Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

Developmental Biology 300 (2006) 2 – 8

DEVELOPMENTAL BIOLOGY

www.elsevier.com/locate/ydbio

Review

Shedding genomic light on Aristotle's lantern

Erica Sodergren *, Yufeng Shen, Xingzhi Song, Lan Zhang, Richard A. Gibbs, George M. Weinstock

Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Alkek N1519, Houston, TX 77030, USA

Received for publication 5 June 2006; revised 4 October 2006; accepted 5 October 2006
Available online 10 October 2006

Abstract

Sea urchins have proved fascinating to biologists since the time of Aristotle who compared the appearance of their bony mouth structure to a lantern in *The History of Animals*. Throughout modern times it has been a model system for research in developmental biology. Now, the genome of the sea urchin *Strongylocentrotus purpuratus* is the first echinoderm genome to be sequenced. A high quality draft sequence assembly was produced using the Atlas assembler to combine whole genome shotgun sequences with sequences from a collection of BACs selected to form a minimal tiling path along the genome. A formidable challenge was presented by the high degree of heterozygosity between the two haplotypes of the selected male representative of this marine organism. This was overcome by use of the BAC tiling path backbone, in which each BAC represents a single haplotype, as well as by improvements in the Atlas software. Another innovation introduced in this project was the sequencing of pools of tiling path BACs rather than individual BAC sequencing. The Clone-Array Pooled Shotgun Strategy greatly reduced the cost and time devoted to preparing shotgun libraries from BAC clones. The genome sequence was analyzed with several gene prediction methods to produce a comprehensive gene list that was then manually refined and annotated by a volunteer team of sea urchin experts. This latter annotation community edited over 9000 gene models and uncovered many unexpected aspects of the sea urchin genetic content impacting transcriptional regulation, immunology, sensory perception, and an organism's development. Analysis of the basic deuterostome genetic complement supports the sea urchin's role as a model system for deuterostome and, by extension, chordate development.

Keywords: Echinoderm; Sea urchin; Genome sequence; Genome annotation; BAC clone

> "All animals whatsoever, whether they fly or swim or walk upon dry land, whether they bring forth their young alive or in the egg, develop in the same way:" (Aristotle, 350 B.C.E.)

Introduction

The turn of this millennium will likely be remembered as the Genome Sequencing Era with the completion of the Human Genome Sequencing Project and rapid accumulation of genome sequences of important model organisms. The sea urchin *Strongylocentrotus purpuratus* now takes its place in this august group of human, mouse/rat and fruit fly with the publication and analysis of its genome sequence. While many genomes of biological interest can be listed as possible targets for genome sequencing, these outweigh available resources, even with the dramatic decrease in cost of sequencing over the past decade. The sea urchin's utility as a model in developmental biology, its evolutionary niche, and the active research community working on the sea urchin were compelling rationales for proceeding with *S. purpuratus* (Cameron and Davidson, 2002). The long, rich history of using the sea urchin to study processes involved in an organism's development, combined with recent insights such as a systems biology paradigm for early development, set the stage for generating one more valuable research resource, the genome sequence. In the larger context of human evolution, the sea urchin, an Echinoderm, would be the first species outside the Chordate branch of Deuterostomia to be sequenced, allowing a fuller description of the basal Deuterostome genetic complement and furthering our understanding of human evolution and biology by comparison.

The Sea Urchin Genome Sequencing Project (SUGSP) was conceived as a high quality draft (HQD) sequence of the ~800 Mb genome (Hinegardner, 1971), to be produced by the Human Genome Sequencing Center at Baylor College of Medicine. The HQD state of a genome sequence refers to nearly

* Corresponding author. Fax: +1 713 798 5741.
E-mail address: ericas@bcm.edu (E. Sodergren).

0012-1606/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ydbio.2006.10.005

complete coverage of a genome (>95% in this case), with high accuracy and contiguity of the sequence. Most genes lie in regions of contiguity and long-range structure can be seen, allowing reliable prediction of gene and protein sequences. A HQD sequence also has limitations: repeated sequences are not necessarily completely represented, gaps are present, some regions may be misassembled by misjoining through repeats, difficult to sequence regions (due to repeats or secondary structures for instance) are not completely resolved, and base errors are present. Nevertheless, a rich picture of the genetic potential of an organism can be inferred from a HQD sequence, allowing a detailed annotation and analysis.

