Chromosomal-Level Genome Assembly of the Sea Urchin *Lytechinus variegatus* Substantially Improves Functional Genomic Analyses

Phillip L. Davidson[1,†], Haobing Guo[2,3,†], Lingyu Wang[1], Alejandro Berrio[1], He Zhang[2,3], Yue Chang[2,3], Andrew L. Soborowski[4,5], David R. McClay[1], Guangyi Fan[2,3,*], and Gregory A. Wray[1,4,5,*]

[1]Department of Biology, Duke University

[2]Beijing Genomics Institute—Qingdao, China

[3]Beijing Genomics Institute—Shenzhen, China

[4]Program in Computational Biology and Bioinformatics, Duke University

[5]Center for Genomic and Computational Biology, Duke University

*Corresponding authors: E-mails: fanguangyi@genomics.cn; gwray@duke.edu.

Accepted: 12 May 2020

†These authors contributed equally to this work.

Data deposition: Sequencing data and genome assembly have been deposited at the China National GeneBank Database (CNGBdb) under the accession CNP0000959. Transcript models, protein models, and genome annotation files are available in supplementary data 1, Supplementary Material online.

Abstract

Lytechinus variegatus is a camarodont sea urchin found widely throughout the western Atlantic Ocean in a variety of shallow-water marine habitats. Its distribution, abundance, and amenability to developmental perturbation make it a popular model for ecologists and developmental biologists. Here, we present a chromosomal-level genome assembly of *L. variegatus* generated from a combination of PacBio long reads, 10× Genomics sequencing, and HiC chromatin interaction sequencing. We show *L. variegatus* has 19 chromosomes with an assembly size of 870.4 Mb. The contiguity and completeness of this assembly are reflected by a scaffold length N50 of 45.5 Mb and BUSCO completeness score of 95.5%. Ab initio and transcript-informed gene modeling and annotation identified 27,232 genes with an average gene length of 12.6 kb, comprising an estimated 39.5% of the genome. Repetitive regions, on the other hand, make up 45.4% of the genome. Physical mapping of well-studied developmental genes onto each chromosome reveals nonrandom spatial distribution of distinct genes and gene families, which provides insight into how certain gene families may have evolved and are transcriptionally regulated in this species. Lastly, aligning RNA-seq and ATAC-seq data onto this assembly demonstrates the value of highly contiguous, complete genome assemblies for functional genomics analyses that is unattainable with fragmented, incomplete assemblies. This genome will be of great value to the scientific community as a resource for genome evolution, developmental, and ecological studies of this species and the Echinodermata.

Key words: *Lytechinus variegatus*, sea urchin, echinoderm, genome, chromosome, gene regulatory network.

Introduction

Lytechinus variegatus, also known as the variegated sea urchin, is a widely distributed western Atlantic species commonly found in seagrass beds and hard-bottomed shallow-water habitats (Watts et al. 2020). The distribution of this warm water species ranges from North Carolina, throughout the Gulf of Mexico, and to southern Brazil (Moore et al. 1963). Due to its abundance and distribution, *L. variegatus* is a focal species for marine ecology and environmental studies, and a popular model for developmental biology because of its high fecundity, translucent embryos, rapid development, amenability to experimental perturbation, and well-studied developmental gene regulatory network (GRN). *Lytechinus variegatus* belongs to the Camarodonta, a large clade of primarily shallow-water sea urchins whose crown group dates to ∼100 Ma (Smith 2005) and includes several other species widely studied by ecologists, developmental biologists, and evolutionary biologists.

Over the past decade, there has been a major impetus to increase the availability of genomic resources available for echinoderms, spearheaded by the late developmental biologist Dr. Eric H. Davidson (reviewed in Cameron et al. 2015). Among echinoderms, *L. variegatus* is among the most commonly studied species, with 45 articles focusing on this species published in the past 5 years (PubMed title and abstract search; accessed March 6, 2020). Although there have been efforts to assemble the genome of *L. variegatus* using Illumina short-read sequencing technology (www.echinobase.org: *Lvar_2.2*; accessed March 14, 2020), a high-quality genome assembly is lacking. Here, we report an annotated, chromosome-level reference assembly for *L. variegatus* (*Lvar_3.0*) which will serve as a powerful genomic resource for the research community.

