Lawrence Berkeley National Laboratory
Recent Work

Title
Chapter 2: The Xenopus tropicalis genome project

Permalink
https://escholarship.org/uc/item/0zp1q843

Journal
Current Genomics, 4

Authors
Richardson, Paul M.
Chapman, Jarrod

Publication Date
2003

eScholarship.org Powered by the California Digital Library
University of California

Current Genomics Hot Topic issue on

***Xenopus* genomics**

CHAPTER 2

The *Xenopus tropicalis* Genome Project

*Paul M Richardson and Jarrod Chapman

US Department of Energy Joint Genome Institute

2800 Mitchell Drive

Walnut Creek, CA 94598, USA

*Corresponding author: PMRichardson@lbl.gov; Telephone: 1 925 296 5851; Fax: 1 925 296 5875

Abstract

The Human Genome Project has resulted in the elucidation of the genomic sequence of a number of model organisms as well as a reference sequence for the human genome. The utility of these available genomes has been demonstrated by researchers throughout the world, and spurred the desire to obtain additional genomic information from a number of sources. The United States Department of Energy's Joint Genome Institute has undertaken a project to sequence the genome of the amphibian *Xenopus (Silurana) tropicalis*. The primary goal of the project is to produce a high-quality genome sequence and annotation to meet the needs of the research community. In March of 2002, a number of *Xenopus* researchers from around the world met at the JGI Production Genomics Facility in Walnut Creek, California to discuss goals and strategies for the project. The project is designed to make use of a whole-genome shotgun approach supplemented with extensive BAC end sequences and shotgun sequence from selected BACs. A high-quality draft genome is desired that will meet minimal criteria for contiguity and long-range linking information. At depths of 6-8X sequence coverage, we expect that a large fraction of features of interest (exons, promoters and regulatory regions) will be covered in large contigs of high sequence quality without gaps. In addition, long-range linking of contigs will be achieved through paired end gap-spanning clones so that contigs are ordered and oriented into large scaffolds with gaps of defined size. These scaffolds typically contain multi megabase-sized regions of the genome. This approach has led to high-quality draft genomes of the pufferfish (*Fugu rubripes*), *Ciona intestinalis* and the mouse. Since there will be extensive coverage of large inserts for *Xenopus* including BAC and Fosmid end sequencing, clones will be readily available for finishing selected regions of the genome.

Key words

EST, Genome, Sequencing, *Xenopus*

List of abbreviations

JGI-Joint Genome Institute

NIH-National Institutes of Health

BAC-Bacterial Artificial Chromosome

EST-Expressed Sequence Tag

RH- Radiation Hybrid

NCBI-National Center for Biotechnology Information

Introduction

The Human Genome Project has resulted in not only the complete reference sequence of a human, but also the genomic sequences for several model organisms including *E.coli*, yeast, *Drosophila*, *C. elegans*, as well as draft sequences for *Fugu rubribes*, *Ciona intestinalis*, mouse and numerous other smaller genomes [1-3]. Significant advancements in technology along the way have led to a large increase in capacity to generate sequence and a concomitant decrease in costs. Chief among these efforts were the consolidation of production sequencing centers that are able to take advantage of automated platforms and economies of scale. The separation of tasks has led to a production line process that has increased efficiency substantially. Technology improvements, especially capillary electrophoresis instruments and reliable fluorescent dye terminator chemistries, have also led to increased throughput. A typical large-production sequencing center can now generate on the order of 20-30 million lanes of sequence per year. This tremendous capacity is now being used to delve deeper into the kingdoms of life to help us better understand organismal differences and evolutionary relationships between phyla.

For years, cell and developmental biologists have relied on the amphibian *Xenopus*. Recent advances in technology have led researchers in these fields to begin to think in new ways about how best to utilize new data, especially those resources created by the revolution in genomics. Meetings and discussions at the National Institutes of Health (NIH, Bethesda, Maryland, USA) in 1999 and earlier led to a report that describes recommendations and priorities for developing *Xenopus*-based genetics and genomics resources (http://www.nih.gov/science/models/*Xenopus*/reports/*Xenopus*_report.pdf). Among the recommendations were development of cDNA libraries and expressed sequence tags (ESTs), a database, microarray facilities and the development of the *Xenopus tropicalis* system for genetics, including insertional and chemical mutagenesis, mapping resources such as BAC libraries, Radiation Hybrid (RH) panels, and microsatellite markers as well as funds for transitioning labs to using *X. tropicalis*. The

background and current progress of these projects is reviewed in Klein et al. [4] and available at http://www.nih.gov/science/models/xenopus/.

