HHS Public Access

Author manuscript

Mar Genomics. Author manuscript; available in PMC 2016 August 01.

Published in final edited form as:

Mar Genomics. 2015 August ; 22: 1–9. doi:10.1016/j.margen.2015.02.004.

Do Echinoderm Genomes Measure Up?

R. Andrew Cameron[a], **Parul Kudtarkar**[a], **Susan M. Gordon**[a], **Kim C. Worley**[b], and **Richard A. Gibbs**[b]

[a]Division of Biology 139-74, California Institute of Technology, Pasadena, CA USA

[b]Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine Houston, TX USA

Abstract

Echinoderm genome sequences are a corpus of useful information about a clade of animals that serve as research models in fields ranging from marine ecology to cell and developmental biology. Genomic information from echinoids has contributed to insights into the gene interactions that drive the developmental process at the molecular level. Such insights often rely heavily on genomic information and the kinds of questions that can be asked thus depend on the quality of the sequence information. Here we describe the history of echinoderm genomic sequence assembly and present details about the quality of the data obtained. All of the sequence information discussed here is posted on the echinoderm information web system, Echinobase.org.

2. INTRODUCTION

Sea urchin gametes and embryos occupied front row seats for many of the innovations that propelled cell and developmental biology over the last 175 years. Using the low resolution microscopes of his day, Derbe (1847) demonstrated the necessity of sperm for development to ensue but he couldn't see sperm-egg fusion. By the 1880s the phase contrast microscope was used to observe pronuclear fusion in sea urchin zygotes (Hertwig, 1876). The requirement of a complete set of chromosomes for development emerged from experiments on sea urchins in the early 1900's (Boveri, 1901). As developmental biologists began to examine cell lineages in embryos, the importance of intercellular communication in development grew out of blastomere recombination experiments in a Mediterranean sea urchin (Horstadius, 1939). The term chemical biology was coined mid-20th century to describe the innovations stemming from the use of cell fractionation by ultracentrifugation and allied techniques which again took advantage of the copious amounts of sperm, eggs and embryos available from sea urchins (Brachet, 1950). The advent of biological radionuclides afforded an opportunity to dissect the mechanisms of DNA replication, transcription and translation during this period and soon thereafter (reviewed in Davidson,

Corresponding author: R. Andrew Cameron Beckman Institute 139-74 California Institute of Technology 1200 East California Blvd. Pasadena, CA 91125 acameron@caltech.edu.

1968). Then, solution hybridization using DNA from sea urchin and other easily available sources became a favorite technique to explore genome structure and the mechanisms of gene expression (Britten and Davidson, 1969). The establishment of recombinant DNA technology that followed launched efforts to understand the mechanisms of gene regulation in development (Davidson, 1968). As molecular biology studies expanded, the sea urchin became a favored system for gene transfer (McMahon, et al, 1984; Colin, 1986). It seemed remarkable that naked DNA constructs could be injected into zygotes where they were amplified along with nuclear DNA and were expressed in a manner identical to the exogenous sequences (Flytzanis et al, 1985).

By the end of the 20th century the catalog of expressed genes was extensive and the focus of gene expression studies had come to lie on the interactions between genes by means of the cis-regulatory modules that control them. In parallel, a community enterprise arose to support the sequencing of the purple sea urchin genome. It was realized that genome assemblies would be ultimately required to fully describe the intricate gene regulatory networks that drive development (http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/SeaUrchin_Genome.pdf).

It is the purpose of this essay to detail the series of sequencing activities that bring us to the assemblies of multiple echinoderm genomes available today. It relates the history over about 10 years of the efforts to construct an accurate draft genome for the purple sea urchin and the rapid expansion in additional species brought about by the disruptive technology of next-generation sequencing. In the process, we hope to give a sense of the experimental nature of the process of genome sequencing and assembly as well as the intellectual expansions and technical limitations that the quality and extent of the genomic information provide.

> “Because of the small number of people producing this resource relative to the large number using it, the nature of the data is, unfortunately, not commonly appreciated …..” (Mardis et al, 2002).

As Elaine Mardis says, the relatively solitary nature of genome sequencing efforts impedes a general appreciation for the quality of the data. Perhaps this essay will remedy this for echinoderm genomes.

3. THE ECHINODERM PHYLUM

3.1 Phylogeny

Echinoderms are bilaterian animals even though their adult body plans exhibit pentameral symmetry. The larval stages are definitely bilateral. Based on embryonic feature and recent molecular data, echinoderms occupy the same branch of the bilaterian tree as the chordates. Together with the hemichordates they form the Ambulacraria which is the sister group to the chordates. Of the five classes of echinoderms, four are the free-living eleutherozoans: echinoids (sea urchins), holothuroids (sea cucumbers), asteroids (sea stars), and ophiuroids (brittle stars). The mouth faces the substrate in these forms while the fifth class, the crinoids, has the mouth on the top surface. There have been two competing hypotheses about the relationships among the eleutherozoan classes. Two recent reports utilizing transcriptome data favor the Asterozoa topology where the asteroids and ophiuroids are a sister group to

the holothuroids and echinoids (Telford et al., 2014; Reich et al, 2014) (Figure 1). The lack of resolution of these relationships until recently is probably due to a paucity of molecular data for some classes and to the rapid divergence of the groups (Pisani et al, 2012). The interval over which they are estimated to diverge is only about 35 million years in the Cambrian period about 500 million years ago.

That the phylogenetic relationships of echinoderm groups extend into deep evolutionary time offers an opportunity to examine histories of changes at a level available in few other places among the bilaterians. Comparisons of genomic structure among these animals have the capacity to reveal the milestones of genomic change that accompany the divergence of echinoderm classes. A common feature of echinoderms is a particular form of skeleton, the stereome, which is found in all of the adult forms. The development of this unique structure thus extends backward 540 million years (Bottjer et al, 2006). The way in which the structural gene batteries and developmental gene regulatory networks may have changed is intriguing. Only sea urchins and brittle stars have prominent skeletal elements in embryonic stages. (Sea cucumbers have small spicules in the developmental stages. These are likely homologous to the sea urchin ones.) Considering the asterozoan topology these structures are either a result of convergent evolution or existed in the common ancestor of the four eleutherozoan classes and were lost in asteroids.

The data are still scarce but one study found no skeletal matrix proteins shared between the well-studied sea urchins and an ophiuroid (Vaughn, et al, 2012). This observation leans the inference toward convergent evolution of larval skeletons.

3.2. Echinoderm sequencing candidates

Representative members of the echinoderm classes were chosen for genome sequencing to complement ongoing research and address some of the evolutionary topics detailed above (Table 1). Due to the extensive body of work on molecular mechanisms of cell and developmental biology, the purple sea urchin, *Strongylocentrotus purpuratus* (Sp) was chosen as the first subject for sequencing. There already existed a suite of resources for genomic studies in this species in the form of arrayed cDNA and genomic DNA libraries (Cameron et al, 2000). An informal network of investigators supported this first project. The cidaroid sea urchin *Eucidaris tribuloides* (Et), is diverged from the reference species by 255 MY and exhibits interesting differences in the mode of skeletal formation. The variegated sea urchin, *Lytechinus variegatus* (Lv) from the east coast of North America is diverged from the common ancestor of the purple sea urchin by about 50 MY. It has been used as a research model for many years and has recently been shown to provide genomic comparisons that reveal conserved non-coding sequences likely to be sites of transcriptional control of protein coding genes.

