

# Handbook of Plant Biotechnology

---

*Editors-in-Chief*

**Paul Christou**

*Fraunhofer Institute for Molecular Biology and Applied Ecology,  
Schmallenberg, Germany*

and

**Harry Klee**

*University of Florida, Horticultural Sciences, Gainesville, FL, USA*

**VOLUME 1**



John Wiley & Sons, Ltd

Copyright © 2004 John Wiley & Sons Ltd,  
The Atrium, Southern Gate,  
Chichester, West Sussex  
PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): [cs-books@wiley.co.uk](mailto:cs-books@wiley.co.uk)  
Visit our Home Page on [www.wileyeurope.com](http://www.wileyeurope.com) or [www.wiley.com](http://www.wiley.com)

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to [permreq@wiley.co.uk](mailto:permreq@wiley.co.uk), or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

#### *Other Wiley Editorial Offices*

John Wiley & Sons Inc., 111 River Street,  
Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street,  
San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12,  
D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road,  
Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,  
Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road,  
Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

#### *Library of Congress Cataloging-in-Publication Data*

Christou, Paul.

Handbook of plant biotechnology/Paul Christou, Harry Klee.  
p. cm.

Includes bibliographical references and index.

ISBN 0-471-85199-X (cloth) – ISBN 0-471-85199-X

1. Plant biotechnology – Handbooks, manuals, etc. 2. Plant genetic engineering – Handbooks, manuals, etc.

I. Klee, Harry. II. Title.

TP248.27.P55C545 2004

660.6–dc22

2003027341

#### *British Library Cataloguing in Publication Data*

A catalogue record for this book is available from the British Library

ISBN 0-471-85199-X

Typeset in 10/12 pt Times by Techbooks, Delhi, India.

Printed and bound in Great Britain by Antony Rowe, Chippenham, UK.

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

# Construction and Application of Genomic DNA Libraries

Hye Ran Kim, Tae Jin Yang, David A. Kudrna and Rod A. Wing

*Arizona Genomics Institute, Department of Plant Sciences, The University of Arizona,  
Tucson, AZ, USA*

## INTRODUCTION

Crop failures, pathogenic outbreaks and famine are serious problems facing society. Understanding an organism's genome will help provide the genetic tools needed to solve these complex problems in a shorter time and with lesser effort. To efficiently study an organism's genome, it can be partitioned into a permanent and stable collection of DNA fragments, called a library. Such libraries provide convenient access to a genome for both laboratory and breeding applications. Genomic libraries can be used as substrates to physically map and sequence entire genomes, clone agriculturally important genes and to investigate gene expression patterns. Further, genomic libraries also provide powerful tools and resources for evaluating germplasm conservation stocks and biological diversity. Since the ongoing explosion of genetic data and molecular clone resources has opened new scientific possibilities with unfamiliar terms and acronyms to researchers venturing across all avenues of applied science, we have included a table of useful terms and abbreviations for this chapter (Table 5.1).

Serious consideration must be given to the various kinds of libraries that can be constructed with respect to different insert sizes, fragmentation methods, vectors, coverage, downstream operations, etc. Though downstream applications

(see the Applications section of this chapter) differ for types of libraries, fragmentation size is the first consideration in genomic library construction. Genomic DNA libraries are classified as shotgun libraries, medium-size insert libraries and large insert libraries depending on insert sizes. A shotgun library, with a smaller insert size (generally 1.5–10 kbp), is made using high-copy plasmid vectors. It is the most common resource for complete sequencing of large genomic DNA clones (e.g. cosmid, PAC, BAC and YAC clones) and large DNA fragments (PCR products and restriction fragments). With the introduction of capillary sequencers for large-scale, high-throughput DNA sequencing, the shotgun library approach is considered a standard method for generating sequence-ready sublibraries for genome projects and positional cloning studies.

Cosmid and fosmid libraries contain medium-sized inserts (35–45 kbp). Cosmids are hybrids of  $\lambda$ -phage and plasmids. They can replicate in the cell like a plasmid due to the plasmid replication origin and be packaged like a phage because of existing *cos* sites. Since most of the  $\lambda$ -phage structure has been deleted, they can carry DNA inserts up to 45 kbp. This type of library is useful for cloning studies because of its simplicity and economic efficiency for construction in comparison with large insert library construction. The fosmid cloning system is similar in size to cosmids and

**Table 5.1** Reference terms and abbreviations used

| Term/abbreviation  | Definition, literature reference                                                                                           |
|--------------------|----------------------------------------------------------------------------------------------------------------------------|
| ABRC               | Arabidopsis Biological Resource Center, <a href="http://www.arabidopsis.org/abrc/">http://www.arabidopsis.org/abrc/</a>    |
| AGI                | Arizona Genomics Institute, <a href="http://www.genome.arizona.edu">http://www.genome.arizona.edu</a>                      |
| BAC                | Bacterial artificial chromosome, Shyzuya <i>et al.</i> (1992)                                                              |
| Chimerism          | Cloning of two or more unlinked DNA fragments in a single molecule                                                         |
| Cos                | Cohesive end in DNA of bacteriophage lambda                                                                                |
| Cosmid             | Cohesive end site ( <i>cos</i> ) of bacteriophage $\lambda$ DNA containing plasmid, Hohn (1979), Hohn <i>et al.</i> (1988) |
| Coverage           | Number of subclones that span the source clone or genome                                                                   |
| CUGI               | Clemson University Genomics Institute, <a href="http://www.genome.clemson.edu">http://www.genome.clemson.edu</a>           |
| F1 factor          | Fertility factor of <i>E. coli</i> , Cavalli-Sforza (1953), Hayes (1952), Willetts and Skurray (1987)                      |
| Fosmid             | F-based cosmid, Kim <i>et al.</i> (1992)                                                                                   |
| Genome             | The total genetic content contained in an organism                                                                         |
| Genomic library    | A set of DNA clones obtained from one original DNA sample                                                                  |
| High copy plasmid  | A circular, double-stranded unit of DNA that replicates >100 copies, independent of chromosomal DNA                        |
| Insert             | Foreign DNA to be inserted into vector                                                                                     |
| PAC                | P1 artificial chromosome, Ioannou <i>et al.</i> (1994)                                                                     |
| PCR                | Polymerase chain reaction                                                                                                  |
| Physical mapping   | Assembly of a whole (or partial) genome using large clones like BACs, YACs or PACs                                         |
| Positional cloning | The isolation of a gene based on its position on a genetic and/or physical map                                             |
| Shotgun            | Random                                                                                                                     |
| Vector             | Agent that transfers genetic material                                                                                      |
| YAC                | Yeast artificial chromosome, Burke <i>et al.</i> (1987)                                                                    |

