Preface

Functional genomics is a young discipline whose origin can be traced back to the late 1980s and early 1990s, when molecular tools became available to determine the cellular functions of genes. Today, functional genomics is perceived as the analysis, often large-scale, that bridges the structure and organization of genomes and the assessment of gene function. The completion in 2000 of the genome sequence of *Arabidopsis thaliana* has created a number of new and exciting challenges in plant functional genomics. The immediate task for the plant biology community is to establish the functions of the approximately 25,000 genes present in this model plant.

One major issue that will remain even after this formidable task is completed is establishing to what degree our understanding of the genome of one model organism, such as the dicot Arabidopsis, provides insight into the organization and function of genes in other plants. The genome sequence of rice, completed in 2002 as a result of the synergistic interaction of the private and public sectors, promises to significantly enrich our knowledge of the general organization of plant genomes. However, the tools available to investigate gene function in rice are lagging behind those offered by other model plant systems. Approaches available to investigate gene function become even more limited for plants other than the model systems of Arabidopsis, rice, and maize.

The challenge to determine the function of the tens of thousands of plant genes, many of them showing no detectable homology to genes for which cellular roles have been identified in bacteria, yeast, or animals, has triggered an avalanche of novel methodologies. The aim of *Plant Functional Genomics: Methods and Protocols* is to provide in a single volume a detailed description and guide to some of the most commonly used approaches to investigating plant gene function. Rather than focusing solely on model organisms, this collection also covers recent efforts devised for investigating a wide variety of plants, plant pathogens, and even some algae.

Plant Functional Genomics: Methods and Protocols is organized into five parts, three of which represent the detailed sequence of steps in which a protocol for the discovery of genes and their functions is carried out. Chapters in the first part describe how to identify genes in complex systems that include large genomes, few cells, and mixed cell systems (such as plant and pathogens together). The second part describes powerful computational and statistical tools to help predict gene function on the basis of comparative genomics or from the

analysis of complex genome sequences. Descriptions in the third part focus on several methods that permit the discovery of gene function by loss-of-function mutant analyses, a classical approach that remains very useful. However, it is evident that the high level of genetic redundancy present in large genomes creates formidable obstacles to the successful identification of mutant phenotypes. Thus, gain-of-function approaches can often be a powerful complement to mutant analyses. Effective methods for gain-of-function studies are covered in the fourth part. Finally, because establishing gene function relies on the identification of phenotypes, chapters in the fifth part expand the concept of phenotypes, including the use of multiple outputs as the ultimate phenotypic result of changes in gene activity.

In its assessment of the quality of both new and established technologies, *Plant Functional Genomics: Methods and Protocols* is aimed at plant biologists with a wide range of interests. The combination of detailed computational, molecular, and genetic protocols focusing on both general and specific problems should allow scientists with little or no experience in the specific areas covered to investigate gene function associated with their particular system of interest, and to do so using the most recent methodologies.

To conclude, I would like to thank all those who helped with revising and compiling the material for this collection, especially Diane Furtney. Special thanks go to the authors of the individual chapters who responded with patience and enthusiasm to my numerous requests for additional information or format changes.

Erich Grotewold

2

Constructing Gene-Enriched Plant Genomic Libraries Using Methylation Filtration Technology

Pablo D. Rabinowicz

Summary

Full genome sequencing in higher plants is a very difficult task, because their genomes are often very large and repetitive. For this reason, gene targeted partial genomic sequencing becomes a realistic option. The method reported here is a simple approach to generate gene-enriched plant genomic libraries called methylation filtration. This technique takes advantage of the fact that repetitive DNA is heavily methylated and genes are hypomethylated. Then, by simply using an *Escherichia coli* host strain harboring a wild-type modified cytosine restriction (McrBC) system, which cuts DNA containing methylcytosine, repetitive DNA is eliminated from these genomic libraries, while low copy DNA (i.e., genes) is recovered. To prevent cloning significant proportions of organelle DNA, a crude nuclear preparation must be performed prior to purifying genomic DNA. Adaptor-mediated cloning and DNA size fractionation are necessary for optimal results.

Key Words

gene-enriched libraries, shotgun sequencing, Mcr, DNA methylation, retrotransposons, gene discovery, repetitive DNA

1. Introduction

Highly accurate full genomic sequencing like that performed for example in *Saccharomyces cerevisiae* ***(1)*** and *Caenorhabditis elegans* ***(2)*** has proven to be an invaluable resource to accelerate all areas of biological research. In particular in plants, the *Arabidopsis thaliana* genome sequence has been deciphered, meeting the highest standards of accuracy ***(3)***. Undoubtedly, the availability of this information had an immense impact not only in the *Arabidopsis* community, but in research in all other plant systems as well. Unfortunately, the production of such a high quality genomic resource is not an easy task. It implies

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

a significant amount of sequence redundancy only achievable by producing a huge number of sequence reads. Such reads are assembled and processed to produce as long contiguous stretches as possible, called contigs. In order to link these contigs in the right order and orientation, a large insert genomic library (using bacterial artificial chromosome [BAC] or P1-derived artificial chromosome [PAC] vectors) needs to be constructed, at least partially sequenced, and physically mapped.

