

## Review Article

# Progress in Understanding and Sequencing the Genome of *Brassica rapa*

Chang Pyo Hong,<sup>1</sup> Soo-Jin Kwon,<sup>2</sup> Jung Sun Kim,<sup>2</sup> Tae-Jin Yang,<sup>3</sup> Beom-Seok Park,<sup>2</sup> and Yong Pyo Lim<sup>1</sup>

<sup>1</sup> Department of Horticulture, College of Agriculture and Life Science, Chungnam National University, Daejeon 305764, South Korea

<sup>2</sup> Brassica Genomics Team, National Institute of Agricultural Biotechnology (NIAB), Rural Development Administration (RDA), Suwon 441707, South Korea

<sup>3</sup> Department of Plant Science, College of Agriculture and Life Sciences, Seoul National University, Seoul 151921, South Korea

Correspondence should be addressed to Yong Pyo Lim, yplim@cnu.ac.kr

Received 13 June 2007; Accepted 21 November 2007

Recommended by Yunbi Xu

*Brassica rapa*, which is closely related to *Arabidopsis thaliana*, is an important crop and a model plant for studying genome evolution via polyploidization. We report the current understanding of the genome structure of *B. rapa* and efforts for the whole-genome sequencing of the species. The tribe *Brassicaceae*, which comprises ca. 240 species, descended from a common hexaploid ancestor with a basic genome similar to that of *Arabidopsis*. Chromosome rearrangements, including fusions and/or fissions, resulted in the present-day “diploid” *Brassica* species with variation in chromosome number and phenotype. Triplicated genomic segments of *B. rapa* are collinear to those of *A. thaliana* with InDels. The genome triplication has led to an approximately 1.7-fold increase in the *B. rapa* gene number compared to that of *A. thaliana*. Repetitive DNA of *B. rapa* has also been extensively amplified and has diverged from that of *A. thaliana*. For its whole-genome sequencing, the *Brassica rapa* Genome Sequencing Project (BrGSP) consortium has developed suitable genomic resources and constructed genetic and physical maps. Ten chromosomes of *B. rapa* are being allocated to BrGSP consortium participants, and each chromosome will be sequenced by a BAC-by-BAC approach. Genome sequencing of *B. rapa* will offer a new perspective for plant biology and evolution in the context of polyploidization.

Copyright © 2008 Chang Pyo Hong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## 1. IMPORTANCE OF BRASSICA GENOMICS

The genus *Brassica* is one of the core genera in the tribe *Brassicaceae* and includes a number of crops with wide adaptation under a variety of agroclimatic conditions. Economically, *Brassica* species are important sources of vegetable oil, fresh, preserved vegetables, and condiments. *B. napus*, *B. rapa*, *B. juncea*, and *B. carinata* provide about 12% of the worldwide edible vegetable oil supply [1]. The *B. rapa* and *B. oleracea* subspecies represent many of the vegetables in our daily diet. In particular, *B. rapa* ssp. *pekinensis* (Chinese cabbage), on which this article focuses, is one of the most widely used vegetable crops in northeast Asia. Moreover, *Brassica* species are important sources of dietary fiber, vitamin C, and anticancer compounds [2].

The genetic relationships among the different diploid and amphidiploid *Brassica* species are described by the U’s triangle [3]. Of the six widely cultivated species of *Brassica*, *B. rapa* (AA,  $2n = 20$ ), *B. nigra* (BB,  $2n = 16$ ), and *B. oleracea*

(CC,  $2n = 18$ ) are monogenomic diploids. The remaining three species, *B. juncea* (AABB,  $2n = 36$ ), *B. napus* (AACC,  $2n = 38$ ), and *B. carinata* (BBCC,  $2n = 34$ ) exhibit stable diploid genetics, but are allotetraploids, which have evolved via hybridization between differing monogenomic diploids [3]. The diploid *Brassica* genomes range from 1.1 pg/2C (529 Mbp/1C) for *B. rapa* to 1.4 pg/2C (696 Mbp/1C) for *B. oleracea* (see Figure 1) [4]. The genomes of the allotetraploids range from 2.2 pg/2C (1,068 Mbp/1C) for *B. juncea* to 2.6 pg/2C (1,284 Mbp/1C) for *B. carinata* (see Figure 1).

The genus *Brassica* is characterized by morphological diversity with regard to inflorescences, leaves, stems, roots, and terminal or apical buds [5]. For example, such morphological diversity can be easily observed in subspecies of *B. oleracea*: the enlarged inflorescences of cauliflower (*B. oleracea* ssp. *botrytis*) and broccoli (*B. oleracea* ssp. *italica*); the enlarged stems of kohlrabi (*B. oleracea* ssp. *gongylodes*) and marrowstem kale (*B. oleracea* ssp. *medullosa*); and the many axillary buds of Brussels sprout (*B. oleracea* ssp. *gemmifera*)

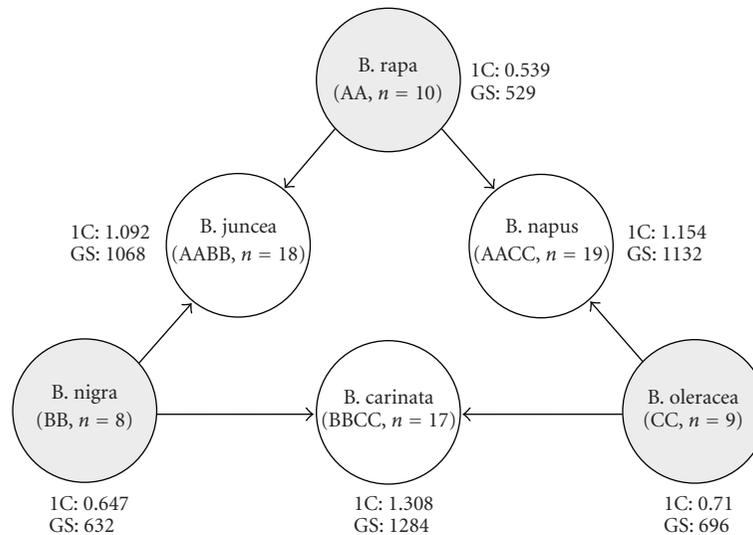


FIGURE 1: Genetic relationship of the different diploid and amphidiploid *Brassica* species. 1C, 1C nuclear DNA content (pg); GS, genome size (Mbp) [3, 4].