the vector is derived from the endogenous *E. coli* F1 factor that maintains inserts in a single-copy state adding to the fidelity and stability of the DNA insert. The fosmid system is useful for easy generation of medium-sized insert (45 kbp) genomic libraries produced from small amounts of source DNA such as flow-sorted chromosomal DNA (Gingrich *et al.*, 1996).

Large-insert genomic DNA libraries are essential for physical mapping, positional cloning and genome sequencing of complex genomes. There are two principal large-insert cloning systems: yeast and bacterial artificial chromosome systems (YACs and BACs). The YAC cloning system was first developed in 1987 (Burke *et al.*, 1987), and uses *Saccharomyces cerevisiae* as the host and maintains large inserts (up to 1 Mb) as linear molecules with a pair of yeast telomeres and a centromere. Although used extensively in the late 1980s and early 1990s, this system has several disadvantages (Anderson, 1993; Zhang and Wing, 1997). The recombinant DNA in yeast can be unstable. DNA manipulation is difficult and inefficient. Most importantly, a high level of chimerism—cloning of two or more unlinked DNA fragments in a single molecule—is inherent within the YAC cloning system. These disadvantages impede the utility of YAC libraries and subsequently this

system has been gradually replaced by the BAC cloning system introduced in 1992 (Shizuya *et al.*, 1992).

The BAC system uses a derivative of the *E. coli* F-factor as a vector and *E. coli* as the host, making BAC cloning and subsequent downstream procedures efficient and easy to perform. Recombinant DNA inserts of up to 200 kb can be efficiently cloned and stably maintained as single-copy plasmids in *E. coli*. BAC libraries have been developed for essentially all major crop plants, as well as model plant species, such as *Arabidopsis thaliana* and rice, and can be accessed from the laboratories that made the libraries or through various stock centres (e.g. AGI, CUGI and ABRC).

In this chapter we will present the details for small- and large-insert library construction and their applications.

### Shotgun Libraries

Shotgun libraries of whole genomes or individual BAC clones are extremely useful for generating a draft or complete sequence of a genome or genomic region. The term shotgun refers to fragmenting the target DNA into small fragments that can be easily cloned and sequenced. For BAC and

**Table 5.2** Procedures, methods and considerations for inserts of shotgun libraries

| Procedure                     | Consideration                                                                                                                                                                                                                                               | Method                                                                                                              | Reference                                                                                                                                                                                                                                                                                                                                       |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1. Target DNA isolation       | <i>E. coli</i> genomic DNA contamination free<br>Quantity (more than 2 µg)<br>Easy and simple for high throughput                                                                                                                                           | (a) CsCl-EtBr gradients method<br>(b) Double Acetate precipitation method<br>(c) Qiagen large construct kit         | (a) Sambrook <i>et al.</i> (2001)<br>(b) <a href="http://www.genome.arizona.edu/information/protocols/BAC_DNA_prep.html">http://www.genome.arizona.edu/information/protocols/BAC_DNA_prep.html</a><br>(c) <a href="http://www1.qiagen.com/literature/handbooks/INT/plkkit.aspx">http://www1.qiagen.com/literature/handbooks/INT/plkkit.aspx</a> |
| 2. Fragmentation              | Reproducibility<br>Narrow range of fragment sizes<br>Fragment length<br>High yield fragmentation<br>Randomness                                                                                                                                              | (a) Partial digestion<br>(b) DNase I<br>(c) Sonication<br>(d) Nebulisation<br>(e) Hydroshear                        | (a) Maniatis <i>et al.</i> (1982)<br>(b) Anderson (1981)<br>(c) Deininger (1983)<br>(d) Bodenteich <i>et al.</i> (1994)<br>(e) Oefner <i>et al.</i> (1996);<br><a href="http://www.genemachines.com/hydroshear/index.html">http://www.genemachines.com/hydroshear/index.html</a>                                                                |
| 3. End repair                 | Less reaction and purification steps to minimize loss of the sample<br>Better ends for ligation<br>T4 DNA polymerase: 5' overhang filling and 3' overhang digestion<br>Klenow enzyme: 5' overhang repair<br>T4 polynucleotide kinase: phosphorylate 5' ends | (a) T4 DNA polymerase, Klenow enzyme and T4 polynucleotide kinase<br>(b) mixed enzyme kit<br>(c) mung-bean nuclease | (a) Wilson and Mardis (1997)<br>(b) Epicentre End-it kit<br><a href="http://www.epicentre.com">http://www.epicentre.com</a><br>(c) Wilson and Mardis (1997)                                                                                                                                                                                     |
| 4. Size selection and elution | High yield<br>Endonuclease free<br>Simplicity for high throughput                                                                                                                                                                                           | (a) Agarose gel electrophoresis and gel elution<br>(b) Column chromatography                                        | (a) Qiagen QIAEX II gel extraction system<br><a href="http://www.qiagen.com">http://www.qiagen.com</a><br>(b) Clontech CHROMA SPIN Columns<br><a href="http://www.clontech.com">http://www.clontech.com</a>                                                                                                                                     |

whole-genome sequencing, two types of libraries are typically constructed: 2–5 kilobase (kb) insert libraries in standard high-copy plasmid vectors, and 10 kb insert libraries constructed in medium-copy vectors. The most effective shotgun library construction method is blunt-end cloning of randomly fragmented DNA. This ensures that the library is as random as possible and it is not dependent upon the frequency of a particular restriction site. Shotgun library construction follows the general scheme below:

(1) BAC or genomic DNA is isolated using standard protocols. The DNA is fragmented to the desired size (e.g. GeneMachine's hydroshear, nebulisation and sonication), end repaired to generate blunt ends, size selected on agarose gels, eluted and concentrated.

For making shotgun libraries from large-insert genomic clones (cosmid/fosmid/BAC/PAC), the clones are first checked for cross contamination by comparing restriction–digestion patterns from

at least ten isolated single colonies, and the clone identity is validated by fingerprinting, hybridisation or BAC-end sequencing. The most common methods for favourable insert preparation are summarised in Table 5.2.

(2) High-copy or medium-copy vectors are digested with a blunt-end restriction endonuclease (e.g. *EcoR* V) and then dephosphorylated using calf intestinal alkaline phosphatase (CIP).

The choice of the shotgun-cloning vector depends on the strategy for sequencing. Shotgun cloning vectors are classified as single-stranded bacteriophages, double-stranded plasmids and phagemids (double-stranded DNA vectors that produce single-stranded products by helper bacteriophages). The features of these various shotgun cloning vectors are summarised in Table 5.3.

The single-stranded DNA approach, using M13 bacteriophage vectors, is a simple method that produces a high-quality sequence-grade template. Although these advantages are utilised in high-throughput approaches, the bacteriophage vector

**Table 5.3** Features of shotgun library vectors

| Class         | Name               | Size*<br>(bp) | Replicon†  | Copy<br>number‡ | Antibiotic<br>selection<br>marker§ | Insert<br>selection<br>gene¶ | Bluntend<br>cloning<br>site | Optimal<br>insert size<br>(kb) | Reference                                                                    |
|---------------|--------------------|---------------|------------|-----------------|------------------------------------|------------------------------|-----------------------------|--------------------------------|------------------------------------------------------------------------------|
| Bacteriophage | M13 series         | 7300          | P15A       |                 | KM                                 | <i>lac Z</i>                 | Sma I                       | 1.5                            | Yanisch-Perron <i>et al.</i> (1985)                                          |
| Phagemid      | pBluescript series | 3000          | pMB1       | 500–700         | Amp                                | <i>lac Z</i>                 | EcoRV                       | <10                            | Stratagene <a href="http://www.stratagene.com">http://www.stratagene.com</a> |
| Plasmid       | pUC series         | 2700          | pMB1       | 500–700         | Amp                                | <i>lac Z</i>                 | Sma I                       | <10                            | Yanisch-Perron <i>et al.</i> (1985)                                          |
| Plasmid       | pZERO series       | 3297          | ColE1      | 500–700         | KM                                 | <i>ccd B</i>                 | EcoRV                       | <10                            | Invitrogen <a href="http://www.invitrogen.com">http://www.invitrogen.com</a> |
| Plasmid       | pCUGIblu21         | 3064          | pMB1 + rop | >25             | KM                                 | <i>lac Z</i>                 | EcoRV                       | <10                            | Yang <i>et al.</i> (2003)                                                    |
| Plasmid       | pCUGIblu31         | 2401          | ColE1      | 500–700         | KM                                 | <i>lac Z</i>                 | EcoRV                       | <10                            | Yang <i>et al.</i> (2002)                                                    |

\* Small vector size is recommended for sequence-finishing phase.

† The pMB1 origin of replication is closely related to that of ColE1.

‡ High copy number plasmid is desirable for high-throughput sequencing and medium-copy number plasmid is desirable for construction of relatively large-insert (upto 10 kbp) library. This medium copy number plasmid is derived from pBR322. The difference in copy number is due to differences in the *rop* gene. Functional *rop* gene maintains the copy number at about 25 copies per cell by preventing access of DNA polymerase I to the RNA II primer, and thus inhibiting DNA replication of the plasmid.

§ Ampicillin in selection media results in undesirable satellite colonies; carbenicillin or penicillin-G is more stable.

¶ *lac Z* gene is a reporter gene to allow selection of positive protein-protein interactions by assaying for galactosidase activity. *ccd B* gene is the lethal *E. coli* gene that allows direct selection of insert via disruption of lethal gene.

performs poorly in regions containing repeated DNA sequence because of clone instability during growth and replication.

Double-stranded vectors provide a better representation with less clone bias, as well as stable maintenance of repetitive regions. With the development of cycle sequence protocols (Gyllenstein, 1989) and simpler, easier plasmid template generation methods (Kim, in preparation), the shotgun approach with a double-stranded vector is preferred.

Phagemid vectors can be used to generate both single- and double-stranded products, which introduce the advantage of stability from double-stranded vectors and high-quality sequence reading from single-stranded vectors. This type of vector has become more useful with the development of high-throughput template generation of plasmids and improved cycling-sequencing chemistry, which is desirable for either single- or double-strand template generation. Since a single-stranded template can be easily generated from the phagemid subclones, which are just like phagemids, they can be more useful for removing low-quality regions that are GC rich or contain a complex secondary structure.