A major obstacle to obtain the complete and accurate sequence of a complex (i.e., eukaryote) genome is the presence of large amounts of repetitive DNA. This DNA is composed of satellite DNA, transposons and retrotransposons, among other repeats, which often show a high degree of sequence conservation. For this reason, the computer software designed to assemble random sequence reads fails to build correct contigs of repetitive sequences, usually assembling most members of a repeat family in a single contig, regardless of their actual location in the genome.

In the early 1980s by the time the idea of sequencing the human genome was opened to discussion for the first time ***(4)***, Putney et al. ***(5)*** reported a method that allowed to discover new genes simply by cDNA sequencing, later called expressed sequence tag (EST) sequencing ***(6)***. This widely used technique allows obtaining gene sequence information getting around the problem of sequencing repetitive DNA. However, the EST approach has two main limitations. The first is the redundancy of cDNA libraries. Some cDNAs are often overrepresented and will be sequenced many times before a cDNA corresponding to a weakly expressed gene is found. The second limitation is the partial representation due to the tissue-specific and developmental regulation of gene expression. Some genes are expressed only in certain tissues or cells, and some are developmentally regulated. In order to recover the corresponding ESTs, libraries from several different tissues and developmental stages need to be constructed. Another although minor, disadvantage of EST sequencing is that repetitive elements are often transcribed and thus included in EST collections.

One way to solve the problem of the redundancy is to use normalized libraries ***(7)***. Normalization techniques are based on reassociation kinetics and have been improved to avoid the elimination of members of gene families. However, it is not trivial to obtain a normalized library where representation is acceptable. Regardless of these limitations, EST projects are being conducted for many organisms and are a key tool for gene discovery, annotation of genes, cross-species comparative analysis, and definition of intron–exon boundaries among many other uses. In particular for plants, ESTs have been the alternative to full genome sequence, because the genomes of many plants, often important crop species, are very large and repetitive. Usually, the genome size (or subgenome size in the case of polyploids) correlates with the proportion of

repetitive DNA. It has been proposed that all diploid higher plant genomes share essentially the same set of genes, called the "gene space" ***(8)***. Then, the bigger the genome, the higher sequencing cost per gene, due to the amount of nongenic (e.g., repetitive) DNA that needs to be sequenced before reaching a gene.

The conservation of coding sequences across different species allows identifying genes simply by comparing two different genomes. Frequently, gene modeling software fails to identify genes that can be spotted with this comparative genomics approach. Furthermore, once the complete genomic sequence is obtained for one organism, it can be compared to a draft (lowly redundant and discontinuous) sequence of a related organism. This approach yields a lot of new information for both species under analysis. The additional advantage of genomic vs cDNA sequencing in terms of representation makes the lowly redundant genomic sequencing a cost-effective process. In the case of plants however, the large genome sizes prevent the pursuit of full or even draft genomic sequencing projects. For these reasons, alternatives to obtain genomic sequences enriched in genes avoiding the repetitive DNA have been developed. In maize for example, the very active transposon *Mutator* ***(9)*** shows a strong bias to insert in low copy DNA (i.e., genes). By generating large *Mutator*-induced insertional mutagenesis, it is possible to collect genomic sequences flanking transposon insertion sites, which will mainly correspond to genes ***(10)***. Although *Mutator* insertions may not be completely at random in the genome, it can be a good complement to an EST project.

Another alternative for gene enriched genomic sequencing of plants is the methylation filtration technique, which takes advantage of the fact that most of the repetitive elements in plants are heavily methylated, while genes are hypomethylated. Because of their methylation status, repeats are sensitive to bacterial restriction-modification systems, in particular the Mcr system ***(11,12)***, which includes two restriction enzymes: McrA and McrBC. McrBC recognizes DNA containing 5-methylcytosine preceded by a purine ***(13)***. Restriction requires two of these sites separated by 40–2000 nucleotides. Such recognition sites are very frequent in any methylated genomic DNA. Thus, by the selecting a *mcrBC*$^{+}$ *Escherichia coli* host strain, repetitive DNA can be largely excluded from genomic shotgun libraries, preserving the low copy DNA. Basically, methylation filtration consists in shearing and size fractionation of genomic DNA to select fragments smaller than the estimated size of the genes. Larger fragments have a high probability of including some portion of repetitive DNA, which would be methylated and thus counter-selected in the filtered library. On the other hand, if fragments are too small, there are more chances to recover small fragments of repetitive DNA with low GC content. Such fragments may be poor in methylated sites susceptible to restriction by McrBC and

then can be frequently recovered in filtered libraries. The selected fragments are then end-repaired and cloned into a standard sequencing vector. Subsequently, the ligation is introduced in a *mcrBC*$^+$ *E. coli* host. The recombinant clones isolated after plating are picked for automatic sequencing. The same ligation mixture can be transformed into a *mcrBC*$^-$ *E. coli* strain to obtain an unfiltered control library.