X. tropicalis is a close relative of the more widely used *Xenopus laevis,* and may be used to complement studies in *X. laevis.* In particular, *X. tropicalis* is much more suited to genetic studies because of its diploid genome and shorter generation time [5-8]. Many of the developmental studies and tools used for *X. laevis* are applicable to *X. tropicalis*. In addition, the estimated genome size of approximately 1.7 billion bases make it the smallest of the frog species, and an ideal candidate for a genome sequencing project. [9]. The amphibians sit in a key phylogenetic position relative to the mammals and fish genomes that have been sequenced. The genome sequence of *X. tropicalis* will enable comparative genomics approaches for studying development, cell biology, and physiology among vertebrates. These comparisons will allow insights into similarities and differences of protein-coding regions as well as regulatory segments of these genomes.

EST sequencing

There are a number of groups throughout the world generating cDNA libraries and end sequences of clones from those libraries (ESTs) that are being deposited in the databases. Currently, genbank contains approximately 300,000 *X. tropicalis* ESTs and over 170,000 *X. (Silurana) tropicalis* ESTs from a variety of sources. Among the groups contributing to these ongoing efforts are the Sanger Center, NIH, the JGI as well as groups in France and Japan. Data is being deposited in public databases and libraries and clones are available through the Image Consortium (http://image.llnl.gov/image/html/iresources.shtml). This data will provide a rich source for annotation of the genome sequence and will also provide researchers with valuable information for designing probes and primers. The National Center for Biotechnology Information (NCBI) has constructed a Unigene set from the *X. laevis* ESTs (http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Xl) and is working on a version for *X. tropicalis*. In addition, this set will lead to the generation of a group of full-length cDNA clones for expression studies that are sequence verified, much like the mammalian

gene collection. There are currently over 1,200 full-length cDNAs in the *Xenopus* gene collection available to the research community (http://xgc.nci.nih.gov/).

Sequencing Strategy

The sequencing project will be based on a whole-genome shotgun approach, similar to the method used to generate the mouse sequence [3]. End sequences generated from a number of different shotgun libraries containing a variety of insert sizes will be utilized for the assembly. Sequence from ongoing EST and cDNA projects will be used for confirmation of assemblies and annotation purposes. In addition, any available genetic and physical mapping information such as BAC fingerprint contigs and microsatellite markers will also be used to place and verify scaffolds on the genome. Whole-genome shotgun libraries constructed from DNA isolated from a single individual will be used for the bulk of the sequencing project. The data will be assembled using JAZZ, the whole-genome assembler developed at the JGI and used for the *Fugu* and *Ciona intestinalis* genome projects [1, 2].

Library Construction

NIH has sponsored the construction of BAC libraries of *X. tropicalis* [4]. A BAC library was constructed at the Institute for Systems Biology (Seattle, Washington, USA) from the DNA of several 5th generation Nigerian *X. tropicalis* individuals. The average insert size of this library is approximately 75kb (S. Qin, personal communication). A second library was also constructed from the DNA of a single sixth generation inbred Nigerian individual that is a sister of the individual whose DNA is being used for whole-genome shotgun sequencing. These BAC clones and filters for screening are available for distribution to the community through the CHORI BAC/PAC distribution facility (http://www.chori.org/bacpac/home.htm). Washington University Genome Sequencing Center will end sequence and fingerprint map a large number of clones from both libraries. These sequences will be deposited in the NCBI trace archive and fingerprint maps will be available through a distributed genome annotation server. End sequences from over 225,000 clones will be used in the whole-genome assembly and will be extremely valuable for providing long-range contiguity to the assembly.

In addition, the JGI has begun a program to shotgun sequence and finish a number of BACs selected to contain genes of interest to the research community. Nominations are encouraged for specific gene sequences to be used for screening BAC libraries to identify clones containing the genes. Researchers may submit requests at the JGI *Xenopus* website: http://genome.jgi-psf.org/xenopus0/bacs/index.html. Nominations are reviewed periodically by the *X. tropicalis* genome project steering committee. A table of genes currently being screened is listed in Table 1 and is updated regularly on the web site.

Random shotgun libraries for sequencing are being constructed at the JGI from DNA isolated from several tissues of a single individual. The sequencing plan calls for the construction of 3 major shotgun libraries: a small-insert library of approximately 3kb average insert size in pUC18; an ~8kb insert library in a low copy vector; and a fosmid library using the pCC1Fos (Epicentre, Madison, WI) inducible vector. These libraries will be sequenced to an approximately 6X sequence coverage in high-quality bases ($\geq$Q20 bases [10]).

