

Supplemental Methods – Naftaly *et al.* 2021

Nuclei Isolation and ATAC-seq library preparation

To isolate nuclei, half of the liver was homogenized in 1X PBS with proteinase inhibitor cocktail (PIC, cOmplete tablets Roche). The homogenized cells were then fixed using 16% formaldehyde and washed twice with PBS+PIC. The cells were lysed using a lysis buffer containing 1M Tris-HCl, pH=8, 0.5M EDTA, 10% NP-40, 50% glycerol/molecular grade H₂O, and 1X PIC. The nuclei were then diluted to 60,000 – 80,000 nuclei prior to starting the ATAC-seq library preparation. ATAC-seq library preparation was performed following Lu *et al.* (Lu *et al.* 2017). To integrate the sequencing adapters, the diluted nuclei were rinsed with 1X TAPS to remove EDTA remaining from the lysis step. Tn5 transposase was then added to the nuclei and the reaction was carried out for 30 minutes at 37 °C. Immediately following the addition of sequencing adapters, the DNA was purified using a New England Biolabs Monarch DNA Cleanup kit (T1030). These library fragments were then amplified using Phusion enzyme (F530N) and Nextera PCR primers (Supplemental Table S12), using the following PCR conditions: 72 °C for 5 minutes, 98 °C for 2 minutes, and thermocycling at 98 °C for 10 seconds, 63 °C for 30 seconds, and 72 °C for 1 minute. Libraries were amplified for 13 cycles and Serapure beads were used to remove fragments smaller than 200 bp. ATAC-seq libraries were sequenced on Illumina NextSeq (2 x 150 bp) (Georgia Genomics & Bioinformatics Core, Athens, Georgia).

Long-read RNA alignment and isoform identification

Circular consensus sequences (CCS) were created from the raw subreads using ccs (SMRTlink, v6; --noPolish --minPasses 1). We ran ccs on each sample separately. cDNA primers (5' AAGCAGTGGTATCAACGCAGAGTACATGGGG and 3' AAGCAGTGGTATCAACGCAGAGTAC) were removed using lima (Iso-Seq3, v3.1; --isoseq -dump-clips --no-pbi). Trimmed circular consensus sequences were classified as full-length reads based on the presence of a poly(A) tail using refine (Iso-Seq3, v3.1; --require-polyA). Full-length reads were clustered using cluster (Iso-Seq3, v3.1; default parameters). Any two full length reads were considered to be part of the same isoform if the 5' overhang was less than 100 bp long, the 3' overhang was less than 30 bp long, and any internal gaps between the reads were less than 10 bp long (<https://github.com/PacificBiosciences/IsoSeq>). These overlapping reads were clustered together. Clustered full-length reads were polished using polish (Iso-Seq3, v3.1; default parameters). SAMtools was used to sort the alignments (Li et al. 2009; Li 2018). Reads that had accuracy scores less than 99% after polishing (low quality reads) were not considered further. The polished high-quality reads were aligned to the threespine stickleback genome (Ensembl build 97; (Jones et al. 2012; Aken et al. 2016) using minimap2 (v2.13) with the following parameters: -ax splice -uf --secondary=no -C5 (Li 2018). We also aligned the high-quality reads using deSALT (v1.5.6) with the following parameters: -x ccs -T (Liu et al. 2019). To remove redundancy among full length isoforms, identical isoforms were removed using cDNA Cupcake (collapse_isoforms_by_sam.py; https://github.com/Magdoll/cDNA_Cupcake). Count information for the reduced isoforms was then calculated using cDNA Cupcake (get_abundance_post_collapse.py, https://github.com/Magdoll/cDNA_Cupcake).

SQANTI characterization was then performed on the isoforms (Tardaguila et al. 2018). SQANTI characterizes isoforms into one of nine categories. Full splice matches (FSM) were defined as isoforms where all splice junctions fully matched an annotated gene. Incomplete

splice matches (ISM) were defined as isoforms where some, but not all splice junctions matched. Isoforms were also categorized as genic intron (isoforms were located fully within an intron of an annotated gene), hereafter called intronic isoforms, genic genomic (isoforms overlapped an exon and an intron of an annotated gene) hereafter called genic, intergenic (isoforms were located outside of any annotated gene), fusion (isoforms spanned two annotated genes), and antisense (isoforms overlapped an annotated gene, but on the complementary strand). Novel isoforms of previously annotated genes were classified as either novel in catalog (NIC) or novel not in catalog (NNC). These categories were defined solely on whether the splice acceptors or donors for the splice junctions were known or novel. The default splice donors and acceptors in SQANTI were used (the most common sequences found in humans: GT-AG, GC-AG, and AT-AC; (Mount 1982; Ohshima and Gotoh 1987; Shapiro and Senapathy 1987; Tardaguila et al. 2018)). Stickleback-specific splice junction sequences are not known. Protein coding predictions were completed using GMST in SQANTI with the Ensembl transcriptome as the reference (Tang et al. 2015; Tardaguila et al. 2018).

