OXFORD

G3 Genes | Genomes | Genetics

G3, 2021, **11(9)**, jkab212
DOI: 10.1093/g3journal/jkab212
Advance Access Publication Date: 18 June 2021
Genome Report

Chromosome-level genome assembly of the blue crab, *Callinectes sapidus*

Tsvetan R. Bachvaroff,[1] Ryan C. McDonald,[1] Louis V. Plough,[2] and J. Sook Chung [1,*]

[1]Institute of Marine and Environmental Technology, University of Maryland Center for Environmental Science, Baltimore, MD 21202, USA
[2]Horn Point Laboratory, University of Maryland Center for Environmental Science, Horn Point, MD 21613, USA

*Corresponding author: Institute of Marine and Environmental Technology, University of Maryland Center for Environmental Science, 701 East Pratt Street, Baltimore, MD 21202, USA. Email: chung@umces.edu

Abstract

The blue crab, *Callinectes sapidus* (Rathbun, 1896) is an economically, culturally, and ecologically important species found across the temperate and tropical North and South American Atlantic coast. A reference genome will enable research for this high-value species. Initial assembly combined 200× coverage Illumina paired-end reads, a 60× 8 kb mate-paired library, and 50× PacBio data using the MaSuRCA assembler resulting in a 985 Mb assembly with a scaffold N50 of 153 kb. Dovetail Chicago and HiC sequencing with the 3d DNA assembler and Juicebox assembly tools were then used for chromosome scaffolding. The 50 largest scaffolds span 810 Mb are 1.5–37 Mb long and have a repeat content of 36%. The 190 Mb unplaced sequence is in 3921 sequences over 10 kb with a repeat content of 68%. The final assembly N50 is 18.9 Mb for scaffolds and 9317 bases for contigs. Of arthropod BUSCO, ~88% (888/1013) were complete and single copies. Using 309 million RNAseq read pairs from 12 different tissues and developmental stages, 25,249 protein-coding genes were predicted. Between *C. sapidus* and *Portunus trituberculatus* genomes, 41 of 50 large scaffolds had high nucleotide identity and protein-coding synteny, but 9 scaffolds in both assemblies were not clear matches. The protein-coding genes included 9423 one-to-one putative orthologs, of which 7165 were syntenic between the two crab species. Overall, the two crab genome assemblies show strong similarities at the nucleotide, protein, and chromosome level and verify the blue crab genome as an excellent reference for this important seafood species.

Keywords: genome assembly; chromosome; synteny; Brachyura; Portunidae; *Callinectes sapidus*

Introduction

The blue crab, *Callinectes sapidus* Rathbun (1896) is a well-studied resident, decapod crustacean species distributed across the Western Atlantic region. The population ranges along the western Atlantic coast from Argentina to Cape Cod in MA, United States (Williams 1974). In some areas, including the Chesapeake Bay, the blue crab is an important fishery species and 97,896 metric tons were harvested globally in 2016 (FAO Fisheries and Aquaculture 2019). In many coastal habitats, the blue crab also plays a key ecological role as a keystone predator on numerous species of smaller invertebrates and serves as prey for sea birds, turtles, and large fish species (Hines 2007; Lipcius *et al.* 2007; Hines *et al.* 2011). Additionally, the blue crab has expanded its territory as a successful invasive species to a number of coastal regions in the world, including in the Mediterranean Sea where the crab poses a serious problem due to its foraging ability and lack of natural predators (Nehring 2011; Katsanevakis *et al.* 2014; Mancinelli *et al.* 2017).

Despite a complicated life cycle involving four major life stages (larval, megalopal, juvenile, and adult) and 27–29 molts (Van Engel 1958), blue crabs have been successfully reared in a hatchery setting (Zohar *et al.* 2008), closing the life cycle in captivity (Figure 1). This enables the blue crab as a tractable, naive experimental animal model for studying growth, reproduction, sexual differentiation, and disease responses of decapod crustaceans (Zmora *et al.* 2009, 2009; Chung *et al.* 2011; Zmora and Chung 2014; Techa and Chung 2015; Huang *et al.* 2016). In captivity, application of reverse genetics such as RNA interference (RNAi) experiments at key life stages led to the discovery and functional description of a novel crustacean female sex hormone (Zmora and Chung 2014).

Another major finding from captivity studies is that a clutch derived from one mother shows huge variation in growth and survival, while females also vary in their reproductive performance (Bembe *et al.* 2017; Maurer *et al.* 2017; Bembe *et al.* 2018). This phenomenon is not isolated to *C. sapidus*, as similar results have been reported in oysters (Lannan 1980; Taris *et al.* 2007), shrimp (Sandifer and Smith 1979), and other decapods (Anger 1998). High levels of genetic variation among blue crab parents could contribute to such outcomes, which in turn demands further studies of the genetic architecture of important hatchery or aquaculture traits.

Received: March 12, 2021. **Accepted:** June 15, 2021

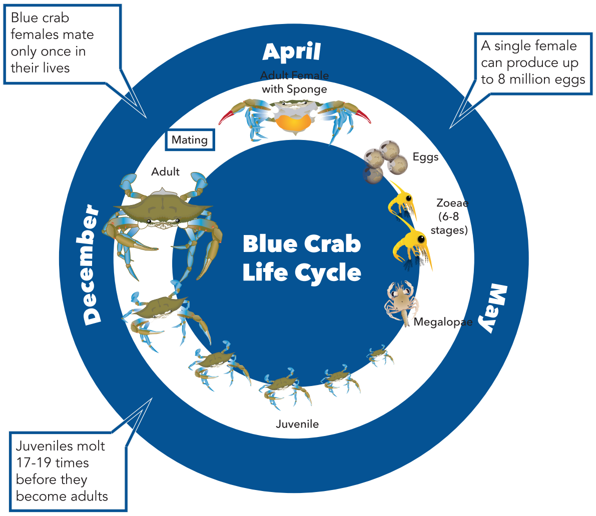
Blue crab females mate only once in their lives
April
Adult Female with Sponge
A single female can produce up to 8 million eggs
Mating
Adult
Eggs
Zoeae (6-8 stages)
Blue Crab Life Cycle
December
May
Megalopae
Juvenile
Juveniles molt 17-19 times before they become adults

Figure 1 The life cycle of *C. sapidus* involves four major life stages. In a hatchery setting, the larvae hatched in April mature to adult stage in December, which then spawn in 3 months and complete their life cycle in a calendar year (J. Sook Chung, unpublished data). The female mates after puberty-terminal molt and travels to higher salinity waters to spawn in the mouth of Chesapeake Bay.

A chromosome-level genome is an important foundational step to provide the expanding blue crab research community a common, shared resource for advancing the science of this species. At present, 67 genome assemblies are available in GenBank for crustaceans including shrimp (Zhang *et al.* 2019), crayfish (Gutekunst *et al.* 2018), Chinese mitten crab, *Eriocheir sinensis* (Song *et al.* 2016; Tang *et al.* 2020a), and the gazami crab, *Portunus trituberculatus* (Tang *et al.* 2020b). In the absence of karyotype or linkage information, the *P. trituberculatus* genome provides an ideal dataset for cross-species comparisons of genome structure, synteny, and orthology. Both species are members of Portunidae or “swimming crab” clade (Evans 2018). The *P. trituberculatus* genome assembly contained 50 chromosome-scale scaffolds, 54% repeat content, and an overall size of 1 billion bases.

For sequencing the blue crab genome, high-coverage short Illumina reads were combined with lower-coverage longer PacBio reads for scaffolding. The reads from the Chicago and HiC were then mapped to the scaffolded assembly and this contact data were used to assemble a chromosome-level genome. Genes were predicted based on comparisons with other species and from 12 blue crab RNAseq datasets from a variety of tissues or life stages. Synteny with *P. trituberculatus* was also used for assembly comparisons. The results provide a chromosome-level draft genome for this important species.