Larger-insert shotgun libraries (e.g. 10 kbp, 50 kbp) require construction and the use of

medium- or low-copy number vectors for stable maintenance of inserts. The pCUGIblu21 vector (Table 5.3), a derivative of the medium-copy plasmid vector pBR322 (25 copies/cell), was created for larger-insert shotgun library construction and is publicly available (Yang *et al.*, 2003).

(3) The size-selected target DNA and linearised vector are ligated together and then electroplated into *E. coli*. The transformation mixture is plated onto a selective medium and grown overnight. Before picking sufficient clones for a library, a quality check is performed as a preliminary test. Templates of clones (e.g. 96 samples) are prepared and sequenced. The sequence is analysed by comparison to the *E. coli* genome sequence and vector sequences using the cross-match program (<http://www.phrap.org/>). In general, a library with less than 5% contamination is considered useful, and would be picked into 384 well-cultured plates containing appropriate media, stored at –80°C prior to convenient sequencing.

### Bacterial Artificial Chromosome Libraries

In 1994 our group was the first to construct a BAC library for a plant using the 750 Mb genome of *Sorghum bicolor* (Woo *et al.*, 1994).

Since then we have constructed a substantial number of deep-coverage BAC libraries, including Arabidopsis (Choi *et al.*, 1995), rice (Chen *et al.*, 2002), melon (Luo *et al.*, 2001), tomato (Budiman *et al.*, 2000), soybean (Tomkins *et al.*, 1999) and barley (Yu *et al.*, 2000), and have provided them to the community for genomics research (<http://www.genome.arizona.edu> and <http://www.genome.clemson.edu>).

Several detailed protocols have been published for the construction of high-quality plant and animal BAC libraries (Amemiya *et al.*, 1996; Birren *et al.*, 1997; Osoegawa *et al.*, 1998; Zhang *et al.*, 1996; Choi and Wing, 1999; Peterson *et al.*, 2000; Luo and Wing, 2003) including four from our laboratory. The construction of BAC libraries is different from that of a shotgun or lambda DNA library used as described above and follows the following general scheme:

1. Megabase-size DNA is isolated by embedding nuclei or protoplasts in agarose followed by DNA purification *in situ* (Wing *et al.*, 1993; Zhang *et al.*, 1995; Birren and Lai, 1993). The result is a 'porous bag' of high molecular weight DNA that has never been subjected to shearing. DNA in this state can be stored indefinitely and can easily be digested with restriction enzymes.
2. Megabase-size DNA is then partially digested with a restriction enzyme (e.g. *Hind*III, *Eco*R1, *Mbo*I) to generate a majority of DNA fragments in the size range of 100–300 kb. DNA fragments in the size range of 100–300 kbs are then size selected using pulsed-field gel electrophoresis in 1% agarose gels.
3. The size-selected DNA fragments are eluted from the gel, concentrated and then ligated to an appropriately digested and dephosphorylated BAC vector. The most commonly used BAC vectors are pBeloBAC11 or pIndigo-BAC536 (Shizuya and Kouros-Mehr, 2001).

Two BAC plant transformation vectors, binary BAC (BIBAC: Hamilton, 1997) and transformation-competent artificial chromosome (TAC: Liu *et al.*, 1999), have been developed, and can be used to integrate large DNA fragments into plant genomes. Libraries in such vectors have the potential to streamline positional cloning projects by eliminating the subcloning of individual candidate genes into

conventional plant transformation vectors (Liu *et al.*, 1999).

4. After ligation, the DNA is electropolated into high-efficiency, competent *E. coli*. The transformation mixture is then plated onto a selective medium and finally, the recombinant clones are individually picked into 384-well microtitre plates, grown overnight and stored at  $-80^{\circ}\text{C}$  until further use.
5. Once a BAC library is constructed it needs to be in a format that is easily accessible to the researcher. One of the most convenient methods available is to grid the library onto  $22\text{ cm} \times 22\text{ cm}$  nylon filters. After colony growth, the colonies are lysed and the DNA is denatured and fixed to the filter. The filters may be probed with single or pooled radio-labelled probes and the positive signals detected by autoradiography.

### Library Coverage

When constructing a genomic DNA library, the number of clones required to obtain adequate coverage can be calculated using the following formula (Clarke and Carbon, 1976):

$$N = \ln(1 - P) / \ln(1 - I/GS)$$

where  $N$  is the number of clones,  $P$  is the probability,  $I$  is the insert size and  $GS$  the genome size.

In general, a 5X genome-equivalent library will give a 95% chance of obtaining a particular clone of interest. In practice, for BAC libraries, a 10X genome-equivalent BAC library is sufficient for most purposes. However, if one is considering using BAC libraries to generate sequence-ready physical maps or to sequence a complete genome, then 2–3 10X BAC libraries should be constructed from two to three different restriction enzymes. Such a collection of libraries will help to ensure that most of the genome is covered with BAC clones.

For shotgun-sequencing libraries, clones are assembled by overlapping sequence reads. Because of the probabilistic nature of the overlaps, high redundancy is required for complete coverage. Coverage is defined as the number of subclone sequence reads that span the large-insert source clone. Theoretically, an eight-fold average redundancy covers 99% of target DNA with 500 bp read

length. This, or even higher redundancy is required for high accuracy and completeness for whole-genome sequencing.

The number of reads for desired coverage can be determined by the formula:

$$N = (CS/L)/R$$

where  $N$  is the number of reads,  $C$  is the desired coverage,  $S$  is the size of the target DNA,  $L$  is the high-quality sequence read length and  $R$  is the sequencing successful ratio (%).