SQANTI contains multiple steps to remove potential artifacts or false positive isoforms from the dataset. During the QC step, isoforms with a signature of reverse transcriptase switching (RTS) or off-priming were identified. RTS results when the reverse transcriptase jumps across templates without stopping DNA synthesis while off-priming produces false cDNA molecules when the primer used in first-strand synthesis extends into intron-lariats or pre-mRNA regions. Next, the SQANTI machine learning classifier labeled artifacts based on features such as isoform length, presence of an open reading frame, and the number of full-length reads per isoform. Then, SQANTI utilized a true positive and a true negative set to set the expectations for isoform artifacts. Because we did not have a true positive or negative set, we used the default behavior of SQANTI, which assigns all the full splice match isoforms (FSM) as the true positives and the fusion isoforms as the true negatives. All potential artifacts identified by SQANTI were then removed.

Optimizing long-read transcriptome pipeline

To survey overall completeness in both transcriptomes, we used a benchmarking universal single-copy orthologs (BUSCO) analysis. BUSCO utilizes essential single copy orthologs that are expected to be present in complete transcriptomes (Simao et al. 2015; Seppey et al. 2019). We compared the existing threespine stickleback Ensembl transcriptome, deSALT transcriptome, and the minimap2 transcriptome to two databases, Metazoa with 954 genes and Actinopterygii with 3,640 genes. The Ensembl transcriptome contained mostly complete orthologs. 93.3% were complete orthologs in Metazoa (single copy orthologs: 630; duplicated orthologs: 260; Supplemental Fig S1) and 85.0% were complete orthologs in Actinopterygii (single copy orthologs: 2,370; duplicated orthologs: 724; Supplemental Fig S1). The minimap2 transcriptome had fewer complete orthologs. 56.3% of the Metazoan gene sets were complete (single copy orthologs: 231; duplicated orthologs: 306; Supplemental Fig S1) and 37.4% of Actinopterygian gene sets were complete (single copy orthologs: 715; duplicated orthologs: 646; Supplemental Fig S1). The deSALT transcriptome had the fewest complete orthologs. 27.0% of the Metazoan gene sets were complete (single copy orthologs: 108; duplicated orthologs: 150; Supplemental Fig S1) and 16.9% of the Actinopterygian gene sets were complete (single copy orthologs: 296; duplicated orthologs: 320; Supplemental Fig S1). The lower level of completeness produced by deSALT seems to be driven, in part, by the identification of fewer protein-coding isoforms. In the deSALT transcriptome, 11,156 isoforms (23.1%) were protein-

coding out of 48,345 total isoforms identified. In comparison, the minimap2 transcriptome contained 18,058 protein coding isoforms (68.3%) out of a total of 26,432 isoforms.

We were also concerned that the full Iso-Seq3 pipeline was too stringent and that we were missing genes and isoforms because of this. Isoforms were identified for the female brain sample using CCS reads. To test this, we identified isoforms using the CCS reads directly, bypassing much of the Iso-Seq3 pipeline, for the female brain as a representative sample. The CCS reads were directly aligned to the threespine stickleback genome (Ensembl build 97; (Jones et al. 2012; Aken et al. 2016) using GMAP (v8.3) (Wu and Watanabe 2005) using the following parameters: `-n 0 --cross-species -max-intronlength-ends 200000 -z sense_force`. Redundant isoforms were removed using cDNA Cupcake as previously outlined. Count information was not collected as this requires information from the classification step of the Iso-Seq3 pipeline. SQANTI was then run on the sample to classify genes and isoforms. We then compared the CCS transcriptome with the full Iso-Seq3 transcriptome. Using the CCS reads, we identified 1,784 genes and 1,982 isoforms (compared to 1,282 genes and 1,355 isoforms with the full Iso-Seq3 pipeline). We used BUSCO to examine the completeness of the female brain CCS transcriptome compared to the female brain full Iso-Seq3 transcriptome. In Metazoa gene sets, the female brain CCS transcriptome contained 0.8% of the complete orthologs (single copy orthologs: 7; duplicated orthologs: 1; Supplemental Fig S2) while the female brain full Iso-Seq3 transcriptome contained 9.1% complete orthologs (single copy orthologs: 84; duplicated orthologs: 3; Supplemental Fig S2). The same pattern was seen in Actinopterygii, where the female brain CCS transcriptome had 0.2% of the complete orthologs (single copy orthologs: 8; duplicated orthologs: 1; Supplemental Fig S2). The female brain full Iso-Seq3 transcriptome had 3.5% of the complete orthologs (single copy orthologs: 120; duplicated orthologs: 7; Supplemental Fig S2). Although using the CCS reads produced slightly more isoforms, the full Iso-Seq3 pipeline produced a transcriptome that contained over 10-fold more complete single copy orthologs.

