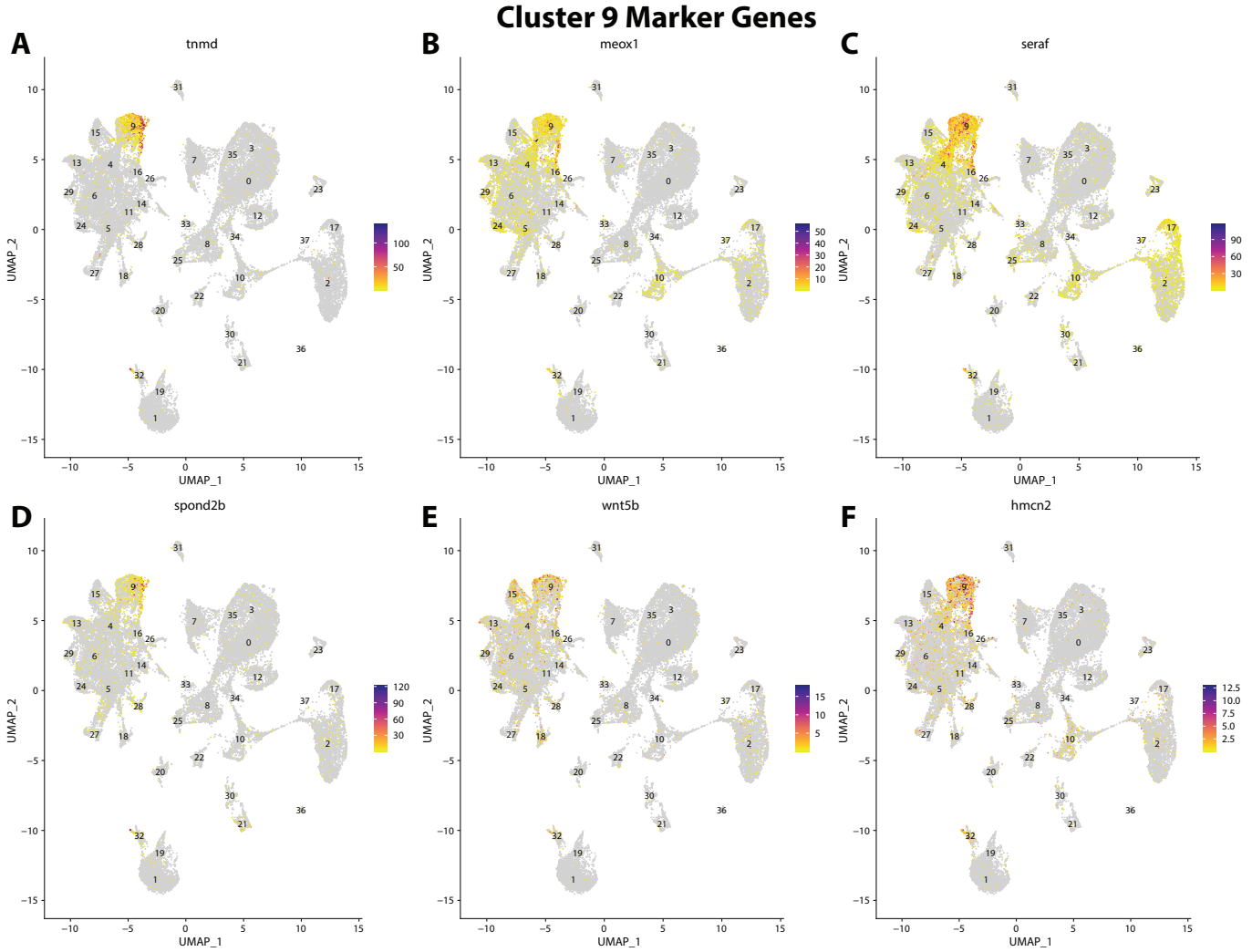


Figure S3.30: *Gulf pipefish cluster 9 markers.* FeaturePlots of top markers for cluster 9 identified from marker gene sets.

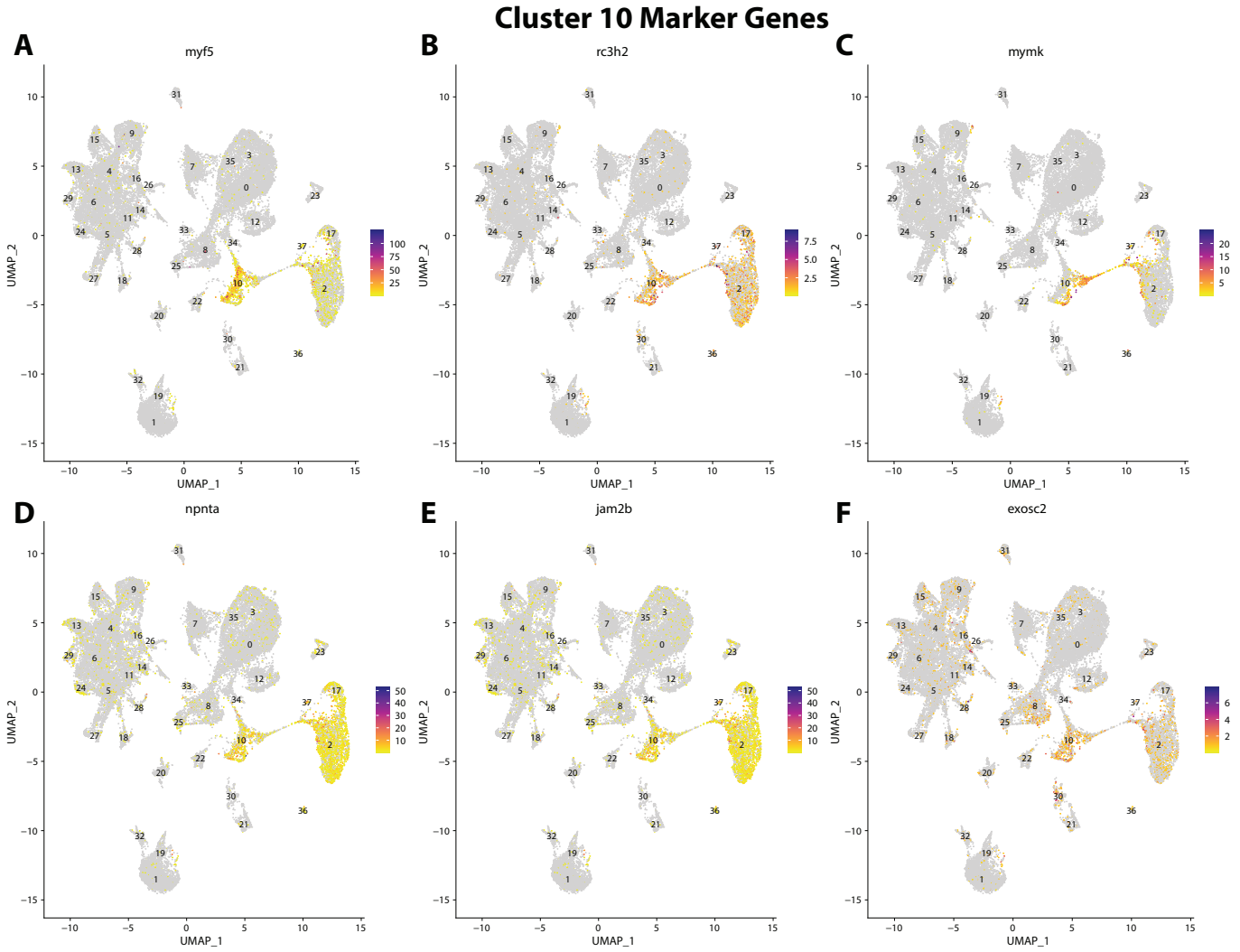


Figure S3.31: *Gulf pipefish cluster 10 markers.* FeaturePlots of top markers for cluster 10 identified from marker gene sets.

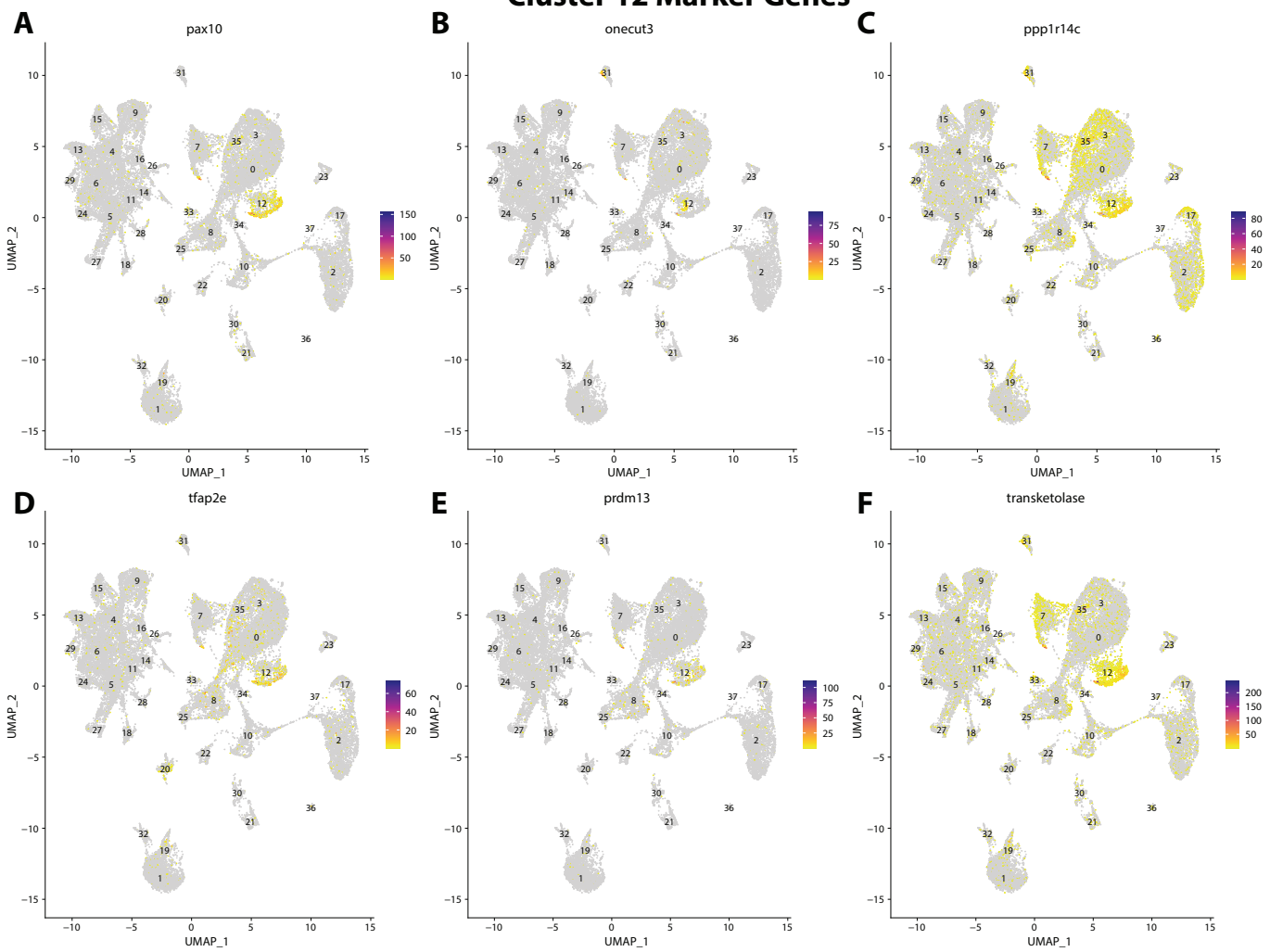


Figure S3.32: *Gulf pipefish cluster 12 markers.* FeaturePlots of top markers for cluster 12 identified from marker gene sets.

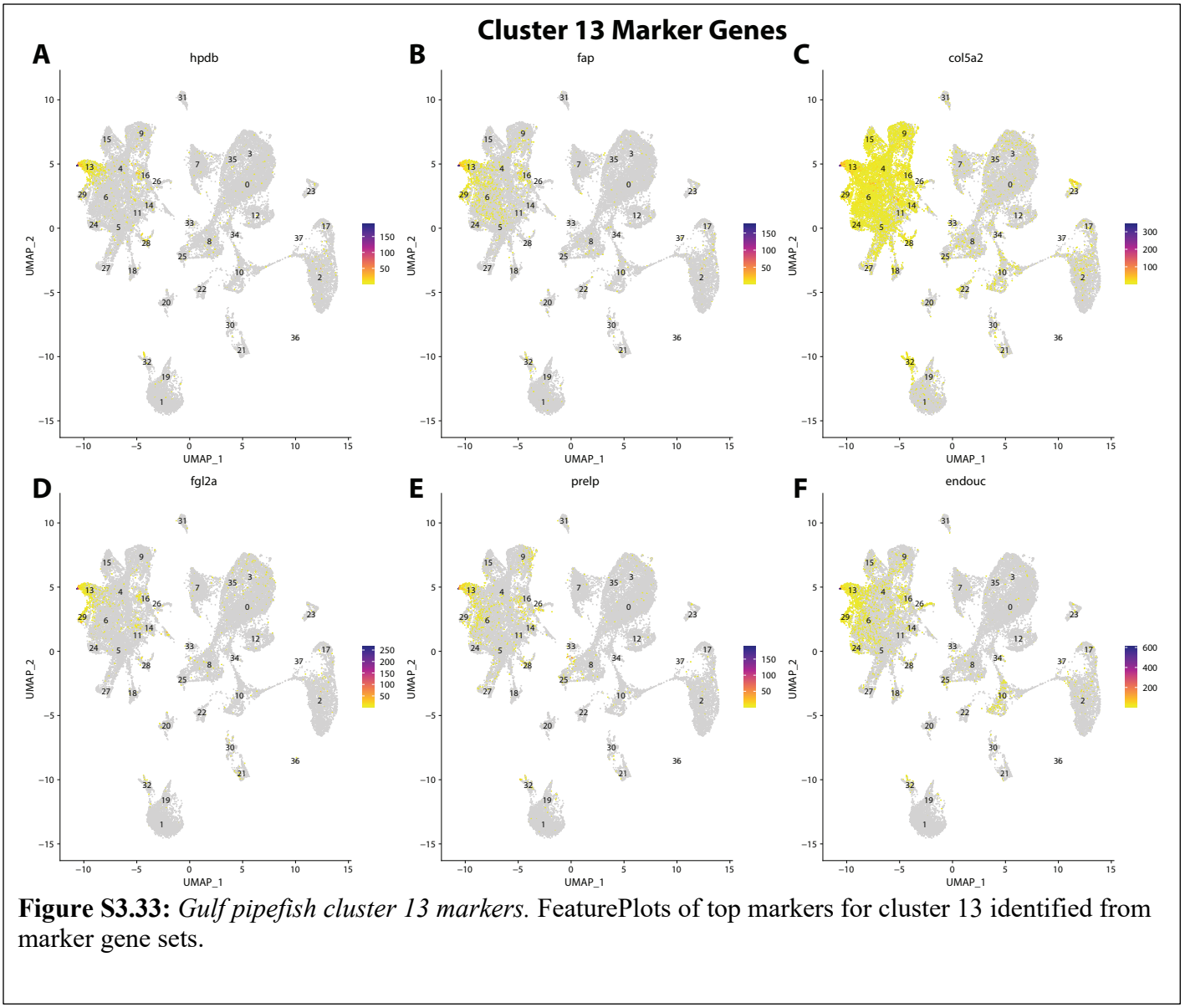


Figure S3.33: *Gulf pipefish cluster 13 markers.* FeaturePlots of top markers for cluster 13 identified from marker gene sets.

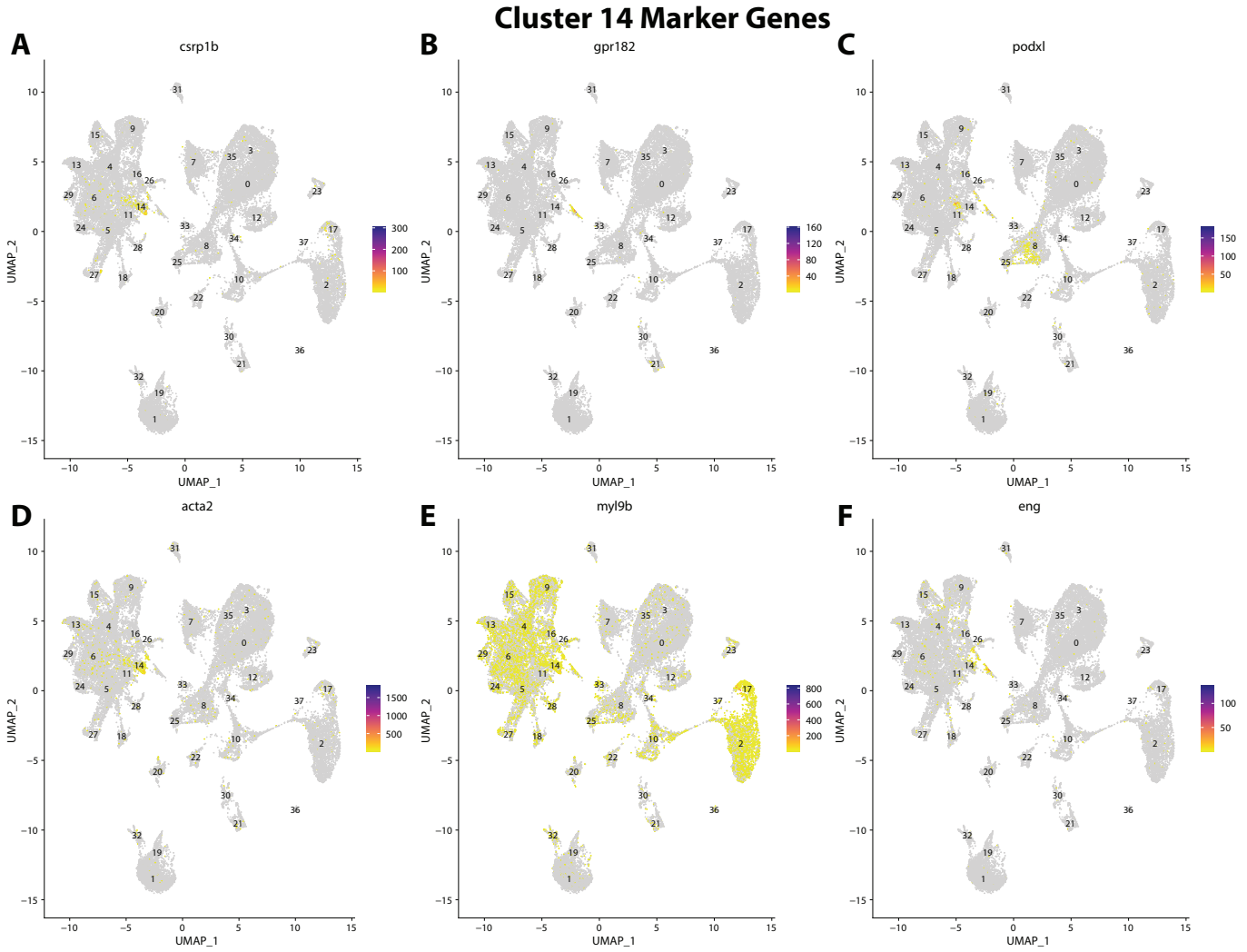


Figure S3.34: *Gulf pipefish cluster 14 markers.* FeaturePlots of top markers for cluster 14 identified from marker gene sets.

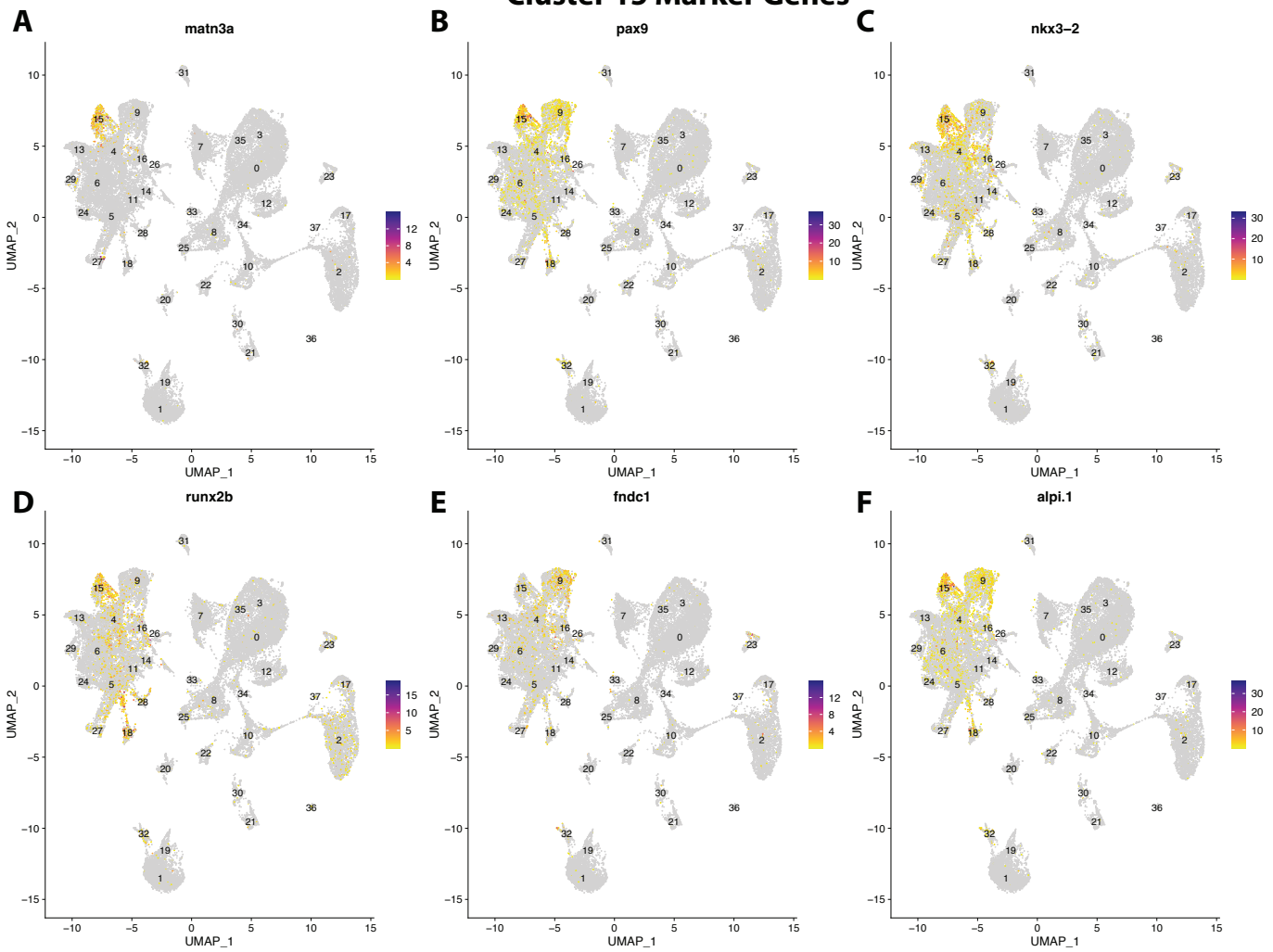


Figure S3.35: *Gulf pipefish cluster 15 markers.* FeaturePlots of top markers for cluster 15 identified from marker gene sets.

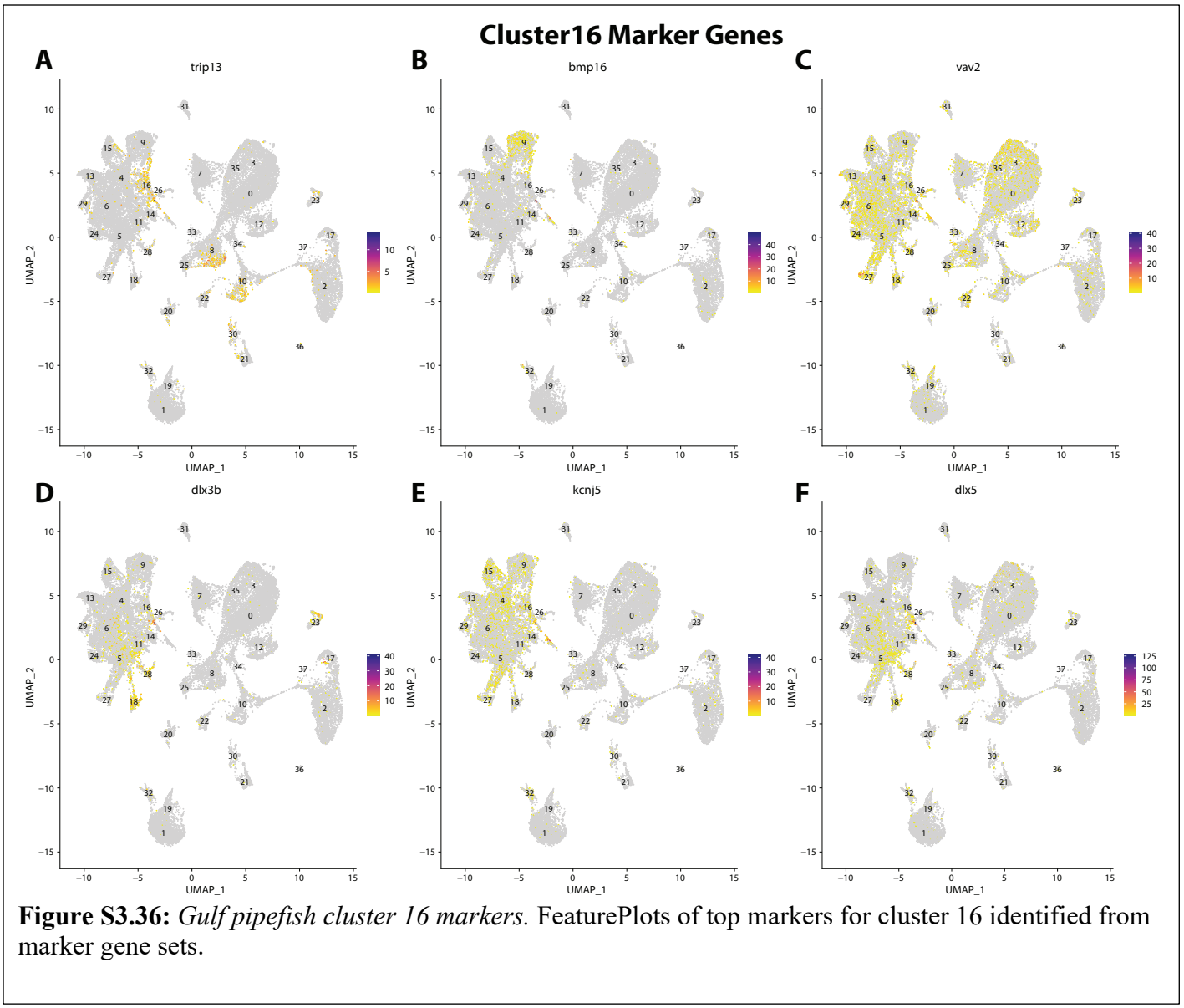


Figure S3.36: *Gulf pipefish cluster 16 markers.* FeaturePlots of top markers for cluster 16 identified from marker gene sets.

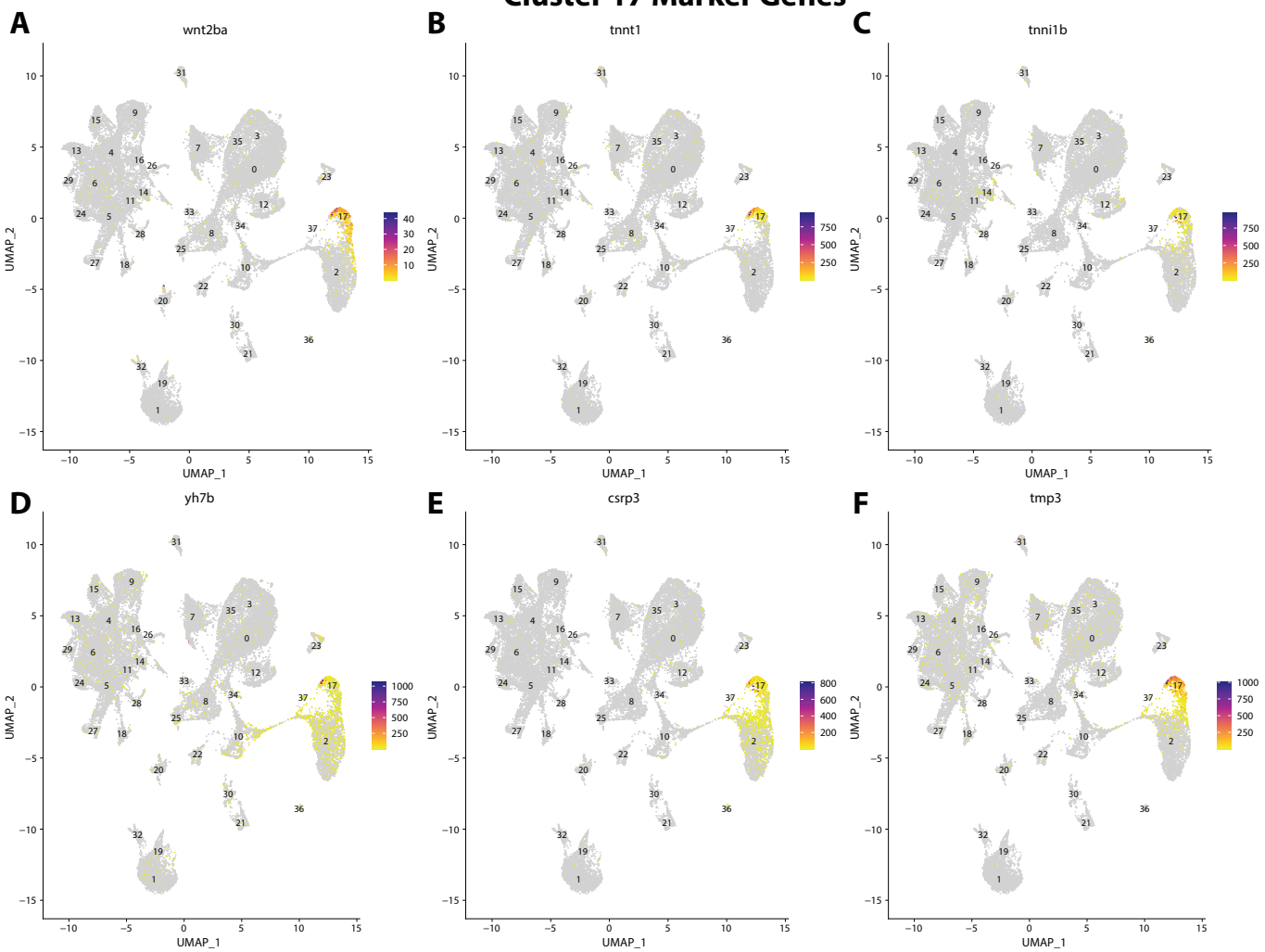


Figure S3.37: *Gulf pipefish cluster 17 markers.* FeaturePlots of top markers for cluster 17 identified from marker gene sets.

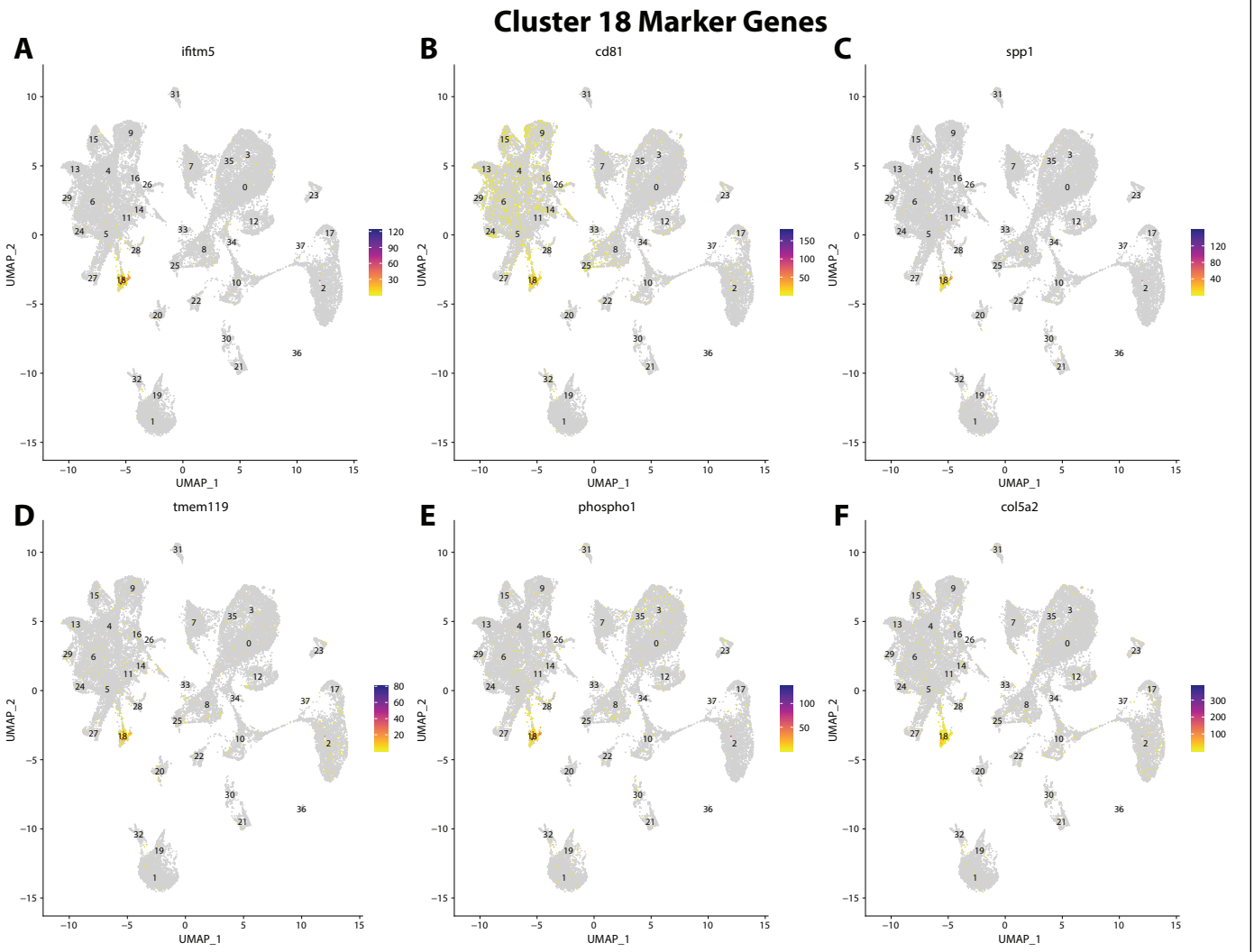


Figure S3.38: *Gulf pipefish cluster 18 markers.* FeaturePlots of top markers for cluster 18 identified from marker gene sets.

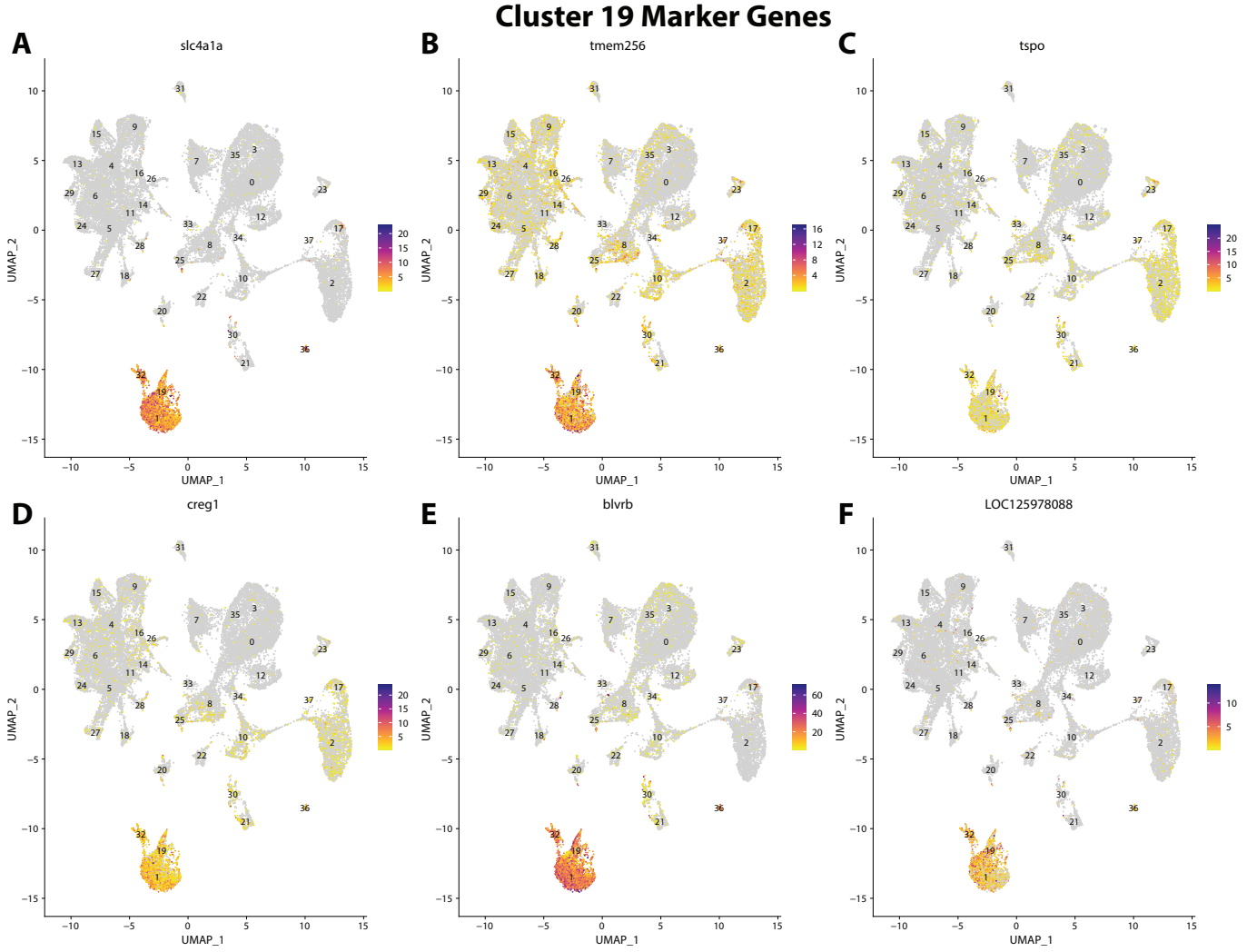


Figure S3.34: *Gulf pipefish cluster 19 markers.* FeaturePlots of top markers for cluster 19 identified from marker gene sets.