The technique works very well for maize ***(14)***, and there is evidence that it works for many other plants (Rabinowicz and Martienssen, unpublished). The advantage of methylation-filtered libraries vs cDNA and transposon insertion libraries is that there is no bias towards a certain region of the genome or a given fraction of the genes. It is possible though, that methylated genes are not recovered in filtered libraries. However, gene methylation is often restricted to defined regions of the gene, mainly the ends ***(15–17)***. This would allow to clone at least most of the coding sequence of methylated genes. Furthermore, genes that are regulated by methylation may become demethylated during different developmental stages. In these cases, the construction of methylation-filtered libraries from a couple of developmental stages of a given plant would likely overcome the problem. For larger scale projects, another problem is posed by the cloning efficiency. In plants with very large genomes, repetitive DNA may account for more than 90% of the nuclear DNA. Then, most of the DNA is likely to be methylated leaving a very small fraction of the genome to be recovered in methylation-filtered libraries. As a result, the number of recombinant clones recovered after plating a filtered library may be <10% of the number of clones obtained in the corresponding unfiltered control library. Furthermore, the proportion of nonrecombinant background (blue colonies) may become significant. The use of adaptors often improves the cloning efficiency in addition to reduce the formation of chimerical clones. The cloning protocol presented here uses three-nucleotide overhang adaptors and a compatible sticky-end vector made by filling in one nucleotide in the four-nucleotide 5' overhang generated by a restriction nuclease ***(18)***. The advantage of using three- vs four-nucleotide overhang is that the nonrecombinant background is highly reduced because the vector ends become incompatible.

2. Materials

2.1. Nuclear DNA Preparation

1. Isolation buffer 1 (IB 1): 25 m*M* citric acid (pH to 6.5 with 1 *M* NaOH), 250 m*M* sucrose, 0.7% Triton® X-100, 0.1% 2-mercaptoethanol (*see* **Note 1**). IB 1 can be prepared at a 5× concentration. 2-Mercaptoethanol should be added immediately before usage.
2. Centrifuge tubes.
3. Liquid N_2.

4. Blender.
5. Polytron (Brinkmann Instruments).
6. Two 15-cm wide funnels.
7. Ring stand and clamps.
8. Cheesecloth (Fisher Scientific).
9. 60-μm Nylon mesh (Millipore).
10. 500-mL Centrifuge bottles with rubber o-ring sealing cap (Nalgene).
11. Isolation buffer 2 (IB 2): 50 m*M* Tris-HCl, pH 8.0, 25 m*M* EDTA, 350 m*M* sorbitol 0.1% 2-mercaptoethanol.
12. 5% Sarkosyl.
13. 5 *M* NaCl.
14. CTAB solution: 8.6% CTAB (Sigma), 0.7 *M* NaCl.
15. Chloroform:octanol (24:1).
16. Isopropanol.
17. 70% ethanol.
18. 10 m*M* Tris-HCl, pH 8.0.
19. Glass rod with bent tip.

2.2. DNA Shearing and End-Repairing

1. Glycerol 50%.
2. 10× Nebulization buffer: 0.5 *M* Tris-HCl, pH 8.0, 150 m*M* $MgCl_2$.
3. 14-mL Falcon® tubes (Becton Dickinson, cat. no. 35–2059).
4. Aero-mist nebulizer (CIS-US; cat. no. CA-209).
5. N_2 gas cylinder with a regulator able to deliver 1–50 psi.
6. Three-sixteenths-inch internal diameter PVC tubing (Fisher Scientific).
7. Parafilm.
8. 5 *M* NaCl.
9. Ethanol.
10. 70% Ethanol.
11. SpeedVac® (Savant Instruments).
12. 5 m*M* Tris-HCl, pH 8.0.
13. dNTPs 0.5 m*M* each (Roche Molecular Biochemicals).
14. T4 DNA polymerase (New England Biolabs).
15. T4 DNA polymerase buffer (New England Biolabs).
16. Klenow enzyme (Roche Molecular Biochemicals).
17. QIAquick™ polymerase chain reaction (PCR) purification kit (Qiagen).
18. T4 Polynucleotide kinase (PNK) (New England Biolabs).
19. T4 PNK buffer (New England Biolabs).
20. 100 m*M* ATP (Roche Molecular Biochemicals).
21. Equilibrated phenol:chloroform (1:1).

2.3. Adaptor Ligation

1. 200 μ*M* Top adaptor oligonucleotide 5'[P]-TAGACGCCTCGAG.
2. 200 μ*M* Bottom adaptor oligonucleotide 5'[OH]-CTCGAGGCGT.
3. 1 *M* NaCl.
4. T4 DNA ligase (Roche Molecular Biochemicals).
5. T4 DNA ligase buffer (Roche Molecular Biochemicals).
6. TEN buffer: 10 m*M* Tris-HCl, pH 7.5, 0.1 m*M* EDTA, 25 m*M* NaCl.
7. cDNA size fractionation columns (Invitrogen, Carlsbad, CA, USA).

