© 2002 Oxford University Press Nucleic Acids Research, 2002, Vol. 30 No. 21 **e118**

BACFinder: genomic localisation of large insert genomic clones based on restriction fingerprinting

Mark L. Crowe, Debashis Rana, Fiona Fraser, Ian Bancroft and Martin Trick*

John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK

Received May 31, 2002; Revised and Accepted August 20, 2002

ABSTRACT

We have developed software that allows the prediction of the genomic location of a bacterial artificial chromosome (BAC) clone, or other large genomic clone, based on a simple restriction digest of the BAC. The mapping is performed by comparing the experimentally derived restriction digest of the BAC DNA with a virtual restriction digest of the whole genome sequence. Our trials indicate that this program identified the genomic regions represented by BAC clones with a degree of accuracy comparable to that of end-sequencing, but at considerably less cost. Although the program has been developed principally for use with *Arabidopsis* BACs, it should align large insert genomic clones to any fully sequenced genome.

INTRODUCTION

Bacterial artificial chromosomes (BACs) are widely used for cloning large fragments of DNA from a variety of sources (1–3). Typically, BACs might be selected for further experimentation from a library in one of two ways: by PCR or hybridisation screening (4,5) or by their ability to complement a specific mutation (6). Once a BAC of interest has been identified, sequencing or other analysis may be carried out to determine the genes contained in the BAC insert. In both cases, significant amounts of work are required to select the clones. Nor is either technique scalable; almost as much work is required to find BACs complementing further genes or hybridising to different probes as is needed the first time. Therefore, neither technique is well suited to predicting the genomic loci or gene content of large numbers of BACs.

An alternative approach to identifying a BAC of interest is that of end-sequencing, where sequence data from the ends of the BAC insert are used to locate the clone within the genome. This is a powerful and efficient technique and can process many BACs relatively quickly. However, although sequencing is becoming cheaper and quicker, a large scale end-sequencing project is still a major undertaking.

We are participants in a functional genomics programme studying the model plant *Arabidopsis thaliana*. One of our roles is to screen BAC libraries to identify clones containing defined sets of genes for functional testing (http://www.york.ac.uk/res/garnet/bancroft.htm). In contrast to the libraries used for genome sequencing, these libraries were constructed using specialised vectors that contain the *cis*-acting sequences necessary for transfer of the clone inserts into the genomes of plants using *Agrobacterium*-mediated transformation. The average clone insert size in the libraries most extensively used is ~80 kb, so libraries of ~13 000 clones are required to provide ~8-fold redundant coverage of the 130 Mb genome of *A.thaliana*. To improve the efficiency of the process, we aimed to develop laboratory and computational methods to permit the systematic alignment of many thousands of clones to the genome of *A.thaliana*, which is almost fully sequenced (7). We achieved this by adapting standard clone fingerprinting methods (8) and developing BACFinder, a software application that uses a ‘virtual digest’ of a complete genome sequence to predict the location of any BAC clone based on alignment of its restriction digest pattern against the predicted digest pattern produced by the genome sequence. Although designed and tested for use with *A.thaliana*, the method is applicable to the many species of plants, animals and (particularly) microorganisms with fully, or almost fully, sequenced genomes.

MATERIALS AND METHODS

Chromosome sequences

We used the chromosome sequences made available by TIGR on 7 January 2002 (available from ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/) for all results described in this paper.

BAC DNA isolation and restriction digestion

We carried out DNA extraction and fingerprinting essentially as described in Marra *et al*. (8). The DNA in each well was digested with restriction enzyme in 15 µl reaction volume for 2 h. Then 2 µl of 6× dye was added to the digest. Digest volume was reduced to 10 µl by centrifuging the plates without the lids at 1550 *g* for 20 min.