Materials and methods

Animal culture

An adult female that completed its entire life cycle in the blue crab hatchery (Aquaculture Research Center, Institute of Marine and Environmental Technology, Baltimore, MD, USA) was chosen for the genome sequencing. In brief, it was produced from a locally caught (Maryland waters of the Chesapeake Bay) wild-broodstock female in the blue crab hatchery and cultured using a closed, artificial-seawater recirculating aquaculture system. Upon reaching adulthood, the female crab (named as MET-the chosen one) was mated with an adult male derived from a different female that was cultured the same as above, leading to spawning 3 months later (Figure 1).

Sequencing and initial assembly

Genomic sequencing using Illumina 150 bp PE on a HiSeq 2500 platform (Illumina, Inc., San Diego, CA, USA) had a target coverage of 200 Gb based on a genome size of 2 Gb (Bachmann and Rheinsmith 1973). Hemocytes were used for DNA isolation for short-read sequencing. First, the hemolymph was collected directly into a 5 ml syringe containing 2.5 ml of ice-cold anticoagulant at a 1:1 volume ratio of hemolymph and anticoagulant (Alvarez and Chung 2015). The hemocytes were concentrated by centrifuging the hemolymph at 800**g* for 10 min at 4°C. DNA was extracted from these cell pellets using a High Pure PCR Template Preparation kit (Roche), and concentration was measured using a Thermo Scientific (Wilmington, DE, USA) NanoDrop 2000C spectrophotometer (Thermo Scientific Wilmington, DE, USA). In total, 1.4 billion 150 base read pairs were sequenced in a forward and reverse direction. A jump library with a target of 8 kb between mate pairs was constructed by Macrogen Inc. (Rockville, MD, USA) and resulted in 220 million 100 base read pairs in a reverse and forward direction. DNA isolation for PacBio long-read sequencing used various tissues and the rest of the body was kept at −80°C for the future use.

The hybrid assembly of Illumina and PacBio data was done with MASurCA (v 3.2.2) (Zimin *et al.* 2017) followed by deduplication using the included script. Genome size estimates were made by using genomescope (http://qb.cshl.edu/genomescope/) after running jellyfish (Marçais and Kingsford 2011) on the paired reads. After assembly, KAT (Mapleson *et al.* 2017) was used to estimate kmer coverage and samtools (Li *et al.* 2009) was used to estimate genome coverage from reads mapped with bwa-mem (Li 2013). To assess genome quality, Quast (v 5.0.2) was used to tabulate all values reported in the manuscript (Mikheenko *et al.* 2018).

Scaffolding with Chicago, HiC reads using Juicebox

Two libraries of the DNA extracted from the gills and muscles of the frozen animal were constructed using the MboI restriction enzyme by Dovetail Genomics (Santa Cruz, CA, USA) using Chicago and Hi-C methods followed by HiRise scaffolding (Putnam *et al.* 2016). To verify the assembly received from Dovetail genomics, the Arima Mapping UserGuide (v A160156_v02) with included perl scripts was followed. The Chicago and Hi-C reads were extracted from the mapped read files with samtools (v 1.9), followed by mapping to the assembly with bowtie2 (v 2.3.4.1) as single reads, sorting with samtools, 3′ end trimming with filter_five_end.pl, then the reads were combined back into pairs with two_read_bam_combiner.pl, picard tools (v 2.22.4) was used to remove PCR duplicates, and merge replicate library mapping files. A total of 283 million Chicago reads and 160 million HiC reads were mapped to the initial assembly. Juicer (v 1.5) was used to create MboI sites from the draft genome and process the mapped Chicago and Hi-C reads into a contact map (Durand *et al.* 2016; Dudchenko *et al.* 2018). For editing the contact map, Juicebox assembly tools scripts and Juicebox_1.11.08.jar from the 3D-DNA assembly pipeline (v 180922) were used.

Gene prediction

After running RepeatModeler with the included LTR pipeline (Flynn *et al.* 2020), the maker pipeline with repeatmasker and repeatrunner was then used for repeat identification with the species-specific

RepeatModeler libraries and the transposable element library release (January 27, 2017). For protein comparison, a set of eight arthropod genomes were used: *Bicyclus anynana* (v1.2_proteins.fa, GCF_000214255.1), *Bombus terrestris* 1.0_protein.faa, *Drosophila melanogaster* (dmel-all-translation-r6.21.fasta), GCF_000517525.1 *Limulus polyphemus* 2.1.2_protein.faa, *E. sinensis*.gene.pep GCF_003789085.1, *Litopenaeus vannamei* SM378908v1_protein.faa, GCA_000611955.2 *Stegodyphus mimosarum*_v1_protein.faa, *P. trituberculatus* prot_cds.fasta.

Twelve RNAseq, paired-end read datasets from *C. sapidus* (Supplementary Table S1) were mapped to the genome with hisat2, followed by StringTie (v2.1.4, coverage parameters: -f 0.05 -c 3 -t -s 6) to create an "EST" or transcript set for Maker annotation (*Campbell et al.* 2014). JBrowse was used for the genome browser (Buels *et al.* 2016). BUSCO (v 4.1.1) with arthropoda_odb10 was used to estimate genome completeness (Sima *et al.* 2015). A *de novo* gene prediction model for Augustus (Keller *et al.* 2011) was created during the BUSCO pipeline and Augustus was incorporated into the Maker pipeline. Annotation used blastp against the SwissProt UniProt database (2020-03) using an e-value cutoff of 1e-6. Protein domain annotation used interproscan.sh (InterProScan-5.46-81.0) *vs* a set of seven protein domain databases.

Gene validation

For more detailed comparison to *P. trituberculatus* (Tang *et al.* 2020a,b), nucleotide comparisons of genome assemblies used D-GENIES (Cabanettes and Klopp 2018) with MashMap (Jain *et al.* 2018). In addition, MashMap v2.0 was independently used to compare assemblies and perform cross-species nucleotide comparisons. The Synima pipeline (Farrer 2017) was used for comparisons of protein-coding genes which incorporates BLAST and orthomcl for orthology and synteny calculations. Ideograms were constructed with Rideogram (Hao *et al.* 2020).

Data availability

The assembled genome and sequences over 10 kb were deposited to GenBank Accession *Callinectes sapidus* IMET-TCO (JAHFWG000000000; BioSample SAMN18290318) and the assembly, repeats, and protein-coding genes are available on figshare https://doi.org/10.25387/g3.14210594.

Results and discussion

Genome size

Based on kmer analysis and coverage with read mapping, the blue crab genome is estimated to be between 800 and 900 Mb. GenomeScope (v 1.3), a kmer spectrum tool, estimated a genome size of 752 Mb with 245 Mb repeated and a heterozygosity estimate of 1.6% (Figure 2A). The KAT analysis and total assembly size were in agreement with a genome size range of 800–900 Mb (Figure 2B). For planning sequencing, a 2 Gb genome size estimate based on fluorescent staining was used (Bachmann and Rheinsmith 1973), but that value is likely an approximately two-fold overestimate. Using read mapping and coverage confirmed a genome size of under 1 billion bases, so that the short-read coverage was approximately 200× rather than the 100× target. BUSCO scores also supported assembly completeness with only 41 of 1014 total arthropod genes missing (Table 1).

Assembly

The initial or super-read assembly stage using paired and mate paired Illumina sequence libraries resulted in 55 million super-reads summing to 4.6 billion bases. A total of 9.8 million PacBio reads with 53 billion bases and an N50 of 14,305 were then combined with the super-reads to create mega-reads. This second assembly stage produced 2.8 million mega-reads with 7.4 billion bases. After assembly of the mega-reads in the MaSurCA Celera Assembler stage and deduplication via the MaSurCA deduplication script, the assembly was 985 Mb in 14,718 scaffolds or contigs over 10 kb, with an N50 of 152 kb, an N content of 16.3% or 160 Mb, a maximum scaffold size of 1.1 Mb, a GC content of 40%, and a contig N50 of 7954 (Table 2).