Based on comparison between five functionally characterized cis-regulatory modules (CRM) from the *S. purpuratus* genome and orthologous regulatory and flanking sequences obtained from a bacterial artificial chromosome genome library of a congener, *S. franciscanus* (Sf), it was observed that large indels are statistically almost absent from cis-regulatory modules at this evolutionary distance of about 20 MY (Cameron et al, 2005a). This metric though probabilistic could be used to help characterize CRMs and it was

decided to sequence the genomes of two species at this close evolutionary distance. Therefore, *S. franciscanus* and *Allocentrotus fragilis* (Af), were selected for limited sequencing.

In order to obtain a broad perspective of echinoderm genomes, representatives of the other classes were also selected (Table 1). Many genomic resources have accumulated for these species as well (Cameron et al, 2000; http://Echinobase.org). The bat star *Patiria miniata* (Pm) is easily available along the Pacific coast of California and is the subject of studies in early development (Hinman et al., 2003). The sea cucumber, *Parastichopus parvamensis* (Pp) is obtainable in the same areas as the bat star. Transcriptomes for this holothuroid have been described (McCauley et al., 2012). Some preliminary work has also been done on the brittle star, *Ophiothrix spiculata* (Os), in southern California. These three species have been included on the roster for sequencing at The Baylor College of Medicine, Human Genome Sequencing Center (BCM-HGSC).

It is notable that these candidates share with the other echinoderms several features that render sequence assembly difficult. They have large genomes that vary from one quarter to several times the size of the human genome. They have a large number of low frequency repeat classes which confound assembly. Solution hybridization experiments reveal an intraspecies genome sequence variation for the purple sea urchin of about 4% (Britten et al, 1978). (This is 50 times the amount of variation found in typical human nuclei which show about 0.1% variation between the two genomes they contain [Antonarakis, 2010]). Recent comparisons of assembled BAC sequences in Sp show that the ratio of base changes due to insertions and deletions versus SNPs is about 3:1 (Britten et al, 2003). It will be necessary to overcome these assembly difficulties in order to make experiments based on non-coding sequence features like CRM characterization efficient and reliable.

4. SEQUENCING THE REFERENCE GENOME, *Strongylocentrotus purpuratus*

The sequencing and assembly of the purple sea urchin genome has been both an end in itself and an opportunity to experiment with emerging strategies for obtaining and assembling large polymorphic genomes. Beginning with a round of whole genome shotgun (WGS) Sanger sequencing, a total of 5 separate episodes of sequencing has been conducted over the years using the DNA from the same single male purple sea urchin. The BCM-HGSC has conducted the individual rounds and assembled the sequences. The first assembly of the whole genome shotgun sequences was posted to Genbank on April 15, 2005 as Spur_0.5. The Center completed the assembly of about 7 million reads taken from several WGS libraries with inserts of 2-6 kb. The quality of the 0.5 assembly version (Table 2) was deemed sufficient to support a gene prediction effort. An estimated gene number of 23,300 (Sea Urchin Sequencing Consortium, 2006) and a genome size of 800Mb (Hinegardner, 1974) yields an intergenic distance of 26kb and in general an individual gene is considered to span about 10kb. For this version 0.5 the N50 of contigs over 1Kb is 10.2 Kb (8 Kb for all contigs) and the number of contigs to N50 is ~19,000 (Sodergren et al, 2006). (Half (50%) of the genome is in assembled sequence pieces of the N50 or greater (90% in pieces >=N90). Thus the majority of genes ought to be contained in usable sequence fragments.

The WGS sequence of version 0.5 produced an assembly with about 15% redundancy when compared to a set of high quality BAC sequences. This redundancy was due in part to the highly polymorphic nature of the sea urchin genome which adds to the difficulty of assembling large genomes (see above). Essentially, regions of assembled reads from each allele appear sufficiently different to the assembly program that it considers them independent fragments. To overcome this difficulty the BCM-HGSC added a BAC sequencing strategy to the mix (Sodergren et al, 2006; Sea Urchin Sequencing Consortium, 2006). Since each BAC insert is of one haplotype, the reads will be partitioned to different BACs and the subsequent fragments will collapse since they are for the most part similar. Using restriction enzyme fingerprinting, a minimum set of BAC clones covering the genome was produced at the Michael Smith Genome Sciences Center in Vancouver and then sequenced to 2X genome coverage in a pooled strategy (Cai et al, 2001; Sea Urchin Sequencing Consortium, 2006). The reads from this set of ~8,000 BACs were deconvolved and the reads were then assembled into individual BACs. An assembler, ATLAS, was developed at BCM-HGSC specifically to handle WGS and BAC sequences (Havlak et al, 2004). The BAC assemblies were enriched with WGS reads and then combined with the BAC end sequences to form bactigs. Paired end sequences from two sets of BAC inserts ranging in size from 30-40 Kb and 130-160 Kb, respectively were also collected and used in the assembly. The subsequent assembly yielded a reduction in the redundancy to about 5% by the previously mentioned measure. Furthermore, the scaffold N50 more than doubled and the contig N50 improved somewhat (Table 2; compare 0.5 and 2.1). The addition of these large BAC inserts and paired BAC end sequences supported the ordering of contigs into scaffolds more readily than lengthening contigs. By aligning sea urchin sequences from the EST database at NCBI it was shown that at least 95% of the genome was represented in this assembly. This is the genome assembly labeled Spur_2.1 (Table 2) submitted to Genbank. It is the version described in the initial sea urchin genome paper in Science (Sea Urchin Sequencing Consortium, 2006).

The various parts of the previous sequencing efforts were all done with Sanger sequencing. The next increment took advantage of the SOLiD sequencing-by-ligation technology commercialized by Applied Biosystems (see Mardis, 2008 for review). The paired-end SOLiD reads provided 18X coverage of the genome. 500 million reads had a length of 25 bp and 46 million were 50 bp long. Of the 273 million clones, 30 million or 11% mapped uniquely to the genome. The 13% (4 million) of the uniquely mapping pairs that span two different scaffolds were used to improve the assembly. From the statistics for version 2.6 (Table 2), one sees that the scaffold N50 increases significantly (from 123 Kb to 168 Kb) while the contig N50 was essentially unchanged. The small proportion of reads that align to the previous contigs is probably due to a combination of the high polymorphism, the error rate and the short length of the SOLiD reads. Nevertheless, an improvement in assembly quality resulted and this version was submitted as Spur_2.6 to Genbank.

The next increment of sequencing used the Illumina platform to produce a genome coverage of 40X using reads with different spacing. Called rainbow libraries, the collection of sequencing libraries consists of a fragment paired end with ~300 bp inserts and mate-pair libraries with 1 Kb, 3Kb and 5-6Kb inserts. The reads were mapped to the Spur_2.6 genome assembly and then used to bridge existing scaffolds as well as gap filling from local

alignments. The scaffold N50 improved from 168Kb to 402 Kb and the contig N50 went from 11.5 Kb to 13.5 Kb. This version is labeled Spur_3.1 and the various versions are listed at Genbank under BioProject PRJNA10736.

