

Supplementary Materials for

A chromosome-scale genome assembly and karyotype of the ctenophore *Hormiphora californensis*

Darrin T. Schultz^{1,2,*}, Warren R. Francis³, Jakob D. McBroome¹, Lynne M. Christianson², Steven H.D. Haddock^{2,3}, Richard E. Green¹

Correspondence to: dts@ucsc.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figures S1 to S15
Tables S1 to S5

Other Supplementary Materials for this manuscript include the following:

Data S1 [SI_Data_S1_LibraryInfo.xlsx]

Materials and Methods

Figure S1

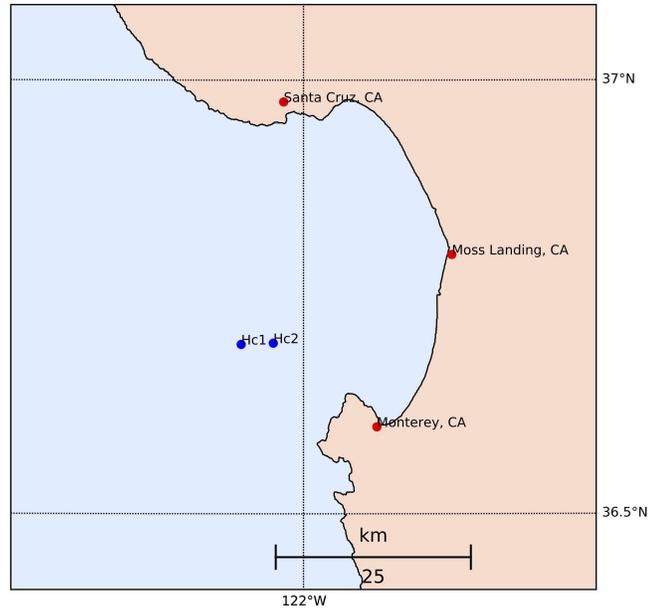


Figure S1. *Hormiphora californensis* and *B. forskalii* sample collection map. *H. californensis* individuals Hc1 and Hc2 were collected within two kilometers of one another three years apart. See collection conditions and parameters in Table S1.

Sequencing data preparation

H. californensis HMW DNA was isolated from individual Hc1 by lysing tissue in CTAB buffer (Dawson *et al.* 1998), then purifying the DNA with a chloroform, phenol:chloroform, chloroform, ethanol precipitation protocol (Sambrook and Russell 2006). Two PacBio SMRT CLR sequencing libraries were constructed and sequenced on three SMRT cells on a PacBio Sequel or Sequel II machine at UCD, yielding 27.4 Gbp of CLR subreads (Figure S2). A Hc1 HMW DNA extract was also used to create three Dovetail Chicago libraries at University of California Santa Cruz (UCSC) (Putnam *et al.* 2016), using either the DpnII or MluCI enzyme. The Chicago libraries were sequenced to a depth of 105 million read pairs. One Hc1 HMW DNA extract was used to construct a 10X chromium library at UCSC, and was sequenced to a depth of 74 million read pairs (Weisenfeld *et al.* 2017). Eight Hi-C libraries for individual Hc1 were constructed using less than 50mg of flash-frozen tissue per prep (Adams *et al.* 2020). Six libraries were made with DpnII, and two were made with MluCI. Four of these libraries were sequenced to a depth of 616.4 million read pairs, with each replicate having at least 95.9 million read pairs. In addition, we prepared one DpnII Hi-C library with tentacle tissue from individual Hc3, sequenced to a depth of 233.9 million read pairs.

Total RNA was isolated from *H. californensis* individual Hc1 by pulverizing 100 mg of frozen tissue under liquid nitrogen, then proceeding with a Trizol RNA isolation protocol (Rio *et al.*

2010). The RNA was assayed at the UC Davis (UCD) DNA Technologies Core. One Illumina TruSeq RNA Library Prep Kit v2 library was constructed from this RNA at UCD. This library was sequenced to a depth of 95 million read pairs. The UCD DNA Technologies Core also prepared an Iso-Seq library and sequenced this library on a single Sequel II SMRT cell.

Lastly, *H. californensis* shotgun libraries were prepared from Hc1 and Hc2 by isolating DNA using the Omega Biotek EZNA Mollusc DNA kit, shearing the DNA using a Bioruptor, and preparing libraries with insert sizes of 400-500bp using the NEB Next Ultra II WGS, NEB Next Ultra II FS, or Illumina TruSeq Nano DNA library prep kits. Hc1 libraries were sequenced to a depth of 120 million read pairs. The Hc2 library was sequenced to a depth of 64 million 100PE reads on a HiSeq 2500 at the University of Utah DNA Sequencing Core Facility.

Trimming raw sequencing data

All Illumina libraries were trimmed with Trimmomatic v0.35 (Bolger *et al.* 2014) using the options ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:1:TRUE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. All Hi-C and Chicago libraries were additionally trimmed by removing the 3' end of reads after the restriction enzyme's junction sequence.

The PacBio Iso-Seq data were converted to circular consensus sequences using ccs v4.0 (github.com/PacificBiosciences/ccs). The Iso-Seq data then had the 5' and 3' cDNA primers removed using lima v1.10.0 (github.com/PacificBiosciences/barcoding), then polyA tails and chimeric sequences were removed with isoseq3 v3.2 (github.com/PacificBiosciences/IsoSeq). We used pauvre marginplot commit 13uhtt7 (github.com/conchoecia/pauvre) to check the overall consensus quality and length of transcripts (Figure S3) (De Coster *et al.* 2018).

Mitochondrial Genome Assembly and Annotation

Mitochondrial genome assembly

To assemble the *H. californensis* mitochondrial genome, we first mapped the PacBio Sequel CCS reads to the corrected *P. bachei* mitochondrial genome (Kohn *et al.* 2012; Arafat *et al.* 2018) using minimap2 v2.17 (Li 2017) with parameters `-ax asm20`. Reads that mapped to the *P. bachei* mitochondrial genome were assembled using canu v2.1.1 (Koren *et al.* 2017) with the options `genomeSize=15kb -pacbio-corrected`. We used Geneious v11 to identify the largest ORF, and used blastn v2.6.0 (Altschul *et al.* 1997) to identify that the ORF encoded COX1. We selected the start codon of the COX1 gene to be the 5'-most position of the mitochondrial genome, as is conventional with previous ctenophore mitochondrial genome annotations (Kohn *et al.* 2012; Pett and Lavrov 2015; Arafat *et al.* 2018). The sequence was trimmed up to the start codon of the COX1 gene on the canu contig. To confirm that the sequence was circular, we mapped the CCS reads to two concatenated copies of the mitochondrial genome using minimap2.

The mitochondrial genome assembly for individual Hc2 was generated by mapping the trimmed Hc2 Illumina WGS reads to the Hc1 mitochondrial assembly using BWA-MEM (Li 2013), then

correcting the reference using pilon (Walker *et al.* 2014). We mapped the reads back to the pilon-corrected reference to verify that it was correct.

The final 12564 bp Hc1 mitochondrial genome assembly was annotated by mapping the rRNA and CDS sequences from the corrected *P. bachei* mitochondrial genome (Arafat *et al.* 2018) to the assembly using Geneious v11. Geneious was then used to predict ORFs using the Mold Protozoan Mitochondrial translation table. ORF start sites that were conserved between Hc1 and Hc2 were used to delimit the beginning of the transcripts.

To annotate the ribosomal RNA boundaries we mapped the untrimmed RNA-seq reads to the final assembly with BWA-MEM (Li 2013). The start and stop sites for each ribosomal RNA were selected by finding positions that had several reads with the same start/stop site followed by a fast attenuation in coverage, also guided by the length of the *P. bachei* ribosomal RNA sequences. I-TASSER was used to predict the protein structure and to find the best structural analogs for the conserved URFs present in the genomes (Yang *et al.* 2015). We used the TMHMM tool to predict transmembrane domains for the URFs (Krogh *et al.* 2001). We used tRNAscanSE and ARWEN to search for mitochondrially-encoded tRNAs (Lowe and Eddy 1997; Laslett and Canback 2008).

Phylogeny construction

Full-length ctenophore 18S sequences were downloaded from NCBI, aligned using MUSCLE, then trimmed such that each sequence had greater than 90% occupancy. This alignment was used in a rapid bootstrapping maximum likelihood (ML) search of 250 trees with the GTR GAMMA model using RAxML v7.2.8 (Stamatakis 2014). A tree for COX1 nucleotide sequences was constructed in the same fashion. The mitochondrial nucleotide alignment was constructed by individually performing translation alignments on the COX1, COX2, COX3, CYTB, ND1, ND2, ND4, and ND5 loci from multiple species using MAFFT v7.388 (Kato *et al.* 2002). The alignments were concatenated, and a RAxML ML tree was constructed using the parameters described above. A Bayesian tree was constructed with the same concatenated protein alignment using MrBayes v3.2.6 (Ronquist and Huelsenbeck 2003), with *Tethya actinia* as an outgroup, the HKY85 substitution model, gamma rate variation, chain length of 30000, 4 heated chains, 0.2 heated chain temp, subsampling frequency every 200 trees, a 2500-tree burn-in, and a random seed of 1910.

Genome assembly

The wtdbg2 assembler v2.4 (Ruan and Li 2019) with parameters `-g 85m -p 0 -k 15 -e 3 -A -S 2 -s 0.05 -L 5000 -R --aln-dovetail 10240` was used to *de novo* assemble the PacBio CLR subreads. The assembly was polished with arrow v2.2 (github.com/PacificBiosciences/gcpp), then with pilon v1.22 (Walker *et al.* 2014) using the Illumina WGS libraries. Haplotigs were removed with Purge Haplotigs v1.0.4 (Roach *et al.* 2018) using parameters `purge_haplotigs cov -l 50 -m 175 -h 600 -j 70 -s 80` and `purge_haplotigs purge -a 30`. We then ran `purge_haplotigs clip` to remove overlapping contig ends.