[5]. The morphological diversity in *Brassica* species may be linked to genomic changes associated with polyploidization [6]. The polyploidization in *Brassica* species has brought about triplication of genomic segments and subsequent rearrangements such as inversions, insertions, deletions, and substitutions [7–16], and these genetic variations may cause novel phenotypic variations for traits among these species [5, 6]. Thus, *Brassica* genomics will provide us with an understanding of the rapid phenotypic evolution of polyploid plants. Additionally, it will help us to understand genomic changes and how they shape the allotetraploid *Brassica* species. For example, a study has been done looking at rapid genomic changes and the effect of nuclear-cytoplasm interaction in synthetic allotetraploid species [17].

Because of the high economic value of *Brassica* species throughout the world and their potential to be models for the study of polyploidization, genome sequencing projects for *Brassica* species, especially *B. rapa* and *B. oleracea*, have recently been initiated (<http://www.brassica.info>) [18–20]. In particular, *B. rapa* ssp. *pekinensis* inbred line Chiffu-401-42, discussed in this article, has been selected for *Brassica*-A genome sequencing in the *Brassica rapa* Genome Sequencing Project (*BrGSP*) (<http://www.brassica.info>), a component of the consortium of the Multinational *Brassica* Genome Project, with the goal of completely sequencing this genome through a BAC-by-BAC approach. The *BrGSP* consortium has developed genomic resources for this purpose and is proceeding with whole-genome sequencing.

## 2. CURRENT UNDERSTANDING OF THE GENOME STRUCTURE OF *B. RAPA*

### 2.1. Karyotype of *B. rapa*

Karyotyping is the starting point for understanding the genome structure of a species. Moreover, it provides insight

into genome evolution. Most of the karyotypic analyses in *B. rapa* have been performed on mitotic metaphase chromosomes [21–24]. However, the analyses are limited in what they can reveal about the cytological structure of the genome because of the low resolution of the technique. For example, different measurements of chromosome lengths and rDNA loci are obtained by this method. Recently, the high-resolution karyotype for the *B. rapa* ssp. *pekinensis* inbred line Chiffu was determined on pachytene chromosomes by using 4'-6-diamino-2-phenylindole dihydrochloride (DAPI) staining and fluorescence in situ hybridization (FISH) of rDNAs and pericentromeric satellite repeats [25]. By DAPI analysis, the mean lengths of ten pachytene chromosomes ranged from 23.7  $\mu\text{m}$  to 51.3  $\mu\text{m}$ , with a total of 385.3  $\mu\text{m}$ , a total length which is 11.9- ~ 17.5-fold longer than that of the mitotic metaphase chromosomes reported by Lim et al. [24] and Koo et al. [25]. In comparison, pachytene chromosome length of *A. thaliana*, *Medicago truncatula*, and tomato was estimated to be about 7.4%, 15%, and 24% of the total pachytene chromosome length, respectively (reviewed in Koo et al. [25]). In *B. rapa*, the pachytene karyotype consists of two metacentric (chromosomes 1 and 6), five submetacentric (chromosomes 3, 4, 5, 9, and 10), two subtetacentric (chromosomes 7 and 8), and one acrocentric chromosome (chromosome 2), with the corresponding centromeric index ranges of 38.8–41.0%, 29.5–36.7%, 17.4–20.2%, and 9.38%, respectively [25]. In the chromosomal structure at pachytene, the total length of pericentromeric heterochromatin regions was estimated to be 38.2  $\mu\text{m}$ , which is approximately 10% of the total chromosome length [25]. In conjunction with chromosomal structure and characteristics, 5S rDNA loci were located on pericentromeric regions of the short arms of chromosomes 2 and 7 as well as the long arm of chromosome 10, while 45S rDNA loci were located on the short arms of chromosomes 1, 2, 4, and 5 as well as the long arm of chromosome 7 [24, 25].

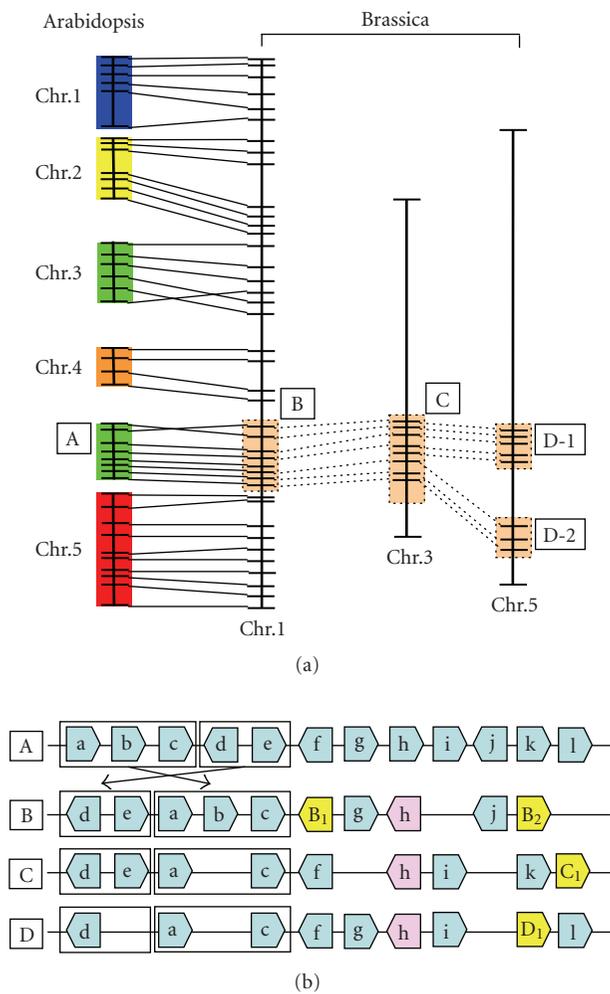


FIGURE 2: An example of a comparative map of *Arabidopsis* and *Brassica*. (a) Collinearity between genomic segments of the two species and genome triplication of *Brassica* revealed by comparative genetic mapping. (b) Synteny of genes in a triplicated genomic region of *Brassica*.

