Genomics & Informatics Vol. 2(4) 174-179, December 2004

Construction of Various Copy Number Plasmid Vectors and Their Utility for Genome Sequencing

Tae-Jin Yang[1,2,*], Yeisoo Yu[2], David A. Frisch[3], Seunghee Lee[2], Hye-Ran Kim[2], Soo-Jin Kwon[1], Beom-Suk Park[1] and Rod A. Wing[2]

[1]Brassica Genomics Team, National Institute of Agricultural Biotechnology (NIAB), RDA, Suwon 441-707 Korea, [2]Arizona Genomics Institute, 303 Forbes building, University of Arizona, Tucson, Az 85721, USA, [3]Genome Center of Wisconsin, 425 Henry Mall, Madison, WI 53706, USA

Keywords: Plasmid, Cloning Vector, Shotgun library, Genomic library, BAC

Abstract

We developed various plasmid cloning vectors that are useful in the construction of genomic and shotgun libraries. Two medium copy vectors, pCUGIblu21 (pCb21) and pAGIblu21 (pAb21), which are resistant to kanamycin (Km^R) and chloramphenicol (Cam^R), respectively, are useful for cloning DNA inserts ranging from 5kb to 15kb. Two high copy vectors, pCUGIblu31 (pCb31) and pAGIblu31 (pAb31), containing Km^R and Cam^R, respectively, are useful for DNA inserts less than 5kb. These vectors are well adapted for large-scale genome sequencing projects by providing choice of copy number and selectable marker. The small vector size is another advantage of these vectors. All vectors contain *lacZα* including multicloning sites that originated from pBluscriptIIsk- for easy cloning and sequencing. Two medium copy vectors contain unique and rare cutting *Swa*I (ATTTAAAT) restriction enzyme sites for easy determination of insert size. We developed two combined vectors, pC21A31 and pC31A21, which are combinations of (pCb21 + pAb31) and (pCb31 + pAb21), respectively. These two vectors provide four choices of vectors such as Km^R and medium, Cam^R and high, Cam^R and medium, and Km^R and high copy vectors by restriction enzyme cutting, dephosphorylation, and gel purification. These vectors were successfully applied to high throughput shotgun sequencing of rice, tomato, and brassica BAC clones. With an example of extremely biased hydro sheared 3 kb shotgun library of a tomato BAC clone, which is originated from cytogenetically defined peri-centromeric region, we suggest the utility of an additional 10 kb library for sequence assembly of the difficult-to-assemble BAC clone.

*Corresponding author: E-mail tjyang@rda.go.kr, Tel +82-31-299-1677, Fax +82-31-299-1672
Accepted 4 November 2004

Introduction

Genome sequencing is achieved by large-scale sequence assembly of various shotgun libraries. There are two approaches in sequencing large genomes: (1) whole-genome shotgun sequencing (WGS) and (2) clone-based shotgun sequencing (CBS). WGS requires whole genome libraries containing various insert sizes such as 50 kb, 10 kb, and 3 kb (Venter *et al.*, 2001). CBS involves obtaining a collection of large-insert clones, such as a bacterial artificial chromosome (BAC) library, covering a genome multiple times. The large insert clones are then assembled into contigs, reproducing each chromosome of the organism. The series of large insert clones covering the whole genome with a minimal overlap are then chosen for sequencing. This approach was successfully used in sequencing the *Arabidopsis* genome (Arabidopsis Genome Initiative, 2000) and rice genome (Yu *et al.*, 2003). Complete sequence assembly of an individual BAC clone is dependent on production of a shotgun library containing equally distributed clones. Whole genome sequencing requires each shotgun libraries of thousands of individual BAC clone on minimum tiled path.