Dovetail Genomics HiRise (v Aug 2019) was used to scaffold the genome first using the Chicago libraries, then using the Hi-C libraries (Putnam *et al.* 2016). We mapped shotgun reads to the contig assembly with BWA-MEM v0.7.17 (Li 2013) and calculated the mean coverage and GC content using BlobTools v1.1.1 (Laetsch and Blaxter 2017). Scaffolds with a mean coverage of less than 100, or having greater than 50% GC, were removed from the assembly. The resulting assembly was gapfilled using LR Gapcloser with the PacBio subreads (commit 156381a) (Xu *et al.* 2019). The assembly was then polished with pilon using the Illumina WGS libraries.

Hi-C heatmap generation

We generated a Hi-C heatmap to check for genome misassemblies. The Hi-C reads were mapped to the genome assembly using BWA-MEM with options `-5SPM` (Li 2013), the BAM was converted to a sorted and deduplicated pair file with pairtools v0.3.0 (github.com/mirnylab/pairtools), the pairs file was indexed with pairix v0.3.7 (github.com/4dn-dcic/pairix), then the pairs file was converted to a normalized mcool file using Cooler v0.8.10 (Abdennur and Mirny 2020). Additionally, we generated a PretextMap Hi-C matrix (github.com/wtsi-hpag/PretextMap commit ee1bf66). To visualize the matrices we used HiGlass v1.10.0 (Kerpedjiev *et al.* 2018) or PretextView v0.1.0 (github.com/wtsi-hpag/PretextView).

Variant Calling

To call variants to be used in phasing and in other analyses, we first mapped the PacBio CLR WGS reads to the genome using minimap2 v2.17 (Li 2017), and mapped the Hc1 Illumina WGS reads to the genome using BWA-MEM and samtools (Li *et al.* 2009; Li 2013). We then called variants using these two BAM files as inputs to the software freebayes and gnu parallel (Tange and Others 2011; Garrison and Marth 2012). We filtered the VCF file to only include diploid calls.

To phase the variants we then marked duplicates in the Hi-C BAM file using Picard v2.25.1 (“Picard Toolkit” 2016), then used HapCUT2 v1.3.1 (Edge *et al.* 2017) extractHairs on the Hi-C, Chicago, and PacBio CCS BAMs. For the PacBio subreads we used the extractHairs parameters `--pacbio 1 --new_format 1 --indels 1`. For the Hi-C reads we used the HapCUT2 extractHairs parameters `--hic 1 --new_format 1 --indels 1`. For the Chicago reads we used the HapCUT2 extractHairs parameters `--maxIS 10000000 --new_format 1 --indels 1`. We then concatenated these fragment files and used them as input to phase the genome using HapCUT2 with the parameters `--hic 1 --outvcf 1` (Edge *et al.* 2017).

Genome annotation

The genome annotation is composed of manually-selected transcripts from several software packages, including BRAKER, GeneMark-ES/ET, AUGUSTUS, Stringtie, pinfish, and the cDNA cupcake pipeline. Blast results to the *Mnemiopsis leidyi* v2.2 proteins or the SwissProt database (Skinner *et al.* 2009; Robinson *et al.* 2011) were also used as additional sources of evidence. To generate the individual annotations, we performed the following:

BRAKER, GeneMark-ES/ET and AUGUSTUS: Illumina RNA-seq reads were aligned to the genome assembly using STAR v2.7.1a (Dobin *et al.* 2013), and the Trinity transcriptome and PacBio Iso-Seq reads were aligned to the assembly using minimap2 with option `-x splice:hq`. AUGUSTUS and GeneMark-ES/ET annotations were generated by running BRAKER v2.14 with the Illumina RNA-seq, PacBio Iso-Seq, and Trinity transcriptome BAM files as inputs (Stanke *et al.* 2004; Lomsadze *et al.* 2014; Hoff *et al.* 2019).

Cupcake: We mapped the full length, non-chimeric (FLNC) PacBio Iso-Seq reads mentioned above in “Sequencing read preparation” to the *H. californensis* genome using minimap2 with the parameters `-ax splice -uf --secondary=no -C5`. We then used the PacBio Cupcake tools to collapse the FLNC reads into transcript models (github.com/Magdoll/cDNA_Cupcake). We generated one set of transcripts containing singletons, and one dataset without singletons, using the command `filter_away_subset.py --fuzzy_junction 5`.

Stringtie: Transcripts were predicted from the BAM file output of the minimap2 FLNC PacBio Iso-Seq-to-genome alignment using StringTie v2.0.4 (Pertea *et al.* 2015). Long parameters were used (`-L`) and the minimum isoform fraction was set to 0.1 (`-f 0.1`), with otherwise default parameters.

Pinfish: Transcripts were also predicted from the long reads using pinfish (github.com/nanoporetech/pinfish), with minimum isoform percentage set to 20, a minimum cluster size of 2 reads (`-p 20 -c 2`) and otherwise default parameters.

Manual inspection of each of the four annotations revealed many genes were erroneously fused or broken, compared to the true isoforms evident from the Iso-Seq data mapped to the reference. Because we found that each of the four annotations described above were imperfect, we chose to manually curate the annotation of the *H. californensis* genome. To ensure that the quality of the manual annotation was consistent across all 110 Mb, we developed a set of rules for difficult-to-annotate genes, like nested genes, gene clusters that appeared to have a trans-spliced leader exon, and how to combine multiple annotations into a single gene. These guidelines are available for download (github.com/conchoecia/hormiphora and Zenodo DOI: 10.5281/zenodo.4074309).

Transcript phasing

We first generated a transcript sequence for each isoform in the genome annotation with gffread (github.com/gpertea/gffread), then non-splice aligned the Illumina RNA-seq and PacBio Iso-Seq reads to the transcripts with BWA-MEM and minimap2 (Li 2013, 2017). We then used freebayes to call variants for each isoform (github.com/ekg/freebayes) (Tange and Others 2011), then phased each isoform with WhatsHap (Patterson *et al.* 2015). A new reference sequence for each haplotype was generated using bcftools consensus (Li 2011), then haplotype-specific Iso-Seq reads were used to correct the new haplotype-specific isoform using pilon v1.22 (Walker *et al.* 2014). These isoforms were then mapped to the reference genome using minimap2 `-ax splice`, phased with WhatsHap, then matched with the whole-genome phase variant phase blocks.

The longest ORFs from the phased and polished transcript isoforms were predicted using prottrans.py using the parameters `-a 50 -r` (bitbucket.org/wrf/sequences/src/master/prottrans.py).

To generate a non-redundant model set of proteins for convenience for analyses, we randomly selected one of the amino acid sequences from one of the haplotypes for each gene isoform. When the amino acid sequence from one haplotype was longer than the amino acid sequence on the other haplotype, we selected the longer one.

P. bachei genome reannotation

As no structural annotation, specifically no GFF file, was provided with the *P. bachei* genome, we created an exon-by-exon annotation file in GFF format from the reported scaffolds and transcripts for use in our whole-genome comparisons with *H. californensis*. The transcripts were mapped to the scaffolds with minimap2, using the options `-x splice --secondary=no`. Based on the mapping positions of each transcript in the BAM file, a GFF file was generated using pinfish (github.com/nanoporetech/pinfish) with the option `-g`. Of the 18950 transcripts, 18947 mapped back to the genome. For many protein comparisons, the proteins and transcripts provided with the *P. bachei* genome were insufficient due to the fragmented nature of the source scaffolds.

Next, we generated gene models using the AUGUSTUS web server (<http://bioinf.uni-greifswald.de/webaugustus/index>) (Hoff and Stanke, 2013) using the transcript models as the training set. This yielded two versions, the “hints” set and the “ab initio” set. As the “hints” version closely matched the transcript models, and likewise any gene fusions or breaks of that dataset, we instead used the “ab initio” set for downstream analyses. Lastly, due to the relatedness between *H. californensis* and *P. bachei*, we examined whether we could simply map the *H. californensis* model transcripts to the *P. bachei* scaffolds using minimap2, with the options `-x splice --secondary=no`. With this strategy, 99% of *H. californensis* transcripts mapped to *P. bachei*. 8000 of the transcripts had an additional mapping, likely due to fragmentation across different scaffolds or matching to both of a pair of uncollapsed haplotigs. We then used pinfish to generate a GTF file, as used above for the transcript model set.

Assessing fragmentation and fusion of genes

Using the *H. californensis* protein set, we used a custom Python script (`compare_hcal_ref_proteins.py`) to examine fragmentation of the *M. leidy* protein set. The script uses the coordinates of local alignments generated by diamond (Buchfink *et al.* 2015) to check whether a protein in *H. californensis* contains multiple non-overlapping alignments to *M. leidy* proteins on the same scaffold. Although this could mutually imply an erroneous fusion of two genes in *H. californensis*, the use of Isoseq reads for annotation makes this scenario unlikely. Nonetheless, out of around 1200 *M. leidy* proteins that were identified as fragmented, we then manually checked a set of 384 genes (all those with 3 or more fragments, as well as others) and found that all of them were indeed fragmented. Most of these had correct isoforms from *de novo* transcriptome assemblies.

Supplementary Text - Extended Results

Figure S2

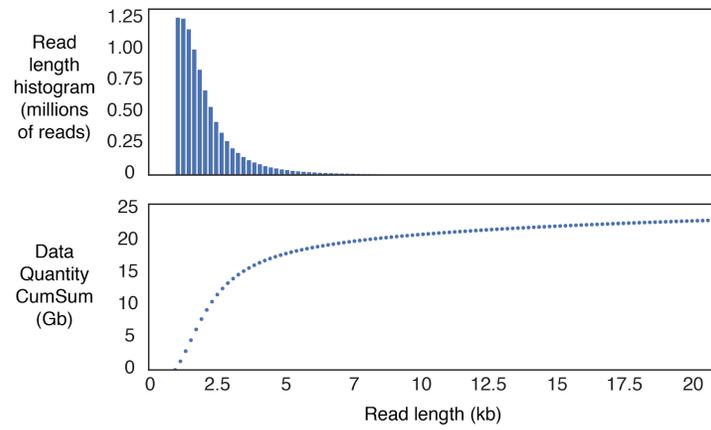


Figure S2. PacBio subread size distribution. Read length distribution (top) and cumulative sum of total basepairs (bottom) of the PacBio Sequel and Sequel II subreads.