DNA from a single sixth-generation inbred Nigerian individual (provided by Rob Grainger, University of Virginia, USA) was isolated from 4 tissues of the adult female frog. High-molecular weight DNA has been used in the construction of all whole-genome shotgun libraries. Several test libraries were constructed by random shearing of genomic DNA in a Hydroshear (GeneMachines, San Carlos, CA). After blunt-end repair using T4 DNA polymerase and Klenow fragment, sheared DNA was size selected on a 1% agarose gel (Fig. 1). Fragments corresponding to approximately 3kb were excised and purified from the gel and ligated into pUC18. Colonies resulting from transformation of electrocompetent DH10B *E. coli* cells with this ligation were checked for inserts using flanking forward and reverse primers (Fig. 1B). Detailed protocols can be found at http://www.jgi.doe.gov/Internal/prots_index.html and in Detter et al [11].

Sample sequencing of the 3kb pUC library was begun in July of 2002. Initial results indicated a low level of insertless clones (~1%) and mitochondria contamination (<0.1%). This library was expanded and extensive sequencing has begun. The reads were screened for vector sequence and trimmed for quality by taking the longest continuous stretch of at least 100 bases with a mean Phred quality of 15 or greater over a

21-base sliding window [10]. The resultant data set used in the following analysis contained 2,980,800 passing reads totaling 1.66 Gigabases of sequence. The mean trimmed read length was 558 bases (Fig. 2). Of the reads that passed vector and quality trimming, 88% have paired passing sister reads. All reads are available for blast comparison or downloading as they are generated (http://genome.jgi-psf.org/xenopus0/xenopus0.info.html). In addition, all raw traces are deposited regularly into the trace archive at NCBI.

We chose 12 BACs from the 75Kb BAC library and shotgun sequenced and assembled with Phrap to serve as reference sequence of the whole-genome libraries. Insert size distribution for the nominal 3Kb pUC library was obtained by blastn alignment of the trimmed reads against contigs in the phrap assemblies of the 12 BACs (total of 788 kb). Inserts with both ends aligning to a nonrepetietive region of a BAC at 99% identity over 99% of the (trimmed) length were used to calculate the distribution of insert sizes. The observed mean insert size was 2.85 $\pm$ 0.45 kb (see Fig. 3A). The read-averaged G+C content was 40.4% and the most-likely value was 37.7% (Fig. 4). For comparison, the G+C content distribution of the mitochondrial DNA averaged over 500 base windows is shown on the same graph.

Mitochondrial Sequence

We have been able to reconstruct the mitochondrial genome from the low level of contamination seen in the genomic DNA prep. The mitochondrial reads were identified by blastn against the *X. laevis* mitochondrial sequence using a 1e-50 evalue cutoff. Reads that matched the *X. laevis* mitochondrial sequence, their paired sister reads, and any additional reads that shared high-quality alignments (as detected by Malign, a module of the Jazz whole genome assembler [1,2]) with these reads, were then assembled together into a single contig using Phrap. This assembly and annotated sequence is available for viewing and download at http://genome.jgi-psf.org/xenopus0/xenopus0.home.html.

Genome Size Estimates

The genome size of *X. tropicalis* has been estimated to be approximately 1.7 Gigabases [9]. We have attempted to estimate the genome size (of clonable DNA) from

sequence information in the data set described above. An all-to-all alignment of the trimmed reads in the dataset was performed using Malign. The mean number of high-quality alignments per read was calculated as a function of read length. For this purpose, high quality was defined as $\geq$ 97% identity, and reads with an exceptionally large number of high-quality alignments were omitted from the averaging. The theoretical functional form of this quantity is given by: N(x) = d*[1+(x-2*L0)/L], where d is the depth of coverage, L0 the minimal detectable alignment length, and L the mean read length. The observed data was fit to the theoretical curve using gnuplot (which employs a nonlinear lest-squares (NLLS) Marquardt-Levenberg algorithm). From this curve fit, the depth is estimated to be (1.2 $\pm$ 0.1). Using the formula that Genome size equals total bases divided by depth results in an estimated genome size of 1.4 $\pm$ 0.1 gigabases (Fig. 5A). A second method has been employed to estimate genome size. All reads were aligned using blastn to the shotgunned BAC contigs obtained with phrap (used above to estimate insert size of the 3 kb library). Using an in-house program to visualize the alignments, the number of reads aligning to non-repetetive BAC sequence at $\geq$ 99% identity over 99% of their trimmed length was determined (Fig. 5B). Total number of reads aligning over the length of contigs was used to estimate the depth at 1.1 $\pm$ 0.1. This figure results in an estimated genome size of 1.6 $\pm$ 0.1 gigabases.

Conclusions

As a result of the investments in the Human Genome Project, several model organisms have been sequenced including *E.coli*, yeast, *Drosophila* and mouse. The JGI has begun to utilize its extensive sequencing capacity to move beyond the human genome and produce sequenced genomes of *Fugu rubribes, Ciona intestinalis* as well as many microbes. The production of the *X. tropicalis* genomic sequence will be a major source of information for *Xenopus* researchers as well as provide a basis for comparative and functional genomics studies.