Many high copy plasmid vectors such as the pUC series (Yanisch-Perron *et al.*, 1985) or pBluscript series (Stratagene, La Jolla, CA) can be used for construction of small insert shotgun libraries having less than 5 kb inserts. Lambda, cosmid or fosmid (Sambrook *et al.*, 1989), bacterial artificial chromosome (BAC) (Shizuya *et al.*, 1992) or P1-derived artificial chromosome (PAC) (Ioannou *et al.* 1994), and yeast artificial chromosome (YAC) (Burke, 1991) vectors carry up to 20 kb, 50 kb, 150 kb, and 500 kb insert DNA, respectively. Tao and Zhang (1998) demonstrate that some plasmid vectors can maintain up to 300 kb insert. The medium copy vector pBR322 was used for the whole genome 10 kb insert library produced for whole genome shotgun sequencing of the human genome (Venter *et al.*, 2001). However, there is little information about the vector used for

Table 1. DNA segments used for plasmid construction - origin, length and primer sequence.

PCR Name[z]	Original Vector	Length(bp)	Priming Site	Primer sequence 5' to 3'
Mc	pUC18	181	340	*ATTTAAATT*AAGTTGGGTAACGCCAGGG
			504	*ATTTAAATT*GTG AGCGGA TAACAATTTC
OKm	pACYC177	2061	674	TTAAAAGGATCTAGGTGAAG
			2734	TAGAAAAACTCATCGAGCAT
Km	pACYC177	899	1867	CATCATGAACAATAAAACTGTCTGC
			2765	CCAGTGTTACAACCAATTAACCAAT
OR	pBR322	1474	1689	CTGGCTACCCTGTGGAACAC
			3162	ACTGAGCGTCAGACCCCGTA
OL	pBCSKII(-)	1504	356	TGATTTATAAGGGATTTTGCCG
			1859	ACGTGAGTTTTCGTTCCACT
L	pBCSKII(-)	695	312	*CCATTTAAAT*TCCAAACTGGAACAACACTCAAC
			987	*CCATTTAAAT*ATTAATGCAGCTGGCACGAC
ORKm	pCUGI21	2373	182	CATCATGAACAATAA AACTGTCTGC
			2554	ACTGAGCGTCAGACCCCGTA
Cm	Tn-Cam^R	750	265	ACTTATTCAGGCGTAGCAAC
			1014	TTGTCGAGATTTTCAGGAGG
LOR	pCb21	2169	312	*CCATTTAAAT*TCCAAACTGGAACAACACTCAAC
			1689	CTGGCTACCCTGTGGAACAC

[z]PCR name represent the abbreviation of PCR fragments and their functions: Mc (multi cloning site), O (Origin of replication), Km (kanamycin resistance), R (*rop* gene which controls copy number of plasmid), L (*LacZ* including multicloning sites), Cm (chloramphenicol resistance).
Note: Italics and underlined sequences are added sequence and *Swa*I restriction site (ATTTAAAT), respectively.

construction of these libraries. Most high copy vectors confer ampicillin resistance by expression of the beta-lactamase gene. Because ampicillin resistance is due to the extracellular degradation of ampicillin by beta-lactamase, satellite colonies lacking the beta-lactamase gene often are present surrounding resistant clones. The appearance of satellite clones can be reduced by growth on high concentrations of carbenicillin. However, it still remains an obstacle for large-scale and high throughput process. Therefore, we developed high and medium copy vectors containing the stable antibiotic selectable markers, kanamycin or chloramphenicol. This study provides a convenient choice of vectors based on copy number or antibiotics resistance for construction of genomic or shotgun libraries.

