Breeding Science 61: 661–664 (2012)
doi:10.1270/jsbbs.61.661

Note

DaizuBase, an integrated soybean genome database including BAC-based physical maps

Yuichi Katayose*[1)], Hiroyuki Kanamori[1,2)], Michihiko Shimomura[3)], Hajime Ohyanagi[3)], Hiroshi Ikawa[2)], Hiroshi Minami[2)], Michie Shibata[1,2)], Tomoko Ito[2)], Kanako Kurita[1,2)], Kazue Ito[1,2)], Yasutaka Tsubokura[1)], Akito Kaga[1)], Jianzhong Wu[1)], Takashi Matsumoto[1)], Kyuya Harada[1)] and Takuji Sasaki[1)]

[1)] *National Institute of Agrobiological Sciences*, 1-2 Ohwashi, Tsukuba, Ibaraki 305-8634, Japan
[2)] *STAFF Institute*, 446-1 Ippaizuka, Kamiyokoba, Tsukuba, Ibaraki 305-0854, Japan
[3)] *Mitsubishi Space Software Co. Ltd.*, Takezono, Tsukuba, Ibaraki 305-0032, Japan

Soybean [*Glycine max* (L) Merrill] is one of the most important leguminous crops and ranks fourth after to rice, wheat and maize in terms of world crop production. Soybean contains abundant protein and oil, which makes it a major source of nutritious food, livestock feed and industrial products. In Japan, soybean is also an important source of traditional staples such as tofu, natto, miso and soy sauce. The soybean genome was determined in 2010. With its enormous size, physical mapping and genome sequencing are the most effective approaches towards understanding the structure and function of the soybean genome. We constructed bacterial artificial chromosome (BAC) libraries from the Japanese soybean cultivar, Enrei. The end-sequences of approximately 100,000 BAC clones were analyzed and used for construction of a BAC-based physical map of the genome. BLAST analysis between Enrei BAC-end sequences and the Williams82 genome was carried out to increase the saturation of the map. This physical map will be used to characterize the genome structure of Japanese soybean cultivars, to develop methods for the isolation of agronomically important genes and to facilitate comparative soybean genome research. The current status of physical mapping of the soybean genome and construction of database are presented.

Key Words: BAC-end sequencing, physical map, database.

Introduction

In 2010, the soybean genome was sequenced and assembled by the Soybean Genome Sequencing Consortium in the USA (Schmutz *et al.* 2010). The genome data are available via databases, phytozome (http://www.phytozome.net/soybean) and Soybase (Grant *et al.* 2010) (http://soybase.org/). Other soybean genomes were sequenced by a next generation sequencer (Kim *et al.* 2010, Lam *et al.* 2010). Soybase is an essential site and tool for soybean researchers to investigate genetics, molecular biology, breeding and genomics. Although this database is important for soybean research, Williams82 genome data are insufficient for Japanese soybean research. We therefore constructed a genome database from the Japanese cultivar Enrei, a common cultivar in Japan. Enrei was selected to construct the physical map and decode the genome sequence.

Communicated by T. Anai
Received October 27, 2011. Accepted December 9, 2011.
*Corresponding author (e-mail: katayose@nias.affrc.go.jp)

BAC library construction

BAC libraries were constructed from nuclear DNA prepared from young leaves of Enrei (Baba *et al.* 2000). Two restriction endonucleases, *Hin*dIII and *Mbo*I, were used for partial digestion of DNA. Partially digested and size-selected DNA (100–180 kb) was ligated into the BAC vector, pIndigoBAC5 (Epicentre Biotechnologies), then transformed into *E. coli*, ElectroMAX DH10B cells (Life Technologies). We picked up 80,000 clones of *Hin*dIII digest, and 100,000 clones of *Mbo*I digest, and designated GMJENa as the *Hin*dIII digest library and GMJENb as the *Mbo*I library. Insert DNAs were 140 and 100 kb for GMJENa and GMJENb libraries, respectively. Each clone was stored in 384-well microplates and kept at −80°C.

End sequencing of BAC clones

Both ends of all clones of GMJENa and 20,000 clones of GMJENb were sequenced by the BigDye Terminator (Life Technologies) method and ABI 3730xl capillary sequencer