The overall quality of a HQD genomic sequence can be improved by including a component of sequenced BAC (*B*acterial *A*rtificial *C*hromosomes) clones, each containing a random sea urchin genomic segment of 145–165 kb. Such clones aid in the assembly process (below) by providing smaller regions to assemble rather than addressing the entire genome *simultaneously*, which is valuable in avoiding assembly errors resulting from joining segments at repeated sequences. In addition, since each BAC insert represents a single haplotype, this can be used to select reads of the same haplotype, simplifying assembly of highly heterozygous genomes, such as the sea urchin. A set of BAC clones that only contain short overlapping regions with each other while as a group cover all "clonable" regions of the genome defines a minimal tiling path (MTP) of BAC clones. The sea urchin project broke new ground by sequencing an entire MTP of BAC clones via a pooling method (Clone-Array Pooled Shotgun Strategy or CAPSS (Cai et al., 2001)) rather than sequencing all clones individually.

The annotation and analysis phase of the project reflected the melding of the rationale for sequencing the sea urchin with the biology of the organism. This phase was also notable in that it drew in the wider research community. Over 200 additional individuals collaborated to curate and analyze over 9000 genes from a master prediction set of 28,945 gene models. During the analysis a number of lines of evidence led to the current estimate of 23,300 genes for *S. purpuratus* (Sea Urchin Genome Sequencing Consortium, 2006).

Sequencing the S. purpuratus genome

The sea urchin genome presented severe challenges to reach the high quality draft grade, principally due to the high frequency of polymorphism. The presence of high genome variation is a consequence of the population structure of marine organisms (Lessios et al., 2001) and the difficulty of producing an inbred sea urchin line (Cameron et al., 1999). Early experiments by Britten et al. (1978) suggested approximately 4–5% sequence divergence between the single-copy DNA of two individual sea urchins. Measurements of sequence variation in the initial *S. purpuratus* assembly revealed at least one single nucleotide polymorphism (SNP) per 100 bases, and a comparable frequency of insertion/deletion (indel) variation. This ratio of SNPs:indels can vary locally (Britten et al., 2003). This means that in a single DNA sequencing read of 800 bases there are on average 8 single base mismatches and 8 indel mismatches between the two haplotypes, or one mismatch per 50 bases with some regions exhibiting much higher variation.

The basic operation in assembling a genome is to correctly align individual reads and use this layout for building the consensus sequence. The challenge lies in distinguishing the true overlaps between reads from "false" overlaps when the reads contain repeated sequences. A mismatch every 50 bases in overlap regions of the two haplotypes is similar to the overlaps observed between divergent repeats, which are rejected to avoid improper joining of sequences. The tendency is for the assembly process to split the genome into two haplotype assemblies, which are both then of lower coverage and accuracy. Thus, rather than an overall sequence coverage of 6×, the result is 3× coverage for each of two haplotypes. For another highly polymorphic marine organism, *Ciona savignyi* (Vinson et al., 2005), with a 190 Mb genome, the approach to solve this problem was to sequence the genome to 12× coverage, assemble each haplotype separately at 6×, and then merge. While this approach can be used for smaller genomes, the high level of coverage is costly for the larger sea urchin genome. A more elegant, economical solution was to use sequencing of large insert BAC clones as well as new assembly algorithms.

The use of BAC clones deserves special mention, since many genomes, such as *Ciona*, were sequenced using a pure whole genome shotgun (WGS) approach. The use of BACs allows each BAC-defined region of the genome, ~1/8000 of the whole, to be assembled individually and then all the BAC sequences can be stitched together for a complete genome. The local assembly helps in dealing with the repeated sequence problem, since the repeat structure of a BAC-size region is simpler than the whole genome. But since each BAC is a single haplotype, they also help with the polymorphism problem. The BCM-HGSC approach, pioneered with the Rat Genome Sequencing Project (Gibbs et al., 2004), is to use a minimal tiling path (MTP) of BAC clones, each sequenced to low (2×) coverage, along with cheaper and faster WGS sequencing to 6× coverage. The Atlas assembly software (Havlak et al., 2004) was developed specifically for combining the WGS and BAC reads, and is unique among whole genome assembly software in this regard. Each set of BAC reads is used as 'bait' to 'fish' for overlapping WGS reads and then the local assembly is performed (see below and Fig. 2 for elaboration). The product is called an enriched BAC (eBAC), the basic unit of the Combined Assembly approach of Atlas. The eBACs are stitched together to form the genome. Because the bait reads from each BAC are a single haplotype, it was possible to distinguish which WGS reads were from the same haplotype and add the reads from the second haplotype at a later step when the assembly was already more clearly defined.