References

[1] Aparicio, S., J. Chapman, et al. (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes." *Science* **297**(5585): 1301-10.

[2] Dehal, P., Y. Satou, et al. (2002). "The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins." *Science* **298**(5601): 2157-67.

[3] Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." *Nature* **420**(6915): 520-62.

[4] Klein, S. L., R. L. Strausberg, et al. (2002). "Genetic and genomic tools for Xenopus research: The NIH Xenopus initiative." *Dev. Dyn.* **225**(4): 384-91.

[5] Hirsch, N., L. B. Zimmerman, et al. (2002). "Xenopus, the next generation: X. tropicalis genetics and genomics." *Dev. Dyn.* **225**(4): 422-33.

[6] Hirsch, N., L. B. Zimmerman, et al. (2002). "Xenopus tropicalis transgenic lines and their use in the study of embryonic induction." *Dev. Dyn.* **225**(4): 522-35.

[7] Khokha, M. K., C. Chung, et al. (2002). "Techniques and probes for the study of Xenopus tropicalis development." *Dev. Dyn.* **225**(4): 499-510.

[8] Amaya, E., M. F. Offield, et al. (1998). "Frog genetics: Xenopus tropicalis jumps into the future." *Trends Genet.* **14**(7): 253-5.

[9] Thiebaud, C. H. and M. Fischberg (1977). "DNA content in the genus Xenopus." *Chromosoma* **59**(3): 253-7.

[10] Ewing, B., Hillieer, L., and P. Green. (1998) "Base-calling of automated sequencer traces using Phred I Accuracy probablilities." *Genome Research* **8**:175-185.

[11] Detter, J. C., J. M. Jett, et al. (2002). "Isothermal strand-displacement amplification applications for high-throughput genomics." *Genomics* **80**(6): 691-8.

Table 1
Genes used as basis of BAC screening.
The list of genes was submitted by researchers hoping to identify BACs containing the full-gene sequence. Overgo probes designed from cDNA and shotgun sequence are used to probe filter sets representing the available BAC libraries. Non-redundant positive clones are selected for shotgun library sequencing. Draft sequence is deposited in Genbank and posted on the *Xenopus* web site immediately upon assembly. Finished sequence and annotation will be displayed as soon as it becomes available.

Figure legends

Figure 1
Construction of shotgun library

(A) Agarose gel of *X. tropicalis* genomic DNA after shearing (middle two lanes). The black rectangle indicates the region where DNA was excised from the gel for cloning into pUC18. Markers (M) on either side are Lambda HindIII digests.
(B) Agarose gel showing PCR-products of 48 individual colonies from the completed library using primers flanking the cloning site in pUC18. Lambda Hind III markers are indicated by (M).

Figure 2
Distribution of read lengths.
Graph of the distribution of trimmed read lengths for individual sequence reads, where trimming was performed by taking the longest continuous stretch of at least 100 bases with a mean Phred quality of 15 or greater over a 21-base sliding window. Reads are represented as a percentage of the total (y-axis) versus read length in bases (x-axis). The mean (558), median (610), and most-likely (666) read lengths are indicated by vertical bars on the graph from left to right, respectively.

Figure 3
Distribution of library insert sizes.
Insert sizes were determined by aligning paired end reads to contigs obtained from assembled BACs as detailed in the text. Graph is the percent of the total trimmed reads (Y-axis) *versus* the observed distance between read pairs in bases (X-axis). The observed insert size for the pUC18 shotgun library is 2.85 kb +/- 0.45 kb.

Figure 4
Percent G+C content.
The read-averaged G+C content (percent, X-axis) was calculated for the complete set of

trimmed reads as a function of the total # of reads (Y-axis). The mean G+C content was 40.4%. For comparison, the G+C content distribution of the mitochondrial DNA averaged over 500 base windows is shown on the same graph (right Y-axis).

Figure 5

Genome Size Estimate

(A) An alignment of all reads to one another was used to determine average depth of coverage using the trimmed read data set. Graph shows the average trimmed read length (X-axis) versus the percent of total reads (rightY-axis). The mean number of high-quality alignments (left Y-axis) are represented as diamonds. From this graph, the depth is estimated to be 1.2 +/- 0.1 resulting in an estimated genome size of 1.4 +/- 0.1 Gigabases.

(B) An alignment of reads to BAC contigs was also used to estimate genome size. Counting the number of high-quality alignments (99% identity over 99% of the read, shown in red), matching contigs representing approximately 250 kb (60 kb shown) of genomic DNA resulted in an estimated depth of 1.1 +/- 0.1. This depth implies a genome size of 1.6+/- 0.1 Gigabases.