## 2.2. Collinearity between genomic segments of *Arabidopsis* and *Brassica*

*Brassica* species are closely related to *A. thaliana*, having diverged 17–18 million years ago (MYA) from their common ancestor [16]. *A. thaliana*, which has been completely sequenced, has a rather small genome (about 146 Mb) with relatively little repetitive DNA and a high gene density [26, 27]. Protein-coding regions of the genomes of *Brassica* species show high sequence conservation with those of *A. thaliana*, with nucleotide sequence similarity in exons between *B. oleracea* and *A. thaliana* ranging from 75% to 90%, compared to <70% for intronic regions [28]. This similarity allows the identification of sets of candidate genes in *Brassica* species and the studying of their genome structures through comparative genomics [29]. Comparative studies between *Arabidopsis* and *Brassica* have revealed the presence of collinear chromosome segments (see Figure 2). Comparative genetic mapping between diploid *Brassica* species and *A. thaliana*

to identify homologous loci have revealed many conserved blocks in their genomes [7, 8, 14, 30]. Comparative physical mapping between *Arabidopsis* and *Brassica* further corroborated the findings. A set of six bacterial artificial chromosomes (BACs), representing a 431-kb contiguous region of *Arabidopsis* chromosome 2, was mapped on chromosomes and DNA fibers of *B. rapa* [31]. Moreover, studies on a 222-kb gene-rich region of *A. thaliana* chromosome 4 and its homologous counterparts in *B. rapa* or *B. oleracea* revealed the collinearity of genes in homologous segments [9, 11, 13]. This finding was supported by sequence analysis of specific homologous genomic segments [15, 16]. However, many structural rearrangements differentiate the *Brassica* and *Arabidopsis* chromosomes (see Figure 2). Comparative genetic mapping between *B. nigra* and *A. thaliana* species revealed that the average length of conserved segments between the two species was estimated at about 8 cM, which corresponds to ~90 rearrangements since the divergence of the two species [7]. In addition, it was found that gene contents in their homologous genomic segments were also variable with interstitial gene losses and insertions [9, 11, 13, 15, 16].

## 2.3. Genome triplication of diploid *Brassica* species

Most of the comparative studies mentioned above demonstrated that *Brassica* species contain extensively triplicated counterparts of the corresponding homologous segments of the *A. thaliana* genome (see Figure 2), thereby suggesting that diploid *Brassica* species may have been derived from a hexaploid ancestor: the genome which was similar to *Arabidopsis*. Consistent with the nature of genome triplication, Yang et al. [16] reported that paralogous subgenomes of diploid *Brassica* species triplicated 13 ~ 17 MYA, very soon after the *Arabidopsis* and *Brassica* divergence that occurred at 17 ~ 18 MYA. In addition, it was reported that after the *Brassica* genomes had triplicated, their subgenomes were rearranged by inversions, translocations [7, 12, 32], extensive interspersed gene loss, as well as gene insertions occurred relative to the inferred structure of the ancestral genome (see Figure 2). Additionally, such genome triplication was extensively found across the tribe *Brassicaceae* [12]. In comparison with the genome of *A. thaliana*, the genome triplication in *Brassica* species has clearly led to an increase in the genome size, resulting in a 3- to 5-fold inflation.

Genome triplication events in *Brassica* species may also have an effect on gene expression of multicopy genes, leading to such phenomena as pseudogenization, subfunctionalization, or neofunctionalization in species [33–38]. For example, the MADS-box transcription factor family, whose members control key aspects of plant vegetative and reproductive development, shapes genetic systems by subfunctionalization [37]. It appears that after polyploid formation, considerable and sometimes very rapid changes in genome structure and gene expression have occurred. Researchers have hypothesized that genomic triplication in *Brassica* species permits mutations in loci that are normally under tight selective constraints in *Arabidopsis*, and may thus result in the observed

TABLE 1: Comparison of gene, TE, and SSR abundances in *B. rapa* and *A. thaliana*.

Contents	<i>B. rapa</i>	<i>A. thaliana</i>	References
Genome size (Mb)	529	146	[4, 27]
Gene number	4300 ~ 63000	26,207	[16, 39, 43]
TE abundance (%) <sup>(1)</sup>	13.8	6 ~ 7	[39, 44]
SSR number	≈110,000 (one SSR/4.8 kb)	≈36,756 (one SSR/3.2 kb)	[39, 45]

<sup>(1)</sup> Coverage of TEs in the genome.

greater phenotypic plasticity in *Brassica* [5]. Studies on expression of duplicated genes in *Brassica* species will provide insight into the role of polyploidization in the phenotypic divergence of the plant genus.

#### 2.4. Survey of the *B. rapa* genome revealed by BAC-end sequence analysis

The *B. rapa* genome was surveyed via the analysis of its 12,017 *Hind*III BAC-end sequences (Table 1) [39]. Analyses of BAC-end sequence or genome survey sequences assist in understanding whole genome structure [39–41]. It was estimated that the *B. rapa* genome might contain about 43000 genes (covering 16.8% of the genome), 1.6 times more than the *A. thaliana* genome. Recently, Yang et al. [16] also estimated the gene content of *B. rapa* to range from 49,000 to 63,000, based on predictions from microsynteny studies. It has been suggested that chromosomal triplication events in *Brassica* have led to an increase in gene number with subsequent gene loss [15, 16, 39, 42].

Transposable elements (TEs) with a predominance of retrotransposons were estimated to occupy approximately 14% of the genome (covering approximately 74 Mb), 8.2 times greater than that observed previously in *A. thaliana* [44]. Zhang and Wessler [44] reported that TEs in *B. oleracea* constituted 20% of the genome, slightly more than what was predicted for the *B. rapa* genome. Of the predicted TEs, LTR retrotransposon families were the most abundant (69.9%), followed by non-LTR retrotransposons (13.4%), DNA transposons (11.4%), and other retrotransposons (5.3%). In particular, *Ty1/copia*-like and *Ty3/gypsy*-like retrotransposons occupied 39.5% and 30.2% of LTR retrotransposon families, respectively. The amplification of TEs in *B. rapa*, especially retrotransposons, may have played a crucial role in both evolution and genomic expansion.