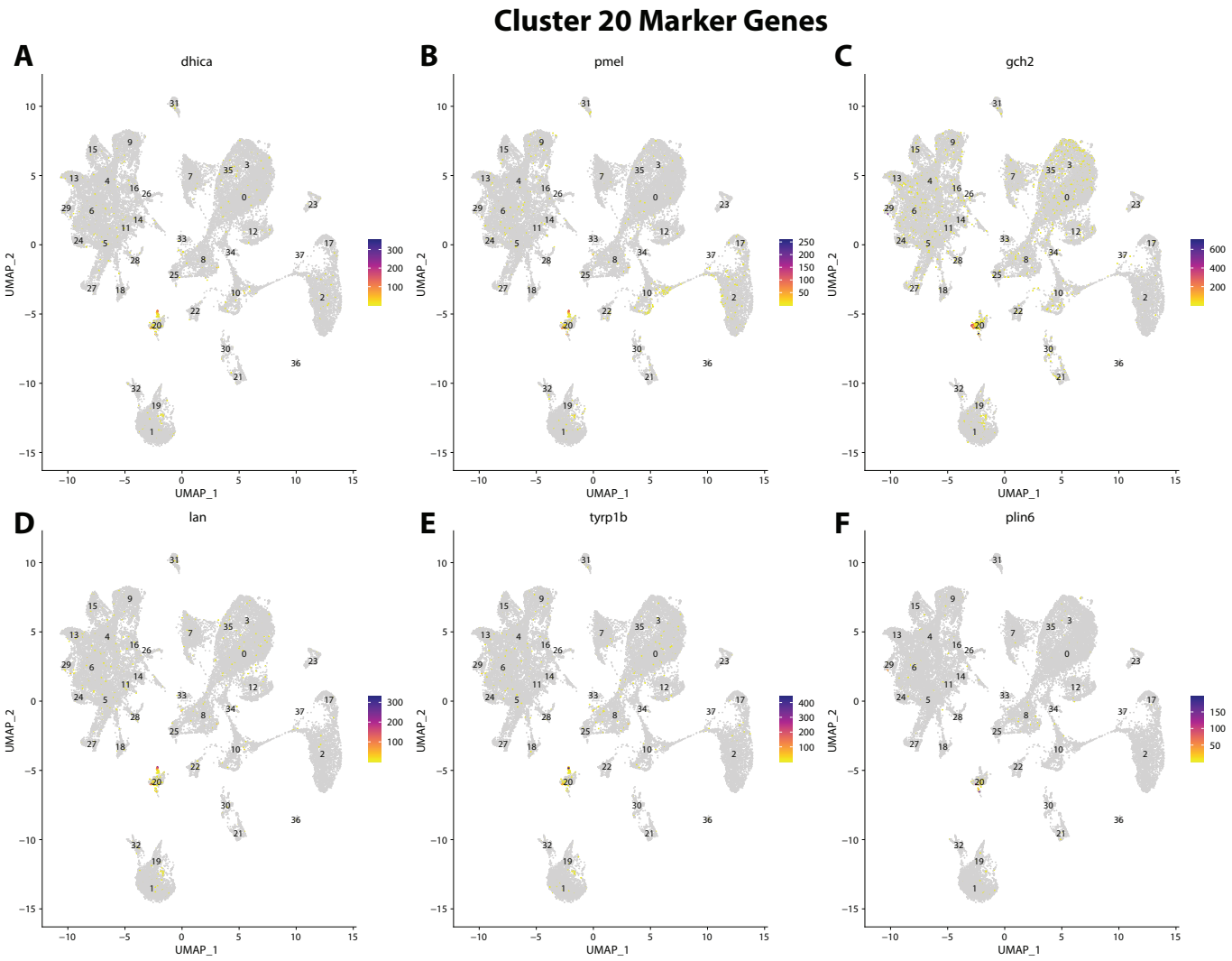


Figure S3.40: *Gulf pipefish cluster 20 markers.* FeaturePlots of top markers for cluster 20 identified from marker gene sets.

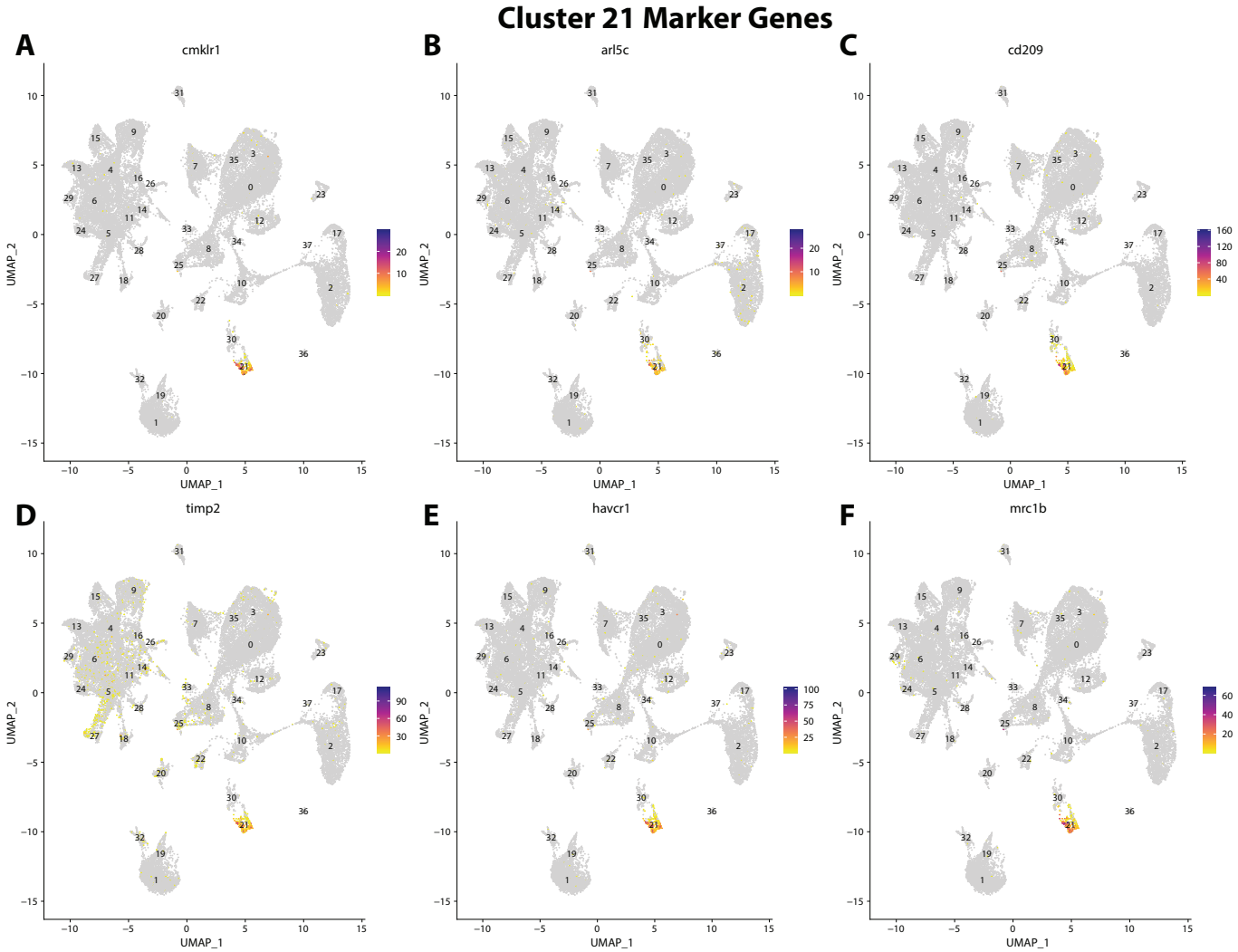


Figure S3.41: *Gulf pipefish cluster 21 markers.* FeaturePlots of top markers for cluster 21 identified from marker gene sets.

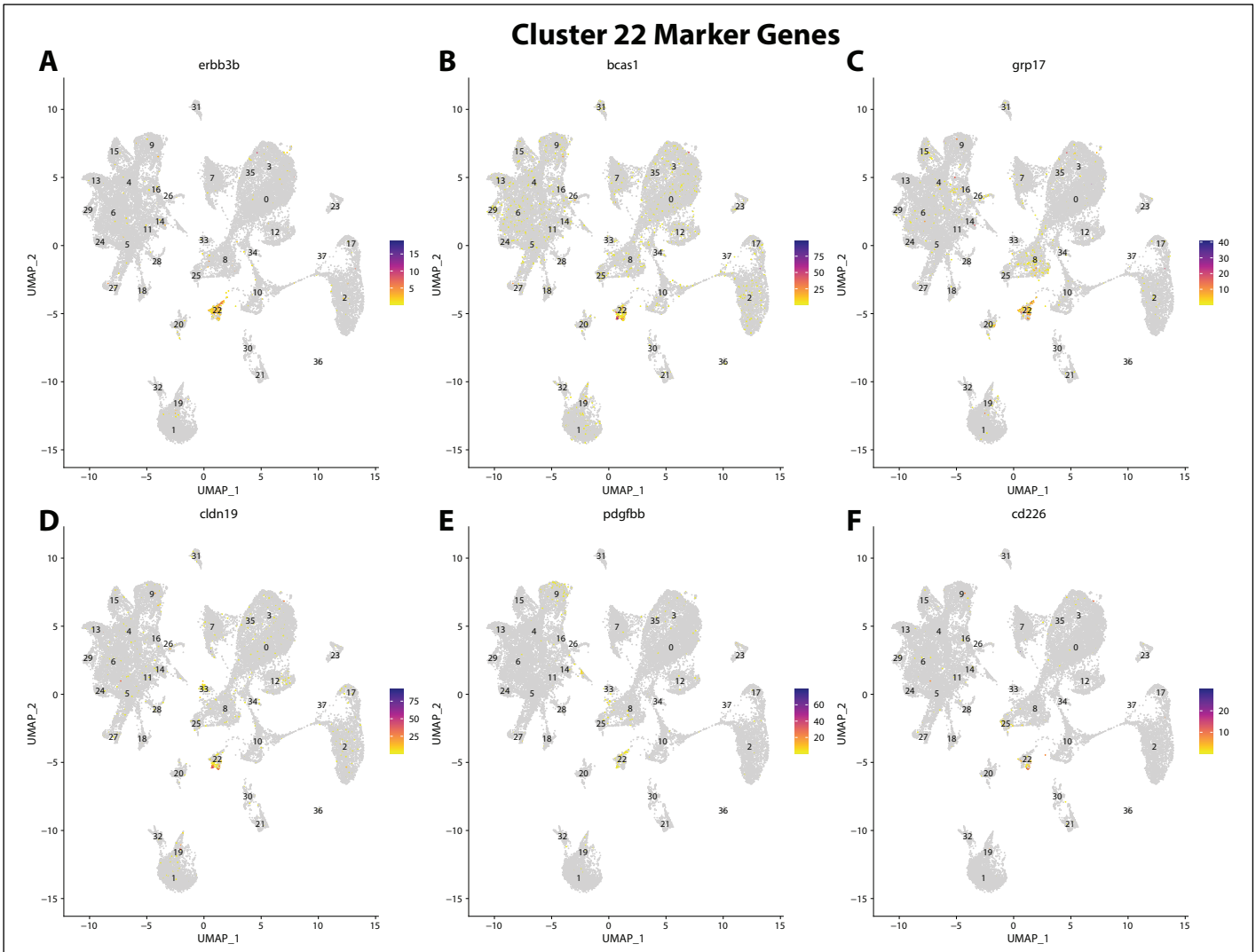


Figure S3.42: *Gulf pipefish cluster 22 markers.* FeaturePlots of top markers for cluster 22 identified from marker gene sets.

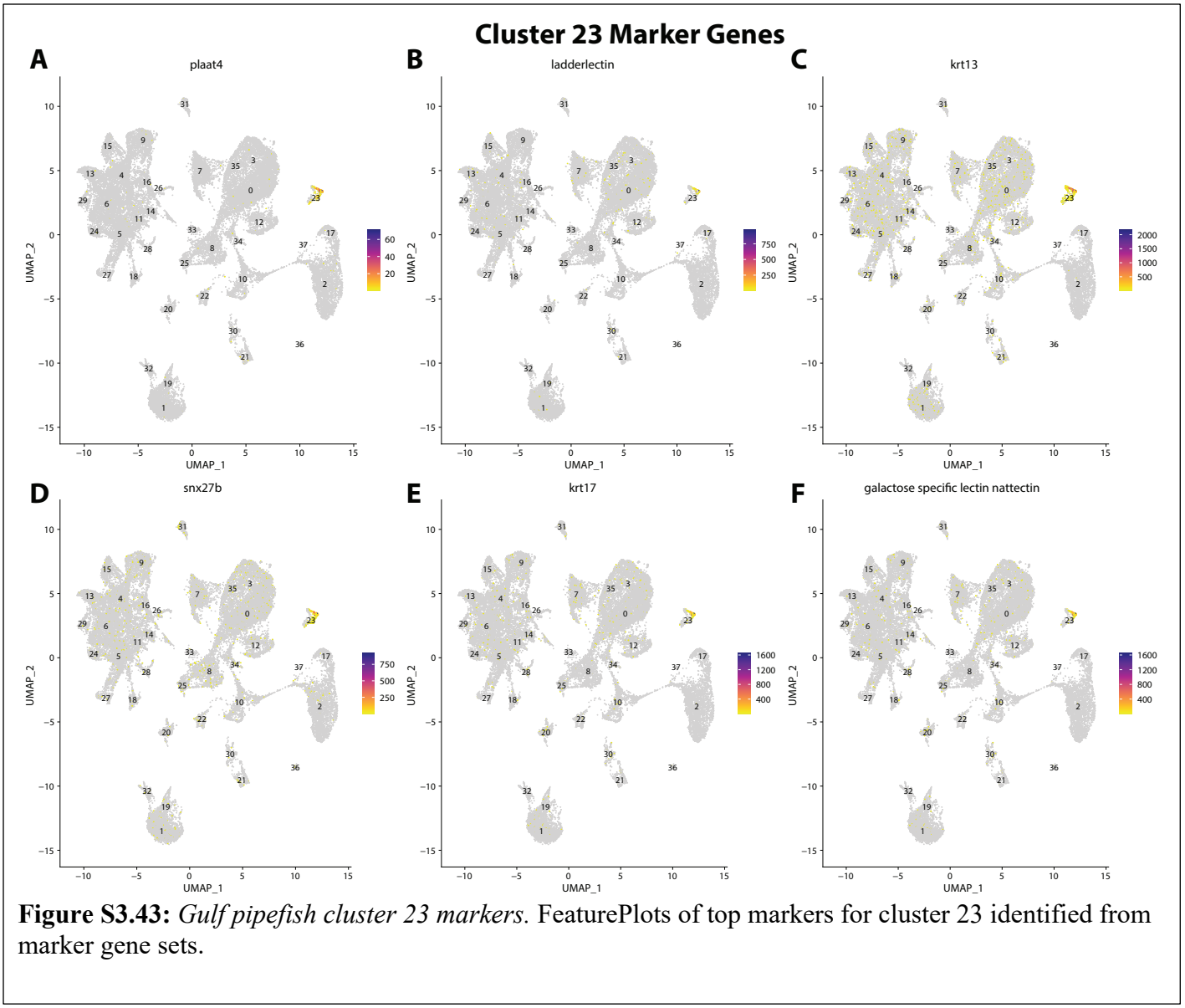


Figure S3.43: *Gulf pipefish cluster 23 markers.* FeaturePlots of top markers for cluster 23 identified from marker gene sets.

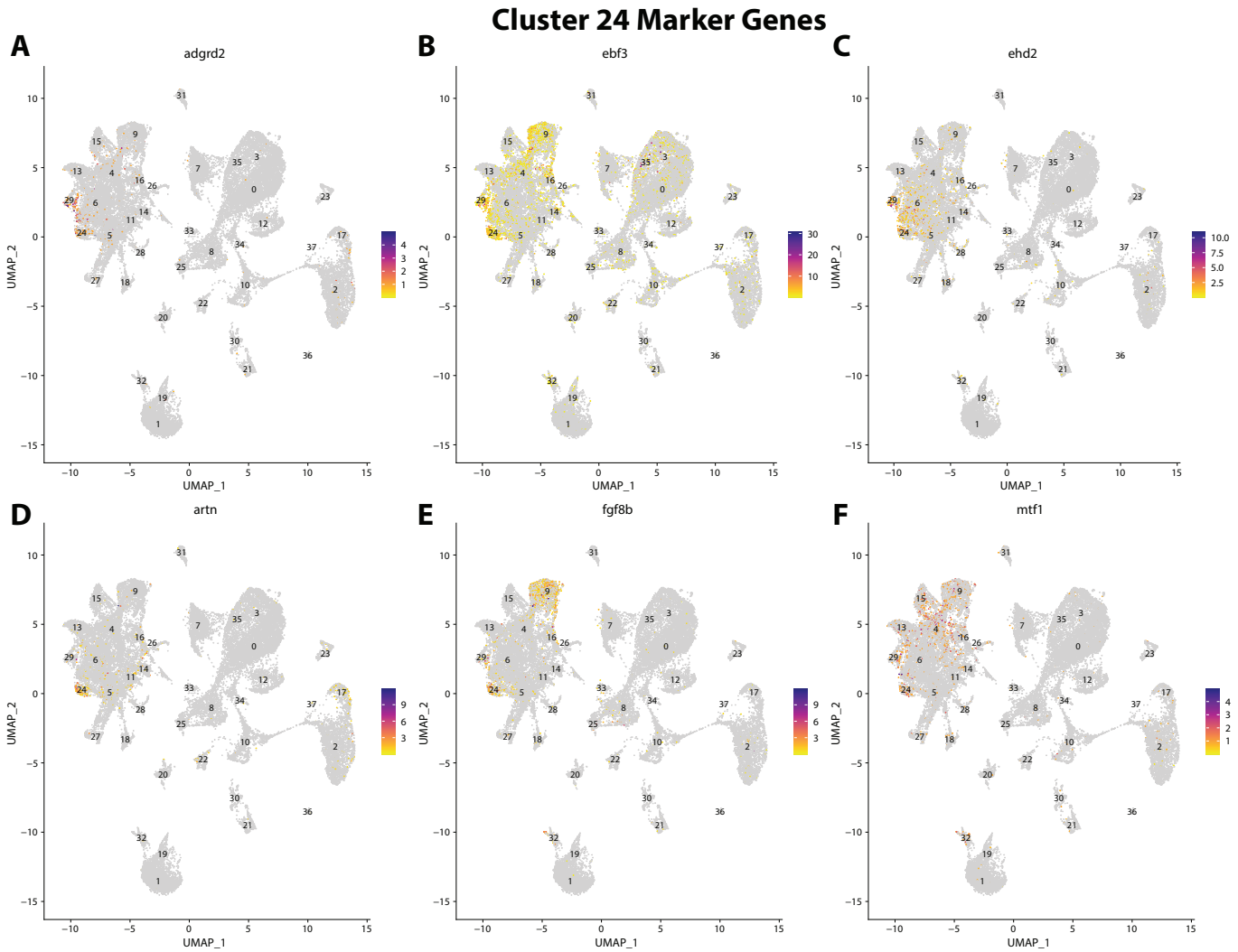


Figure S3.44: *Gulf pipefish cluster 24 markers.* FeaturePlots of top markers for cluster 24 identified from marker gene sets.

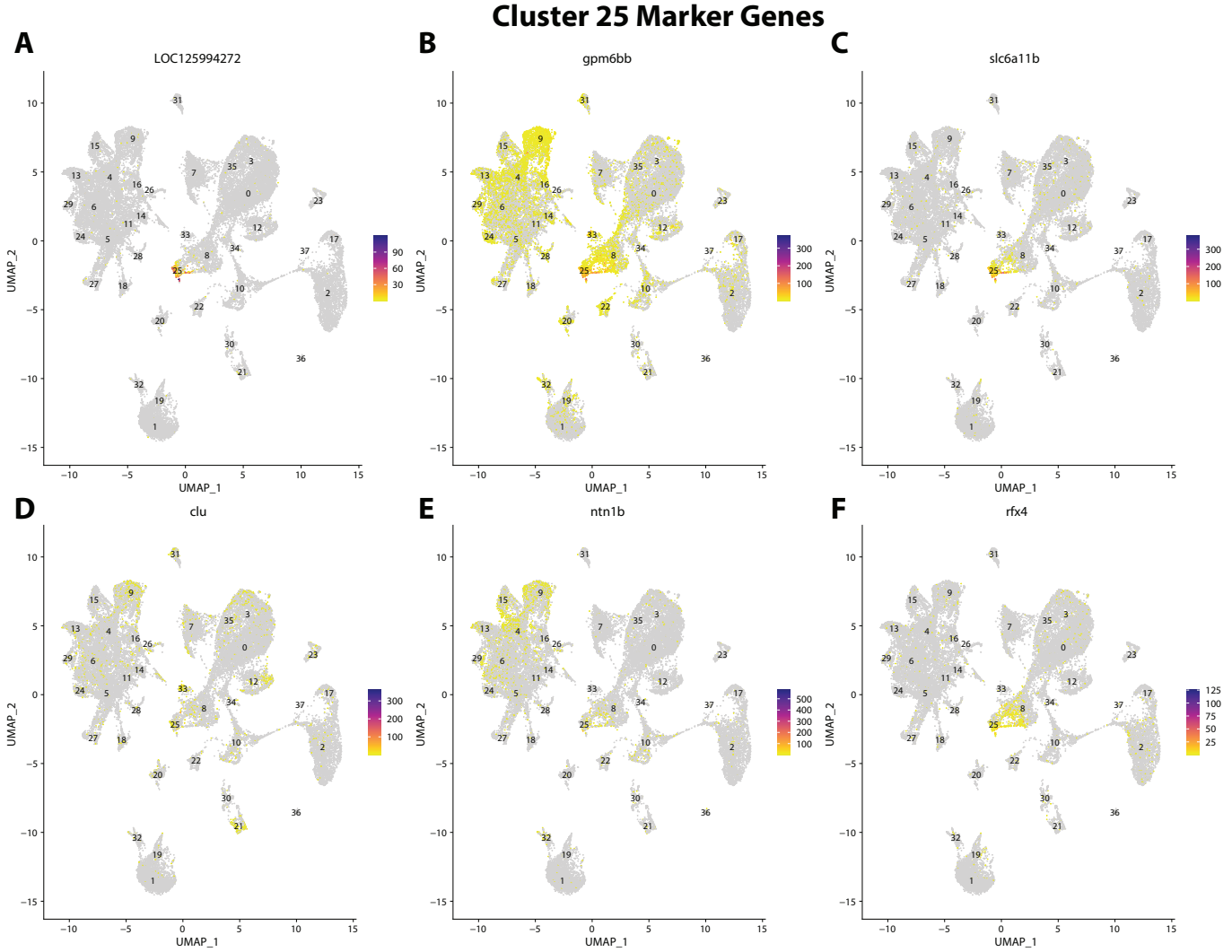


Figure S3.45: *Gulf pipefish cluster 25 markers.* FeaturePlots of top markers for cluster 25 identified from marker gene sets.

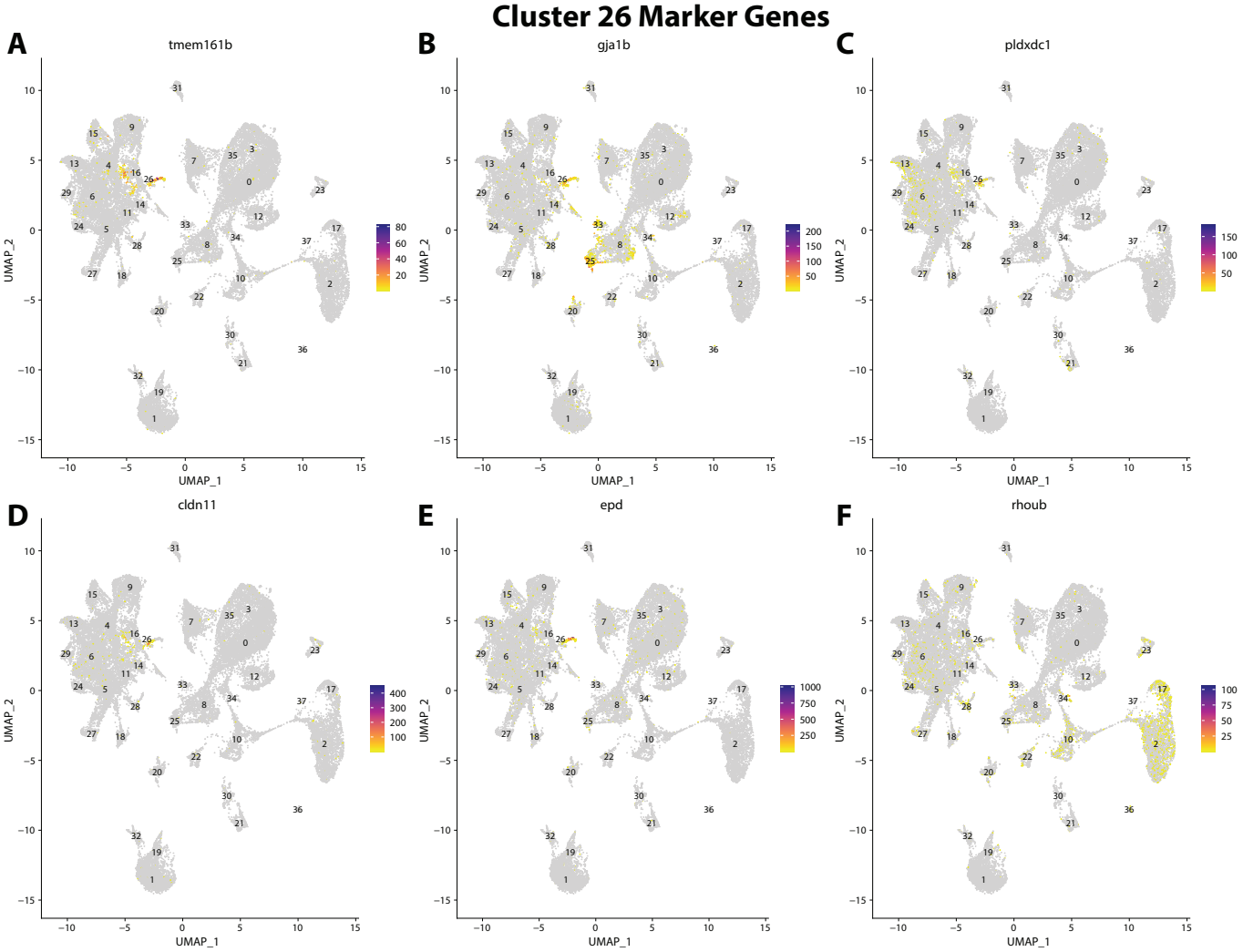


Figure S3.46: *Gulf pipefish cluster 26 markers.* FeaturePlots of top markers for cluster 26 identified from marker gene sets.

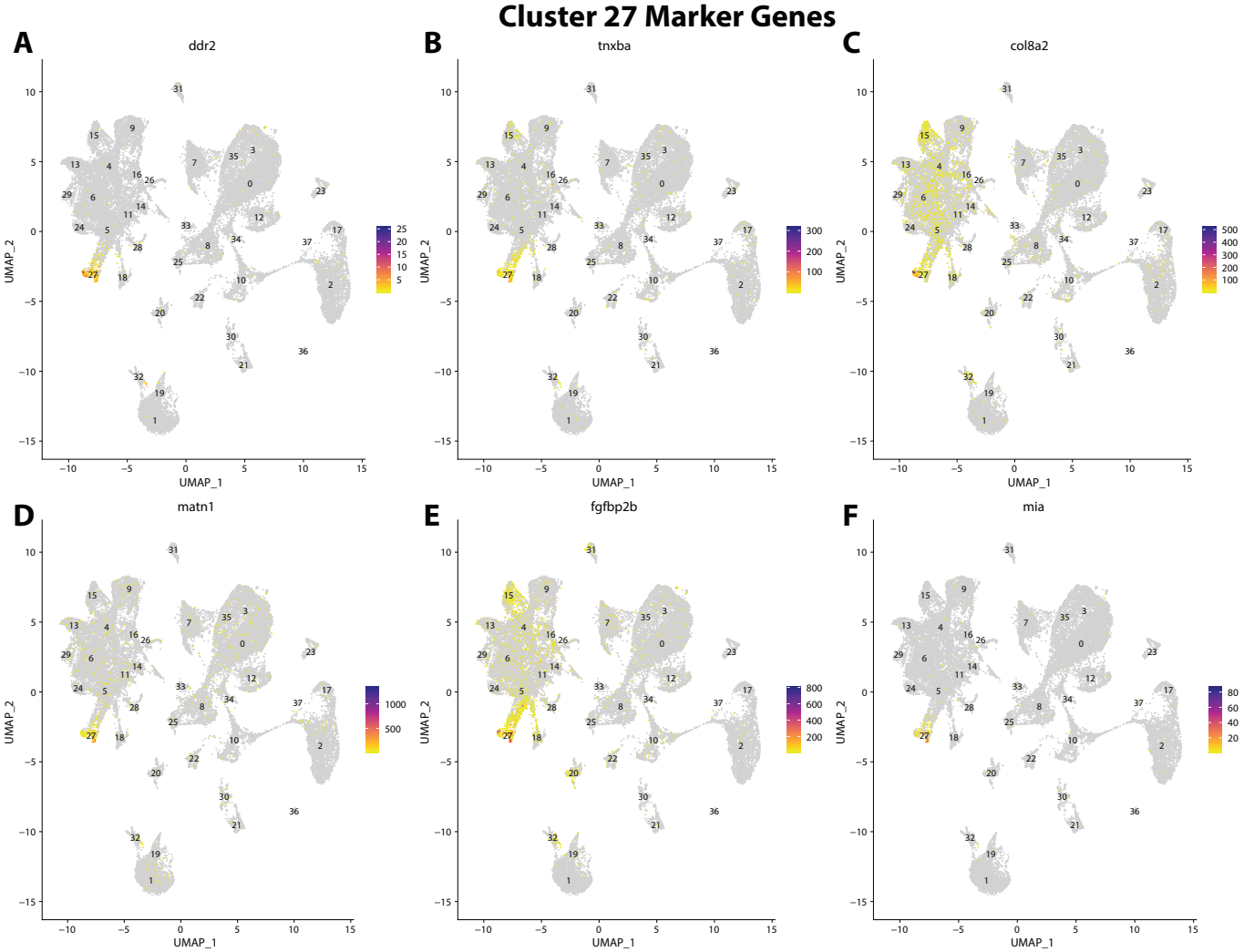


Figure S3.47: *Gulf pipefish cluster 27 markers.* FeaturePlots of top markers for cluster 27 identified from marker gene sets.

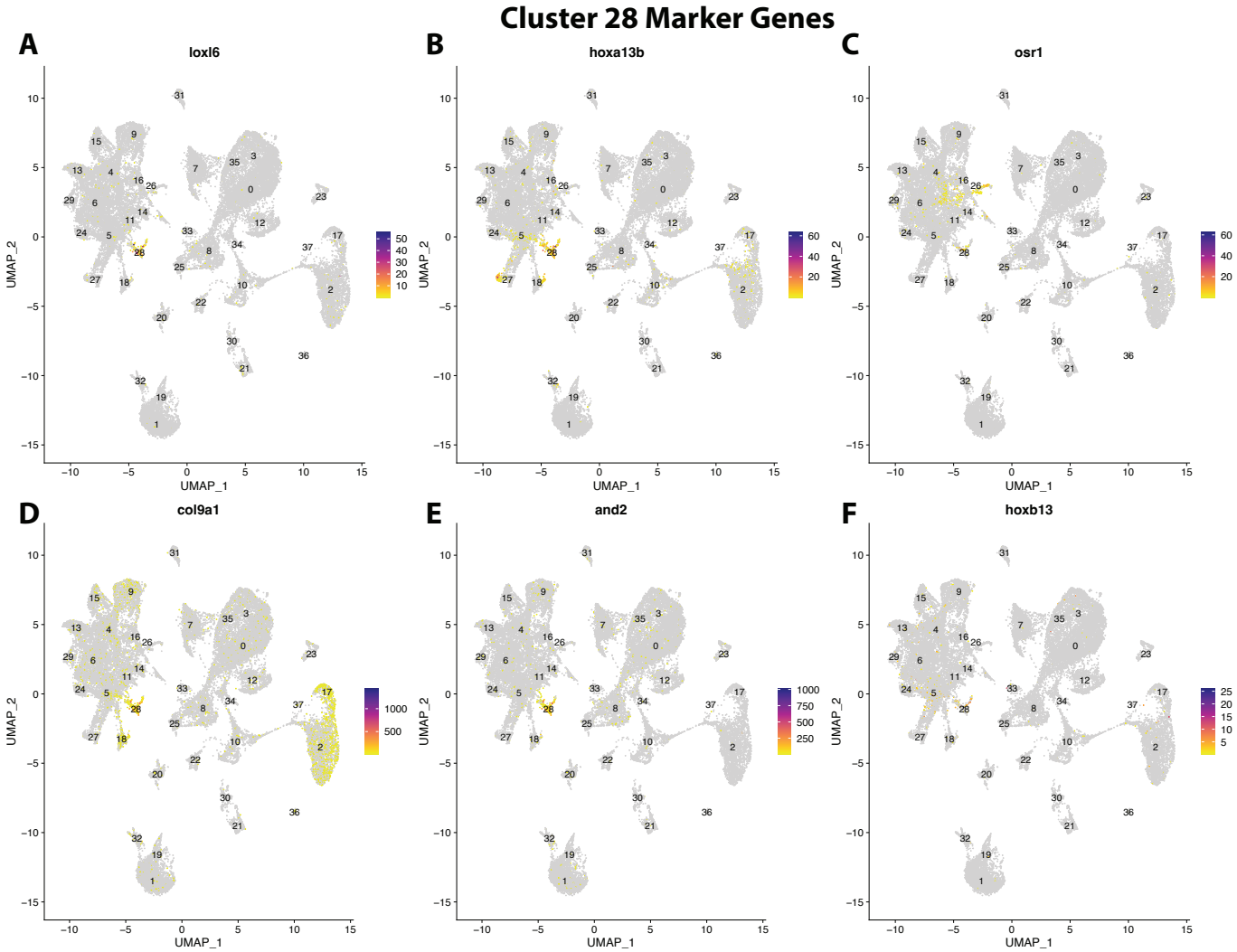


Figure S3.48: *Gulf pipefish cluster 28 markers.* FeaturePlots of top markers for cluster 28 identified from marker gene sets.

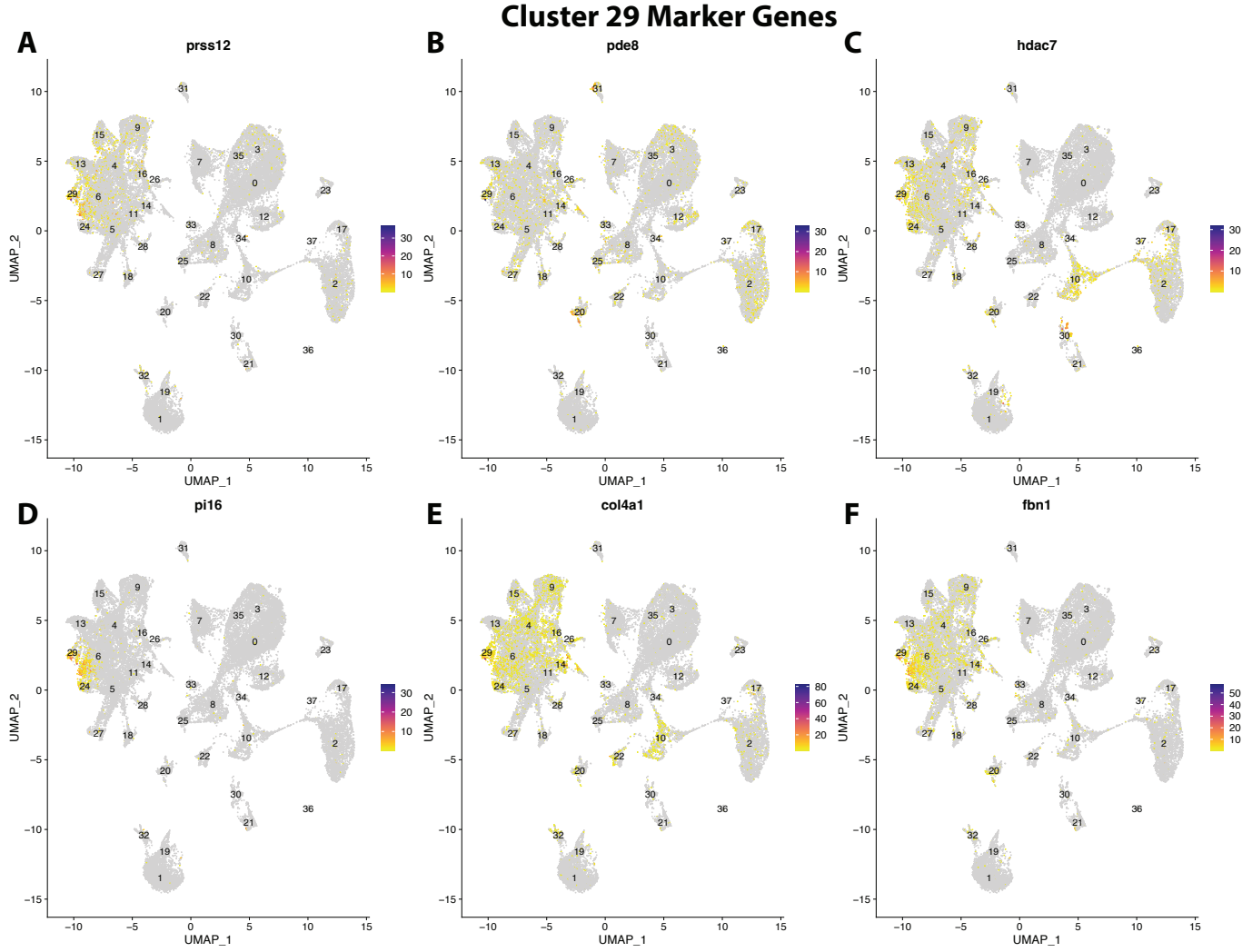


Figure S3.49: *Gulf pipefish cluster 29 markers.* FeaturePlots of top markers for cluster 29 identified from marker gene sets.