Materials and Methods

Vector construction

Vectors were developed by recombination of functional DNA fragments obtained by PCR amplification. Commercially available plasmid vectors, pBR322 (Amersham Bioscience), pUC18 (Amersham Bioscience), pBluescriptIIsk(-) (Stratagene, La Jolla, CA), and pACYC177 (GenBank accession no. X06402) and Entranceposon-Cam of template generation system (Finnzyme Co. Finland) were used as PCR templates. Oligonucleotide primers were synthesized by MWG- iotech (http://www.mwg-iotech.com/). PCR was performed with *Pfu* DNA polymerase (Stratagene, La Jolla, CA) using the primers and templates in Table 1. The modeling of new vectors and designing of primers was performed using vector NTI software (Invitrogen). Some PCR fragments (labeled as -p in Table 2) were phosphorylated using T4 polynucleotide kinase (Promega, Madison, WI) and gel purified using Qiaex DNA fragment purification kit. The same amount of the purified DNAs were combined such as the combinations in Table 2 using T4 DNA ligase (Promega, Madison, WI) for the new vector construction and transformed into DH10B T1-resistant competent cells (Invitrogen, Carlsbad, CA).

Genomic and Shotgun library construction

Large amounts of plasmid DNA were prepared using Qiagen plasmid midi kit (Qiagen, Valencia, CA). The dephosphorylated linearized plasmid vector was prepared as described by Kim *et al.* (2004) and Yang *et al.* (2003). The vectors were tested for construction of tens of genomic and shotgun libraries using various insert DNAs, such as randomly hydrosheared or restriction enzyme digested 10 kb rice nuclei DNA (Yang *et al.*, 2003) or hydro sheared BAC shotgun DNAs ranging from 300 bp to 10 kb.

Table 2. Summary of new vector series and their DNA fragments used for in vitro combinations.

Vector Name	Size (bp)	Combination of PCR Fragments[z]	Replicon	Copy /Cell	Markers	UniqueSites
pCUGI11	2242	Mc^p + OKm	p15A	10	Km^R	*Swa*I
pCUGI21	2554	Mc^p + Km + OR^p	pMB1	25	Km^R	-
pCUGIblu11	2757	L^p + pCUGI11-*Swa*I	p15A	10	Km^R	*Swa*I
pCUGIblu21	3069	L^p + ORKm	pMB1+rop	25	Km^R	*Swa*I
pCUGIblu31	2401	OL^p + Km	pMB1	500	Km^R	*Pvu*II
pAGIblu21	2922	Cm^p + LOR	pMB1+rop	25	Cam^R	*Swa*I
pAGIblu31	2254	Cm^p + OL	pMB1	500	Cam^R	*Pvu*II

[z]PCR fragments correspond to the PCR name in Table 1 and the DNA fragments phophorylated before ligation were labeled as^p.

Sequence of the vectors

Orientation and cloning sites of the vectors were evaluated by restriction digestion with several enzymes. Complete sequences of the vectors were obtained using 24 transpositioned clones of each vector (template generation system; Finnzyme, Finland) and additional reactions for finishing such as Yang *et al.* (2003).

Results and Discussion

Construction of vectors

Seven vectors summarized in Table 2 were constructed via in vitro combination of various PCR fragments with the functions presented in Table 1. Two vectors, pCUGI11 and pCUGI21, were developed for direct cloning methods such as the double adaptor method (Andersson *et al.*, 1996) which does not require the blue/white screening because there is no *lac*Zα gene. The pCUGIblu11 contains p15A origin of replication, which is an intermediate copy (around 10 copies per cell) and compatible with pMB1 origin of replication. Nakano *et al.* (1995) constructed several intermediate copy number vectors containing p15A *ori*. However, we did not used the pCUGIblu11 for library construction because the yield of plasmid is not quite enough for high throughput sequencing. Thus, four vectors were selected: pCUGIblu21 (pCb21), pCUGIblu31 (pCb31), pAGIblu21 (pAb21), and pAGIblu31 (pAb31) for the final practical use as shown in Fig. 1. The origins of PCR fragments were shown as a gray, dot, and white in the vector maps (Fig. 1). The gray regions represent the selectable markers, Km^R or Cam^R, which originated from pACYC177 or transposon DNA harboring Cam^R (template generation system of Finnzyme), respectively. The dotted regions in high copy vectors are originated from pBluescriptIIsk(-) and the white region in medium copy vectors are originated from the pBR322. The high copy vectors were combined with two PCR fragments, ori and *lac*Z including multicloning sites (dotted) and selectable markers (gray) (Fig. 1, top left). Medium copy vectors were combined with three PCR fragments such as *ori* (white), *lac*Z including multi cloning sites (dotted), and antibiotics resistance (gray) (Fig. 1, top right).