Gel preparation and loading

Agarose (SeaKem LE, FMC Bioproducts, Rockland, ME, US) gels (1%) were run in Gator Wide Format System model A3-1 (Owl Scientific, US). Special 121-well gel combs were made locally by following the Sanger Centre design. Marker DNA (Analytical Marker DNA, Wide Range, catalogue no. DG1931) was loaded every fifth lane starting from the first lane. Two microlitres of restriction enzyme-digested BAC

*To whom correspondence should be addressed: Tel: +44 1603 450557; Fax: +44 1603 450045; Email: martin.trick@bbsrc.ac.uk

DNA was loaded in each well. The gels were run in 1× TAE in a cold room for 16–18 h. Gels were stained in 250 ml of 20 mM Tris and 0.1 mM EDTA containing 25 µl Vistra Green (Amersham Life Sciences, UK) for at least 1 h. Gels were then scanned in a Molecular Dynamics FluorImager 595 scanner.

Analysis of gel images and contig assembly

Gel images were imported into Image v3.11 developed at the Sanger Centre (http://www.sanger.ac.uk/software/Image). Edited bands were then transferred to FPC v4.7, also developed at the Sanger Centre (http://www.sanger.ac.uk/Software/fpc). Where they could be conclusively identified the vector bands were removed during the transfer process. Upon transfer, a file containing the size data of individual bands of each BAC was generated. This file was then used as input for the BACFinder program.

In silico chromosome digest

BACFinder works by identifying a region or regions of the genome that are predicted to generate the same restriction pattern as that observed for the BAC being mapped. Therefore the initial step is the generation of a virtual restriction map of the genome. ChrDigest, an accompanying program to BACFinder, performs this function. ChrDigest processes the chromosomes in a linear fashion, identifying each restriction site in turn. The length of the fragment from the previous restriction site is calculated, and these predicted fragment lengths, together with their start and stop coordinates, are stored in an array in the order that they occur in the chromosome. These chromosome models are then combined to produce an overall genome model, which can be used by BACFinder. Fragments containing regions of unknown sequence (defined as ⩾20 consecutive Ns) are considered to have a possibly inaccurate or arbitrary length; such fragments are assigned a null length and will never match.

The settings of ChrDigest can be easily modified to use different restriction enzymes and other source genomes. Also the length of unknown sequence allowed before a fragment is classed as null can be changed.

BAC mapping

BACFinder reads in the genome model generated by ChrDigest and the list of bands identified from each BAC, one BAC at a time. Matches are identified in a two-step process. The first step is the identification of a region of the chromosome where all the predicted restriction sites are either the same size (to within the accuracy allowed) as a band observed on the gel or are outside the range of accurate size definition of the gel. Once such a region has been identified, BACFinder then tests whether all the bands observed on the gel can be accounted for by predicted fragments in that genomic region. If both tests are successful, the smallest number of consecutive fragments that can account for all the observed bands is identified, to identify the minimum as well as the maximum possible matched region. The match is then recorded and the rest of the genome is processed.

If mismatches are allowed, and the initial attempt to match the BAC fails, then the process will be repeated; this time, however, the second step will be modified in that one of the observed bands need not match. If this too fails, then further attempts will be made, each time increasing the number of mismatches allowed until the preset limit is reached; in practice, a maximum of two mismatches seems sufficient, and more would probably be undesirable due to a resulting increase in the risk of false matches.

Once either match is found or the mismatch limit is reached, the result is output to either the screen or a text file, and the next BAC is processed.

Sequencing

We grew BAC clones as 100 ml cultures (LB + 25 µg/ml kanamycin) and isolated BAC DNA using the Qiagen Maxi 500 kit. Plasmid yield was quantified by comparing 1 µl of the preparation against a standard (200 ng pGEM vector) on a 1% agarose, 1× TBE gel. We used Terminator Cycle Sequencing with an ABI 3700 sequencer to sequence the BACs, using the Big Dye Version 3 kit (Applied Biosystems) to set up 20 µl sequencing reactions [8 µl Big Dye mix, 2 µl primer (5 pmol/µl), 5 µl BAC DNA (~1200 ng), 5 µl water]. Reaction conditions were an initial 5 min at 95°C, then 45 cycles of 95°C for 30 s, 55°C for 20 s and 60°C for 4 min, holding at 4°C on completion. The sequencing primers used were based on the sequence of the pYLTAC17 vector, from which the BAC library was constructed, and were derived from Liu *et al.* (9): (i) forward 5′-CTAGATCATGATCGGTACCTTTG-3′; (ii) reverse 5′-GTTCATGTCTCCTTCTGTATGTAC-3′.