All of these assemblies were made with the Atlas suite of software tools developed at BCM-HGSC. They were specifically designed to assemble large genomes from combinations of whole genome shotgun strategies and BAC-based strategies. The tools include Atlas-GapFill which maps reads to gaps and then assembles them locally with a variety of assemblers like Newbler, Phrap and Velvet. Another associated tool is Atlas-Link which rapidly orders genome contigs using mate-pair information. Documentation and downloads are available at https://www.hgsc.bcm.edu/software

An as yet un-submitted assembly of purple sea urchin (version 4.0) adds reads from SMRT Sequencing System from Pacific Biosciences (PacBio) to the mix. A total of 8.5 Gb of read sequence yielding about 11X genome coverage was produced and the PBJelly2 program was used to add these to the assembly. The contig N50 increased from 13.5 kb to 17.6 kb. The scaffold N50 increased from 402 kb to 431 kb.

Each increment of genome sequencing added to the existing data improved the assembly in a slightly different way (Table 2). The addition of BAC sequences in version 2.1 dealt with the problems brought on by the high polymorphism in the Sp genome and reduced the sequence redundancy from 15% to 5% (Sodergren et al., 2006). This is reflected in the decrease of total sequence length by 10%. Simultaneously contig and scaffold numbers were decreased and sizes increased as the BAC sequences served as a platform for larger assembled fragments. Even though only a small portion of the SOLiD sequence could be mapped to the previous assembly, the size of scaffolds increased by about 25%. The nearly identical total sequence value between version 2.1 and version 2.6 infers that the existing sequence data and assembly software have reached the limit in the capacity to reduce the redundancy due to polymorphism. The addition of the Illumina sequence made a huge improvement in scaffold size and number with much less change in the contig values. Thus the relatively short, paired end sequences from the Illumina platform mostly joined contigs to improve the scaffolding. The long read Pacific Biosciences data is effective at spanning and filling gaps within scaffolds and can also join existing pieces into larger scaffolds.

5. ADDITIONAL ECHINODERMS WITH DRAFT GENOME ASSEMBLIES

The economy of next-generation sequencing opened the way for additional echinoderm genome sequencing efforts (Table 1). The four eleutherozoan classes of echinoderms diverged rapidly over a 35 million year span about 500 million years ago. This well characterized group of animals with good fossil records back to the beginning of the Cambrian exhibit a variety of developmental modes. Comparative analysis of the gene regulatory networks underlying development in these forms will highlight the possibilities and constraints that lead from the genome to the form of these animals. Thus the main rationale here is the opportunity to describe and compare gene regulatory networks (GRN) in a suite of species that encompass over 500 million years of evolutionary history.

5.1 Recently diverged species: *Strongylocentrotus fransciscanus* and *Allocentrotus fragilis*

To expand the data used in comparisons to *S. purpuratus* at a short evolutionary distance (20 MY), the genomes of two species of sea urchins were chosen for genome sequencing at a low coverage of slightly less than 2X. For *S. fransciscanus* a total of 11 Roche 454 sequencing platform runs were combined to yield 935.9 Mb of sequence. These sequences are available under the NCBI BioProject PRJNA20313. Similarly, six runs were accomplished for *A. fragilis* yielding 475.7 Mb of sequence under project PRJNA20317. The sequences were mapped to the purple sea urchin reference genome and displayed on the Echinobase genome browser (http://Echinobase.org).

5.2 *Lytechinus variegatus*

The next echinoid species chosen for sequencing was *Lytechinus variegatus*, a temnopleurid sea urchin from the east coast of North America. This species is an often-used model that shared a common ancestor with the purple sea urchin about 50 MY ago. The requirement for a suitable genome sequence for gene discovery was one motivation to have the genome of this species sequenced. Another important reason stems from the use of comparative genomics to identify and characterize cis regulatory modules. Unless they are functional, non-coding sequences will have changed from those of the common ancestor through genetic drift over this divergence time. Well over 50 sequence comparisons around genes involved in developmental gene regulatory networks had been made from BAC sequences derived from these two species by 2006. Many fragments were identified in this manner and a majority of the active cis-regulatory modules (CRM) for a gene under study will be found in the patches of conserved sequences identified by an un-gapped comparison using a sliding window technique (Brown et al, 2005).

The *L. variegatus* draft genome sequence was assembled from about 13X Roche 454 reads determined from fragment and 2.5 Kb insert paired ends and approximately 21X Illumina reads The 454 data was assembled using the CABOG assembler to produce 716 Mb of contigs and 429 Mb of degenerate sequences. The assembly and degenerate reads were chopped into fake reads of about 10X coverage and re-assembled using the Newbler assembler. Both the 454 and Illumina reads were mapped onto this assembly using BLAT and bwa respectively. From the mapping positions of paired ends, the contigs were ordered and oriented into scaffolds by ATLAS-Link. Then ATLAS-GapFill was used to assemble the reads locally and to fill gaps. This data set was submitted to NCBI on December 22, 2011 (Accession GCA_000239495.1).

The genome assembly is 835 Mb in length with a contig N50 size of 6.05kb and scaffolds N50 size of 39.17 kb. These parameters are a bit less that the Sp version 0.5 but still adequate for preliminary gene predictions. The latest version of the Lv genome (v2.2) is an upgrade of version 0.4 using PacBio reads with an N50 length of 2.9kb. The long reads at a coverage of 16.5X were employed to fill gaps using the software program PBJelly2 (English et al, 2012) and the further improvement was accomplished using existing sequence with ATLAS-Link (https://www.hgsc.bcm.edu/software/atlas-link). These manipulations produced an improved assembly with a contig N50 of 9.7 kb and a scaffold N50 of 46 Kb.

Similar to the Sp genome assembly, the PacBio improvement was mainly in the length of contigs with the scaffold size distribution hardly changing.

5.3 *Patiria miniata*

The third echinoderm genome to be sequenced was the sea star *Patiria miniata*, a common member of the intertidal community of the Pacific coast of the US. It was chosen as a contrast to the echinoids sequenced because recent molecular details of development reveal interesting differences that inform the evolution of cis-regulatory modules and gene regulatory networks. The assembly used a combination of 15X coverage of Roche 454 reads (fragment and 2.5 kb paired ends) and 70X coverage of Illumina reads (300 bp insert and 2.5 kb paired ends). The 454 reads were assembled at low stringency using CABOG. Both the contig and degenerate 454 reads were then chopped into fake reads and assembled by Newbler. The 454 reads and the Illumina reads were then mapped to the assembly by BLAT and bwa respectively. The resultant contigs were ordered and oriented using ATLAS-Link. Local assemblies of reads using ATLAS-GapFill were used to fill gaps among the contigs within the scaffolds. The final result was an assembly containing 770.2 Mb (811.2 Mb gapped length) of sequence with a contig N50 size of 9.5 kb and a scaffold N50 size of 50.3 kb. It was submitted to NCBI on July 6, 2012 (Accession GCA_000285935.1).

The quality of the *P. miniata* assembly approximates that of the *S. purpuratus* (Sp) 0.5 version. It was this Sp version that was used to computationally identify gene models (Sodergren et al, 2006). For both Pm.v.1.0 and the Sp.v.0.5, the N50 scaffold size is twice the average intergenic distance of 23kb, a number calculated from a deeply sequenced Sp transcriptome (see below). Obviously a genome assembly at this level of completion is quite adequate for gene discovery efforts.