The overall approach for the SUGSP is shown in Fig. 1. DNA came from a single male and was used to prepare a variety of clone resources: small insert (2–6 kb) plasmids produced at BCM-HGSC and medium insert (30–50 kb) and large insert (130–160 kb) BACs produced at Cal Tech (Cameron et al., 2000). A fingerprint map and tiling path of BAC clones was constructed in work done at the Michael Smith Genome

Sciences Centre in Vancouver (Sea Urchin Genome Sequencing Consortium, 2006). Each BAC's DNA was cut by the restriction enzyme EcoRI, the fragments sized on a gel (FPC: (Soderlund et al., 2000; Soderlund et al., 1997)) and a region of shared sequence was recognized by the presence of a group of common restriction fragments. This information was then used to order the BACs throughout the genome. All of the DNA sequence data was generated by the BCM-HGSC: (1) the ~6-fold sequence coverage generated by WGS reads, (2) the clone-end sequences from the medium and large insert BAC libraries used to order and orient assembled sequence and (3) the 2-fold sequence coverage generated from sequencing tiling path BACs. An initial assembly was performed with the pure WGS reads, and this assembly was adequate for generating a gene list and accelerating the annotation and analysis phase (below).

The BAC skimming was performed using another innovation developed by the BCM-HGSC as part of the SUGSP: the use of BAC pooling (Fig. 2). The BAC tiling path of the sea urchin is 8248 clones, which traditionally would require a separate DNA preparation and shotgun library for each. The Clone-Array Pooled Shotgun Sequencing (CAPSS) method (Cai et al., 2001) allows BAC DNA and shotgun libraries to be prepared from pools of BACs, reducing the overall workload, time, and cost. The BACs are arrayed so each is present in two pools, 576 clones in a 24 column by 24 row format per array, and this information is used to deconvolute the mixture of reads to individual BACs (Shen et al., 2006). A particular clone's reads will display sequence overlaps between reads from the row pool and column pool whose intersection define its unique position in the array. The reads identified in this manner are referred to as deconvoluted reads and the resulting BAC sequences as deconvoluted BACs in order to draw the distinction from the traditional direct sequencing of individual BAC clones. These deconvoluted BAC reads were used as 'bait' to 'fish' for overlapping WGS reads to produce eBACs (enriched BAC assemblies) by Atlas. The eBACs are joined together according to their layout in the MTP to form BACtigs, the major component of the assembly. Inevitably there are some gaps between BACtigs, representing regions for which there was not a BAC clone in the tiling path or the original BAC library. For various reasons, the WGS reads usually cover the genome more comprehensively than the large BAC clones. Therefore, WGS contigs that cover locals ignored by eBAC assemblies were identified. These unique WGS contigs are either merged with the BACtigs with which they overlap to fill in the gaps or added to the genome assembly as separate contigs. Thus, the ultimate Combined Assembly of the genome is a composite of ordered eBAC assemblies complemented with WGS contigs not represented in the eBACs.

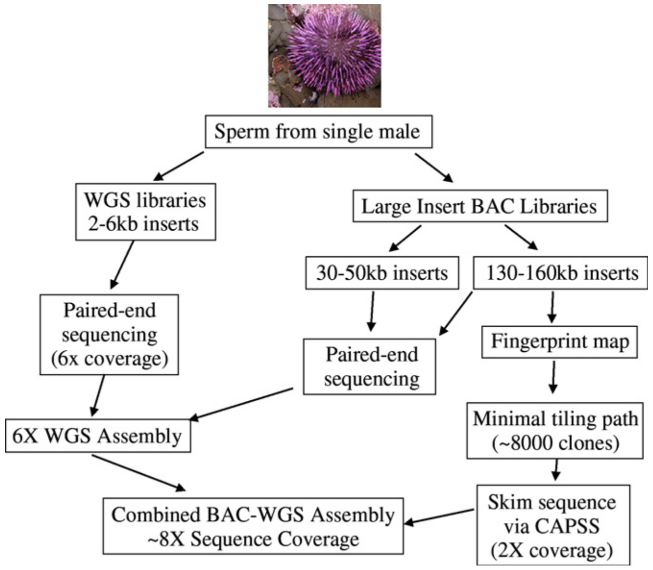


Fig. 1. Approach for sequencing the sea urchin *S. purpuratus* genome. A single male individual's DNA was used to create all of the genomic libraries utilized. Two data production approaches were implemented coincidentally. (1) The left side of the diagram describes the path to the initial WGS assembly. (2) The right side describes the generation of sequence from a minimal tiling path set of BAC clones utilizing the CAPSS strategy. In addition, end sequences from both large insert libraries were used in organizing assemblies. Ultimately, all data were integrated into the final combined BAC–WGS assembly.

