Rice Research: Open Access
ISSN: 2375-4338

Short Communication Open Access

# Sequencing the Rice Genome

**Artemi Cerda***

*Crop Science Department, Federal University of Santa Catarina, Florianópolis, SC, Brazil*

## Introduction

Rice is one of the most important crops in the world. Rice, wheat, and maize together account for about half of the world's food production, and rice itself is the principal food of half of the world's population. Rice is the obvious choice for the first whole genome sequencing of a cereal crop. The rice genome is well mapped and well characterized, and it is the smallest of the major cereal crop genomes at an estimated 400 to 430 Mb. The next largest genome of an important cereal crop is that of sorghum, at 750 to 770 Mb, and the wheat genome is ~37 times the size of the rice genome at close to 16,000 Mb. Grass genomes, including those of rice, wheat, maize, barley, rye, and sorghum, share a large degree of synteny, making rice an excellent model cereal .Rice is also the easiest of the cereal plants to transform genetically. A genome size of 430 Mb nonetheless represents a daunting task for whole genome sequencing. The rice genome is 3.5 times the size of the Arabidopsis genome and the third largest public genome project undertaken to date, behind the human and mouse genomes.

## Description

The International Rice Genome Sequencing Project (IRGSP) began in September 1997, at a workshop held in conjunction with the International Symposium on Plant Molecular Biology in Singapore. Scientists from many nations attended the workshop and agreed to an international collaboration to sequence the rice genome. As a result, representatives from Japan, Korea, China, the United Kingdom, and the United States met six months later in Tsukuba to establish the guidelines [1]. The participants agreed to share materials and to the timely release of physical maps and annotated DNA se-quince to public databases. The IRGSP has evolved to include 11 nations, and the IRGSP Working Group, composed of a representative from each participating nation, formulates IRGSP policies and finishing standards. The recent interim IRGSP meeting at Clemson University (September 19 and 20, 2000) in South Carolina was the largest rice genome meeting to date and was attended by more than 70 scientists and administrators from Japan, Taiwan, Thailand, Korea, China, India, Brazil, France, Canada, and the United States [2]. The meeting was organized by Rod Wing, U.S. IRGSP Representative (Clem-son University), and chaired by Ben Burr, IRGSP Coordinator (Brookhaven National Laboratory, New York), and Talkie Sasaki, Program Director of the Rice Genome Research Program (RGP) in Japan. Major players in the project include the RGP; the CCW, collaboration between the Clemson University Genomics Institute (CUGI), Cold Spring Harbour Laboratory, and the Washington University Genome Sequencing Center; the Institute for Genome Research (TIGR) in Rockville, MD; and the Plant Genome Initiative at Rutgers University (PGIR). Various additions and/or changes in IRGSP members were noted at the meeting. Brazil became the newest member and was represented by Antonio Costa de Oliveira of the Universidad Federal de Pelotas, who proposed to work on chromosome 12. Canada representative Thomas Bureau of McGill University proposed switching from work on chromosome 2 to coordinating activities on chromosome 9 with Thailand. India, previously an unfunded member of the IRGSP [3], has a new Rice Genome Program (represented by Akhilesh Tyagi of the University of Delhi and Nagendra Singh of the Indian Agricultural Research Institute) and will begin work on chromosome 11. A full list of participating countries and institutions, including URLs of sites offering information relevant to the IRGSP, is provided.

Physical mapping and the construction of "sequence-ready" minimal tiling paths Fingerprinting and physical mapping are used to create minimal tiling paths and to anchor BAC clones to physical positions along the length of a chromosome. A sequence-ready BAC coting is a contiguous set of minimally overlapping BAC clones that has been anchored to a position along the length of a particular chromosome [4]. Presentations during the opening session on physical mapping and a physical mapping workshop provided a good overview of the various techniques and procedures used during this process. Garnet Presting (formerly of CUGI and now at the Novartis Agricultural Discovery Institute in San Diego) described the CUGI physical mapping project (now led by Eric Fang and others at CUGI). The project involves fingerprinting of HindIII and EcoRI BAC libraries, assembling the fingerprinted BACs into contigs, anchoring of the BACs onto the physical map with DNA gel restriction fragment length polymorphism (RFLP) and BAC end sequence analysis, and connecting and extending of contigs by chromosome walking. Another project involves the mapping of plant ESTs onto the rice physical map. Information on the ESTs is being integrated into the rice physical map and made accessible on the CUGI website.