Combined vector

Two vectors were combined based on copy number and selectable markers (Fig.1, bottom) to improve the convenience of vector preparation. The pC21A31 plasmid was constructed by combining pCb21, Km^R and medium copy plasmid, and pAb31, Cam^R and high copy plasmid. Both are digested with *Bam*HI and ligated each other (Fig. 1). The pC31A21 is obtained by combining pCb31 (Km^R and high copy plasmid) and pAb21 (Cam^R and medium copy plasmid). The individual vector can be recovered by digesting with a restriction enzyme as

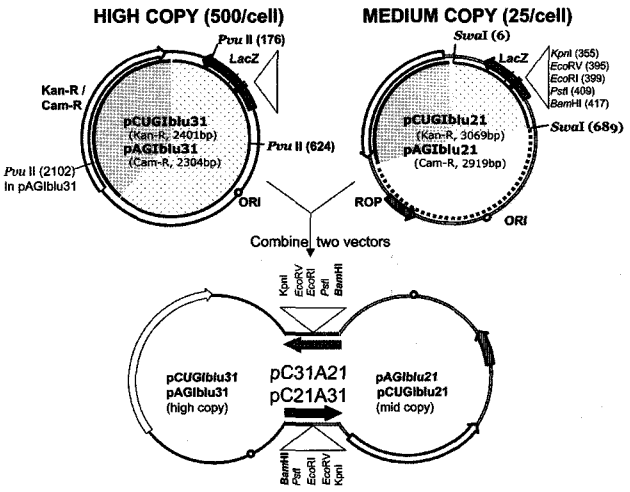

Fig. 1. Representative maps of vectors. The high copy vectors, pCb31 (2401 bp) and pAb31 (2304 bp) are represented at top left. The medium copy vectors, pCb21 (3069 bp) and pAb21 (2919 bp) are represented at the top right. The origin of DNA fragments are represented as grayed, dotted, and white region for pACYC177 (or Entranceposon-Cam), pBluscriptIIsk-, and pBR322, respectively.

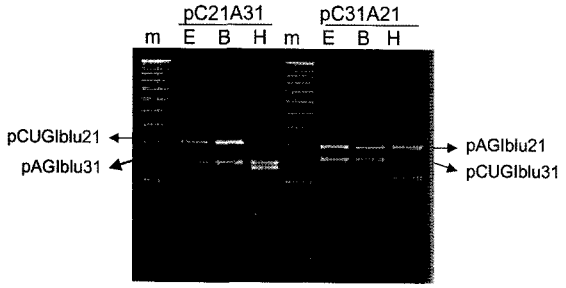

Fig. 2. The combined vectors, pC21A31 and pC31A21. Two vectors were digested with *Eco*RV (E), cohesive ending *Bam*HI (B), or *Hin*dIII (H). The digestion with the blunt end restriction enzyme *Eco*RV or the cohesive end *Bam*HI provides two independent vectors, one medium copy vector (upper band) and one high copy vector (lower band). These two combined vectors provide four choices of linearized ready-to-use vectors by restriction cutting and subsequent dephosphrylation and gel purification. The pC21A31 provide pCb21, the Km^R and medium copy vector, (upper band) and pAb31, the Cam^R and high copy vector, (lower band). The pC31A21 provide pCb31, the Km^R and high copy vector (lower band) and pAb21, the Cam^R and medium copy vector (upper band). The DNA marker (m) is the 1kb ladder.

shown in Fig. 2. Two different linearized ready-to-use vectors, one medium and one high copy vector, can be prepared by enzyme digestion of a combined vector and subsequent dephosphorylation and gel purification.

