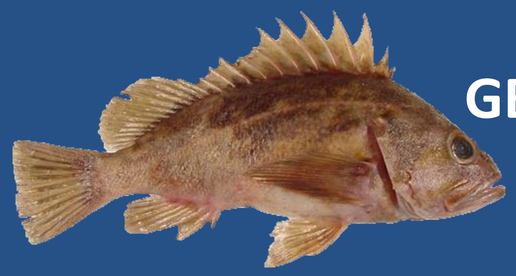


FINDING THE TOOLS: USING MULTIPLE ASSEMBLERS TO GENERATE A COMPLETE TRANSCRIPTOME FROM A NON-MODEL LARVAL MARINE FISH, SEBASTES AURICULATUS

Green, J.M*[1], Cline, A.J.[1], Bernardi, G.[2], Jue, N.K.[1], Logan, C.A.[1]

[1] School of Natural Sciences, California State University Monterey Bay, Seaside, CA, USA

[2] Department of Ecology & Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, USA

Why RNAseq?

- Ecologists and evolutionary biologists use RNAseq to study population genomics and differential gene expression in non-model species[1]
- Without a genome, a de novo transcriptome assembly must be created from short sequence reads to create full length genes[2]
- **Goal:** Use RNAseq data from Sebastes auriculatus (brown rockfish) to identify transcriptome assembly program strength and weaknesses in non-model fish species.
- **Approach:** Build single and merged assemblies from 3 programs (Oases, Transabyss & Trinity) using different parameters (coverages and k-mer sizes).[1,2]

Comparing Methodologies

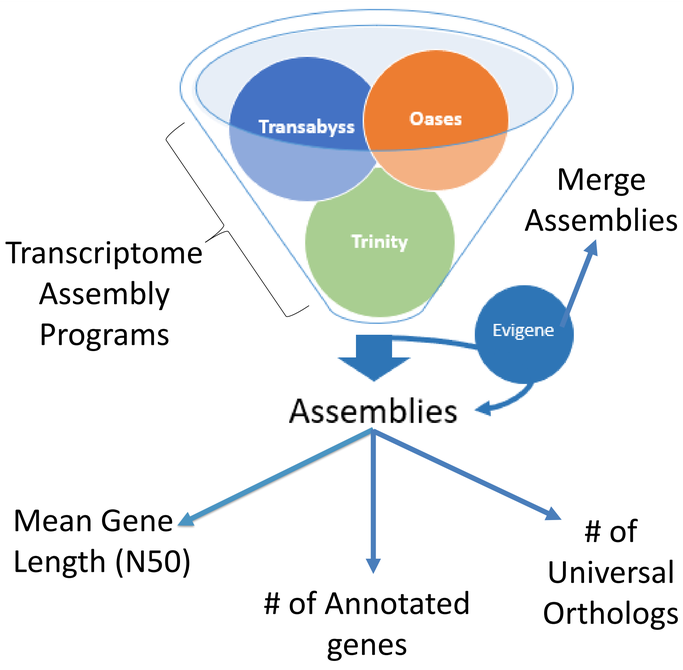


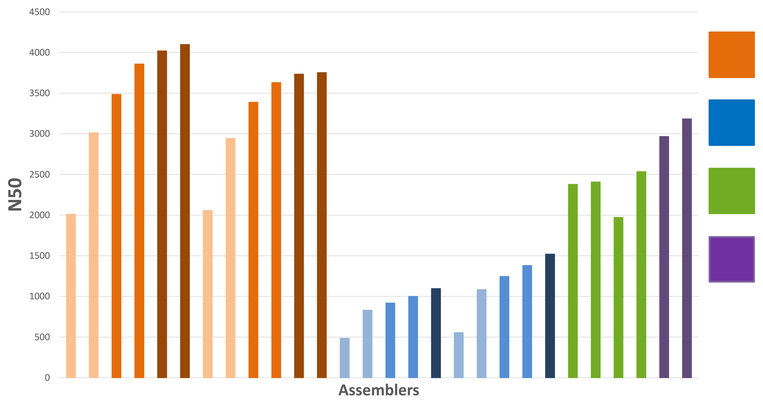
Figure 1: Our workflow used three de novo assemblers: Oases, Transabyss, and Trinity as well as the program used to merge assemblies (Evigene). Assemblies were assessed and ranked by by highest (A) average gene length (N50), (B) # of annotated genes, (C) # of full length conserved orthologs present.[3]

- To test the effect of Kmer size (size of gene pieces) and coverage (# of gene pieces required to assemble a gene) on quality, multiple assemblies were made in each program by varying these parameters
- Two "merged" assemblies were built from the best individual assemblies (Evigene)

Assembly Rankings

A) Mean gene length (N50):

Oases > Merged > Trinity > Transabyss



B) % of Annotated Genes:

Trinity > Transabyss > Oases

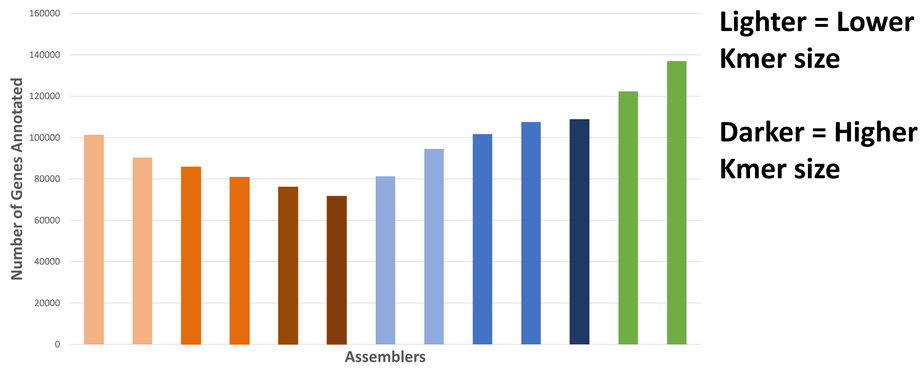


Figure 2: A) Mean gene length (N50) and B) number of annotated genes (Uniprot). Darker shades are assemblies with higher k-mer size (light to dark). Two sets of the same color bars represent two coverage sizes (left=100; right=30) from the same assembly program. Merged assemblies are not shown in B).

C) % of full length orthologs:

Transabyss > Oases > Trinity > Merged

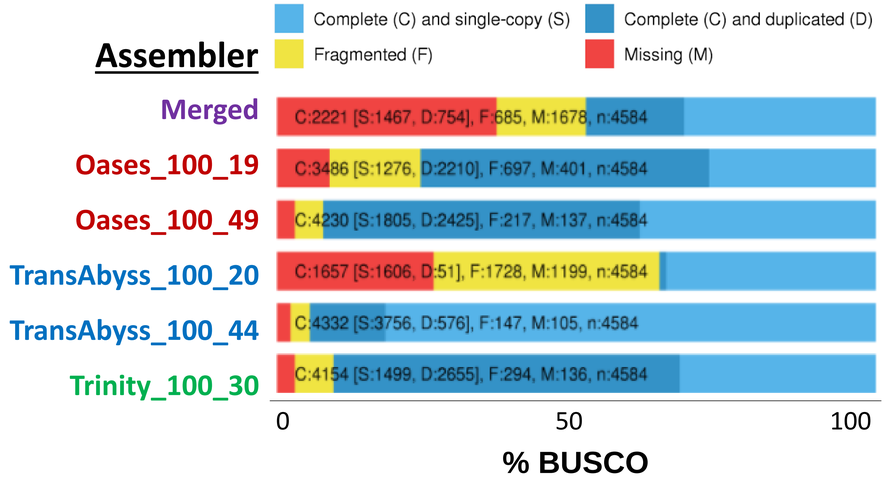


Figure 3: Benchmark Universal Single-Copy Orthologs (BUSCO) identifies commonly shared genes across taxa using the Actinopterygii database. Only highest performing assemblies based on N50 value and annotation rate are presented. Assembler program label is followed by coverage and k-mer size used.

Results

- Each assembly program has its own strengths:
  - **Oases → longest genes (highest N50)**
  - **Trinity → highest annotation rates**
  - **TransAbyss → most complete orthologs**
- Longer k-mers (size of gene pieces) used to assemble genes consistently resulted in longer genes
- Higher coverage (# of gene pieces required to assemble a gene) also resulted in longer genes
- **Take Home Message**. No one-size-fits-all "best" assembly program was identified, but higher coverage and k-mer size parameters generally produced better assemblies.

Future Work

- Compare assembly programs used here to determine if there is an effect on downstream differential gene expression data
- Test and compare RNAseq assemblies from other rockfish species to determine if results shown here are generalizable across related species

Literature Cited

[1] Bradnam, K.R. et al. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*. 2:10.

[3] Honaas, L.A., Wafula, E.K., Wickett, N.J., Der, J.P., Zhang, Y., Edger, P.P., et al. (2016) Selecting superior de novo transcriptome assemblies: lessons learned by leveraging the best plant genome. PLoS ONE 11(1): e0146062. doi:10.1371/journal.pone.0146062.

[3] Brown, C.T., Cohen, L.J., Pierce, T., Khudyakov, J. (2017) DIBSI Nonmodel mRNASeq Workshop. UC Davis. https://github.com/dib-lab/dibsi.

Funding & Acknowledgements

We sincerely thank NOAA SWFSC for aquarium space and staff time to support our time course experiments. We also thank Dr. Scott Hamilton's lab at MLML for fish collections, and Dr. Kristy Krocker's lab at UCSC for carbonate chemistry analysis. Logan Lab at CSUMB, BIO-430 students, and Knights of Castle Black coding team.

Southwest Fisheries Science Center

Sea Grant California

CALIFORNIA PRE-DOCTORAL PROGRAM
Sally Casanova Scholar