An Advanced Strategy for \textit{Brassica} Genome Sequencing Using Comparative Genomics with \textit{Arabidopsis}

Tae-Jin Yang$^1$, JungSun Kim$^1$, Ki-Byung Lim$^1$, Soo-Jin Kwon$^1$, Jin-A Kim$^1$, Mina Jin$^1$, Jee Young Park$^1$, Myung-Ho Lim$^1$, Yong-Moon Jin$^1$, Ho-II Kim$^1$, Jang-Ho Hahn$^1$, Yong-Pyo Lim$^2$ and Beom-Seok Park$^1$

1Brassica Genomics Team, National Institute of Agricultural Biotechnology, RDA, Suwon, 441-707, Korea
2Chungnam National University, Daejeon, 305-764, Korea

Keywords: BAC end sequence, pericentromeric heterochromatin, homeologous chromosome, retrotransposon, in silico, CentBr

Abstract

The complete genome sequence of an organism provides unlimited information on the sequencing of other organisms as well as on the related taxa. According to the guidance of the Multinat ional Brassica Genome Project (MBGP), the Korea Brassica Genome Project (KBGP) is aiming to complete the sequencing of chromosome 1. We have selected 48 seed BACs on chromosome 1 using EST genetic markers and FISH analysis. Among them, around 30 BAC clones are sequenced and the comparative genome sequence analysis with their homeologous partner regions on the \textit{Arabidopsis} genome revealed a high resolution comparative map between \textit{Brassica} chromosome 1 and the \textit{Arabidopsis} genome. Most BACs showed collinearity with the homeologous regions of \textit{Arabidopsis}, having an average of 81\% sequence similarity. Sequence comparison of five homeologous BAC clones and their counterparts of \textit{Arabidopsis} indicated that the \textit{Brassica} genome had undergone triplication and subsequent gene losses after speciation with \textit{Arabidopsis}. Based on the information, we developed an advanced method for complete sequencing using a comparative physical map with \textit{Arabidopsis}, a model plant genome. This strategy was applied successfully for in silico chromosome walking and for clone validation on chromosome 1 using BAC end sequences. Additionally, we have defined the (peri)centromeric heterochromatin blocks with centromeric tandem repeat of \textit{Brassica} (CentBr), rDNA, and centromeric retrotransposons of \textit{Brassica} (CRB). The CentBr, rDNA, and CRB appeared to occupy more than 40\% of the \textit{Brassica} genome based on their distribution on 10,284 BAC ends.

INTRODUCTION

The complete genome sequence of an organism provides a large amount of information for biological studies not only on the sequenced organisms but also on closely related taxa. In plants, the genome of the model plant \textit{Arabidopsis} has been sequenced completely by an international collaboration (The Arabidopsis Genome Initiative, 2000).

The genus \textit{Brassica} is one of the core genera including many important vegetables in \textit{Brassicaceae}. Six \textit{Brassica} species are cultivated worldwide; three diploids such as \textit{B. rapa} (AA, 2n=20), \textit{B. nigra} (BB, 2n=16) and \textit{B. oleracea} (CC, 2n=18), and three amphidiploids (allotetraploids) such as \textit{B. juncea} (AABB, 2n=36), \textit{B. napus} (AACC, 2n=38) and \textit{B. carinata} (BBCC, 2n=34) (U, 1935). Chinese cabbage (\textit{B. rapa} ssp. \textit{pekinesis}) is one of the most important vegetables in Korea and in northeast Asia.

The haploid genome equivalent of \textit{B. rapa} is about 500-550 Mb and is closely related with \textit{Arabidopsis} (125 Mb haploid genome equivalent), the completely sequenced model plant. \textit{Arabidopsis} and \textit{Brassica} are speciated around 14.5-20.4 million years ago from a common ancestor (Bowers et al., 2003). Comparative genetic mapping has revealed collinear chromosome segments (Schmidt et al., 2001) in the \textit{Brassicaceae} family and conserved linkage arrangements between \textit{Arabidopsis} and \textit{Brassica}. The genomes of \textit{Brassica} species have triplicated counterparts of the corresponding
homoeologous segments of the *Arabidopsis* genome (O'Neill and Bancroft, 2000).