Further scaffolding with Chicago and Hi-C sequencing approaches by Dovetail genomics produced an assembly of nearly 1 Gb with an N50 of 7.7 Mb, with half the data in the 12 largest scaffolds and a maximum scaffold size of 108 Mb (Table 2). Two sets of reads using the Chicago approach and another two sets using the HiC approach were obtained. Using the Arima Mapping pipeline with the two sets of Chicago reads to the dovetail assembly resulted in 78% and 80% of reads mapped. For the two Hi-C libraries 95% of reads mapped for both sets of reads. When viewed in Juicebox the contact map of the dovetail assembly suggested over-assembly because chromosome boundaries did not correspond with blocks of contact. In addition, the small chromosome number and dramatic size difference between chromosomes was not supported by: (1) preliminary linkage map data (L. V. Plough., personal communication), (2) attempts at karyotyping which did not suggest small chromosome numbers (SC, personal communication) or >4 fold differences in chromosome sizes, and (3) comparisons with the *P. trituberulatus* genome assembly. The 3D-DNA pipeline was then used followed by editing with Juicebox assembly tools (JBAT), resulting in the final *C. sapidus* genome draft assembly of 998 Mb (Figure 3) with 3971 scaffolds over 10 kb and an N50 of 12 Mb at 23 sequences (Table 2, Supplementary Additional file 1).

The genome assembly contains 50 sequences over 1.5 Mb or candidate chromosomal scaffolds with sizes ranging up to 37 Mb (Figure 4). The large scaffolds sum to 810 Mb with a median chromosome size of 14.5 b wiSupplementary Table S2). The remaining approximately 190 Mb of sequence was not placed into the chromosomal assembly. The unplaced smaller sequences often had abundant Hi-C contacts between sequences, but these contacts were ambiguous and not diagnostic or unique for placement into a specific large scaffold (Figure 3). In particular, HiC_scaffold_46 in the lower right of the contact map has many blocks of contact within the small unplaced sequences in the lower right corner. Similar contact patterns have also been observed when microchromosomes are present (Cheng *et al.* 2021). In *C. sapidus*, these contacts are general and often match many chromosomes. The overall contact map is 1 Gb, but the unplaced 190 Mb span of sequence corresponds well with gaps within the scaffolds which are up to 20% of some chromosomes. This then suggests the contact map is a ∼20% overestimate of the genome size.

Repeat annotation

As the first step in annotation, RepeatModeler (dfam/tetools: 1.2) identified 1935 families of repeats in the genome (Supplementary Additional files 2 and 3). RepeatMasker used these families and existing databases to classify 401 Mb of sequence as repeats covering 36% of the genome (Figure 5, Supplementary Table S2). The largest set of repeats spanned 188 . The largest set of repeats spanned 188sed these families and existing databases to classify 401 Mb of sequence as repeats covering 36% of the genome (with gaps within the scaffolds whssified by RepeatMasker, long interspersed nuclear elements (LINEs) were most common, with

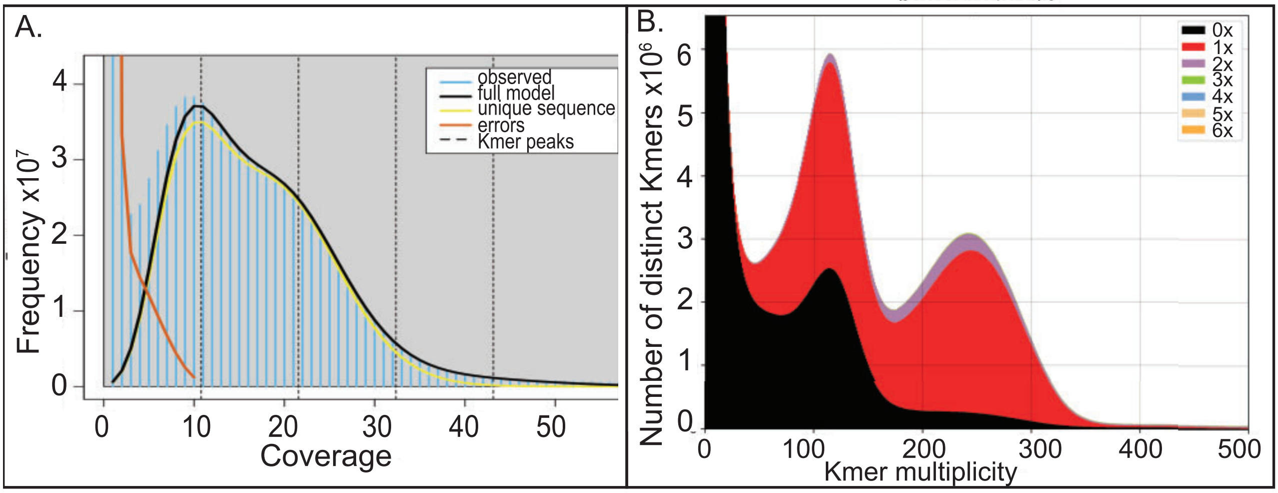

Figure 2 Genome size estimates for *C. sapidus* using Kmer-based methods and coverage of the assembly. (A) GenomeScope 24mer profile using merged 150 base Illumina reads calculated genome size of 750 Mb with 245 Mb of repeats. (B) After assembly the KAT Kmer multiplicity with 100× coverage Illumina paired reads agreed with the genome size estimates.

Table 1 Universal single copy ortholog results using BUSCO version 4.1.1 and arthropoda_odb10

	Number	Percent
Complete BUSCOs	946	93
Complete and single-copy BUSCOs	888	88
Complete and duplicated BUSCOs	58	6
Fragmented BUSCOs	26	3
Missing BUSCOs	41	4
Total BUSCO groups searched	1013	

Table 2 Assembly statistics for different stages of genome assembly

	MASuRCA	Dovetail	JBAT assembly
# contigs	90,964	89,369	85,343
Total length	1,156,251,922	1,113,662,720	1,115,648,567
# contigs (≥10,000 bp)	14,718	8006	3971
# contigs (≥25,000 bp)	6025	3569	1371
# contigs (≥50,000 bp)	3703	2089	892
Total length (≥10,000 bp)	984,995,868	996,110,707	998,050,073
Total length (≥25,000 bp)	851,874,722	925,280,978	961,688,798
Total length (≥50,000 bp)	772,162,490	873,473,002	944,714,467
Largest contig	3,463,967	108,362,586	37,712,751
GC (%)	40.28	40.14	40.14
N50	152,832	7,699,975	14,840,478
N75	22,206	74,275	500,000
L50	1711	12	23
L75	6678	1463	79
# N's per 100 kbp	16,272	20,091	20,234

47.5 Mb. Within the LINEs the L2/CR1/Rex group was 21 Mb while RTE/BovB covered 14 Mb, 6.4 Mb of which was attributed to a single RepeatModeler family-190. Repeats classified as DNA transposons only covered 12.5 Mb. Simple repeats were the second largest class of repeats with over 109 Mb and the most common dinucleotide repeats were TC and the complimentary AG which together spanned 28 Mb. Strongly AT or GC-biased repeats were not common with only 3.3 Mb of AT repeats, but these were about one hundred times more common than the 26.5 kb of GC repeats. Overall, the simple repeats were not strongly GC or AT biased, similar to the overall genome bias for AT (40.32% GC).

When the repeats were mapped to individual scaffolds a striking pattern emerged (Figure 6). Of the 810 Mb of sequence in the 50 largest scaffolds, 291 Mb or 36% was annotated as repetitive while for the remaining 190 Mb of smaller sequences 110 Mb were repetitive, or 58%. This suggests that repeat sequences were more difficult to place into large scaffolds and may be underestimated in the larger chromosome scale assembly because of ambiguity in placement based on Hi-C contact. Simple repeats accounted for 8–10% of the sequence in 42 of the 50 largest scaffolds but there were nine scaffolds with less than 8% simple repeat content (Figure 6A). Similarly, most scaffolds had 20–40% interspersed repeats, but there were seven scaffolds with more than 50% interspersed repeat content including scaffold_50 with 81% (Figure 6B). Intriguingly, seven of the nine scaffolds with the highest interspersed repeat content also had lower simple sequence repeats using the values described above. Overall, this agrees with the data from the contact map where the lower right corner of higher numbered scaffolds have abundant contact with the smaller sequences and often contact with several chromosomes (Figure 3).