Because several functional genomic data sets exist for *L. variegatus*, we also took the opportunity to investigate the impact of a high-contiguity, well-annotated reference assembly on read mapping and quantifying genomic features. We report quality and informatic metrics from aligning bulk RNA-seq, single-cell RNA-seq, and ATAC-seq reads to our new assembly, *Lvar_3.0*, and to the older *Lvar_2.2* assembly. These results provide valuable information to investigators when deciding how much resources to invest in their own genome projects.

Materials and Methods

Tissue Collection and DNA Extraction

Lytechinus variegatus adults were collected near the Duke University Marine Lab in Beaufort, North Carolina, USA. The interpyramidal muscle of Aristotle's lantern (the sea urchin's feeding apparatus), tube feet, and the ovarian tissue were dissected from a single female individual for DNA extraction and sequencing.

Karyotyping

To make chromosome preparations, 16- and 32-cell stage embryos of *L. variegatus* were suspended in filtered seawater and colchicine (1 mg/ml) for 45 min, then filtered through a Nylon cell strainer (100 μm; Falcon) to remove the fertilization membrane and detach cell clusters. Cells were centrifuged to form a pellet and resuspended in a hypotonic solution of sodium citrate (7% w/v) for 5 min. Cells were fixed using Carnoy's fixative (ethanol and glacial acetic acid, 3:1), and chromosome slides were prepared using a dropping technique (Camargo et al. 2006). Images were captured using a Zeiss Axioskop with a 64× oil immersion objective and analyzed in imageJ2 (Rueden et al. 2017).

DNA Sequencing

A third-generation DNA library was sequenced on a PacBio sequel II CLR platform, generating 105.8 Gb of data with an N50 read length of 26.2 kb. DNA from the same individual was also used to construct linked-reads (10× Genomics) and Hi-C libraries, which were sequenced on a BGI-SEQ 500 platform, generating 188.53 and 204.42 Gb of data, respectively. Jellyfish v2.2.6 (Marcais and Kingsford 2011) and GenomeScope v1.0.0 (Vurture et al. 2017) were deployed to conduct a k-mer-based survey of genome composition using linked-read sequencing data based on 17-mer frequency distribution to estimate the genome size and heterozygosity of *L. variegatus* (supplementary fig. 1, Supplementary Material online).

Assembly Strategy

PacBio sequencing data were employed to assemble a de novo contig-level genome assembly using Canu v1.8 (minReadLength = 1,200; minOverlapLength = 1,000) (Koren et al. 2017). Subsequently, HaploMerger2 v3.6 (Huang et al. 2017) was used to create breakpoints in the contigs where potential misjoins have occurred by aligning allelic contigs via Lastz v1.02.00 (Harris 2007). From these fragmented contigs, the longest of each allelic pair was identified and selected using Redundans v0.14a (Pryszcz and Gabaldon 2016), resulting in a near-haploid level genome assembly. The output of this pipeline was polished using Pilon v1.23 (Walker et al. 2014) with 10× sequencing data to improve assembly quality and accuracy at single-base resolution. Lastly, contigs were assembled into scaffolds by mapping Hi-C read pairs to the polished assembly with HiC-Pro (Servant et al. 2015), resulting in ∼33.34% valid Hi-C reads pairs. Juicer v1.5 (Durand et al. 2016) and 3D-DNA v180419 (Dudchenko et al. 2017) were used to correct and finalize the construction of *L. variegatus* chromosomes (supplementary fig. 2, Supplementary Material online).

Repeat Identification

Genomic repetitive elements were identified with RepeatModeler v2.01 (Smit and Hubley 2010) to generate an *L. variegatus*-specific repeat element library. Prior to masking, repeat libraries were filtered via BlastN v2.3.0 (Camacho et al. 2009) for significant hits to gene models of the well-studied sea urchin *Strongylocentrotus purpuratus* (www.echinobase.org/Echinobase/SpAbout, accessed February 26, 2020) to prevent unintentional masking of genic regions (91 of 2,377 repeat families removed). Repeats were mapped to the genome with RepeatMasker v4.0.9 (Smit et al. 2015) using the most sensitive setting (-s) to identify the location of