Table 1. Statistics of "Enrei" BAC-based physical map base on 20 chromosomes

Chromosome	BAC	BAC contig	Single BAC contig	Total length (bp)	Covered length (bp)	Total gap length (bp)	Cover rate
Gm01 (D1a)	4,110	44	6	55,915,595	53,637,206	2,278,389	96
Gm02 (D1b)	3,179	72	10	51,656,713	46,459,754	5,196,959	90
Gm03 (N)	2,462	62	7	47,781,076	43,124,153	4,656,923	90
Gm04 (C1)	2,882	56	6	49,243,852	45,507,841	3,736,011	92
Gm05 (A1)	2,852	39	4	41,936,504	38,979,257	2,957,247	93
Gm06 (C2)	2,760	64	8	50,722,821	45,066,672	5,656,149	89
Gm07 (M)	2,695	48	7	44,683,157	41,367,378	3,315,779	93
Gm08 (A2)	2,763	56	7	46,995,532	43,208,178	3,787,354	92
Gm09 (K)	3,101	40	3	46,843,750	44,090,053	2,753,697	94
Gm10 (O)	3,077	60	6	50,969,635	45,376,931	5,592,704	89
Gm11 (B1)	2,447	49	4	39,172,790	35,810,276	3,362,514	91
Gm12 (H)	2,430	41	5	40,113,140	35,646,507	4,466,633	89
Gm13 (F)	1,992	70	12	44,408,971	36,659,143	7,749,828	83
Gm14 (B2)	3,774	45	6	49,711,204	45,751,866	3,959,338	92
Gm15 (E)	3,117	49	3	50,939,160	47,368,637	3,570,523	93
Gm16 (J)	2,392	49	11	37,397,385	33,708,594	3,688,791	90
Gm17 (D2)	2,363	55	11	41,906,774	37,992,668	3,914,106	91
Gm18 (G)	3,714	63	3	62,308,140	57,128,821	5,179,319	92
Gm19 (L)	2,994	51	5	50,589,441	46,460,750	4,128,691	92
Gm20 (I)	3,893	45	4	46,773,167	43,610,445	3,162,722	93
Total	58,997	1,058	128	950,068,807	866,955,130	83,113,677	91

BAC clones mapped on other scaffolds are not shown.
BAC number: number of BAC clones mapped on each chromosome.
BAC contig: number of contigs on each chromosome.
Single BAC contig: number of contigs, consists of one BAC clone.
Total length: base-pair of each chromosome.
Covered length: size of BAC-covered regions.
Total gap length: size of no BAC regions.
Cover rate: (covered length)/(total length) × 100 (%).

(Life Technologies) (Katagiri *et al.* 2004). The obtained sequence data were analyzed by PhredPhrap software (Ewing and Green 1998, Ewing *et al.* 1998). After exclusion of low-quality (Phred <30) bases, the average read-length of BAC-end sequences was 650 bases.

Mapping of BAC clones and construction of physical map

To identify the physical positions of each sequenced clone, end sequences were analyzed by Blastn with the Williams82 genome assembly (Glyma1.09). After sequencing, end-sequenced BAC clones were mapped on each chromosome of the Williams82 genome assembly. Finally, 59361 BAC clones (58997 clones were mapped on 20 chromosomes, 364 clones were mapped on other scaffolds) were mapped on the Williams82 genome and 91% of the genome was covered by Enrei BAC clones (Table 1). We detected differences between Enrei BAC-end sequences and the Williams82 genome assembly. The mismatch rate was 0.2–0.5%, and the deletion rate was less than 0.1% for each chromosome.

DaizuBase

We constructed an integrated soybean genome database, DaizuBase (http://daizu.dna.affrc.go.jp). This database consists of Gbrowse, Unified map and blast search. The Gbrowse page shows BAC-based physical map, unified map page shows linkage map and DNA markers, both are based on Williams82 genome assembly. Gbrowse provides a tracking function for DNA sequence, BAC-end, BAC contigs, GC contents, ESTs, full-length cDNAs (Umezawa *et al.* 2008), DNA markers (Fig. 1). And also, DaizuBase has a sequence, keyword and position search systems.

The prospects

Using the Roche/454 next generation sequencer, GS-FLX Titanium (Margulies *et al.* 2005), 10 equivalent size of the genome of Japanese soybean cultivar, Enrei, has already been sequenced. After analyzing the data, we will upload genome data for Enrei into DaizuBase.

The database will provide SNPs and In/Dels data for Enrei and Williams82 genomes.

Enrei genome data will be useful to distinguish domestic soybean genomes and isolate important genes. Furthermore,