Protein-coding gene prediction and annotation

For gene prediction from 12 RNAseq datasets from different developmental stages or tissues (Supplementary Table S1), 309 million read pairs were mapped to the assembly using hisat2, which mapped 89.46% of reads. The StringTie software then was used to make a sequence dataset of 63,487 transcripts. These were used as an Expressed Sequence Tag input for the maker pipeline. In addition, the protein dataset of genomes described in Methods and an Augustus gene prediction model based on BUSCO gene searching was also provided to the Maker pipeline.

The gene prediction pipeline resulted in 25,249 predicted protein-coding genes of which 25,067 were unique with a mean length of 1813 bases and 6 introns per gene (Figure 7) (Supplementary Additional file 4). Of the 156,740 predicted exons, the mean length was 292 bases. Of the predicted genes, just under half had a BLAST hit with an e-value cutoff of 1e−6 or less, and this is reflected in the UniProt annotation results for different protein domains (Table 3).

The distribution of exons across the larger scaffolds (Figures 4 and 6C, Supplementary Table S2) is also highly variable and for

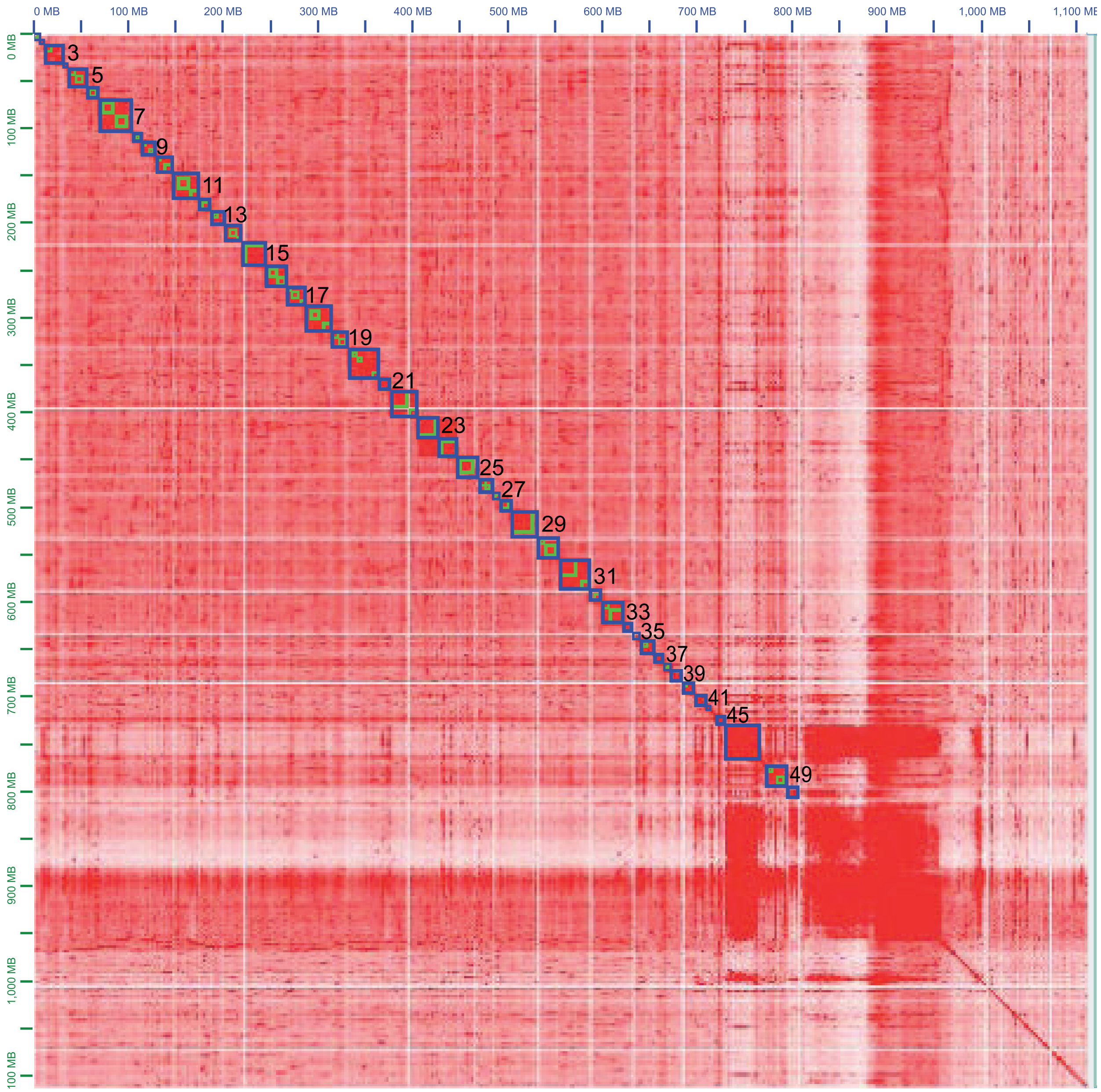

Figure 3 Contact map: The Juicebox Assembly Tools contact map of Chicago and Hi-C sequence data for *C. sapidus* with initial scaffold boundaries in green and final scaffold boundaries in purple. Every other scaffold is labeled. The assembly can be roughly divided into the 810 Mb in 50 large scaffolds and 190 Mb of unplaced sequence in the lower right. There is a set of sequence between 870 Mb and 960 Mb that has strong contact with most of the assembled data.

some scaffolds inversely correlated with repeat abundance. In the large scaffolds, the predicted gene content was as high as 848 genes per scaffold, with mean of 27.2 genes per Mb (Figure 6C) and the large scaffolds accounted for 21,588 or 86% of protein-coding genes. The 50 scaffolds could be divided into two categories, one with 41 sequences containing an average of 30.2 (range 24–50) predicted genes per Mb. These scaffolds contained 10% simple sequence repeats and 23% interspersed repeats. The second group with nine scaffolds contained 969 protein-coding genes with an average of only 13.6 (1.5–34.5) genes per Mb (Figure 6C). As described above these scaffolds had on average 60% interspersed repeat content and 4% simple repeats.

Synteny with *Portunus trituberculatus*

Comparisons at the amino acid and nucleotide level of the genomes from *C. sapidus* and *P. trituberculatus* assemblies revealed strong approximate nucleotide matches for 41 chromosomes based on D-GENIES (Figure 8, Supplementary Table S2). Many specific regions of high identity were found between the genomes (Figure 9). However, of a total of 50 large chromosome-scale scaffolds in both genomes, nine did not strongly match between the two species. The scaffolds in *C. sapidus* that did not have strong matches to *P. trituberculatus* were scaffolds, 4, 41, 43, 44, 46, 47, 48, 49, and 50, while in *P. trituberculatus* scaffolds 1, 2, 4, 5, 6, 7, 8, 22, and 26 did not match to *C. sapidus*. As noted above, the higher

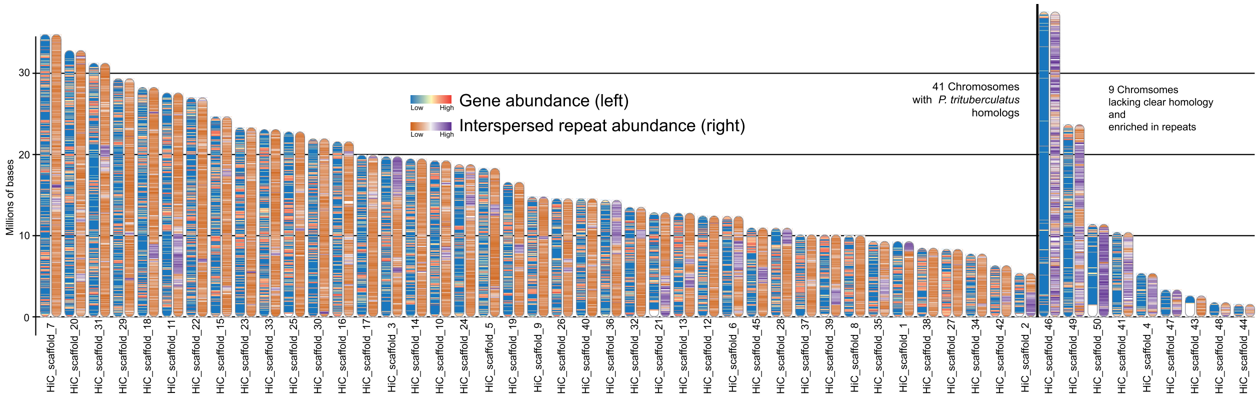

Figure 4 Ideogram of the 50 large scaffolds showing protein-coding gene abundance (exon density on left) and repeat masker interspersed repeat abundance (on right) for each of the 50 large scaffolds. Scaffolds with clear homologs in *P. trituberculatus* are shown on the left sorted by size, while on the right scaffolds with no obvious homology are also sorted by size. Scaffold numbering as based on the contact map from Figure 3.