Figure S3

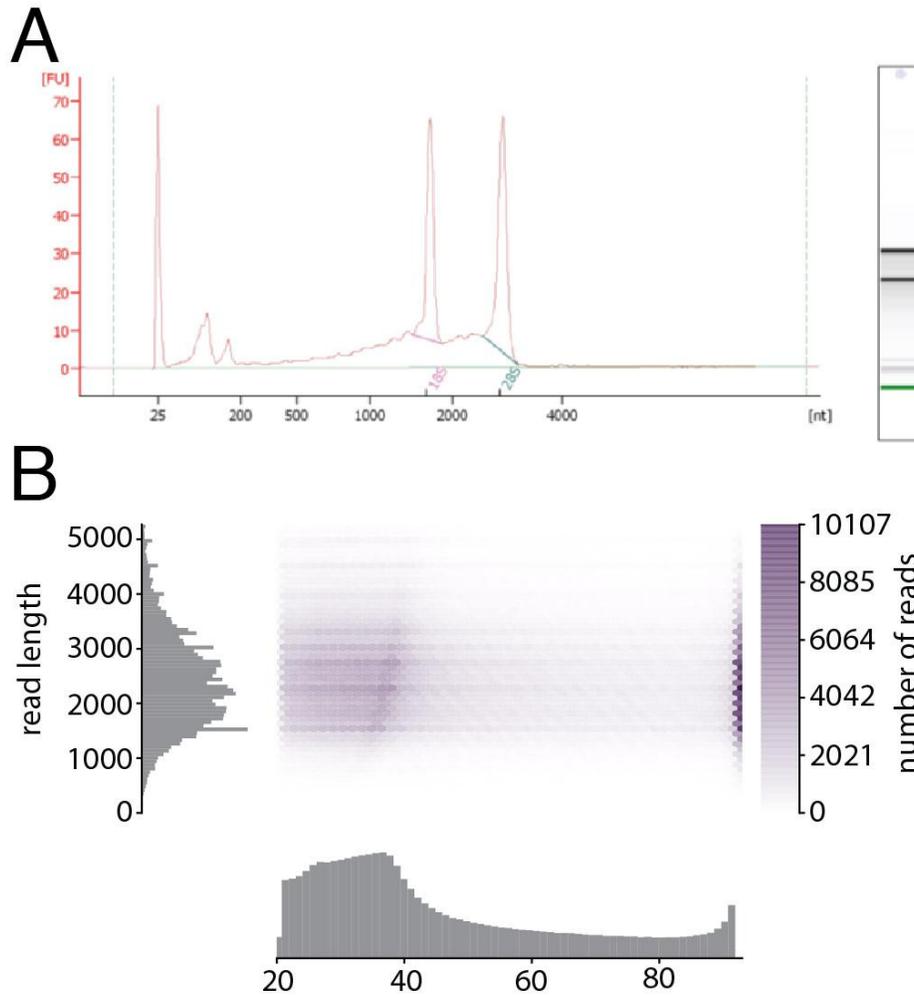


Figure S3. RNA and IsoSeq size distribution. (A) The Agilent Bioanalyzer trace of the RNA used to create the PacBio Iso-Seq library Hc1_lib18_run1_PB_Iso-Seq (SRR10403581 and SRR10403849). The RNA used for the library was largely intact. (B) A heatmap of the Iso-Seq reads after consensus calling with the ccs software and filtering to retain full-length, non-chimeric sequences. The read length histogram roughly resembles the RNA size distribution in Panel A.

Figure S4

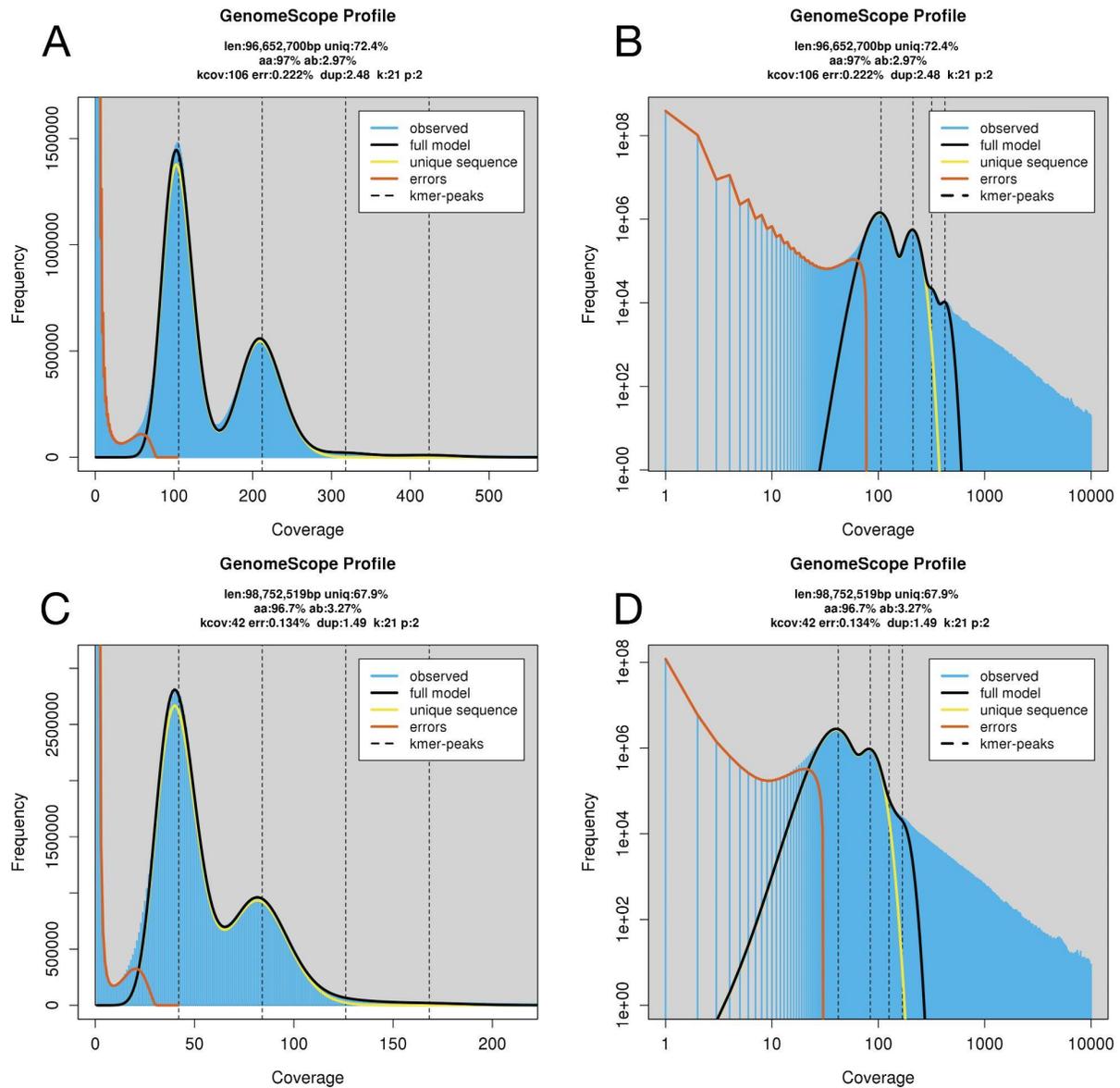


Figure S4. *H. californensis* k-mer based genome size prediction. (A,C) The haploid genome size estimate from the GenomeScope2 for Hc1 (A) was 96.6 Mb, and for Hc2 (C) was 98.72 Mb. Altogether, the Illumina WGS libraries from Hc1 had 212x genome coverage, and the Hc2 Illumina WGS library had 82x genome coverage. The *H. californensis* genome appears to be diploid from the k-mer spectrum based on the presence of two peaks in both A and C. Panels B and D are log-log plots of panels A and C.

Figure S5

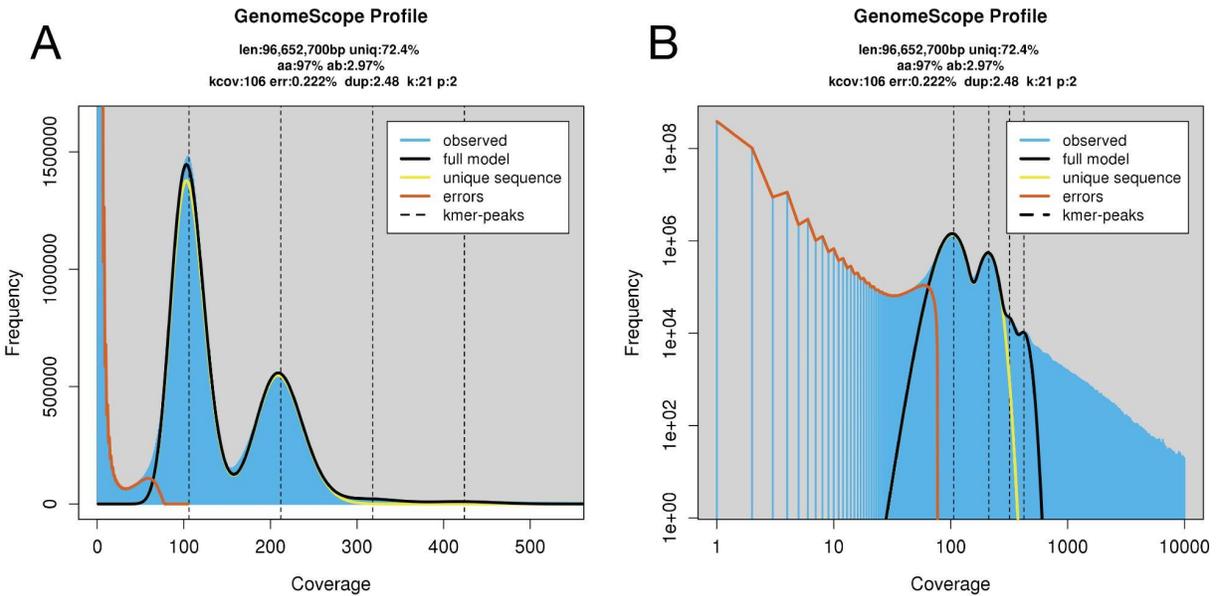


Figure S5. *P. bachei* k-mer based genome size prediction. We predicted the *P. bachei* genome size using publicly-available single-individual WGS data from SRA, the jellyfish k-mer counter, and GenomeScope2. (A) The predicted haploid size was 97.57 Mb. This predicted size is very close to the predicted size of the *H. californensis* genome (96-98 Mb). Altogether, the shotgun libraries had approximately 250x coverage of the genome. Similar to the *H. californensis* k-mer spectrum, this plot suggests that the animal is diploid. (B) is the log-log version of panel A.