We removed vector sequence from the end-sequence data, and used WU-BLASTN with default parameters (W. Gish, http://blast.wustl.edu) to align the remaining (insert) sequence against the complete genome sequence model for *A.thaliana*. We defined the end position of the BAC insert as the chromosomal location of the first matching base of the insert sequence.

RESULTS

Training BACFinder

We performed *Hin*dIII digests on 81 BACs and attempted to map them against the *Arabidopsis* genome using a developmental version of BACFinder. Based on this initial trial, we selected 24 BACs, which had given a range of results from 0 to 30 predicted loci, for end-sequencing. We used BLAST alignment of the end-sequences against the genome sequence to identify the correct locus of each BAC, and in this way found the loci of 18 BACs, although in one case only one end was matched. The remaining six BACs either failed to produce sufficient quality sequence data for a locus to be identified or they did not match the published genome sequence.

Improvements were made to the program and parameters set according to a comparison of the predicted against the sequence-determined loci.

Band size accuracy

From observation of the agarose gels, we estimated that band sizes of between 1 and 12 kb could be determined accurately using the gel system available. We verified this by comparing the band size predictions given by the image analysis software with the actual band sizes, determined from the genomic sequence of the BAC locus as identified by end-sequence data. This confirmed that band size calling was accurate to within

4% in this 1–12 kb range. Relatively few BACs generated bands of >12 kb, while smaller bands could not be reliably sized due to the resolution of the gel and the intensity of staining. We therefore decided to discount bands outside the 1–12 kb size range from both the observed data and the predicted genomic digest pattern when performing the locus predictions.

Mismatch allowance

Three of the BACs were found to produce extra bands on the gel that did not correspond to fragments predicted from the genome sequence. It is not clear at this stage whether these bands were artefacts, perhaps resulting from an incomplete digest of the BAC or chimaeric BACs, or whether they are genuine bands that are not predicted due to errors in the genome sequence. We have introduced the capacity to allow for these extra bands, but at the cost of a higher risk of generating a false match. If mismatches are required to find a matching genomic locus, the output from BACFinder will indicate how many mismatches were needed for that prediction.

No examples were found where there were predicted genomic bands that were not observed on the gel. Since clones that are missing predicted genomic bands are probably the result of deletions (and should be excluded from further studies), and because of the increased risk of producing false matches, we chose not to introduce measures to allow for this case.

If a BAC contains two or more bands that differ in size by only a few percent, it is unlikely that they could be distinguished as separate on an agarose gel. Hence, it is a fundamental part of the program that a single gel band can match many genomic fragments. However, the reverse is not true: no fragment can match more than one band, since one genomic fragment could not produce two bands on a gel.

Minimum band requirement

In the initial trials, several hundred possible loci had been predicted for some BACs. All of these BACs had given very few bands on the gel, and in some cases none at all, within the allowable size range of 1–12 kb. We therefore introduced a requirement for a minimum number of bands before a valid match can be obtained, which prevents multiple false matches by such BACs. The choice of the cut-off values for the minimum band number is described below. If mismatches are allowed, the probability of mislocating a BAC is increased, both as a consequence of having fewer bands used for the prediction and because rather than a single band pattern, there are many possible combinations of bands allowed (for example with 10 bands and two mismatches, there are 56 possible band combinations). Therefore the minimum number of bands required is incremented as more mismatches are permitted.