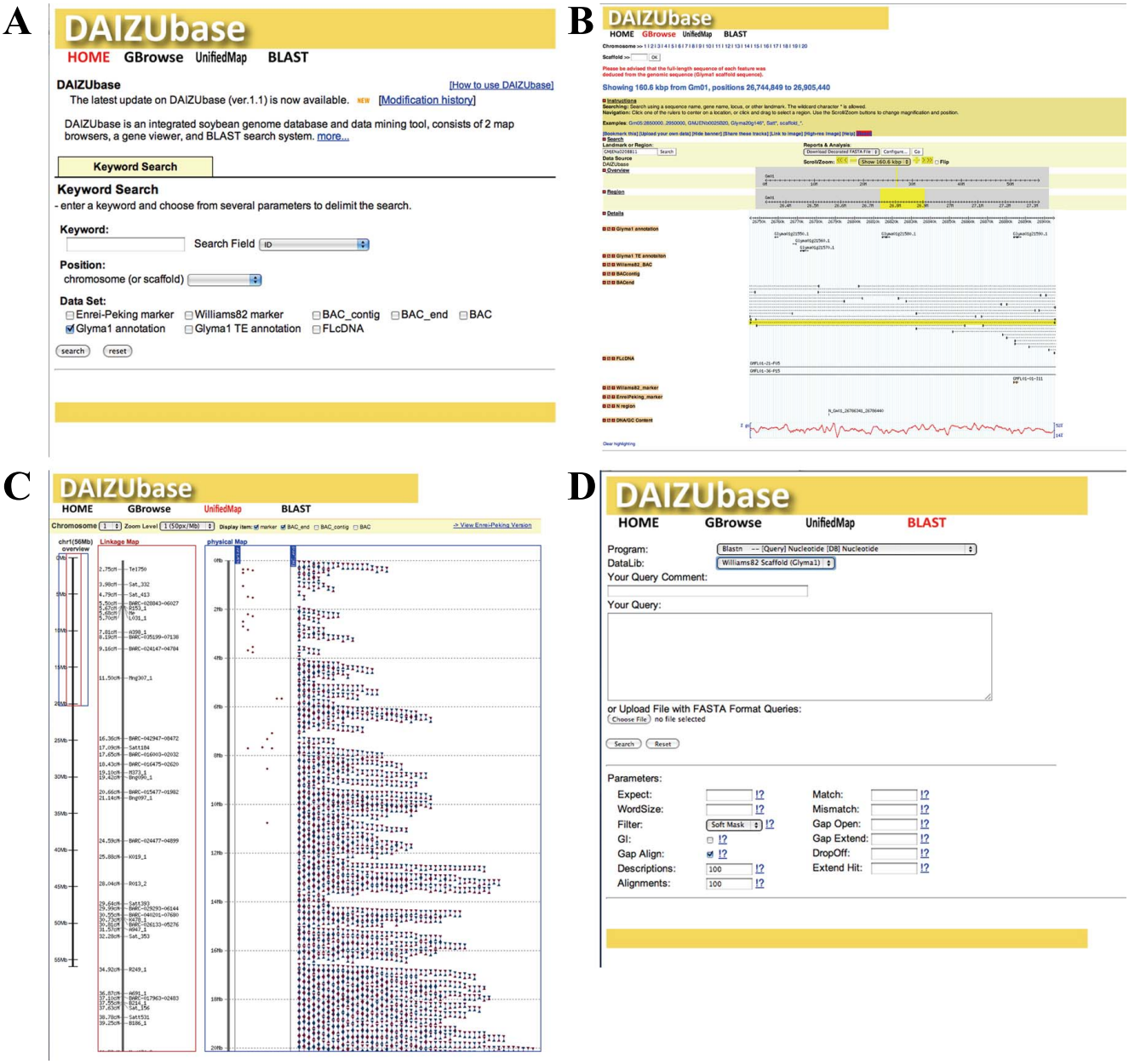

Fig. 1. Browsing DaizuBase. A) DaizuBase top page with links to Gbrowse, Unified Map and Blast search. B) Gbrowse shows BAC-based physical map data. C) Unified Map shows relationships among the linkage map, DNA markers and BAC end sequences. D) Sequence search systems using BLAST.

sequencing of various Japanese cultivar genomes is progressing using the next generation sequencer. These genomic data will be useful for establishing DNA markers for Japanese cultivars.

Acknowledgements

We thank Dr. Naoki Katsura, President of the STAFF Institute, for encouragement and continuous support of the Soybean Genome Research. This work was supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics for Agricultural Innovation, DD-1020 and SOY1001).

Literature Cited

Baba, T., S. Katagiri, H. Tanoue, R. Tanaka, Y. Chiden, S. Saji, M. Hamada, M. Nakashima, M. Okamoto, M. Hayashi *et al.* (2000) Construction and characterization of Rice genomic libraries: PAC library of Japonica variety, Nipponbare and BAC library of Indica variety, Kasalath. Bulletin of NIAR 14: 41–49.

Ewing, B. and P. Green (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194.

Ewing, B., L. Hiller, M.C. Wendel and P. Green (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8: 175–185.

Grant, D., R.T. Nelson, S.B. Cannon and R.C. Shoemaker (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38: D843–D846.

Katagiri, S., J. Wu, Y. Ito, W. Karasawa, M. Shibata, H. Kanamori, Y. Katayose, N. Namiki, T. Matsumoto and T. Sasaki (2004) End sequencing and chromosomal *in silico* mapping of BAC clones derived from an *indica* rice cultivar, Kasalath. Breed. Sci. 54: 273–279.

Kim, M.Y., S. Lee, K. Van, T.-H. Kim, S.-C. Jeong, I.-Y. Choi, D.-S. Kim, Y.-S. Lee, D. Park, J. Ma *et al.* (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. Proc. Natl. Acad. Sci. USA 107: 22032–22037.

Lam, H.-M., X. Xu, X. Liu, W. Chen, G. Yang, F.-L. Wong, M.-W. Li, W. He, N. Qin, B. Wang *et al.* (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature Genet. 42: 1053–1061.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen *et al.* (2005) Genome sequencing in microfabricated high-density picoliter reactors. Nature 437: 376–380.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng *et al.* (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183.

Umezawa, T., T. Sakurai, Y. Totoki, A. Toyoda, M. Seki, A. Ishiwata, K. Akiyama, A. Kurotani, T. Yoshida, K. Mochida *et al.* (2008) Sequencing and analysis of approximately 40 000 soybean cDNA clones from a full-length-enriched cDNA library. DNA Res. 15: 333–346.