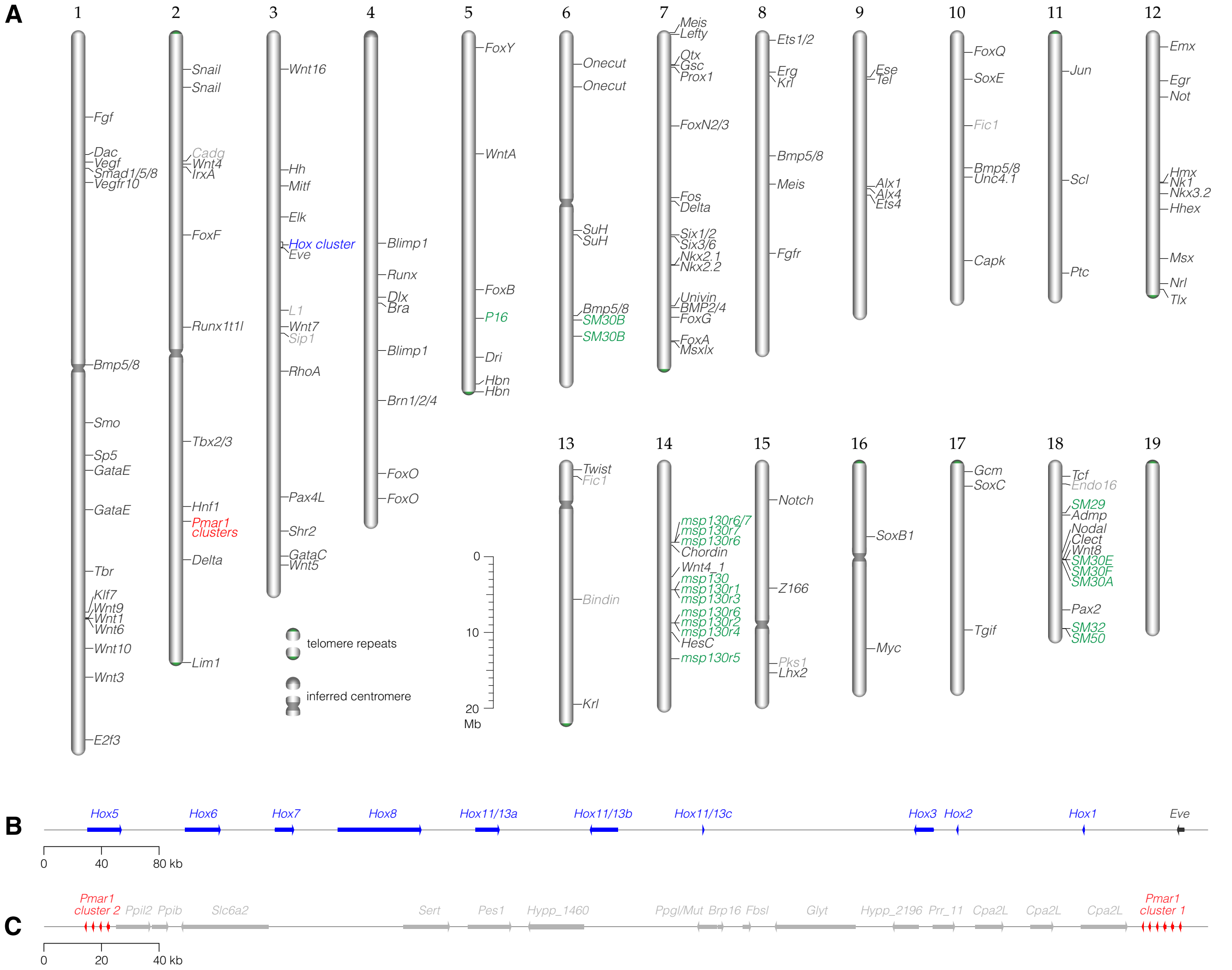

Fig. 1.—Ideogram and distribution of developmental genes in *Lytechinus variegatus*. (*A*) Physical distribution of developmentally significant genes. Black, regulatory genes encoding components of the GRN (gene regulatory network); blue, genes encoding proteins of the biomineral matrix of the endoskeleton; green, *Hox* genes; red, *Pmar1* genes; gray, other genes. Inferred centromere location is shown for seven chromosomes (see Materials and Methods and supplementary table 1, Supplementary Material online, for criteria). (*B*) *Hox* cluster organization (chromosome 3). (*C*) The two clusters containing ten *Pmar1* paralogs (chromosome 2). In panels (*B*) and (*C*), arrowheads indicate direction of transcription and colors are the same as in panel (*A*).

repetitive elements. As centromeres and flanking DNA tend to be very gene-poor, repeat-rich, and difficult to assemble, we predict candidate centromere locations for a subset of chromosomes based on regions with low gene density and high repeat density (supplementary fig. 3, Supplementary Material online; see fig. 1*A*) as described in Weighill et al. (2019). These regions often correspond with increased density of gap sequence as well.