Simple sequence repeats (SSRs) have been estimated to occur with a frequency of approximately one per 4.8 kb within the *B. rapa* genome, as compared to approximately one per 3.2 kb within the *A. thaliana* genome [39]. Of SSRs identified, trinucleotides were the most abundant repeat type, constituting about 37% of all SSRs, a percentage similar to those reported in other plant genomes [39, 45]. Comparison of SSR densities in different genomic regions demonstrated that SSR density was greatest immediately in 5'-flanking regions of predicted genes [45]. SSRs were also preferentially associated with gene-rich regions, with pericentromeric heterochromatin SSRs mostly associated with retrotransposons [45], suggesting that the distribution of SSRs in the genome is nonrandom [39, 45].

#### 2.5. Structure of (peri)centromeres of *B. rapa*

The centromere is a dynamic and rapidly evolving structure and consists largely of highly repetitive DNA sequences, especially tandem satellite repeats and retrotransposons [46, 47]. Centromeric repeats characterized in plant genomes are composed of 155 ~ 180-bp tandem repeat motifs, including the 180-bp pAL1 satellite in *A. thaliana* [48–50], the 155 ~ 165-bp CentO satellite in rice [51, 52], the 156-bp CentC satellite in maize [53] and the 169-bp satellite in *Medicago truncatula* [54, 55]. Centromeric satellite repeats of *Brassica* species, except for those of *B. nigra*, are represented by the 176-bp CentBr [24, 25, 56–59]. The CentBr repeats in the *B. rapa* genome belong to two classes which have 82% sequence similarity. The two classes are chromosome-specific, with CentBr1 found on eight chromosomes (chromosomes 1, 3, and 5–10) and CentBr2 on two chromosomes (chromosomes 2 and 4) [24, 25, 39]. Such distribution of the CentBr family may reflect the predominance of CentBr1 in the *Brassica* genome [39]. The CentBr repeats have also undergone rapid evolution within the *B. rapa* genome and have diverged among the related species of *Brassicaceae* [39]. Recently, Lim et al. [59] identified and characterized the major repeats in centromeric and pericentromeric heterochromatin of *B. rapa*. The region contains CentBr arrays, 238-bp degenerate tandem repeat (TR238) arrays, rDNAs, centromere-specific retrotransposons of *Brassica* (CRB), and pericentromere-specific retrotransposons (PCRBr). In particular, CRB was a major component of all centromeres in three diploid *Brassica* species and their allotetraploid relatives, and PCRBr and TR238 were A-genome-specific [59].

### 3. PROGRESS OF *B. RAPA* GENOME SEQUENCING

#### 3.1. Genomic resources

The development of genomic resources is a prerequisite to undertaking genome sequencing in any crop species. Genomic resources, including reference mapping populations, DNA libraries, and DNA sequences have been developed for *B. rapa* ssp. *pekinensis* inbred line Chiffu-401-42 (Table 2). Two reference mapping populations were derived from two *B. rapa* ssp. *pekinensis* inbred lines, Chiifu-401-42 and Kenshin-402-43 (CK), and comprise 78 double haploid (DH) lines (the CKDH population) and 201 recombinant inbred (RI) lines (the CKRI population). These mapping populations have been used for construction of reference genetic maps for genome sequencing [20]. The bacterial artificial chromosome (BAC) system, commonly used for

TABLE 2: Genomic resources for whole-genome sequencing of *B. rapa*.

Genomic resources	Source material	Number
Mapping populations		
DH line	Chiffu-401-42 × Kenshin-402-43	78 lines (F <sub>2</sub> generation)
RI line	Chiffu-401-42 × Kenshin-402-43	201 lines (F <sub>8</sub> generation)
BAC libraries		
<i>Hind</i> III (KBrH)	Chiffu-401-42	56592 clones (115 kb <sup>(1)</sup> )
<i>Bam</i> HI (KBrB)	Chiffu-401-42	50688 clones (124 kb <sup>(1)</sup> )
<i>Sau</i> 3AI (KBrS)	Chiffu-401-42	55296 clones (100 kb <sup>(1)</sup> )
cDNA libraries		
22 cDNA libraries	Different tissues of Chiffu-401-42 and Jangwon including leaves, roots, cotyledons, stems, seedlings, ovules, siliques, anthers	—
BAC-end sequences	KBrH, KBrB, and KBrS clones	200017 sequences
ESTs	22 cDNA clones	129928 sequences
BAC shotgun sequences	KBrH, KBrB, and KBrS clones	on-going <sup>(2)</sup>

<sup>(1)</sup>Average insert size (kb).

<sup>(2)</sup>Of BACs sequenced, 511 BACs have been deposited in GenBank.

TABLE 3: Genetic linkage maps of *B. rapa* developed since 1990.

Mapping population	Population type	Population size	No. of loci	Type of markers	Total length of map (average interval)	References
Michihili × Spring broccoli	F <sub>2</sub>	95	280	RFLP	1850 cM (6.6 cM)	[62]
Per (winter turnip rape) × R500 (spring yellow sarson)	F <sub>2</sub>	91	139	RFLP	1785 cM (13.5 cM)	[63]
Per (winter turnip rape) × R500 (spring yellow sarson)	F <sub>6</sub> RI	87	144	RFLP	890 cM (6.0 cM)	[64]
Developed from Chinese cabbage F <sub>1</sub> cultivar Jangwon	F <sub>2</sub>	134	545	RFLP, SSR	1287 cM (2.4 cM)	[65]
G004 (CR <sup>(a)</sup> DH line) × A9709 (CS <sup>(b)</sup> DH line)(cultivars of Chinese cabbage)	F <sub>2</sub>	94	262	RFLP, SSR, RAPD	1005 cM (3.7 cM)	[66]
Chiffu-401-42 × Kenshin-402-43	DH	78	556	AFLP, SSR, RADP, ESTP, STS, CAPS	1182 cM (2.83 cM)	[67]