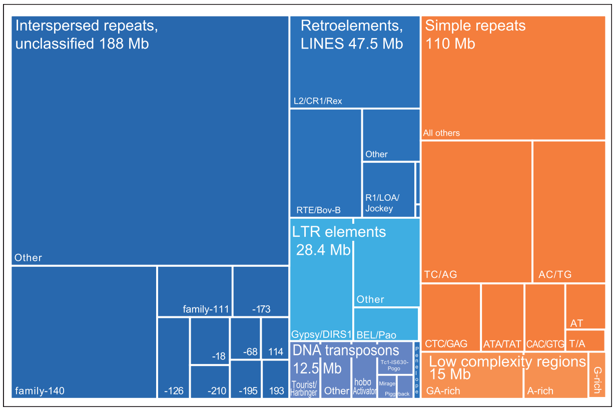

Figure 5 An area proportional treemap of the different repeat categories found using RepeatModeler coupled with RepeatMasker on the *C. sapidus* genome assembly. The left side in dark blue with 188 Mb of data represents sequence data masked based on *C. sapidus*-specific repeat families, with the largest 10 families shown and the remaining families aggregated into one category. In the middle in blue are repeats classified by RepeatModeler into LINES, LTR, and DNA transposons. On the right simple repeats and low complexity regions are shown in orange with the most abundant di- and trinucleotides shown as individual categories.

numbered scaffolds in *C. sapidus* often had abundant contact with the repeat-dense unplaced sequence in the contact map (Figure 3) and were often enriched in repeat sequence and had lower gene content (Figures 5 and 6). This could be due to a combination of genuine differences between the genomes including rearrangement and mis-assembly. In most comparisons at the nucleotide and amino acid level there were multiple rearrangements between the genomes of the two species. The example of scaffold 11 in Figure 8 shows several likely major rearrangements.

The synteny comparison of predicted protein-coding genes from *C. sapidus* with *P. trituberculatus* revealed a similar pattern to the nucleotide comparison where there was high amino acid identity between individual genes. The orthomcl pipeline identified 9423 genes that were one-to-one predicted orthologs and 733 were duplicated once in *C. sapidus*, while 359 were duplicated once in *P. trituberculatus*. Synteny of at least two colinear genes between the two genomes using the Synima pipeline accounted for 7165 or 76% of the putative orthologs (Figures 7 and 8). Syntenic genes accounted for between 9% and 50% of the overall predicted protein-coding genes in homologous scaffolds between species (Figure 6C, Supplementary Table S2). Overall, these results also reflect possible overestimates in predicted protein-coding genes for *C. sapidus* with a core gene number predicted by the Synima and orthomcl pipeline between the two species of 17,553 in *C. sapidus* and 14,420 in *P. trituberculatus*.

Conclusions

Callinectes sapidus, like *P. trituberculatus*, *E. sinensis*, and *L. vannamei* (Zhang *et al.* 2019) is a high-value, culturally and ecologically important marine crustacean. The blue crab is also a good candidate for commercial aquaculture, hatchery programs to support the fishery, and monitoring of wild population genetic diversity, all of which can benefit from a complete draft genome. Comparison of the genome sequences of these other marine crustacean species to the blue crab revealed common elements including chromosome numbers >40, highly repetitive genomes (36–56%), and high predicted gene content of 15–25 thousand genes. In a phylogentic context, the astacidea clade close to brachyra, have karyotype estimates of 94 chromosomes for the crayfish, *Procambrus clarkii* (Salvadori *et al.* 2014) and 66 for lobster *Homarus americanus* (Hughes, 1982), even higher than brachyruans and *L. vannamei*. *Homarus americanus* also has a variable number of small supernumary chromosomes which have not yet been described in *C. sapidus*, but which might partially explain the abundant non-specific contacts in the contact map and non-matching scaffolds in comparisons with *P. trituberculatus*.

Of the 50 chromosome scale scaffolds in the *C. sapidus* genome assembly, only 41 had clear homology to *P. trituberculatus* based on nucleotide comparisons and syntenic protein-coding genes. The other related decapods with genome assemblies do not yet have chromosome-scale data currently available. Strikingly, both of the crab assemblies had 50 large scaffolds and nine orphan scaffolds that were not strong matches between species. For *C. sapidus*, in particular, these nine scaffolds were enriched for repetitive elements, had reduced simple sequence repeats, and reduced protein-coding potential when compared to chromosomes with homology to *P. trituberculatus*. Sequencing *P. trituberculatus* relied on Nanopore long reads combined with Illumina sequencing while *C. sapidus* used shorter PacBio sequencing and the initial scaffolding used different assemblers. Both the *P. trituberculatus* and *C. sapidus* genome assembly methods used contact maps at the last assembly stage and this may have resulted in artifactual assembly of repeat-rich scaffolds with widespread contacts due