Gene Annotation and Prediction Strategy

Repetitive regions were soft-masked prior to gene annotation and prediction. The developmental transcriptome of *L. variegatus* was retrieved from Israel et al. (2016) and curated with EvidentialGene (Gilbert 2013) under default parameters to remove duplicate sequences and select the longest open reading frame for transcripts representing the same gene. In addition, UniProt Swiss-Prot (Bateman 2019) and *S. purpuratus* v5.0 protein models (www.ncbi.nlm.nih.gov/assembly/GCF_000002235.5, accessed February 26, 2020) were incorporated in the gene annotation and prediction pipeline.

Maker v2.31.10 (Campbell et al. 2014) was used to create an initial set of gene annotations by incorporating the previously mentioned resources under default parameters, except split_hit = 20,000. Augustus v3.3.3 (Stanke et al. 2006) and SNAP v2006-07-28 (Korf 2004) gene prediction tools were trained on high-confidence gene models from this run, plus an additional 1,000 bp of 5′ and 3′ flanking sequence. Trained gene prediction parameters from these two programs were incorporated into a second round of annotation with Maker to generate improved gene models. This cycle was repeated iteratively until gene model quality stopped improving (four iterations).

Assembled protein models were functionally annotated using BlastP v2.3.0 (Camacho et al. 2009) with three pre-existing protein databases: *S. purpuratus* v4.2 protein models (models frequently referenced by the sea urchin development community), UniProt Knowledgebase Swiss-Prot protein models (Bateman 2019), and RefSeq invertebrate protein models (O'Leary et al. 2016) with *S. purpuratus* excluded (e-value: 1e-5). In addition, protein domains were identified and annotated using InterproScan v5.38-76.0 (Jones et al. 2014). These annotations are available in supplementary data 1, Supplementary Material online. Lastly, BUSCO v3.0.2 (Seppey et al. 2019) was utilized as a measure of the completeness of the genome assembly (parameters: –long, –db meatazoa_odb9).

RNA-Seq and ATAC-Seq Analyses

Three replicates of bulk RNA-seq data from *L. variegatus* 32-cell embryos and prism stage larvae (retrieved from Israel et al. [2016]) were mapped to gene models of each assembly (*Lvar_3.0* and *Lvar_2.2*) using bowtie2 v2.3.5.1 (parameter: –very-sensitive) (Langmead and Salzberg 2012). These alignments were quantified via Salmon v0.14.1 (parameters: –seqBias, –gcBias) (Patro et al. 2017). Single-cell RNA-seq reads from a 10× Chromium data set of 3- and 24-h postfertilization *L. variegatus* embryos were mapped with STAR (Dobin et al. 2013) via the Cell Ranger v3.0.1 software suite. Lastly, three replicates of 32-cell and two-arm pluteus larvae ATAC-seq data were aligned with Stampy v1.0.28 (parameter: –sensitive) (Lunter and Goodson 2011). ATAC-seq alignments were filtered for *L. variegatus* mitochondrial sequences (Bronstein and Kroh 2019) and required an alignment quality score of at least 5 using samtools v1.9 (Li et al. 2009). Peaks were called from these filtered alignments using the MACS2 v2.1.2 (Zhang et al. 2008) *callpeak* function (parameters: –nomodel, –keep-dup=auto, –shift 100, –extsize 200).