2.4. Vector Preparation

1. Supercoiled pUC 19 DNA.
2. *Xba*I (Roche Molecular Biochemicals).
3. H buffer (Roche Molecular Biochemicals).
4. L buffer (Roche Molecular Biochemicals).
5. 10 mg/mL bovine serum albumin (BSA) (New England Biolabs).
6. 1 m*M* dCTP (Roche Molecular Biochemicals).
7. Klenow enzyme (Roche Molecular Biochemicals).
8. Calf intestinal phosphatase (CIP) (Roche Molecular Biochemicals).
9. CIP buffer (Roche Molecular Biochemicals).
10. 0.5 *M* EDTA.
11. Equilibrated phenol:chloroform (1:1).
12. QIAquick PCR purification kit.
13. Chloroform.
14. 5 *M* NaCl .
15. Ethanol.
16. 70% Ethanol.
17. 10 m*M* Tris-HCl, pH 8.0.

2.5. Preparation of Electrocompetent Cells

1. SOB medium without magnesium: 20 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 2.5 m*M* KCl, and 0.5 g/L NaCl (pH 7.0 with NaOH, autoclaved).
2. 10% Glycerol (autoclaved).
3. Sterile 250-mL centrifuge bottles with rubber o-ring sealing cap.
4. Sterile 14-mL centrifuge tubes.

2.6. Electroporation

1. Electroporation cuvettes 0.1 cm (Bio-Rad).
2. Electroporator (Bio-Rad).
3. SOC medium: 20 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 2.5 m*M* KCl, and 0.5 g/L NaCl (pH 7.0 with NaOH, autoclaved, sterile 2 *M* $MgCl_2$, and 1 *M* glucose are added to a final concentration of 10 and 20 m*M*, respectively, after cooling down).
4. Sterile 14-mL centrifuge tubes.

5. Isopropyl β-D-thiogalactopyranoside (IPTG) 200 mg/mL.
6. 5-Bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) 20 mg/mL in dimethylformamide.
7. LB-ampicillin agar plates: 10 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 10 g/L NaCl (pH 7.0 with NaOH); agar is added to a final concentration of 1.5%, autoclaved, cooled to 55°C, ampicillin is added to a final concentration of 100 μg/mL, and plates are poured).

2.7. Ligation

1. Ligation buffer.
2. Ligase (Roche Molecular Biochemicals).
3. 10 m*M* NaCl.
4. QIAquick PCR purification kit.

2.8. Checking the Average Library Insert Size by Colony PCR

1. 10× PCR buffer (Qiagen).
2. dNTP mixture (10 m*M* each dNTP) (Qiagen).
3. *Taq* DNA polymerase 5 U/μL (Qiagen).
4. 10 μ*M* M13/pUC sequencing (–40) primer (New England Biolabs).
5. 10 μ*M* M13/pUC reverse sequencing (–24) primer (New England Biolabs).
6. 250 μL PCR tubes or 8-strips (MJ Research).

3. Methods

3.1. Nuclear DNA Preparation

Plastids are very abundant, not only in green tissues, and their DNA is unmethylated. Thus, if chloroplast DNA is present in a DNA sample, it will be selected during the filtering process. For this reason, it is important to purify nuclei from the rest of the cell organelles before purifying the genomic DNA. The protocol used here is a modification of those reported by Kiss et al. and Wagner et al. ***(19,20)***.

1. In a cold room, prepare a ring stand with two funnels attached with clamps, one on top of the other, so that the top funnel drains inside the bottom one. Cover the upper funnel with four 30 × 30 cm layers of cheese cloth and the lower one with one 30 × 30 cm layer of 60-μm nylon mesh. Put a 500-mL centrifuge bottle under the lower funnel to collect the liquid.
2. Grind 50–100 g of frozen tissue in liquid N_2 (*see* **Note 2**).
3. Transfer to a blender containing 6–8 vol of IB 1.
4. Homogenize 3× at maximum speed for 10 s each time.
5. Transfer to a plastic beaker and further homogenize 3× with a polytron, 5 s each time (*see* **Note 3**).
6. Slowly pour the slurry into the top funnel.
7. When it stops dripping, squeeze the liquid out of the cheese cloth using gloves.

8. Centrifuge at 2000*g* for 15 min at 4°C.
9. Carefully discard the supernatant and resuspend the nuclear pellet in 0.1–0.5 vol of IB 1.
10. Transfer to 14- or 50-mL centrifuge tubes and centrifuge at 2000*g* for 15 min at 4°C.
11. Resuspend in 5–20 mL of IB 2.
12. Add one-fifth vol of 5% Sarkosyl.
13. Mix gently and incubate 15 min at room temperature.
14. Add one-seventh vol of 5 *M* NaCl and mix gently.
15. Add one-tenth vol of CTAB solution preheated to 60°C.
16. Mix gently and incubate for 30 min at 60°C, mixing by inversion every 2–4 min.
17. Add 1 vol of chloroform:octanol and mix well by inversion (do not vortex mix).
18. Centrifuge at 6000*g* for 15 min at 4°C.
19. Transfer upper phase to a new centrifuge tube.
20. Add two-thirds vol of isopropanol and mix slowly by inversion.
21. Hook the DNA with a glass rod bent in the tip to help preventing the DNA from falling off (*see* **Note 4**).
22. Wash the nuclear DNA by immersing the glass rod in 70% ethanol.
23. Air-dry the DNA for a few minutes.
24. Immerse the DNA in 0.5–1 mL 10 m*M* Tris-HCl, pH 8.0, and shake it quickly until it falls off the glass rod.
25. Let the DNA resuspend overnight at 4°C.