5.4 Other species planned or in progress

A cidaroid sea urchin, *Eucidaris tribuloides* was selected for sequencing (Table 1) because it represents a very different mode of skeletal development. This species has few to zero micromeres and the secondary mesenchyme cells that emerge later from the tip of the archenteron construct the skeleton (Wray and McClay, 1988). A total of 23x coverage of Roche 454 reads (fragment and 2.5kb insert paired ends) and 23X Illumina reads (300bp insert and 2.5kb insert pair ends) were obtained from a single male. The 454 reads were assembled at a stringency lower than the default value from the program. Both the contigs and degenerate sequences were chopped into fake reads and re-assembled with the Newbler assembler. Then the Illumina and 454 paired end reads were mapped on to the assembly using ATLAS-Link to order and orient the contigs. A local assembly of reads was obtained using ATLAS-GapFill in an attempt to fill the gaps in the scaffolds. The final result is a total size of 1.75 GB with a contig N50 of 2.8kb and a scaffold N50 of 28.2Kb. This nascent assembly is a good candidate for the addition of PacBio reads, a step that is currently in progress.

To complete the suite of eleutherozoan genomes, a brittle star (*Ophiothrix spiculata*, Op) and a sea cucumber (*Parastichopus parvamensis*, Pp) were chosen. These species provide two more branches of the phylogenetic tree that diverged during the Cambrian and fit the

rationale for comparative genome sequencing in the echinoderms. These are the first echinoderm genomes derived solely from Illumina sequence (Table 3). For each of these two species the BCM-HGSC constructed paired end sequencing libraries at a range of sizes. Although several of the longer ones failed, a useful amount of sequence was obtained. Preliminary assemblies yielded contig N50 sizes of less than 10Kb, too small to predict genes adequately. A second issue with the data emerged from a kmer analysis that showed almost 2/3 of the kmers were present more than once. This is likely due to a large amount of repetitive sequences in the genome of this brittle star. These data are available in GenBank under accessions JXUT00000000 and JXSR00000000. Probably an additional round of Pac Bio sequencing would extend the contig sizes to useful lengths.

5.5 Non-vertebrate Deuterostomes

Over the span of time that the echinoderm genomes were being sequenced, those of other non-vertebrate deuterostome species have also been posted or published (Table 4.). The genomes of the urochordates, *Ciona* and *Oikopleura*, are significantly smaller than the other non-vertebrate deuterostomes. They show evidence of both gene loss and non-gene sequence reduction (Denoeud et al, 2010). Clearly, the smaller size contributes to an increase in assembly quality; due in part to the reduction in the diversity of repeat sequences such as transposable elements (Denoeud et al, 2010). In contrast the colonial ascidian *Botryllus schlosseri* has a relatively large genome. It has been estimated at 725 Mb using flow cytometry (DeTomaso et al, 1998). The genome sequencing project used a novel approach of pool sequencing and produced about a 580 Mb assembly. The contig size and scaffold size and number are nearly identical indicating there are hardly any multi-contig scaffolds. This assembly is adequate for gene identification but less useful for non-coding sequence analysis. It is notable that *B. schlosseri* does not appear to have a compacted genome as seen in the solitary ascidians or Oikopleura.

The cephalochordate (*Branchiostoma floridae*) genome was sequenced to 8.1X genome coverage of the 575 Mb genome assembly (Putnam et al, 2008). It is notably large residing in only 398 scaffolds. This is due to a strategy where first two separate haplotypes were assembled and then integrated into one mosaic genome sequence. This sort of strategy was first used with the *Ciona savignyi* genome assembly (Vinson et al, 2005; Kerrin et al, 2007). The hemichordate *Saccoglossus kowalevskii* (Sk) genome assembly is of a quality similar to the echinoderms, its sister group. The Sk genome was assembled from about 7.0X coverage of the ~800 Mb genome at BCM-HGSC using ABI platform (Sanger) sequences.

Considering the complexities of genome size, sequencing strategy and assembly approaches the echinoderm genomes are assembled to comparable levels as the other non-vertebrate deuterostome genomes.

6. TRANSCRIPTOMES AND GENE MODELS

A major justification for sequencing echinoderm genomes is the discovery of genes functionally important to processes in cell and developmental biology (see whitepaper link above). Transcriptome sequencing to complement the genome sequences and aid in gene identification in these species and annotation is described below.

6.1 Strongylocentrotus purpuratus

Even before a genome assembly was available high-throughput sequencing schemes in individual tissues of the purple sea urchin were undertaken to identify gene sequences. The first effort identified clones from an arrayed cDNA library derived from activated coelomocytes (Smith et al, 1996). Subsequent studies focused on individual cDNA libraries including the primary mesenchyme cells that produce skeleton (Zhu et al, 2001) and the unfertilized egg (Poustka et al, 1999). Finally for the arrayed libraries, a full set of libraries covering embryonic and larval development was sequenced and clustered to produce a gene catalog (Poustka et al, 2003). These expressed sequence tags were later used in the gene prediction process.

The gene models derived from the version 0.5 genome assembly were computed using four different de novo gene prediction programs. The results were then combined using a latent class analysis method called GLEAN (Elsik, Mackey et al, 2007). A total of 28944 models emerged from this pipeline. Compared to the original 4 prediction programs, the GLEAN predictions had an intermediate number of genes and the best match to a set of ESTs and sequenced cDNAs not included in the prediction pipelines (Sodergren et al, 2006). The gene models were posted on a web site and about 10,000 of them were manually annotated through a community effort by over 200 members of the sea urchin consortium from 73 institutions in 10 countries (Sodergren et al, 2006). To preserve these community annotations, the original gene model set was subsequently mapped to each succeeding genome assembly version. The annotations form the basis for the information about Sp gene models first displayed at SpBase (Cameron et al, 2009) and later at Echinobase (http://Echinobase.org)

A more accurate set of gene models was determined from a deeply sequenced transcriptome project using Illumina 76 bp paired-end reads (Qiang et al, 2012). This RNA-seq effort employed 22 samples covering embryonic, larval and adult stages as well as adult tissues and yielded 784 million reads. The transcripts were assembled using the Bowtie-TopHat-Cufflinks pipeline (Langmead et al. 2009; Trapnell et al. 2009; Trapnell et al. 2010). To filter for false positives those transcripts that met one of three criteria were retained: 1) length of the model must be larger than 400bp; 2) the FPKM of the model must be over 0.5 (which equates to a coverage of ~50X); and 3) the model should have evidence of protein coding capacity. The latter criterion could be a GLEAN model match, a significant SwissProt match or an open reading frame longer that 500 bp with more than one exon. The filtered data set included 21,092 transcripts and over half of the models were supported by all three kinds of protein-coding evidence.

Considering that the gene model annotations have not been systematically reviewed since the version 0.5 genome and the gene models have been mapped intact to each new version of the assembly, additional manual annotation was in order. The re-examination of the gene models took advantage of the more accurate RNA-seq transcriptome. In total, 2785 gene annotation entries were modified through a variety of evidence sources. One hundred seventy four histone gene models present in the Spur_3.1 assembly were found to be duplicates and retired. More recently, the genes on the largest 100 scaffolds (Scaffold1 to Scaffold100) were updated using the WebApollo annotation tool (Lee et al, 2013) and then

integrated into the gene information system on Echinobase. These WebApollo annotation results included merging 351 models with either overlapping or adjacent ones. Consolidation of redundant sequences with improved genome assemblies allowed us to remove 27 exact duplicate models. In addition, 9 new models were created via splitting of a preexisting model and will receive identifiers. Another 79 gene models had structural adjustments made by addition, removal or boundary modification of exons. The shuffling of scaffold sequences with each improved version forced the renaming of 71 models which were not altered structurally but renamed to indicate that they are components of another gene. This process is incomplete as there are still overlapping models remaining. In summary a net loss of 369 models resulted from this last effort. By this view the gene models are reasonably accurate since the number of deprecated models is relatively small compared to the 4014 gene models located on these first 100 scaffolds.