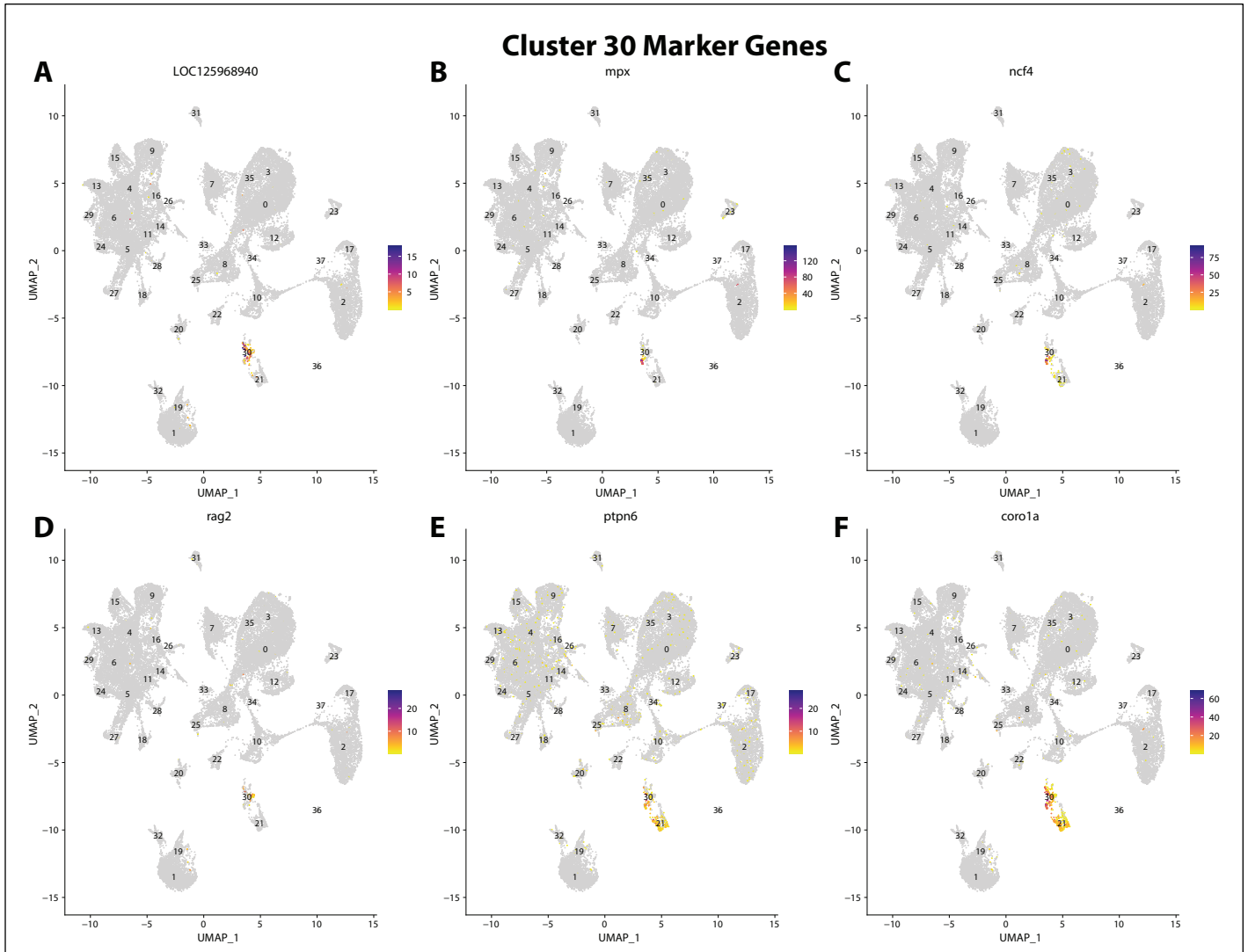


Figure S3.50: *Gulf pipefish cluster 30 markers.* FeaturePlots of top markers for cluster 30 identified from marker gene sets.

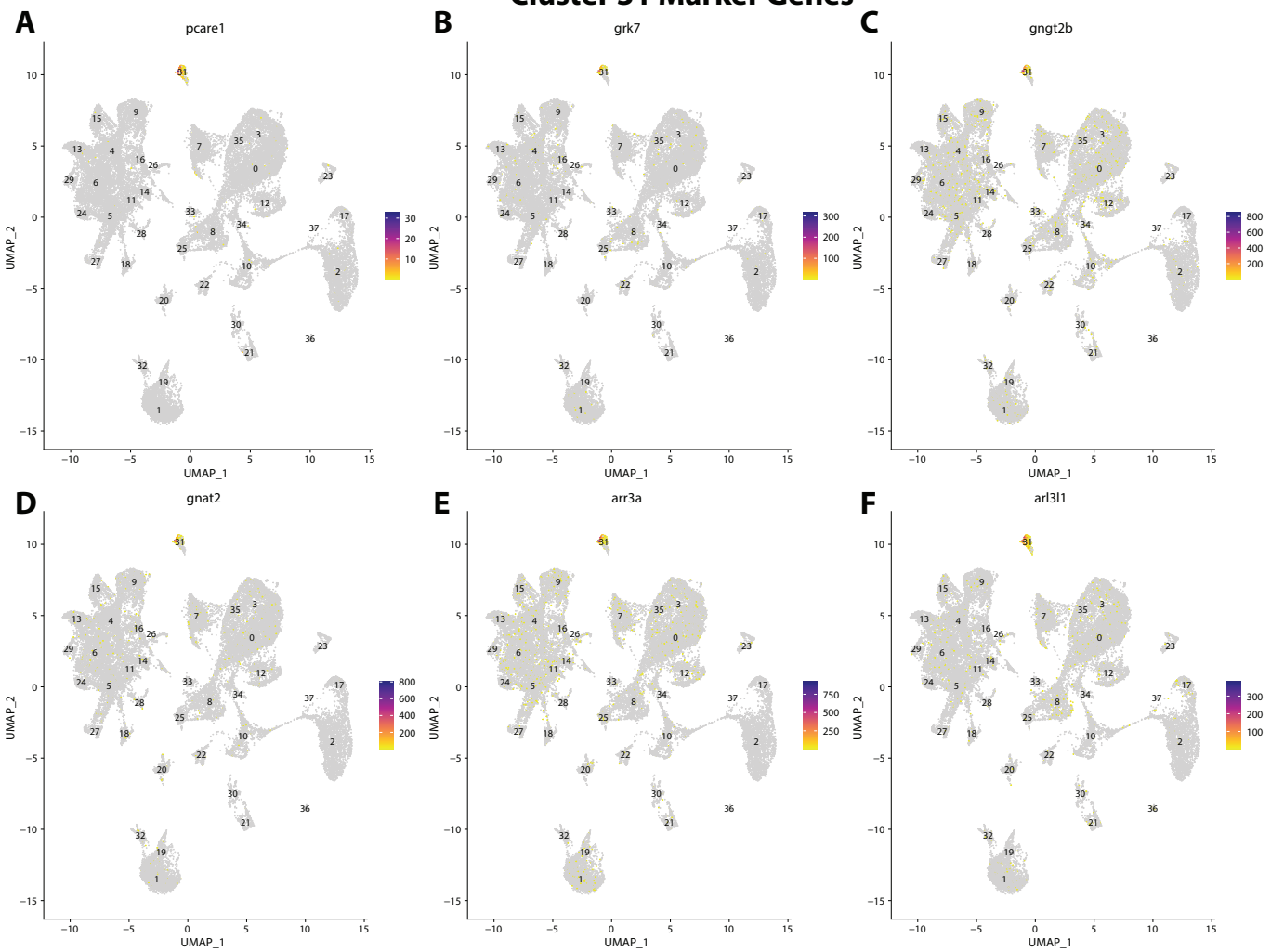


Figure S3.51: *Gulf pipefish cluster 31 markers.* FeaturePlots of top markers for cluster 31 identified from marker gene sets.

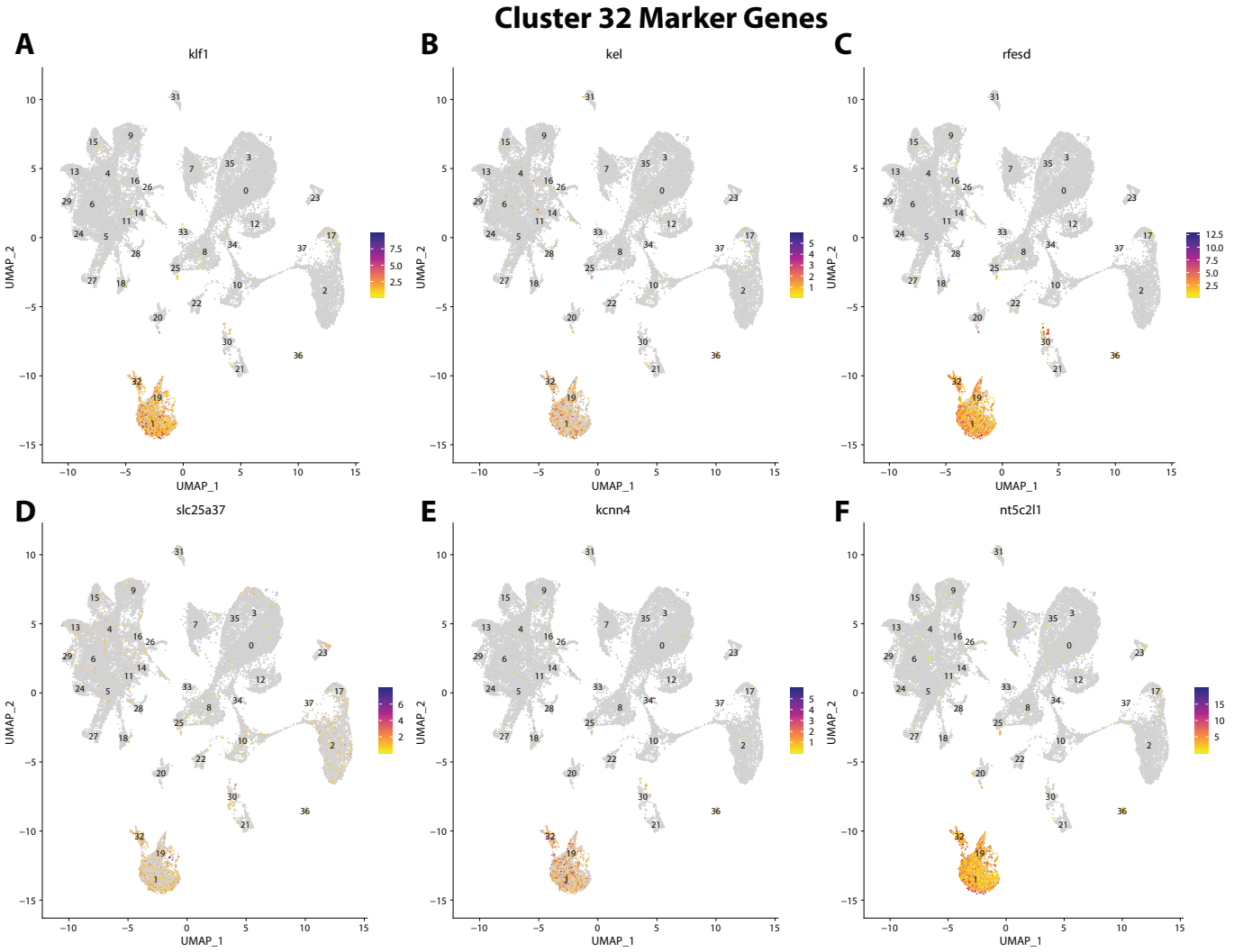


Figure S3.52: *Gulf pipefish cluster 32 markers.* FeaturePlots of top markers for cluster 32 identified from marker gene sets.

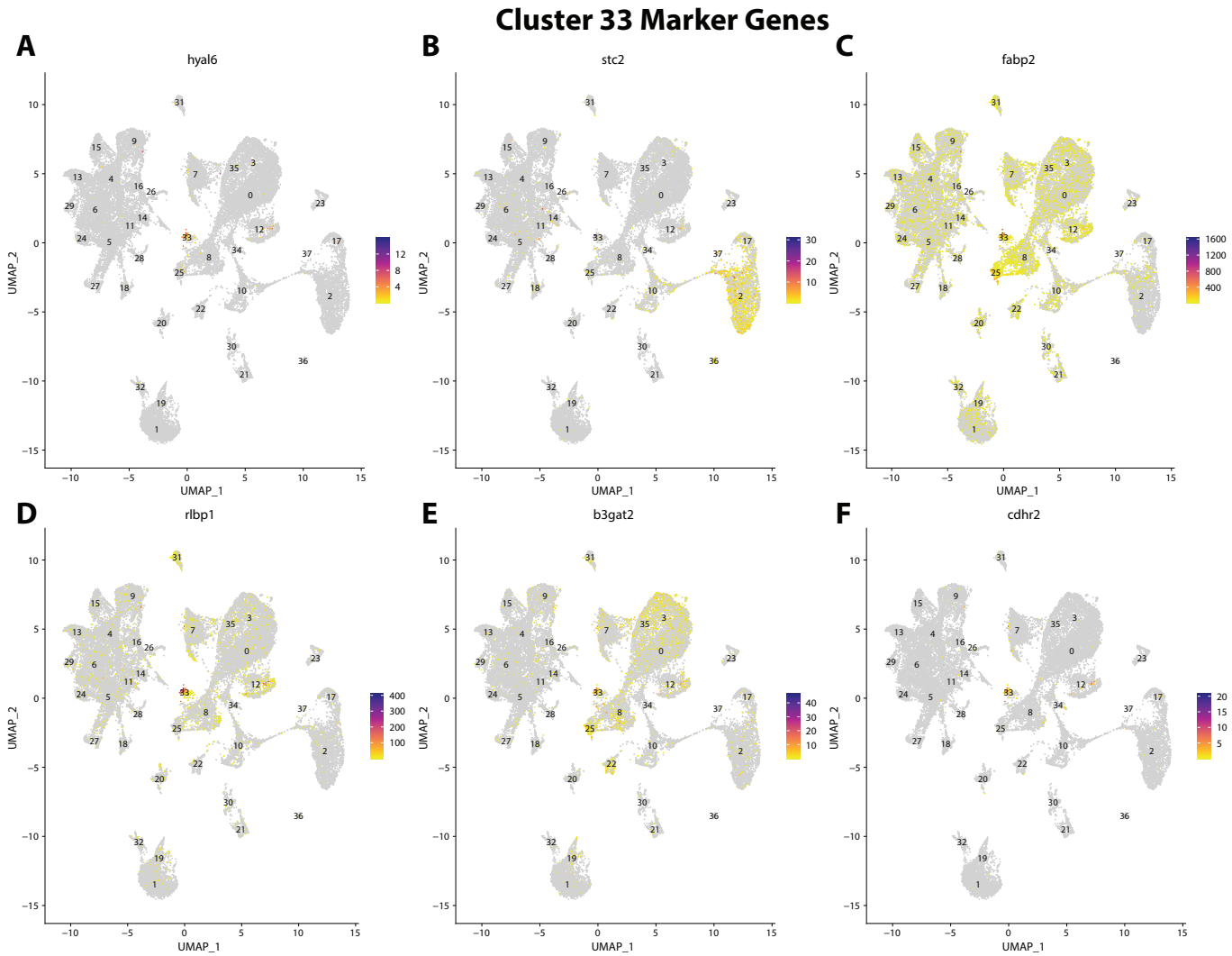


Figure S3.53: *Gulf pipefish cluster 33 markers.* FeaturePlots of top markers for cluster 33 identified from marker gene sets.

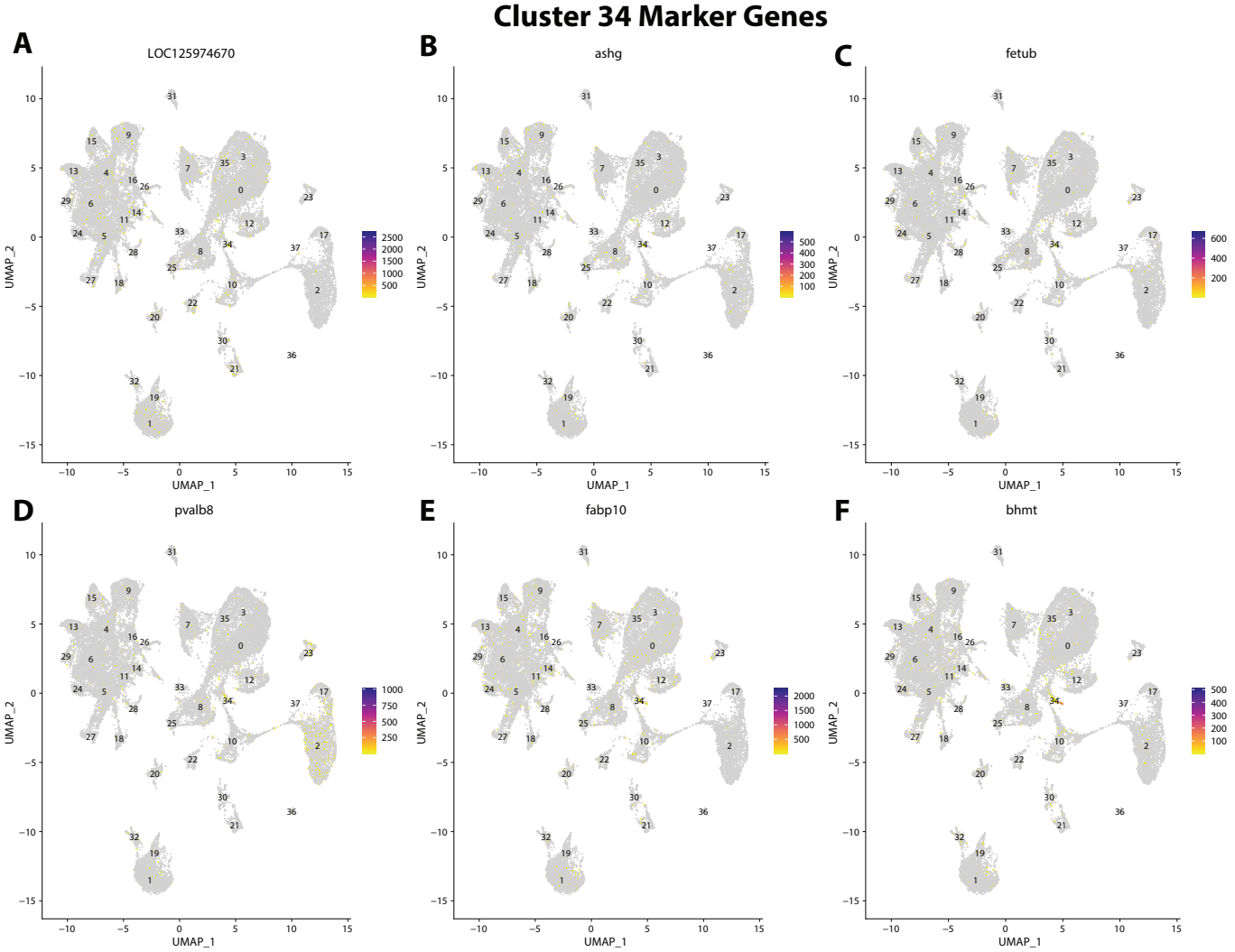


Figure S3.54: *Gulf pipefish cluster 34 markers.* FeaturePlots of top markers for cluster 34 identified from marker gene sets.

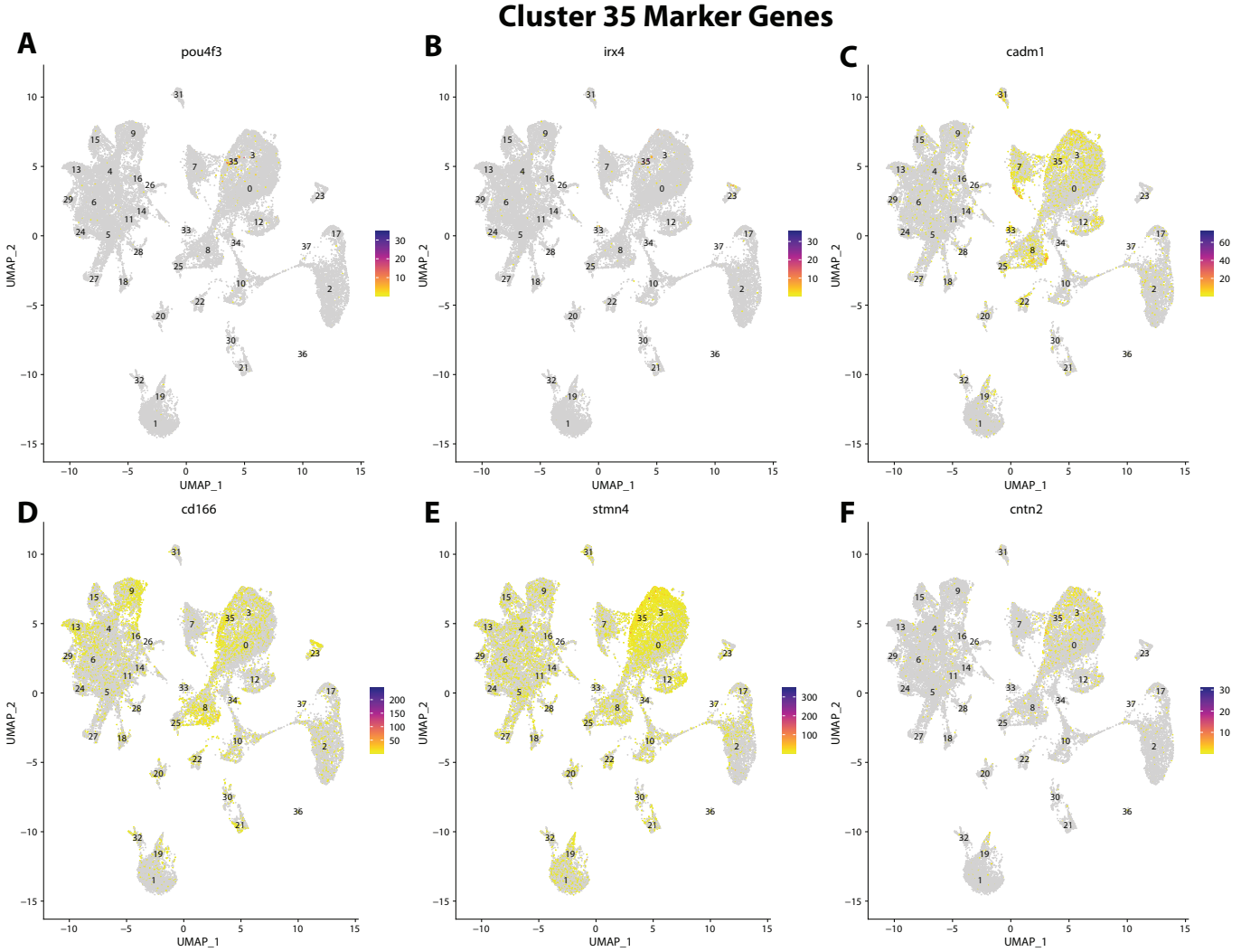


Figure S3.55: *Gulf pipefish cluster 35 markers.* FeaturePlots of top markers for cluster 35 identified from marker gene sets.

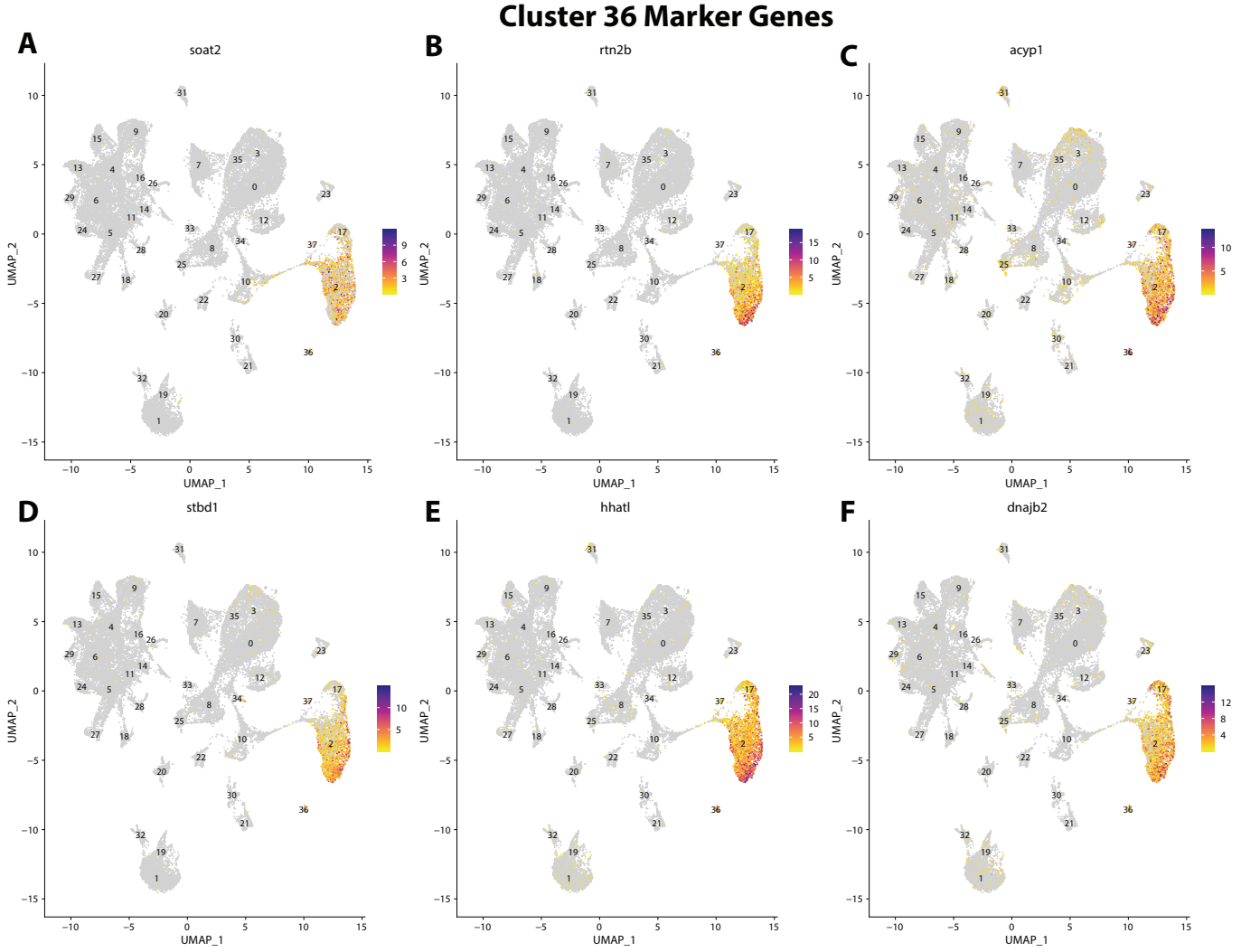


Figure S3.56: *Gulf pipefish cluster 36 markers.* FeaturePlots of top markers for cluster 36 identified from marker gene sets.

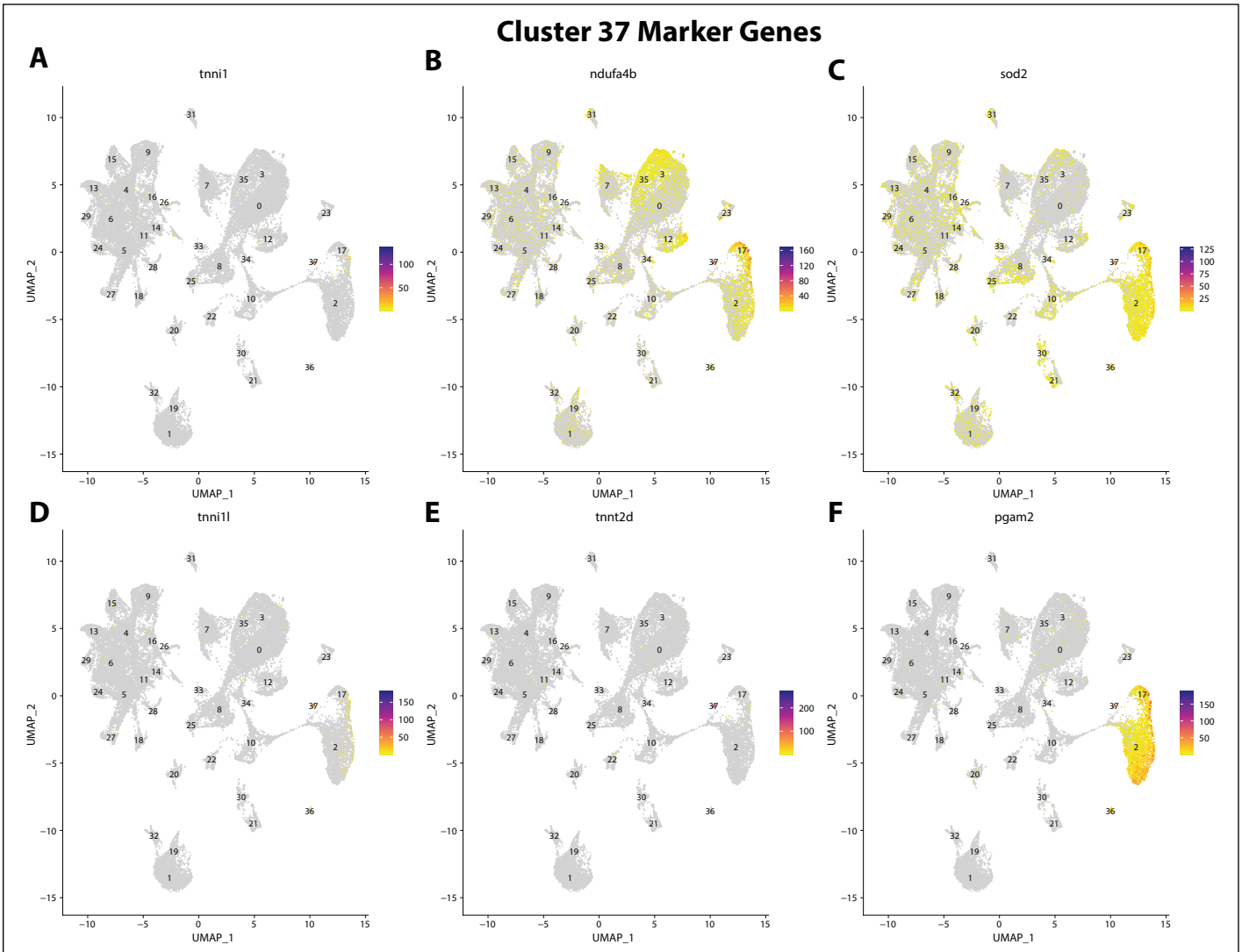


Figure S3.57: *Gulf pipefish cluster 37 markers.* FeaturePlots of top markers for cluster 37 identified from marker gene sets.

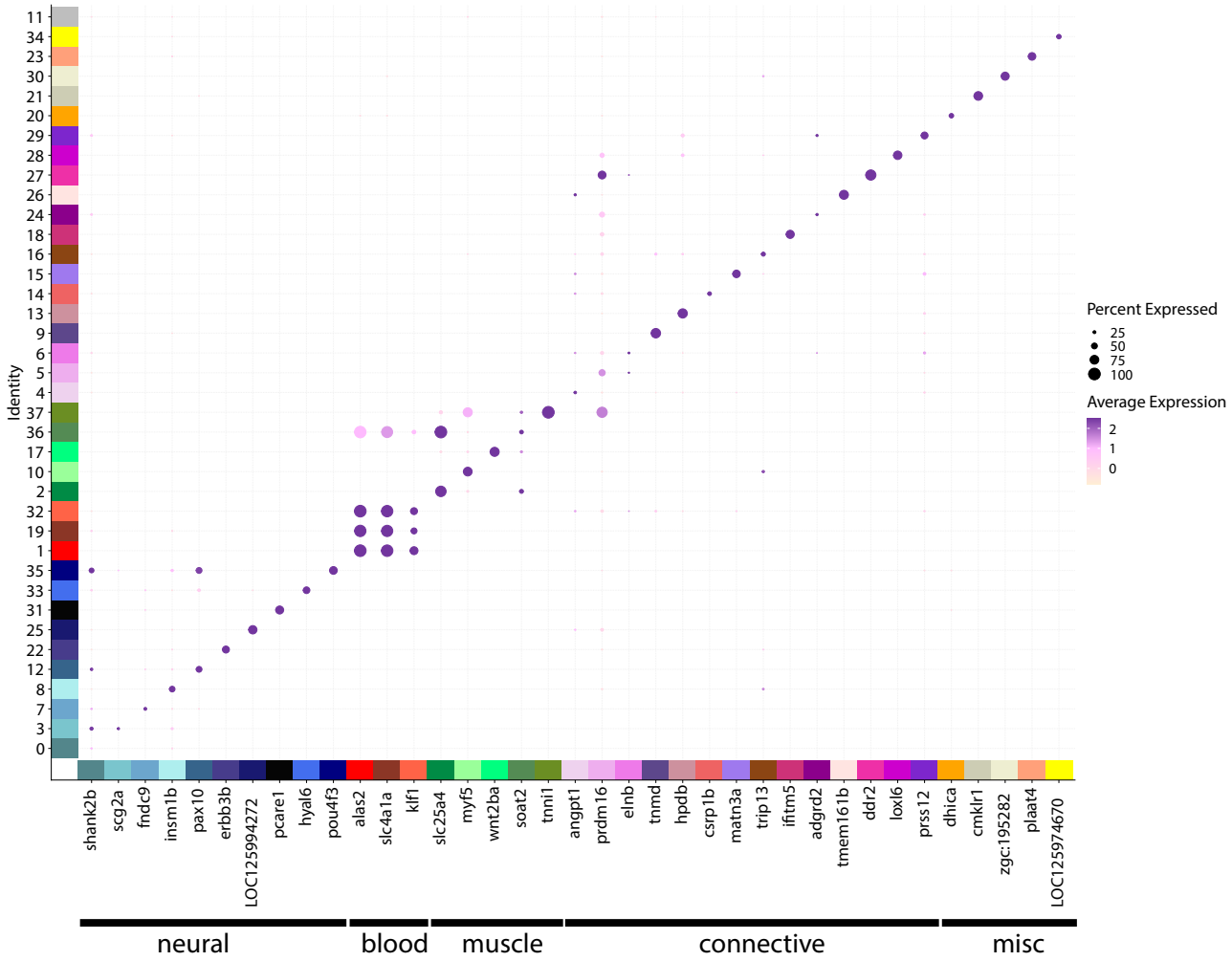


Figure S3.58: *Marker genes define cell identity and are a resource for cell cluster exploration*. This dotplot shows the marker genes on the x-axis, and the cell cluster on the y-axis. The size of each dot indicates the percentage of cells that express the gene in each cluster. The darkness of each dot represents the average expression of the gene.

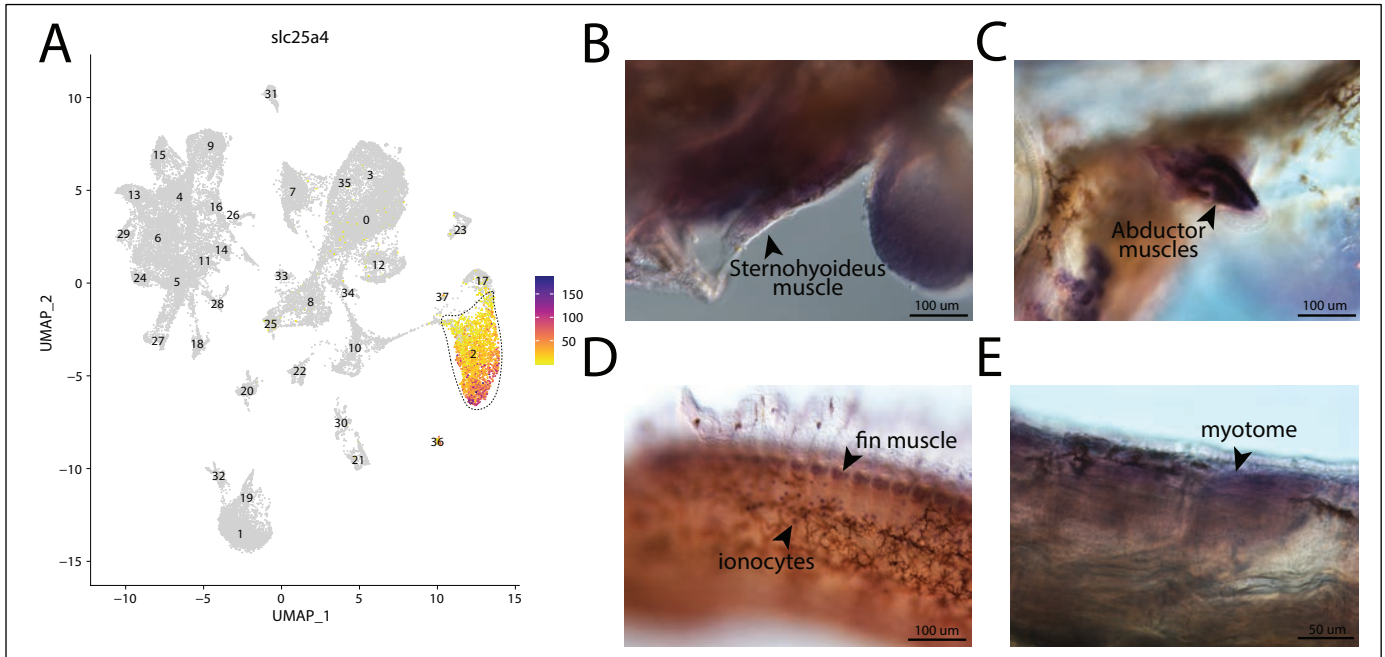


Figure S3.59: *slc25a4, a marker for cluster 2 is expressed in muscle cells in bay pipefish.* Atlas expression is limited to cluster 2 (A). *In situ* hybridization experiments revealed muscle staining of *slc25a4* in bay pipefish (B-D). Wild caught bay pipefish from the onset of craniofacial elongation were used for the experiment. Embryos were mounted laterally to highlight expression domains.

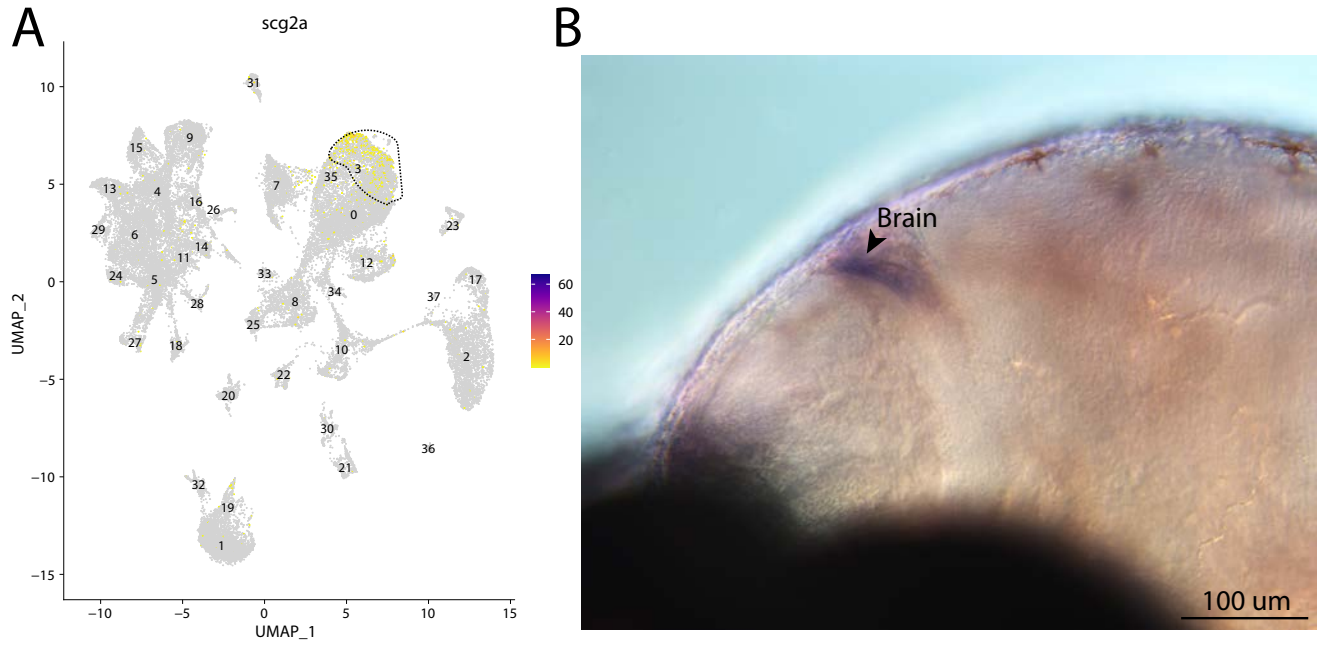


Figure S3.60: *scg2a, a marker for cluster 3 is expressed in the brain of bay pipefish.* Atlas expression is limited to cluster 3 (A). *In situ* hybridization experiments revealed brain staining of *scg2a* in bay pipefish (B-D). Wild caught bay pipefish from the onset of craniofacial elongation were used for the experiment. Embryos were mounted laterally to highlight expression domains.