developing large-insert DNA libraries, is an invaluable resource for structural and functional genomics. Three Chiffu BAC libraries were constructed by using restriction enzymes: *Hind*III, *Bam*HI, and *Sau*3AI, and designated as KBrH, KBrB, and KBrS. These libraries consist of 56592, 50688, and 55296 clones with an average insert size of 115 kb, 124 kb, and 100 kb, respectively. These BAC libraries cover approximately 36 genome equivalents, assuming that the genome size of Chinese cabbage is 529 Mb. Using these BAC clones, the *Br*GSP community has recently generated a total of 200017 BAC-end sequences. In combination with BAC fingerprinting data, the BAC-end sequences will give insight into the structure of the genome, be a resource for development of genetic markers, and aid in finding the BAC clones that correspond to the minimal tiling paths in genome sequencing [19, 60, 61]. For functional genomics of *B. rapa*, 22 cDNA libraries from different tissues, including leaves, roots, cotyledons, stems, seedlings, ovules, siliques, and anthers of Chiffu, have been constructed, and a total of 128582 expressed sequence tags (ESTs) have been generated from these cDNA li-

braries (GenBank accession number CO749247 ~ CO750684 and EX015357 ~ EX142500). Currently, the ESTs have been used for construction of *B. rapa* unigene set and gene expression microarray (<http://www.brassica-rapa.org>).

### 3.2. Genetic and physical mapping

Some genetic linkage maps of *B. rapa*, on which genetic markers were distributed over ten linkage groups, have been constructed since 1990 [62–67] (summarized in Table 3). The distances of genetic linkage maps ranged from 890 cM to 1850 cM. However, the genetic linkage maps may not provide direct and accurate genetic information for the Chiffu genome sequencing because of genetic variation between the mapping populations. For that reason, the *Br*GSP community has constructed the CK genetic linkage map. Using the 78 CKDH lines, a reference genetic linkage map has been constructed [67]. The map consists of a total of 556 markers, including 278 AFLPs, 235 SSRs, 25 RAPDs, and 18 ESTPs/STS/CAPS markers. Ten linkage groups were

TABLE 4: The correspondence between genetic linkage groups of *B. rapa* ssp. *pekinensis* based on *B. napus* reference linkage maps.

Genetic linkage map of <i>B. napus</i> [68]	Choi et al. [67]	Kim et al. [65]	Suwabe et al. [66]
A1 (N1)	R1	R1	LG6
A2 (N2)	R2	R2	LG8
A3 (N3)	R3	R3	LG1
A4 (N4)	R4	R4	LG10
A5 (N5)	R5	R5	LG3
A6 (N6)	R6	R6	LG2
A7 (N7)	R7	R7	LG4
A8 (N8)	R8	R8 </td <td>LG7</td>	LG7
A9 (N9)	R9	R9	LG5
A10 (N10)	R10	R10	LG9

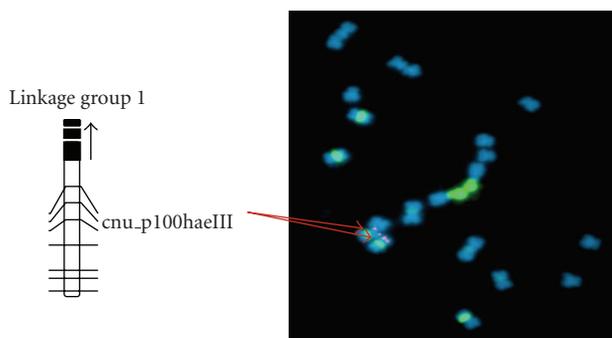


FIGURE 3: An example of an alignment of linkage group 1 in the reference genetic map to the corresponding chromosome 5 through FISH using locus-specific BAC clones.

identified and designated as R1 to R10 via mapping with SSR markers derived from the reference linkage map of *B. napus* reported previously [68] (Table 4). The total length of the linkage map was 1182 cM with an average interval of 2.83 cM between adjacent loci. Recently, for high-resolution genetic mapping, the community has set a goal of developing more than 1,000 SSR markers derived from BAC-end sequences, ESTs, and BACs. Moreover, based on the sequence-tagged site (STS) markers, a CKRI genetic linkage map has been constructed to be complementary to the CKDH one. The linkage groups in these genetic maps may not correspond to the chromosomes assigned in the cytogenetic map. Therefore, it is important to align the linkage groups on the genetic map with chromosomes of the cytogenetic map. All ten linkage groups of a reference genetic map of *B. rapa* are being assigned to the corresponding chromosomes through fluorescence in situ hybridization (FISH) using locus-specific BAC clones as probes (see an example in Figure 3, unpublished data).

The fingerprinted BAC map (so-called “physical map”) makes it possible to select clones for sequencing that would ensure comprehensive coverage of the genome and reduce sequencing redundancy [69]. In addition, the clone-based

map also enables the identification of large segments of the genome that are repeated, thereby simplifying the sequence assembly. To construct a deep-coverage BAC physical map of the *B. rapa* genome, all BAC clones from the three BAC libraries were fingerprinted using restriction enzyme digestion and SNaPshot [70] methodologies, and then BAC contigs have been assembled by FingerPrinted Contigs (FPC) software (<http://www.agcol.arizona.edu/software/fpc/>). This data will be open to the *Brassica rapa* genome sequencing consortium.