6.2 *Lytechinus variegatus* genes and transcriptomes

Two RNA pools, a pre-gastrula developmental stage pool and one of post-gastrular embryonic stages, were sequenced on the 454 platform as part of the original genome project. These were used to assess the completeness of the sequence assembly. To facilitate gene discovery by the community the two pools of reads were separately assembled using Velvet (Zerbino and Birney, 2008) and the assembled transcripts were mounted at Echinobase as both a blast database and a query database. The query database reports matches to the Sp gene models.

Given a genome assembly for *L. variegatus* of sufficient quality to yield a useful estimate of predicted gene models, the Maker2 program (Holt and Yandell, 2011) was used to derive gene models. The models were generated by masking all repeats, using a training set generated from the transcripts derived above and protein evidence alignments to the genome scaffolds. This information was combined with ab initio gene predictions using the SNAP and AUGUSTUS programs. The Maker2 pipeline parameters were adjusted to yield a set of 28,204 protein coding gene models. We used a reciprocal best BLAST (RBB) comparison to compare the Pm and Sp genes. This is a very strict comparison and it doesn't take into account duplicate gene history, for example. A RBB against the Sp transcriptome proteins yielded 11,727 protein matches (Table 5). The mean size of the predicted gene sequences was 342 bp and the maximum was 12, 662 bp. The percent identity to the purple sea urchin genes was 18%. Using a BLAST expected value threshold of 1×10^{-3}, 407 Lv predicted proteins matched the total of 539 Sp transcription factor proteins. NCBI lists 240 protein models but these have not been mapped to the Echinobase gene models. In a separate comparison, about 68 Lv matches were found to the 253 toll-like receptor sequences in Sp (Buckley and Rast, 2012). These statistics indicate that the genome assembly was of adequate quality to derive a useful set of gene predictions.

6.3 *Patiria miniata* genes and transcriptomes

As in the case of Lv, two pooled samples of embryonic RNA were sequenced on the 454 platform to aid in assessment of the genome sequencing project. Designated "before gastrulation" and "after gastrulation", the reads from these samples were assembled with Velvet and posted as both a BLAST database and a query tool reporting matches against the

Sp RNA-seq transcriptome gene models. Other samples from ovary and testes have been recently submitted to Genbank. However, we did not use them in this analysis

The Maker2 gene prediction pipeline was also used to predict genes from the *P. miniata* genome assembly. Gene models were generated in a similar manner with training sets and the same *ab initio* predictors. The parameters chosen produced 29,697 protein coding gene predictions. Of these, 8,995 yielded a RBB result with the Sp RNA-seq transcripts (Table 5). Similarly, the reciprocal best BLAST matches between Pm and Lv were 8,934, nearly the same number of matches as those between Pm and Sp. Among these reciprocal best BLAST instances between Pm and Sp, 376 matched the 539 annotated transcription factor models from *S. purpuratus*. The mean size of the gene models was 355 bp and ranged up to 7,965 bp. The percent identity compared to the sea urchin genes was a mean of 16% and ranged from 0.3 to 98%.

6.4 Echinoderm transcriptomes without genome assemblies

Assessing the extent of echinoderm transcriptome projects that do not accompany a genome project is a little more difficult. There are a number of reasons to sequence transcriptomes besides first order gene discovery including population studies and other sorts of gene variation analysis. As of this writing there are 219 echinoderm projects in the Short Read Archive (SRA) database at NCBI. Excluding those labeled as population studies or genome sequences, 180 are transcriptome projects of various sorts. A total of 34 species are represented in this collection. By far the most abundant are echinoids closely followed by asteroids. There is only one crinoid transcriptome from ovaries of *Oxycomanthus japonicus* (Reich, et al, 2014).

7. CONCLUSIONS

The variety of genome assembly versions from different echinoderm species display the changes in sequencing technology and computational approaches that have developed over the last 15 years or so. The repeated addition of new sequencing reads for the purple sea urchin derived from new platforms clearly illustrates the unique way that the individual datasets improve the overall quality of the assembly. Short read paired end sequences did little to improve contig length but the long read sequences from PacBio contributed a large increment in contig length. In contrast, the addition of SOLiD sequences made a fractional improvement in scaffold length and the Illumina sequences increased this metric by more than 2-fold. For Lv, the initial combination of 454 and Illumina sequences yielded an assembly similar to the Sp WGS plus BAC skims. Then the contiguity, measured by contig N50, of the assembly of the Lv genome doubled after the addition of PacBio sequence.

While this is not the place to discuss the aspects of various assembler pipelines, it is clear that the best results emerge from a combination of software packages tailored to the kinds of sequencing reads to combine and align. The most recent improvement is PBJelly2 which uses PacBio long reads to efficiently fill gaps and improve scaffolding in existing assemblies (English et al, 2012). The ATLAS package designed and implemented at BCM-HGSC has proved useful on top of a number of initial assemblers.

The refinement of the draft echinoderm sequences reflected in each increment of the genome assembly supports different kinds of questions. The whole genome shotgun sequence from the Sanger platform for the purple sea urchin was sufficient to give a useful approximation of the genes in that genome. Gene set features like the remarkable proliferation of TLR innate immune receptors and scavenger receptors was revealed in these first gene predictions (Sea Urchin Sequencing Consortium, 2006, Hibino et al., 2006). Similarly combinations of 454 and Illumina short read sequences yielded a similar level of completion and supported adequate gene predictions. However, the complete purple sea urchin Hox cluster which covers about 800 kb of genome sequence (Cameron et al, 2005b) was not assembled until version 3.1. Neither Lv nor Pm assemblies contain a complete Hox cluster. If the unusual gene order seen in the purple sea urchin is also found in other echinoderms, the Hox5 gene will be adjacent to one encoding acetylcholinesterase and the Hox1 gene will be adjacent to one encoding a homolog of the even-skipped transcription factor (Cameron et al, 2005b). The scaffolds containing Hox gene paralogs are too short in both Pm and Lv assemblies to confirm this gene arrangement.

Identifying the genes in the echinoderm genomes is an obvious goal of these sequencing projects. Ideally, an exhaustive RNA-seq transcriptome would be aligned to a perfectly assembled genome sequence and the identified transcripts described. Unfortunately, sufficient gene sequence information is seldom available and computational approaches are required. At this point, the GLEAN and Maker2 pipelines have been used to predict genes from echinoderm genomes. Both of these methods use transcribed sequences as training sets. The several individual gene prediction programs used in the GLEAN pipeline for Sp overlapped by about 85% when compared to a “gold standard” of gene sequences not included in the computational sequence data sets (Elsik, unpub). While the GLEAN process on the Sp genome yielded about 28,000 gene sequences, the RNA-seq gene set included about 21,000 that passed a rigorous filtering process. It is likely that the RNA-seq transcriptome is more accurate that the predicted one since it is derived from expressed sequences. However, the computational steps in RNA-seq alignment to the genome sequence are prone to some kinds of errors as well (Garber et al, 2011). Furthermore, the shorter RNAseq sequences are subject to errors because the sequences are not long enough to span more than one splice site resulting in poorer connectedness. This may be less of a problem in genomes where alternative splicing is less frequent. Nevertheless, gene models are often a single consensus rather than a full set of alternate transcripts.