The statistics for the WGS and Combined assemblies are shown in Table 1. Here N50 is the length of a scaffold (or contig) sequence such that half of the genome is in sequences of the N50 length or longer (e.g., half of the genome is represented by scaffolds 65.6 kb or larger in the WGS assembly). Contigs are the stretches of ungapped sequence produced from contiguous overlapping reads. Scaffolds are contigs that have been linked to each other but with short gaps separating them. Both ends of each clone (plasmid or BAC) are sequenced and the resulting read pairs sometimes fall in different contigs, allowing these contigs to be linked into scaffolds, with the intervening gap size estimated from the insert size of the read paired clone. Also noted is the amount of 'redundancy' in the assembly. This is measured by comparing the structure of test regions from the HQD assembly that have been independently taken to a more complete sequencing grade (an upgrade of the HQD that includes targeted sequencing across gaps and low quality regions that can be identified by automated algorithms). Specifically, each of the two assemblies was compared to 25 BAC clones that had been individually sequenced to a very high quality, providing a benchmark to determine how well these 25 regions had been assembled in the WGS and Combined draft sequences. When the draft sequences are aligned to these BAC test sequences, redundancy is observed as contigs in the draft sequences that overlap in the alignment. For the WGS assembly ~15% of the sequence was found to lie in these overlapping regions. This redundancy was reduced in the final Combined HQD sequence to ~5% and is largely due to regions from the two haplotypes that overlap but have not been merged due to extreme sequence differences. Taking into account the redundancy, the size of the genome represented in the assembly is comparable to previous measurements of 800 Mb (Hinegardner, 1971).

The robust quality of the WGS assembly is of particular note since it allowed the annotation and analysis phase of the project to begin while the BAC sequencing and assembly segment was still underway. Given that the size of a typical sea urchin gene is