### 3.3. Approach to genome sequencing

Seed BACs for genome sequencing have been selected through in silico allocation of *B. rapa* BAC-end sequences onto counterpart locations of *Arabidopsis* chromosomes [19]. Of 91000 BAC-end sequences, a total of 45232 showed significant sequence similarity with unique *Arabidopsis* sequences, and 4317 BAC clones were allocated on *Arabidopsis* chromosomes by significant matching with both ends within 30–500 kb intervals, which span 93 Mb of *Arabidopsis* euchromatin regions (covering 78.2% of the *Arabidopsis* genome). However, approximately 9.4 Mb of euchromatin regions and 16.6 Mb pericentromeric heterochromatin regions of the *Arabidopsis* genome were not covered by the *B. rapa* BAC span (span is considered by best hit of paired ends). Based on the physical map of *B. rapa* and the in silico comparative map of its BAC-ends onto *Arabidopsis* chromosomes, 629 seed BACs have been selected spanning 86 Mb of *Arabidopsis* euchromatin regions and scattered throughout the *B. rapa* genome (<http://www.brassica-rapa.org>), and the BACs have been mapped on *B. rapa* chromosomes by STS mapping and FISH analysis. The seed BACs which are anchored and sequenced will be used as stepping stones for sequencing of the ten chromosomes.

Considering the large genome size and the possibility of international cooperation, a chromosome-based approach was suggested. Of ten chromosomes (or linkage groups), eight have been allocated to the participating countries as follow: Korea (R3 and R9), Canada (R2 and R10), UK and China (R1 and R8), USA (R6), and Australia (R7). However, R4 and R5 have remained unassigned. Progress of chromosome sequencing will be reported soon by each country.

## 4. CONCLUSIONS

*Brassica* species are economically important crops and serve as model plants for studying phenotypic evolution associated with polyploidization. The *Brassica* genomes have extensively triplicated and undergone subsequent genome rearrangements with sequence variations. This has significantly affected their genome structure and may underline phenotypic diversity. Genome sequencing of *B. rapa* can pave the way for elucidation of the relationship between genome evolution and phenotypic diversity. Moreover, it enables us to search for genes and develop molecular markers associated with agricultural traits, thereby establishing a molecular breeding system contributing to improvement of *Brassica* species economically.

## ACKNOWLEDGMENTS

This research was supported by grants from Rural Development Administration (BioGreen 21 Program), National Institute of Agricultural Biotechnology (Project no. 04-1-12-2), and Korean Science and Engineering Foundation (R21-2004-000-10010-0), South Korea.

## REFERENCES

- [1] K. S. Labana and M. L. Gupta, "Importance and origin," in *Breeding Oilseed Brassicas*, K. S. Labana, S. S. Banga, and S. K. Banga, Eds., Springer, Berlin, Germany, 1993.
- [2] J. Fahey and P. Talalay, "The role of crucifers in cancer chemoprotection," in *Phytochemicals and Health*, D. L. Gustine and H. E. Florens, Eds., pp. 87–93, American Society of Plant Physiologists, Rockville, Md, USA, 1995.
- [3] N. U, "Genome analysis in Brassica with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization," *Japanese Journal of Botany*, vol. 7, no. 3, pp. 389–452, 1935.
- [4] J. S. Johnston, A. E. Pepper, A. E. Hall, et al., "Evolution of genome size in Brassicaceae," *Annals of Botany*, vol. 95, no. 1, pp. 229–235, 2005.
- [5] A. H. Paterson, T. H. Lan, R. Amasino, T. C. Osborn, and C. Quiros, "Brassica genomics: a complement to, and early beneficiary of, the *Arabidopsis* sequence," *Genome Biology*, vol. 2, no. 3, pp. 1011.1–1011.4, 2001.
- [6] L. N. Lukens, P. A. Quijada, J. Udall, J. C. Pires, M. E. Schranz, and T. C. Osborn, "Genome redundancy and plasticity within ancient and recent Brassica crop species," *Biological Journal of the Linnean Society*, vol. 82, no. 4, pp. 665–674, 2004.
- [7] U. Lagercrantz, "Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that *Brassica* genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements," *Genetics*, vol. 150, no. 3, pp. 1217–1228, 1998.
- [8] T. H. Lan, T. A. Delmonte, K. P. Reischmann, et al., "An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*," *Genome Research*, vol. 10, no. 6, pp. 776–788, 2000.
- [9] C. M. O'Neill and I. Bancroft, "Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*," *The Plant Journal*, vol. 23, no. 2, pp. 233–243, 2000.
- [10] D. Babula, M. Kaczmarek, A. Barakat, M. Delseny, C. F. Quiros, and J. Sadowski, "Chromosomal mapping of *Brassica oleracea* based on ESTs from *Arabidopsis thaliana*: complexity of the comparative map," *Molecular Genetics and Genomics*, vol. 268, no. 5, pp. 656–665, 2003.
- [11] D. Rana, T. van den Boogaart, C. M. O'Neill, et al., "Conservation of the microstructure of genome segments in *Brassica napus* and its diploid relatives," *The Plant Journal*, vol. 40, no. 5, pp. 725–733, 2004.
- [12] M. A. Lysak, M. A. Koch, A. Pecinka, and I. Schubert, "Chromosome triplication found across the tribe Brassiceae," *Genome Research*, vol. 15, no. 4, pp. 516–525, 2005.
- [13] J. Y. Park, D. H. Koo, C. P. Hong, et al., "Physical mapping and microsynteny of *Brassica rapa* ssp. *pekinensis* genome corresponding to a 222 kbp gene-rich region of *Arabidopsis* chromosome 4 and partially duplicated on chromosome 5," *Molecular Genetics and Genomics*, vol. 274, no. 6, pp. 579–588, 2005.
- [14] I. A. P. Parkin, S. M. Gulden, A. G. Sharpe, et al., "Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*," *Genetics*, vol. 171, no. 2, pp. 765–781, 2005.
- [15] C. D. Town, F. Cheung, R. Maiti, et al., "Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy," *The Plant Cell*, vol. 18, no. 6, pp. 1348–1359, 2006.
- [16] T. J. Yang, J. S. Kim, S. J. Kwon, et al., "Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of *Brassica rapa*," *The Plant Cell*, vol. 18, no. 6, pp. 1339–1347, 2006.
- [17] L. N. Lukens, J. C. Pires, E. Leon, R. Vogelzang, L. Oslach, and T. Osborn, "Patterns of sequence loss and cytosine methylation within a population of newly resynthesized *Brassica napus* allopolyploids," *Plant Physiology*, vol. 140, no. 1, pp. 336–348, 2006.
- [18] M. Ayele, B. J. Haas, N. Kumar, et al., "Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis*," *Genome Research*, vol. 15, no. 4, pp. 487–495, 2005.
- [19] T. J. Yang, J. S. Kim, K. B. Lim, et al., "The Korea Brassica genome project: a glimpse of the Brassica genome based on comparative genome analysis with *Arabidopsis*," *Comparative and Functional Genomics*, vol. 6, no. 3, pp. 138–146, 2005.
- [20] Y. P. Lim, P. Plaha, S. R. Choi, et al., "Towards unraveling the structure of *Brassica rapa* genome," *Physiologia Plantarum*, vol. 126, no. 4, pp. 585–591, 2006.
- [21] B. F. Cheng, W. K. Heneen, and B. Y. Chen, "Mitotic karyotypes of *Brassica campestris* and *Brassica alboglabra* and identification of the *B. alboglabra* chromosome in an addition line," *Genome*, vol. 38, no. 2, pp. 313–319, 1995.
- [22] K. Fukui, S. Nakayama, N. Ohmido, H. Yoshiaki, and M. Yamabe, "Quantitative karyotyping of three diploid *Brassica* species by imaging methods and localization of 45s rDNA loci on the identified chromosomes," *Theoretical and Applied Genetics*, vol. 96, no. 3–4, pp. 325–330, 1998.
- [23] R. J. Snowdon, T. Friedrich, W. Friedt, and W. Köhler, "Identifying the chromosomes of the A- and C-genome diploid *Brassica* species *B. rapa* (syn. *campestris*) and *B. oleracea* in their amphidiploid *B. napus*," *Theoretical and Applied Genetics*, vol. 104, no. 4, pp. 533–538, 2002.
- [24] K. B. Lim, H. de Jong, T. J. Yang, et al., "Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassica rapa*," *Molecules and Cells*, vol. 19, no. 3, pp. 436–444, 2005.
- [25] D. H. Koo, P. Plaha, Y. P. Lim, Y. Hur, and J.-W. Bang, "A high-resolution karyotype of *Brassica rapa* ssp. *pekinensis* revealed by pachytene analysis and multicolor fluorescence in situ hybridization," *Theoretical and Applied Genetics*, vol. 109, no. 7, pp. 1346–1352, 2004.
- [26] The Arabidopsis Genome Initiative, "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*," *Nature*, vol. 408, no. 6814, pp. 796–815, 2000.
- [27] M. Bevan and S. Walsh, "The *Arabidopsis* genome: a foundation for plant research," *Genome Research*, vol. 15, no. 12, pp. 1632–1642, 2005.
- [28] C. F. Quiros, F. Grellet, J. Sadowski, T. Suzuki, G. Li, and T. Wroblewski, "Arabidopsis and Brassica comparative genomics: sequence, structure and gene content in the *ABI1-Rps2-Ck1* chromosomal segment and related regions," *Genetics*, vol. 157, no. 3, pp. 1321–1330, 2001.
- [29] J. A. Kim, T. J. Yang, J. S. Kim, et al., "Isolation of circadian-associated genes in *Brassica rapa* by comparative genomics