Assembly of the fingerprinted BACs into counties is greatly aided by the use of software called Fingerprinted Coting's (FPC). Sunderland described how she and others at CUGI are working on developing ways of using FPC in conjunction with data from markers (DNA gel blot analysis), fluorescence in situ hybridization (FISH) and optical mapping (when available), and the BAC end sequence (STC) database to improve the efficiency and reliability of creating sequence-ready BAC counties [5]. Eric Fang of CUGI described techniques for extending BAC counties and closing gaps as part of the process to create a sequence-ready framework for the entire genome. One method described was the use of overgo probes. In this procedure, a pair of 24-bp sequences that contain an 8-bp overlap is designed from BAC end sequences. The 24-bp sequences are joined to create a 40-bp "overgo,"

FISH and optical mapping are two other methods that may be used to aid genome sequencing projects. Jiming Jiang (Department of Horticulture, University of Wisconsin) presented data to illustrate the application of FISH to rice physical mapping. The length of the pachytene chromosome structure in rice makes it particularly amenable to FISH

***Corresponding author:** Artemi Cerda, Laboratory of Plant Pathology, Crop Science Department, Federal University of Santa Catarina, SC, Brazil, Email: artemi@cerda.br

**Received:** 2-May-2022, Manuscript No:rroa-22-63940 **Editor assigned:** 4-May -2022, Pre QC No: rroa-22-63940(PQ), **Reviewed:** 18-May-2022, QC No: rroa -22-63940 **Revised:** 24-May-2022, Manuscript No: rroa-22-63940(R), **Published:** 30-May-2022, DOI: 10.4172/2375-4338.1000302

**Citation:** Cerda A (2022) Sequencing the Rice Genome. J Rice Res 10: 302.

and resolution and sensitivity are comparable to results that have been obtained from human chromosomes. Jiang showed how FISH could be used to easily identify rice chromosomes, to determine the chromosome location of uncertain clones, and to determine the physical nature of large linkage gaps, which could facilitate sequence closing at chromosome ends and telomere regions. Jiang's group previously reported that this technique would be valuable for characterizing BAC clones that contain complex repetitive DNA sequences such as those found in rice. Sally Leong (United States Department of Agriculture Agricultural Research Service Plant Disease Resistance, Madison, WI) described the optical mapping technique developed by David Schwartz and colleagues at the University of Wisconsin. This technique uses fluid flow capillary action to extend and align DNA molecules onto a specially prepared glass surface. DNA that is extended and fixed to the surface is then digested with restriction enzymes, and fluorescence microscopy imaging is used to map an ordered array of fragments. Schwartz's group has used optical mapping to create whole genome restriction maps of the microorganisms, and they received a National Science Foundation Plant Genome Award in 1999 to support the construction of an optical restriction map of the rice genome. FISH and optical mapping could significantly enhance physical mapping and sequencing of difficult regions, such as centromeres and regions containing highly repetitive sequence, although their expense currently limits the extent to which they are being used in genome sequencing.

### Sequencing, finishing, and annotation

The Rice Genome Research Program (Japan) uses a shotgun approach to sequence PAC or BAC clones. With this procedure, individual PAC/BAC clones (100 to 200 kb) from a sequence-ready counting are shattered by sonication or nebulization, and the fragments are sub cloned to produce a shotgun library with an average insert size of 1 to 3 kb. Clones from the shotgun library are then sequenced at random to provide the desired degree of "coverage" of the total sequence. For example, to provide for fourfold coverage of a 120-kb BAC, at least 1200 clones from a shotgun library would be sequenced at random (assuming 400 bp per sequence read).

After sequencing, software such as PHRED and PHRAP is used to order the sub clone sequences and reassemble the entire BAC sequence. PHRED, developed by Phil Green and Brent Ewing at the University of Washington, reads DNA sequencer trace data, calls bases, and assigns quality values to the bases; PHRAP is a program developed by Phil Green for assembling shotgun DNA sequence data. Of course, it is rare that an entire BAC will be assembled without gaps from shotgun sequences. Doug Johnson (Washington University Genome Sequencing Center), Melissa de la Bastide (Cold Spring Harbor Laboratory), Robin Buell (TIGR), and Apichart Vanavichit (Kasetsart University, Thailand) presented information on the finishing process and discussed ways to deal with problem regions and filling gaps. Problem regions in sequencing include highly repetitive sequence and AT-rich and GC-rich regions. The rice genome carries a large amount of repetitive sequence. Problems in these areas often can be overcome by switching the sequencing chemistry; de la Bastide presented a list of various sequencing kits and reagents that have been used successfully at Cold Spring Harbour Laboratory to sequence through difficult areas. She discussed two other ways of dealing with particularly recalcitrant regions: the use of small insert libraries and transposons. Small insert libraries are created by physical shearing and sub cloning of a template that spans the region of difficulty. Transposons also may be used to break up and thereby aid the sequencing of a difficult region, which is achieved by random insertion of transposons into a difficult clone.