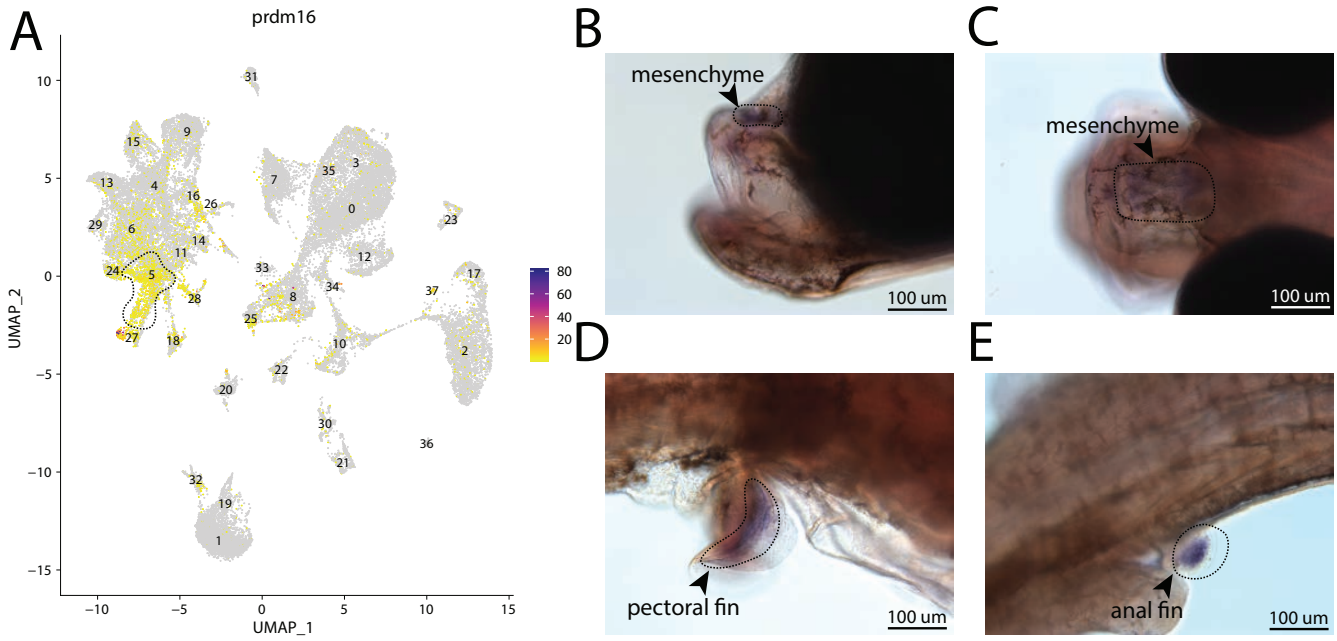


Figure S3.61: *prdm16, a marker for cluster 5 is expressed in the head and fins of bay pipefish.* Strong expression of *prdm16* is observed in cluster 5, osteochondrogenic mesenchyme (A). *In situ* hybridization experiments revealed mesenchymal staining of *prdm16* in the fins and face of bay pipefish (B-D). Wild caught bay pipefish from the onset of craniofacial elongation were used for the experiment. Embryos were mounted laterally (B, D, E) and with a dorsal view (C) to highlight expression domains.

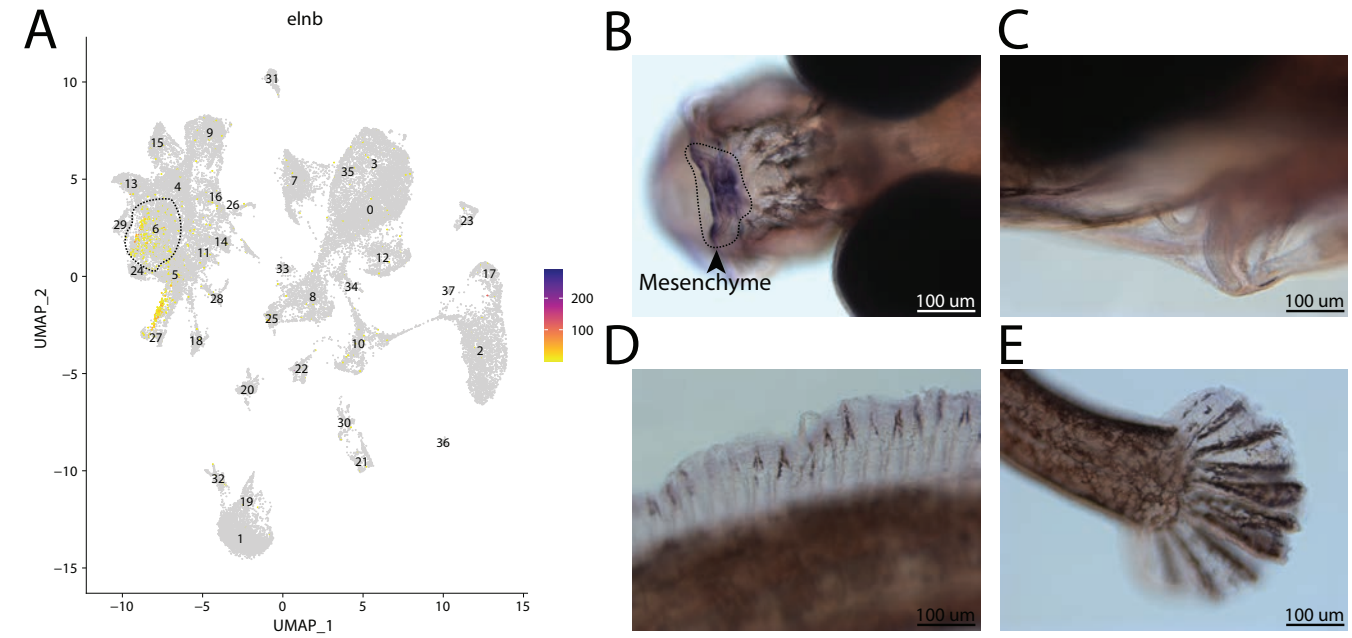


Figure S3.62: *elnb, a marker for cluster 6 is expressed in the head of bay pipefish.* Strong expression of *elnb* is observed in cluster 6, osteochondrogenic mesenchyme (A). *In situ* hybridization experiments revealed mesenchymal staining of *elnb* in the face of bay pipefish (B-D). Wild caught bay pipefish from the onset of craniofacial elongation were used for the experiment. Embryos were mounted dorsally (B) and laterally (C-E) to show the lack of staining in the ceratohyal, dorsal fin, and caudal fin.

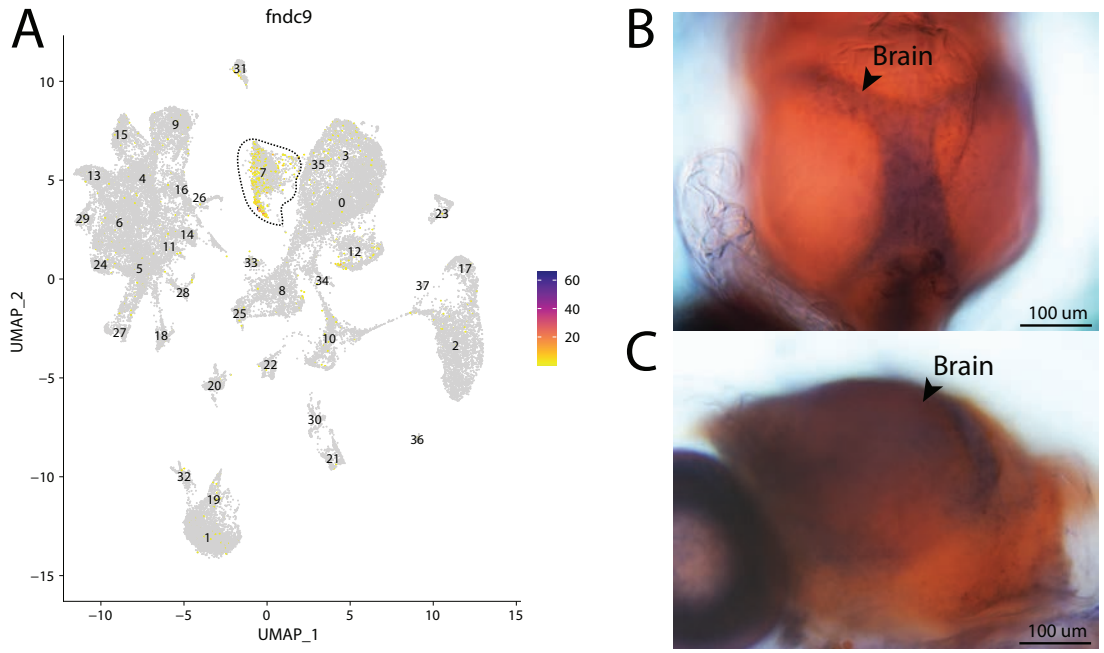


Figure S3.63: *fndc9, a marker for cluster 7 is expressed in the brain of bay pipefish.* Expression of *fndc9* is limited to cluster 7, brain cells, in the atlas (A). *In situ* hybridization experiments revealed brain staining of *fndc9* in Gulf pipefish (B-C). One day post spawn Gulf pipefish larvae were used for the experiment. Embryos were mounted to highlight two views of the brain staining, one dorsal (B) and one lateral (C).

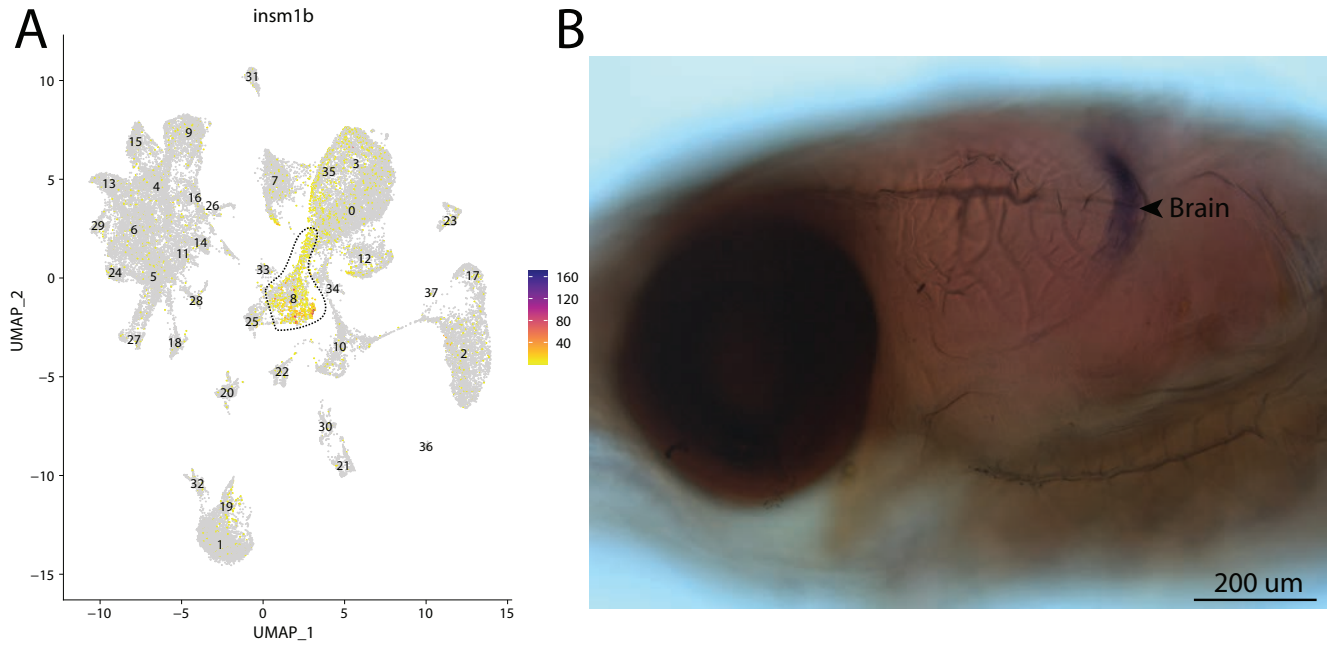


Figure S3.64: *insm1b, a marker for cluster 8 is expressed in the brain of bay pipefish.* Strong expression of *insm1b* is observed in cluster 8, brain progenitor cells, in the atlas (A). *In situ* hybridization experiments revealed brain staining of *insm1b* in Gulf pipefish (B-D). Five-day post spawn Gulf pipefish were used for the experiment. Embryos were mounted to highlight the brain staining.

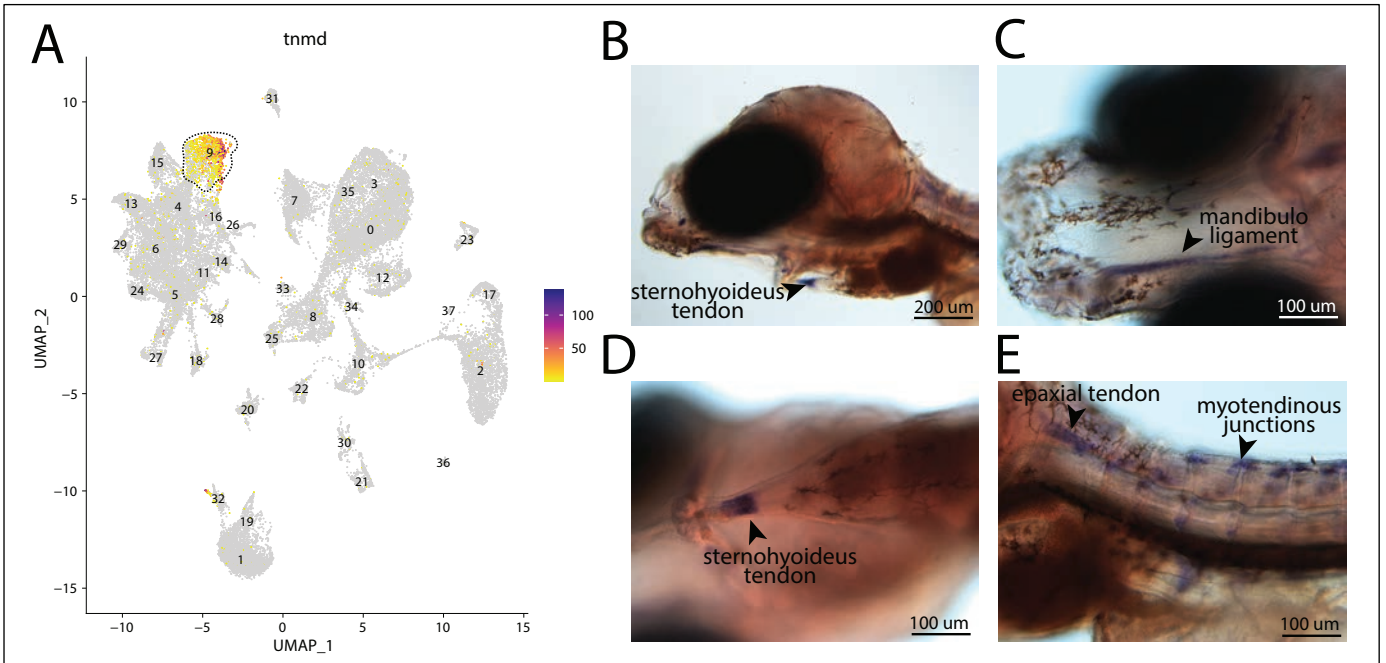


Figure S3.65: *tnmd, a marker for cluster 9 is expressed in tendons and ligaments of bay pipefish.* Strong expression of *tnmd* is observed in cluster 9, tenocyte and ligament cells, in the atlas (A). *In situ* hybridization experiments revealed tendon and ligament staining of *tnmd* in bay pipefish (B-E). Wild caught bay pipefish embryos from the onset of craniofacial elongation were used for the experiment. Embryos were mounted laterally (B, E) and ventrally (C, D) to showcase tendon and ligament staining.

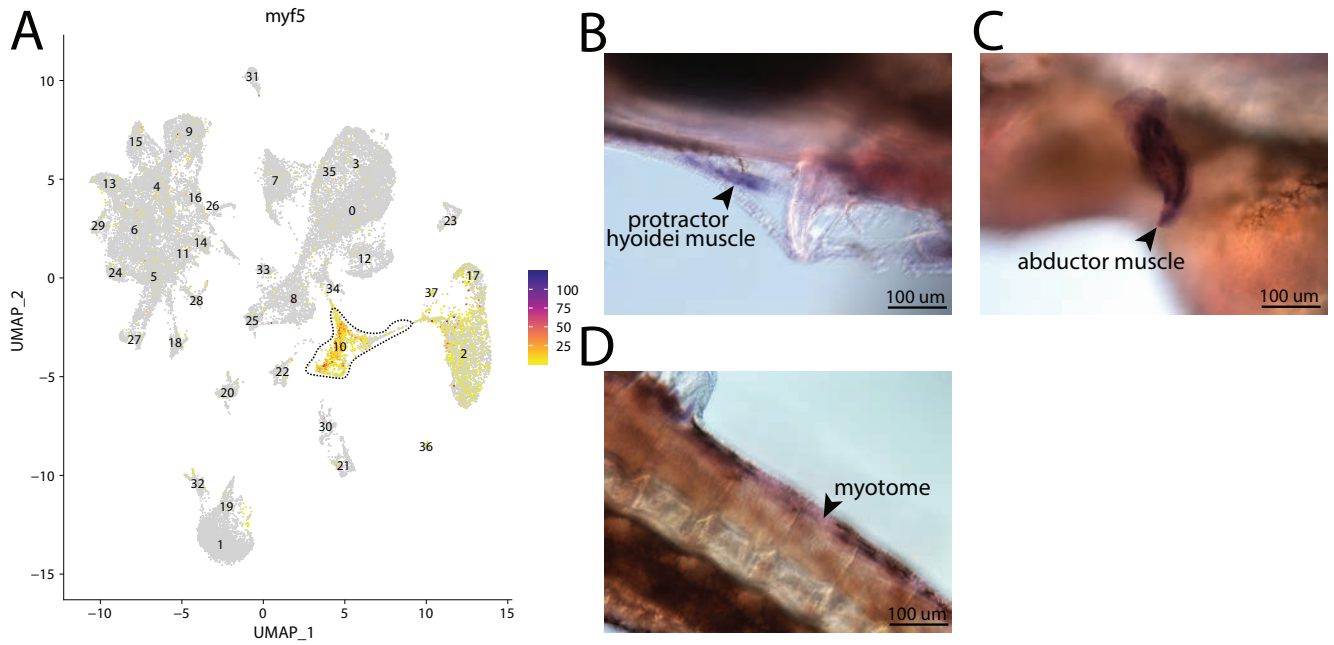


Figure S3.66: *myf5, a marker for cluster 10 is expressed in the muscle of bay pipefish.* Strong expression of *myf5* is observed in cluster 10, muscle progenitor cells, in the atlas (A). *In situ* hybridization experiments revealed muscle staining of *myf5* in bay pipefish (B-D). Wild caught bay pipefish embryos from the onset of craniofacial elongation were used for the experiment. Embryos were mounted laterally to highlight muscle staining.

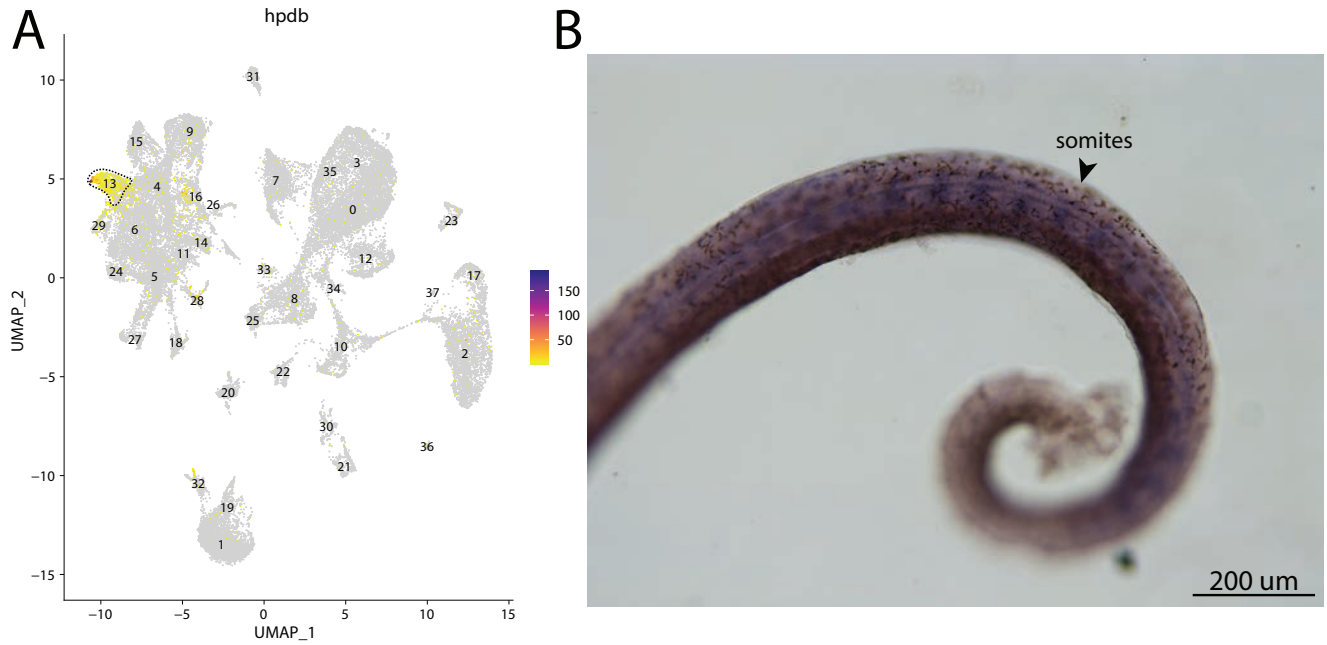


Figure S3.67: *hpbd, a marker for cluster 13 is expressed in somites of bay pipefish.* Strong expression of *hpbd* is observed in cluster 13, fibroblasts, in the atlas (A). *In situ* hybridization experiments revealed somite staining of *hpbd* in bay pipefish (B). Wild caught embryos from bay pipefish at the mid-somitogenesis stage were used for the experiment. Embryos were mounted laterally to highlight muscle staining.

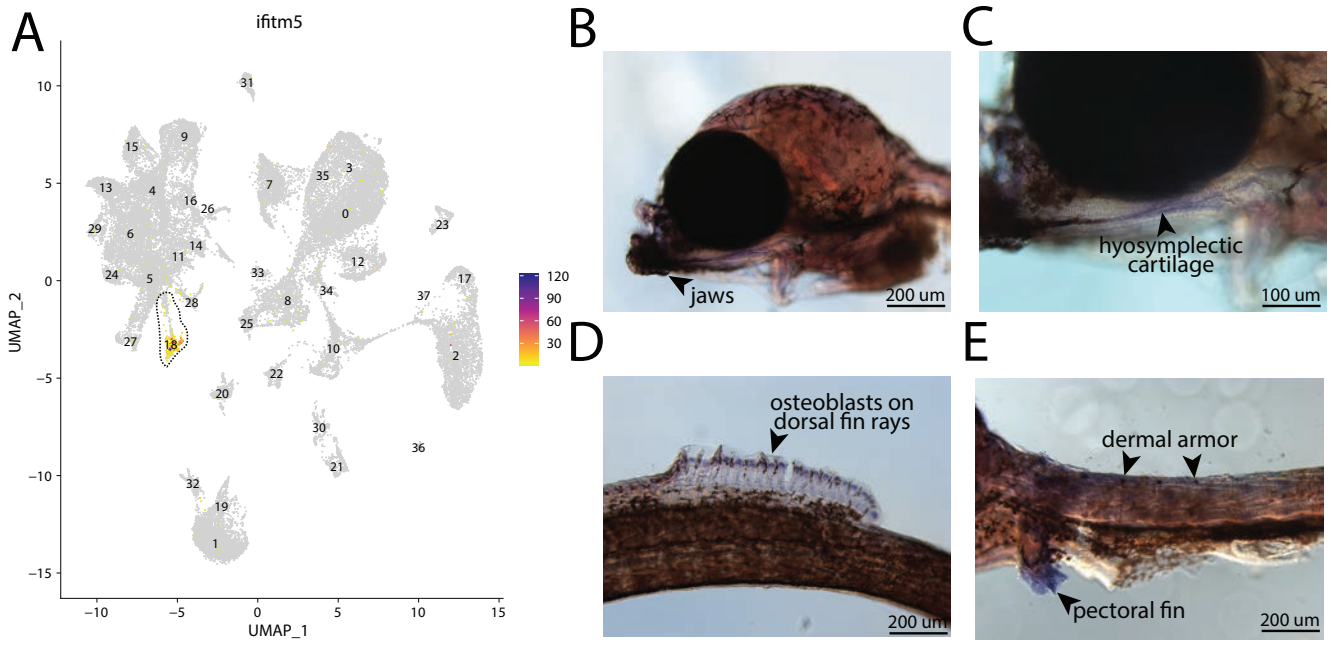


Figure S3.68: *ifitm5, a marker for cluster 18, is expressed in bay pipefish osteoblasts.* Specific expression of *ifitm5* is observed in cluster 18, osteoblasts, in the atlas (A). *In situ* hybridization experiments revealed bone staining of *ifitm5* in bay pipefish (B-E). Wild caught embryos from bay pipefish at the onset of craniofacial elongation were used for the experiment. Embryos were mounted laterally to highlight muscle staining.

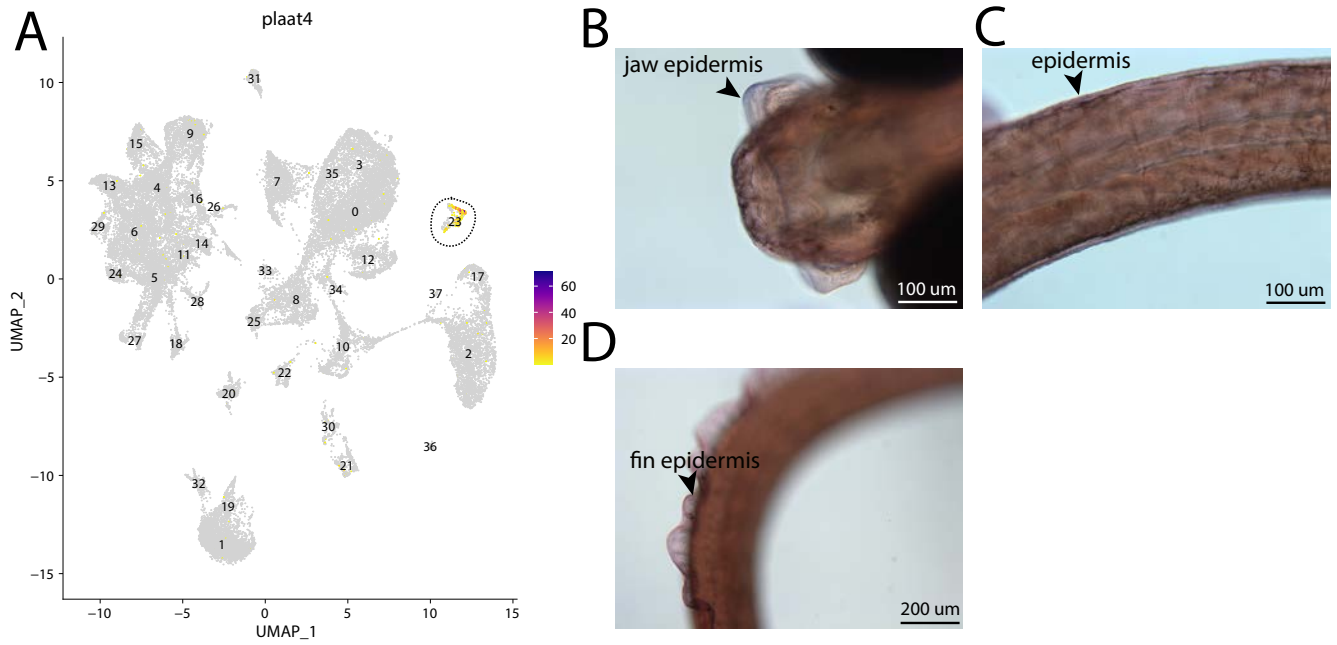


Figure S3.69: *plaat4, a marker for cluster 23, is expressed in weakly expressed in bay pipefish epidermal cells.* Specific expression of *plaat4* is observed in cluster 23, epidermal cells, in the atlas (A). *In situ* hybridization experiments revealed epidermal staining of *plaat4* in bay pipefish (B-D). Wild caught embryos from bay pipefish at the onset of craniofacial elongation were used for the experiment. Embryos were mounted ventrally (B) and laterally (C, D) to epidermal muscle staining.

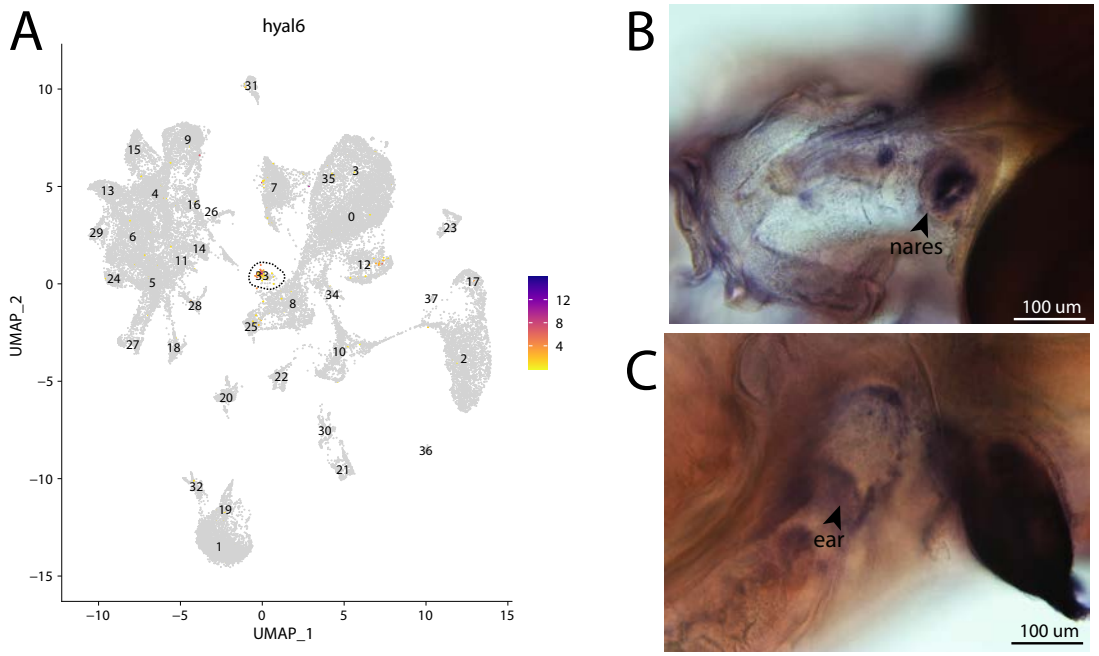


Figure S3.70: *hyal6, a marker for cluster 33, is strongly expressed in bay pipefish hair cells.* Specific expression of *hyal6* is observed in cluster 33, hair cells, in the atlas (A). *In situ* hybridization experiments revealed staining of *hyal6* in bay pipefish nares and ears (B-C). Wild caught embryos from bay pipefish from one day post spawn larvae were used for the experiment. Embryos were mounted dorsally (B) and laterally (C) to highlight staining in nares and ears respectively.

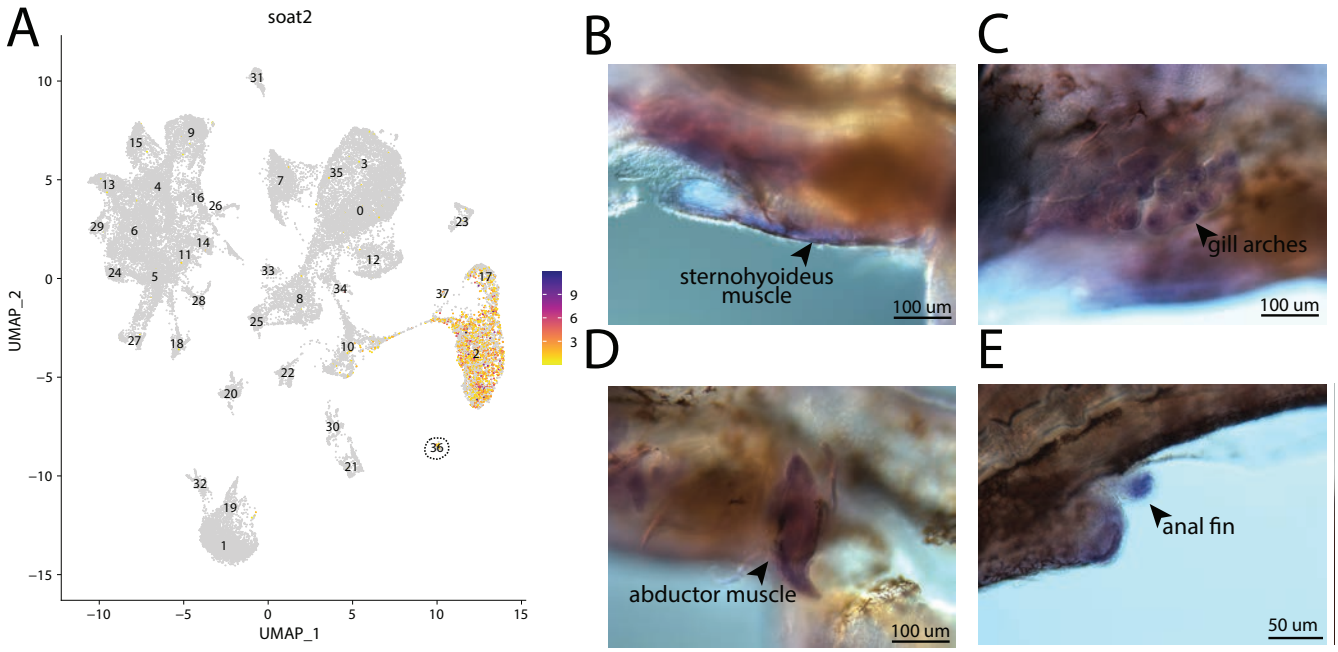


Figure S3.71: *soat2, a marker for cluster 36, is strongly expressed in bay pipefish muscle cells.* Expression of *soat2* is observed in cluster 36, muscle cells, in the atlas (A). *In situ* hybridization experiments revealed staining of *soat2* in bay pipefish muscle and fins (B-E). Wild caught embryos from bay pipefish at the onset of craniofacial elongation were used for the experiment. Embryos were mounted laterally to highlight staining in muscle and fins.

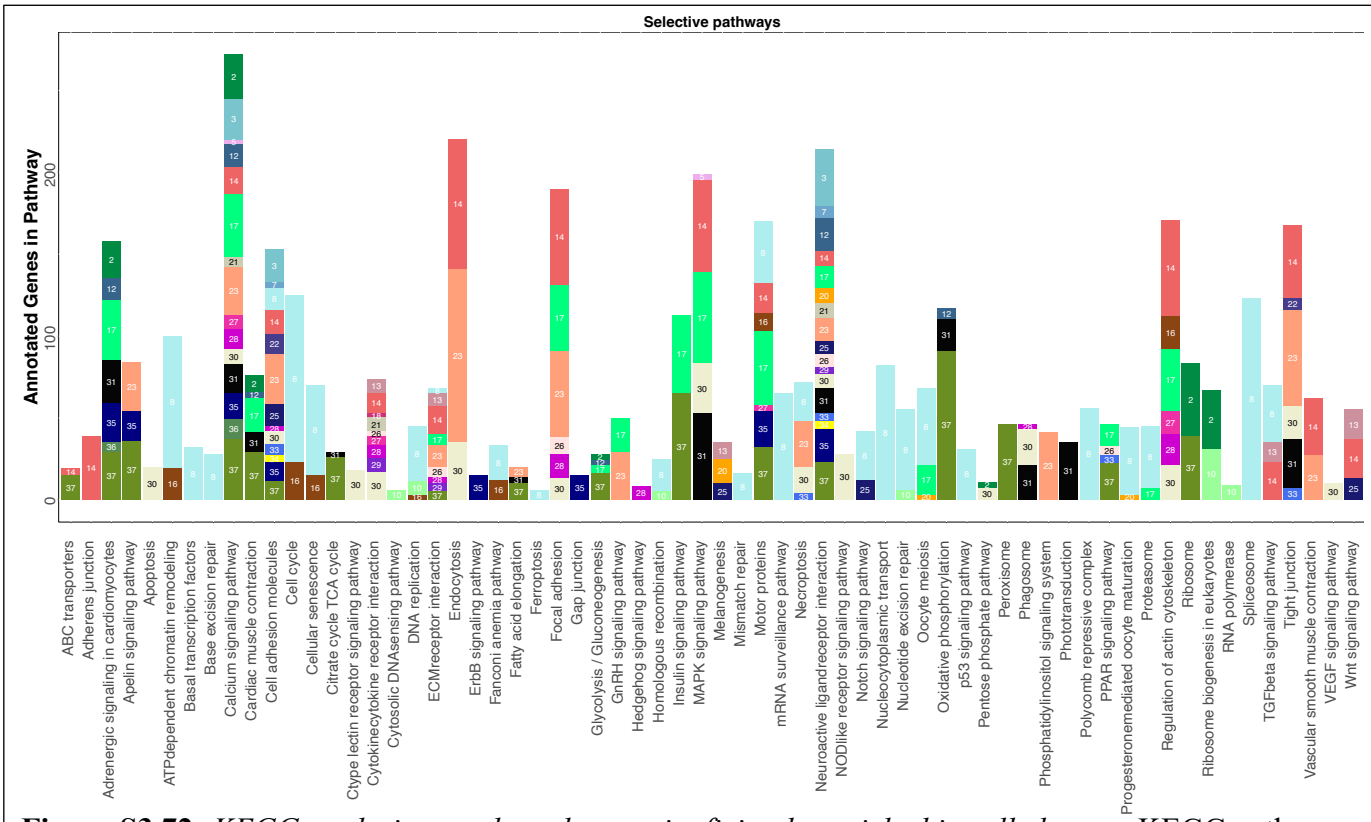


Figure S3.72: *KEGG analysis reveals pathways signficiantly enriched in cell clusters.* KEGG pathways are listed on the x-axis. For each cluster, the number of genes that belong to the specific KEGG pathway is represented by the height of each bar. Cell clusters are labeled by both color and their designated number. For clarity, only cell clusters with more than three annotated genes in each pathway are shown and certain pathways involving infection and metabolism are removed.