End prediction

BACFinder allows observed bands to match multiple predicted fragments, and for fragments outside the valid size range to match by default. Consequently, the predicted end coordinates (the chromosomal positions of the ends of the insert) of the BAC are sometimes a significant distance beyond the true ends of the BAC. Conversely, a minimal match, taking the shortest run of consecutive fragments required to match all bands, may predict end coordinates well within the BAC. In view of this, we have set up BACFinder to use both approaches and produce a range within which the end coordinates of the BAC are predicted, rather than a single value. In the best case, this range may be a single base pair, in which case the maximum and minimum extents of the predicted BAC location are the same; conversely, the range may be several tens of kilobases in size. The latter case might occur if, for example, the minimal prediction were flanked by a very large restriction fragment, outside the range of resolution of the gel. Since it is not possible to distinguish from the gel whether the band is present or not, it is also impossible to definitively include or exclude it from the prediction. These occasional large ranges of prediction are therefore unavoidable.

Table 1. The proportion of random artificial digest patterns with different numbers of bands that falsely map to a genomic locus

Number of bands	% predicting locus	
	0 mismatches	2 mismatches
4	77.66	n.d.
5	34.09	100
6	11.32	99.79
7	3.51	96.1
8	1.59	78.54
9	0.58	49.96
10	0.15	27.42
11	0.22	14.15
12	0	8.12
13	n.d.	4.67
14	n.d.	2.66
15	n.d.	1.81

This demonstrates the increased reliability of prediction with increasing band numbers. n.d., not done.

Testing for false matches

To measure the probability of incorrectly locating BACs, we generated test sets of 10 000 'pseudo-BACs' consisting of fragments selected at random from all of those from the full predicted digest that were within the range of valid sizes, i.e. 1–12 kb. All members of each test set had a specific number of 'bands' (from 4 to 15). None of these pseudo-BACs should generate a real match, since they should not correspond to any real region of the genome. Any predicted matches would therefore be a result of chance, and the frequency of these matches would give an indication of the likelihood of false matches being obtained with real data. The results of these trials are shown in Table 1; in summary, under the conditions used to map the real BACs, there is a probability of <5% that a false match would be obtained from a BAC with seven or more bands with no mismatches, or from a BAC with 13 or more bands with two mismatches. As the number of bands increases, so the likelihood of a false match falls further.

As a further indication of the low probability of generating false matches, no matches were found for any of 171 *Brassica rapa* BAC fingerprints tested against the *Arabidopsis* genome.

Success of locus prediction

Based on the training results for BACFinder, the following settings were used as providing the optimum accuracy of match for the most BACs: (i) band size accuracy, 4%; (ii) band

Table 2. Comparison of the BACFinder predictions against the exact end coordinates as established from end-sequence data

BACFinder prediction			Sequence results		
Chr[a]	Start	Stop	Chr[a]	Start	Stop
2	9899119 ± 0	9986278 ± 5161	2	9899119	9981117
1	8053033 ± 4336	8153675 ± 179	1	8048697	8153854
1	761702 ± 2558	856414 ± 368	1	763713	856405
3	11037406 ± 1864	11106256 ± 5970	3	11035543	11100286
5	7836419 ± 7852	7909825 ± 9061	5	7842030	7916622
2	14735721 ± 0	14801917 ± 0	2	14735721	14801917
3	20141369 ± 66	20172986 ± 1122	3	20141435	20161865
2	14249690 ± 5289	14341659 ± 3212	2	14248001	14344259
5	15753958 ± 110	15857598 ± 9683	n.d.	n.d.	n.d.
2	15116675 ± 9116	15196563 ± 671	2	15125469	15195892
1	8731444 ± 178	8762217 ± 594	1	8731622	8762049
5	15217182 ± 2084	15284457 ± 0	5	n.d.	15284457
5	24925300 ± 8391	25044481 ± 17205	5	24924067	25027992
2	9416929 ± 8788	9507503 ± 12866	2	9425717	n.d.
4	5366000 ± 7358	5435340 ± 0	4	5373358	5435340
4	10249271 ± 434	10317398 ± 219	4	10248838	10317617
3	20025431 ± 1209	20106104 ± 652	3	20026640	20105453
5	25362196 ± 4146	25434953 ± 3487	5	25365968	25431846
2	623502 ± 106	694817 ± 1326	2	623608	693492
5	898875 ± 4525	966771 ± 9030	**2**	**2642547**	**2710725**

