Warren J. Baker Endowment
for Excellence in Project-Based Learning
Robert D. Koob Endowment *for Student Success*

FINAL REPORT

Final reports will be published on the Cal Poly Digital Commons website(http://digitalcommons.calpoly.edu).

I. Project Title

Unveiling mechanisms of regeneration in a tunicate model using genome and transcriptome sequencing

II. Project Completion Date

January 15, 2020

III. Student(s), Department(s), and Major(s)

(1) Jack Sumner, Biological Sciences Department, Biological Sciences Major with Molecular and Cellular Biology Concentration

(2) Jenna Landy, Statistics Department, Statistics Major with Data Science Minor

(3) Charlie Liou, Mathematics Department, Mathematics Major with Data Science Minor

IV. Faculty Advisor and Department

Jean Davidson, Biological Sciences Department

V. Cooperating Industry, Agency, Non-Profit, or University Organization(s)

California Polytechnic State University, Stanford University

VI. Executive Summary

Through Baker-Koob funding, we obtained Cal Poly's first long read sequencer; our subsequent research marked the first use of next generation sequencing at Cal Poly and resulted in the first-ever draft sequence of the *B.* violaceus genome. Whole genome sequencing for B. violaceus was completed using a dual platform approach, implementing Illumina short read and Nanopore long read sequencing technologies for single read and hybrid assembly pipelines (Fig. 1). Approximately 7.5 Gb of Illumina and 900 Mb of Nanopore data were acquired.

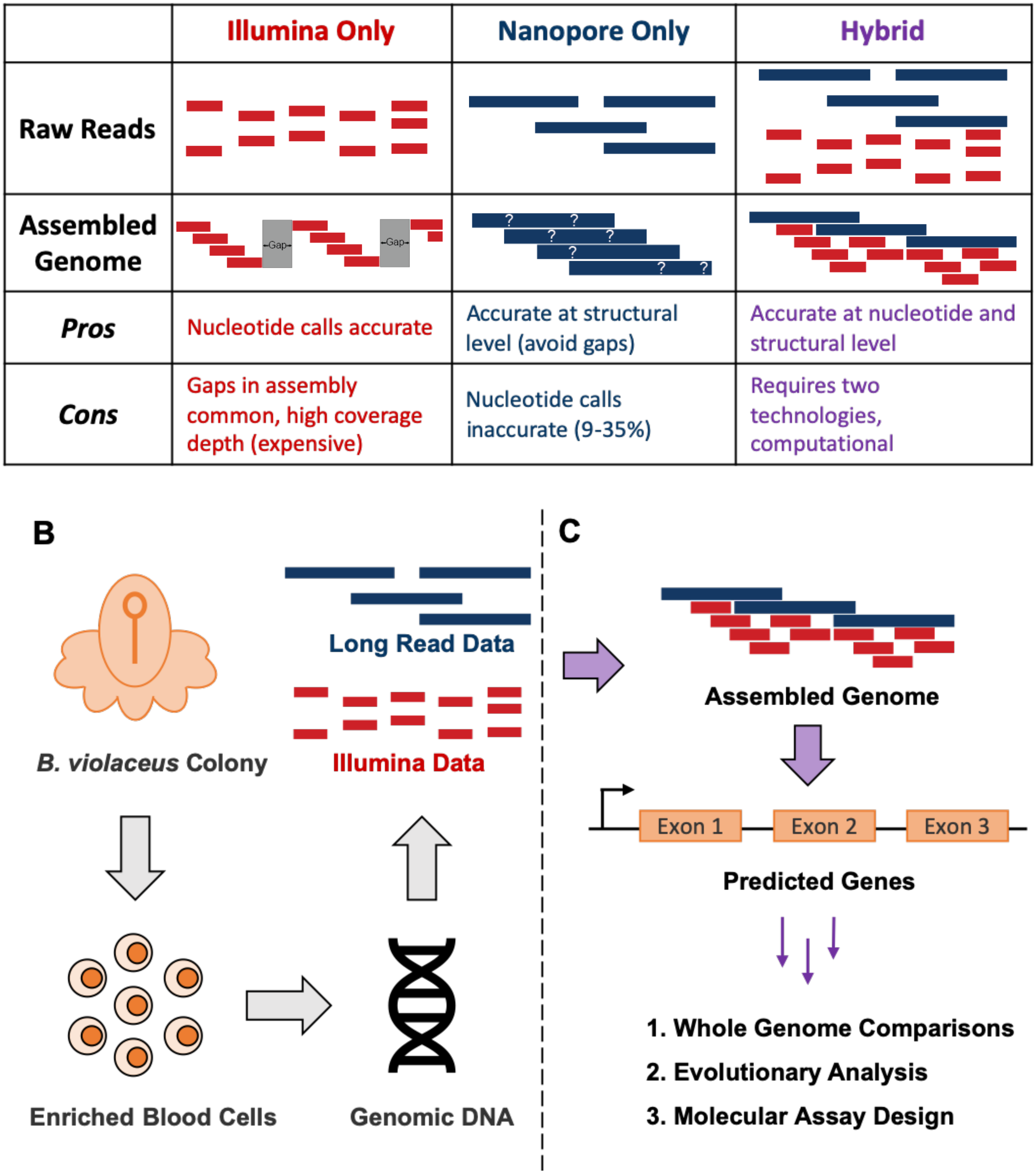
Assembly efforts have been broken down into two phases: single read assembly and hybrid assembly. Short reads were deconvoluted using Kraken V2.0.1 and assembled with String Graph Assembler V1.0.0 on Illumina's cloud compute service (i.e. BaseSpace) to develop the first draft genome for *B. violaceus* (Fig 2).
Using comparative bioinformatics algorithms such as Augustus V3.3.3 and GhostKOALA V2.2.0, we have annotated thousands of putative genes to further unveil the evolutionary origins of colonial ascidians and