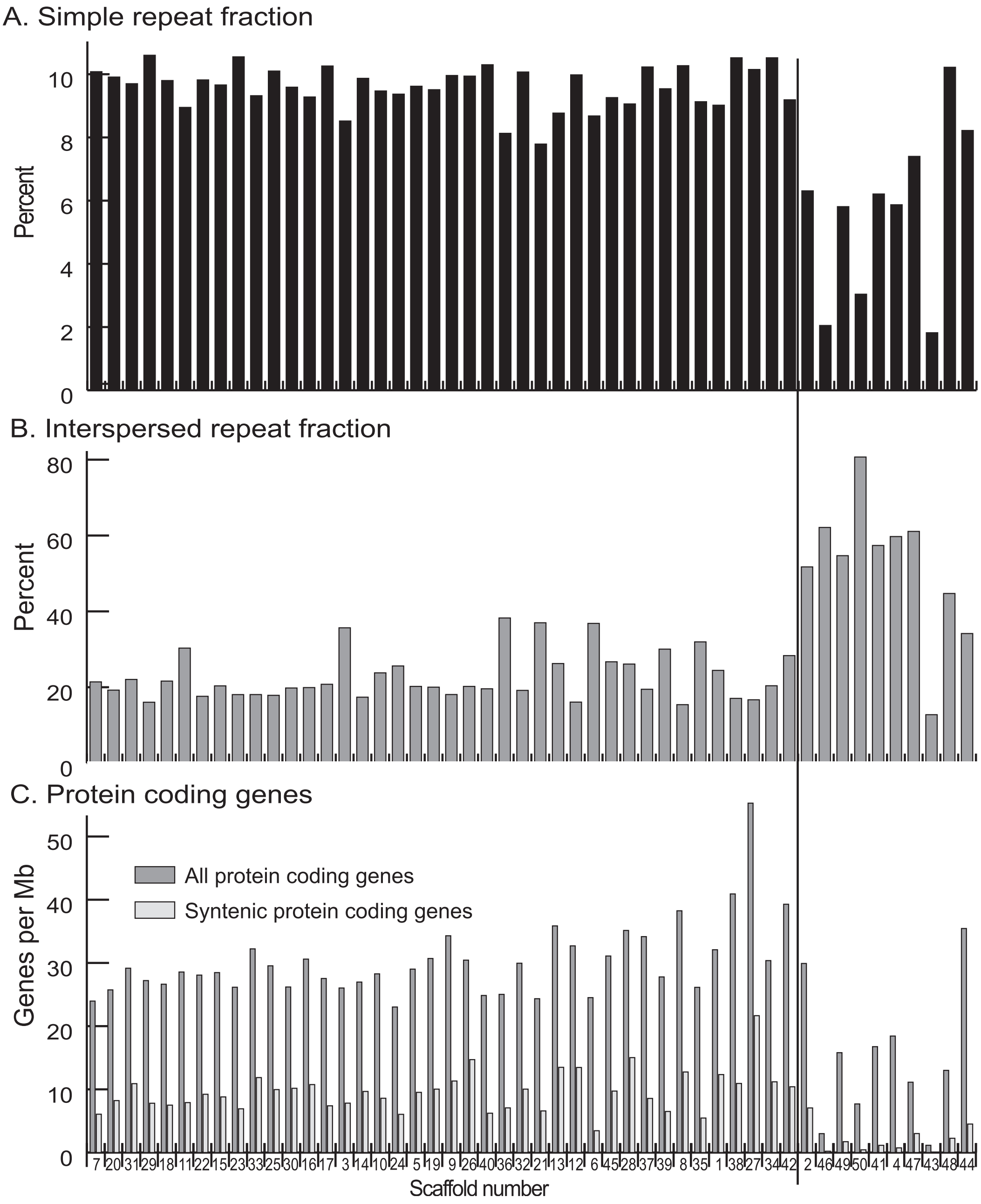

Figure 6 Annotation of the 50 large scaffolds ordered as in Figure 4. (A) Simple sequence repeats, (B) Interspersed repeats, and (C) Protein-coding genes including all predicted genes and genes that were syntenic with *P. trituberculatus*.

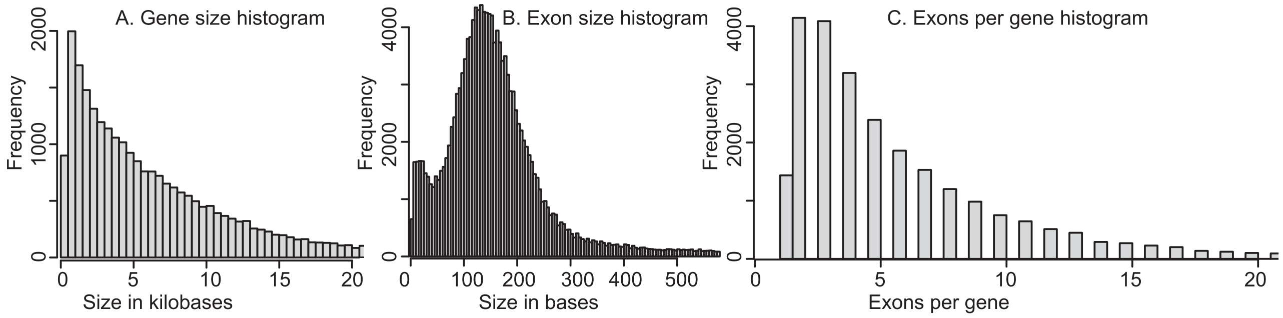

Figure 7 Predicted protein-coding genes based on a combination of RNAseq from 12 *C. sapidus* tissues or life stages, 8 arthropod genomes and *ab initio* prediction with Augustus. (A) Size of predicted protein-coding genes. (B) Size of exons. (C) Number of exons per gene.

Table 3 Annotation based on interproscan and blastp

Database	No. of genes	No. of GO terms for given database
Blastp to SwissProt/UniProt	12,925	NA
Pfam	12,479	11,853
TIGRFAM	498	493
CDD	5278	1778
PRINTS	2448	7684
SMART	5680	7791
SUPERFAMILY	11,036	4145
PANTHER	14,353	5155

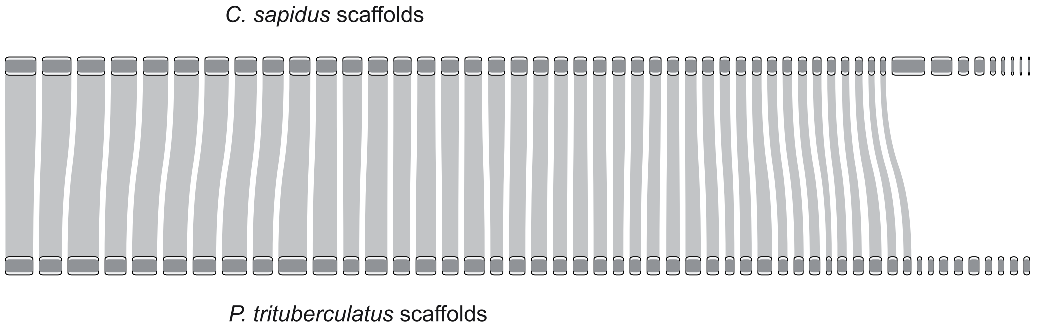

Figure 8 Overall identity between the 50 largest scaffolds from *C. sapidus* and *P. trituberculatus* based on D-GENIES with MashMap mapping. The scaffold order is as in Figure 4.

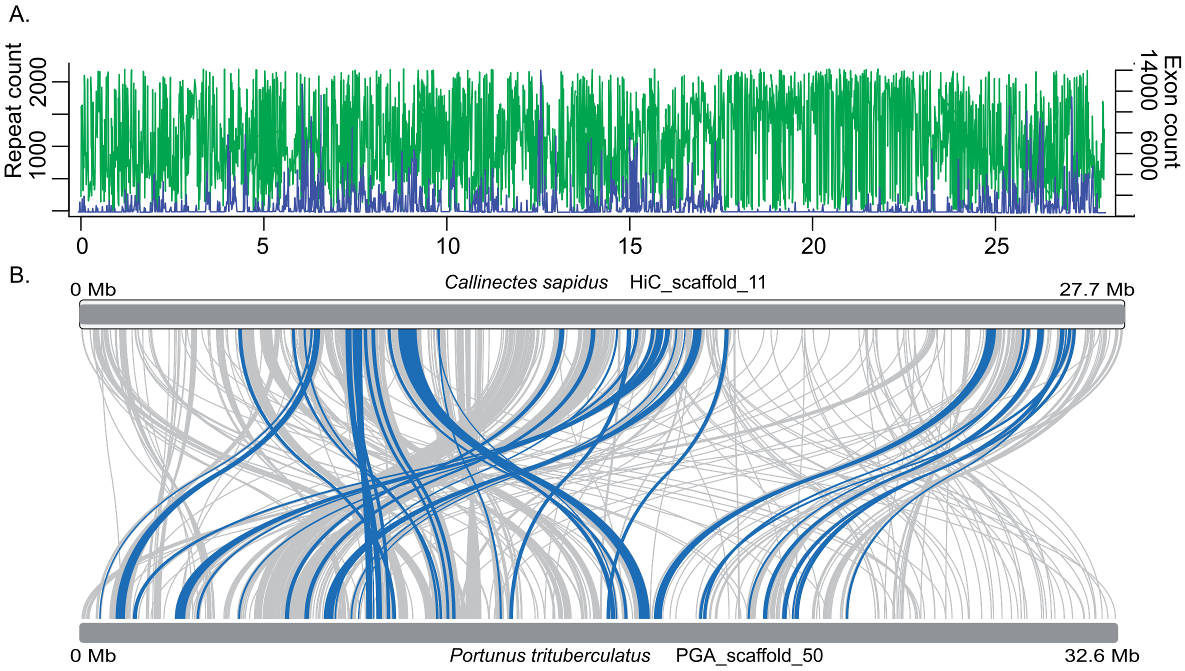

Figure 9 Detail of a single 27.7 Mb *C. sapidus* putative chromosome sized scaffold 11. (A) Repeat content in a 10 kb sliding window is shown in green, while predicted exon content is shown in blue for the same window size. (B) Below is a comparison with the 32.6 Mb homologous chromosome from *P. trituberculatus*, scaffold 50. The nucleotide matches over 5 kb between the chromosomes are shown with gray lines and syntenic protein-coding genes are shown in blue lines.

to the relatively high chromosome number, high repeat density, and limits of the contact map technology and analysis. However, it is also possible that at least some of these repeat-rich chromosomes may represent genuine, rapidly diverging regions between the two species.