Results and Discussion

Genome Assembly

The *Lvar_3.0* genome assembly contains 19 chromosome-scale scaffolds with an N50 length of 45.6 Mb, which cumulatively constitute over 99.9% of the 870.4 Mb assembly. The chromosomal-scale scaffolds range in size from 23.4 to 96.7 Mb. Moreover, we found that the number of scaffolds >20 Mb (19) matches the haploid mean number of chromosomes from our karyotypic analyses (supplementary figs. 4 and 5, Supplementary Material online). The scaffolded genome assembly is comprised 466 contigs (N50 length: 5.9 Mb), 180 kb of total gap sequence (0.02% of assembly), and has a GC content of 36.1%. This represents a substantial improvement upon the most recently available annotated *L. variegatus* genome assembly *Lvar_2.2*, which has a scaffold N50 of 46.3 kb, contig N50 of 9.7 kb, and 5.4% of unresolved (*N*) gap sequence. Repetitive elements make up an estimated 45.4% of the *Lvar_3.0* assembly but vary from 38.0% to 56.3% across the 19 chromosomes (supplementary table 1, Supplementary Material online). The completeness of the assembly is reflected in a BUSCO "complete" score of 95.5%, including only 0.6% duplicated hits. About 85 scaffolds with an average length of 10.2 kb could not be assigned to any of the 19 chromosomal scaffolds. See table 1 and supplementary figure 6, Supplementary Material online, for a detailed summary of genome assembly statistics, including a comparison with the *Lvar_2.2* assembly.

Gene Annotation

We identified 27,232 genes with an average length of 12.6 kb (including UTR regions), which cumulatively make up 39.5% of the genome. An average of 31.5 genes per Mb were annotated on each chromosome, but this figure varied from 28.5 (chromosome 3) to 33.6 (chromosome 14) (supplementary fig. 7, Supplementary Material online), indicating an unequal distribution of genic content among the chromosomes. Of the 27,232 gene models, 93.4% had an identifiable start (ATG) codon, whereas 90.3% had both a start codon and a stop codon (TGA|TAA|TAG). 24,886 (91.4%) gene models had significant hits to *S. purpuratus* protein models, 19,312 (70.9%) to Uniprot Swiss-Prot proteins, and 22,130 (81.3%) to non-*S. purpuratus* RefSeq invertebrate proteins. 1,553 (5.7%) gene models did not have a significant hit to any of these three protein databases. In addition, 24,169 (88.8%) gene models had a significant hit to the InterproScan suite of protein function and domain databases, including 18,872 (69.3%) with a hit to Pfam protein families (El-Gebali et al. 2019). 26 genes were modeled on 18 of the 85 unplaced scaffolds, but were of poor annotation quality (average length: 1.7 kb).

Chromosomal Mapping of Key Developmental Genes

The ability to map gene locations at the scale of entire chromosomes reveals several noteworthy features of the *L. variegatus* genome. We mapped >100 developmental regulatory genes (fig. 1*A*, red). Of the 12 genes encoding *Wnt*family ligands, five are located within a <90 Mb region on chromosome 1. The ten *Hox* genes form a tight cluster within a region of <700 kb that is devoid of other genes, as in many other marine invertebrates. Although the order of *Hox* genes is highly conserved among metazoan phyla, an inversion of the first three genes is present in the sea urchin *S. purpuratus* (Cameron et al. 2006). We find that this inversion is present in *L. variegatus* as well (fig. 1*B*), suggesting that it predates the origin of the Camarodonta. The *Pmar1* gene

Table 1
Comparison of *Lvar_3*.0 and *Lvar_2.2* Assemblies

	Lvar_2.2	Lvar_3.0	Fold-Change
Assembly			
Assembly size	1061.2 Mb	870.4 Mb	−1.22
No. scaffolds	322,794	104	−3103.79
N50 scaffold length	0.046 Mb	45.6 Mb	991.30
Longest scaffold	0.55 Mb	96.7 Mb	175.82
No. scaffolds >10 Mb	0	19	NA
No. contigs	452,418	466	−970.85
N50 contig length	0.0097 Mb	5.85 Mb	603.09
No. contigs >1 Mb	0	285	NA
N (%)	5.38	0.02	−269.00
GC (%)	36.34	36.31	−1.00
BUSCO			
Complete	86.4	95.5	1.11
Duplicated	6.3	0.6	−10.50
Fragmented	7.2	0.8	−9.00
Missing	6.4	3.4	−1.88
Annotations			
No. genes	22,105	27,232	1.23
Average gene length	7.7 kb	12.6 kb	1.64
% Start codon	75.8	93.4	1.23
% Start and stop codon	41.4	90.3	2.18
Bulk RNA-seq			
% Reads aligned to gene model	28.7	42.9	1.49
Average counts per sample	10,061,423	15,052,961	1.50
Mean counts per transcript	1,673.6	3,315.9	1.98
scRNA-seq			
% Reads mapped to exons	10.3	27.0	2.62
Median genes per cell	878	1520	1.73
Median UMIs per cell	1,579	3235	2.05
Total cells	1,002	1038	1.04
ATAC-seq			
% Reads mapped to nuclear genome	39.11	38.0	−1.03
No. peaks	103,937.0	65,263.5	−1.59
Average pileup per peak	21.1	23.4	1.11
% of peaks on scaffold with gene	70.4	99.9	1.42