Table S1

Assembly Step	% (Ill./ PB) WGS reads mapping	Number of Contigs	Number of Scaffolds	Assembly Size	contig N50	scaffold N50	BUSCO stats				
							(C)	(S)	(D)	(F)	(M)
wtdbg2	(97.85 / 94.08)	1769	1769	113.14 Mb	144 kb	143 kb	58.8	58.1	0.7	18.8	22.4
arrow + pilon	(97.85 / 95.14)	1769	1769	113.15 Mb	144 kb	143 kb	88.8	86.8	2	5.3	5.9
purge_haplotigs	(98.03 / 94.57)	1309	1309	106.89 Mb	152 kb	152 kb	87.5	85.8	1.7	5.6	6.9
blobtools	(/ 94.06)	1283	1283	106.44 Mb	153 kb	152 kb	87.5	85.8	1.7	5.6	6.9
HiRise Chicago	(/)	1334	287	106.55 Mb	150 kb	822 kb	88.4	87.1	1.3	4.6	7
HiRise Hi-C	(/)	1340	44	106.57 Mb	150 kb	8.14 Mb	87.8	86.1	1.7	5.3	6.9
PBjelly	(/)	975	44	110.67 Mb	204 kb	8.54 Mb	88.4	87.1	1.3	5.3	6.3
LRGC	(/ 95.16)	355	44	110.67 Mb	581 kb	8.54 Mb	88.5	86.8	1.7	5.3	6.2
pilon	(98.24 / 95.32)	355	44	110.66 Mb	580 kb	8.54 Mb	89.4	88.1	1.3	4.6	6

Table S1. Statistics through the *H. californensis* genome assembly stages. Each row of this table shows various statistics after each step of the assembly. The percent of Illumina and PacBio WGS reads that map to the genome, the contig N50, the scaffold N50, and the BUSCO nucleotide mode completeness scores increase with subsequent assembly steps.

Table S2

Annotation Step	Number of non-Protein Coding Genes	Protein-coding genes				Total Number of Genes
		Number of Protein-Coding Genes	Number of proteins with hits >1e-5 to nr	Number of proteins without hits >1e-5 to nr	Number of Proteins that do not appear in Mnemiopsis or Pleurobrachia genomes	
1. Genes added from Iso-Seq	248	12,987	8,420	4,567	619	13,235
2. Genes added from AUGUSTUS	38	1,170	585	585	95	1,208
3. Genes added from Pleurobrachia transcripts	23	108	20	88	10	131
Totals	309	14,265	8,945	5,320	714	14,574

Table S2. Genome Annotation Steps. This table includes the total number of genes added at each annotation step. There were 13,236 genes that had Iso-Seq read support. There were 12,987 that were protein-coding and 249 that were not protein-coding. For each step we also included the number of protein coding genes that had significant hits to nr, and the number of protein-coding genes that did not appear in the *Mnemiopsis* or *Pleurobrachia* genomes' proteins, but appeared in the transcriptomes of other ctenophores. The total number of protein-coding genes that we identified was 14,265. The total number of genes that we identified, including non-protein-coding genes, was 14,574.

Table S3

Dataset	Complete	Complete + Partial	Number of missing core genes	Average number of orthologs per core genes	% of detected core genes that have more than 1 ortholog	BUSCO string
Protein Models	281 (92.74%)	291 (96.04%)	12 (3.96%)	1.18	10.68%	C:92.7%[S:82.8%,D:9.9%],F:3.3%,M:4%
Genome	270 (89.11%)	286 (94.39%)	17 (5.61%)	1.01	0.74%	C:89.1%[S:88.4%,D:0.7%],F:5.3%,M:5.6%
IsoSeq FLNC	290 (95.71%)	296 (97.69%)	7 (2.31%)	101.57	96.55%	C:95.7%[S:3.3%,D:92.4%],F:2.0%,M:2.3%
Illumina <i>de novo</i> Transcrip-tome	299 (98.68%)	300 (99.01%)	3 (0.99%)	2.35	71.57%	C:98.7%[S:28.1%,D:70.6%],F:0.3%,M:1.0%

Table S3. BUSCO scores. These BUSCO protein mode scores were calculated using gVolante (Nishimura *et al.* 2017).

Figure S6

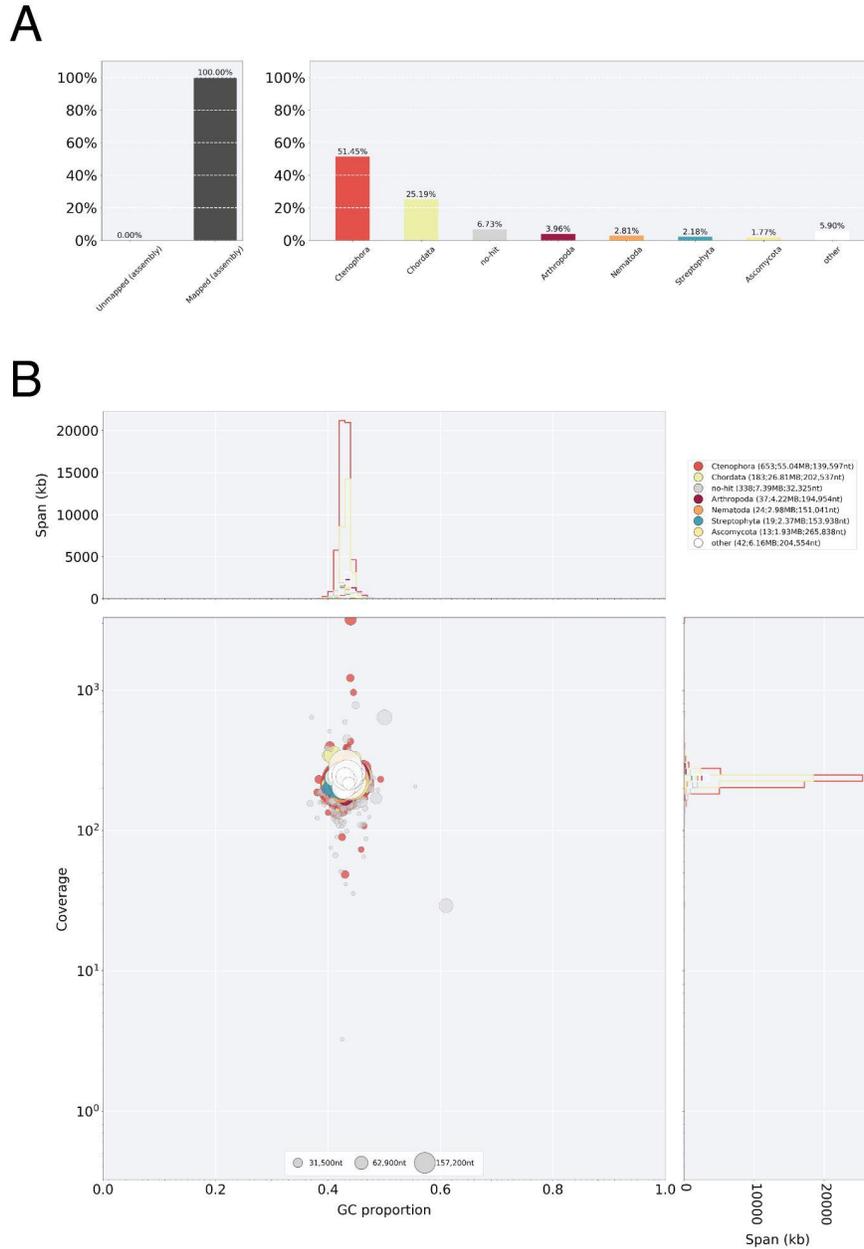


Figure S6. *H. californensis* assembly intermediate blobtools plot. (A) While the BlobTools taxonomic classification suggests that there is a large amount of contamination in the DNA sequencing libraries, the GC and read depth coverage plot (B) suggests otherwise. Most contigs had close to 42% GC and had a mean read depth coverage around the haploid k-mer coverage of 212.