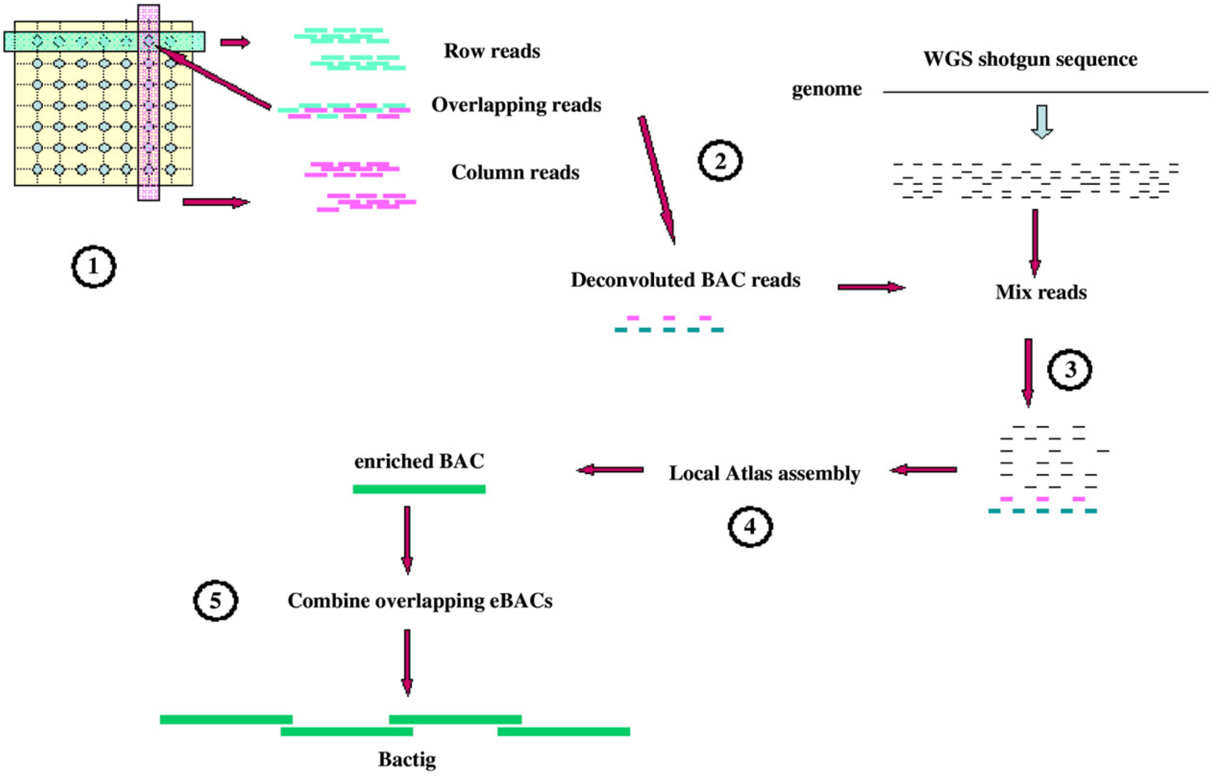


Fig. 2. Implementing the CAPSS strategy in the BAC-based assembly. Cartoon (1) depicts an array (in this case a 7 × 7 array) from which aliquots of each BAC culture in the top row are pooled prior to DNA isolation and shotgun library production. Aliquots of each BAC culture in Column 6 are also pooled. DNA sequence reads from each pool are generated. Reads that have sequence overlaps, other than by repeated sequences, define the region covered by the BAC at position R1C6. Steps (2) and (3) illustrate how these deconvoluted reads define a single BAC and can be used to identify overlapping reads produced in the whole genome shotgun path of the project. Step (4) is the Atlas assembly of this BAC-defined region that results in the eBAC assembly. Overlapping eBACs are then stitched together in a longer Bactig.

about 10 kb, the size of contigs (N50 of 10.2 kb) and scaffolds (N50 of 65.6 kb) from the WGS assembly was adequate for gene predictions. Moreover, assessing the completeness of the assembly by measuring how many of the sea urchin ESTs available in dbEST gave matches to the sequence indicated at least 95% of the genome was represented in the assembly.

Table 1
Statistics of sea urchin genome assemblies

Statistic	WGS assembly	Combined BAC–WGS assembly version 2.1
Assembly size [a]	980 Mb	847 Mb
% Redundancy [b]	15	5
Total # scaffolds >1 kb [c]	77,484	54,960
N50 scaffolds >1 kb [c, d]	65.6 kb	142 kb
# Scaffolds to N50 [d]	2819	1374
Total # contigs >1 kb [c]	121,006	105,692
N50 contigs >1 kb [c, d]	10.2 kb	18.5 kb
# Contigs to N50 [d]	18723	13575
Total # reptigs >1 kb	56,554	N/A
N50 reptigs >1 kb	1.52 kb	N/A

[a] Bases contained in all scaffolds (plus reptigs, contigs assembled using stringent criteria from repeated sequences, for WGS assembly).

[b] Determined by measuring sequence redundancy in comparison to the assemblies of 25 Phase2 BACs (2.9 Mb) that had been individually sequenced.

[c] Scaffolds/contigs greater than 1000 bp in length.

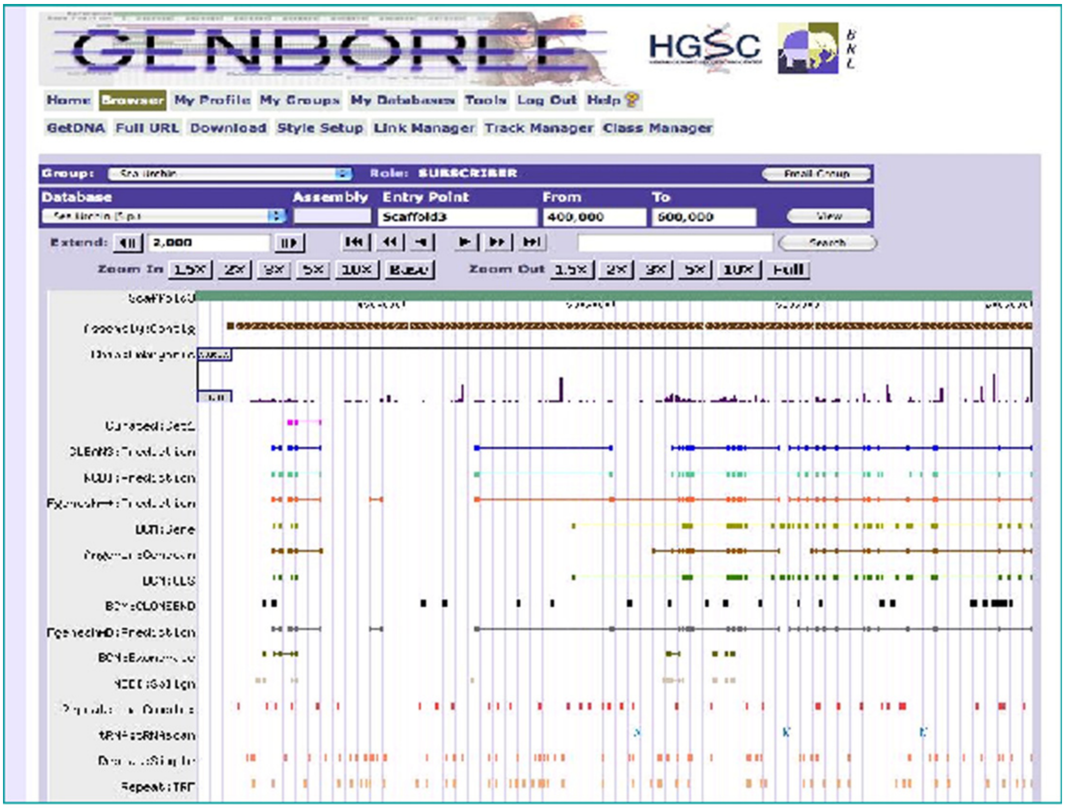
[d] The N50 nomenclature refers to the minimum number of scaffolds/contigs that contain 50% of the genome. All scaffolds/contigs are ordered by size and one starts with the largest representative and sums all scaffolds/contigs until 50% of the genome size is reached. The N50 is the size of the last scaffold/contig added to the aggregate, and the number to the N50 is the total number of scaffolds/contigs in the aggregate.