- with *Arabidopsis thaliana*,” *Molecules and Cells*, vol. 23, no. 2, pp. 145–153, 2007.
- [30] S. P. Kowalski, T. H. Lan, K. A. Feldmann, and A. H. Paterson, “Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization,” *Genetics*, vol. 138, no. 2, pp. 499–510, 1994.
- [31] S. A. Jackson, Z. Cheng, M. L. Wang, H. M. Goodman, and J. Jiang, “Comparative fluorescence in situ hybridization mapping of a 431-kb *Arabidopsis thaliana* bacterial artificial chromosome contig reveals the role of chromosomal duplications in the expansion of the *Brassica rapa* genome,” *Genetics*, vol. 156, no. 2, pp. 833–838, 2000.
- [32] P. A. Ziolkowski, M. Kaczmarek, D. Babula, and J. Sadowski, “Genome evolution in *Arabidopsis/Brassica*: conservation and divergence of ancient rearranged segments and their breakpoints,” *The Plant Journal*, vol. 47, no. 1, pp. 63–74, 2006.
- [33] M. Lynch and J. S. Conery, “The evolutionary fate and consequences of duplicate genes,” *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.
- [34] A. Lawton-Rauh, “Evolutionary dynamics of duplicated genes in plants,” *Molecular Phylogenetics and Evolution*, vol. 29, no. 3, pp. 396–409, 2003.
- [35] Z. Gu, S. A. Rifkin, K. P. White, and W.-H. Li, “Duplicate genes increase gene expression diversity within and between species,” *Nature Genetics*, vol. 36, no. 6, pp. 577–579, 2004.
- [36] K. L. Adams and J. F. Wendel, “Novel patterns of gene expression in polyploid plants,” *Trends in Genetics*, vol. 21, no. 10, pp. 539–543, 2005.
- [37] R. C. Moore and M. D. Purugganan, “The evolutionary dynamics of plant duplicate genes,” *Current Opinion in Plant Biology*, vol. 8, no. 2, pp. 122–128, 2005.
- [38] W.-H. Li, J. Yang, and X. Gu, “Expression divergence between duplicate genes,” *Trends in Genetics*, vol. 21, no. 11, pp. 602–607, 2006.
- [39] C. P. Hong, P. Plaha, D.-H. Koo, et al., “A survey of the *Brassica rapa* genome by BAC-end sequence analysis and comparison with *Arabidopsis thaliana*,” *Molecules and Cells*, vol. 22, no. 3, pp. 300–307, 2006.
- [40] C. P. Hong, S. J. Lee, J. Y. Park, et al., “Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences,” *Molecular Genetics and Genomics*, vol. 271, no. 6, pp. 709–716, 2004.
- [41] J. Messing, A. K. Bharti, W. M. Karlowski, et al., “Sequence composition and genome organization of maize,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 40, pp. 14349–14354, 2004.
- [42] J. E. Bowers, B. A. Chapman, J. Rong, and A. H. Paterson, “Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events,” *Nature*, vol. 422, no. 6930, pp. 433–438, 2003.
- [43] B. J. Haas, J. R. Wortman, C. M. Ronning, et al., “Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release,” *BMC Biology*, vol. 3, p. 7, 2005.
- [44] X. Zhang and S. R. Wessler, “Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 15, pp. 5589–5594, 2004.
- [45] C. P. Hong, Z. Y. Piao, T. W. Kang, et al., “Genomic distribution of simple sequence repeats in *Brassica rapa*,” *Molecules and Cells*, vol. 23, no. 3, pp. 349–356, 2007.
- [46] M. Ventura, N. Archidiacono, and M. Rocchi, “Centromere emergence in evolution,” *Genome Research*, vol. 11, no. 4, pp. 595–599, 2001.
- [47] J. Jiang, J. A. Birchler, W. A. Parrott, and R. K. Dawe, “A molecular view of plant centromeres,” *Trends in Plant Science*, vol. 8, no. 12, pp. 570–575, 2003.
- [48] K. Nagaki, P. B. Talbert, C. X. Zhong, R. K. Dawe, S. Henikoff, and J. Jiang, “Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres,” *Genetics*, vol. 163, no. 3, pp. 1221–1225, 2003.
- [49] E. K. Round, S. K. Flowers, and E. J. Richards, “*Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure,” *Genome Research*, vol. 7, no. 11, pp. 1045–1053, 1997.
- [50] H. Thompson, R. Schmidt, A. Brandes, J. S. Heslop-Harrison, and C. Dean, “A novel repetitive sequence associated with the centromeric regions of *Arabidopsis thaliana* chromosomes,” *Molecular Genetics and Genomics*, vol. 253, no. 1–2, pp. 247–252, 1996.
- [51] Z. Cheng, F. Dong, T. Langdon, et al., “Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon,” *The Plant Cell*, vol. 14, no. 8, pp. 1691–1704, 2002.
- [52] Y. Zhang, Y. Huang, L. Zhang, et al., “Structural features of the rice chromosome 4 centromere,” *Nucleic Acids Research*, vol. 32, no. 6, pp. 2023–2030, 2004.
- [53] E. V. Ananiev, R. L. Phillips, and H. W. Rines, “Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 22, pp. 13073–13078, 1998.
- [54] O. Kulikova, G. Gualtieri, R. Geurts, et al., “Integration of the FISH pachytene and genetic maps of *Medicago truncatula*,” *The Plant Journal*, vol. 27, no. 1, pp. 49–58, 2001.
- [55] O. Kulikova, R. Geurts, M. Lamine, et al., “Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*,” *Chromosoma*, vol. 113, no. 6, pp. 276–283, 2004.
- [56] X. Xia, G. Selvaraj, and H. Bertrand, “Structure and evolution of a highly repetitive DNA sequence from *Brassica napus*,” *Plant Molecular Biology*, vol. 21, no. 2, pp. 213–224, 1993.
- [57] X. Xia, P. S. Rocha, G. Selvaraj, and H. Bertrand, “Genomic organization of the canrep repetitive DNA in *Brassica juncea*,” *Plant Molecular Biology*, vol. 26, no. 3, pp. 817–832, 1994.
- [58] G. E. Harrison and J. S. Heslop-Harrison, “Centromeric repetitive DNA sequences in the genus *Brassica*,” *Theoretical and Applied Genetics*, vol. 90, no. 2, pp. 157–165, 1995.
- [59] K. B. Lim, T. J. Yang, Y. J. Hwang, et al., “Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species,” *The Plant Journal*, vol. 49, no. 2, pp. 173–183, 2007.
- [60] J. C. Venter, H. O. Smith, and L. Hood, “A new strategy for genome sequencing,” *Nature*, vol. 381, no. 6581, pp. 364–366, 1996.
- [61] G. G. Mahairas, J. C. Wallace, K. Smith, et al., “Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 17, pp. 9739–9744, 1999.
- [62] K. M. Song, J. Y. Suzuki, M. K. Slocum, P. M. Williams, and T. C. Osborn, “A linkage map of *Brassica rapa* (syn. *campestris*) based on restriction fragment length polymorphism loci,”