Values highlighted in bold indicate those incorrectly positioned by BACFinder. n.d., not determined.

[a]Chr, chromosome number.

size range, 1–12 kb; (iii) maximum mismatches, 2; (iv) minimum band number, 7; (v) minimum band increment, +3 per mismatch.

Using these settings, we re-analysed all 81 BACs. BACFinder predicted single loci for 60 (~74%). The majority of the remaining BACs (15/22) were rejected as having too few bands for a valid prediction. No loci could be found for the remaining six. In no case were multiple loci predicted, which suggests that BACFinder is robust against misidentifying regions of duplicated sequence.

We selected 20 further BACs, from the initial 81, which had not been used in the training of the program and for which BACFinder had predicted a single locus. The ranges of the predicted values for the end coordinates of these 20 BACs varied from 0 to 34 410 bp, with a mean of 7460 bp. We were able to identify complete genomic loci for 17 of these BACs from end-sequence data, and the location of one end for two more. Poor sequence data prevented identification of loci for the remainder. Of the 36 BAC end coordinates thus identified by sequencing, 34 were within the range predicted by BACFinder (Table 2). The two failures were due to the locus of a single BAC being misidentified by BACFinder; this was the single case out of the 20 BACs tested where two mismatched bands had been required to find a match, so this probably reflects the known risk of mislocation when mismatches are allowed. This was accentuated by the restriction digest for this BAC having the minimum number of bands necessary to pass the cut-off point for a valid prediction.

Further analysis of this mislocated BAC showed that it, in fact, had only a single observed band that did not correspond to a genomic fragment, but critically, a genomic fragment which had not been observed on the gel. Since no allowance is made for mismatched genomic fragments in the program, a correct match could not have been made.

Speed of locus prediction

The speed of identifying a locus for an individual BAC can vary greatly; factors affecting the processing time include the number of bands, the number of mismatches required, and obviously the computer being used to run BACFinder. However, on a 933 MHz Pentium PC with 512 MB RAM, it took us ~150 s to map 81 BACs. A successful match requiring no mismatches typically took <1 s, while failures (testing with zero, one and two mismatches) took up to 15 s. Larger and smaller genomes would, of course, require more and less time respectively.

Other restriction enzymes

The above results are based on data from *Hin*dIII restriction digests. We used *Hin*dIII in these initial experiments because it had been used both for the initial partial digest of the genome and for cloning into the BAC vector. Therefore, once we had identified and eliminated the vector band from the gel analysis, all remaining bands should have represented genuine genomic *Hin*dIII fragments.

However, this will not be possible in all cases, for example where a library is generated using a different enzyme to that used to clone the insert. Therefore we have also re-tested these BACs using different restriction enzymes, namely *Bam*HI and *Xba*I. The main problem of using restriction enzymes different to those used for producing the library is the generation of vector/insert hybrid bands. These will be formed from each end of the insert where the enzyme cuts some way into the insert and then somewhere in the vector, rather than cleanly at the end of the insert. These hybrid bands cannot be automatically removed when calling band sizes because they are of variable size and, since they are unlikely to correspond in size to the genuine genomic fragment of which they partially consist, they will cause problems with mismatches. We have considered two approaches to deal with such fragments. The

Table 3. Comparison of BACFinder predictions obtained from digests with three different restriction enzymes