identify potential gene regulatory networks. BUSCO assessment (V3.0) of this draft estimate 60-80% completion via observation of conserved eukaryotic and metazoan orthologs in putative gene predictions. Furthermore, the Keeling Lab at Cal Poly aims to apply genomic information on *B. violaceous* to expand available experimental techniques; designing molecular assays will provide empirical insight into the underlying mechanisms of regeneration in colonial ascidians. This genome sequence will play a critical role in the advancement of ascidian stem cell biology and promote the use of an underestimated non-model organism.

Hybrid assembly is a computationally complex task, requiring advanced computing resources currently unavailable at Cal Poly. This challenge has stalled hybrid assembly efforts; to overcome this obstacle, we have initiated collaborations with various budding Cal Poly resources (i.e. Digital Transformations Hub, Amazon Web Services [AWS]). These collaborations will pioneer "big data" analysis at Cal Poly and provide the foundation for accessible genomic pipelines to the institution's research community.

In May 2019, Jack Sumner presented these data at the Bay Area Stem Cell Conference in Asilomar, CA hosted by UCSF and Stanford. As the only undergraduate presenter, this conference provided unique, otherwise-inaccessible insight into methods necessary to overcome current research challenges. Early collaborations with Stanford faculty focused on tunicate genomics have started because of this conference and generous funding by the Baker-Koob Endowment.

Now that initial sequencing efforts have been completed, we are switching our focus to improving the assembly and annotation of the genome, in order to complete additional biological experimentation. Already, the Keeling lab has taken advantage of this new genome to study various WNT family genes. We are hoping to improve our genome with the addition of more long-read sequences, both from our in-house Nanopore and potentially collaborating with sequencing companies. We are excited to continue the work initiated with this grant – it has truly jumpstarted an exciting research agenda.



	Illumina Only	Nanopore Only	Hybrid
Raw Reads			
Assembled Genome	Gap Gap	? ? ? ? ? ? ?	
Pros	Nucleotide calls accurate	Accurate at structural level (avoid gaps)	Accurate at nucleotide and structural level
Cons	Gaps in assembly common, high coverage depth (expensive)	Nucleotide calls inaccurate (9-35%)	Requires two technologies, computational

Figure 1: Assembly methods and computational pipelines implemented in *B. violaceus* genome project. **(A)** Comparison of single read and hybrid assembly pipelines. **(B)** Experimental procedure to obtain high quality DNA and sequencing data. Grey arrow represents molecular and cellular biology techniques (order: Cell Enrichment, DNA Isolation, Nanopore and Illumina Sequencing). **(C)** Computational pipelines executed for assembly, annotation, and application of genome data. Purple arrows represent data manipulation to synthesize and analyze genomic information (order: Assembly, Annotation, Various Programs).