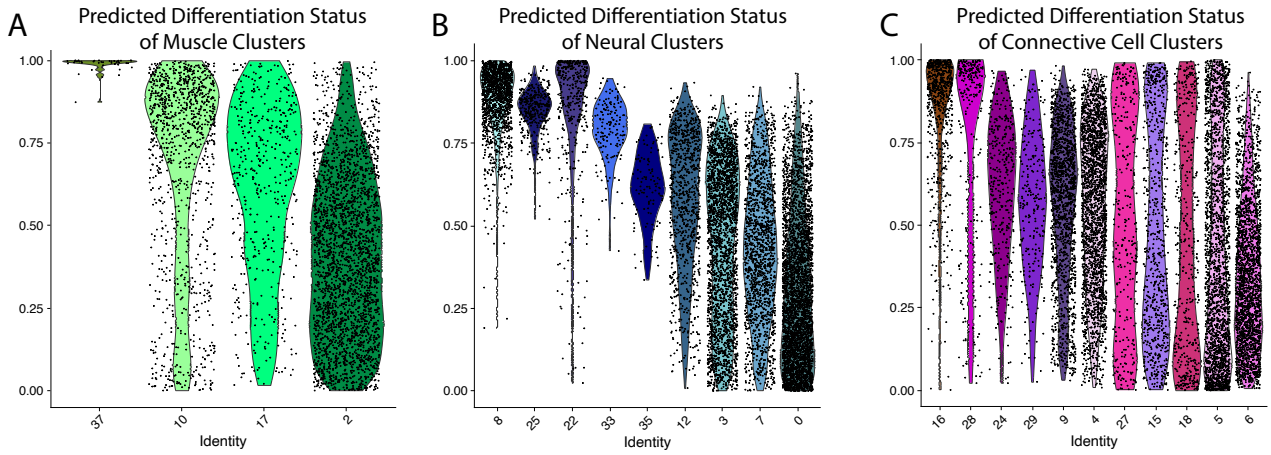


Figure S3.73: *Clusters 8, 10, and 16 are composed of largely undifferentiated cells.* Differentiation state of cell clusters was measured using CytoTRACE, where an output of 1 indicates an undifferentiated cell and a value of 0 suggests a cell is fully differentiated. Cells are plotted on a violin plot with individual cell predications shown as dots. Panels show results from muscle (A), neural (B), and connective (C) clusters.

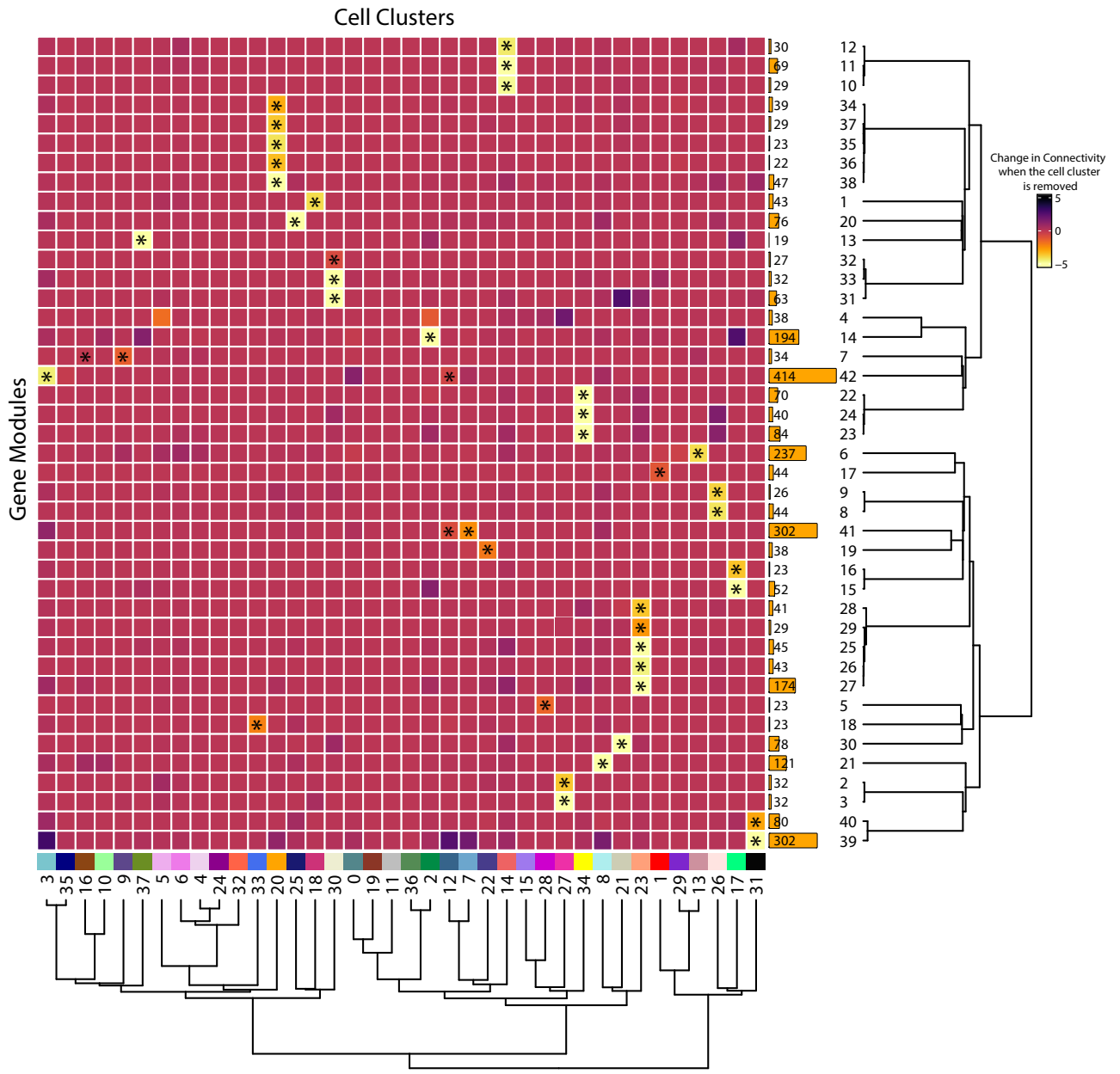


Figure S3.74: *Cell clusters drive gene module connectivity.* The heatmap shows the change in gene network connectivity when individual cell clusters are removed. Cell clusters are in the columns and gene modules are in rows. The coloration scale is based on the change in connecitvity when a cell cluster is removed, with yellow indicating a large change in connectivity. The modules and clusters are clustered using the Pearson distance method. Asterisks indicate cluster-module pairs that are significant using a one-way permutation test and FDR correction. The size of the gene modules are shown to the right.

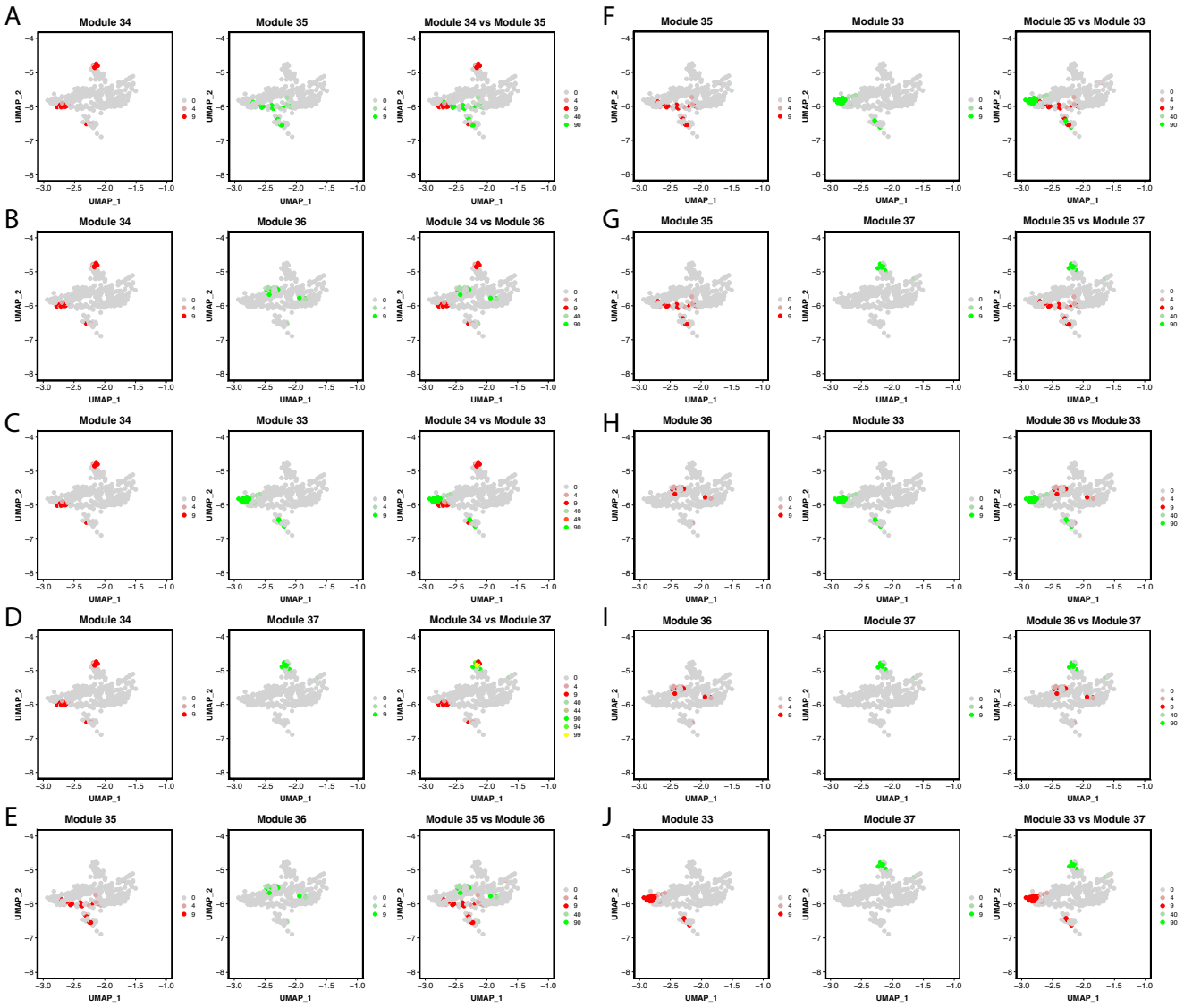


Figure S3.75: *Gene module expression comparison with the Pigment Cluster reveals potential elusive cell types within the cluster.* Genes within the network are combined into one feature then plotted on a feature plot. In this case, UMAP dimensions one and two are the x and y-axis accordingly, dots represent each pigment cell, and the color of the dots represents the expression of the module. Cells that express gene modules are either colored red or green (when modules are plotted on separate plots). When plotted on the same plot, cells that express both gene modules are yellow. Panels A through J show each gene module comparison.

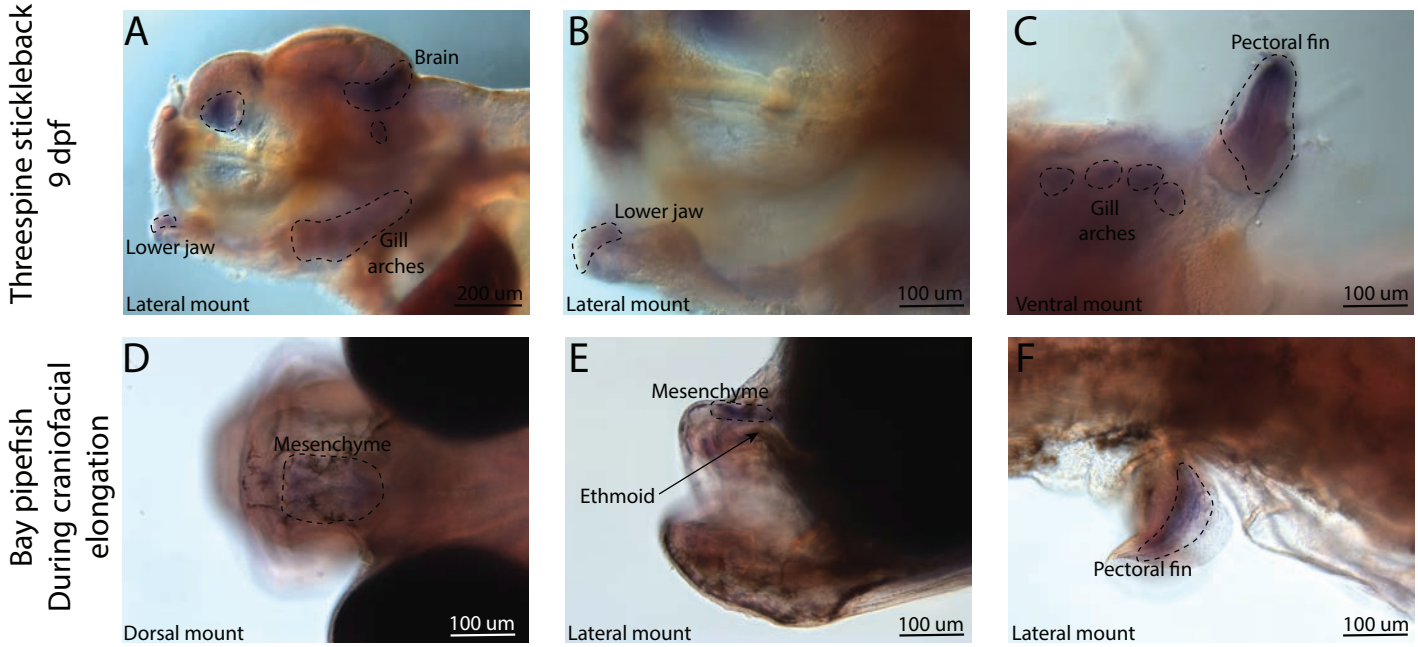


Figure S3.76: *Conserved fin prdm16 expression domains in stickleback and pipefish, but fishes have expression in different regions of the head. prdm16* expression was probed using *in situ* hybridization. Embryos used in this assay were similarly staged (9dpf in threespine stickleback and mid-craniofacial elongation in bay pipefish). Stickleback (A-C) expressed *prdm16* in the lower jaw, brain, gill arches, and fins. Bay pipefish also have fin staining (F), but craniofacial expression is limited to the mesenchyme above the ethmoid region.

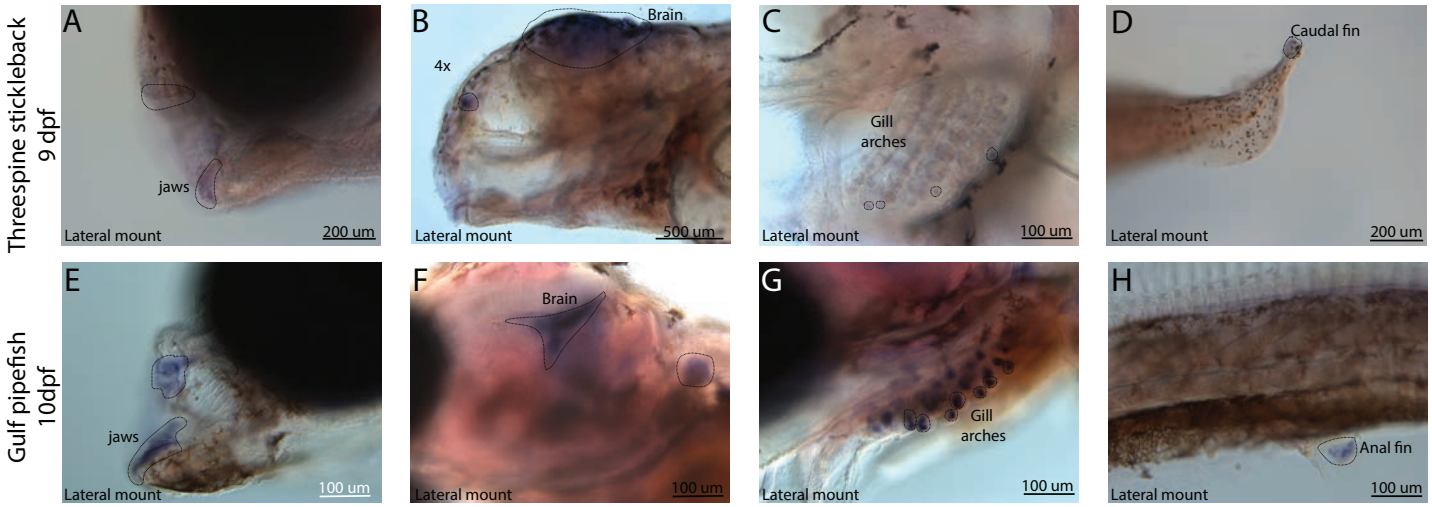


Figure S3.77: *Conserved expression domains of fgf22 in pipefish and stickleback fish. fgf22* expression was probed using *in situ* hybridization. Embryos used in this assay were similarly staged (9dpf threespine stickleback and 10dpf Gulf pipefish). Staining in the jaws (A, E), brain (B, F), gill arches (C, G), and fins (D, H) suggests similar *fgf22* expression domains in both species.

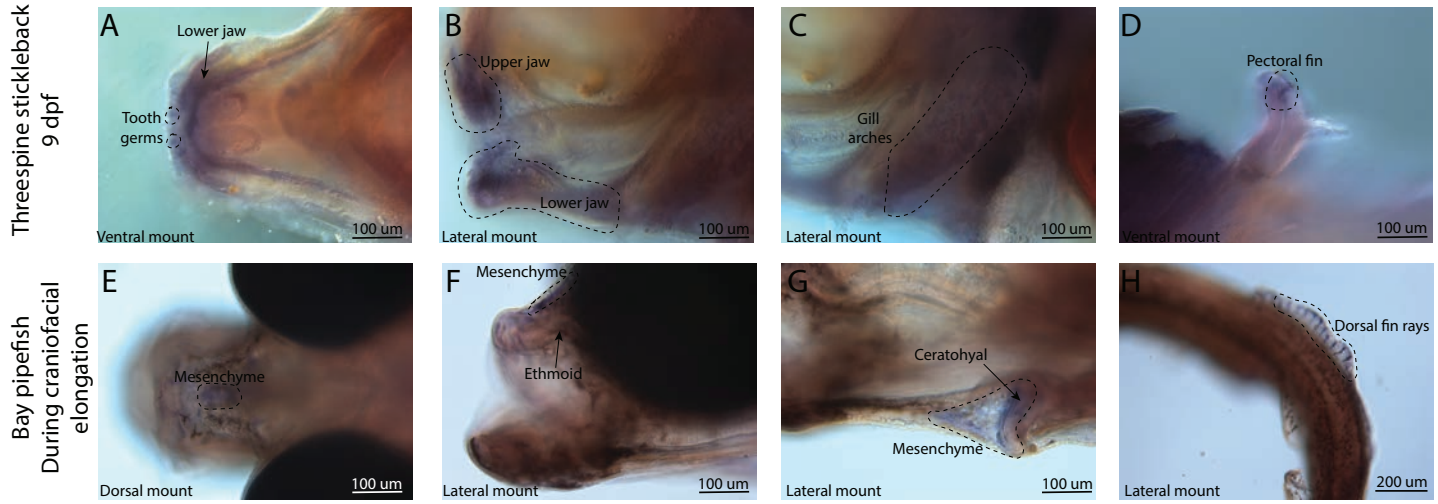


Figure S3.78: *Broad bmp4 craniofacial expression in stickleback, but bmp4 is restricted to ethmoid and ceratohyal regions in pipefish. bmp4* expression was probed using *in situ hybridization.* Embryos used in this assay were similarly staged (9dpf threespine stickleback and mid-craniofacial elongation staged bay pipefish). Stickleback have *bmp4* staining in tooth germs (A), the jaws (A-B), and gill arches (C). However, *bmp4* expression is restricted in bay pipefish to mesenchyme above the ethmoid and ceratohyal.

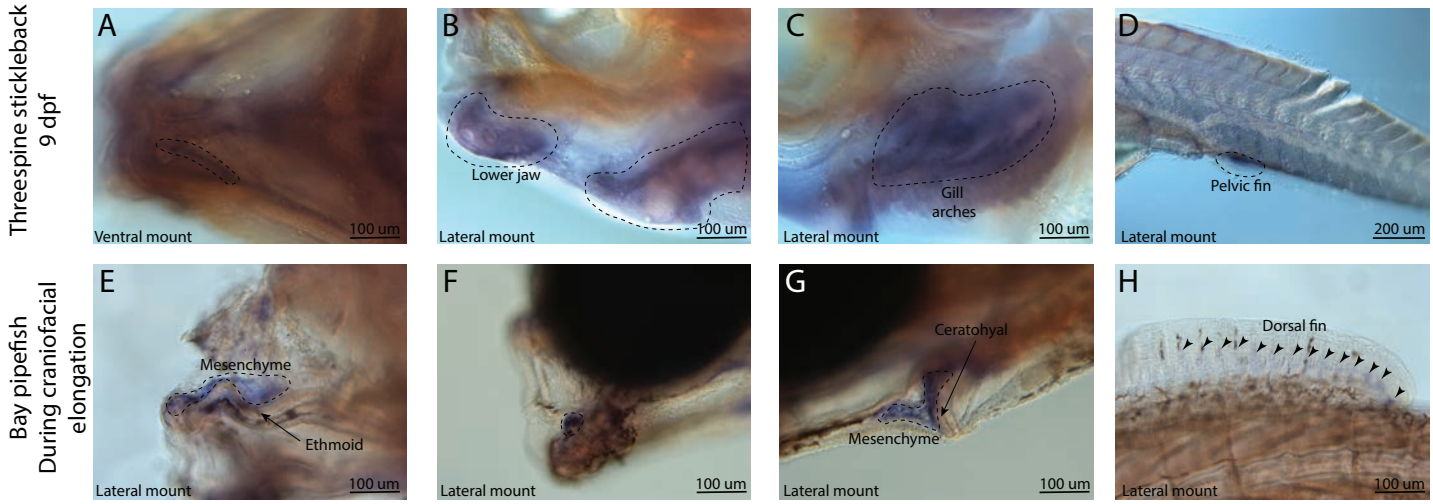


Figure S3.79: *sfrp1a has conserved fin expression in pipefish and stickleback, but craniofacial expression differs between the species. sfrp1a* expression was probed using *in situ* hybridization. Embryos used in this assay were similarly staged (9dpf threespine stickleback and mid-craniofacial elongation staged bay pipefish). Both species had *sfrp1a* staining in the fins (D, H). However, stickleback have broad craniofacial expression (in lower jaws, A-B, and gill arches, C) while expression is limited to mesenchyme above the ethmoid (E), the jaw joint (F), and around the ceratohyal (G) in bay pipefish.

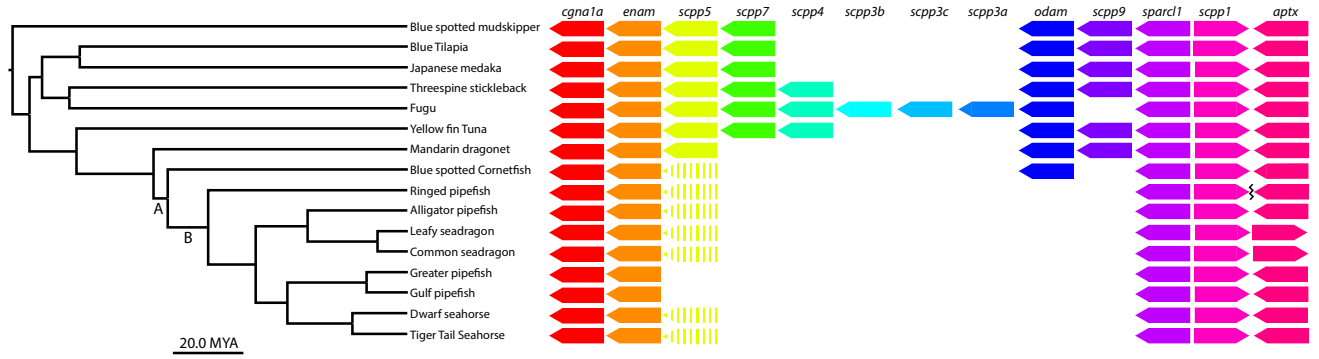


Figure S3.80: *Scpp gene cluster analysis identifies most Scpp losses are not unique to syngnathids*. A) denotes the Syngnathoidei lineage which contains syngnathids and blue spotted cornetfish (among other fish) to highlight that this clade contains various tooth alterations. Syngnathids lack teeth completely and cornetfish have small teeth limited to the lower jaw. B) denotes solely the syngnathid clade, which completely lacks teeth. The phylogenetic tree was time calibrated using Hughes et al. 2018 and Stiller et al. 2022.

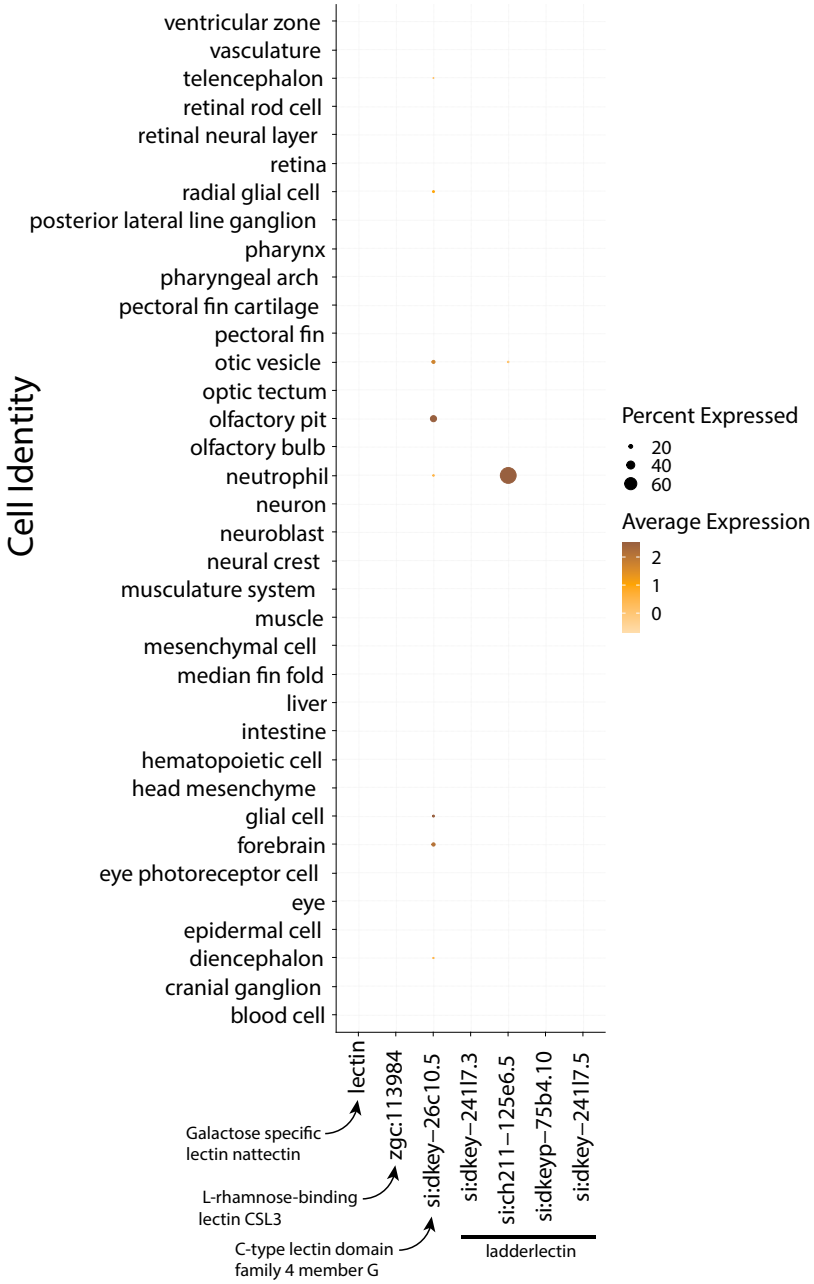


Figure S3.81: *3dpf Zebrafish do not express lectin genes in epidermal cells.* This dotplot shows the marker genes on the x-axis and y-axis, organized by cell type. The size of the dot represents the percentage of cells that express the gene in each cluster. The darkness of the dot represents the average expression of the genes.

APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER IV

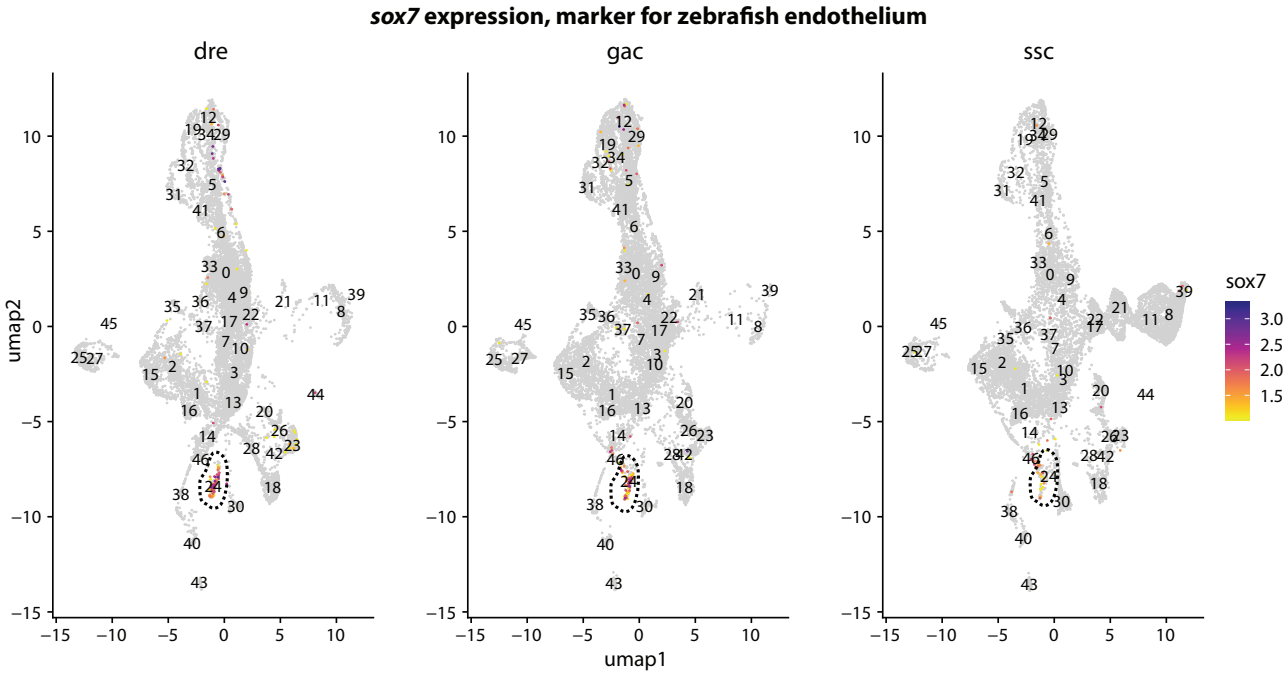


Figure S4.1: *sox7 expression marks zebrafish endothelium in Daniocell as well as zebrafish, stickleback, and pipefish endothelial cells (cluster 24) in our dataset.* Expression of *sox7* is shown in each fish via feature plots. Cluster 24 is outlined with dashed lines to highlight the endothelial cluster.

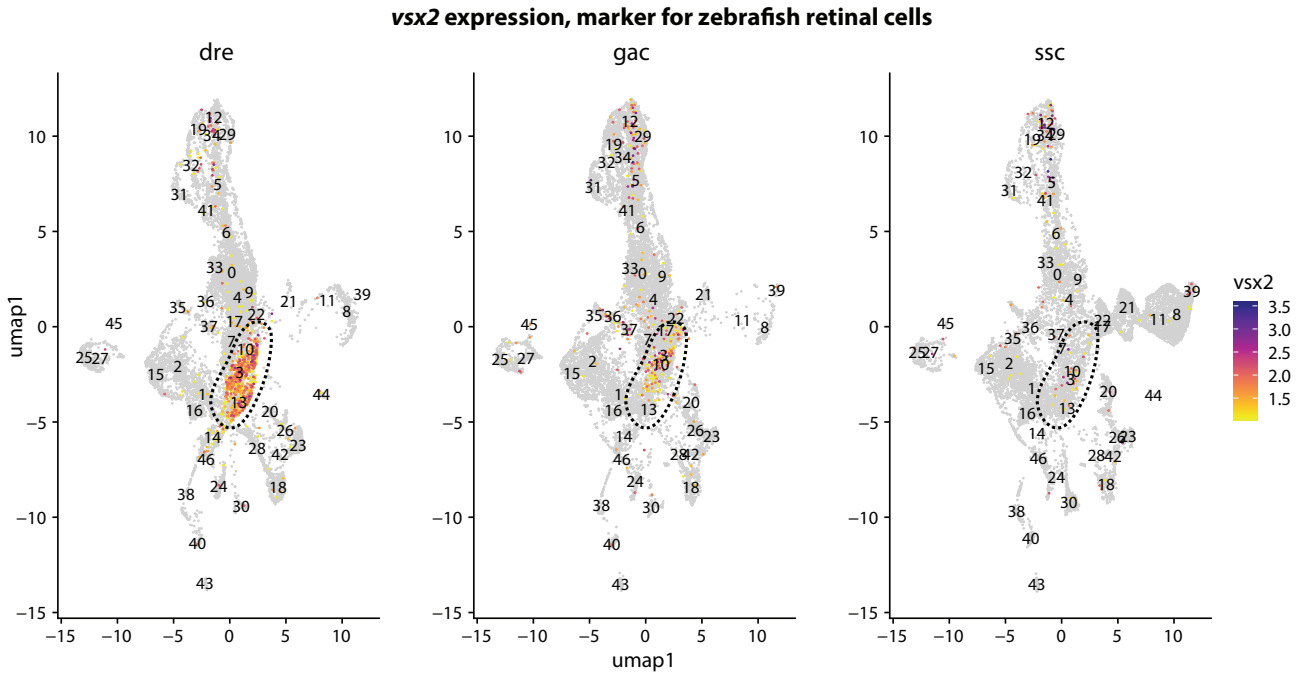


Figure S4.2: *vsx2 expression marks zebrafish retinal cells in Daniocell as well as zebrafish, stickleback, and pipefish retinal cells (clusters 3 and 13) in our dataset.* Expression of *vsx2* is shown in each fish via feature plots. Clusters 3 and 13 are outlined with dashed lines to highlight retinal cell clusters.

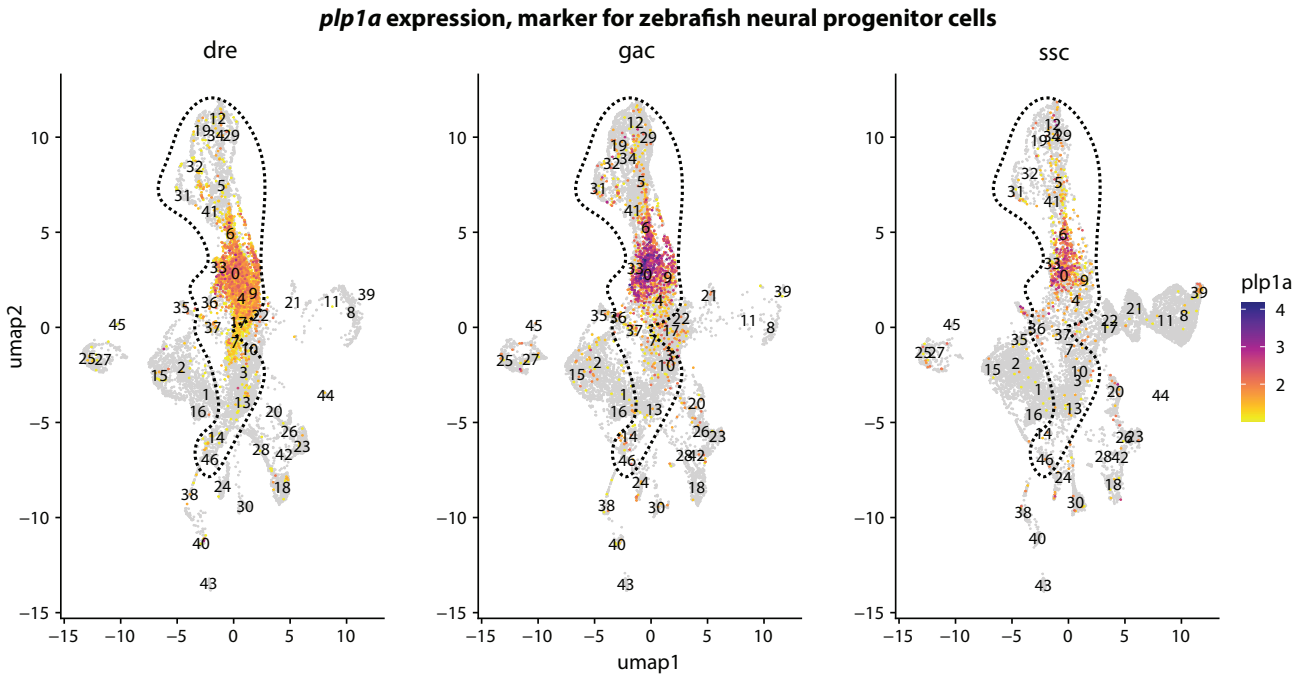


Figure S4.3: *plp1a expression marks zebrafish neural progenitor cells in Daniocell as well as zebrafish, stickleback, and pipefish neural cells in our dataset.* Expression of *plp1a* is shown in each fish via feature plots. Neural cell clusters are outlined with dashed lines.