Enzyme	BACFinder prediction		Chr[a]	Sequence results	
	Start	Stop		Start	Stop
*Hind*III	24925300 ± 8391	25044481 ± 17205	5	24924067	25027992
*Xba*I	24925646 ± 8614	25039044 ± 2299			
*Bam*HI	24927207 ± 10482	25037036 ± 11891			
*Hind*III	10249271 ± 434	10317398 ± 219	4	10248838	10317617
*Xba*I	10248412 ± 0	10315819 ± 1673			
*Bam*HI	10251955 ± 0	10324042 ± 8164			
*Hind*III	25362196 ± 4146	25434953 ± 3487	5	25365968	25431846
*Xba*I	25369039 ± 4607	25427437 ± 3435			
*Bam*HI	25378177 ± 9065	25430051 ± 3269			

Sequence result positions are for the location of the *Hind*III sites at the extreme ends of the insert.
[a]Chr, chromosome number.

first, and probably better, solution is to use an enzyme that does not cut the vector at all. If the vector is larger than the maximum band cut-off size, this will mean that the vector and the two fragment partial ends will be ignored in the analysis. The second is to allow the mismatch facility of BACFinder to deal with the hybrid bands. In the latter case, because mismatches are being introduced from the beginning, it may be necessary to increase the number of mismatches allowed in predicting a locus.

Because there were no suitable restriction enzymes that did not cut our vector at all, we used the second approach in our tests with alternative enzymes. A selection of the results is shown in Table 3 and demonstrates good agreement between the predictions based on the three different enzymes as well as the end-sequence data. All 81 initial BACs were re-tested using *Bam*HI and *Xba*I digests and, in all cases where a location for the BAC was predicted, agreement was seen between the results from the different digests. Unfortunately, neither extra digest generated enough bands to predict a location for the BAC that had been incorrectly positioned using the *Hin*dIII data.

One feature to note about the predictions based on different restriction digests is that the end predictions do not necessarily overlap. The predicted end coordinates are actually the coordinates of the outermost occurrences of the restriction enzyme site in the insert, rather than necessarily those of the full insert. For *Hin*dIII, the enzyme with which the library was generated and cloned, the two are the same. However, for the *Bam*HI and *Xba*I digests, the outermost restriction site may be some distance from the end of the insert.

In addition to acting as confirmation of the initial predictions, a second benefit of performing extra digests is that loci may be predicted for BACs that were not previously mapped. In our case, four additional BACs could be mapped with the *Bam*HI and *Xba*I data for which no match had been found using *Hin*dIII.

DISCUSSION

Of the 81 BACs initially tested, we have predicted single genomic loci for 64, or 79%, with an apparent accuracy of >90%. During testing of the accuracy of the matches, we have end-sequenced 44 BACs and, of these 88 sequences, we have identified genomic loci for 71, or 81%. Therefore, BACFinder gives a comparable rate of success of locus prediction as end-sequencing, albeit with a somewhat lower degree of accuracy. When the much reduced cost and labour required for mapping BACs using BACFinder is considered in addition (96 BACs can be processed in 2 days with very low consumables costs), we consider that BACFinder is a useful alternative to end-sequencing in mapping large numbers of BACs. Even in cases where knowing the exact end coordinates of a BAC insert is necessary, BACFinder can be a useful supplementary tool to rapidly screen for suitable target BACs, which can then be used for end-sequencing.

Although BACFinder was written principally to locate BACs within the *Arabidopsis* genome, we believe that it could be used just as easily for any other fully sequenced genome. However, we predict that as the genome size increases, it is likely that the minimum number of matched bands required to prevent false locations being predicted would also increase. Preliminary trials using the test datasets on individual *Arabidopsis* chromosomes suggest that the percentage of false matches generated for a digest of a given number of bands increases linearly as a function of genome size. However, since many BACs may generate 30 or more bands on digestion, BACFinder should still work well for such organisms provided that the average size of the BAC library inserts is relatively large. An alternative approach would be to perform parallel digests of the BACs with more than one restriction enzyme. Even if multiple loci were predicted from both digests, the correct locus would be evident by its presence in both sets of results, whereas the false predictions would occur in only one or the other.