Note.—Assembly summary statistics (top) and analysis metrics of three functional genomics data sets (bottom) mapped onto *Lvar_3.0* and the previous assembly, *Lvar_2.2*. For each functional genomics analysis, numbers reflect the sample average from two developmental timepoints: 32-cell embryo and early stage larva.

family, which encodes a paired-box transcription factor, is represented by ten tandemly repeated genes in two tight clusters separated by ~432 kb and 17 unrelated genes on chromosome 2 (fig. 1*C*). Although expression of *Pmar1* is critical for cell fate specification in the early embryo (Oliveri et al. 2002), its high copy number has prevented an accurate reconstruction of the number and organization of orthologs in any sea urchin species until now. Also of interest are genes encoding proteins associated with the calcite endoskeleton, an autapomorphy for the phylum (Brusca and Brusca 2003), that are highly expressed in the developing larva (fig. 1*A*, blue). The ten genes of the *msp130* family that encode a Ca2+ ion transporter critical for skeletogenesis are all located on chromosome 14 in three small tandem clusters and one singleton. Six genes encoding skeletal matrix proteins (*SM* family) are located on chromosome 18, with two highly similar copies on chromosome 6. Genomic organization thus suggests that nonhomologous recombination was primarily responsible for the expansion of the *Pmar1*, *msp130*, and *SM* gene families.

Comparison of Functional Genomics Data between *Lvar_3.0* and *Lvar_2.2*

We aligned three types of functional genomics data (bulk RNA-seq, single-cell RNA-seq, and ATAC-seq) to this genome assembly as well as to *Lvar_2.2* to compare how chromosomal-scale assemblies improve acquisition of biological information over less contiguous ones. For each data type, two developmental stages (32-cell embryo and early larva) were analyzed under identical parameters (results summarized in table 1). First, whole embryo RNA-seq was aligned to each assembly's gene models (CDS+UTR). On an average, read mapping rate and counts per sample increased by nearly 50% and mean counts per transcript almost doubled by mapping to the gene models of the *Lvar_3.0* assembly over the older *Lvar_2.2* models. Next, scRNA-seq data from these two stages were aligned to each genome assembly. Relative to *Lvar_2.2*, mapping to *Lvar_3.0* increased exon mapping rate by 16.7% (2.6-fold increase), which contributed an 1.7-fold increase in median genes per cell, a 2.0-fold increase in median UMIs (unique molecular identifiers) per cell, and a 3.6% increase in total number of cells confidently identified. Lastly, ATAC-seq data were mapped to the nuclear genomes from each assembly, which resulted in slightly higher alignment rate (+2.7%) and many more peaks (+37.2%) being called in *Lvar_2.2* relative to *Lvar_3.0*. This is likely an artifact of the larger assembly size, higher duplication rate, and increased fragmentation of the *Lvar_2.2* assembly. Consistent with this, ATAC-seq data aligned to *Lvar_3.0* showed a significant signal-to-noise improvement with fewer peaks and marked increase in the average read pileup per peak (+10.9%). Importantly, 99.99% of peaks were located on a scaffold with a gene model (as opposed to 70.4% in *Lvar_2.2*).

These results demonstrate that a high-quality, chromosomal-scale genome assembly can add substantial value to functional genomic data sets—in this case, more than doubling some informatic metrics. Although there are trade-offs between cost, sequencing strategy, and ultimately genome assembly quality, researchers will likely reap long-term value from a larger initial investment in a high-quality genome assembly, particularly in cases where additional population and functional genomics data sets for the organism are anticipated. Moving forward, as sequencing technologies and assembly strategies improve, this trade-off will be increasingly important to researchers considering the assembly or reassembly of more species' genomes.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by National Institutes of Health (HD14483 to D.R.M.) and National Science Foundation Division of Integrative Organismal Systems (1929934 to G.A.W.). We thank Esther Miranda for assistance with karyotyping of sea urchin cells and the Wray Lab members for reviewing the manuscript.