The Sp RNA-seq transcriptome quality was measured by hand annotation and comparison of the genes encoding transcription factors with those same models from the GLEAN set (Qiang et al, 2012). About 45% of these transcription factor genes were essentially consistent between the two data sets. Two-thirds of this fraction differed in the UTR sequences which are more complete in the RNA-seq models. Another third differed in a manner requiring significant revision. GLEAN models with no support from RNA-seq transcripts are 11% and another 5% bore insufficient sequence identity to be matched uniquely to a specific transcriptome model. This history seems to indicate that the original gene models are sufficiently accurate to align to homologous forms but not enough to reveal the fine detail of the gene structure. Given the genome sequence polymorphism and the

precision of the gene finding programs it remains possible that more than one model for an actual gene may exist in this data set.

Homology or sequence similarity remains the best metric for assessing the extent and accuracy of the predicted protein-coding gene models. The public databases of protein-coding gene sequences are enormous and probably cover the vast majority of gene sequences in existence. Of course, truly novel genes will be missed and even widely divergent ones may not be retained. Another approach is the comparison to a more complete gene set from a related organism, in this case Sp (Table 5). The reciprocal BLAST matches between Sp and Lv or Pm are almost exactly the same suggesting that the prediction methods are at least consistent and many of the gene models are identifiable. As expected on phylogenetic grounds the two sea urchins (Sp and Lv) have higher number of matches to each other than either do to the sea star set (Pm). All three echinoderm gene sets match the vertebrate sets (*Homo sapiens* and *Mus musculus*) better than they match the urochordate (*Ciona intestinalis*) reflecting the gene loss and rapid divergence of the urochordates. Indeed, the urochordate matches are nearer to those of the protostomes (*Caenorhabditis elegans* and *Drosophila melanogaster*) a more distant clade.

An important focus for the organization and presentation of the genome sequence information has remained its utility to bench scientists investigating cell and developmental biology at the molecular level. The design of PCR primers and other sequenced based reagents depends on high accuracy in the genome sequence. This is especially true for Sp the most used research model. The current genome assemblies are a mosaic of two haplotypes that may differ by as much as 4% in the case of Sp. Much anecdotal information derived from hand annotated genome sequences of Sp suggests that errors do exist but no systematic assessment has yet been done.

The information on echinoderm genome sequence assemblies is still young. As described above only three draft assemblies are posted at this time and several others are in process. It is likely that additional sequencing will be done as long read technologies become cheaper and more available. It is unlikely that these assemblies will ever be designated other than permanent draft. Many interesting research questions are accessible now and more will emerge as the quality increases. The possibility of describing ancient conservation in structure and function of these genomes which have diverged from each other in deep time is still on the near horizon.

ACKNOWLEDGEMENTS

We thank Ung-Jin Kim and David Felt for work on the Echinobase website and the data therein. We acknowledge Eric Davidson and many members of the Davidson laboratory for sharing sequence data. Professor Susan Ernst kindly read the manuscript and made thoughtful suggestions. Ann Cutting helped with images and attribution. Genome sequencing was supported by the National Human Genome Research Institute, National Institutes of Health U54 HG003273 to RAG. Some of the bioinformatics work described in this review was supported by P41HD071837 to RAC.

REFERENCES

Antonarakis, SE. Human Genome Sequence and Variation. In: Speicher, m.; Antonarakis, SE.; Motulsky, editors. Vogul and Matulsky's Human Genetics. Springer; Berlin: 2010. ISBN 978-3-540-37653-8

Bottjer DJ, Davidson EH, Peterson KJ, Cameron RA. Paleogenomics of Echinoderms. Science. 2006; 134:956–960. [PubMed: 17095693]

Boveri T. Über die polarität des seeigeleies. Verh. d. Psy-med. Ges. Würzburg, N.F. 1901; 34:145–176.

Brachet, J. Chemical Biology. Wiley Interscience; New York: 1950.

Britten R, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969; 165:349–358. [PubMed: 5789433]

Britten RJ, Cetta A, Davidson EH. The single copy sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus*. Cell. 1978; 15:1175–1186. [PubMed: 728997]

Britten RJ, Rowen L, Williams J, Cameron RA. Majority of divergence between closely related DNA samples is due to indels. Proc. Natl. Acad. Sci. USA. 2003; 100:4661–4665. [PubMed: 12672966]

Brown CT, Xie Y, Davidson EH, Cameron RA. Paircomp, FamilyRelationsII and Cartwheel: tools for interspecific sequence comparison. BMC Bioinformatics. 2005; 6:70. [PubMed: 15790396]

Buckley KM, Rast JP. Dynamic evolution of toll-like receptor multigene families in echinoderms. Front Immunol. 2012; 3:136. doi: 10.3389/fimmu.2012.00136. [PubMed: 22679446]

Cai WW, Chen R, Gibbs RA, Bradley A. A clone-array pooled shotgun strategy for sequencing large genomes. Genome Res. 2001; 11:1619–23. [PubMed: 11591638]

Cameron RA, Mahairas G, Rast JP, Martinez P, Biondi TR, Swartzell S, Wallace JC, Poustka AJ, Livingston BT, Wray GA, Ettensohn CA, Lehrach H, Britten RJ, Davidson EH, Hood L. A sea urchin genome project: sequence scan, virtual map, and additional resources. P Natl Acad Sci USA. 2000; 97:9514–9518.

Cameron RA, Chow SH, Berney K, Chiu T-Y, Yuan Q-A, Krämer A, Helguero A, Ransick A, Yun M, Davidson EH. An evolutionary constraint: Strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. P Natl Acad Sci USA. 2005a; 102:11769–11774.

Cameron RA, Rowen L, Nesbitt R, Bloom S, Rast JP, Berney K, Arenas-Mena C, Martinez P, Lucas S, Richardson PM, Davidson EH, Peterson KJ, Hood L. Unusual Gene Order and Organization of the Sea Urchin Hox Cluster. J. Exp. Zool. Part B. 2005b; 304B:1–14.

Cameron RA, Samanta M, Yuan A, He D, Davidson E. SpBase: the sea urchin genome database and web site. D750-D754 Nucleic Acids Research. 2009; Vol. 37 2009. doi:10.1093/nar/gkn887.

Colin, AM. Rapid Repetitive Microinjection. Chapter 32 in: Methods in Cell Biology. In: Schroeder, TE., editor. Echinoderm Gametes and Embryos. Vol. Volume 27. Academic Press; Orlando: 1986.

Davidson, EH. Gene Activity in Early Development. Academic Press; New York: 1968.

Denoeud F, et al. Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate. Science. 2010; 330:1381–1385. [PubMed: 21097902]

Derbes M. Observations sur le méchanisme et les phénomènes qui accompagnent la fornmation de l'embryon chez l'oursin comestible. Ann. Sci. Natur. Zool. 1847; 8:80–98.