For both of the above approaches, the choice of restriction enzyme used to perform the digest has a major impact on the quality of the results produced. The ideal enzyme will produce a large number of bands within the size range that can be resolved accurately on a gel, while at the same time it will cut the vector rarely, if at all: although vector-specific bands can be removed on band-size calling, there is a risk of these masking genuine insert bands on the gel. These features become even more important if analysis is carried out with an enzyme that was not used for library construction and cloning, when the generation of vector-insert hybrid products must be considered. These hybrids can be dealt with automatically by BACFinder, but it is best if they can be eliminated altogether by using an enzyme that does not cut the vector, or one that at least generates a large number of insert restriction fragments to reduce the risk of false predictions caused by these mismatching bands.

In addition to being usable on other genomes, BACFinder is not restricted to identifying loci for BACs. Any large DNA molecule could equally well be mapped, provided that all bands used in the mapping correspond to full-length genomic restriction fragments, and that contaminating vector bands can be identified and removed.

The software is available as a Java application and can be downloaded from http://jic-bioinfo.bbsrc.ac.uk/BACFinder/. A Perl version is also available, but it is not the recommended version and is provided principally for users unable to run the appropriate version of Java, or with specific restriction site choices. The majority of settings used by BACFinder are user-adjustable, so for example if a different gel system had a different range of band resolution, BACFinder could be tuned to give the best results from that digest data. BACFinder is accompanied by a second application, ChrDigest, which takes a genome sequence and generates a virtual digest suitable for use by BACFinder. We will continue to develop BACFinder, and up-to-date information on modifications and improvements will also be found on the website.

ACKNOWLEDGEMENT

This work was supported by BBSRC grant 208/IGF12424.

REFERENCES

1. Shizuya,H., Birren,B., Kim,U.J., Mancino,V., Slepak,T., Tachiiri,Y. and Simon,M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA*, **89**, 8794–8797.
2. Goodman,H.M., Ecker,J.R. and Dean,C. (1995) The genome of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **92**, 10831–10835.
3. Woo,S.S., Jiang,J., Gill,B.S., Paterson,A.H. and Wing,R.A. (1994) Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. *Nucleic Acids Res.*, **22**, 4922–4931.
4. Gispert,S., Dutra,A., Lieberman,A., Friedlich,D. and Nussbaum,R.L. (2000) Cloning and genomic organization of the mouse gene Slc23a1 encoding a vitamin C transporter. *DNA Res.*, **7**, 339–345.
5. Carpten,J.D., Makalowska,I., Robbins,C.M., Scott,N., Sood,R., Connors,T.D., Bonner,T.I., Smith,J.R., Faruque,M.U., Stephan,D.A. *et al.* (2000) A 6-Mb high-resolution physical and transcription map encompassing the hereditary prostate cancer 1 (HPC1) region. *Genomics*, **64**, 1–14.
6. Antoch,M.P., Song,E.J., Chang,A.M., Vitaterna,M.H., Zhao,Y.L., Wilsbacher,L.D., Sangoram,A.M., King,D.P., Pinto,L.H. and Takahashi,J.S. (1997) Functional identification of the mouse circadian Clock gene by transgenic BAC rescue. *Cell*, **89**, 655–667.
7. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
8. Marra,M.A., Kucaba,T.A., Dietrich,N.L., Green,E.D.,Brownstein,B., Wilson,R.K., McDonald,K.M., Hillier,L.W., McPherson,J.D. and Waterston,R.H. (1997) High throughput fingerprint analysis of large-insert clones. *Genome Res.*, **7**, 1072–1084.
9. Liu,Y.G., Shirano,Y., Fukaki,H., Yanai,Y., Tasaka,M., Tabata,S. and 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**, 6535–6540.