However, for positional cloning, when the primary purpose of a shotgun library is to identify and isolate single-candidate coding exons, the precision and completeness of the sequencing become less important. Cleverie (1994) reported that 2.5-fold coverage is sufficient to determine approximately 92% of a target sequence of any length which will allow most coding exons to be discovered in a positional cloning experiment.

## Applications

### Positional Cloning

Positional cloning can be defined as the isolation of a gene based on its position on a genetic map and usually involves four basic steps: (a) fine genetic mapping of the target gene with molecular markers; (b) determination of the relationship between genetic and physical distance (PFGE, FISH); (c) chromosome walking or landing using a BAC library; and (d) gene identification (e.g. genetic complementation, shotgun sequencing).

Over 130 plant genes have been positionally cloned in the past 10 years with a majority of examples using BAC clones.

### Physical Mapping

Since high-quality deep-coverage BAC libraries exist for the majority of crop plants, it is now practical to consider the development of whole-genome physical maps for these crops. Whole-genome physical maps are constructed by generating restriction fragment fingerprints for each BAC clone of a library and assembling the fingerprints into a contig based on shared bands using software like FPC (Fingerprint Contig: Soderlund *et al.*,

1997). Additional information can be added by sequencing the ends of all the BAC clones to generate a sequence-tagged connector database.

Whole-genome physical maps have been published for Arabidopsis and rice and several others are in progress including maize, sorghum, tomato, wheat, cotton, medicago and soybean.

For rice, Chen *et al.* (2002) fingerprinted over 72,000 BAC clones from two libraries representing 20X the genomic coverage. The fingerprints were assembled into 458 contigs, half of which could be easily anchored to the rice genome.

A new method of fingerprinting called high-information content fingerprinting technique (HICF) (Ding *et al.*, 2001) is emerging. This method employs a double restriction enzyme digestion (a type II 4-base cutter and a type IIS 6-base cutter) leaving one 5' overhang available for a fill-in reaction with fluorescently labelled dNTP. Chromatograms are analysed following electrophoresis on a fluorescence-detecting DNA analyser. Fragment sizes, and the identities of the labelled base at the 5' end, are uploaded into an adapted FPC program called multiFPC. Assembly is five- to seven-fold more efficient relative to agarose and the average number of fragments per clone increases from 25 for agarose to 90 with HICF. Probability ratios are also dramatically reduced by three- to fivefold. Utilisation of high-throughput DNA capillary sequence analysers also will streamline this improved method. Data handling is accomplished completely by mechanical means as opposed to a semi-manual scoring of fragments of gel images, thus allowing for digital files of sizes and tagged bases at markedly improved throughput rates with significant reductions in errors.

As fingerprinted genomes become available, greatly reduced numbers of BACs may be tiled together to represent a genome. These minimum tiling libraries consist of identified BACs from a fingerprinted library such that one BAC overlaps the next BAC and so forth, encompassing an entire chromosome. These minimum-tile physical mapping libraries will provide researchers with concise resources to reduce time and costs.

### Shotgun Sequencing

Shotgun sequencing is a standard approach for complete DNA sequencing of large-insert genomic

clones (e.g. cosmids, PACs and BACs) as well as entire genomes (e.g. Arabidopsis, human, yeast, fruitfly and many microbial genomes). More than 1000 completely sequenced organisms, and those whose sequencing is in progress (e.g. rice and corn), are summarised in the Entrez Genome database. (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>).

There are two strategies for large-scale genomic DNA sequencing using the shotgun approach. The first is clone-by-clone shotgun sequencing (Figure 5.1) which was used for yeast, *C. elegans* (Goffeau *et al.*, 1996; The *Caenorhabditis elegans* Sequencing Consortium, 1998), human (The International Human Genome Mapping Consortium, 2001) and Arabidopsis genomes (The Arabidopsis Genome Initiative, 2000). This strategy requires a complete sequence-ready physical map. A sequence-ready physical map is a set of cloned DNA fragments whose position in the genome, relative to each other, is known. The selected clones are randomly fragmented and subcloned into an appropriate vector. The randomly

fragmented shotgun libraries are sequenced to 8X coverage, assembled and finished (remove gaps and ambiguous or low-quality regions). Some disadvantages of this method are that a physical map is required and the procedure is slow and has significant overlaps between adjacent clones. However, it is commonly used as a standard procedure for eukaryotic genomes that have many repeats or heterochromatin regions.

The second approach is the whole-genome shotgun (WGS) approach (Venter *et al.*, 1996). This procedure, outlined in Figure 5.1, utilises whole genome libraries with a variety of insert sizes (microbial genome: Fleischmann *et al.*, 1995, Fraser and Fleischmann, 1997; *Drosophila* genome: Adams *et al.*, 2000; human genome: Weber and Myers, 1997; Venter *et al.*, 2001). Genomic DNA is fractionated into three sized fractions (e.g. 2 kbp, 10 kbp and 50 kbp) and used to construct plasmid libraries to cover the entire genome in an unbiased manner, thereby reducing potential misassemblies. These libraries are used for sequencing, generating roughly ten-fold coverage and assembled using