The Multinational Brassica Genome Project (MBGP) is aiming to completely sequence the genome of *Brassica rapa* inbred line ‘Chiifu’ (http://www.brassica.info). The Korea Brassica Genome Project (KBGP) aimed to sequence the genome of *Brassica rapa* inbred line ‘Chiifu’ (http://www.brassica.info). We have adopted a BAC clone based shotgun sequencing strategy. A conventional physical map anchored on chromosome 1 is not yet ready despite ongoing efforts for fingerprinting of BAC libraries. However, the comparative sequence analysis of the sequenced BAC clones revealing overall collinearity with a homeologous region of the *Arabidopsis* genome suggests that we can use the *Arabidopsis* genome as backbone of *Brassica* physical map if we align BAC end sequences on it. Here we propose an efficient clone validation method for chromosome specific random clone sequencing using comparative physical map and BAC end sequences.

**BRASSICA GENOME SEQUENCING**

*Brassica* Genome Underwent Triplication and was Followed by Subsequent Gene Losses

We have sequenced five BAC clones containing flowering locus C (FLC) gene. Five BACs are located on different chromosomal regions (KBrH052O08 and KBrH117M18 on chromosome 2 long arm, KBrH004D11 on chromosome 6 long arm, KBrH080A08 on chromosome 9 long arm, and KBrH080C09 on chromosome 6 short arm). In-depth sequence analysis of the five BAC clones showed that four (KBrH052O08, KBrH117M18, KBrH004D11, and KBrH080A08) had an overall collinearity as indicated by 82% sequence similarity with an *Arabidopsis* genome sequence, chromosome 5: 3.0-3.3 Mbp. The other one (KBrH080C09) is collinear with another *Arabidopsis* genome sequence, chromosome 5: 25.8-26.2 Mbp. Sequence similarities among four BAC clones are 82-84% except between two BAC clones, 52O08 and 117M18. Both show high level sequence similarity (over 97%) and are located on almost identical chromosomal region, distal region of chromosome 2 long arm based on genetic mapping and FISH analysis, representing they are locally duplicated recently.

Comparison of the common sequence represented in four BAC clones and *Arabidopsis* (chromosome 5: 3.1-3.2 Mbp) reveals that the collinear *Brassica* DNA segments shrank up to 40% of the common sequence (125 kb) of *Arabidopsis*. A total of 36 genes are detected in the 125 kb of *Arabidopsis* sequence. However, only 24, 17, 13, and 13 genes remained in BAC clones, 80A08, 4D11, 52O08, and 117M18, respectively. A total of three genes remained in all four BAC clones with 77-96% a.a. similarity. Newly emerged (inserted) genes including transposons are detected 6, 3, 2, and 1 times in each of the BAC clones, respectively. Overview of the comparative genomic analysis suggests that the *Brassica* genome underwent triplication, followed by subsequent gene losses due to deletions after speciation with *Arabidopsis*.

Characterization of Pericentromeric Heterochromatin Blocks in the *Brassica* Genome

Previous reports represent that 176 bp tandem repeats compose centromeric regions of the *Brassica* genome (Harrison and Heslop-Harrison, 1995). Further sequencing analysis revealed that the 176 bp centromeric tandem repeats subdivided into two classes, named CentBr1 and CentBr2, based on sequence similarity (82-84% between two classes and over 92% in each class) and supporting FISH analysis (Lim et al., 2005). Blast to our BAC end sequence database (10,204 BAC ends of the KBrH library) and to TIGR B. oleracea shotgun sequence database also demonstrated that abundant amount of these repeats exist as tandem arrays and are the major component of centromeric and pericentromeric sequences. We estimated the amount of CentBr based on BAC end sequence data at 30.2% for the whole genome, which is consistent with our FISH showing that CentBr occupy 30.5% of the chromosome area. Furthermore, we have identified several families of centromere specific retrotransposons of *Brassica* (CRB)
from shotgun sequence of two centromeric BAC clones (unpublished data). Among them two long terminal repeat (LTR) retrotransposons, a Ty3-gypsy type (PCRB; 9135 bp with 2047 bp LTR) and a Ty1-copia type (CRB; 6010 bp with 597 bp LTR) are predominantly located at the peri-centromeric heterochromatin of the *Brassica* genome based on FISH analysis.

Based on the BAC end sequence information and FISH analysis, we assume that four (peri) centromeric repeats occupy most of pericentromeric heterochromatin blocks ranging more than 37% of the *Brassica* genome: CentBr (30.2%), rDNA (over 2%), PCRB (1.5%), and CRB (3.5%). These heterochromatin blocks are hard to sequence and remained as gaps even after intensive sequencing efforts in *Arabidopsis* and rice. Therefore, high resolution optical mapping with each of the (peri) centromeric repeats will be the first subject for the heterochromatin region of the *Brassica* genome and then sequencing of euchromatin regions characterized will be accelerated.