Figure S7

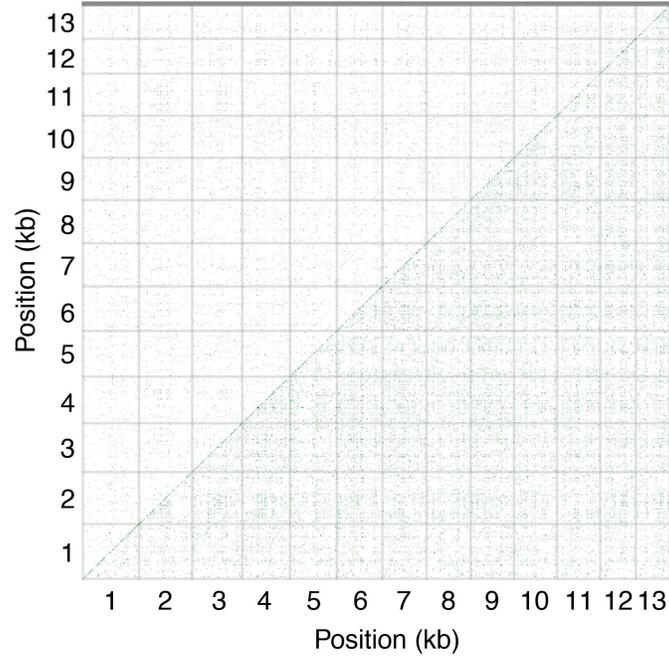


Figure S7. D-genies genome dotplot. A dot-plot of the entire genome aligned against itself, showing that very few regions are duplicated, and there are no large segmental duplications. Light-green lines indicate matches (below diagonal), purple lines indicate reverse-complement matches (above diagonal). Short lines appear as dots.

Figure S8

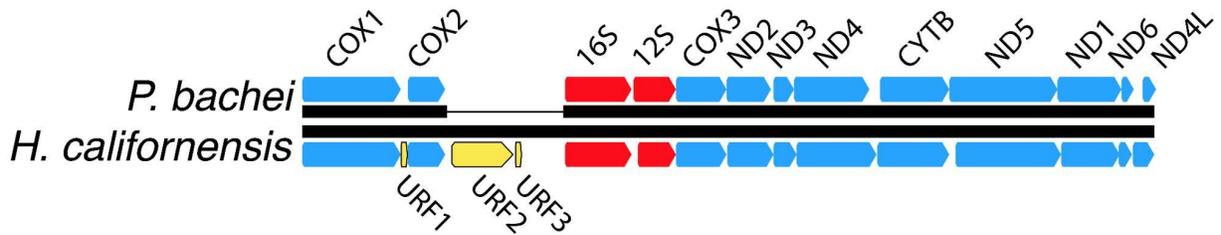


Figure S8. A synteny plot of *P. bachei* and *H. californensis* mtDNA. These two species share the same gene order, except that *H. californensis* has a large insertion between the COX2 gene and the 16S gene. The insertion in the *H. californensis* mtDNA contains two URFs. URF1, URF2, and URF3 occur in both Hc1 and Hc2.

Figure S9

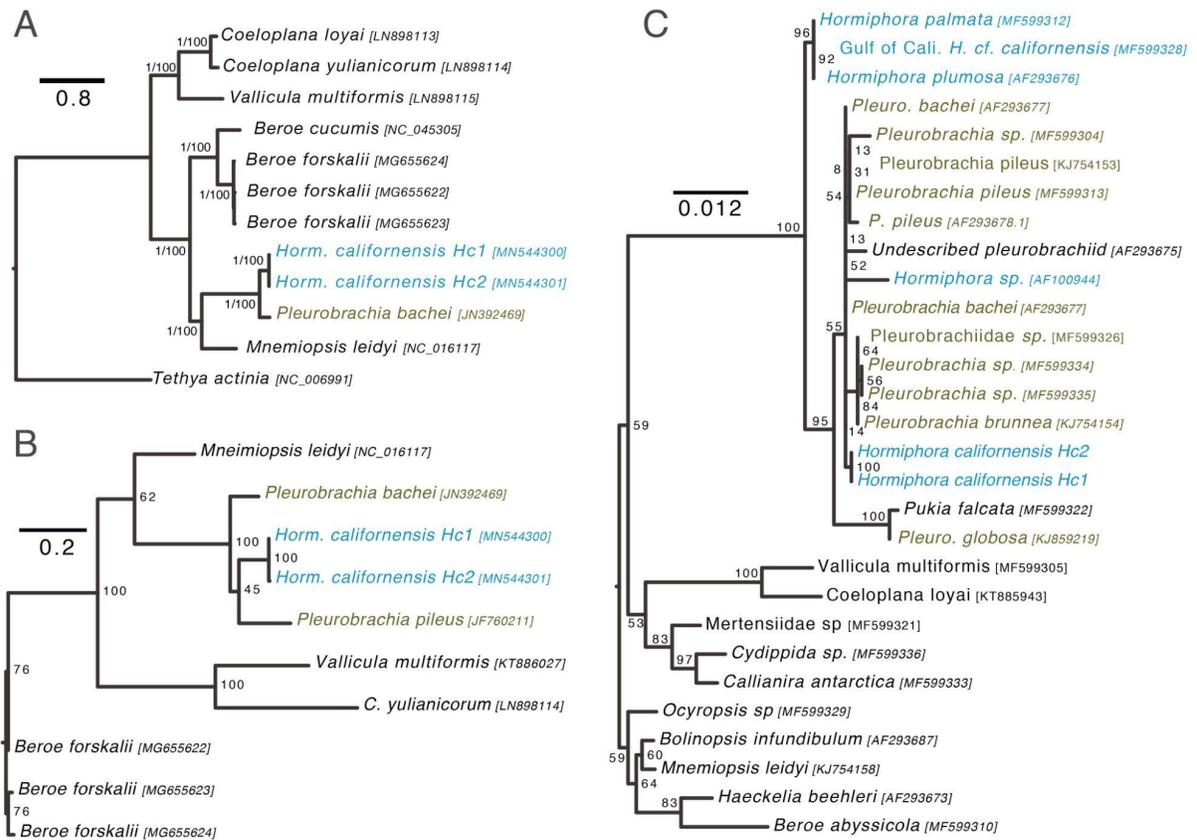


Figure S9. Phylogenetic position of Hc1 and Hc2. (A) Ctenophore mitochondrial protein tree, including the COX1, COX2, COX3, CYTB, ND1, ND2, ND4, and ND5 loci. Node labels are posterior probability from the Bayesian tree, and the bootstrap value from the maximum likelihood tree. All nodes had a posterior probability of 1 and a bootstrap value of 100. (B) A COX1 nucleotide tree using additional COX1 sequences from NCBI. Node labels are bootstrap values. Samples Hc1 and Hc2 are in a clade within the genus *Pleurobrachia*. (C) An 18S ctenophore tree. Node labels are bootstrap values. Samples Hc1 and Hc2 lie within a polytomy of other *Pleurobrachia* species, but are distinct from *H. palmata*, *H. plumosa*, and a *H. californensis* sample from the Gulf of California.

Figure S10

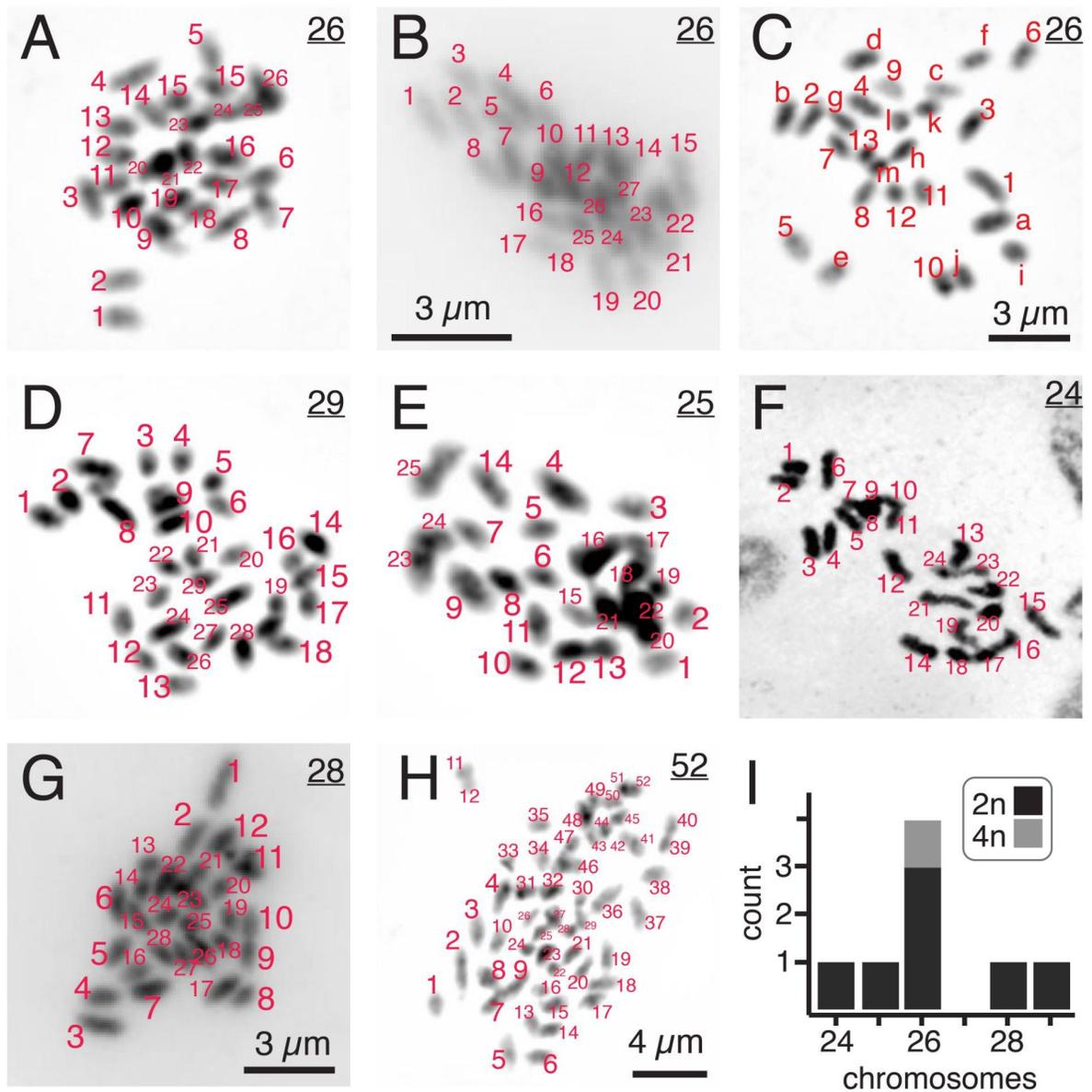


Figure S10 *Hormiphora californensis* karyotyping results. Panels (A-H) are the karyotyping results from individual embryos stained with DAPI, image color inverted and grayscale. Each chromosome is numbered 1-N. Numbers in black and underlined are the total number of chromosomes estimated in that panel. Panel (C) is numbered in pairs using the same scheme as Figure 1. (I) shows a histogram of the number of times that each chromosome count was observed. There is one 4n count included in the 26 chromosomes bin, as the count was 52 and likely corresponded to a 2n of 26. A 2n of 26 corresponds to 13 pairs of chromosomes.