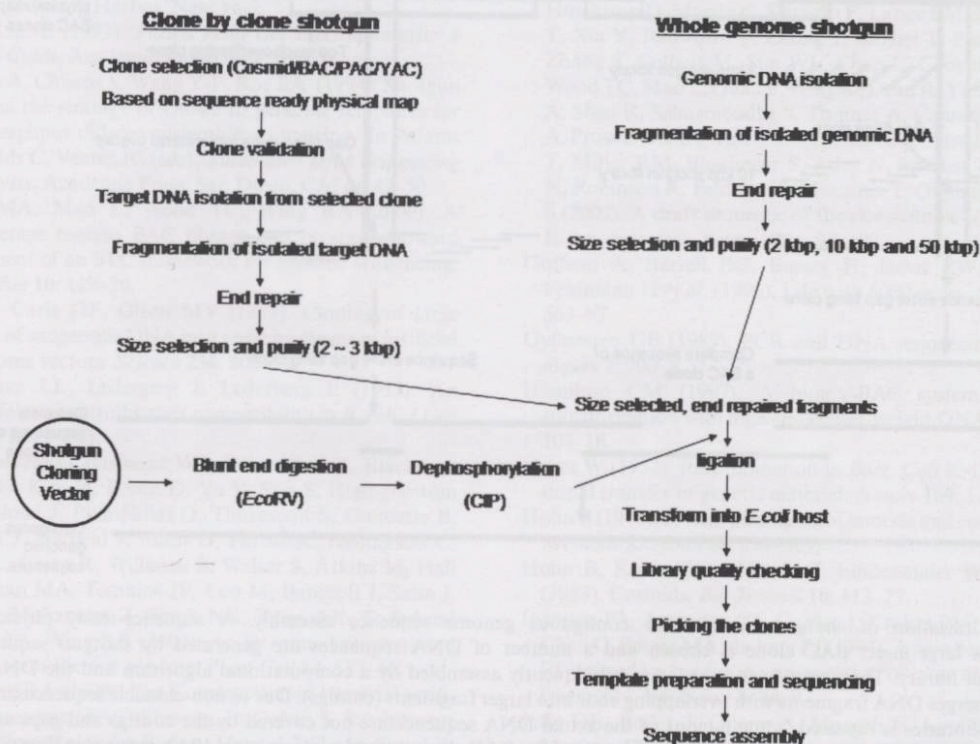


Fig. 5.1 Strategies of clone-by-clone, shotgun and whole-genome shotgun libraries used for large-scale genomic DNA sequencing

software (Celera Assembler for *Drosophila* WGS, Myers *et al.*, 2000; ARACHINE Assembler for mouse WGS, Jaffe *et al.*, 2003; JAZZ Assembler for *Fugu rubripes* WGS, Aparicio *et al.*, 2002; RePS Assembler for rice, *indica*, WGS, Yu *et al.*, 2002; Myriad Assembly Program for rice, *japonica*, WGS, Goff *et al.*, 2002; Phusion Assembler for mouse WGS, and *C. briggsae* WGS, Mullikin and Ning, 2003). The WGS approach is a rapid and economic strategy for sequencing some genomes because it does not require any preliminary data. However, it is not a good approach for complex eukaryotic genomes with repeats and heterochromatin regions that will cause sequence misassemblies and/or poor coverage.

For both sequencing approaches, different sets of libraries (different enzymes or sizes) are required for gap filling or coverage with lesser bias. Ten kilobase shotgun libraries (shotgun library with 10-kbp insert size) have become popular as addi-

tional resources for large DNA sequencing due to their ease and efficient construction in comparison with larger insert libraries (e.g. BAC library or cosmid and fosmid library). The insert size, 10 kb, is ideal for filling sequencing gaps identified after an assembly of small-insert library sequences and for filling clone gaps with minimal overlap, adding to its cost effectiveness (Figure 5.2). Moreover, whole-genome 10 kb libraries are valuable resources for an isolation of complete gene structures because most of the predicted genes from higher organisms are less than 8 kb in size. These libraries may be used for small-scale comparative genome studies by orthologous gene comparison or for identification of full gene structure with expressed sequence tags (ESTs).

Shotgun approaches are also important tools for positional gene cloning. Because many agriculturally important traits are controlled by genes, whose products are unknown, the final step in positional