The utility of the vectors

For the choice of copy numbers, three vector classes are available: intermediate (pCb11), medium (pCb21 and pAb21), and high copy vectors (pCb31 and pAb31) for 10, 25, and 500 copies per cell, respectively. For the choice of selectable markers, two antibiotics resistance genes, Km^R and Cam^R, are available with different copy numbers. The pCUGI series such as pCb11 (low copy), pCb21 (medium copy), and pCb31 (high copy) contain Km^R and the pAGI series such as pAb21 (medium copy) and pAb31 (high copy) contain Cam^R.

The high copy vectors are useful for construction of shotgun library harboring insert DNA less than 5 kb. The medium copy vectors can be used for the construction of libraries containing medium size inserts ranging 5 kb up to 15 kb. Two medium copy vectors, pCb21 and pAb21, were successfully used to construct rice whole genomic libraries covering 8.8x rice genome with an average insert size of 11 kb (Yang *et al.*, 2003) and 10 kb shotgun libraries (carrying inserts ranging 6 kb to 10 kb) from rice and tomato BAC DNA.

The vectors are well adapted for large-scale genome sequencing projects by providing a choice of copy number and selectable marker for use in construction of

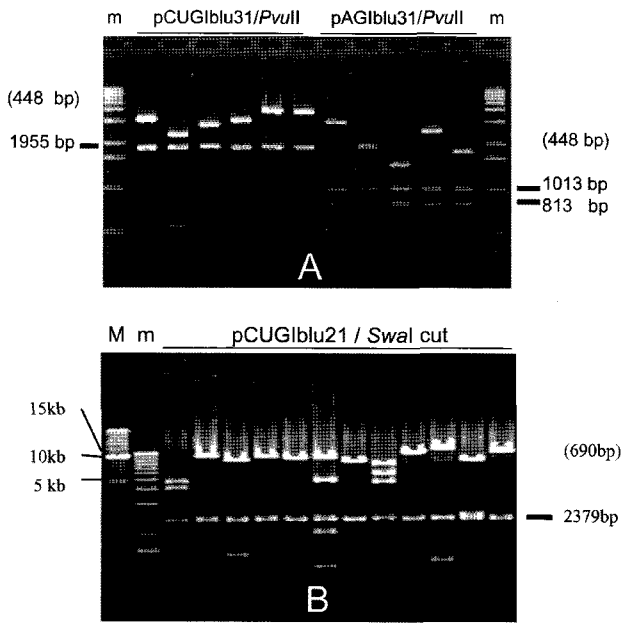

Fig. 3. The shotgun libraries constructed using different vectors, pCb31, pAb31, and pCb21. (A) Two small insert libraries in two high copy vectors were digested with *PvuII*. The pCb31 shows a single1955bp vector band and the pAb31 shows two vector bands: 1013 bp and 813 bp. The various sizes of inserts contain 448bp of vector DNA. (B) One 10 kb library using the pCb21, a medium copy vector, shows a 2379 bp vector and various sizes (up to 15 kb) of inserts by *SwaI* digestion. The *SwaI* digested inserts contain 690 bp of vector sequence. DNA markers, m and M, represent 1 kb ladder and 5 kb ladder, respectively.

whole genome or BAC shotgun libraries. The vectors are also successfully used for other cloning purpose such as cloning of PCR fragments or construction of the shuttle library suggested by McMurray *et al.* (1998). PCR fragments are easily cloned into the vectors after amplification with proofreading *Pfu* polymerase and phosphorylation.