Annotation and analysis

A number of different sets of gene predictions were produced from the WGS assembly. These included gene sets from the BCM-HGSC using software developed at Ensembl (Potter et al., 2004; Sea Urchin Genome Sequencing Consortium, 2006) and installed at BCM-HGSC, NCBI (gnomon software; Souvorov et al., 2004), Softberry (FgenesH software; Salamov and Solovyev, 2000; Solovyev, 2001), and a Genescan approach (Angerer Lab Gene List, 2006). These four gene sets were combined using the GLEAN program (Elsik et al., 2006a) which uses Latent Class Analysis to estimate accuracy and error rates for each source of gene evidence, and then constructs a consensus prediction based on the patterns of agreement between sources. As a measure of completeness and accuracy, the various gene sets were compared to a 'gold standard' set of genes and gene fragments. The analysis/annotation teams submitted ~600 cDNA, EST, QPCR and protein sequences that had been generated in their laboratories but were not as yet in the public domain. This data was not used by any of the gene prediction programs and was thus an appropriate set to evaluate gene prediction sets. Based on these comparisons, the merged gene set produced by GLEAN was selected as the best gene set. This best set contained 28,945 gene models which is an overestimate given the ~15% redundancy in the version of the assembly that was used.

The annotation and analysis of the sea urchin data built on the approach used for the analysis of the honeybee genome (Elsik et al., 2006b) but extended the functionality of the software and extent of gene model analysis. To facilitate easy access and comparison of all the available data types, two tools were built at the BCM-HGSC: (1) an Annotation Database that provided pre-computed comparisons of the GLEAN prediction set as well as captured manual curations of the gene models and (2) an online genome browser, Genboree (www.genboree.org), for visual comparison. The Annotation Database contained pre-computed determinations of PFAM motifs, top 10 alignments to human, mouse and *Ciona* and the genomic coordinates along with DNA and protein sequence of the gene model. Annotators were asked to validate/modify the gene model and provide additional information such as revised DNA/protein sequence, common gene names, expression information and available protein multiple-alignments. The Genboree display allowed one to view all the information available for a region, i.e. a scaffold (Fig. 3). Informative tracks for a scaffold contain the component contigs, embryonic gene expression data (Samanta et al., 2006), curated genes, GLEAN gene predictions and supporting gene prediction sets and EST data. The embryonic expression data resulted from a chip containing unique 50-mers spaced 10 bases apart on the genome (both strands) that was interrogated with embryonic mRNA (Samanta et al., 2006). This data was precise enough for the 3′UTR regions to be inferred for an expressed gene and included in the Annotation Database as pre-computed information. Different displays of information were also available at the NCBI (NCBI genomes, 2006) and the genome browser at the University of California at Santa Cruz (UCSC genome browser, 2006).