Benchmarking universal single-copy orthologs (BUSCO)

For BUSCO analyses, we used the Metazoan (954 genes) and Actinopterygii (3,640 genes) lineages for comparison (OrthoDB v10). The Actinopterygii database was used because threespine stickleback fish are teleosts, the largest infraclass of Actinopterygii. The predicted amino acid sequences for all protein coding genes for the full transcriptome after SQANTI filtering were inspected (BUSCO gene set assessment). Protein coding genes from build 97 from Ensembl were also assessed. The zebrafish (*Danio rerio*) was set as the default species and all other parameters were left at the default settings.

Assessing the completeness of each transcriptome

To further assess whether our samples were sequenced to an adequate depth, we utilized a subsampling approach (Workman et al. 2018). If our samples were not sequenced sufficiently, the full set of isoforms may not have been captured. CCS reads were subsampled at 5%, 15%, 25%, 35%, 50%, 65%, 75%, 85%, and 95% of each individual sample using Picard (v2.16; DownsampleSam VALIDATION_STRINGENCY=SILENT) (<http://broadinstitute.github.io/picard>). The subsampled CCS reads were compared to the nucleotide sequences from the full tissue transcriptome using BLAST (v2.2.6, BLASTN, default parameters) (Altschul et al. 1990; Altschul et al. 1997; Camacho et al. 2009). The BLAST results for each sample were filtered using custom Python scripts. All BLAST alignments that covered at least 50% of the subsampled CCS read and at least 50% of an isoform from the full sample

transcriptome were retained. The total proportion of isoforms detected in each subsample compared to the full sample transcriptome was calculated.

Short read RNA analysis

For short read RNA sequencing, low quality regions and residual adapters were removed using Trimmomatic (v0.36) with default parameters including ILLUMINACLIP (Bolger et al. 2014). Trimmed reads were aligned to the threespine stickleback genome (Ensembl build 97 (Jones et al. 2012; Aken et al. 2016) using HISAT2 (v2.1) with the following parameters: phred33, rna-strandness FR (Kim et al. 2015). SQANTI uses short reads to examine the overall expression of individual isoforms and verify splice junctions (Tardaguila et al. 2018). Gene expression matrices with TPM (transcripts per million) were calculated using kallisto (v0.46; quant function) (Bray et al. 2016) for all transcriptomes. Normalized read counts were calculated using DESeq2 (Love et al. 2014). Read coverage across splice junctions was calculated using STAR (v2.7, default parameters) (Dobin et al. 2013). STAR and kallisto output were then used as input for sqanti_qc.py.

ATAC-seq genome coverage at TSSs

Residual adapter sequences from the Nextera primers were trimmed using Trimmomatic (v0.36) with PE and ILLUMINACLIP (keeping both reads) and the following parameters: LEADING:0 TRAILING:0 MINLEN:30 (Bolger et al. 2014). Trimmed reads were aligned to the revised threespine stickleback genome (Nath et al. 2021), including the mitochondrion and unplaced scaffolds, using Bowtie 2 (v2.3.5; -X 1000 -no-unal) (Langmead and Salzberg 2012). Reads with a mapping quality less than 20 were filtered from the alignments using SAMtools (v1.10) (Li et al. 2009). PCR duplicates were removed using MarkDuplicates from Picard (v2.21.6, default parameters). The read coverage per bp was calculated using BEDTools (v2.26, genomecov -d) (Quinlan and Hall 2010). Custom Python scripts were used to average the read coverage across a 4kb window surrounding the Ensembl and Iso-Seq TSSs.