Overall, the comparison between these first draft genomes of two closely related decapods strongly validates the *de novo* approach to genome assembly at the chromosome scale for economically important non-model species. Both genomes of the blue crab and *P. trituberculatus* contain high and novel repeat content, are composed of 40–50 chromosomes, and have a genome size near a billion bases, adding to the assembly challenge. Therefore, the result that 41 of 50 large chromosome scale scaffolds are homologous between the species is a strong validation of this approach. Having these two genomes also enables more accurate gene prediction, defines a core genome for this subset of decapod crustaceans, and unlocks novel research into population diversity, disease resistance, and adaptation to the environment for *C. sapidus*.

Callinectes sapidus has not had a long-term breeding program to create an inbred line for sequencing and so the assembled genome had to account for substantial heterozygosity. The lack of recombination-based linkage data or other cytogenetic information such as karyotypes for blue crab means that there is little outside reference data to guide the assembly. However, the chromosome level-genome assembly, together with that of *P. trituberculatus*, will permit genomic work assessing genetic diversity and species identification across the range (Place and Plough 2017; Plough 2017; Windsor *et al.* 2019; Lee *et al.* 2020). Future work to advance blue crab aquaculture through selective breeding could make use of the genome through a genomic selection approach (Meuwissen *et al.* 2001; Heffner *et al.* 2009), which has recently been applied to aquaculture species like shrimp and mollusks (Hollenbeck and Johnston 2018; Zenger et al. 2019; Houston *et al.* 2020). The blue crab genome will also facilitate genomic and phylogenomic comparisons among marine crustacea in general (Wolfe et al. 2019) and contribute to the refinement of the phylogeny of Brachyuran crabs specifically, which are a highly speciose group and comprise most of the world's commercially-fished crab species (Ng *et al.* 2008; Davie *et al.* 2015).

Acknowledgments

We are grateful to the staff of aquaculture research center (ARC, Institute of Marine and Environmental Technology, Baltimore, MD) for maintaining the closed recirculating aquaculture system during the crab culture and of Zooquatic Laboratory for daily water analysis. We also thank Mr. Ben Decre for setting up the genome browser and searchable database for sequence searches.

Funding

Funding was generously provided by a number of citizens of Maryland: Mike and Trish Davis, Jim Albrecht, Don and Cathy MacMurray, Bert and Debbie Winchester, Arnie and Alison Richman, Dick Franyo, Jim Connell, Mitch Neitzey, Russell Hill, J. Sook Chung, Jim and Jenny Corckran, David Balcom, and Nick Hammond. Foundation support from the following foundations is greatly appreciated: the Vetlesen Foundation, the Edward St. John Foundation. Additional funding was provided by Sea Grant College of Maryland, National Science Foundation grant (no. 1146774) to J.S.C., and National Science Foundation BioOce Grant (no. 1658466) to L.V.P.

Conflicts of interest

None declared.

Author contributions

J.S.C. cultured the animals, extracted DNA and RNA, participated in genome validation, RNAseq analysis and wrote and edited the manuscript. T.R.B. assembled the genome, did other bioinformatic analysis, wrote and edited the manuscript. R.C.M. did genome comparisons, annotation, and edited the manuscript and drafted figures. L.V.P. wrote and edited the manuscript and assisted with validation of the assembly.

Literature cited

Alvarez JV, Chung JS. 2015. The involvement of hemocyte prophenoloxidase in the shell-hardening process of the blue crab, *Callinectes sapidus*. PLoS One. 10:e0136916.

Anger K. 1998. Patterns of growth and chemical composition in decapod crustacean larvae. Invertebr Reprod Dev. 33:159–176.

Bachmann K, Rheinsmith EL. 1973. Nuclear DNA amounts in pacific Crustacea. Chromosoma. 43:225–236.

Bembe S, Dong L, Chung JS. 2017. Optimal temperature and photoperiod for the spawning of blue crab, *Callinectes sapidus*, in captivity. Aquac Res. 48:5498–5505.

Bembe S, Zmora N, Williams E, Place Ar AR, Liang D, *et al.* 2018. Effects of temperature and photoperiod on hemolymph vitellogenin levels during spawning events of the blue crab, *Callinectes sapidus*, in captivity. Aquac Res. 49:2201–2209.,

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, *et al.* 2016. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 17:66.

Cabanettes F, Klopp C. 2018. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. PeerJ. 6e:e4958.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinform. 48:4.11.1–4.11.39.

Cheng P, Huang Y, Lv Y, Du H, Ruan Z, *et al.* 2021. The American paddlefish genome provides novel insights into chromosomal evolution and bone mineralization in early vertebrates. Mol Biol Evol. 38:1595–1607.

Chung JS, Manor R, Sagi A. 2011. Molecular cloning of the full length cDNA encoding an insulin-like androgenic gland factor (IAG) from the androgenic gland of adult male blue crab, *Callinectes sapidus:* an implication eyestalk neuropeptide(s) involvement of IAG expression. Gen Comp Endocrinol. 173:4–10.

Davie PJF, Guinot D, Ng PKL. 2015. Phylogeny of Brachyura. In: P Castro, P Davie, D Guinot, FR Schram, J von Vaupel Klen, editors. The Crustacea. Leiden: Brill. p. 921–979.

Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, *et al.* 2018. The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv. doi:10.1101/254797.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, *et al.* 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3:99–101.

Evans N. 2018. Molecular phylogenetics of swimming crabs (Portunoidea Rafinesque, 1815) supports a revised family-level classification and suggests a single derived origin of symbiotic taxa. PeerJ. 6:e4260.

FAO Fisheries and Aquaculture. 2019. *Callinectes sapidus* species fact sheet.

Farrer RA. 2017. Synima: A Synteny imaging tool for annotated genome assemblies. BMC Bioinformatics. 18:507.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, *et al.* 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 117:9451–9457.

Gutekunst J, Andriantsoa R, Falckenhayn C, Hanna K, Stein W, *et al.* 2018. Clonal genome evolution and rapid invasive spread of the marbled crayfish. Nat Ecol Evol. 2:567–573. doi:10.1038/s41559-018-0467-9.

Hao Z, Lv D, Ge Y, Shi J, Weijers D, *et al.* 2020. RIdeogram: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci. 6:e251.

Heffner EL, Sorrells ME, Jannink JL. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12.

Hines AH, Johnson EG, Darnell MZ, Rittschof D, Miller TJ, *et al.* 2011. Predicting effects of climate change on blue crabs in Chesapeake Bay. In: GH Kruse, GL Eckert, RJ Foy, RN Lipcius, B Sainte-Marie, *et al.* Biology and Management of Exploited Crab Populations under Climate Change. Alaska SeaGrant Program University of Fairbanks. pp. 109–127.

Hines AH. 2007. Ecology of juvenile and adult blue crabs. In: VS Kennedy, LE Cronin, editors. Blue Crab: Callinectes Sapidus. College Park, MD: MD Sea Grant College Press. p. 575–665.

Hollenbeck CM, Johnston IA. 2018. Genomic tools and selective breeding in molluscs. Front Genet. 9:253–215.

Houston RD, Bean TP, Macqueen DJ, Gundappa MK, Jin YH, *et al.* 2020. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat Rev Genet. 21:389–409.

Huang X, Bae S, Bachvaroff TR, Schott EJ, Ye H, *et al.* 2016. Does a blue crab putative insulin-like peptide binding protein (ILPBP) play a role in a virus infection? Fish Shellfish Immunol. 58: 340–348.