3.2. DNA Shearing and End-Repairing

1. In a 14-mL Falcon centrifuge tube, mix 20 µg of nuclear DNA with 1 mL of 50% glycerol and 0.2 mL of nebulization buffer. Add water up to a final vol of 2 mL.
2. Seal the bottom nebulizer inlet with parafilm.
3. Remove the nebulizer screw-cap and transfer the DNA mixture to the bottom of the nebulizer.
4. Put the nebulizer cap and attach N_2 gas tubing in the bottom inlet. Close the upper nebulizer outlet with the Falcon tube cap.
5. While holding the cap, apply N_2 gas at 8–10 psi for 2 min (*see* **Note 5**).
6. Remove the tubing and spin down the nebulizer 1 min at 1500*g* (*see* **Note 6**).
7. Precipitate the DNA with one-fiftieth vol of 5 *M* NaCl and 2 vol of ethanol.
8. Keep at –20°C overnight.
9. Centrifuge at 12,000*g* for 30 min at 4°C.
10. Add 3 mL of 70% ethanol and centrifuge at 12,000*g* for 10 min at 4°C.
11. Dry in speedVac (*see* **Note 7**) and resuspend in the necessary vol of 5 m*M* Tris-HCl, pH 8.0, to reach a final vol of 100 µL after adding the reagents of the next step.
12. Transfer to a 1.5-mL tube and add 10 µL of dNTPs (0.5 m*M* each), 20 U T4 DNA polymerase, and 10 µL T4 DNA polymerase buffer.
13. Incubate 15 min at 30°C.
14. Add 6 U Klenow enzyme.

15. Incubate 15 min at 30°C.
16. Clean up through a QIAquick column (*see* **Note 8**).
17. Elute with 50 µL of 10 m*M* Tris-HCl, pH 8.0 (EB buffer; Qiagen).
18. Collecting the liquid in the same tube, re-elute with the necessary vol of water to reach a final vol of 100 µL after adding the reagents of the next step.
19. Add 5 U T4 PNK, 10 µL T4 PNK buffer, and 2 µL ATP 100 m*M*.
20. Incubate 30 min at 37°C.
21. Add 100 µL of water and extract with 200 µL of phenol:chloroform by vortex mixing and centrifuging at 12,000*g*.
22. Transfer the upper phase to a new tube and extract with 200 µL of chloroform by vortex mixing and centrifuging at 12,000*g*.
23. Transfer the upper phase to a new tube and precipitate with one-fiftieth vol of 5 *M* NaCl and 2 vol of ethanol.
24. Leave at –20°C overnight.
25. Centrifuge at 12,000*g* for 30 min at 4°C.
26. Add 400 µL of 70% ethanol and centrifuge at 12,000*g* for 10 min at 4°C.
27. Dry and resuspend in 20 µL of 10 m*M* Tris-HCl, pH 8.0.

3.3. Adaptor Ligation

1. In a 1.5-mL tube, mix 10 µL of top adaptor oligonucleotide and 10 µL of bottom adaptor oligonucleotide (*see* **Note 9**).
2. Add 0.5 µL of 1 *M* NaCl.
3. Incubate 2 min at 75°C and anneal for at least 2 h by cooling down very slowly to 4°C.
4. In a new 1.5-mL tube, mix 10 µL of end-repaired DNA, 20 µL of annealed adaptor, 4 µL of T4 DNA ligase buffer, 10 U of T4 DNA ligase, and water to a final vol of 40 µL.
5. Incubate 24 h at 12°C (*see* **Note 10**).
6. Add 60 µL of TEN buffer (*see* **Note 11**).
7. Place the size fractionation column in a support and remove first the top and then the bottom cap (*see* **Note 12**).
8. Drain the liquid by gravity.
9. Wash the column by adding 800 µL of TEN buffer and allowing to drain completely.
10. Repeat the wash three more times.
11. Label 20 1.5-mL tubes and align them in a rack.
12. Add the adapted DNA to the upper frit of the column and allow to drain completely into the first 1.5-mL tube.
13. Add 100 µL of TEN buffer and collect the effluent in the second tube.
14. Add another 100 µL of TEN buffer and begin to collect a single drop per tube until complete drain.
15. Repeat the last step until 18 drops have been collected.
16. Run 3 µL of each fraction in an agarose gel.
17. Pool the first three fractions where DNA can be detected in the gel (*see* **Note 13**).