Gene number and synteny with other ctenophores

For *M. leidy*, the ML2.2 annotation and protein set were compared to the *H. californensis* proteins with the script scaffold_synteny.py (Zenodo DOI: 10.5281/zenodo.4074309). We examined gene positions for 450 of the longest scaffolds in *M. leidy*, accounting for 110Mb, whereby the smallest scaffold examined was 114kb in length. Of the 8685 query proteins with matches to any *M. leidy* protein, 6422 gene matches were retained after filtering for quality and matches to the longest scaffolds.

We then examined the collinearity of genes between the two genomes, again based on unidirectional BLAST hits, requiring at least 3 genes in a row, allowing up to 5 intervening genes. This identified 571 blocks containing 2258 total genes, though 439 of these blocks contained either 3 or 4 genes, suggesting that collinearity is limited between the two species. As the script that identified these blocks allows for two tandem genes to hit the same query, false positives from fragmented genes may account for some of these. For instance, we found 279 cases where the gene in *H. californensis* spans two or more genes in the ML2.2 annotation.

The *P. bachei* genome size prediction using GenomeScope2 was 97.57 Mb (Figure S5) - only 62.5% of the size of the published assembly, 156.1 Mb (Moroz *et al.* 2014). The predicted *P. bachei* genome size of 97.57 Mb is very close to the predicted genome size of *H. californensis*. Based on the mean read mapping depth per-scaffold, it appeared that haplotypes were collapsed for 5310 scaffolds, but that over half of the *P. bachei* scaffolds were unmerged haplotypes. If the remaining 16669 scaffolds were collapsed into a haploid representation this would yield a final estimated genome size of 107Mb, close to the size of the *H. californensis* genome. This suggests that only one third of the *P. bachei* assembly represents a haploid assembly. This may also account for the additional ~7000 proteins predicted in the *P. bachei* genome compared to *H. californensis*.

Therefore, we used two approaches to estimate colinearity between *H. californensis* and *P. bachei*. First, we tried an analysis of only the 59Mb of scaffolds that had a mean coverage close to the haploid k-mer coverage of 250x. Of these scaffolds, the longest was only 221kb, therefore broad scale synteny could not be effectively analyzed. Of the original 18950 *P. bachei* transcripts, 7076 mapped to one of the 5310 haplotype-collapsed *P. bachei* scaffolds. We used this geneset for microsynteny analyses with *H. californensis*. In total, 299 putative collinear gene blocks of at least 3 genes were identified, accounting for 1280 genes. Overall, the high number of scaffolds in the *P. bachei* genome hampered our ability to detect microsynteny between *P. bachei* and *H. californensis*. Despite their relatedness, this was lower than the detectable synteny between the more phylogenetically distant *M. leidy* and *H. californensis*.

Next, we tried reanalyzing the *P. bachei* scaffolds using an ab initio annotation from AUGUSTUS. This program had predicted 32683 total proteins across the *P. bachei* assembly, though the density is much higher than the v1 transcripts. For example, on the longest Pbac scaffold of 320kb, there are 12 mapped transcripts but 31 AUGUSTUS genes are predicted. *Ab initio* gene predictions have difficulty resolving nested genes, which is disabled by default in AUGUSTUS, thus many of these predictions are likely to be fragments of larger genes that are split by nested genes. We analyzed microsynteny between *H. californensis* and the *P. bachei*

AUGUSTUS annotation, using *H. californensis* as the query. This had identified 983 blocks for 5025 genes, more than twice the count from the original transcript annotation. If the *P. bachei* AUGUSTUS predictions were instead used as the query, this identified 1648 blocks with 7803 genes, in many cases spanning the entire scaffold. Because multiple query genes were allowed to map to a single target gene, this increase of almost 3000 genes is likely due to the fragmented AUGUSTUS predictions. Nonetheless, it is evident that there is substantial synteny between *H. californensis* and *P. bachei*.

Comparison to ML2 assembly and annotation

The *M. leidy* ML2 annotation (Ryan *et al.* 2013) had 16545 proteins, almost 2000 more than the *H. californensis* v1 annotation from this study. We sought to explain the large difference in protein number using synteny and orthology information from blast searches.

We found 1200 neighboring ML2 proteins that were bridged by a single *H. californensis* protein, suggesting that either the *M. leidy* proteins are falsely split, or the *H. californensis* protein is a false fusion. The majority of these cases only had two neighboring *M. leidy* genes, though there were 8 cases of 4 or 5 neighboring *M. leidy* genes that were bridged by a single *H. californensis* protein. In all cases, these transcripts were supported with single Iso-Seq reads in *H. californensis*.

We manually corroborated these 8 cases by comparing the *M. leidy* genes to the *H. californensis* ortholog, matches in publicly-available transcriptomes of other ctenophores (Francis *et al.* 2015; Whelan *et al.* 2015, 2017), and orthologs in other animals. This analysis revealed that all 8 proteins appear to be fragmented in *M. leidy* and the *H. californensis* version appears to be complete. Generally these genes were large, and many included nested intronic genes. These included homologs of Midasin (4284AAs), Pecanex (2096AAs), Dynein heavy chain 14 (4735AAs), Piezo (2335AAs), a possible homolog of Centriolin (2141AAs), glycogen synthase (1214AAs), oxysterol binding protein (894AAs), and a putative homolog of SZT2 (3031AAs). Large genes such as dynein heavy chain required manual reannotation in *H. californensis* as well, as only 2/17 dynein genes were correctly annotated in the Iso-Seq-based Stringtie annotation.

Figure S11

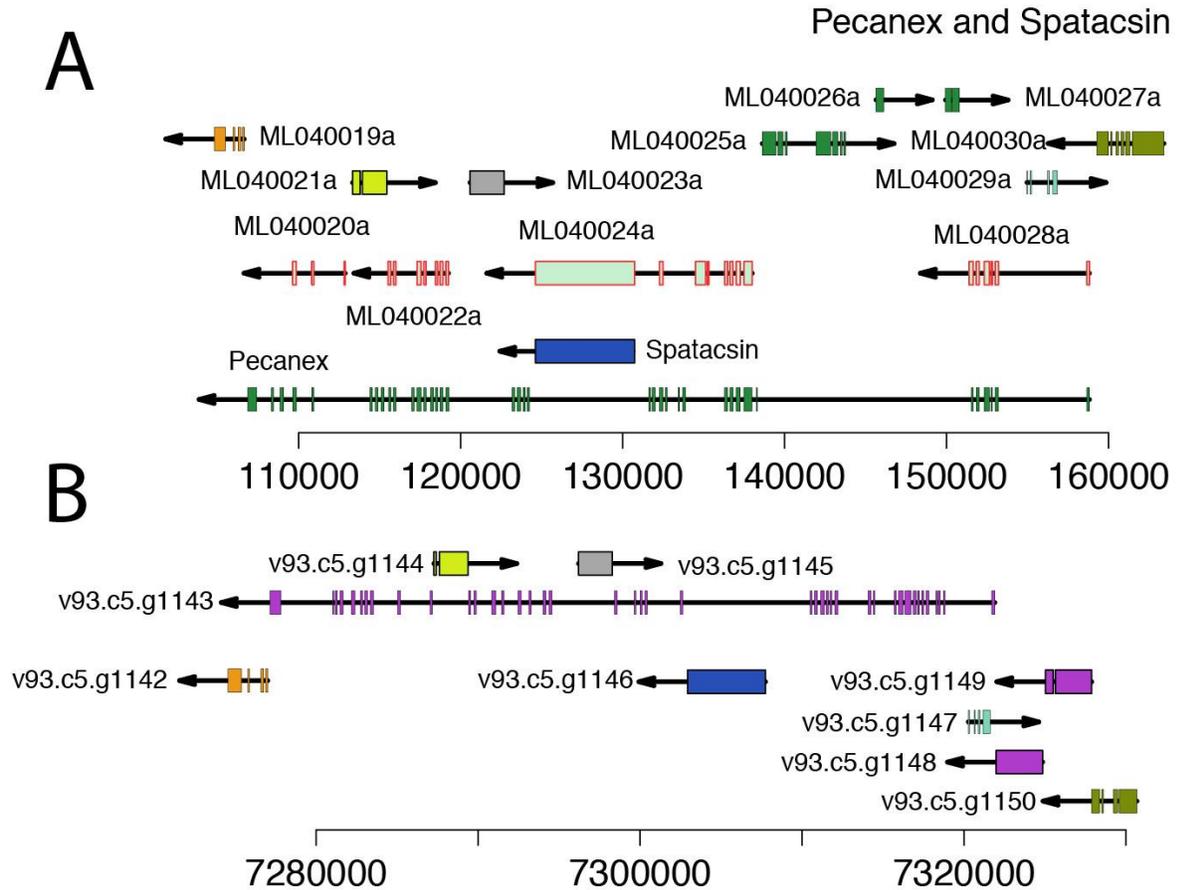


Figure S11. Pecanex and Spatacsin loci. Loci of the homolog of pecanex in *M. leidy* (A) and *H. californensis* (B). In *M. leidy*, the full-length gene joins 4 genes from the ML2 annotation, and contains 6 nested intronic genes, one of which was falsely fused. Four of these genes have homologs in *H. californensis* in the orthologous introns. The gene ML040024a fuses the single-exon homolog of spatacsin, though this is not supported by the transcripts or de novo assembly. Many of the surrounding or nested genes are homologous between the two species, as ML040019a, ML040021a, ML040023a, ML040029a, and ML040030a, match with *H. californensis* c5.g977, c5.g979, c5.g980, c5.g982, and c5.g984, respectively, and are colored pairwise.