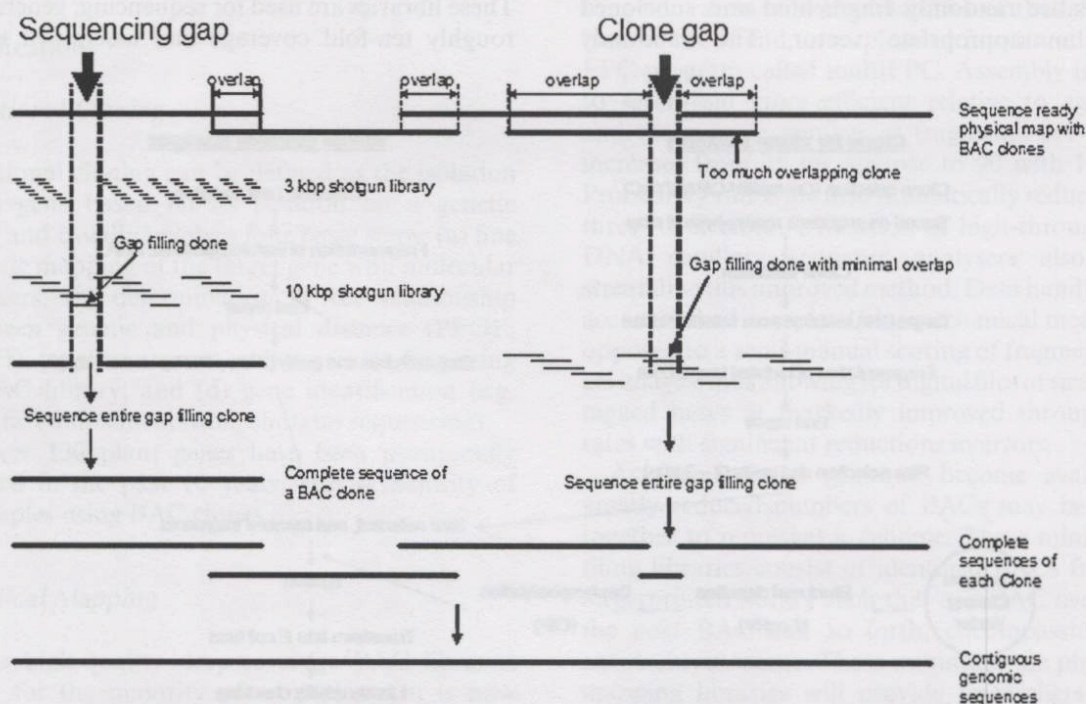


**Fig. 5.2** Utilisation of shotgun libraries for contiguous genome sequence assembly. A sequence-ready physical map is produced, a large-insert BAC clone is chosen and a number of DNA sequences are generated by shotgun sequencing of a 3 kb shotgun library. These sequencing reads are subsequently assembled by a computational algorithm and the DNA sequence assembler merges DNA fragments with overlapping ends into larger fragments (contigs). Due to non-clonable sequences, randomness of shotgun libraries is repeated, some regions of the actual DNA sequence are not covered by the contigs and gaps are observed (sequencing gaps). The sequence-ready physical map from the BAC clones contains clone gaps. 10 kb insert size library is ideal for filling sequencing gaps identified after assembly of small-insert library sequences and/or for filling clone gaps with minimal overlap thus adding to its cost effectiveness in comparison to whole BAC sequencing

cloning is isolating the target gene from a larger genomic DNA insert clone. Subclone shotgun libraries that are created from a differentially fragmented target gene, a containing large clone, provide a systematic approach to final gene identification and sequencing.

## REFERENCES

- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA *et al.* (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–95.
- Amemiya CT, Ota T, Litman GW (1996). Construction of P1 artificial chromosome (PAC) libraries from lower vertebrates. In Birren B, Lai E (eds), *Nonmammalian Genomic Analysis: A Practical Guide*. Academic Press: New York; pp 223–56.
- Anderson C (1993). Genome shortcut leads to problems. *Science* **259**: 1684–87.
- Anderson S (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res* **9**: 3015–27.
- Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit AF (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**: 1301–10.
- Birren B, Green ED, Klapholz S, Myers RM, Roskams J (1997). *Analyzing DNA*, Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.
- Birren B, Lai E (1993). *Pulsed Field Gel Electrophoresis: A Practical Guide*, Academic Press: New York.
- Bodenteich A, Chisoe S, Wang Y-F, Roe BA (1994). Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing. In Adams MD, Fields C, Venter JC (eds), *Automated DNA Sequencing and Analysis*, Academic Press: San Diego, CA; pp 42–50.
- Budiman MA, Mao L, Wood TC, Wing RA (2000). A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Res* **10**: 129–36.
- Burke DT, Carle GF, Olson MV (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**: 806–12.
- Cavalli-Sforza LL, Ledergerg J, Lederberg E (1953). An infective factor controlling sex compatibility in *B. coli*. *J Gen Microbiol* **8**: 89–103.
- Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, Higingbottom S, Phimpilalai J, Phimpilalai D, Thurmond S, Gaudette B, Li P, Liu J, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, Williams B, Walser S, Atkins M, Hall C, Budiman MA, Tomkins JP, Luo M, Bancroft I, Salse J, Regad F, Mohapatra T, Singh NK, Tyagi AK, Soderlund C, Dean RA, Wing RA (2002). An integrated physical and genetic map of the rice genome. *Plant Cell* **14**: 537–45.
- Choi S, Wing RA (1999). The Construction of bacterial artificial chromosome (BAC) libraries. In Gelvin S, Schilperoort R (eds), *Plant Molecular Biology Manual*, 2nd edn. Suppl. 4, Kluwer Academic: The Netherlands; pp 1–32.
- Choi SD, Creelman R, Mullet J, Wing RA (1995). Construction and characterisation of a bacterial artificial chromosome library from *Arabidopsis thaliana*. *Weeds World* **2**: 17–20.
- Clarke L, Carbon J (1976). A colony bank containing synthetic col EI hybrid plasmids representative of the entire *E. coli* Genome. *Cell* **9**: 91–9.
- Claverie JM (1994). A streamlined random sequencing strategy for finding coding exons. *Genomics* **23**: 575–81.
- Deininger PL (1983). Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal. Biochem* **129**: 216–23.
- Ding Y, Johnson MD, Chen WQ, Wong D, Chen YJ, Benson SC, Lam JY, Kim YM, Shizuya H (2001). Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics* **74**: 142–54.
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–512.
- Fraser CM, Fleischmann RD (1997). Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18**: 1207–16.
- Gingrich JC, Boehrer DM, Garnes JA, Johnson W, Wong BS, Bergmann A, Eveleth GG, Langlois RG, Carrano AV (1996). Construction and characterisation of human chromosome 2-specific cosmid, fosmid and PAC clone libraries. *Genomics* **32**: 65–74.
- Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92–100.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H *et al.* (1996). Life with 6000 genes. *Science* **274**: 563–67.
- Gyllenstein UB (1989). PCR and DNA sequencing. *Biotechniques* **7**: 700–8.
- Hamilton CM (1997). A binary-BAC system for plant transformation with high-molecular-weight DNA. *Gene* **200**: 107–16.
- Hayes W (1952). Recombination in *Bact. Coli* K-12: unidirectional transfer of genetic material. *Nature* **169**: 118–9.
- Hohn B (1979). *In vitro* packaging of lambda and cosmid DNA. *Methods Enzymol* **68**: 299–309.
- Hohn B, Koukolikova-Nicola Z, Lindenmaier W, Collins J (1988). Cosmids. *Bio Technol* **10**: 113–27.
- Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H, Chen C, Batzer MA, de Jong PJ (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat Genet* **6**: 84–9.
- Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES (2003). Whole-genome sequence assembly for mammalian genomes: arachne 2. *Genome Res* **13**: 91–6.