Hughes JB. 1982. Variability of chromosome number in the lobsters, *Homarus americanus* and *Homarus gammarus*. Caryologia. 35: 279–289.

Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S. 2018. A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics. 34:i748–i756.

Katsanevakis S, Wallentinus I, Zenetos A, Leppäkoski E, Çinar ME, *et al.* 2014. Impacts of invasive alien marine species on ecosystem services and biodiversity: a pan-European review. Aquat Invasions. 9:391–423.

Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 27:757–763.

Lannan JE. 1980. Broodstock management of *Crassostrea gigas*: I. Genetic and environmental variation in survival in the larval rearing system. Aquaculture. 21:323–336.

Lee BB, Schott EJ, Behringer DC, Bojko J, Kough A, *et al.* 2020. Rapid genetic identification of the blue crab *Callinectes sapidus* and other *Callinectes* spp. using restriction enzyme digestion and High Resolution Melt (HRM) assays. Front Mar Sci. 7:633.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.* 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079. doi:10.1093/bioinformatics/btp352.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv (preprint posted 26 May 2013).

Lipcius RN, Eggleston EB, Heck KL, Seitz RD, Van Montfrans J. 2007. Post-settlement abundance, survival, and growth of postlarvae and young juvenile blue crabs in nursery habitats. In: VS Kennedy, LE Cronin, editors. Blue Crab: Callinectes Sapidus. College Park, MD: MD Sea Grant College Press. p. 535–566.

Mancinelli G, Chainho P, Cilenti L, Falco S, Kapiris K, *et al.* 2017. The Atlantic blue crab *Callinectes sapidus* in southern European coastal waters: distribution, impact and prospective invasion management strategies. Mar Pollut Bull. 119:5–11. doi: 10.1016/j.marpolbul.2017.02.050.

Mapleson D, Accinelli GG, Kettleborough G, Wright J, Clavijo BJ. 2017. Sequence analysis KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 33: 574–576. doi:10.1093/bioinformatics/btw663.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27: 764–770. doi:10.1093/bioinformatics/btr011.

Maurer L, Liang D, Chung JS. 2017. Effects of prey densities and dietary supplementation on the larval development of the blue crab *Callinectes sapidus* Rathbun, 1896 (Brachyura: Portunidae). J Crustac Biol. 37:674–682. doi:10.1093/jcbiol/rux079.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 34:i142–150. doi:10.1093/bioinformatics/bty266.

Nehring S. 2011. Invasion history and success of the American Blue Crab *Callinectes sapidus* in European and adjacent waters. In: B Galil, PF Clark, JT Carlton, editors. The Wrong Place—Alien Marine Crustaceans: Distribution, Biology, and Impacts. US: Springer. p. 607–624.

Ng PKL, Guinot D, Davie PJ. 2008. Systema Brachyurorum: part 1. An annotated checklist of extant brachyuran crabs of the world. Raffles Bull Zool. 17:1–286.

Place AR, Plough LV. 2017. The genetic enablement of the blue crab *Callinectes sapidus*. J. Shellfish Res. 36:227–229. doi:10.2983/035.036.0125.

Plough LV. 2017. Population genomic analysis of the blue crab *Callinectes sapidus* using genotyping-by-sequencing. J Shellfish Res. 36:249–261. doi:10.2983/035.036.0128.

Putnam NH, Connell BO, Stites JC, Rice BJ, Hartley PD, *et al.* 2016. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. Genome Res. 26:342–350. doi:10.1101/gr.193474.115.Freely.

Salvadori S, Coluccia E, Deidda F, Cau A, Cannas R, *et al.* 2014. Karyotype, ribosomal genes, and telomeric sequences in the crayfish *Procambarus clarkii* (decapoda: cambaridae). J Crust Biol. 34:525–531.

Sandifer PA, Smith TJ. 1979. Possible significance of variation in the larval development of palaemonid shrimp. J Exp Mar Biol Ecol. 39:55–64.

Sima FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. Genome analysis BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212. doi:10.1093/bioinformatics/btv351.

Song L, Bian C, Luo Y, Wang L, You X, *et al.* 2016. Draft genome of the Chinese mitten crab, *Eriocheir sinensis*. Gigascience. 5:5–6. doi: 10.1186/s13742-016-0112-y.

Tang B, Wang Z, Liu Q, Zhang H, Jiang S, *et al.* 2020a. High-quality genome assembly of *Eriocheir japonica sinensis* reveals its unique genome evolution. Front Genet. 10:1340.doi:10.3389/fgene.2019.01340.

Tang B, Zhang D, Li H, Jiang S, Zhang H, *et al.* 2020b. Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (*Portunus trituberculatus*). Gigascience. 9:10. doi:10.1093/gigascience/giz161.

Taris N, Batista FM, Boudry P. 2007. Evidence of response to unintentional selection for faster development and inbreeding depression in *Crassostrea gigas* larvae. Aquaculture. 272:S69–S79. doi: 10.1016/j.aquaculture.2007.08.010.

Techa S, Chung JS. 2015. Ecdysteroids regulate the levels of Molt-Inhibiting Hormone (MIH) expression in the blue crab. Plos One. 10:e0117278.doi:10.1371/journal.pone.0117278.

Van Engel WA. 1958. The blue crab and its fishery in Chesapeake Bay. Part 1. Reproduction, early development, growth and migration. Commer Fish Rev. 20:6.

Williams AB. 1974. The swimming crabs of the genus *Callinectes* (Decapoda: Portunidae). Fish Bull./U.S. Dept. Commer. Natl. Ocean. Atmos. Adm. Natl. Mar. Fish. Serv. 72:685–798.

Windsor AM, Moore MK, Warner KA, Stadig SR, Deeds JR. 2019. Evaluation of variation within the barcode region of Cytochrome c Oxidase i (COI) for the detection of commercial *Callinectes sapidus* Rathbun, 1896 (blue crab) products of non-US origin. PeerJ. 7: e7827.doi:10.7717/peerj.7827.

Wolfe JM, Breinholt JW, Crandall KA, Lemmon AR, Lemmon EM, *et al.* 2019. A phylogenomic framework, evolutionary timeline and genomic resources for comparative studies of decapod crustaceans. Proc Biol Sci. 286:20190079.

Zenger KR, Khatkar MS, Jones DB, Khalilisamani N, Jerry DR, *et al.* 2019. Genomic selection in aquaculture: application, limitations and opportunities with special reference to marine shrimp and pearl oysters. Front Genet. 9:693.doi:10.3389/fgene.2018.00693.

Zhang X, Yuan J, Sun Y, Li S, Gao Y, *et al.* 2019. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. Nat Commun. 10:356. doi:10.1038/s41467-018-08197-4.

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, *et al.* 2017. The first near-complete assembly of the hexaploid bread wheat genome. Gigascience. 6:1–7.

Zmora N, Chung JS. 2014. A novel hormone is required for the development of reproductive phenotypes in adult female crabs. Endocrinol. 155:230–239. doi:10.1210/en.2013-1603.

Zmora N, Sagi A, Zohar Y, Chung JS. 2009. Molt-inhibiting hormone stimulates vitellogenesis at advanced ovarian developmental stages in the female blue crab, *Callinectes sapidus* 2: novel specific binding sites in hepatopancreas and cAMP as a second messenger. Saline Syst. 5: 6. doi:10.1186/1746-1448-5-6.

Zmora N, Trant J, Zohar Y, Chung JS. 2009. Molt-inhibiting hormone stimulates vitellogenesis at advanced ovarian developmental stages in the female blue crab, *Callinectes sapidus* 1: an ovarian stage dependent involvement. Saline Syst. 5:7.doi:10.1186/1746-1448-5-7.

Zohar Y, Hines AH, Zmora O, Johnson EG, Lipcius RN, *et al.* 2008. The Chesapeake Bay blue crab (*Callinectes sapidus*): a multidisciplinary approach to responsible stock replenishment. Rev. Fish. Sci. 16: 24–34. doi:10.1080/10641260701681623.

Communicating editor: J. M. Yáñez