De Tomaso AW, Saito Y, Ishizuka KJ, Palmeri KJ, Weissman IL. Mapping the genome of a model protochordate. I. A low resolution genetic map encompassing the fusion/histocompatibility (Fu/HC). Genetics. 1998; 149:277–287. [PubMed: 9584102]

Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Creating a honey bee consensus gene set. Genome Biol. 2007; 8:R13. doi:10.1186/gb-2007-8-1-r13. [PubMed: 17241472]

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE. 2012; 7(11):e47768. doi:10.1371/journal.pone.0047768. [PubMed: 23185243]

Flytzanis CN, McMahon AP, Hough-Evans BR, Katula KS, Britten RJ, Davidson EH. Persistence and integration of cloned DNA in postembryonic sea urchins. Dev. Biol. 1985; 108:431–442. [PubMed: 3000855]

Garber M, Grabherr MG, Guttman M, Trapnel C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods. 2011; 8:469. DOI:10.1038/NMETH.1613. [PubMed: 21623353]

Grula JW, Hall TJ, Giugni TD, Graham GJ, Davidson EH, Britten RJ. Sea urchin DNA sequence variation and reduced interspecies differences of the less variable DNA sequences. Evolution. 1982; 36:665–676.

Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song X-Z, Weinstock GM, Gibbs RA. The Atlas Genome Assembly System. Genome Res. 2004; 14:721–732. [PubMed: 15060016]

Hertwig O. Beitragezurkenntnis der bildung, befruchtung und theilung des thierischen eies. Morph. Jb. 1876; 1:347–432.

Hinegardner R. Cellular DNA content of the Echinodermata. Comp Biochem Physiol. 1974; 49B:219–226.

Hibino T, Loza-Coll M, Messier C, Rast JP, et al. The immune gene repertoire encoded in the purple sea urchin genome. Dev. Biol. 2006; 300:349–365. [PubMed: 17027739]

Hinman V, Nguyen A, Cameron RA, Davidson EH. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. P Natl Acad Sci USA. 2003; 100:13356–13361.

Holt C, Yandell M. MAKER2: an annotation pipeline and genome database management tool for second generation genome projects. BMC Bioinformatics. 2011; 2011:12–491.

Horstadius S. The mechanics of sea urchin development studied by operative methods. Biol Rev. 1939; 14:132–179.

Small KS, Brudno M, Hill MM, Sidow A. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. Genome Biology. 2007; 8:R41. 2007. (doi:10.1186/gb-2007-8-3-r4). [PubMed: 17374142]

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. doi: 10.1186/gb-2009-10-3-r25. [PubMed: 19261174]

Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013; 14:R93. [PubMed: 24000942]

Mardis E, McPherson J, Martienssen R, Wilson RK, McCombie WR. What is Finished, and Why Does it Matter. Genome Res. 2002; 12:669–671. 2002. [PubMed: 11997333]

Mardis ER. Next-Generation DNA Sequencing Methods. Annu Rev Genom Hum G. 2008; 9:387–402.

McCauley BS, Wright EP, Exner C, Kitazawa C, Hinman VF. Development of an embryonic skeletogenic mesenchyme lineage in a sea cucumber reveals the trajectory of change for the evolution of novel structures in echinoderms. EVODEVO. 2012; 3:17. DOI: 10.1186/2041-9139-3-17. [PubMed: 22877149]

McMahon AP, Novak TJ, Britten RJ, Davidson EH. Inducible expression of a cloned heat shock fusion gene in sea urchin embryos. Proc. Natl. Acad. Sci. USA. 1984; 81:7490–7494. [PubMed: 6594699]

Pisani D, Feuda R, Peterson KJ, Smith AB. Resolving phylogenetic signal from noise when divergence is rapid: A new look at the old problem of echinoderm class relationships. Mol Phylogenet Evol. 2012; 62:27–34. [PubMed: 21945533]

Poustka AJ, Herwig R, Krause A, Henning S, Meier-Ewert S, Lehrach H. Toward the Gene Catalogue of Sea urchin Development: The Construction and Analysis of an Unfertilized Egg cDNA Library Highly Normalized by Oligonucleotide Fingerprinting. Genomics. 1999; 59:122–133. [PubMed: 10409423]

Poustka AJ, Groth D, Hennig S, Thamm S, Cameron RA, Beck A, Reinhardt R, Herwig R, Panopoulou G, Lehrach H. Generation, annotation, evolutionary analysis, and database integration, of 20,000 unique sea urchin EST clusters. Genome Res. 2003; 13:2736–2746. [PubMed: 14656975]

Putnam NH, Butts T, Ferrier DEK, Furlong RF, et al. The amphioxus genome and the evolution of the chordate karyotype. Nature. 2008; 453:1064–1071. [PubMed: 18563158]

Reich, A.; Dunn, C.; Akasaka, K.; Wessel, G. Phylogenomic analyses of Echinodermata support the sister groups of Asterozoa and Echinozoa. 2014. 2014submitted

Satou Y, Mineta K, Ogasawara M, Sasakura Y, Shoguchi E, Ueno K, Yamada L, Matsumoto J, Wasserscheid J, Dewar K, Wiley GB, Macmil SL, Roe BA, Zeller RW, Hastings KE, Lemaire P, Lindquist E, Endo T, Hotta K, Inaba K. Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. Genome Biol. 2008; 9:R152. [PubMed: 18854010]

Sea Urchin Genome Sequencing Consortium. The genome of the sea urchin Stongylocentrotus purpuratus. Science. 2006; 314:941–952. [PubMed: 17095691]

Smith LC, Chang L, Britten R, Davidson EH. Sea Urchin Genes Expressed in Activated Coelomocytes Are Identified by Expressed Sequence Tags. J. Immunol. 1996; 156:593–602. [PubMed: 8543810]

Sodergren E, Shen Y, Song X, Zhang L, Gibbs RA, Weinstock GM. Shedding genomic light on Aristotle's lantern. Dev. Bio. 2006; 300:2–8. [PubMed: 17097628]

Telford MJ, Lowe CJ, Cameron CB, Ortega-Martinez O, Aronowicz J, Oliveri P, Copley RR. Phylogenomic analysis of echinoderm class relationships supports Asterozoa. Proc. R. Soc. B. 2014; 281:20140479. 2014.

Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]

Tu Q, Cameron RA, Worley KC, Gibbs RA, Davidson E. Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. Genome Res. 2012 doi: 10.1101/gr.139170.112.