Armed with the above assortment of tools and enormous enthusiasm, the extended analysis community of over 200 individuals from 73 institutions formed 22 working groups defined by major themes of interest such as sensory systems, ciliogenesis, biomineralization, immunological capabilities, signal transduction, transcriptional regulation, cytoskeleton, reproduction, the adhesome and the evolving Metazoan genome. The gene list was used to evaluate the genetic constituents of these themes of interest with the concomitant manual analysis of over 9000 genes. Open discussion and dissemination of information between groups was maintained through weekly conference calls, a list-serv and a sea urchin ftp site. Additional conversations occurred between members of a group and between groups of overlapping interests as the need arose. The Annotation Database facilitated cross group interactions: as individuals found they were not the only one



Group: Sea Urchin
Role: SUBSCRIBER

Database: Sea Urchin (S.p.)
Assembly:
Entry Point: Scaffold3
From: 400,000
To: 600,000

Extend: 2,000

Genboree browser with gene predictions and annotations.

Fig. 3. Visualization of annotation resources. All of the information for a region is displayed on a set of tracks below the assembly scaffold and contigs in the lined window. Dialogue boxes allow one to retrieve scaffold coordinates as well as additional information on any feature. Signal strength from the embryonic expression array chip lies directly under the contigs. The Curated track is dynamically updated as human curations are added to the Annotation Database. The official Glean gene set track is followed by a track for each of the underlying gene prediction sets. Available EST alignments occupy the Exonerate and Splign tracks.

interested in a particular gene, they could contact the person who had already begun annotating a gene and share results. The European community went one step further and held a week-long annotation workshop in Naples, Italy (Ina Arnone and Michael Thorndyke organizers). This highly collaborative adventure has culminated in the variety of biological insights found in the articles presented in this special issue of Developmental Biology and elsewhere (Sea Urchin Genome Sequencing Consortium, 2006).

Future directions

There are several key follow up activities to this initial sequence and analysis of the sea urchin genome. The assembly itself can be further refined without extensive addition of new data. In particular, it is possible to separate the two haplotype sequences which may be of utility in clarifying genetic structure, particularly when one haplotype contains coding sequence mutations. Another critical aspect for the project's future is development of a sea urchin database to collect work done on the genome and continue to update the annotations. This would include not only refinements to the sequences of specific genes done in research labs, but also comparisons to other genomes as they are sequenced, including hemichordates and chordates. Comparisons to other sea urchin genomes should also be of interest, particularly in view of the lower density of indels in regions that may be involved in regulation as well as coding sequences (Cameron et al., 2005). Such comparisons can further refine the annotation of regulatory sequences in addition to coding regions. Looking somewhat further afield, the wealth of SNP markers present in the genome could be of use in detailing the population structure of this marine organism. And, of course, the many doors that have been opened to the biology of the sea urchin, described in the other articles in this issue, will ensure that the genome sequence will continue to be mined in the future.

Acknowledgments

We would like to thank Eric Davidson and Andy Cameron for sharing their wealth of sea urchin history and knowledge as well as making available the large insert libraries. We would also like to thank others who were supportive in more defined areas: Marco Marra, Jacquie Schein and the team at the Michael Smith Genome Sciences Centre prepared the FPC map and BAC tiling path. Aaron Mackey ran the Glean analysis and shared this unpublished software. Kim Worley shared insights learned in the annotation of the honeybee genome. While much of Aleksander Milosavljevic's group supports the Genboree browser, Andrew Jackson advised us as new types of data appeared for display. This work was supported by the NHGRI via grant #5 U54 HG003273.

References

Angerer Lab Gene List, 2006. http://urchin.nidcr.nih.gov/blast/index.html.

Aristotle, 350 B.C.E. The History of Animals, Vol. BookVII Part 7.

Britten, R.J., Cetta, A., Davidson, E.H., 1978. The single-copy DNA sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus*. Cell 15, 1175–1186.

Britten, R.J., Rowen, L., Williams, J., Cameron, R.A., 2003. Majority of divergence between closely related DNA samples is due to indels. Proc. Natl. Acad Sci. U. S. A. 100, 4661–4665.

Cai, W.W., Chen, R., Gibbs, R.A., Bradley, A., 2001. A clone-array pooled shotgun strategy for sequencing large genomes. Genome Res. 11, 1619–1623.

Cameron, R.A., Chow, S.H., Berney, K., Chiu, T.Y., Yuan, Q.A., Kramer, A., Helguero, A., Ransick, A., Yun, M., Davidson, E.H., 2005. An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. Proc. Natl. Acad Sci. U. S. A. 102, 11769–11774.