For the reliable and convenient insert size evaluation, two unique and rare cutting *Swa*I sites (ATTTAAAT) were added, providing a convenient way to evaluate rice whole genome 10 kb libraries (Yang *et al.*, 2003). The *Swa*I sites are highly valuable for characterization of genome library from GC rich prokaryote genome. The *Pvu*II sites occurring two and three times in high copy vectors, pCb31 and pAb31, respectively, were applied for insert evaluation without ambiguity such as in Fig. 3A.

The vectors have the following useful properties: (1) Relatively small size compared to commercially available vectors; (2) Choice of copy numbers based on insert size; (3) Choice of antibiotics based on insert properties; (4) The combined vector provide two

independent vectors having different copy numbers and antibiotics resistance; (5) Unique and rare cutting restriction enzyme sites (*Swa*I and *Pvu*II) for the convenient insert size evaluation; and, (6) Multicloning site of pBluscriptII enabling cloning and rapid sequencing with universal primers.

Improvement of shotgun sequence assembly using various insert size libraries

We have successfully used each vector for shotgun sequencing of tens of BAC clones from rice, tomato, and brassica; two high copy vectors, pCb31 and pAb31, for 3 kb insert library construction and two medium copy vectors, pCb21 and pAb21, for 10 kb insert library construction. The sequence assembly of a BAC clone relies on sequencing of a randomly sheared 3kb library (Bouck *et al.*, 1998). It is suggested that using two kinds of shotgun libraries (such as a 10 kb and a 3 kb shotgun library) improves the sequence assembly of a BAC clone, especially in the finishing phase (Kim *et al.*, 2004). We found an extreme example of bias in a hydro sheared small insert library in a specific BAC clone among tens of BAC clones tested. A hydro sheared 3 kb shotgun library of tomato BAC clone, Le121H12, showed an extreme bias after assembly of more than 6x coverage of shotgun clones as shown in Fig. 4C. The 3 kb shotgun library does not cover more than 50% of BAC DNA sequence, demonstrating that there is no way to finish the complete sequence of the BAC clone if only the 3 kb shotgun library is used. On the other hand, 3x coverage of 10 kb clones harboring *Hae*III partial digested inserts are distributed randomly throughout the entire BAC clone (Fig. 4D). The complete 152 kb sequence was obtained by subsequent sequencing of bridging 10 kb clones spanning sequence gaps after sequence assembly of two libraries. The bridging 10 kb clones are completely sequenced using transposon insertion method (Biery *et al.*, 2000; Devine *et al.*, 1997; Yang *et al.*, 2003) or partially sequenced by primer walking. The BAC clone was originated from a cytogenetically defined centromeric region (Budiman *et al.*, 2004; Yang *et al.*, Manuscipt in preparation). It is hypothesized that the bias of 3 kb shotgun library was due to the unequal distribution of AT rich, low complexity regions (such as shown in Fig. 4B) resulting in uneven mechanical breakdown of the BAC DNA in the process of hydro shearing.

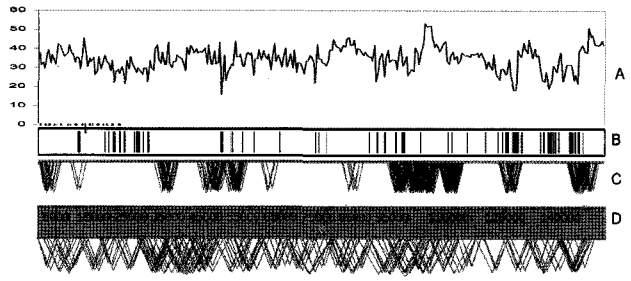

Fig. 4. An example of a BAC shotgun sequence assembly using two different shotgun libraries, 3 kb library and 10 kb library. Tomato BAC clone LeH121H12 was determined to be 151758 bp (phred value > 3.0) after complete sequence assembly and finishing. The nucleotide positions were represented in the gray box of (D) which is imported from the window of a program 'phrapview'. The GC composition (A) and AT rich low complexity regions (B) are presented as percentage in every 50 bp window and black lines, respectively. The shotgun clones are represented as red and pink lines based on their forward and reverse sequence positions and directions. The pink lines represent overlaps of clones at the same positions (C, D). The random sheared 776 shotgun clones (eight 96 plates) of 3 kb library representing 6x converge are extremely biased and represented less than 50% of BAC sequence (C). The 384 clones (four 96 plates) from the *Hae*III partial digested 10 kb library (3x coverage) were evenly dispersed through the entire region and cover the entire BAC sequence. The 3 kb and 10 kb libraries are cloned into a high copy vector (pCb31) and a medium copy vector (pCb21), respectively.