**The Strategy for the *Brassica* Genome Sequencing Project**

The Multinational Brassica Genome Project (MBGP) is aiming to sequence the complete genome of *Brassica rapa* using the Korean variety Chiifu (http://www.brassica.info). The Korea Brassica Genome Project (KBGP) is aiming to complete the sequencing of chromosome 1. We have adopted BAC clone based shotgun sequencing strategy using three BAC libraries (*HindIII*, *BamHI*, and *Sau3AI*; All are available through KBGP: NIAB, http://www.brassica-rapa.org and Chungnam National University; http://www.brassicagenome.org). A conventional physical map based on the fingerprints of every BAC clone is the prerequisite for map based sequencing. The *Brassica* physical mapping is ongoing by fingerprinting of KBrH (*HindIII*) and KBrB (*BamHI*) BAC libraries (http://www.brassicagenome.org/ace/brassica_rapa.html). However anchoring of the fingerprints polymorphism contigs (FPC) on the chromosome is another obstacle for physical mapping.

Therefore, we have adopted random clone sequencing starting from EST genetic markers. We have selected 48 seed BACs on chromosome 1 using EST genetic markers and FISH analysis. Among them, around 30 BAC clones are sequenced and they show collinearity with a counterpart homeologous region of *Arabidopsis*, having 82% sequence similarity. The comparative genome sequence analysis with their homeologous partner regions on the *Arabidopsis* genome revealed a high resolution comparative map between *Brassica* chromosome 1 and *Arabidopsis* genome.

Based on the comparative physical map of *Brassica* chromosome 1, we have proposed an efficient and novel clone validation method for sequencing in advance to the complete physical map, which is coined as Clone Validation through In Silico Comparative Physical Mapping (CVSCM). The CVSCM requires a comparative map with the sequenced genome such as *Arabidopsis* instead of its physical map. In addition to the comparative map, BAC end sequences and advanced technique for FISH analysis or genetic mapping with BAC clones selected by in silico physical mapping are prerequisite steps. The comparative map between *Arabidopsis* and *B. rapa* using EST markers and sequenced BACs reveals the collinear chromosomal regions between the two species.

Moreover, we allocated the *Brassica* BAC clones on to the *Brassica* chromosome 1 counterpart in *Arabidopsis* chromosomes based on the BAC end sequences of each of the clones. This comparative in silico physical map provides enormous information for chromosome walking and clone validation of seed BAC clones on chromosome 1 of *B. rapa*. BAC-FISH and STS mapping using BAC end sequences on the comparative in silico physical map allowed us to select BAC clones on chromosome 1 in which no available genetic markers are present. At least one of three BAC clones is confirmed to be on the expected region of chromosome 1 due to triplicated *Brassica* genome. All the sequenced BAC clones provide further starting point for selection of seed BAC clones extending to flanking sides with minimum overlap based on sequence tagged connectors (STC).
CONCLUSION

The use of “model” species in biological research is based on the assumption that many of their features are shared among a wide range of related taxa. The investigation of macrosynteny and microsynteny requires sequences of genomic DNA enabling direct comparison of the sequences using various computational tools (Eckardt, 2001). Thus, the complete Arabidopsis genome sequence and growing lists of genomic resources for other plants have been an incredible boon to comparative genomics research.

Arabidopsis exhibits extensive conserved synteny with species from the closely related genera Brassica (O'Neill and Bancroft, 2000) and Capsella (Acarkan et al., 2000). The Brassica genome seemed to be triplicated from the ancient polyploid, which is evolved to a diploid state similar to the Arabidopsis genome (O'Neill and Bancroft, 2000). Complete sequencing of Brassica rapa will give opportunity to the enlargement of understanding about evolution between genera and rearrangement after polyploidization. Moreover, we propose a novel sequencing strategy for the Brassica genus using comparative genomics with the model plant, Arabidopsis. It may be the biggest fruit and model for sequencing of the highly duplicated genome using comparative genomics with the Arabidopsis genome sequence. Analysis of paralogous genes in duplicated chromosomal regions will give a lot of information to understand the gene dosage effects, heterosis, and functional regulations in the polyploidized genome. Comparative genome analysis between Brassica and Arabidopsis will give great impacts on understanding the evolutionary process in the polyploidized genome. Furthermore, it will give a great opportunity on agricultural aspects, especially for breeding and molecular farming through finding novel or useful genes. The genus Brassica is one of the core genera including many important vegetables in Brassicaceae. Therefore the genome sequence of Brassica rapa will give great chance to understand other important crops in the genus Brassica.

ACKNOWLEDGEMENTS

This work was partly supported by a grant from BioGreen 21 Program, RDA, Korea.

Literature Cited