Figure S12

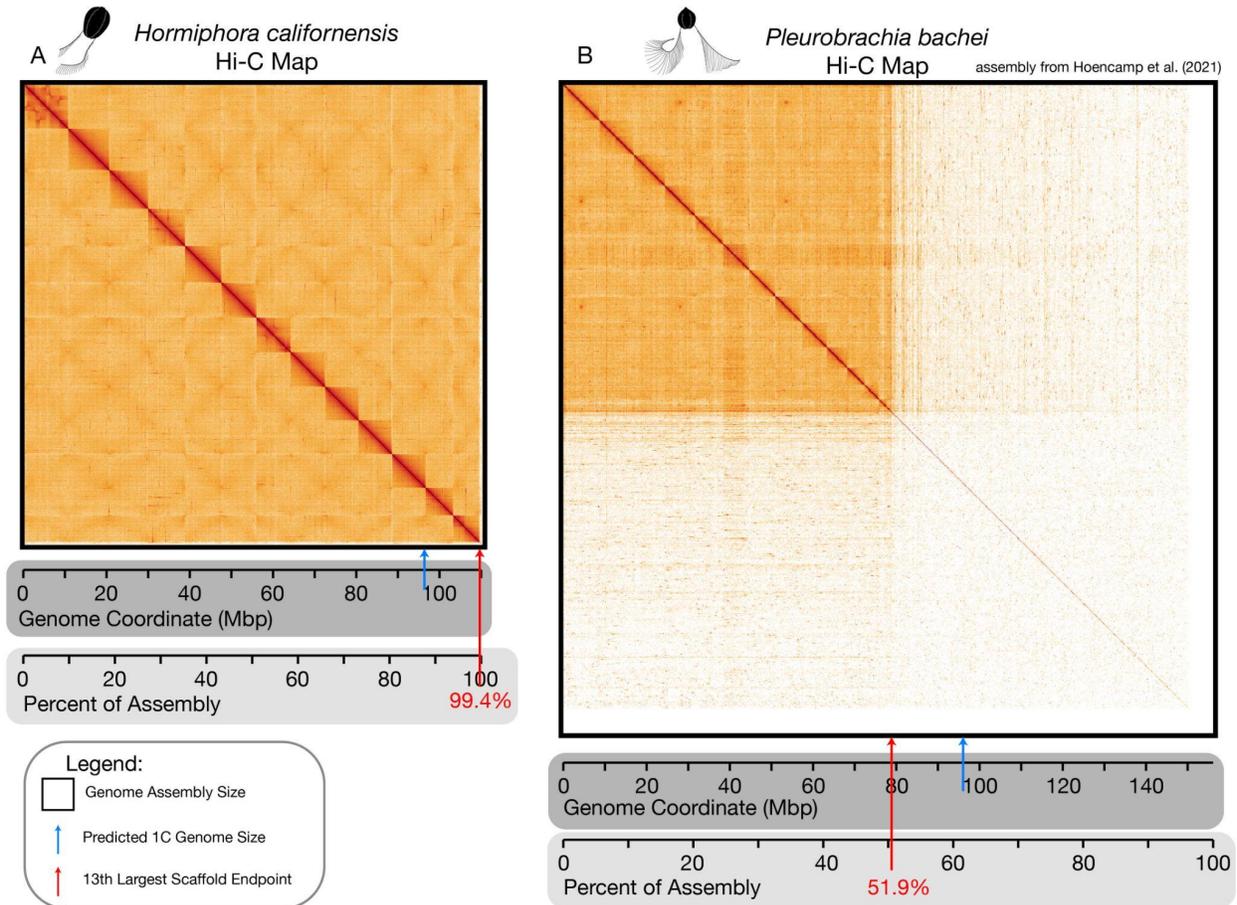


Figure S12. Hi-C map of *H. californensis* and *P. bachei*. These are the Hi-C maps of *H. californensis* and *P. bachei*, shown without individual lines separating scaffolds. The x-y scale of megabase pairs (Mbp) in both plots is the same. The genome assembly sizes are shown with a black bounding border. The predicted genome size for both species based on k-mer spectra, 96.6 Mbp, is shown with a blue arrow. The amount of the genome in the 13 largest scaffolds is shown with a red arrow, and the percent of the assembly in those 13 scaffolds is shown in red text. (A) The Hi-C map for *H. californensis*. The largest 13 scaffolds contain 99.4% of the total bases in the assembly. (B) The Hi-C map for *P. bachei* from Hoencamp et al (2014). The largest 13 scaffolds contain 51.9% of the assembly. 48.1% of the genome is not in chromosome-scale scaffolds, yet has Hi-C connections to the chromosome-scale scaffolds.

Figure S13

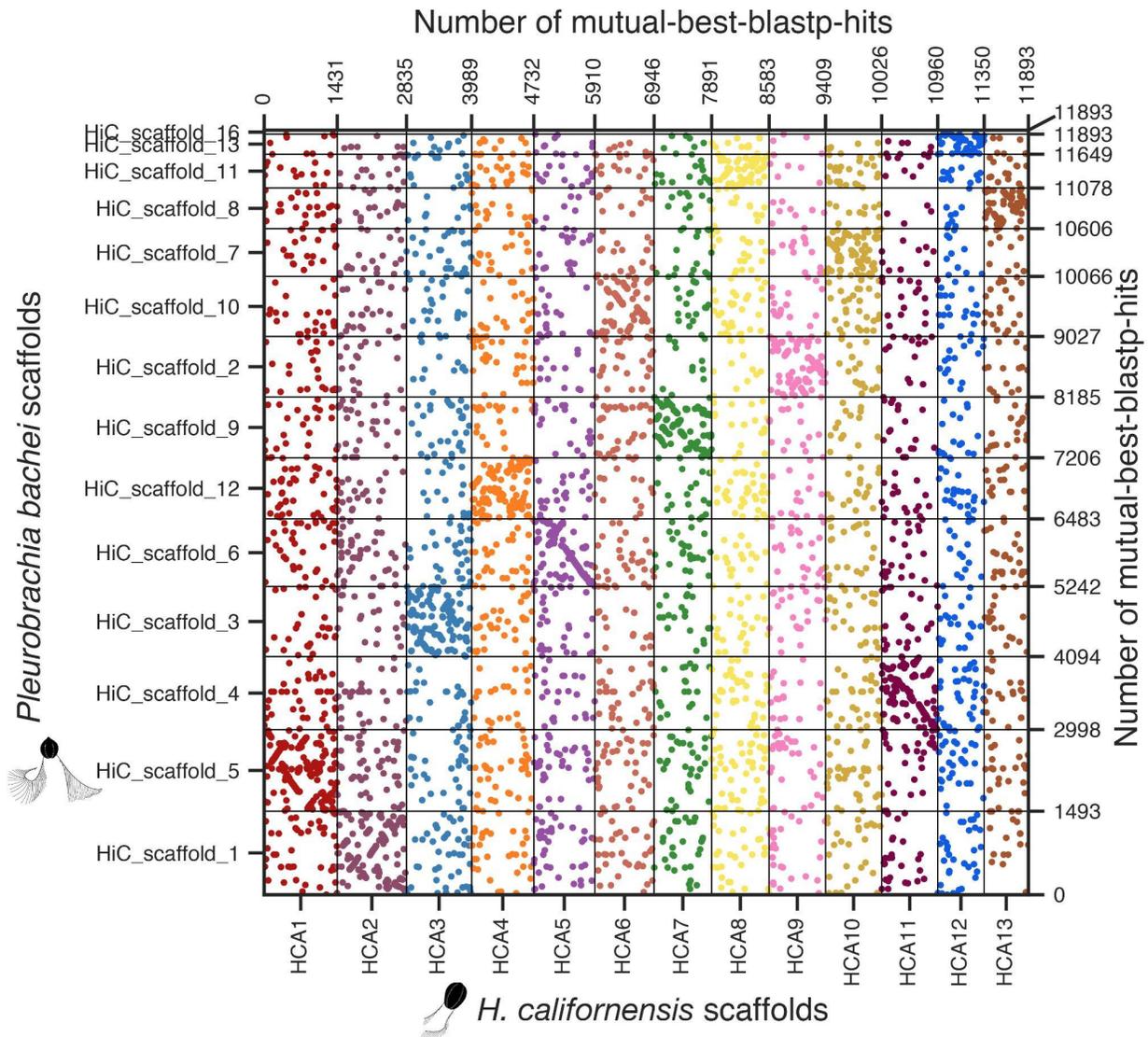


Figure S13. *Pleurobrachia-Hormiphora* Oxford dot plot. This plot shows the coordinates of mutual best blastp hits when comparing the proteins in the genomes of *P. bachei* to *H. californensis*, and *H. californensis* to *P. bachei*. Only the first 13 *H. californensis* scaffolds, and the largest 14 *P. bachei* scaffolds, are plotted. One dot is one putatively orthologous protein shared by the two species. The dots are colored by *Hormiphora* chromosome. This plot shows that each *H. californensis* chromosomal scaffold has a homologous chromosomal scaffold in *P. bachei*. For example, *H. californensis* 1 predominantly shares genes with *P. bachei* HiC_scaffold_5. Moreover, this plot shows that while shared chromosomes 5 and 11 have large regions with gene colinearity, most of the other homologous chromosomes are highly rearranged between *H. californensis* and *P. bachei*. Lastly, we see that only the 13 largest *P. bachei* scaffolds have enough information to assign them to homologous *H. californensis* scaffolds. The 14th-largest *P. bachei* scaffold has no proteins that had reciprocal best matches to the 13 chromosomal *H. californensis* scaffolds.