3.4. Vector Preparation

1. In a 1.5-mL tube, mix 2 µg of pUC 19 DNA, 30 U of *Xba*I, 6 µL of buffer H, and water up to 60 µL (*see* **Note 14**).
2. Incubate 2 h at 37°C.
3. Inactivate the enzyme incubating 20 min at 65°C.
4. Chill on ice and add 4 µL of buffer L, 2 µL of 10 mg/mL BSA, 4 µL of 1 m*M* dCTP, 8 U of Klenow enzyme, and water up to a final vol of 100 µL.
5. Incubate 30 min at 30°C.
6. Inactivate the enzyme incubating 15 min at 65°C.
7. Clean up the DNA through a QIAquick column.
8. Elute with 50 µL of 10 m*M* Tris-HCl, pH 8.0.
9. Re-elute in the same tube with 39 µL of water.
10. Add 10 µL of CIP buffer and 1 µL of 2 U/µL CIP.
11. Incubate 30 min at 37°C.
12. Add 2 µL 0.5 *M* EDTA and incubate 15 min at 65°C.
13. Add 100 µL water.
14. Extract with 200 µL of phenol:chloroform.
15. Extract with 200 µL of chloroform.
16. Precipitate with one-fiftieth vol of 5 *M* NaCl and 2 vol of ethanol.
17. Leave overnight at –20°C.
18. Centrifuge at 12,000*g* for 30 min at 4°C.
19. Add 500 µL of 70% ethanol and centrifuge at 12,000*g* for 10 min at 4°C.
20. Dry and resuspend in 100 µL of 10 m*M* Tris-HCl, pH 8.0 (*see* **Note 15**).

3.5. Preparation of Electrocompetent JM107 or JM107MA2 Cells

This protocol was modified from the manual by Sambrook and Russell ***(21)*** (*see* **Note 16**).

1. Use one JM107 or JM107MA2 colony from a fresh plate to inoculate 3 mL of LB medium. Incubate at 37°C overnight with shaking.
2. Take 2 mL of the overnight culture to inoculate 500 mL of SOB medium without magnesium. Incubate at 37°C shaking at 250–300 rpm until reaching an OD_{550} of 0.6–0.7.
3. Chill the culture on ice for 20 min and transfer to two 250-mL centrifuge bottles. Centrifuge at 2500*g* at 4°C for 15 min.
4. Repeat the wash in 10% glycerol. Discard the supernatant and resuspend each pellet in 10 mL of chilled 10% glycerol.
5. Transfer to two 14-mL centrifuge tubes.
6. Centrifuge at 2500*g* at 4°C for 15 min.
7. Resuspend both pellets in a total of 2 mL of chilled 10% glycerol.
8. Transfer 100 to 200-µL aliquots of the cells suspension to chilled sterile 1.5-mL microcentrifuge tubes. Freeze the cells in liquid N_2 and store at –70°C (*see* **Note 17**).

3.6. Ligation

1. In a 1.5-mL tube, mix 5–10 ng of vector, 10–100 ng of adapted and size fractionated genomic DNA (**step 17** from **Subheading 3.3.**), 1 µL of ligation buffer, 1 U of ligase, and take to a final vol of 10 µL with water.
2. Incubate 16 h at 12°C.
3. Add 90 µL of 10 m*M* NaCl.
4. Clean up the reaction using a QIAquick column, eluting in 50 µL of 10 m*M* Tris-HCl, pH 8.0.

3.7. Electroporation

1. Thaw electrocompetent cells in ice.
2. Mix 30 µL of cells with 1–3 µL of cleaned up ligation reaction in a chilled 1.5-mL tube.
3. Transfer the mixture to a chilled 0.1-cm gap electroporation cuvette and electroporate at 1.8 kV. Immediately add 750 µL of SOC medium and transfer to a sterile 14-mL centrifuge tube.
4. Incubate cells at 37°C for 45 min with gentle shaking.
5. Plate aliquots of approx 200 µL of cells together with 50 µL IPTG and 50 µL X-gal in LB-ampicillin plates.
6. Incubate overnight at 37°C.

3.8. Checking the Average Library Insert Size by Colony PCR

1. In a 1.5-mL tube, mix 60 µL of 10× PCR buffer, 30 µL of 10 µ*M* M13/pUC sequencing (–40) primer, 30 µL of 10 µ*M* M13/pUC reverse sequencing (–24) primer, 12 µL of dNTP mixture, 6 µL of 5 U/µL *Taq* DNA polymerase, and 462 µL of water (*see* **Note 18**).
2. Transfer 20 µL of the mixture to each of 30 250-µL PCR tubes.
3. Using an automatic pipet set in 5 µL, pick one white colony into the first PCR tube and pipet up and down a few times.
4. Repeat the last step for the rest of the tubes using a new tip each time.
5. Put the tubes in a PCR machine under the following program: 5 min at 95°C, then 25 cycles of: 30 s at 95°C, 45 s at 55°C, 3 min 30 s at 72°C, 10 min at 72°C, then forever at 4°C.
6. Run 10 µL of each reaction in an agarose gel.
7. Estimate the average insert size taking into account that the PCR fragments include 30–60 bp of vector sequence in each end. The proportion of clones containing repetitive DNA can be estimated as well (*see* **Note 19**).