Author Contributions

L.W., D.R.M., G.F., and G.A.W. designed and conceived the study. L.W. and Y.C. collected tissue samples and performed the DNA extraction. P.L.D., H.G., and H.Z. assembled the genome. A.B. performed the karyotyping. A.L.S. carried out the repeat identification. P.L.D. performed the gene modeling and prediction. P.L.D. and G.A.W. performed the gene annotations. P.L.D. performed the functional genomics analyses. P.L.D. and G.A.W. wrote the article.

Literature Cited

Bateman A. 2019. Uniprot: a universal hub of protein knowledge. Protein Sci. 28:32–32.

Bronstein O, Kroh A. 2019. The first mitochondrial genome of the model echinoid *Lytechinus variegatus* and insights into Odontophoran phylogenetics. Genomics 111(4):710–718.

Brusca GJ, Brusca RC. 2003. Invertebrates. Sunderland (MA): Sinauer Associates Inc.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.

Camargo M, Duque-Correa MA, Berrio A. 2006. A micro-spreading improvement for spermatogenic chromosomes from Triatominae (Hemiptera-Reduviidae). Mem Inst Oswaldo Cruz. 101(3): 339–340.

Cameron RA, et al. 2006. Unusual gene order and organization of the sea urchin hox cluster. J Exp Zool B Zool. 306B(1):45–58.

Cameron RA, Kudtarkar P, Gordon SM, Worley KC, Gibbs RA. 2015. Do echinoderm genomes measure up? Mar Genomics. 22:1–9.

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

Dobin A, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21.

Dudchenko O, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356(6333):92–95.

Durand NC, et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3(1):95–98.

El-Gebali S, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427–D432.

Gilbert D. 2013. Gene-omes built from mRNA-seq not genome DNA. 7th Annual Arthropod Genomics Symposium. Notre Dame, IN.

Harris RS. 2007. Improved pairwise alignment of genomic DNA [PhD thesis]. State College (PA): Pennsylvania State University.

Huang S, Kang M, Xu A. 2017. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. Bioinformatics 33(16):2577–2579.

Israel JW, et al. 2016. Comparative developmental transcriptomics reveals rewiring of a highly conserved gene regulatory network during a major life history switch in the sea urchin genus *Heliocidaris*. PLoS Biol. 14(3):e1002391.

Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240.

Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5(1):59.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–U354.

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079.

Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21(6):936–939.

Marcais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6): 764–770.

Moore HB, Jutare T, Bauer J, Jones J. 1963. The biology of *Lytechinus variegatus*. Bull Mar Sci. 13:23–53.

O'Leary NA, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44:D733–D745.

Oliveri P, Carrick DM, Davidson EH. 2002. A regulatory gene network that directs micromere specification in the sea urchin embryo. Dev Biol. 246(1):209–228.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 14(4):417–419.

Pryszcz LP, Gabaldon T. 2016. 016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44(12): e113.

Rueden CT, et al. 2017. ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics 18(1):529.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol. 1962:227–245.

Servant N, et al. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16(1):259.

Smit AFA, Hubley R. RepeatModeler Open-1.0. 2008–2015. Available from: http://www.repeatmasker.org. Accessed February 6, 2020.

Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. Available from: http://www.repeatmasker.org. Accessed February 6, 2020.

Smith AB. 2005. The pre-radial history of echinoderms. Geol J. 40(3):255–280.

Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34(Web Server):W435–W439.

Vurture GW, et al. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33(14):2202–2204.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.

Watts S, McClintock JB, Lawrence JM. 2020. *Lytechinus*. In: Lawrence JM, editor. Sea urchins: biology and ecology. San Diego (CA): Academic Press. p. 661–680.

Weighill D, et al. 2019. Wavelet-based genomic signal processing for centromere identification and hypothesis generation. Front Genet. 10:487.

Zhang Y, et al. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9(9):R137.

Associate editor: Adam Eyre-Walker