Figure S14

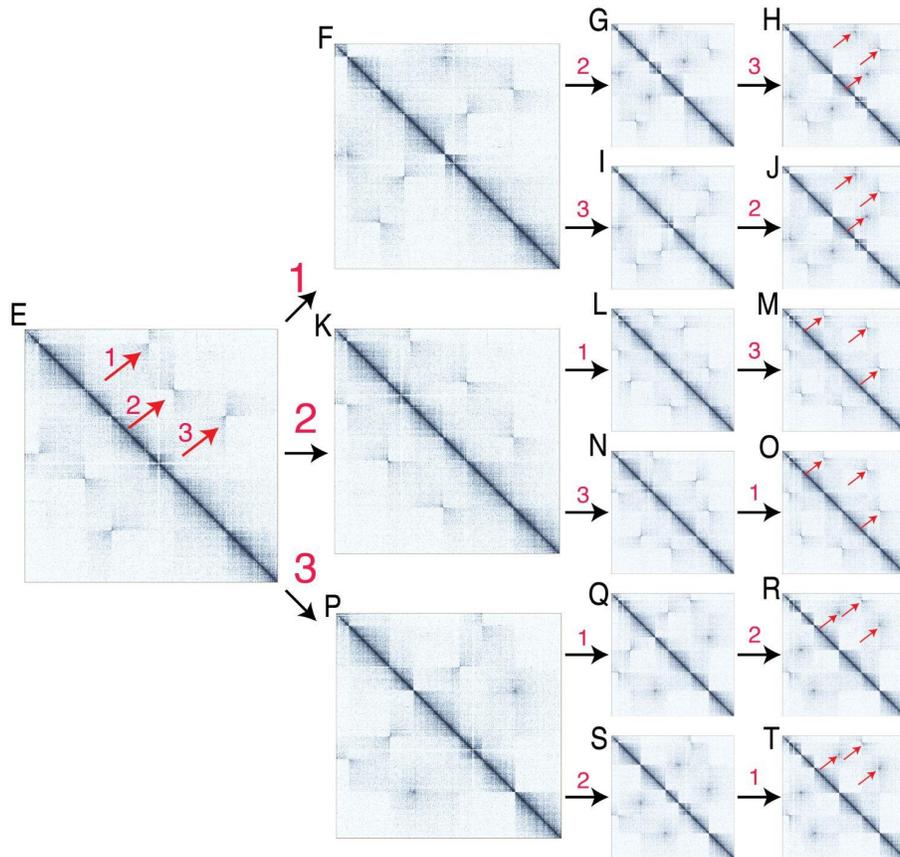
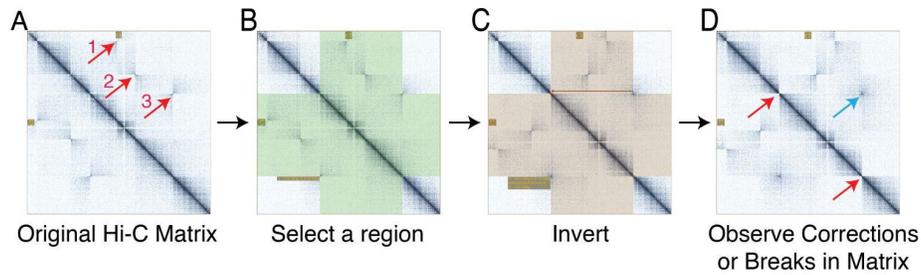


Figure S14. Scaffold 1 heterozygous inversion. Off-diagonal hotspots in Hi-C contact matrices (A, red arrows) indicate assembly errors, or heterozygous inversions. One method of determining if off-diagonal Hi-C hotspots are misassemblies was to manually invert the assembly at the suspect break points (B,C). The manipulation will result in removing the off-diagonal signal while preserving the diagonal signal, or preserving the off-diagonal signal while degrading the diagonal signal (D, red arrows). Panels (E-T) show all the possible combinations of rearrangements to attempt to correct the off-diagonal signal. Red numbers above black arrows indicate which off-diagonal signal was inverted. The right-most panels show that the

off-diagonal signals remain after manipulating the heatmaps, and the continuity of the diagonal signal is interrupted. Therefore, this signal is likely from heterozygous inversions.

Figure S15

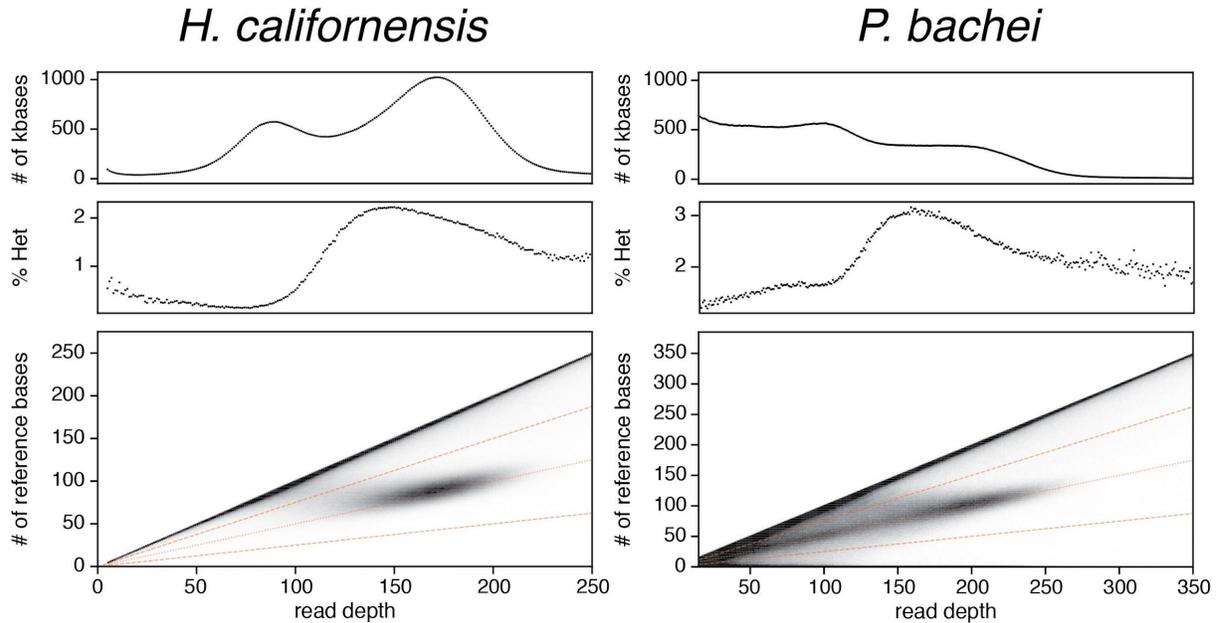


Figure S15. Plots pertaining to the heterozygosity of *H. californensis* and *P. bachei*. The bottom-most panels show a heat map of the number of positions in the genome that have X-number of reads with the reference allele when the total read depth at that position is Y. A smear at 1x sequencing depth coverage (x-axis) with only 50% of bases matching the reference allele, shows that the animals are diploid. The top-most panel is a histogram of the total number of positions in the genome (Y-axis) that have X number of reads at that position. This plot is useful to visualize the proportion of bases that are either located on uncollapsed haplotigs, or are indels present in the assembly. The middle panel shows the heterozygosity at each read depth. The most reliable window for calculating heterozygosity is at the mode of the mapping depth where reads from both haplotypes map to the reference. This point is approximately 160x read depth for *H. californensis* and 205x read depth for *P. bachei*. The top panel of the *P. bachei* analysis shows that there are many positions in the genome that have reads mapped from only one haplotype, indicated by the peak around 102x read depth.

Table S4

Individual	Acc. Number	Species	Method	k-mer size	% SNV Het (min)	% SNV Het (max)
SAMN00216730	SAMN00216730	<i>P. bachei</i>	mpileup	NA	2.63%	NA
			angsd	NA	2.40%	NA
			vcftools	NA	0	NA
			GenomeScope2	21	4.20%	4.25%
			GenomeScope2	41	3.03%	3.08%
Hc1	SAMN12924379	<i>H. californensis</i>	mpileup	NA	2.00%	NA
			angsd	NA	1.65%	NA
			vcftools	NA	1.51%	NA
			GenomeScope2	21	2.95%	2.98%
			GenomeScope2	41	2.36%	2.39%
Hc2	SAMN12924380	<i>H. californensis</i>	angsd	NA	1.85%	NA
			vcftools	NA	1.56%	NA
			GenomeScope2	21	3.25%	3.28%
			GenomeScope2	41	2.55%	2.58%

Table S4. Estimated heterozygosity of *H. californensis* and *P. bachei*. We measured the heterozygosity of *P. bachei* SAMN00216730 and *H. californensis* Hc1 using the mpileup method (Saremi *et al.* 2019). In this table, the mpileup method only measures the single-nucleotide heterozygosity. In addition we measured the heterozygosity using angsd, vcftools, and GenomeScope2 (Danecek *et al.* 2011; Korneliussen *et al.* 2014; Ranallo-Benavidez *et al.* 2020). The k-mer size used and the window of heterozygosity values were reported for the GenomeScope method. Vcftools reported zero heterozygous sites for the *P. bachei* individual, which we attribute to a software error given the results of the mpileup and angsd analyses.

Table S5.

Species	Genome accession used	SRA accessions used
<i>T. wilhelma</i>	(Mills <i>et al.</i> 2018)	SRR2163223
<i>T. adhaerens</i>	GCF_000150275.1	SRX6204530 through SRX6204554
<i>N. nomurai</i>	GCA_003864495.1	SRR6298213
<i>D. melanogaster</i>	GCF_000001215.4	SRR10512945
<i>S. purpuratus</i>	GCF_000002235.5	SRR7211988
<i>H. sapiens</i>	GRCh38	(Zook <i>et al.</i> 2016)

Table S5. Genome samples used in heterozygosity measurements. These genome assemblies and SRAs were used as a comparison for genome heterozygosity measurements compared to *H. californensis*.