Characterizing Non-coding RNAs by size and genome location

Non-coding RNAs (ncRNAs) were characterized by size and genome location using custom Python scripts. All ncRNAs did not have detectable protein coding potential. ncRNAs are generally classified based on overall length: short ncRNAs are less than 200 bp and long ncRNAs are greater than 200bp (Jacquier 2009; Pauli et al. 2011). We then compared the location of each ncRNA to all annotated Ensembl transcripts and the newly identified Iso-Seq isoforms. We separated ncRNAs into these two length categories as well as three main classes: intergenic, intronic, or antisense (Pauli et al. 2012). Intergenic ncRNAs did not overlap with any genes. Intronic ncRNAs fell completely within an intron of another gene and shared no overlap with the surrounding exons. Antisense ncRNAs overlapped with a gene but were on the opposite strand. Any remaining ncRNAs were added to an unknown category.

Novel genes compared to six different fish species.

To determine if novel genes were threespine stickleback specific, we compared isoforms to several diverged fish genomes. These genomes included ninespine stickleback (*Pungitius pungitius*) (Varadharajan et al. 2019), fugu or Japanese puffer (*Takifugu rubripes*) (Aparicio et al. 2002), medaka (*Oryzias latipes*) (Kasahara et al. 2007), Atlantic cod (*Gadus morhua*) (Star et al. 2011), zebrafish (*Danio rerio*) (Howe et al. 2013), and spotted gar (*Lepisosteus oculatus*)

(Braasch et al. 2016) as the outgroup. BLAST was used to compare the novel isoforms to each genome (v2.2.6, BLASTN, default parameters) (Altschul et al. 1990; Altschul et al. 1997; Camacho et al. 2009). The BLAST results were filtered using custom Python scripts. Positive alignments were counted when at least 50% of the query sequence matched the subject sequence.

Gene Ontology Analysis

Gene IDs and GO terms were downloaded from Ensembl using BioMart (Smedley et al. 2015). GO terms of the novel genes that were identified through InterProScan were also added to the complete list of GO terms for threespine stickleback. The total number of occurrences for each GO term was calculated for several analyses. GO term enrichment in each set was compared against 10,000 random permutations of the same sample size randomly drawn from the total set of genes.

References

- Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, Garcia Giron C, Hourlier T et al. 2016. The Ensembl gene annotation system. *Database (Oxford)* **2016**.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol* **215**: 403-410.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.
- Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**: 1301-1310.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.
- Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* **48**: 427-437.
- Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525-527.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.
- Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**: 498-503.
- Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55-61.
- Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**: 714-719.
- Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357-360.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.
- Liu B, Liu Y, Li J, Guo H, Zang T, Wang Y. 2019. deSALT: fast and accurate long transcriptomic read alignment with de Bruijn graph-based index. *Genome Biol* **20**: 274.
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

- Lu Z, Hofmeister BT, Vollmers C, DuBois RM, Schmitz RJ. 2017. Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Res* **45**: e41.
- Mount SM. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res* **10**: 459-472.
- Nath S, Shaw DE, White MA. 2021. Improved contiguity of the threespine stickleback genome using long-read sequencing. *G3 (Bethesda)* **11**.
- Ohshima Y, Gotoh Y. 1987. Signals for the selection of a splice site in pre-mRNA. *J Mol Biol* **195**: 247-259.
- Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A et al. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22**: 577-591.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.
- Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. In *Methods in Molecular Biology*, Vol 1962 (ed. M Kollmar). Humana, New York, NY.
- Shapiro MB, Senapathy P. 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res* **15**: 7155-7174.
- Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210-3212.
- Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* **43**: W589-598.
- Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrom M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**: 207-210.
- Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**: e78.
- Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, 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*.
- Varadharajan S, Rastas P, Loytynoja A, Matschiner M, Calboli FCF, Guo B, Nederbragt AJ, Jakobsen KS, Merila J. 2019. A High-Quality Assembly of the Nine-Spined Stickleback (*Pungitius pungitius*) Genome. *Genome Biol Evol* **11**: 3291-3308.
- Workman RE, Myrka AM, Wong GW, Tseng E, Welch KC, Jr., Timp W. 2018. Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. *Gigascience* **7**: 1-12.
- Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**: 1859-1875.