Vaughn R, Garnhart N, Garey JR, Thomas WK, Livingston BT. Sequencing and analysis of the gastrula transcriptome of the brittle star *Ophiocoma wendtii*. EvoDevo. 2012; 3:19. 2012. [PubMed: 22938175]

Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, Mesirov JP, Satoh N, Satou Y, Nusbaum C, Birren B, Galagan J, Lander ES. Assembly of polymorphic genomes: Algorithms and application to Ciona savignyi. Genome Res. 2005; 15:1127–1135. [PubMed: 16077012]

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Ko h W. Passarelli B, Fan HC, Mantalas GL, Palmeri KJ, Ishizuka KJ, Gissi C, Griggio F, Ben-Shlomo R, Corey DM, Penland L, White RA, Weissman IL, Quake SR. The genome sequence of the colonial chordate, *Botryllus schlosseri*. Elife. Jul 2.2013 2:e00569. 2013. doi: 10.7554/eLife.00569. [PubMed: 23840927]

Wray GA, McClay DR. The origin of spicule-forming cells in a 'primitive' sea urchin (*Eucidaris tribuloides*) which appears to lack primary mesenchyme cells. Development. 1988; 103:305–315. [PubMed: 3066611]

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–9. [PubMed: 18349386]

Zhu XD, Mahairas G, Illies M, Cameron RA, Davidson EH, Ettensohn CA. A large-scale analysis of mRNAs expressed by primary mesenchyme cells of the sea urchin embryo. Development. 2001; 128:2615–2627. [PubMed: 11493577]

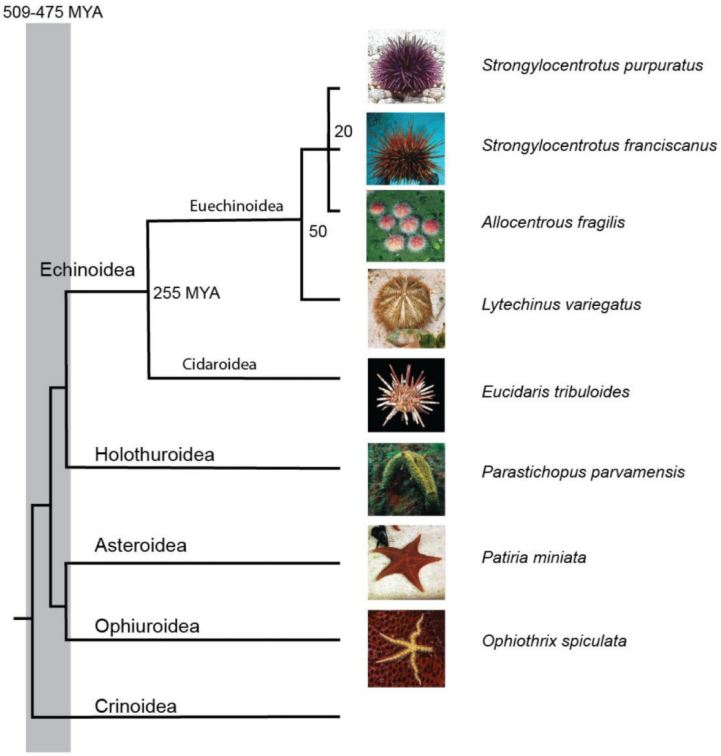

Figure 1.

The phylogeny of echinoderms following the asterozoan topology. Time axis not to scale. The four pictured classes: Echinoidea, Holothuroidea, Asteroidea and Ophiuroidea make up the Eleutherozoa. Redrawn using information from Telford et al, 2014; Pisani et al, 2012 and Reich, et al., 2014. Photo credits: Stronglyocentrotus purpuratus © Andy Cameron, California Institute of Technology. Stronglyocentrotus franciscanus Channel Islands NMS. Allocentrus fragilis Ed Bowlby, NOAA/Olympic Coast NMS; NOAA/OAR/Office of Ocean Exp. Lytechinus variegatus © Hans Hillewaert / CC-BY-SA-3.0 . Eucidaris tribuloides © Ann Cutting Caltech. Parastichopus parvamensis U.S. federal government. Patiria miniata © Ann Cutting KML. Ophiothrix spiculata Jerry Kirkhart from Los Osos, Calif.

Table 1

Sequencing progress for echinoderm genomes. The eight genome projects conducted by the Baylor College of Medicine, Human Genome Sequencing Center and the stage of completion of each.

Species	Status
Strongylocentrotus purpuratus v3.1	mature draft
Strongylocentrotus franciscanus	2x skim coverage
Allocentrotus fragilis	2x skim coverage
Lytechinus variegatus v2.2	improved draft
Eucidaris tribuloides	in assembly
Patiria miniata v1.0	first draft
Parastichopus parvamensis v1.0	first draft
Ophiothrix spiculata v1.0	first draft

Table 2

Genome assembly quality statistics for the purple sea urchin versions.

Assembly version	0.5	2.1	2.6	3.1	4.0
Total seq length (Kb)	1,095,825	907,070	912,546	936,580	1,032,044
Number of scaffolds	187,612	114,222	75,034	32,009	31,879
Scaffold N50 (Kb)	55	123	168	402	431
Number of contigs	278,688	195,154	196,827	174,773	146,491
Contig N50 (Kb)	8	11.7	11.5	13.5	17.6
Total Contig bp (Mb)		804	806	816	902
Sanger	6X	8.3X	8.3X	8.3X	8.3X
SOLiD			18X	18X	18X
Illumina				40X	40X
PacBio					10.6X

Table 3

A synopsis of the draft genome assembly statistics for all of the echinoderm projects now in process.

Species	Contig N50 (Kb)	Scaffold N50 (Kb)	Total Contigs (Mb)	Sanger	454	SOLiD	Illumina	PacBio	Notes
S. pur.	13.5	402	816	8.3x	13x	18x	40x		v3.1
S pur.	17.6	431	954	8.3x	13x	18x	40x	10.6x	v4.0
L. var.	6.2	42.6	823		23x		21x		v0.4
L. var.	9.7	46.	1,004		23x		21x	13x	v2.2
O. spi.	4.5	43	1,900				160x?		v1.0
P. par	7.1	40	707		15x		140x?		v1.0
E. trib	2.78	28.2	1,750				23x		v1.1
P. min.	9.5	52.6	811				70x		v1.0

Table 4

A synopsis of the genome assembly statistics for non-vertebrate deuterostome genome assemblies.

Species	Total Length (Mb)	Scaffold Number	Scaffold N50 (Kb)	Contig Number	Contig N50 (Kb)	Reference
Strongylocentrotus purpuratus v4.0	1,032	31,879	431	140,454	17.6	In process
Lytechinus variegatus v2.0	1061	322,936	46.3	481,804	9.7	In process
Patiria miniata v1.0	811.0	60,183	52.6	179,756	9.4	Direct submission
Saccoglossus kowalevskii v1.1	775.8	54,120	245.8	135,721	10.1	Direct submission
Branchiostoma floridae v2	521.9	398	2,586	41,925	27.9	Putnam et al, 2008
Oikopleura dioica	70.5	1,260	395.4	5,917	24.9	Denoeud et al, 2010
Ciona intestinalis vKH	115.2	1,272	5,153	6,381	37.1	Satou et al, 2008
Botryllus schlosseri	579.6	120,139	7.2	120,124	6.9	Voskoboynik, et al., 2013
Ciona savignyi	174	374	1,779	4,620	116	Small et al, 2007

Table 5

Reciprocal best blast gene comparisons. Each gene model set was compared reciprocally to echinoderm gene sets and public gene sets of other species (see text). Abbreviations: Sp, *S. purpuratus*; Lv, *L. variegatus*; Pm, *P. miniata*; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Ci, *Ciona intestinalis*; Hs, *Homo sapiens*; Mm, *Mus musculus*, Nv, *Nematostella vectensis*

	Sp	**Lv**	**Pm**	**Ce**	**Dm**	**Nv**	**Ci**	**Hs**	**Mm**
Sp		11727	8995	4982	5528	7131	6311	7594	7594
Lv	11727		8934	4044	4884	6434	5625	6170	6686
Pm	8995	8934		4370	5177	6734	5909	6963	6916