- Kim HR. A high-throughput and high quality plasmid extraction procedure, in preparation.
- Kim U-J, Shizuya H, de Jong P, Birren B, Simon M (1992). Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res* **20**: 1083-85.
- Liu YG, Shirano Y, Fukaki H, Yanai Y, Tasaka M, Tabata S, Shibata D (1999). Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. *Proc Natl Acad Sci USA* **96**: 6535-40.
- Luo M, Wang Y-H, Frisch D, Joobeur T, Wing RA, Dean RA (2001). Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon *Fusarium* wilt (*Fom-2*). *Genome* **44**: 154-62.
- Luo M, Wing RA (2003). An improved method for plant BAC library construction. In Grotewold E (ed), *Methods in Molecular Biology, Vol 236 Plant Functional Genomics: Methods and Protocols*, Humana Press: Totowa, NJ (in press).
- Maniatis T, Fritsch EF, Sambrook J (1982). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.
- Mullikin JC, Ning Z (2003). The phusion assembler. *Genome Res* **13**: 81-90.
- Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA (2000). A whole-genome assembly of *Drosophila*. *Science* **287**: 2196-204.
- Oefner PJ, Hunnicke-Smith SP, Chiang L, Dietrich F, Mulligan J, Davis RW (1996). Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. *Nucleic Acids Res* **24**: 3879-86.
- Osoegawa K, Woon PY, Zhao B, Frengen E, Tateno M, Catanese JJ, de Jong PJ (1998). An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52**: 1-8.
- Peterson DG, Tomkins JP, Frisch DA, Wing RA, Paterson AH (2000). Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. *J Agr Genomics* **5**: <http://www.ncgr.org/jag>.
- Sambrook J, Fritsch EF, Maniatis T (2001). *Molecular Cloning: A Laboratory Manual*, 3rd edn. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.
- Shizuya H, Birren B, Kim U-J, Mancino V, Slepak T, Tachiiri Y, 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**: 8794-997.
- Shizuya H, Kourou-Mehr H (2001). The development and applications of the bacterial artificial chromosome cloning system. *Keio J Med* **50**: 26-30.
- Soderlund C, Longden I, Mott R (1997). FPC: a system for building contigs from restriction fingerprinted clones. *CABIOS* **13**: 523-35.
- The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.
- The *Caenorhabditis elegans* Sequencing Consortium (1998). Genome sequence of the nematode *Caenorhabditis elegans*. A platform for investigating biology. *Science* **282**: 2010-8.
- The International Human Genome Mapping Consortium (2001). A physical map of the human genome. *Nature* **409**: 934-41.
- The Phred/Phrap/Consed System Home Page. <http://www.phrap.org/>.
- Tomkins JP, Mahalingam R, Smith H, Goicoechea JL, Knap HT, Wing RA (1999). A bacterial artificial chromosome library for soybean PI 437654 and identification of clones associated with cyst nematode resistance. *Plant Mol Biol* **41**: 25-32.
- Venter JD, Adams MD, Myers EW, Li PW, Mural RJ, Sutton SS, Smith HO, Yandell M, Evans CA, Holt RA *et al.* (2001). The sequence of the human genome. *Science* **291**: 1304-51.
- Venter JD, Smith HO, Hood L (1996). A new strategy for genome sequencing. *Nature* **381**: 364-6.
- Weber JL, Myers EW (1997). Human whole-genome shotgun sequencing. *Genome Res* **7**: 401-9.
- Willetts N, Skurray R (1987). Structure and function of the F factor and mechanisms of conjugation. In FC Neidhardt *et al.* (eds), *Escherichia coli and Salmonella typhimurium, Vol 2*. American Society for Microbiology: Washington, DC; pp 1110-33.
- Wilson RK, Mardis ER (1997). Shotgun sequencing. In Birren B, Green E, Klapholz S, Myers RM, Roskams J (eds), *Genome Analysis: A Laboratory Manual, Vol 1*. Analysing DNA, Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York; pp 397-454.
- Wing RA, Rastogi V, Zhang H-B, Paterson AH, Tanksley SD (1993). An improved method of plant megabase DNA isolation in agarose microbeads suitable for physical mapping and YAC cloning. *The Plant J* **4**: 893-8.
- Woo S-S, Jiang J, Gill BS, Paterson AH, Wing RA (1994). Construction and characterisation of a bacterial artificial chromosome library of sorghum bicolor. *Nucleic Acid Res* **22**: 4922-31.
- Yang TJ, Yu Y, Frisch D, Wing RA (2002). Two series of plasmid vectors for shotgun library. *Plant & Animal Genome Conference X* p 24.
- Yang TJ, Yu Y, Nah GJ, Atkins M, Lee S, Frisch D, Wing RA (2003). Construction and utility of 10 kb shotgun libraries for efficient clone gap closure for rice genome sequencing. *Theor Appl Genet* (in press).
- Yanisch-Perron C, Vieira J, Messing J (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**: 103-9.
- Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79-92.
- Yu Y, Tomkins JP, Waugh R, Frisch DA, Kudrna D, Kleinhofs A, Brueggeman RS, Muehlbauer GJ, Wise RP, Wing RA (2000). A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. *Theor Appl Genet* **101**: 1093-9.
- Zhang HB, Wing RA (1997). Physical mapping of the rice genome with BACs. *Plant Mol Biol* **35**: 115-27.
- Zhang HB, Woo SS, Wing RA (1996). BAC, YAC & Cosmid Library Construction. In Foster G, Twell D (eds), *Plant Gene Isolation: Principles and Practice*, John Wiley & Sons: New York. pp 75-99.
- Zhang HB, Zhao XP, Patterson AH, Wing RA (1995). Preparation of megabase DNA from plant nuclei. *The Plant J* **7**: 175-84.