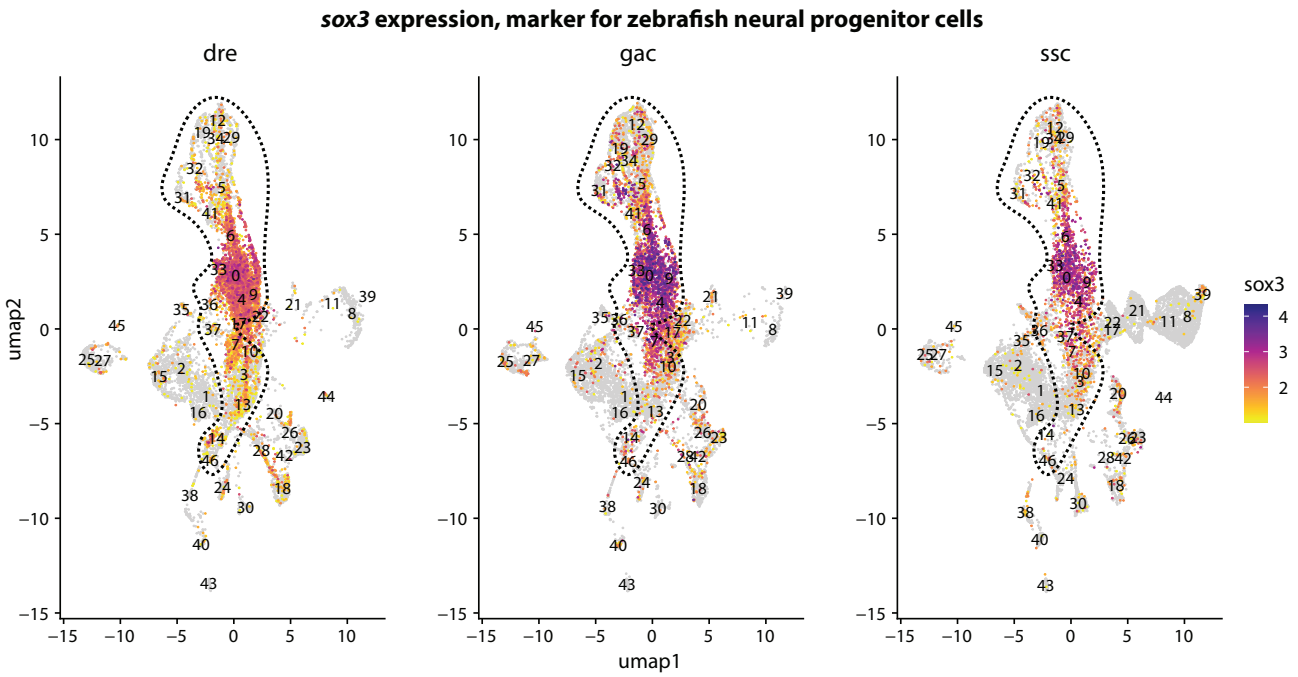


Figure S4.4: *sox3 expression marks zebrafish neural progenitor cells in Daniocell as well as zebrafish, stickleback, and pipefish neural cells in our dataset.* Expression of *sox3* is shown in each fish via feature plots. Neural clusters are outlined with dashed lines.

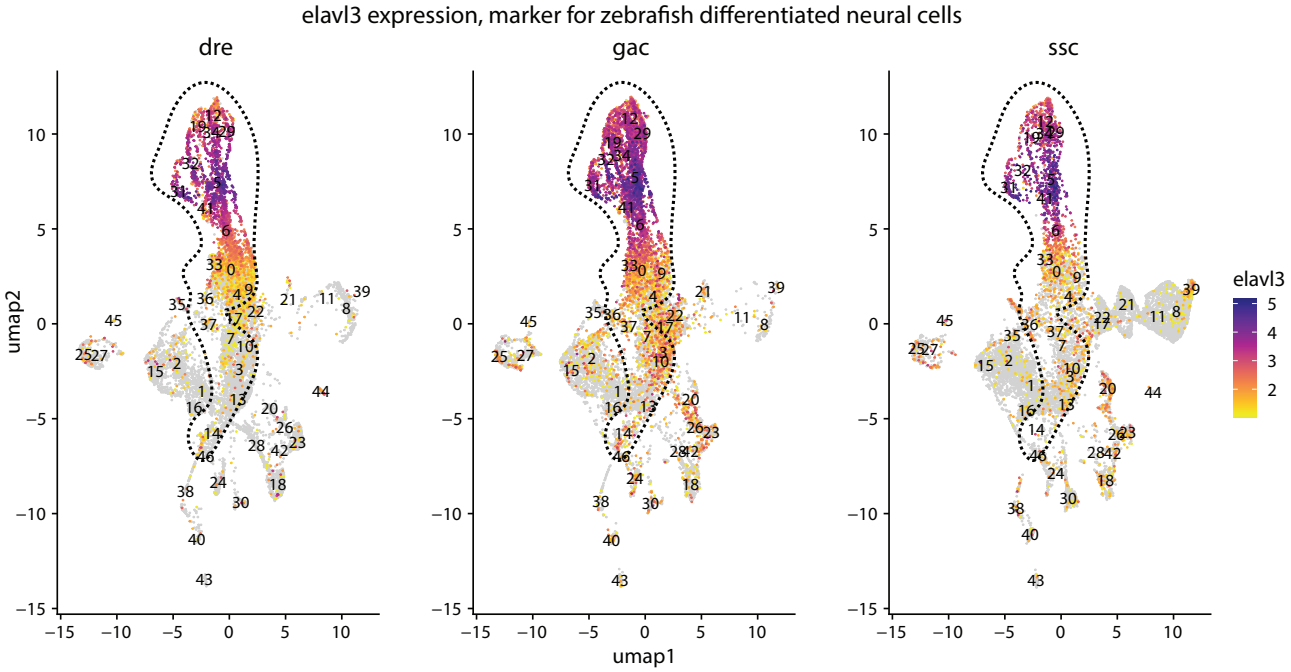


Figure S4.5: *elavl3 expression marks zebrafish differentiated neural cells in Daniocell as well as zebrafish, stickleback, and pipefish neural cells in our dataset.* Expression of *elavl3* is shown in each fish via feature plots. Neural clusters are outlined with dashed lines.

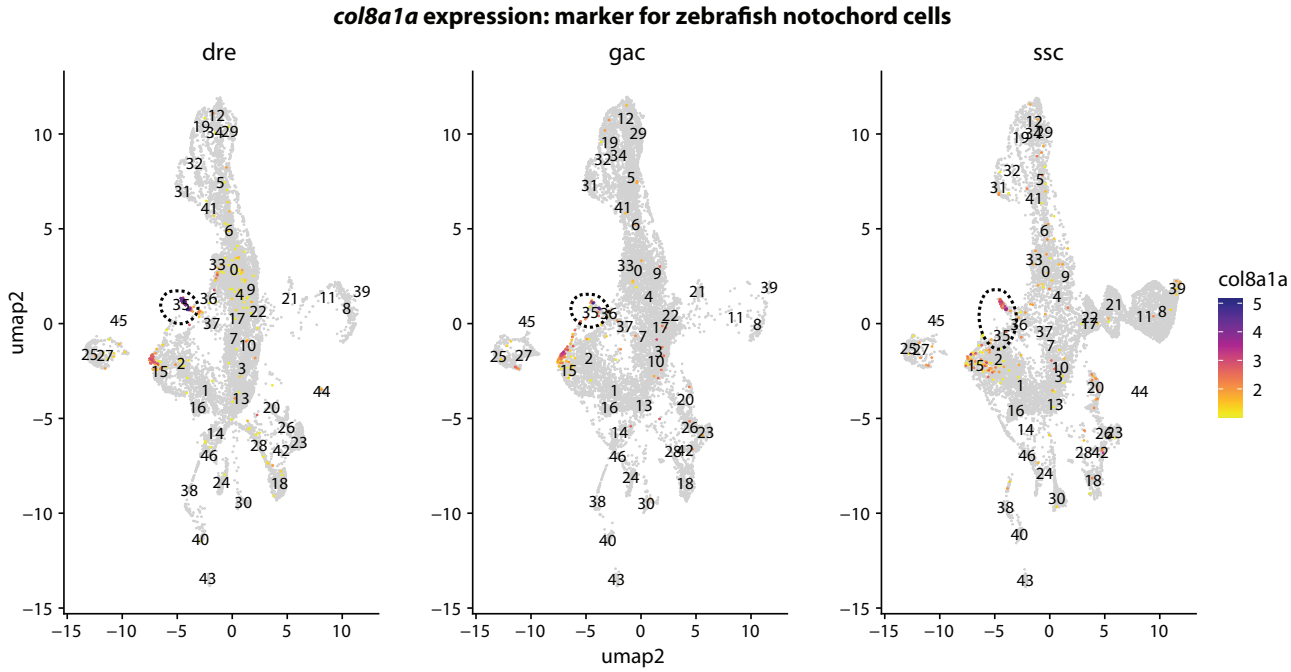


Figure S4.6: *col8a1a expression marks zebrafish notochord in Daniocell as well as zebrafish, stickleback, and pipefish notochord cells (cluster 35) in our dataset.* Expression of *col8a1a* is shown in each fish via feature plots. Cluster 35 is outlined with dashed lines to highlight the notochord cluster.

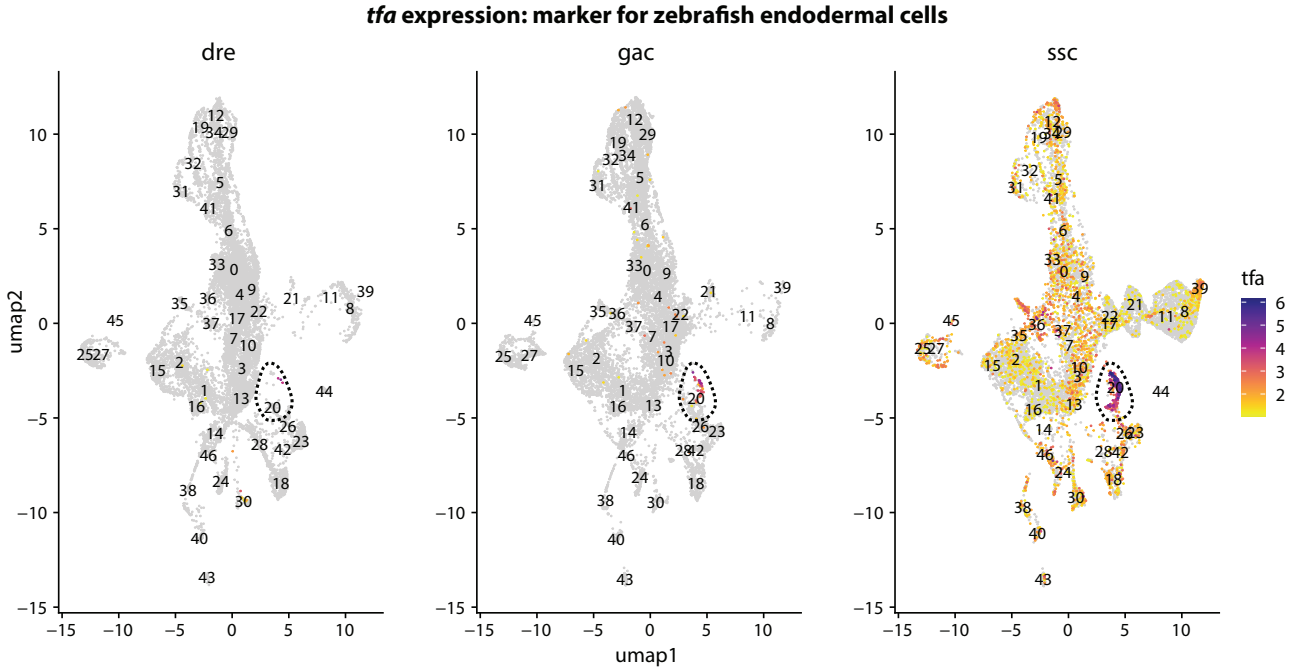


Figure S4.7: *tfa expression marks zebrafish endodermal in Daniocell as well as zebrafish, stickleback, and pipefish endodermal cells (cluster 20) in our dataset.* Expression of *tfa* is shown in each fish via feature plots. Cluster 20 is outlined with dashed lines to highlight the endodermal cluster.

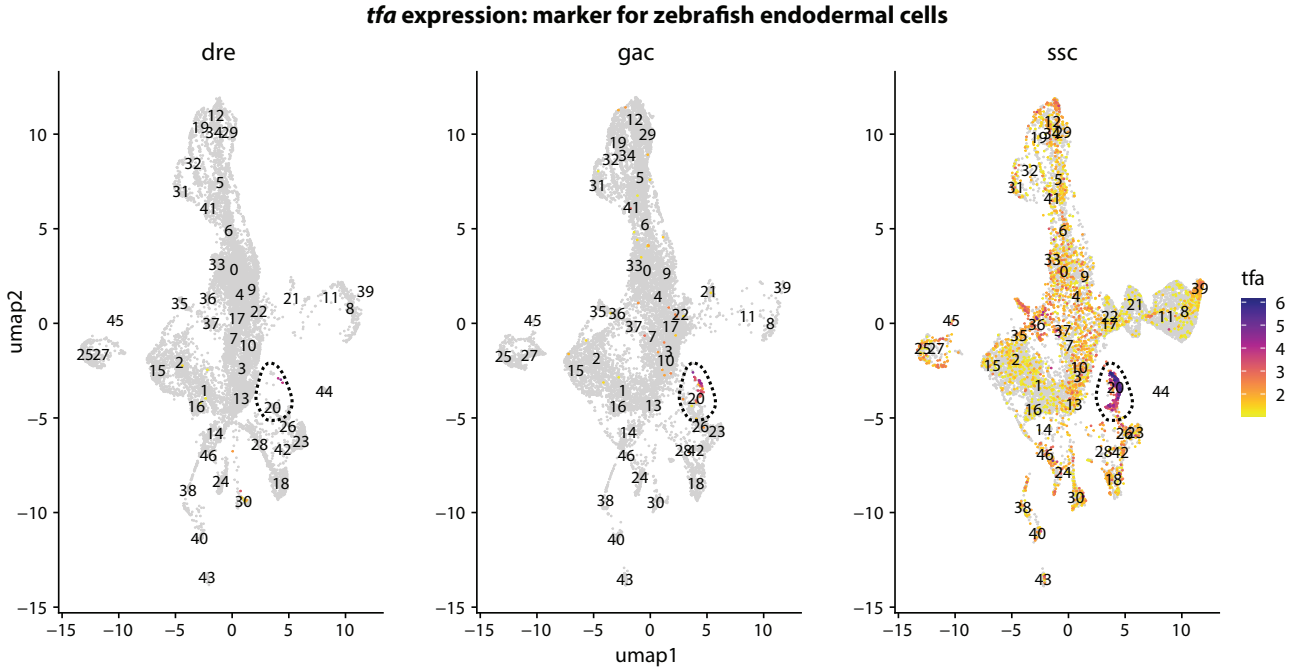


Figure S4.8: *tfa expression marks zebrafish endodermal in Daniocell as well as zebrafish, stickleback, and pipefish endodermal cells (cluster 20) in our dataset.* Expression of *tfa* is shown in each fish via feature plots. Cluster 20 is outlined with dashed lines to highlight the endodermal cluster.

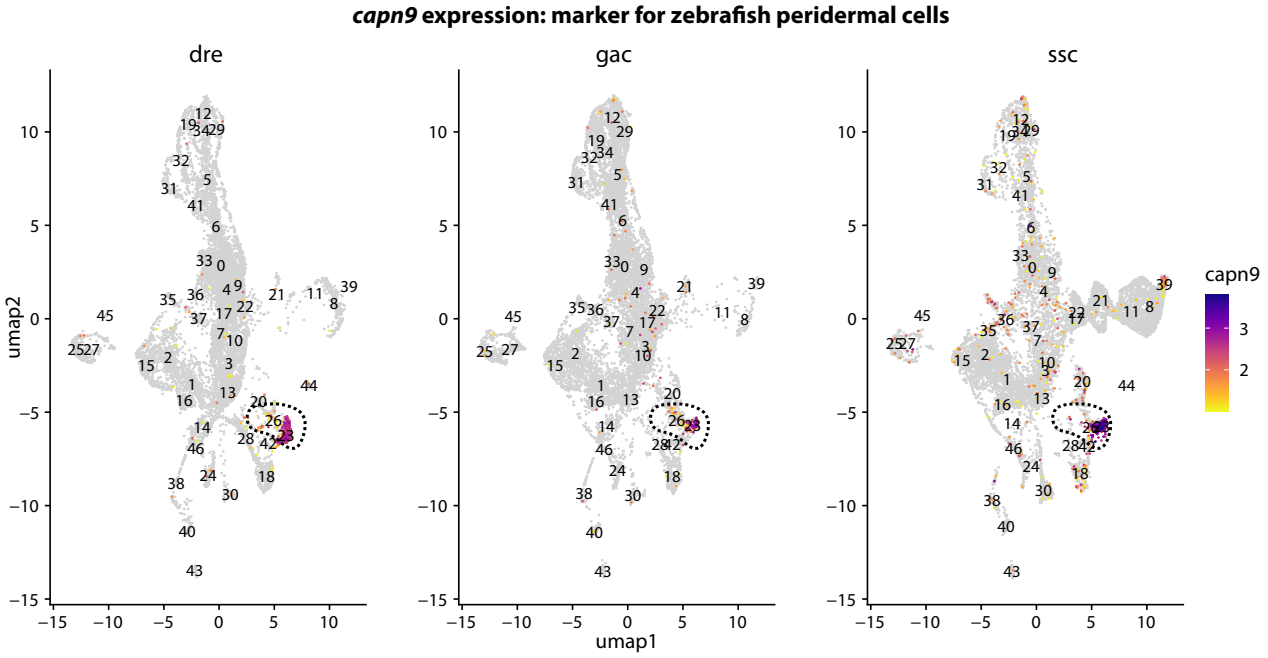


Figure S4.9: *capn9 expression marks zebrafish periderm in Daniocell as well as zebrafish, stickleback, and pipefish peridermal cells (clusters 26 and 28) in our dataset.* Expression of *capn9* is shown in each fish via feature plots. Clusters 26 and 28 are outlined with dashed lines to highlight the peridermal clusters.

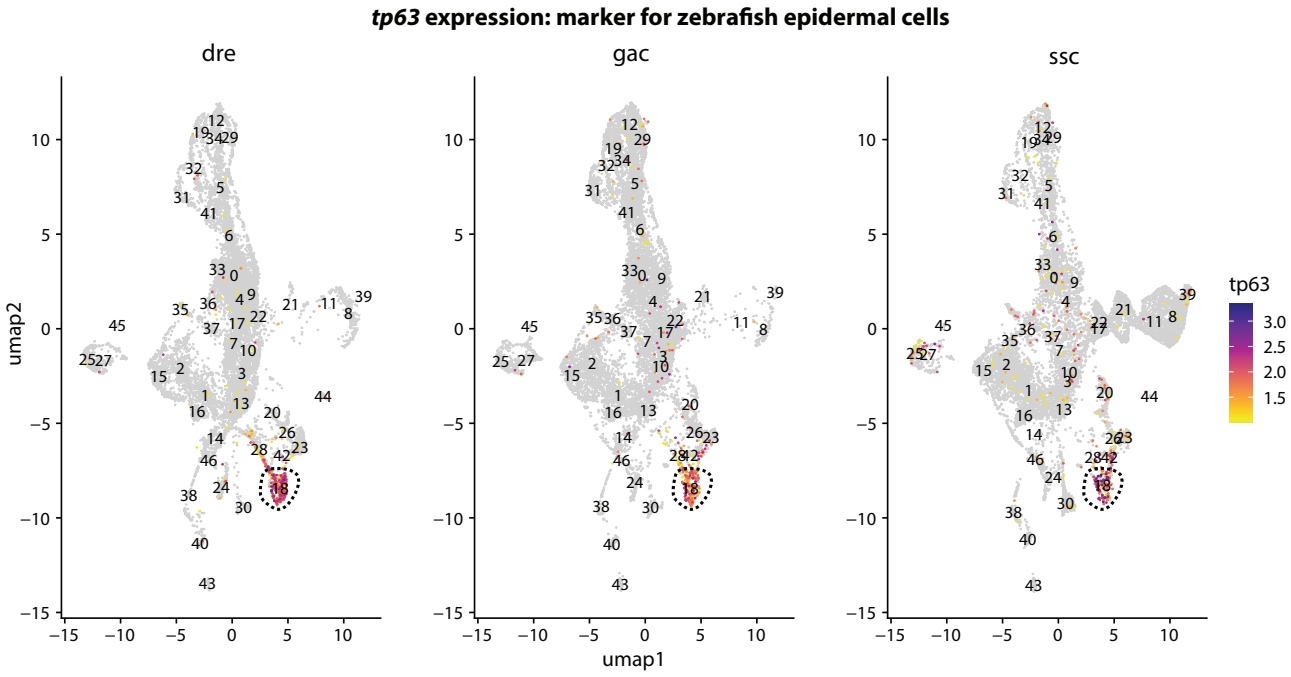


Figure S4.10: *tp63 expression marks zebrafish epidermal cells in Daniocell as well as zebrafish, stickleback, and pipefish endothelial cells (cluster 18) in our dataset.* Expression of *tp63* is shown in each fish via feature plots. Cluster 18 is outlined with dashed lines to highlight the epidermal cluster.

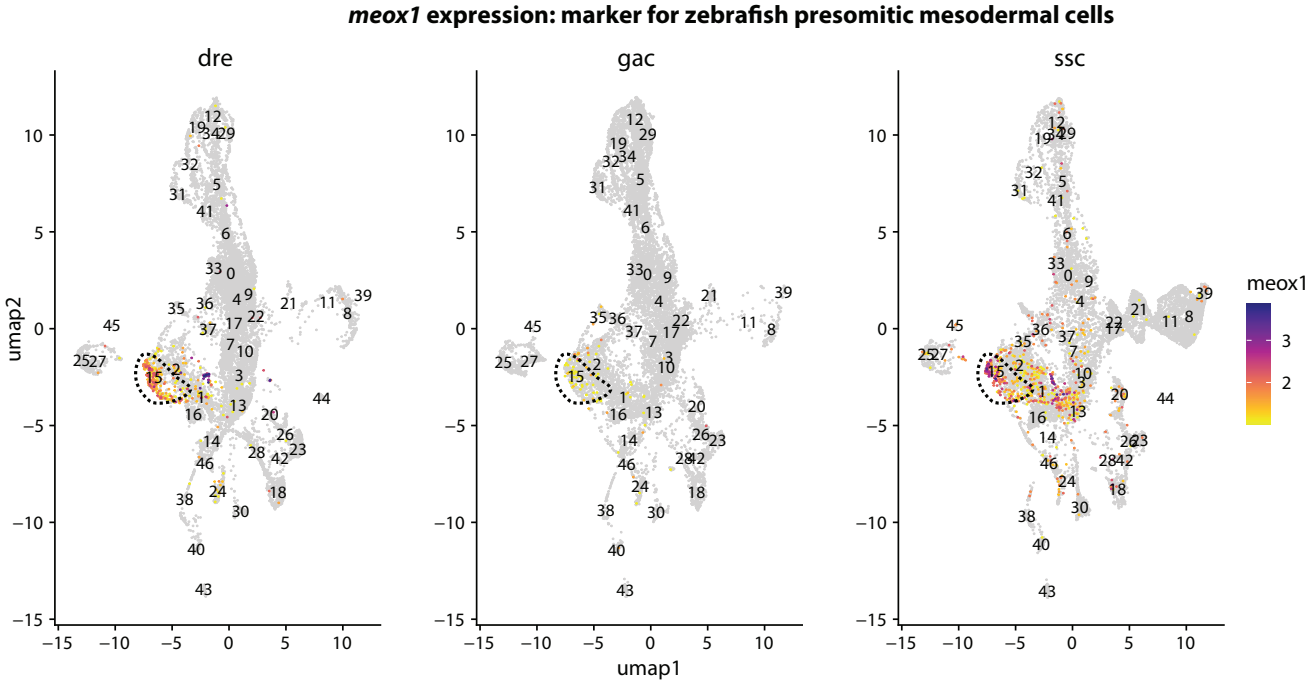


Figure S4.11: *meox1 expression marks zebrafish presomitic mesoderm in Daniocell as well as zebrafish, stickleback, and pipefish these cells (cluster 15) in our dataset.* Expression of *meox1* is shown in each fish via feature plots. Cluster 15 is outlined with dashed lines to highlight the presomitic mesoderm cluster.

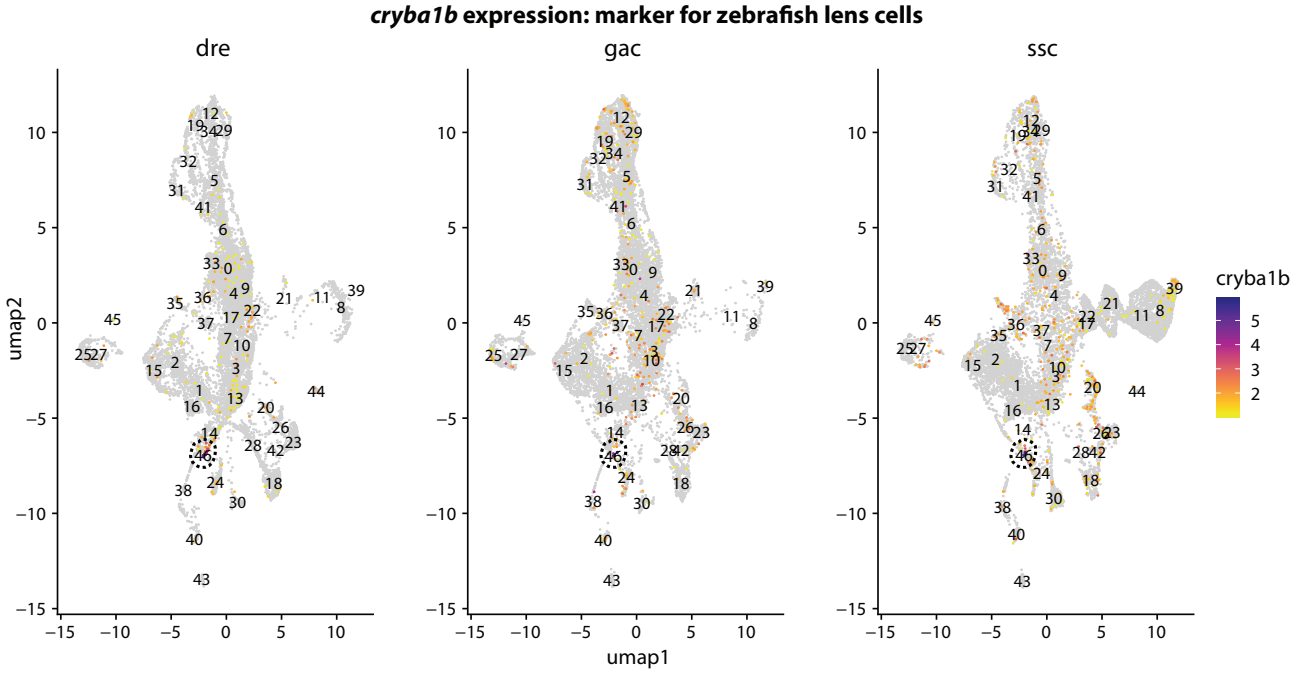


Figure S4.12: *cryba1b expression marks zebrafish lens cells in Daniocell as well as zebrafish, stickleback, and pipefish lens cells (cluster 46) in our dataset.* Expression of *cryba1b* is shown in each fish via feature plots. Cluster 46 is outlined with dashed lines to highlight the lens cluster.

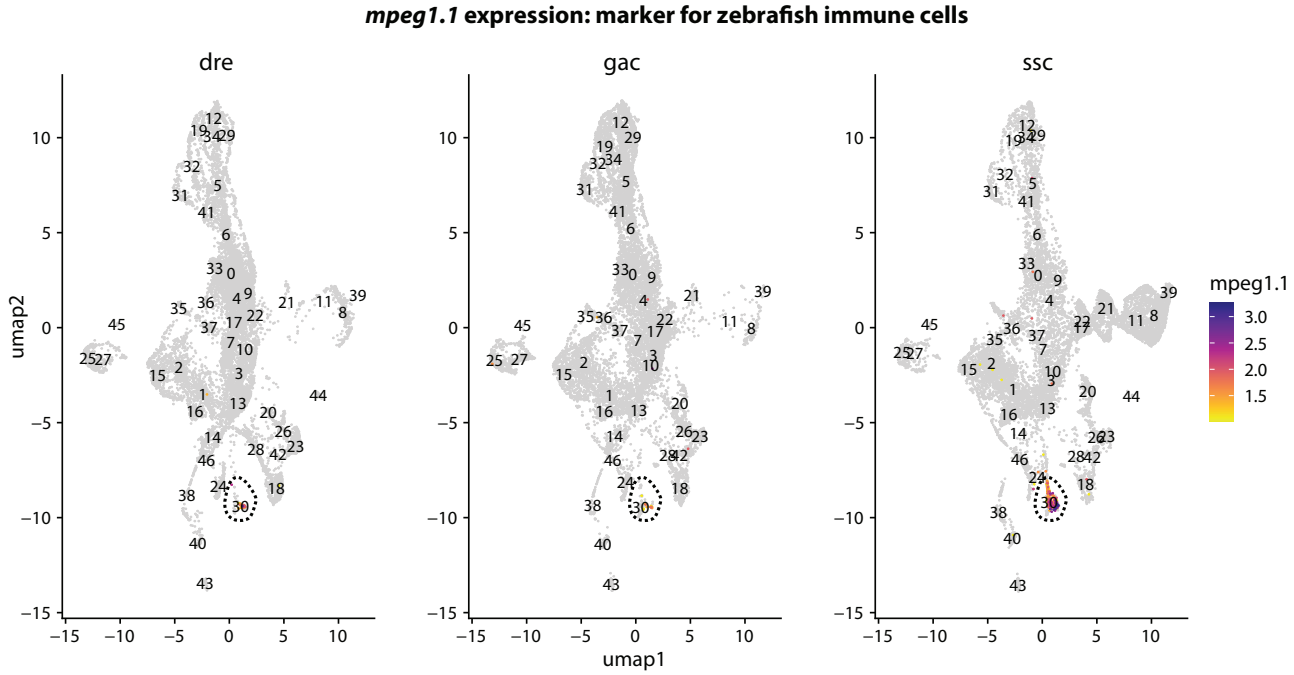


Figure S4.13: *mpeg1.1 expression marks zebrafish immune cells in Daniocell as well as these cells in zebrafish, stickleback, and pipefish (cluster 30) in our dataset.* Expression of *mpeg1.1* is shown in each fish via feature plots. Cluster 30 is outlined with dashed lines to highlight the immune cluster.

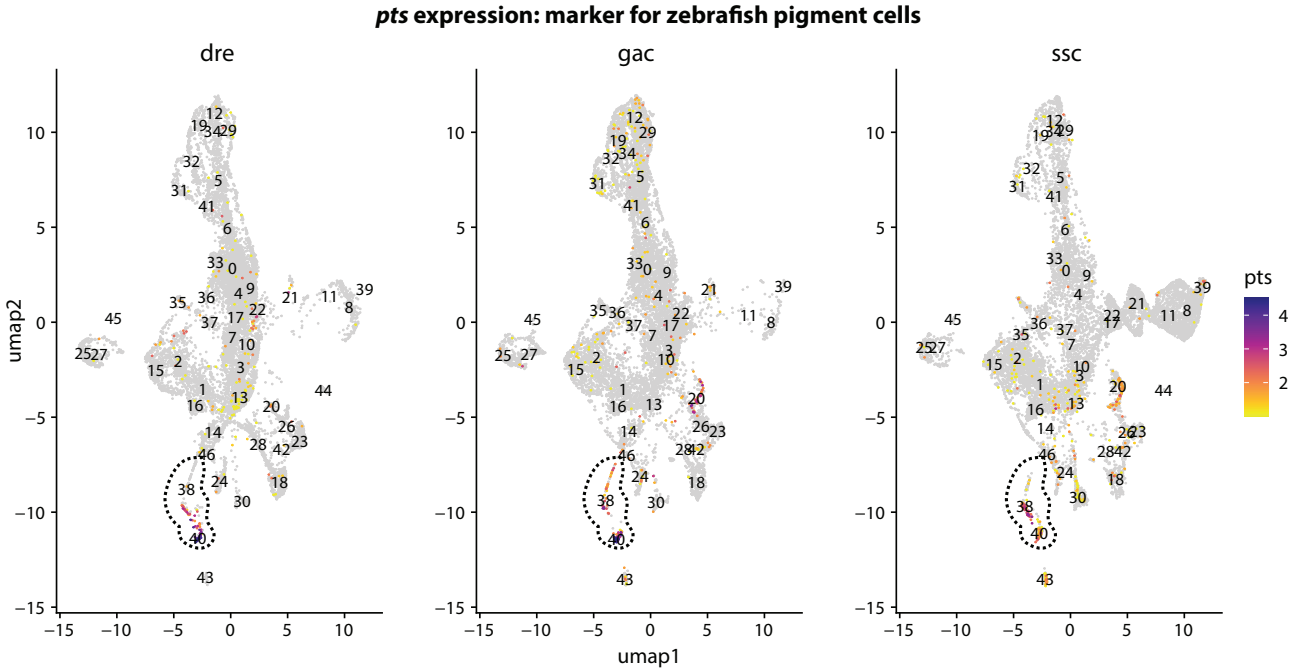


Figure S4.14: *pts expression marks zebrafish pigment in Daniocell as well as these cells in zebrafish, stickleback, and pipefish (clusters 38 and 40) in our dataset.* Expression of *pts* is shown in each fish via feature plots. Clusters 38 and 40 are outlined with dashed lines to highlight the pigment clusters.

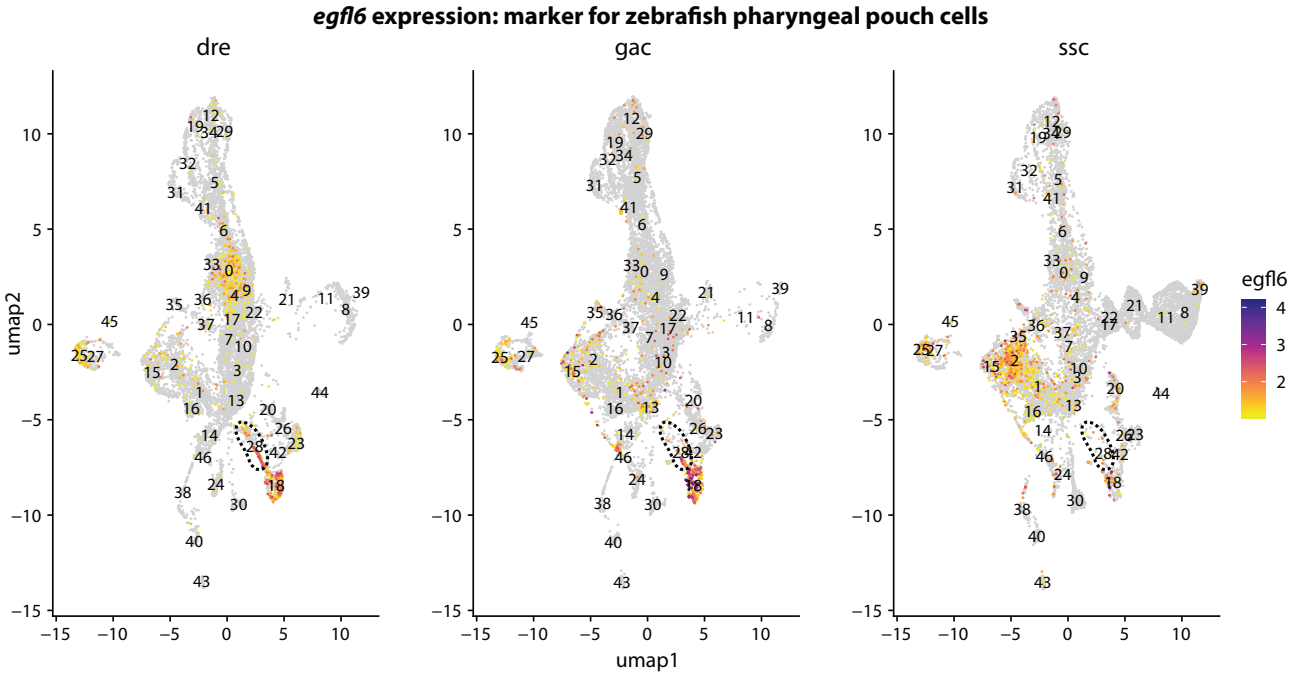


Figure S4.15: *egfl6 expression marks zebrafish pharyngeal pouch cells in Daniocell as well as these cells in zebrafish, stickleback, and pipefish (cluster 28) in our dataset.* Expression of *egfl6* is shown in each fish via feature plots. Cluster 28 is outlined with dashed lines to highlight the pharyngeal pouch cluster.

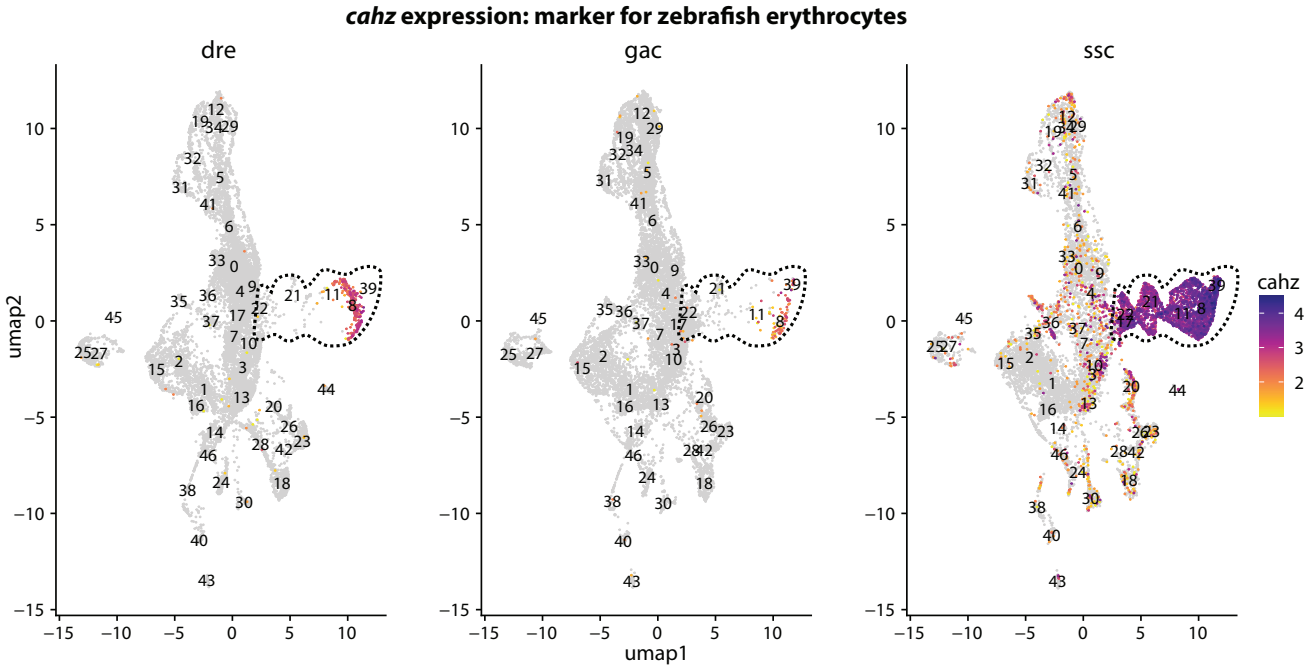


Figure S4.16: *cahz expression marks zebrafish red blood cells in Daniocell as well as zebrafish, stickleback, and pipefish erythrocytes in our dataset.* Expression of *cahz* is shown in each fish via feature plots. Erythrocytes are outlined with dashed lines.