Cameron, R.A., Davidson, E.H. 2002. Sea Urchin White Paper, Vol. www.genome.gov/Pages/Research/Sequencing/SeqProposals/SeaUrchin_Genome.pdf.

Cameron, R.A., Leahy, P.S., Britten, R.J., Davidson, E.H., 1999. Microsatellite loci in wild-type and inbred *Strongylocentrotus purpuratus*. Dev. Biol. 208, 255–264.

Cameron, R.A., Mahairas, G., Rast, J.P., Martinez, P., Biondi, T.R., Swartzell, S., Wallace, J.C., Poustka, A.J., Livingston, B.T., Wray, G.A., Ettensohn, C.A., Lehrach, H., Britten, R.J., Davidson, E.H., Hood, L., 2000. A sea urchin genome project: sequence scan, virtual map, and additional resources. Proc. Natl. Acad Sci. U. S. A. 97, 9514–9518.

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S., Weinstock, G.M., 2006a. Creating a honey bee consensus gene set. Genome Biol. (in press).

Elsik, C.G., Worley, K.C., Zhang, L., Milshina, N.V., Jiang, H., Reese, J.T., Childs, K.L., Venkatraman, A., Dickens, C.M., Weinstock, G.M., Gibbs, R.A., 2006b. Community annotation: procedures, protocols, and supporting tools. Genome Res. 16, 1329–1333.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R.A., Adams, M.D., Amanatides, P.G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C.A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C.L., Nguyen, T., Pfannkoch, C.M., Sitter, C., Sutton, G.G., Venter, J.C., Woodage, T., Smith, D., Lee, H.M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R.B., Dunn, D.M., Green, E.D., Blakesley, R.W., Bouffard, G.G., De Jong, P.J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C.M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W.C., Havlak, P.H., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.Z., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K.C., Cooney, A.J., D'Souza, L.M., Martin, K., Wu, J.Q., Gonzalez-Garay, M.L., Jackson, A.R., Kalafus, K.J., McLeod, M.P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D.A., Zhang, Z., Bailey, J.A., Eichler, E.E., et al., 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493–521.

Havlak, P., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.Z., Weinstock, G.M., Gibbs, R.A., 2004. The Atlas genome assembly system. Genome Res. 14, 721–732.

Hinegardner, R.T., 1971. An improved fluorometric assay for DNA. Anal. Biochem. 39, 197–201.

Lessios, H.A., Kessing, B.D., Pearse, J.S., 2001. Population structure and speciation in tropical seas: global phylogeography of the sea urchin Diadema. Evol. Int. J. Org. Evol. 55, 955–975.

NCBI genomes, 2006. www.ncbi.nlm.nih.gov/Genomes/.

Potter, S.C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S.M., Stabenau, A., Storey, R., Clamp, M., 2004. The Ensembl analysis pipeline. Genome Res. 14, 934–941.

Salamov, A.A., Solovyev, V.V., 2000. Ab initio gene finding in Drosophila genomic DNA. Genome Res. 10, 516–522.

Samanta, M.P., Tongprasit, W., Istrail, S., Cameron, R.A., Davidson, E.H., Stolc, V., 2006. The Transcriptome of the Sea Urchin Embryo. Science 314 (in press).

Sea Urchin Genome Sequencing Consortium, 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314 (in press).

Shen, Y., Sodergren, E.J., Weinstock, G.M., Gibbs, R.A., 2006. Personal communication.

Soderlund, C., Humphray, S., Dunham, A., French, L., 2000. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. 10, 1772–1787.

Soderlund, C., Longden, I., Mott, R., 1997. FPC: a system for building contigs from restriction fingerprinted clones. Comput. Appl. Biosci. 13, 523–535.

Solovyev, V.V., 2001. Statistical approaches in Eukaryotic gene prediction. In: Balding, D.E.A. (Ed.), Handbook of Statistical Genetics. John Wiley and Sons, Ltd, pp. 83–127.

Souvorov, A., Tatusova, T., Lipman, D.J., 2004. Genome annotation with Gnomon—A multi-step combined gene prediction tool. ISMB, p. 125.

UCSC genome browser, 2006. genome.ucsc.edu/cgi-bin/hgGateway?clade=deuterostome and org=S.+purpuratus and db=0 and hgsid=72701798.

Vinson, J.P., Jaffe, D.B., O'Neill, K., Karlsson, E.K., Stange-Thomann, N., Anderson, S., Mesirov, J.P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J.E., Lander, E.S., 2005. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. Genome Res. 15, 1127–1135.