References

Andersson, B., Wentland, M.A., Ricafrente, J.Y., Liu, W., and Gibbs, R.A. (1996). A "double adaptor" method for improved shotgun library construction. *Anal. Biochem.* 236, 107-113.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796-815.

Biery, M.C., Stewart, F.J., Stellwagen, A.E., Raleigh, E.A., and Craig, N.L. (2000). A simple in vitro Tn7-based transposition system with low target site selectivity for genome and gene analysis. *Nucleic Acids Research* 28, 1067-1077.

Bouck, J., Miller, W., Gorrell, J.H., Muny, D., and Gibbs, R.A. (1998). Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Research* 8, 1074-1084.

Budiman, M.A., Chang, S.B., Lee, S., Yang, T.J., Zhang, H.B., de Jong, H., and Wing, R.A. (2004). Localization of jointless-2 gene in the centromeric region of tomato chromosome 12 based on high resolution genetice and physical mapping. *Theor. Appl. Genet.* 108, 190-196.

Burke, D.T. (1991). The role of yeast artificial chromosome clones in generating genome maps. *Curr. Opin. Genet. Dev.* 1, 69-74.

Devine, S.E., Chissoe, S.L., Eby, Y., Wilson, R.K., and Boeke, J.D. (1997). A transposon-based strategy for sequencing repetitive DNA in eukaryotic genomes. *Genome Res.* 7, 551-563.

Ioannou, P.A., Amemiya, C.T., Granes, J., Kroisel, P.M., Shyzuya, H., Chen, C., Batzer, M.A., and de Jong, P.J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genet.* 6, 84-89.

Kim, H.R., Yang, T.J., Kudna, D.A., and Wing, R.A. (2004). Construction and application of genomic DNA libraries. In *Handbook of plant biotechnology*, P. Christou and H. Klee, ed. (John Wiley & Sons, Ltd) pp.71-80.

McMurray, A.A., Sulston, J.E., and Quail, M.A. (1998). Short-Insert Libraries as a method of problem solving in genome sequencing. *Genome Res.* 8, 562-566.

Nakano, Y., Yoshida, Y., Yamashita, Y., and Koga, T. (1995). Construction of a series of pACYC-derived plasmid vectors. *Gene* 162, 157-158.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), Sect. 1.3, pp. 3-5.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri,Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* 89, 8794-8797.

Tao, Q. and Zhang, H.B. (1998). Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. *Nucleic Acids Res.* 26, 4901-4909.

Venter, J.D., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, S.S., Smith, H.O., Yandell, M., Evans, C.A., and Holt, R.A. *et al.* (2001). The sequence of the human genome. *Science* 291, 1304-1351.

Yang, T.J., Yu, Y., Nah, G.J., Atkins, M., Lee, S., Frisch, D., and Wing, R.A. (2003). Construction and utility of 10kb shotgun libraries for efficient clone gap closure for rice genome sequencing. *Theor. Appl. Genet.* 107, 652-660.

Yanisch-Perron, C., Vieira, J., and Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* 33, 103-109.

Yu, Y., Rambo, T., Currie, J., Saski, C., Kim, H.R., Collura, K., Thompson, S., Simmons, J., Yang, T.J., Park, G.N., and Patel, A.J. (2003). The Rice Chromosome 10 Sequencing Consortium. In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* 300, 1566-1569.