Available online at www.sciencedirect.com

ScienceDirect

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