- Theoretical and Applied Genetics*, vol. 82, no. 3, pp. 296–304, 1991.
- [63] R. A. Teutonico and T. C. Osborn, “Mapping of RFLP and qualitative trait loci in *Brassica rapa* and comparison to the linkage maps of *B.napus*, *B.oleracea*, and *Arabidopsis thaliana*,” *Theoretical and Applied Genetics*, vol. 89, no. 7-8, pp. 885–894, 1994.
- [64] C. Kole, P. Kole, R. Vogelzang, and T. C. Osborn, “Genetic linkage map of a *Brassica rapa* recombinant inbred population,” *Journal of Heredity*, vol. 88, no. 6, pp. 553–557, 1997.
- [65] J. S. Kim, T. Y. Chung, G. J. King, et al., “A sequence-tagged linkage map of *Brassica rapa*,” *Genetics*, vol. 174, no. 1, pp. 29–39, 2006.
- [66] K. Suwabe, H. Tsukazaki, H. Iketani, et al., “Simple sequence repeat-based comparative genomics between *Brassica rapa* and *Arabidopsis thaliana*: the genetic origin of clubroot resistance,” *Genetics*, vol. 173, no. 1, pp. 309–319, 2006.
- [67] S. R. Choi, G. R. Teakle, P. Plaha, et al., “The reference genetic linkage map for the multinational *Brassica rapa* genome sequencing project,” *Theoretical and Applied Genetics*, vol. 115, no. 6, pp. 777–792, 2007.
- [68] I. A. Parkin, A. G. Sharpe, D. J. Keith, and D. J. Lydiate, “Identification of the A and C genomes of amphiploid *Brassica napus* (oilseed rape),” *Genome*, vol. 38, pp. 1122–1133, 1995.
- [69] The International Human Genome Mapping Consortium, “A physical map of the human genome,” *Nature*, vol. 409, no. 6822, pp. 934–941, 2001.
- [70] M. C. Luo, C. Thomas, F. M. You, et al., “High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis,” *Genomics*, vol. 82, no. 3, pp. 378–389, 2003.



**Hindawi**

Submit your manuscripts at  
<http://www.hindawi.com>