4. Notes

1. For all buffers and solutions all Milli-Q® water (Millipore) is used.
2. When possible, it is preferable to use a tissue with low plastid content (i.e., maize immature ears). This would reduce the chloroplast DNA contamination. If the methylation status of a certain kind of gene is known to change with development, it should be taken into account at the moment of choosing the tissue for preparing DNA.

3. The use of a Polytron can be omitted if the blender properly homogenizes the tissue. In the case of hard tissue like pine needles, the Polytron may be necessary.
4. If the amount of starting material is small, DNA fibers may not be formed after adding isopropanol. In this case, the DNA can be recovered by centrifugation at 12,000*g* for 30 min.
5. The nebulization time and pressure need to be calibrated. Aliquots of DNA can be taken at different nebulization times and checked in agarose gels. The optimal nebulization conditions should break down the DNA to fragments mainly between 1 and 4 kbp.
6. As nebulizers are not designed for centrifugation, a rotor must be adapted to hold them. For example, the Sorvall® GSA rotor (NEN® Life Science Products) can be used if the bottoms of the wells are cushioned with paper towels.
7. The pellet is often loose and hard to see. It is advisable not to remove all the 70% ethanol and dry it for a longer time in the SpeedVac.
8. If a phenol extraction followed by ethanol precipitation is performed instead of the column clean up, a very hard to dissolve pellet is formed.
9. After annealed, the adaptor looks like this:

```
5'(P)-TAGACGCCTCGAG-3'
      | | | | | | | | | |
   3'-TGCGGAGCTC-5'
```

10. The 3-nucleotide overhang adaptor works very well. However, if necessary, cloning efficiency can be improved by using a double adaptor method ***(22)***.
11. Instead of using a column, the DNA can be size-fractionated by agarose gel electrophoresis. In this case, fragments ranging from 1–4 kbp must be eluted from the gel. One disadvantage of this approach is that a melting step needs to be performed by heating, which may denature the adaptor whose shorter oligonucleotide is not covalently linked. Using high quality low melting point agarose like SeaPlaque GTG agarose (BioWhittaker Molecular Applications) and the QIAquick gel extraction kit allows to melt the agarose at room temperature, which helps to overcome the problem. Alternatively, the shorter oligonucleotide can be added to the vector ligation reaction to improve the ligation efficiency.
12. To avoid the formation of bubbles inside the column, it is advisable to use a needle to make a hole in the top cap before removing it.
13. Taking the first 3 to 4 fractions in which DNA can be observed in the agarose gel usually works well. The next fractions may contain unligated adaptors and small DNA fragments, although they are not visible in the sample loaded in the gel. If no or few small insert clones are detected after estimating the library insert size (*see* **Subheading 3.8.**), the inclusion of more elution fractions can be considered for future construction of filtered libraries.
14. pUC 19 and *Xba*I are used as an example. Other vectors and restriction enzymes can be used as well. However, the protocols must be adapted accordingly in terms of selective antibiotic, adaptor sequence, host strain requirements, etc.
15. Before using a vector for library construction, some controls must be performed

by *E. coli* transformation: *(i)* vector with no ligase; *(ii)* self-ligated vector; and *(iii)* vector ligated to a control insert. The first two controls should yield no or very few blue colonies only. The third one should yield no or very few blue colonies and a large number of white colonies. In this case, the control insert is made by annealing the longer oligonucleotide used to make the adaptor and another 13-mer oligonucleotide: 5'(P)-TAGCTCGAGGCGT-3'. When annealed it looks like this:

```
5'(P)-TAGACGCCTCGAG-3'
      | | | | | | | | |
3'-TGCGGAGCTCGAT-(P)5'
```

16. JM107 ***(23)*** and JM107MA2 ***(24)*** are shown as examples of filtering and unfiltering strains, respectively. Other strains can be used, e.g., DH5α-E (*mcrBC*$^+$) and DH10B (*mcrBC*$^-$), both of which are available as electrocompetent from Invitrogen. If commercial strains are used, the protocols should be adapted to any special requirements of a particular *E. coli* strain. However, among *mcrBC*$^+$ strains, variations in filtering efficiency has been observed ***(14)***. Thus, both the transformation and filtering efficiencies need to be considered when choosing the strain to approach a large-scale methylation filtration project.
17. After a batch of competent cells is prepared, it must be tested by transforming a known amount of supercoiled plasmid. Usually the transformation efficiency is $>1 \times 10^{10}$ colonies/μg of plasmid DNA. Also, cells must be tested for any plasmid contamination by doing an electroporation without DNA, which should yield no colonies in selective medium.
18. The amount of PCR mixture can be increased to compensate for pipeting errors and to include some useful PCR controls like a blue colony, vector DNA, a water control, single primer controls, etc. This is a robust PCR assay and any commercially available PCR reagents should work as well as any combination of M13 forward and reverse primers. Instead of using PCR, insert sizes can be checked by doing plasmid minipreps of white colonies and subsequent restriction enzyme digestion and agarose gel electrophoresis.
19. An easy way to estimate the number of clones containing repetitive DNA is to bind a number of clones to a hybridization membrane and hybridize it against total labeled genomic DNA. In this labeled sample, only the repetitive DNA will be present in high enough proportion to produce a hybridization signal. Low copy DNA will be too diluted to show any hybridization. In this way, the high copy DNA containing clones can be identified as hybridizing clones. The proportion of high vs low copy clones can be compared to that in a control unfiltered library to estimate the filtering efficiency of the cloning process. The unfiltered library is constructed simply by transforming the same ligation mixture used for the filtered library into a *mcrBC*$^-$ *E. coli* strain. The hybridization can be performed on one to a few hundred clones from each library by colony hybridization ***(21)***. For example, for maize, where 80–90% of the genome is composed of repetitive DNA, a 5- to 10-fold decrease in the proportion of repetitive clones is expected in

