De novo assembly and preliminary annotation of the *Schizocardium californicum* genome

Gregory T. Concepcion[1], Paul Peluso[1], Paul Bump[2], Paul Gonzalez[2], Chris Lowe[2], Dan Rokhsar[3], David Rank[1]

1. PacBio, 1305 O'Brien Drive, Menlo Park, CA
2. Hopkins Marine Station of Stanford University, Pacific Grove, CA
3. University of California, Berkeley, CA

Introduction

Animals in the phylum Hemichordata have provided key understanding of the origins and development of body patterning and nervous system organization. However, efforts to sequence and assemble the genomes of highly heterozygous non-model organisms have proven to be difficult with traditional short read approaches. Long repetitive DNA structures, extensive structural variation between haplotypes in polyploid species, and large genome sizes are limiting factors to achieving highly contiguous genome assemblies.

Here we present the highly contiguous *de novo* assembly and preliminary annotation of an indirect developing hemichordate genome, *Schizocardium californicum,* using SMRT Sequencing long reads.

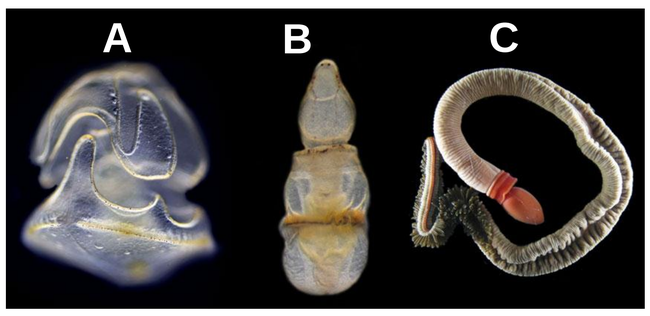


Figure 1. **Life stages of *S. californicum*. A)** Tornaria larval "swimming head" phase **B)** Early juvenile **C)** Mature adult

Sample Prep & Sequencing

High quality genomic DNA was isolated from hemichordate sperm from a single individual. RNA was collected from individuals spanning several life history stages. Long insert and Iso-Seq libraries, respectively, were constructed and SMRT Sequencing was performed on a PacBio Sequel System.

Figure 2. WGS raw subread length distribution Extracted gDNA was used to generate large insert (>30 kb) libraries for subsequent SMRT Sequencing.

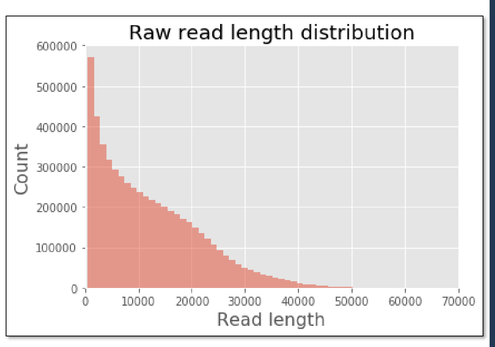
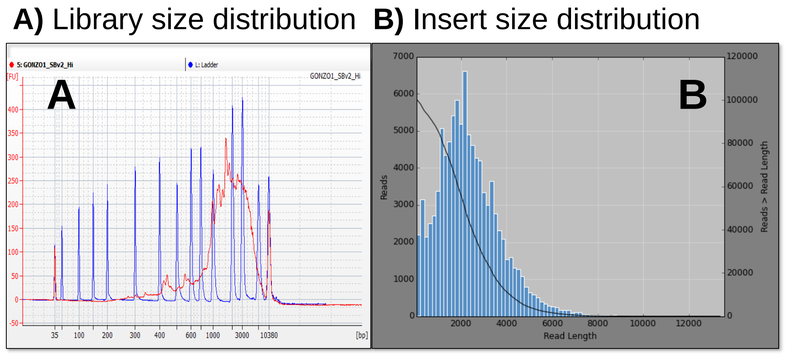


Figure 3. Iso-Seq library and insert sizing

A) Library size distribution **B)** Insert size distribution



Due to the RNA data having originated from several individuals unrelated to the gDNA specimen, the sequenced Iso-Seq transcripts were identified using blastx and used purely for annotation purposes.

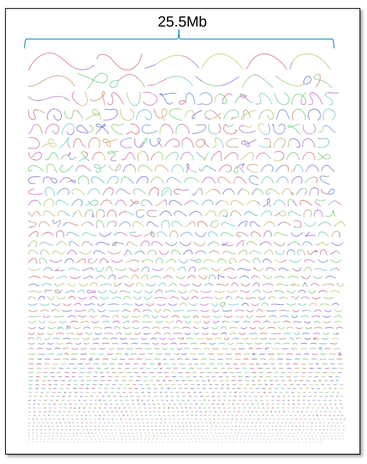
Bioinformatics

Using 113 Gbp of raw data for input and the Canu[2] assembler, a genome of approximately 1.7 Gb in total length with a contig N50 of ~758 kb spread across 5,692 contigs was generated. The 1.7 Gb assembly represents roughly double the expected genome size.

Table 1. Assembly results

Characteristic	Stat
Data amount	113 Gbp
Raw read N50	19.5 Kb
Contigs	5,692
Max contig length	5.5 Mb
Contig N50	758 Kb
Total assembly size	1.75 Gbp

Figure 4. **Bandage plot of Canu assembly unitigs** Plot showing the distribution of unitig lengths relative to the rest of the assembly. For scale, combined length of first 6 contigs is 25.5 Mb.