### Genoscope proposal: whole genome shotgun sequencing of the rice genome

Clones would be distributed among IRGSP sequencing groups on a basis roughly proportional to their current chromosome claim, and each group would submit sequence to a common public database accessible to all groups. Either in parallel with the shotgun sequencing or after it was completed, group members would return to the current BAC-by- BAC approach to complete various chromosomal regions. IRGSP Coordinator Ben Burr stated that there would be on-going discussion of the proposal among members of the working group and indicated that it may be feasible, and desirable, to integrate various aspects of the proposal into IRGSP policy.

### Policy discussion and revision of finishing standard

The goal of the RGP is a complete and accurate sequence of the entire genome. "Complete" was originally defined as less than one error in 10,000 bases, consistent with the Bermuda standards. The measure of completeness was previously considered to be a PHRED score (quality value) of 40 or greater. Sasaki presented empirical evidence showing a quality value of 30 to be consistent with this level of accuracy. There was general agreement on revising the finishing standards to this value to speed the release of sequence while maintaining high-quality data. There was considerable debate regarding a number of other revisions. One of these was whether or not small gaps (such as occur in GC-rich regions) could be left in "completed" sequence representing a single contig. On this point, Sasaki presented evidence from the RGP that these regions are likely to contain open reading frames and that every effort should be made to close gaps before stating that a contig is complete. Others argued that it may be more desirable to release large contigs with small gaps, because closing the gaps is these regions is likely to take a considerable amount of time and effort. Revision of the finishing standards was still under discussion after the meeting. Finally, Ben Burr noted that IRGSP currently is not in line with the Bermuda standards in that groups are only encouraged and not required to release preliminary sequence information. A compromise was discussed that would require submission of phase II data and encourage phase I release. The IRGSP working group is preparing a modified data release policy that will be finalized and released by early 2001.

Robin Buell (TIGR) presented data on the effect of fourfold rather than eightfold coverage on the quality of HTGS phase II sequence released to the public before closure. Her group compared the quality of sequence that would be obtained from fourfold versus eightfold coverage by using data from three rice BACs that have been completed. There was a substantial difference in the quality of the data. The fourfold coverage yielded 42 counties, and the largest counting extended 14.8 kb; whereas the eightfold coverage of the same region yielded only 14 contigs, and the largest coting extended nearly 35 kb. For HTGS submissions, fourfold coverage was predicted to miss, on average, 16% of the sequence that would be obtained through eightfold coverage. Thus, release of fourfold instead of eightfold coverage data could be an important factor for end users before closure and makes a convincing argument for the release of combined Monsanto and IRGSP data at phase II (which together represent at least eightfold coverage).

## Conclusion

Sequencing of the rice genome is a monumental task. To date, ~3.5% of the genome has been completed (15 of 430 Mb) and another 3 to 5% is in production. Nonetheless, the data that have been released have already provided valuable information on genome structure and

organization (see, e.g., Mao et al., 2000), much of which will apply to other cereal crops and to monocots in general. A major part of the nuclear genomes of most plants, and indeed many eukaryotes, is composed of repetitive DNA elements. Repetitive DNA is estimated to constitute at least 50% of the rice genome and as much as 70% of the maize genome (Nagano et al., 1999). Complete sequencing of the rice genome will provide valuable information on the effect of repetitive elements on genome organization and evolution in plants. The IRGSP also constitutes a proving ground for sequencing and finishing methods for complex genomes, which will provide excellent resources for other eukaryotic genome sequencing projects in the future.

## Acknowledgement

I would like to thank my Professor for his support and encouragement.

## Conflict of Interest

The authors declare that they are no conflict of interest

### References

1. Lai Z (1999) A shotgun optical map of the entire Plasmodium falciparum genome. Nat Genet 23: 309-313.
2. Lin J, Qi R, Aston C, Jing J, Anantharaman TS, et al. (1999) Whole genome shotgun optical mapping of Deinococcus radiodurans. Science 285: 1558-1562.
3. Mahairas GG, Wallace JC, Smith K, Swartzell S, HolzmanT, et al. (1999) Sequence-tagged connectors: A se-quence approach to mapping and scanning the human genome. Proc Natl Acad Sci 96: 9739-9744.
4. Mao L, Wood TC, Yu Y, Budiman MA, Woo SS,et al. (2000) Rice transposable elements: A survey of 73,000 sequence-tagged-connectors (STCs). Genome Res 10: 982-990.
5. Nagano H, Wu L, Kawasaki S, Kishima Y, Sano Y (1999) Genomic organization of the 260 kb surrounding the waxy locus in a japonica rice. Genome 42: 1121-1126.