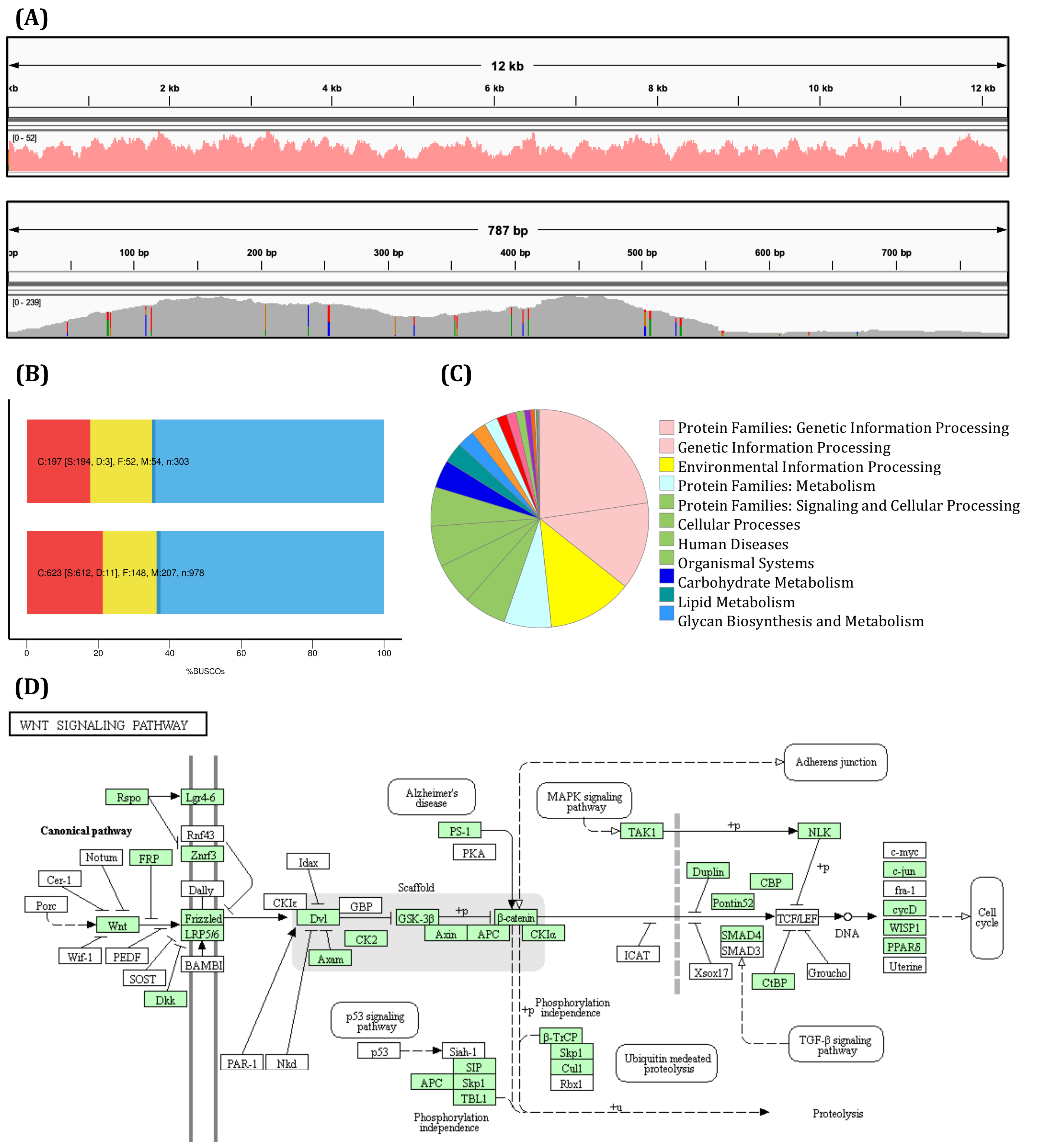


Figure 1: Bioinformatic analysis of *B. violaceus* draft genome emphasizing quality and gene annotation. (A) Assembly coverage visualization and possible repetitive region taken from Integrated Genomics Viewer V2.1.2 with reads mapped using Burrows Wheeler Aligner V1.1.4. *(Top)* Coverage scale ranges from 0-52 and shows relatively consistent coverage. *(Bottom)* Coverage scale ranges from 0-239 and shows extremes in coverage; this discrepancy is possibly due to mis-assembly from highly repetitive region. **(B)** BUSCO Analysis of the *B. violaceus* genome using Assembly A using BUSCO's eukaryote (top) and metazoan (bottom) databases. **(C)** Ortholog analysis of approximately nineteen-thousand putative genes; approximately seven thousand were paired with an ortholog and thus annotated. **(D)** KEGG reconstruction of canonical Wnt/B-catenin signaling pathway. Green boxes represent orthologs identified in *B. violaceus*. White boxes indicate genes that are observed in other organisms but are not observed in our draft assembly for *B. violaceus*.

VII. Major Accomplishments

(1) Premier application of next generation sequencing at Cal Poly

(2) Assembled and analyzed first draft of *B. violaceus* genome

(3) Disseminated data at regional conference hosted by Stanford and UCSF; Bay Area Stem Cell Conference, May 2019 (poster presentation)

VIII. Expenditure of Funds

Travel and registration for Bay Area Stem Cell conference	$500
Sequencing reagents	$3000
DNA isolation and quantification reagents	$1000
Data analysis	$500

IX. Impact on Student Learning

At the intersection of molecular biology and computer science, this multi-disciplinary project has engaged students from diverse academic backgrounds in the pursuit of thought-provoking science. Between the Davidson Lab and Keeling Lab, over ten students have actively synthesized their complementary skills sets to investigate, experiment, and analyze the underlying genomics of regeneration. Pipelines developed for computational analysis of the *B. violaceus* genome will be critical in providing accessible resources for future research efforts at Cal Poly. Projects in the Keeling lab, and moreover in any lab using *B. violaceus* as a model, will extend off this draft sequence to provide a wealth of information for years to come. Involved students have developed unique skills that cross traditional academic boundaries. For instance, Jack Sumner, an undergraduate studying biological science, now actively collaborates with Stanford faculty as a bioinformatician because of the skills he gained from this experience. Furthermore, students involved with this project have gained necessary computational and experimental skills to succeed in graduate education.