a filtered vs a control library. There may be some variations due to the frequent methylcytosine to thymine transition. This mutation occurs frequently in silent repetitive DNA that is not under selective pressure. For this reason, some decayed repeats can be recovered in filtered libraries. Sequencing and Basic Local Alignment Search Tool (BLAST) analysis ***(25)*** of a few hundred clones from each library is an independent way to estimate how well the technique is working.

References

1. Goffeau, A., Barrell, B. G., Bussey, H., et al. (1996) Life with 6000 genes. *Science* **274,** 546–567.
2. The *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018.
3. The *Arabidopsis* Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815.
4. Blattner, F. R. (1983) Biological frontiers. *Science* **222,** 719–720.
5. Putney, S. D., Herlihy, W. C., and Schimmel, P. (1983) A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302,** 718–721.
6. Adams, M. D., Kelley, J. M., Gocayne, J. D., et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–1656.
7. Bento Soares, M. and Bonaldo, M. F. (1998) Constructing and screening normalized cDNA libraries, in *Genome Analysis. A Laboratory Manual. Vol. 2. Detecting Genes.* (Birren, B., Green, E. D., Klapholz, S., Myers, R. M., and Roskams, J., eds.), CSH Laboratory Press, Cold Spring Harbor, NY, pp. 49–158.
8. Barakat, A., Matassi, G., and Bernardi, G. (1998) Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. *Proc. Natl. Acad. Sci. USA* **95,** 10044–10049.
9. Chandler, V. L. and Hardeman, K. J. (1992) The *Mu* elements of *Zea mays*. *Adv. Genet.* **30,** 77–122.
10. Raizada, M. N., Nan, G. L., and Walbot, V. (2001) Somatic and germinal mobility of the *RescueMu* transposon in transgenic maize. *Plant Cell* **13,** 1587–1608.
11. Raleigh, E. A. and Wilson, G. (1986) *Escherichia coli* K-12 restricts DNA containing 5-methylcytosine. *Proc. Natl. Acad. Sci. USA* **83,** 9070–9074.
12. Dila, D., Sutherland, E., Moran, L., Slatko, B., and Raleigh, E. A. (1990) Genetic and sequence organization of the *mcrBC* locus of *Escherichia coli* K-12. *J. Bacteriol.* **172,** 4888–4900.
13. Sutherland, E., Coe, L., and Raleigh, E. A. (1992) McrBC: a multisubunit GTP-dependent restriction endonuclease. *J. Mol. Biol.* **225,** 327–348.
14. Rabinowicz, P. D., Schutz, K., Dedhia, N., et al. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* **23,** 305–308.

15. Walker, E. L. and Panavas, T. (2001) Structural features and methylation patterns associated with paramutation at the *r1* locus of *Zea mays*. *Genetics* **159,** 1201–1215.
16. Walbot, V. and Warren, C. (1990) DNA methylation in the *Alcohol dehydrogenase-1* gene of maize. *Plant Mol. Biol.* **15,** 121–125.
17. Patterson, G. I., Thorpe, C. J., and Chandler, V. L. (1993) Paramutation, an allelic interaction, is associated with a stable and heritable reduction of transcription of the maize *b* regulatory gene. *Genetics* **135,** 881–894.
18. Povinelli, C. M. and Gibbs R. A. (1993) Large-scale sequencing library production: an adaptor-based strategy. *Anal. Biochem.* **210,** 16–26.
19. Kiss, T., Toth, M., and Solymosy, F. (1985) Plant small nuclear RNAs. Nucleolar U3 snRNA is present in plants: partial characterization. *Eur. J. Biochem.* **152,** 259–266.
20. Wagner, D. B., Furnier, G. R., Saghai-Maroof, M. A., Williams, S. M., Dancik, B. P., and Allard, R.W. (1987) Chloroplast DNA polymorphisms in lodgepole and jack pines and their hybrids. *Proc. Natl. Acad. Sci. USA* **84,** 2097–2100.
21. Sambrook, J. and Russell, D. W. (eds.) (2001) *Molecular Cloning. A Laboratory Manual*. CSH Laboratory Press, Cold Spring Harbor, NY.
22. 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.
23. 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**, 103–119.
24. Blumenthal, R. M., Gregory, S. A., and Cooperider, J. S. (1985) Cloning of a restriction-modification system from *Proteus vulgaris* and its use in analyzing a methylase-sensitive phenotype in *Escherichia coli*. *J. Bacteriol.* **164**, 501–509.
25. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.