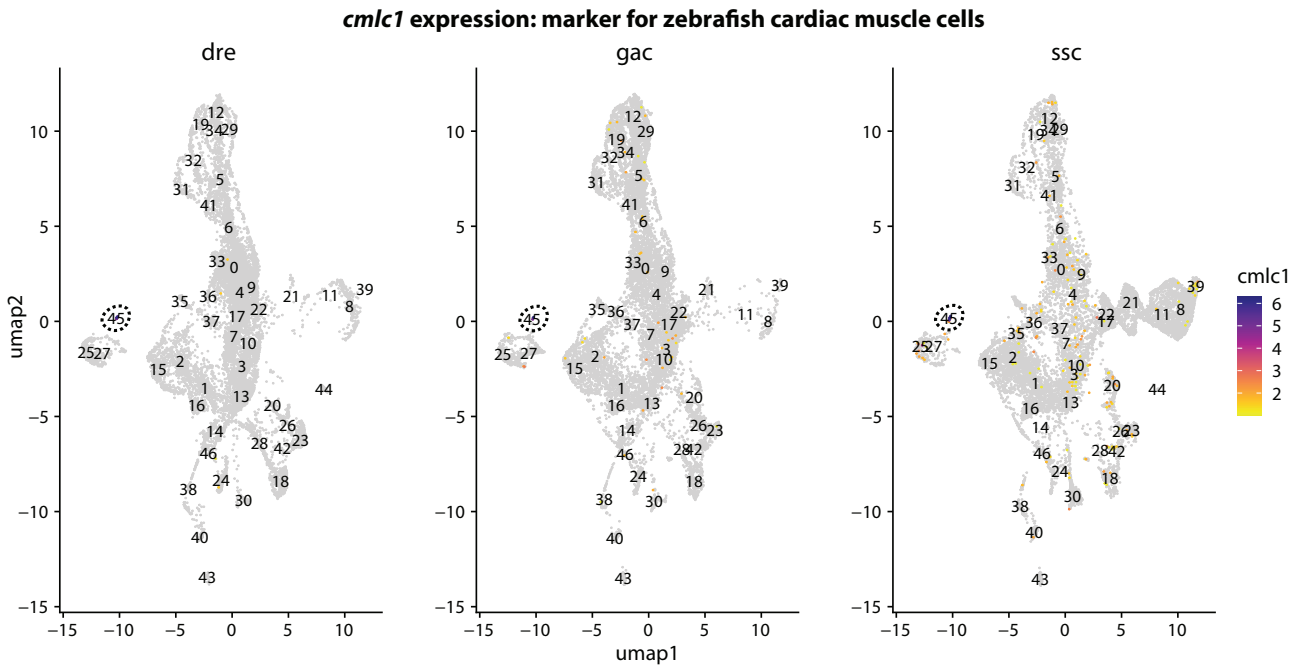


Figure S4.17: *cmlc1 expression marks zebrafish cardiac muscle cells in Daniocell as well as these cells in zebrafish, stickleback, and pipefish (cluster 45) in our dataset.* Expression of *cmlc1* is shown in each fish via feature plots. Cluster 45 is outlined with dashed lines to highlight the cardiac muscle cluster.

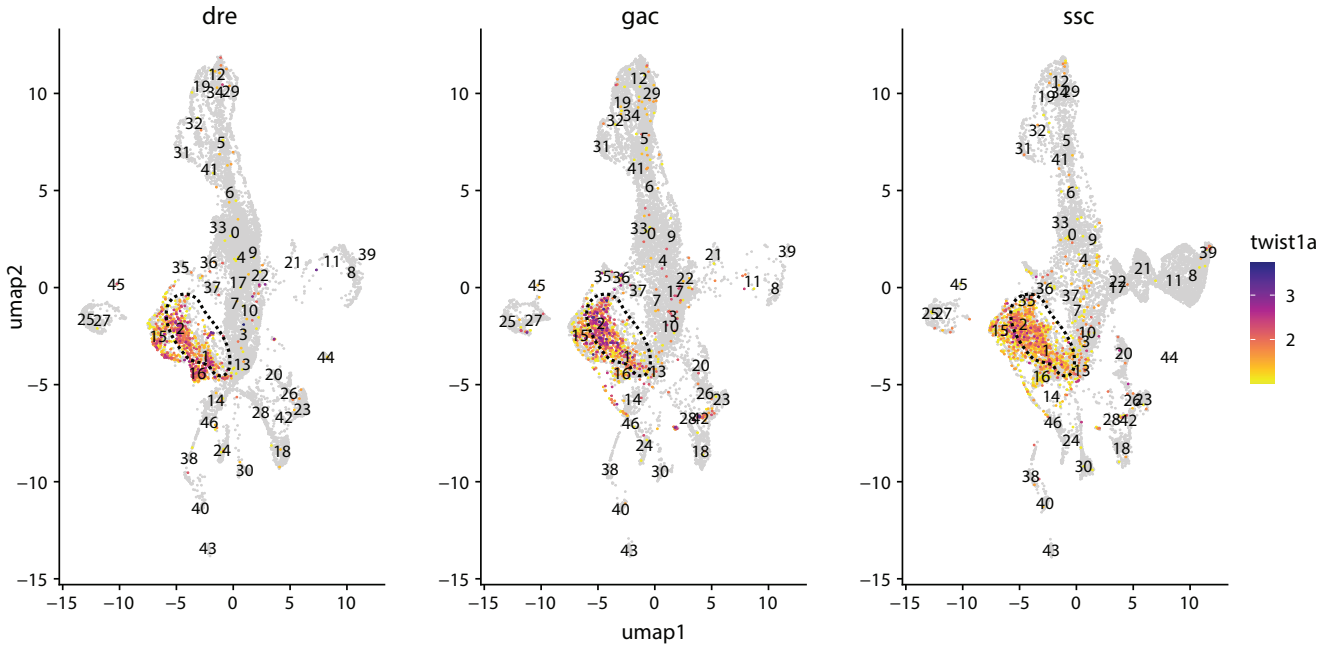


Figure S4.18: *twist1a expression marks zebrafish mesenchyme in Daniocell as well as zebrafish, stickleback, and pipefish mesenchymal cells (clusters 1 and 2) in our dataset.* Expression of *twist1a* is shown in each fish via feature plots. Clusters 1 and 2 are outlined with dashed lines to highlight the mesenchymal cluster.

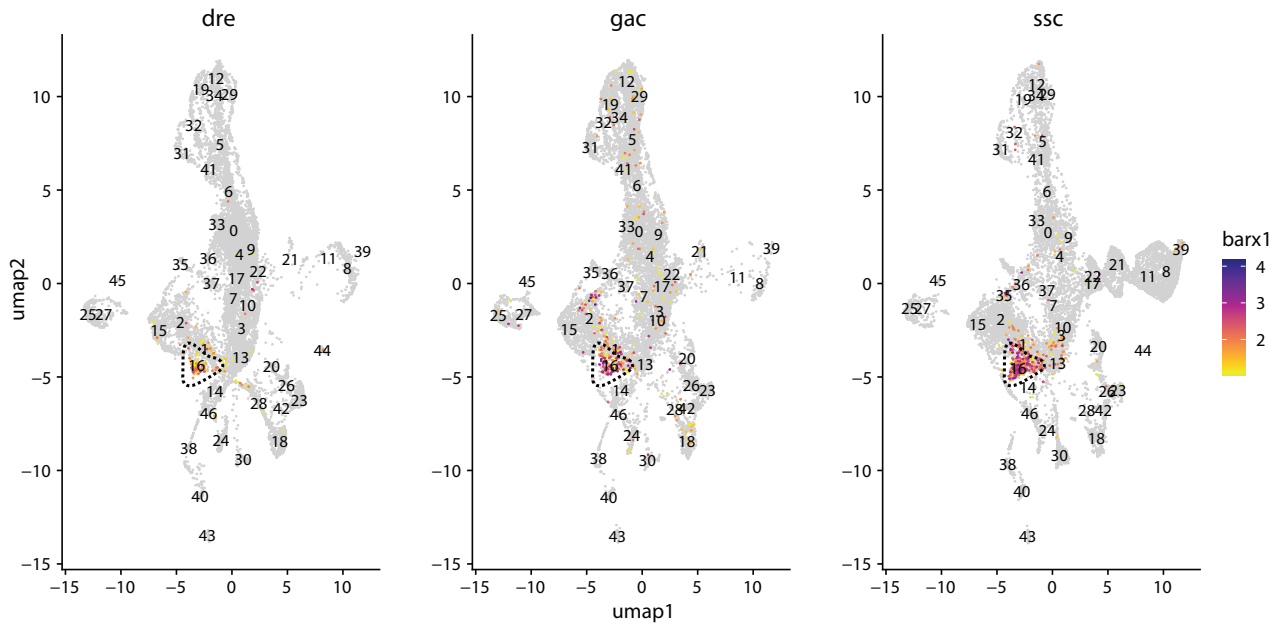


Figure S4.19: *barx1 expression marks zebrafish pharyngeal arch cells in Daniocell as well as these cells in zebrafish, stickleback, and pipefish (cluster 16) in our dataset.* Expression of *barx1* is shown in each fish via feature plots. Cluster 16 is outlined with dashed lines to highlight the pharyngneal arch cluster.

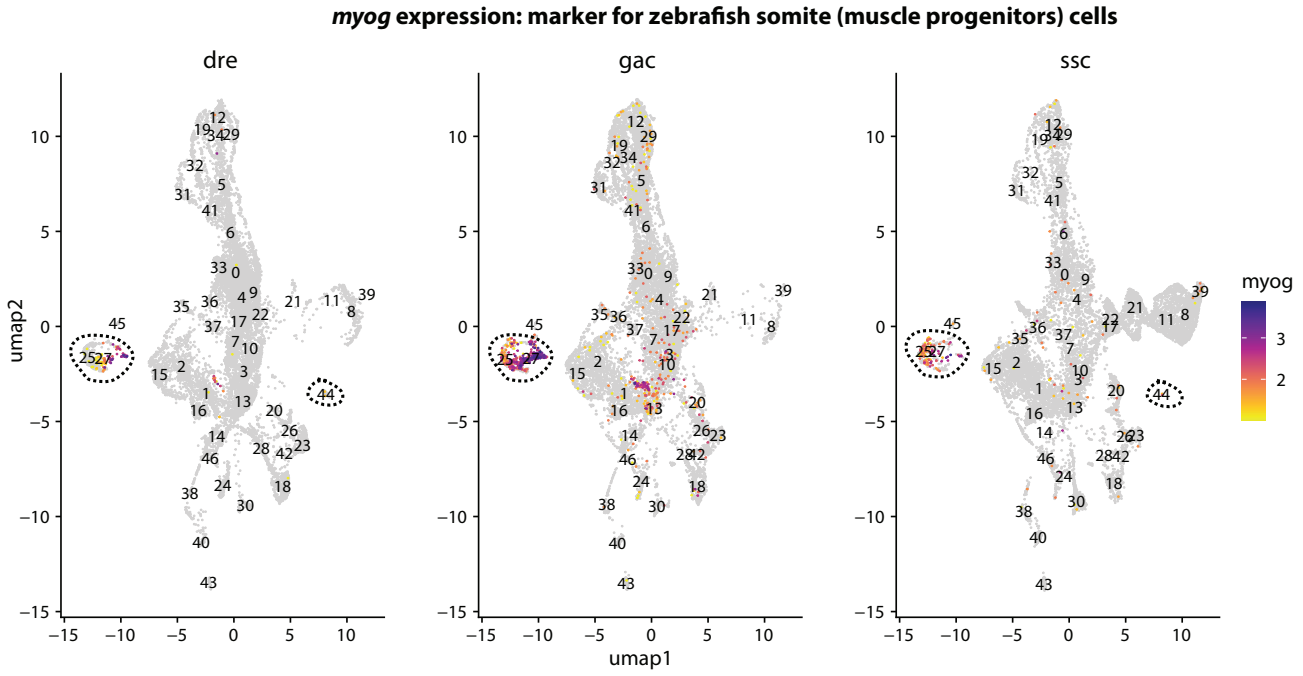


Figure S4.20: *myog expression marks zebrafish somites in Daniocell as well as zebrafish, stickleback, and pipefish somite, muscle progintor, cells (clusters 25 and 27) in our dataset.* Expression of *myog* is shown in each fish via feature plots. Clusters 25 and 27 are outlined with dashed lines to highlight the somite clusters.

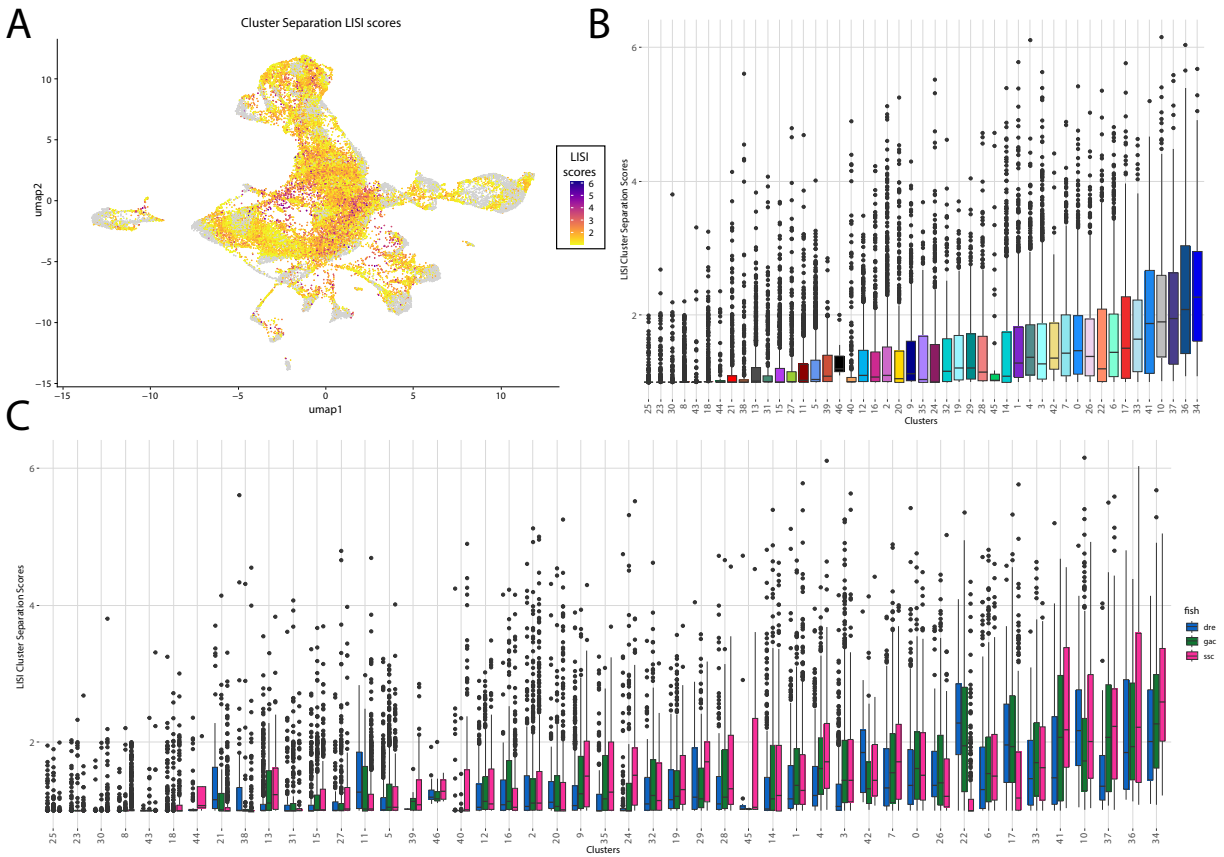


Figure S4.21*: High level of cluster uniqueness in the comparative scRNAseq atlas.* To evaluate the overlap of cells between clusters, LISI scores were quantified. Scores of 1 indicate cells that are close in UMAP space with only cells of the same cluster. LISI values above 1 represent cells that are nearby cells from other clusters in UMAP. Scores are relatively similar across the UMAP space (A), with the highest degree of cluster mixing in neural clusters. Cells are plotted in a FeaturePlot in panel A and are colored by the LISI cluster mixing score. Pharyngeal arch cells (#16) form a distinct cell cluster (B) with a low amount of mixing with other clusters. Cluster distinctness minimally varies by species (C). Cluster data is plotted individually in panels B and C, and species data is separated in panel C.

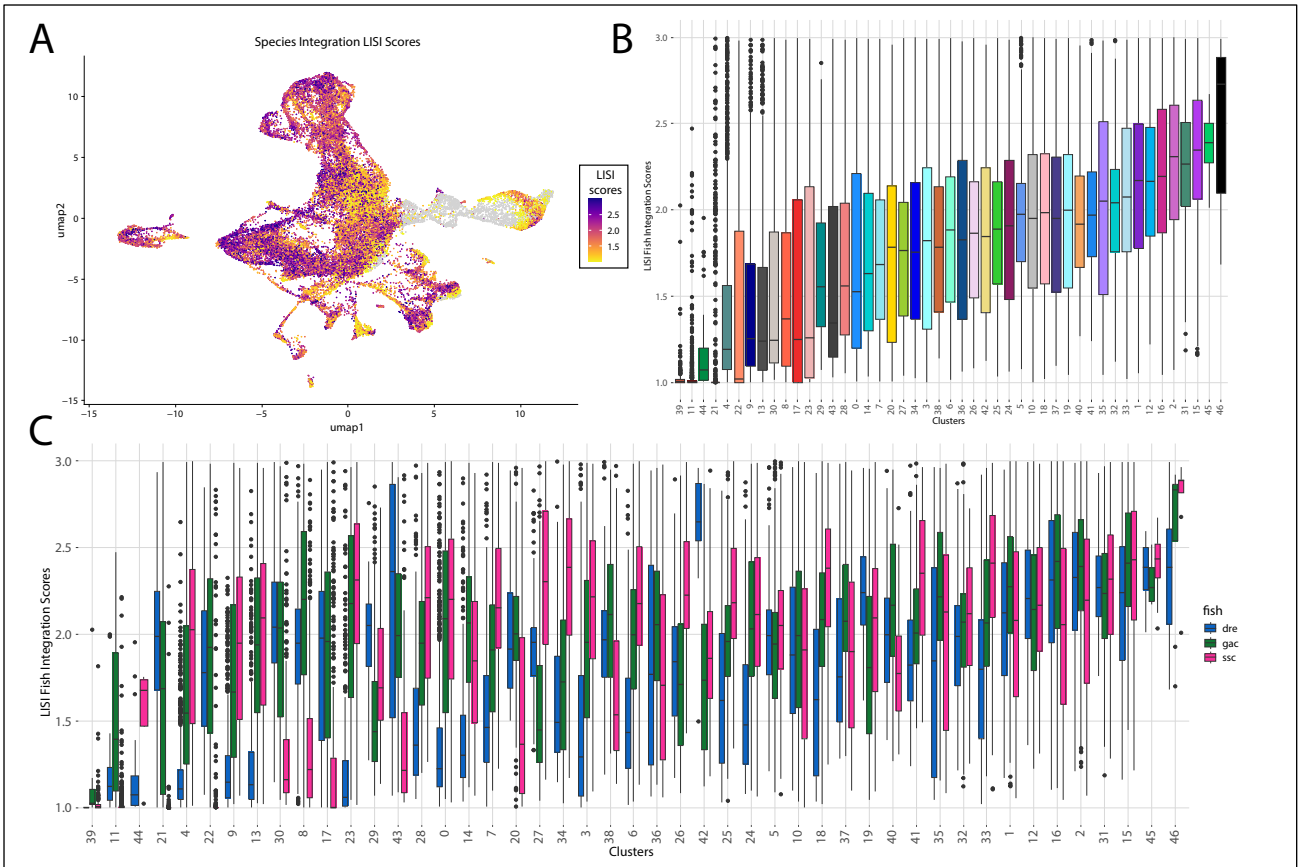


Figure S4.22*: High level of species integration except in erythrocyte clusters marks successful comparative scRNAseq atlas.* To evaluate the extent of cellular integration between species, LISI scores were quantified. Scores of 1 indicate cells are close in UMAP space with only cells of the same species. LISI values of 3 represent cells that are nearby cells of all three species in UMAP dimensionality. Scores are relatively similar across the UMAP (A) except for cells in the erythrocyte clusters. Cells are plotted in a FeaturePlot in panel A and are colored by the LISI fish integration score. Pharyngeal arch cells (#16) have a high level of integration (B) compared to other cell types. Although integration scores vary by species, pharyngeal arch scores are relatively similar (C) between the three species. Cluster data is plotted individually in panels B and C, and species data is separated in panel C.

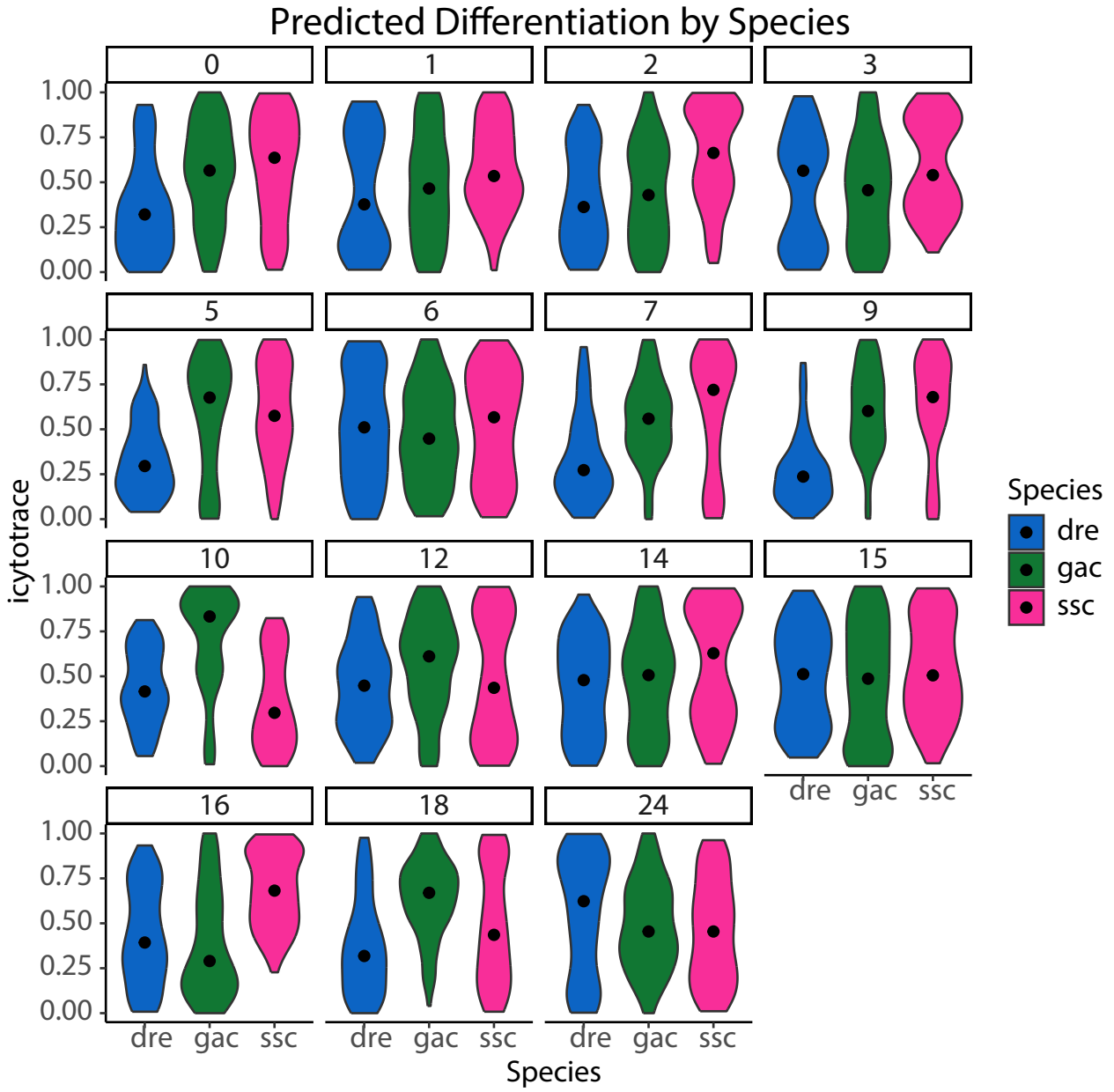


Figure S4.23: *Species differentiation state varied by cluster though no clear evidence for developmental stage mismatch.* Cellular differentiation was tested using integrative CytoTRACE. Species differentiation states for each cell in a cluster are plotted in violin plots. The CytoTRACE scale ranges from 0 to 1 whereby 1 represents undifferentiated cells.

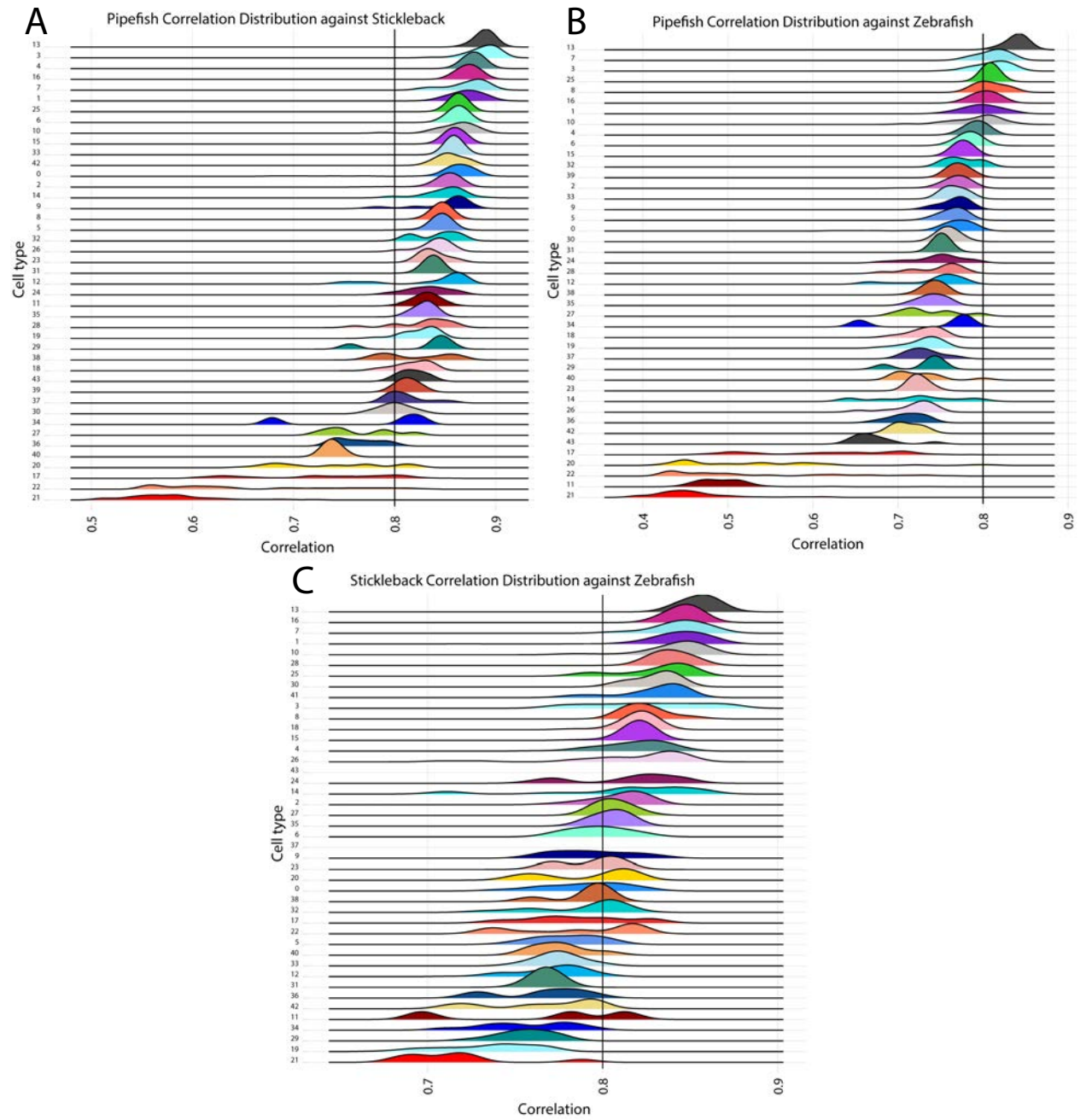


Figure S4.24: *Pharyngeal arch cells express highly conserved core genetic program.* In panels A-C, pairwise correlation analyses demonstrate that pharyngeal arch cells have maintained core gene expression profile in all species.

Table S4.25: *Correlation medians for species comparisons illustrate high conservation of PA cells*. Median correlation scores from the correlation analysis are listed by species comparison. These assessments are directional and completed for every cluster. The highest score for each species contrast is bolded.

Cluster	Zebrafish vs Stickleback	Zebrafish vs Pipefish	Stickleback vs Pipefish	Stickleback vs Zebrafish	Pipefish vs Stickleback	Pipefish vs Zebrafish
0	0.81	0.78	0.86	0.79	0.86	0.77
1	0.85	0.82	0.88	0.85	0.87	0.80
2	0.81	0.78	0.86	0.81	0.85	0.77
3	**0.86**	0.81	0.88	0.83	**0.89**	0.81
4	0.83	0.79	0.88	0.82	0.88	0.79
5	0.78	0.76	0.84	0.78	0.85	0.77
6	0.81	0.78	0.86	0.80	0.86	0.79
7	0.85	0.81	0.88	0.84	0.88	0.81
8	0.84	0.82	0.85	0.82	0.85	0.81
9	0.82	0.77	0.86	0.79	0.86	0.77
10	**0.86**	0.82	0.86	0.85	0.87	0.80
11	0.81	0.73	0.86	0.78	0.83	0.49
12	0.78	0.75	0.85	0.78	0.86	0.75
13	**0.86**	**0.84**	**0.89**	**0.86**	**0.89**	**0.84**
14	0.84	0.79	0.86	0.82	0.85	0.72
15	0.82	0.77	0.86	0.82	0.86	0.78
16	0.85	0.82	0.88	0.85	0.87	0.80
17	0.83	0.77	0.80	0.78	0.74	0.63
18	0.82	0.74	0.83	0.82	0.82	0.74
19	0.77	0.74	0.83	0.74	0.83	0.74
20	0.78	0.60	0.77	0.81	0.74	0.52
21	0.75	0.59	0.63	0.71	0.58	0.45
22	0.81	0.76	0.82	0.79	0.62	0.49
23	0.76	0.69	0.84	0.80	0.84	0.72
24	0.79	0.75	0.83	0.82	0.84	0.75
25	0.83	0.80	0.85	0.84	0.86	0.81
26	0.81	0.71	0.83	0.83	0.84	0.73
27	0.81	0.78	0.71	0.80	0.75	0.72
28	0.84	0.76	0.84	0.84	0.84	0.75
29	0.76	0.73	0.82	0.76	0.84	0.74
30	0.84	0.78	0.83	0.84	0.80	0.76
31	0.77	0.75	0.84	0.77	0.84	0.75
32	0.80	0.76	0.83	0.80	0.85	0.77
33	0.79	0.77	0.85	0.77	0.86	0.76
34	0.78	0.77	0.78	0.75	0.81	0.77
35	0.80	0.73	0.84	0.81	0.83	0.74
36	0.78	0.68	0.75	0.77	0.76	0.71
37	0.78	0.75	0.85	0.79	0.81	0.73
38	0.78	0.74	0.87	0.80	0.82	0.74
40	0.78	0.74	0.74	0.78	0.74	0.72

Table S4.26: *Correlation medians for binned species comparisons illustrate high conservation of PA cells*. Median correlation scores from the correlation analysis are listed by species comparison. These assessments are directional and completed for binned clusters. The highest score for each species contrast is bolded.

	Neural	Conn	RBC	Unk	CNC PA	Epit	Endo	Endot	Musc	Immu	Pigm
Zebrafish vs Stickleback	0.83	0.83	0.84	**0.86**	0.85	0.82	0.78	0.79	0.83	0.84	0.78
Zebrafish vs Pipefish	0.79	0.80	**0.82**	**0.82**	**0.82**	0.74	0.65	0.75	0.79	0.78	0.74
Stickleback vs Pipefish	0.86	0.87	0.81	0.86	**0.88**	0.84	0.81	0.83	0.75	0.83	0.76
Stickleback vs Zebrafish	0.79	0.82	0.80	**0.85**	**0.85**	0.82	0.76	0.82	0.81	0.84	0.80
Pipefish vs Zebrafish	0.78	0.78	0.80	0.80	0.80	0.73	0.54	0.75	**0.81**	0.76	0.74
Pipefish vs Stickleback	0.86	0.86	0.84	**0.87**	**0.87**	0.83	0.77	0.84	0.86	0.80	0.79

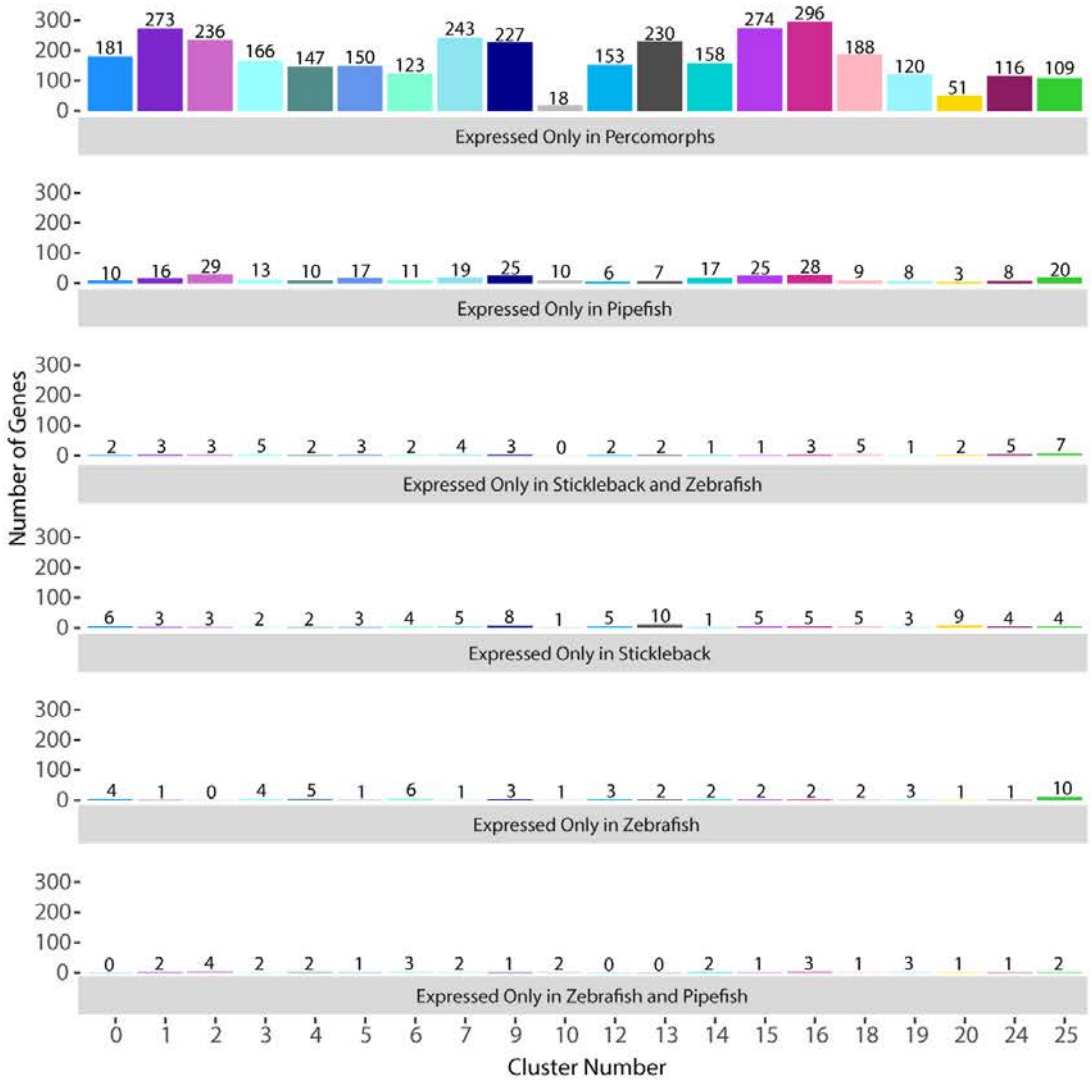


Figure S4.27: *Unique expression patterns of genes by species and groupings shows high number of Percomorph specific differences.* In these plots, the x-axis contains the cluster numbers, and the y-axis is the number of uniquely expressed genes. The data is displayed in a bar plot with the number of genes specific to a cluster displayed.

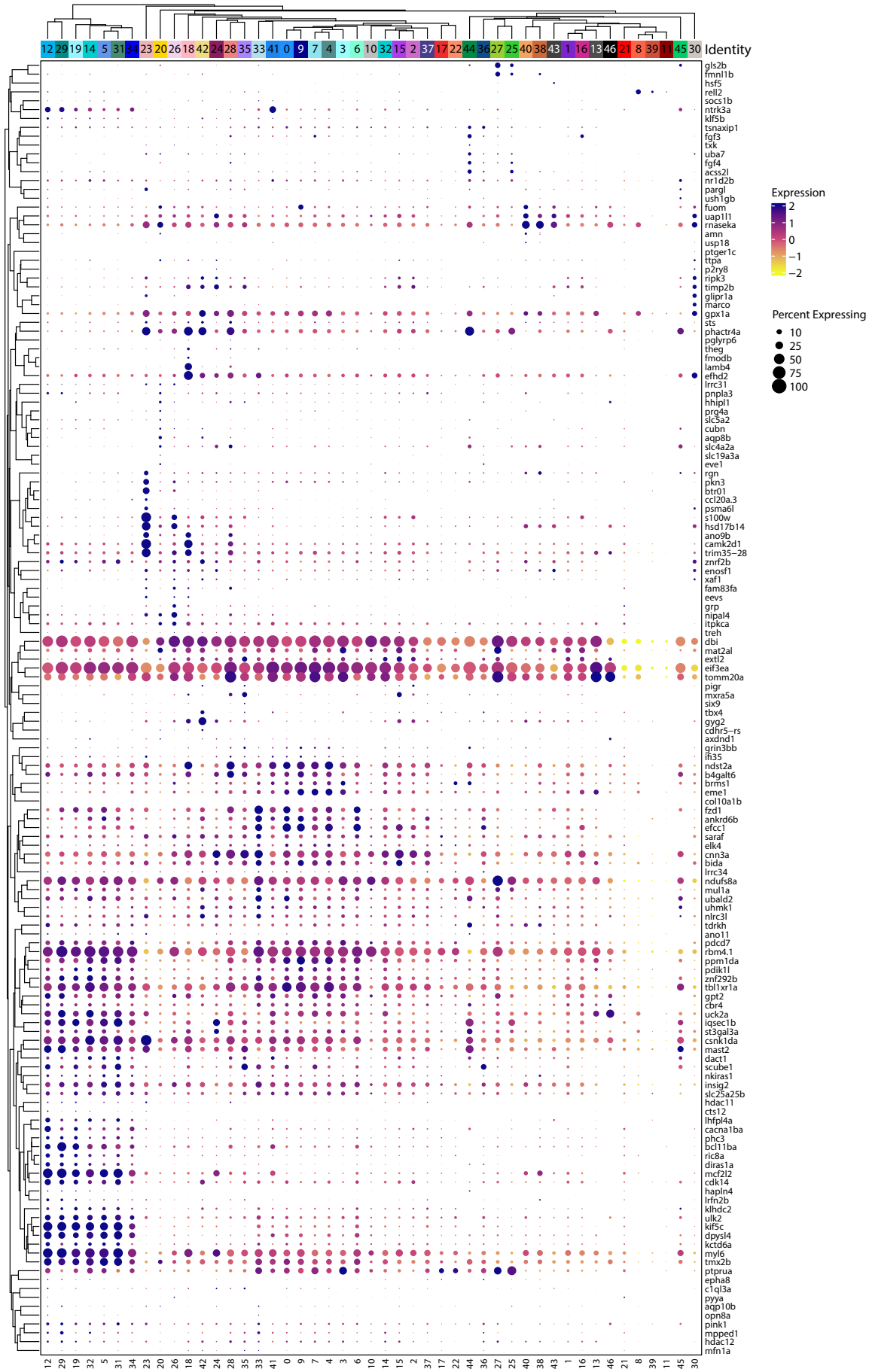


Figure S4.28: *Expression patterns of genes not identified in syngnathid fishes.* Genes are plotted in a clustered dot plot where the cluster numbers are on the x-axis, gene names are on the y-axis, and both are clustered based on similarity of expression. The size of the dot correlates with the percentage of cells that express it and the brightness corresponds to its expression level.