Assembly Completeness

BUSCO[3] analysis of assembly content shows that nearly 83% of the BUSCOs are duplicated, indicating assembly of both haploid phases in the assembly output .

BUSCOs (metazoa)	Mammal Assembly	*S. californicum* Canu
Complete BUSCOs	890	956
Complete Single Copy	819	163
Complete Duplicated	71	793
Fragmented	37	0
Missing	51	22
Total	978	978

Table 2. Assembly completeness Table comparing BUSCO results for high-quality mammalian and *S. schizocardium* assemblies

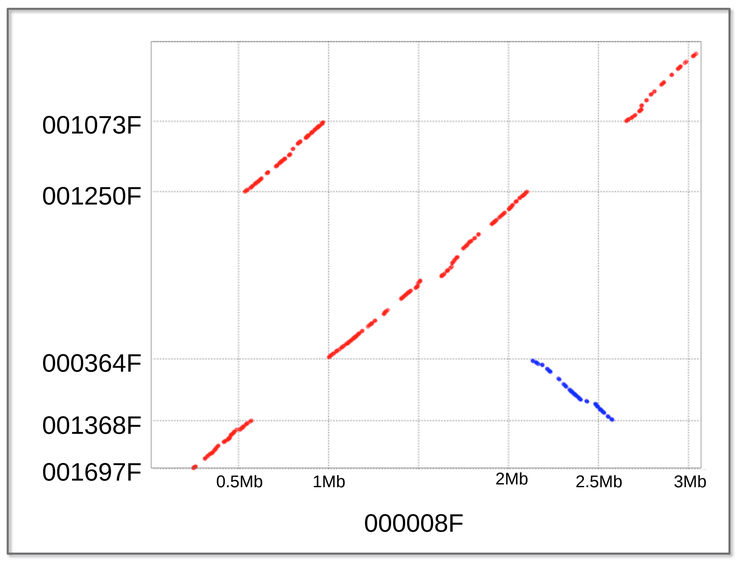


Figure 5. **Alignment overlap in primary contigs** Dotplot showing 5 primary contigs span nearly the entire length of one of the larger primary contigs.

Aligned Transcripts in IGV

Iso-Seq transcripts were aligned back to the Canu assembly using STAR (Spliced Transcript Aligner to Reference) and visualized in Integrative Genome Viewer[4] (IGV).

Figure 6. **Iso-Seq transcripts aligned to Canu assembly** A 740 kb region showing the annotation and transcript diversity in 5 gene regions

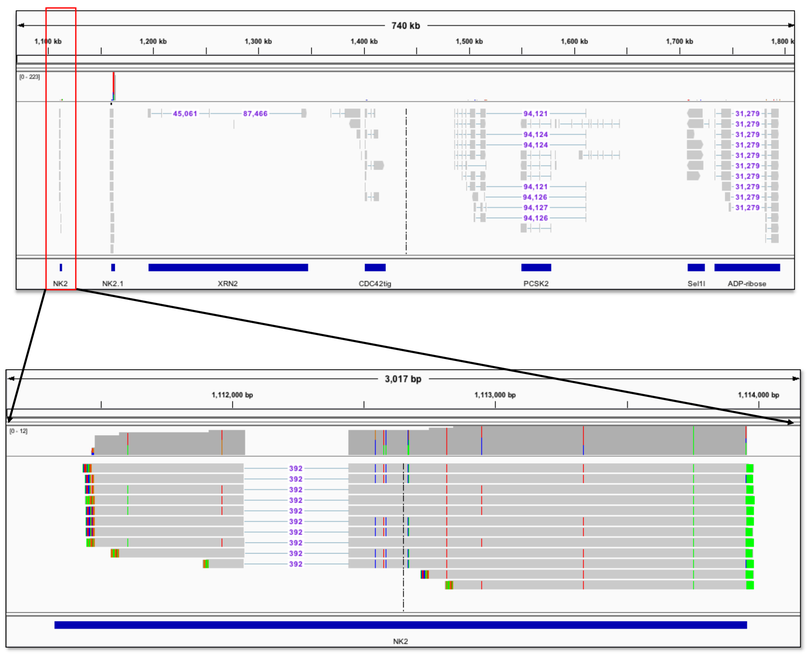


Figure 7. **Isoform diversity in NK2.1 gene region**

A 3,000 bp region showing transcript diversity from one particular locus.

Polymorphisms between transcripts and splice junctions are easily visualized using IGV.

Conclusions

- High quality genomic DNA that is suitable for long read sequencing is obtainable from hemichordate sperm samples.
- SMRT Sequencing of DNA and RNA enables both *de novo* assembly and annotation of highly heterozygous non-model organism genomes.
- High levels of heterozygosity contribute to assembly of the majority of both parental haplotypes in *S. californicum.*

References

1. Gonzalez, P., et al. The Adult Body Plan of Indirect Developing Hemichordates Develops by Adding a Hox-Patterned Trunk to an Anterior Larval Territory. *Curr. Biol.* 27, 87–95 (2017).
2. Koren S, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* (2017).
3. Dobin, A., et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21 (2013).
4. Simão, F.A., et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210-3212 (2015).
5. Thorvaldsdóttir, H., et al. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14, 178-192 (2013).