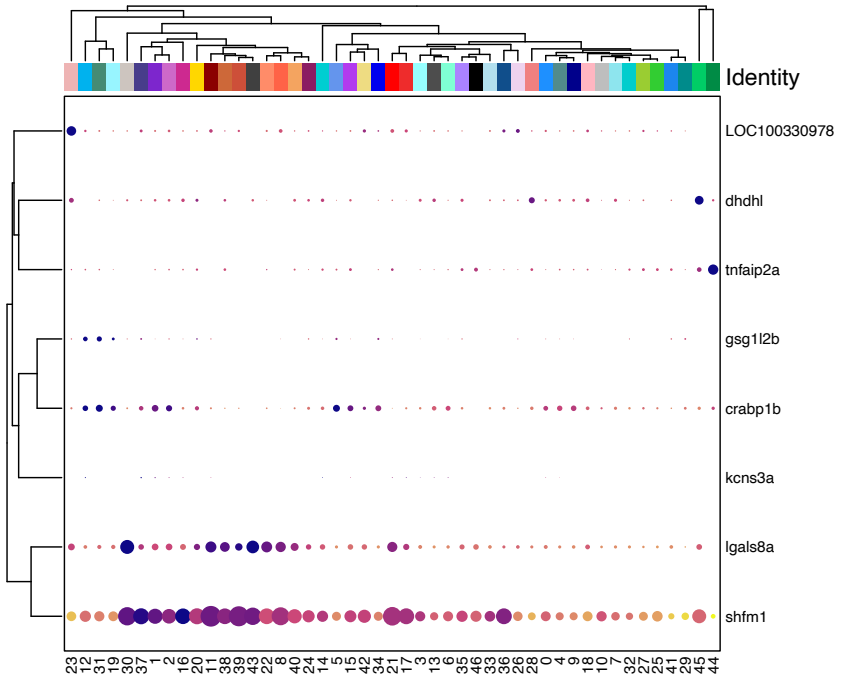


Figure S4.29: *Expression patterns of genes not identified in perciformes fishes.* Genes are plotted in a clustered dot plot where the cluster numbers are on the x-axis, gene names are on the y-axis, and both are clustered based on similarity of expression. The size of the dot correlates with the percentage of cells that express it and the brightness corresponds to its expression level.

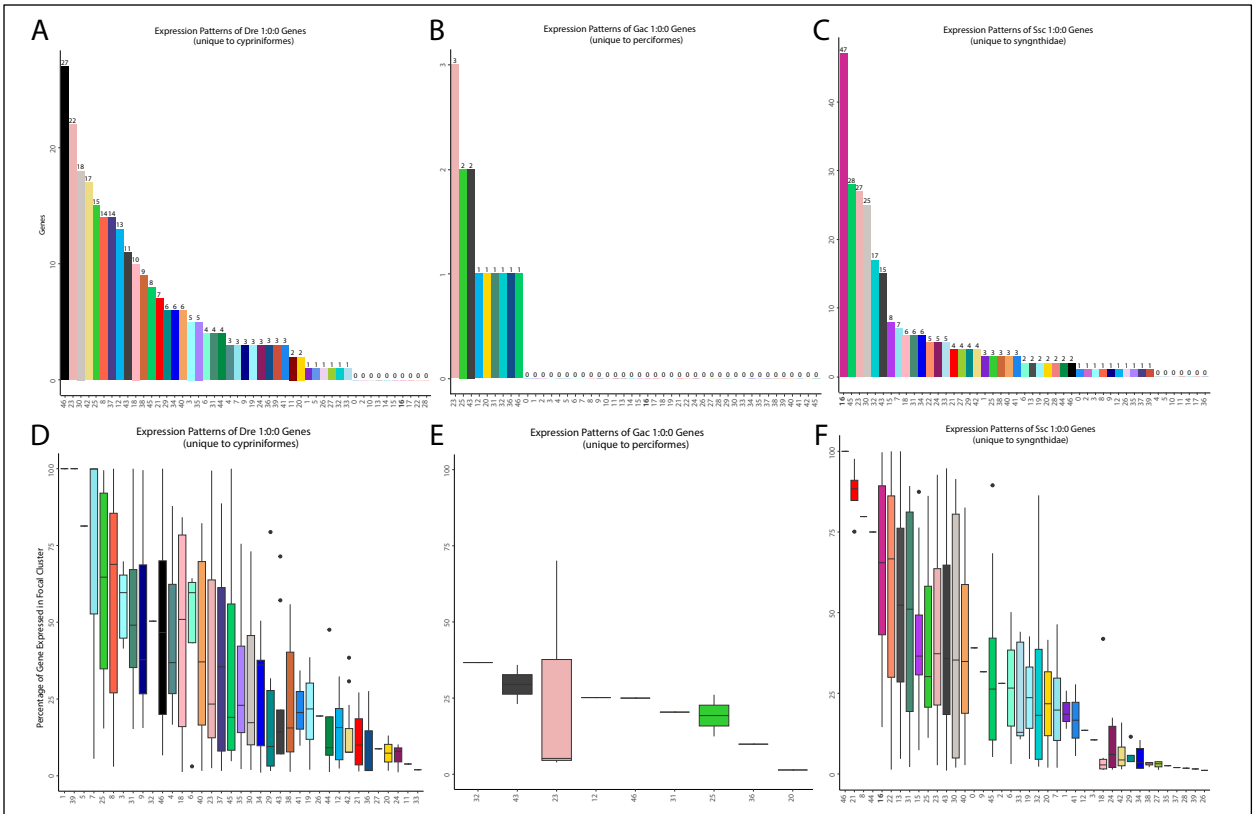


Figure S4.30: *Syngnathids have the greatest number of uniquely identified genes in the pharyngeal arch.* Genes that were solely identified in one clade via the GENESPACE analysis are plotted above. The cluster which has the highest expression of each unique gene is plotted in a barplot of panels A-C. The percentage of the unique gene expressed in that cluster is shown in panels D-F.

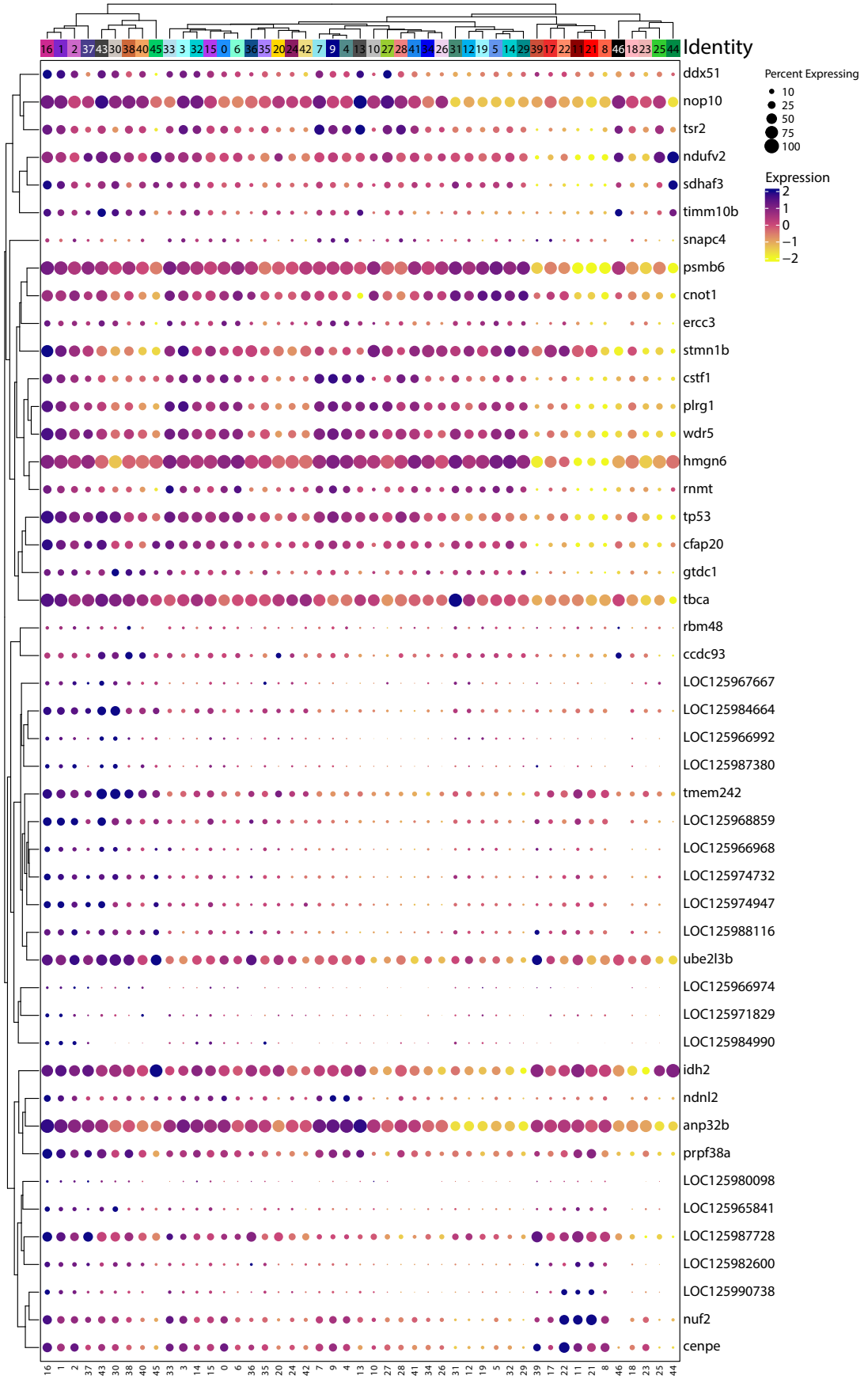


Figure S4.31: *Expression patterns of genes only identified in syngnathid fishes that had the highest percentage of cells from the PA.* Genes are plotted in a clustered dot plot where the cluster numbers are on the x-axis, gene names are on the y-axis, and both are clustered based on similarity of expression. The size of the dot correlates with the percentage of cells that express it and the brightness corresponds to its expression level.

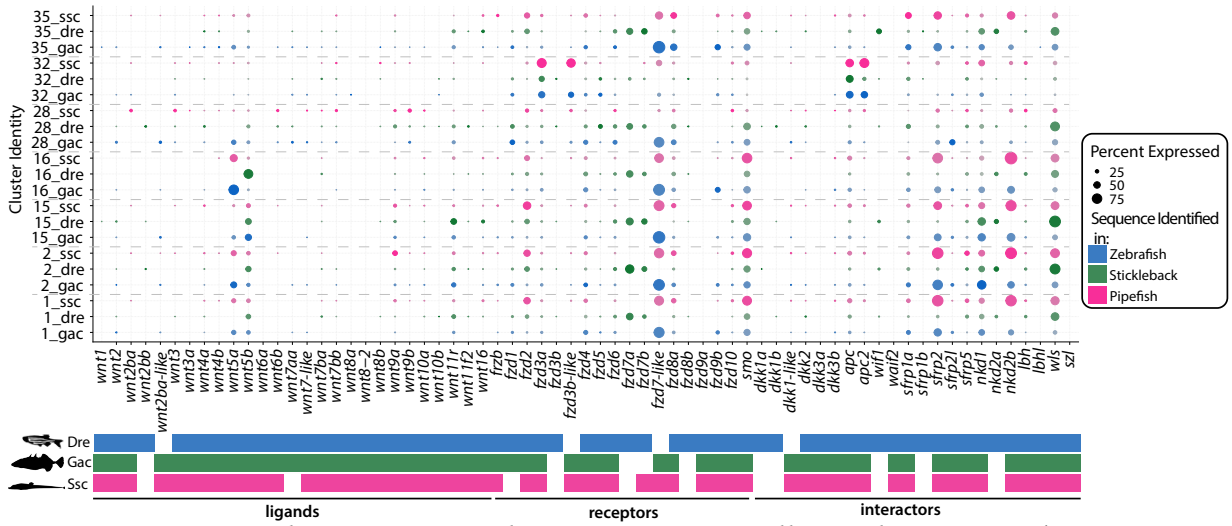


Figure S4.32: *Wnt pathway in CNC and connective tissue cell types by species.* Clusters are listed on the y-axis and are broken down by species. Wnt ligands, receptors, and interactors are distinguished on the x-axis and are organized by function. Dots represent the percentage of cells for a specific cluster that express each gene. The presence/absence of genes in each species are distinguished in the colored bars below the graph. Empty boxes represent a protein coding gene loss in a species.

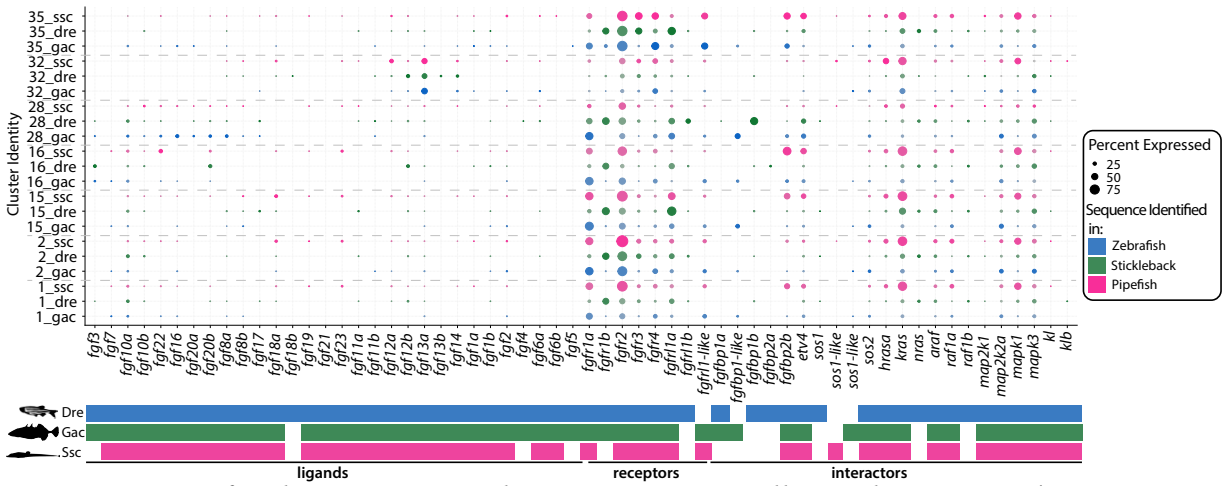


Figure S4.33: *Fgf pathway in CNC and connective tissue cell types by species*. Clusters are listed on the y-axis and are broken down by species. Fgf ligands, receptors, and interactors are distinguished on the x-axis and are organized by function. Dots represent the percentage of cells for a specific cluster that express each gene. The presence/absence of genes in each species are distinguished in the colored bars below the graph. Empty boxes represent a protein coding gene loss in a species.

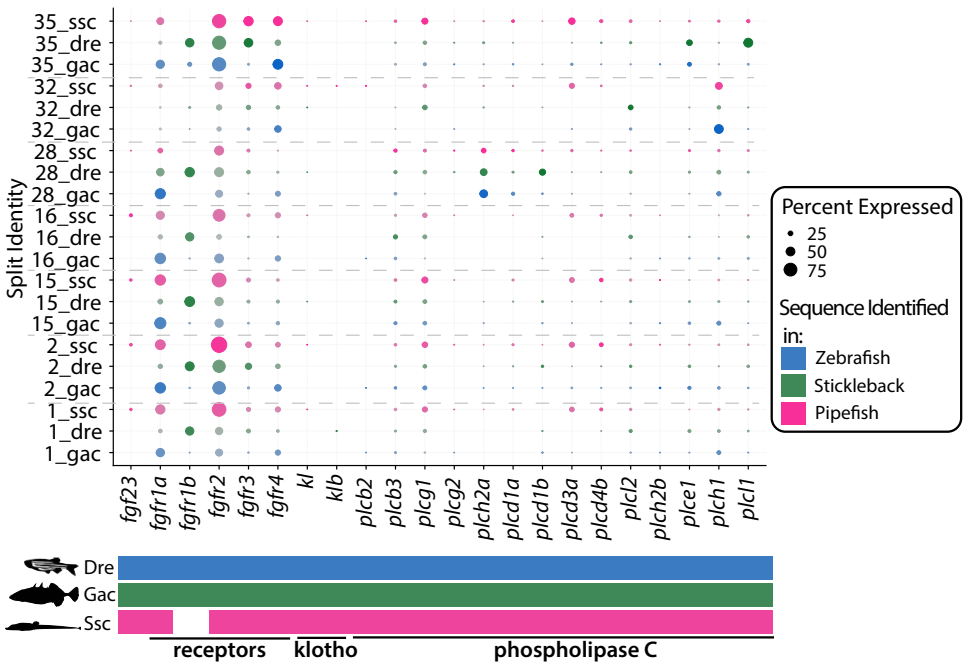


Figure S4.34: *fgf23, receptors, and potential co-factors in CNC and connective tissue cell types by species.* Clusters are listed on the y-axis and are broken down by species. Fgf receptors and *fgf23* co-factors are distinguished on the x-axis and are organized by function. Dots represent the percentage of cells for a specific cluster that express each gene. The presence/absence of genes in each species are distinguished in the colored bars below the graph. Empty boxes represent a protein coding gene loss in a species.

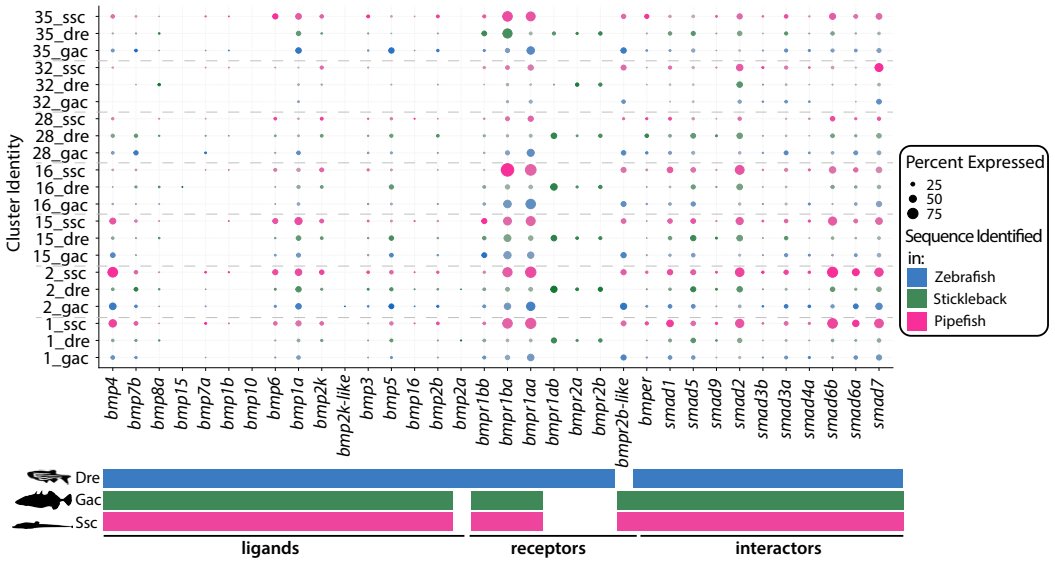


Figure S4.35: *Bmp pathway in CNC and connective tissue cell types by species*. Clusters are listed on the y-axis and are broken down by species. Bmp ligands, receptors, and interactors are distinguished on the x-axis and are organized by function. Dots represent the percentage of cells for a specific cluster that express each gene. The presence/absence of genes in each species are distinguished in the colored bars below the graph. Empty boxes represent a protein coding gene loss in a species.

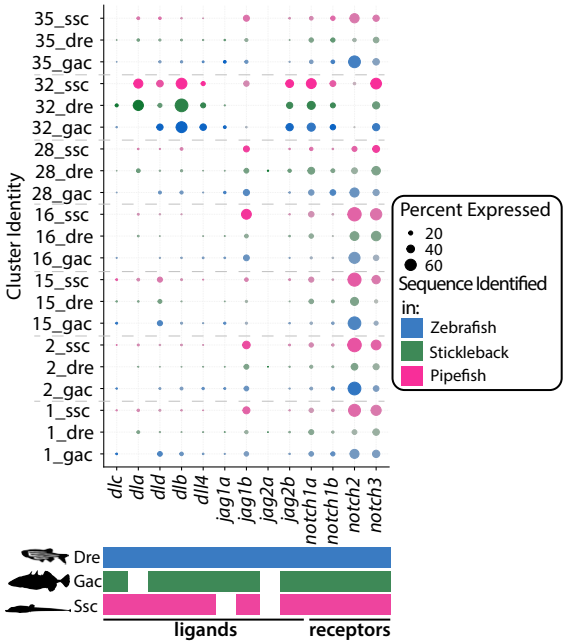


Figure S4.36: *Notch pathway in CNC and connective tissue cell types by species.* Clusters are listed on the y-axis and are broken down by species. Notch ligands, receptors, and interactors are distinguished on the x-axis and are organized by function. Dots represent the percentage of cells for a specific cluster that express each gene. The presence/absence of genes in each species are distinguished in the colored bars below the graph. Empty boxes represent a protein coding gene loss in a species.

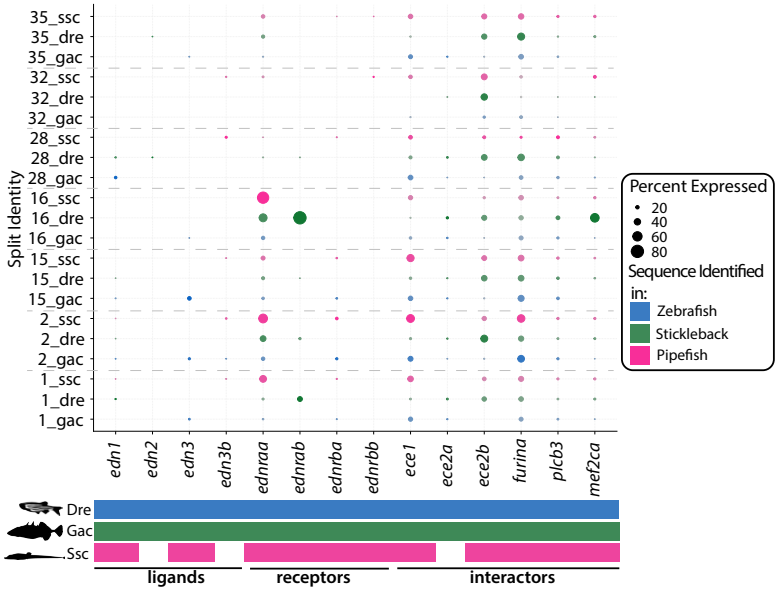


Figure S4.37: *Endothelial pathway in CNC and connective tissue cell types by species.* Clusters are listed on the y-axis and are broken down by species. Endothelian ligands, receptors, and interactors are distinguished on the x-axis and are organized by function. Dots represent the percentage of cells for a specific cluster that express each gene. The presence/absence of genes in each species are distinguished in the colored bars below the graph. Empty boxes represent a protein coding gene loss in a species.

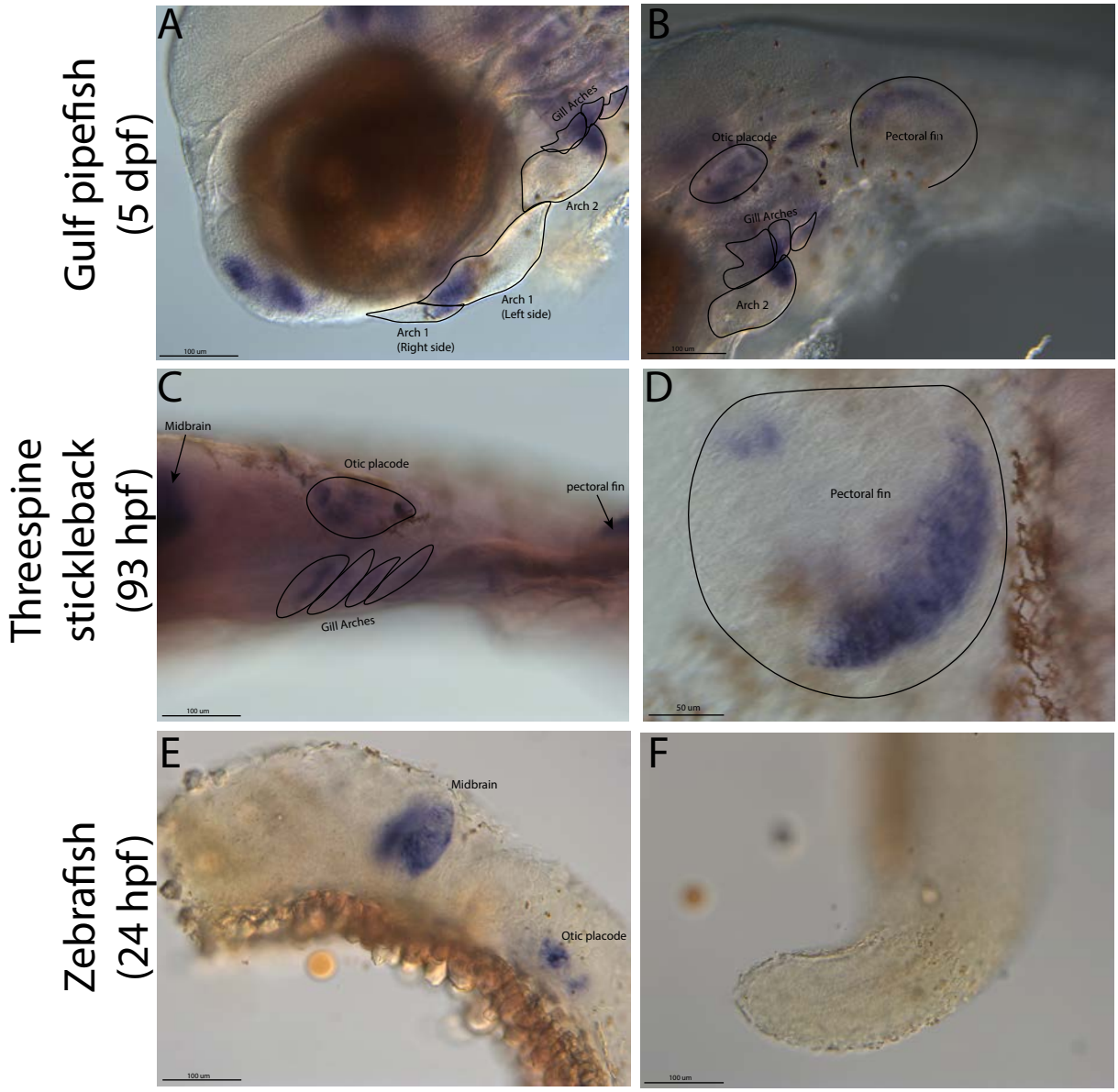


Figure S4.38: *Fgf22 has gained CNC pharyngeal arch and fin expression in percomorphs. In situ* hybridizations from Gulf pipefish, threespine stickleback, and zebrafish reveal spatial gene expression of *fgf22*. All three fishes had midbrain and otic placode staining (A, B, C, E). However, stickleback and pipefish had pharyngeal arch and fin staining (B, D, F). Pharyngeal arch staining of *fgf22* was not detected in stickleback via *in situs* at 70hpf, perhaps due to minimal expression. Fishes were mounted laterally.

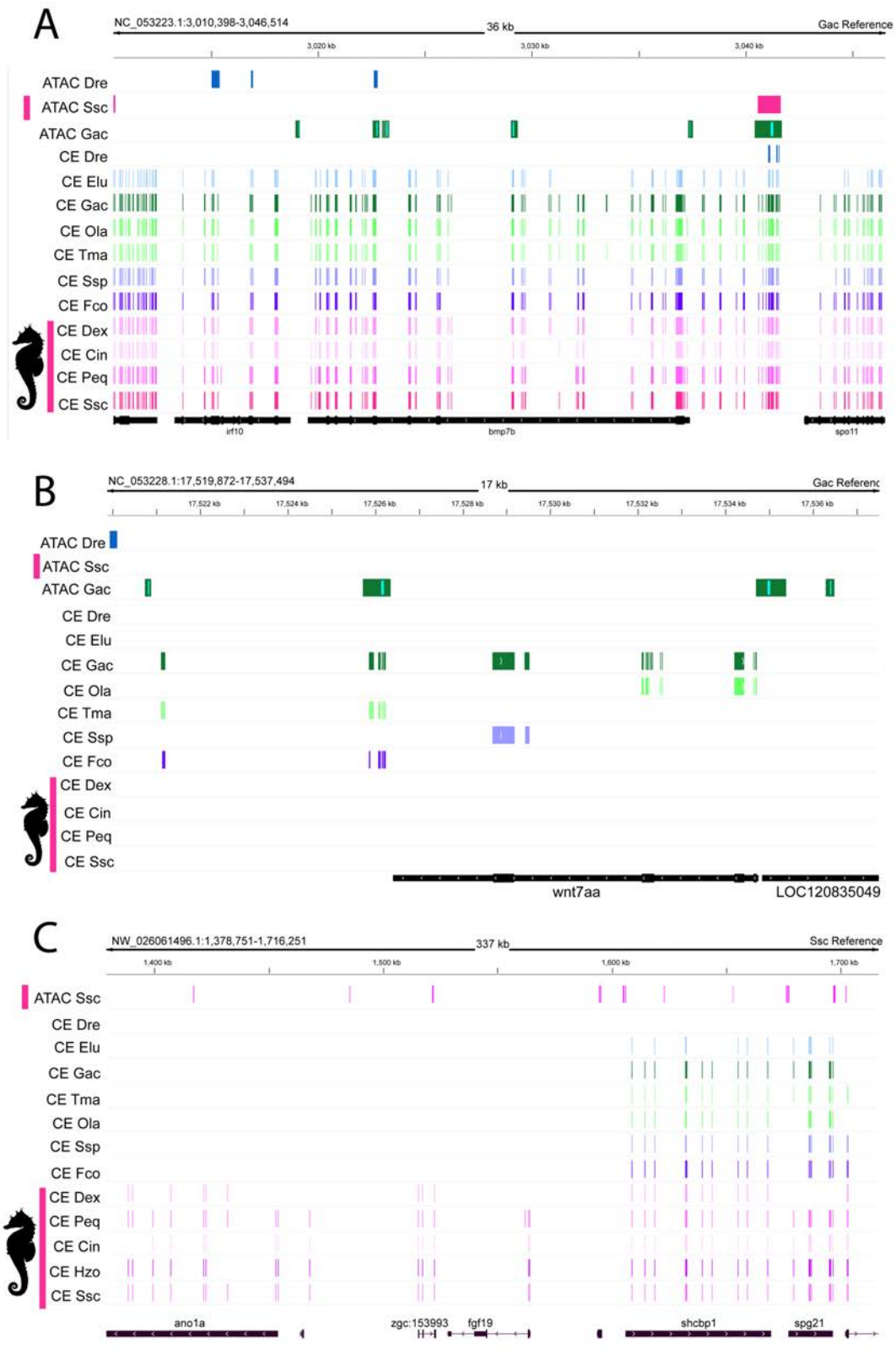


Figure S4.39: *Potential changes in gene sequence and regulation may lead to Gulf pipefish expression differences*. Panels A-C illustrate identified conserved elements (CE) and ATACseq peaks. Pink bars and seahorses to the left of the species names indicate syngnathids. Gene models are shown at the base of the graph, whereby thicker regions show exons and thinner regions show introns. Plots illustrate changes in genomic regions in genes that have altered expression or content in Gulf pipefish.

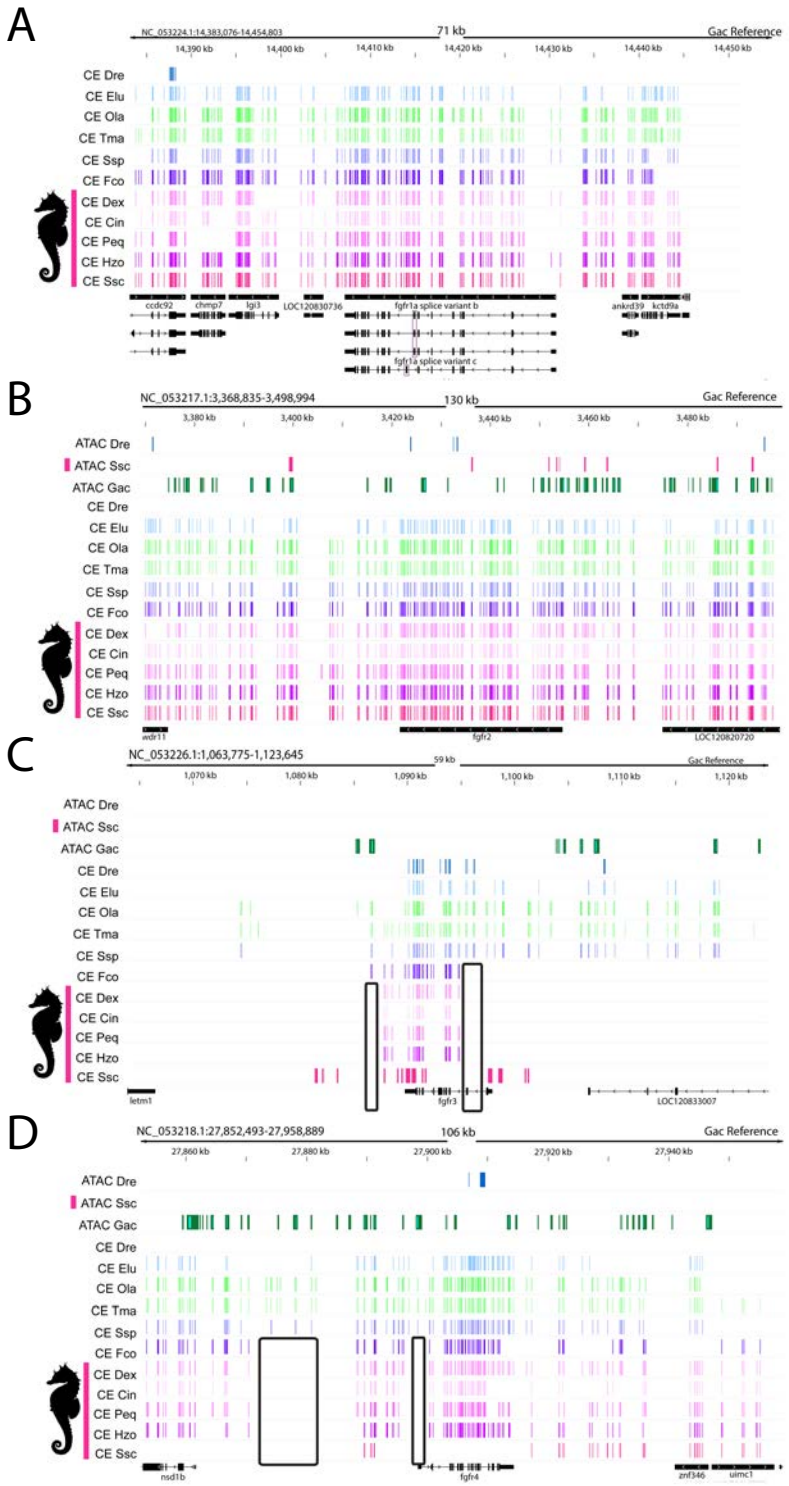


Figure S4.41: *Potential changes in fgfr gene regulation may lead to Gulf pipefish expression differences*. Panels A-D illustrate identified conserved elements (CE) and ATACseq peaks. Pink bars and seahorses to the left of the species names indicate syngnathids. Gene models are shown at the base of the graph, whereby thicker regions show exons and thinner regions show introns. Stickleback is the reference genome for all the plots.

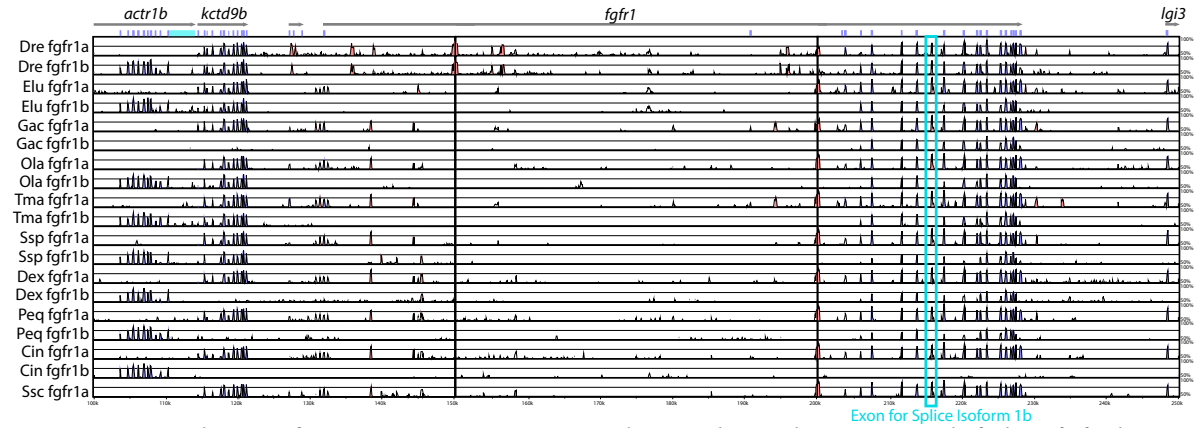


Figure S4.42: *Splice isoform 1B is missing in northern pike and percomorph fishes fgfr1b.* Species sequence conservation is plotted on a VISTA plot. Exons are shown in dark blue, conserved non-coding elements in orange, and UTRs in light blue. The exon for splice isoform 1b is boxed in cyan. The teleost species are plotted against spotted gar, which has not experienced the teleost whole genome duplication event.