Chromosome walking with BAC clones as a method of genome mapping

Z. Kubát

Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic

ABSTRACT

Current sequencing projects are often based on random sequencing of genomic libraries followed by contig assembly by means of bioinformatics tools. This approach is convenient for whole genome sequencing projects. Chromosome walking described here is suitable for mapping and sequencing of short genomic regions in species where whole genome sequencing is not possible or for cloning gene from its closest known marker. This method is based on searching for overlapping BAC clones specific for the genomic region of interest.

Keywords: BAC library; sequencing; physical mapping

Genome sequencing is an important part of current genomics. Genome sequence knowledge facilitates all subsequent analyses of genome structure, organization and function. Libraries of large fragments, mainly BAC (bacterial artificial chromosome) libraries (Schizuya et al. 1992), are mostly used in genome projects (see Kejnovsky 2000 for review). BAC clones are highly stable in the host *Escherichia coli* but show a lower stability if vector contains tandem repeats (Song et al. 2001). Searching for overlapping BAC clones described here is the main point of contig assembly based on the principle firstly described by Bender et al. (1983). The first major use of chromosome walking was finding the Cystic fibrosis gene in 1989 (Riordan et al. 1989). Chromosome walking is used in moderate modifications in cloning and sequencing projects in plants (Stein et al. 2000, McCubbin et al. 2004), fungi (Broggini et al. 2007) and animals (Aerts et al. 2003).

The chromosome walking starts with a BAC clone called "seed BAC clone". Based on its end sequence, called "BAC end sequence" (BES), probe is amplified and used for the screening of the whole BAC library. The specificity of BAC clones selected by hybridization is verified by PCR with specific primers to exclude false positive hybridizing clones. Then, insert sizes of positive clones are estimated and BAC clones with largest inserts are preferentially selected to accelerate walking. Their BAC ends are then sequenced and this information is used for confirmation of overlaps with seed BAC by PCR. When seed BAC sequence is already known, more primers designed inside the overlapping area may be used for multiple confirmations followed by sequence alignment. If the seed BAC sequence is not known, sequencing of PCR products is necessary for verification of 100% identity of homologous regions in overlapping BAC clones. When overlap is confirmed, is located near the BAC end and the BAC insert is large enough, this new BAC is sequenced and used for other walking as a seed BAC.

BAC library screening

BAC libraries (Schizuya et al. 1992) usually have several times coverage of the genome. The widely used vector is pBelo BAC 11 (Kim et al. 1996); its insert size capacity is up to 400 kb. The starting point for the finding of BAC clones overlaps in

genomic library are BAC end sequences (BES) or ideally complete sequences of the seed BAC clone. Probes used for BAC library screening originate from BESs, on which the primers are designed. The goal of library screening is to find clones with an overlap as short as possible, to the limit of 40–50 kb but not shorter than 1 kb. Probes have to be amplified on seed BAC DNA template under highly specific PCR conditions to prevent contamination by closely related but not identical PCR products. PCR products must be purified (Quiagen PCR purification kit), because template DNA or BAC vector DNA contamination may cause unspecific cross-hybridization. Hybridization method does not have to be too sensitive (non-radioactive labeling can be used, e.g. AlkPhos direct labeling kit, Amersham Biosciences) because probe consists of highly concentrated short PCR products.

PCR confirmation

Glycerol stocks from all positive clones are prepared to store BAC clones apart of BAC library plates. Frozen glycerol stocks are stored at –20°C and can be used as PCR templates during next 2 to 3 months; if stored at –70°C, stocks can be used for several years. PCR confirmation of all positive clones with primers used previously for probe amplification has to be performed to exclude false positive clones. Only the clones producing PCR products of the same size as the PCR product of seed BAC are considered as really positive.

BAC insert size estimation

There is a high probability that more than one overlapping clone will be found in the library. In order to prevent work on too short clones, the insert sizes are estimated in selected BAC clones. Sequencing of long BACs will accelerate walking. Low amounts of BAC DNAs are sufficient to estimate insert size by Pulse Field Gel Electrophoresis (PFGE). BAC DNA isolated by plasmid miniprep is then digested with restriction enzymes surrounding the cloning site of BAC vector. Digested BAC DNA is loaded onto PFGE gel with DNA size marker. BAC clones of minimal size 80–100 kb are chosen for further work.
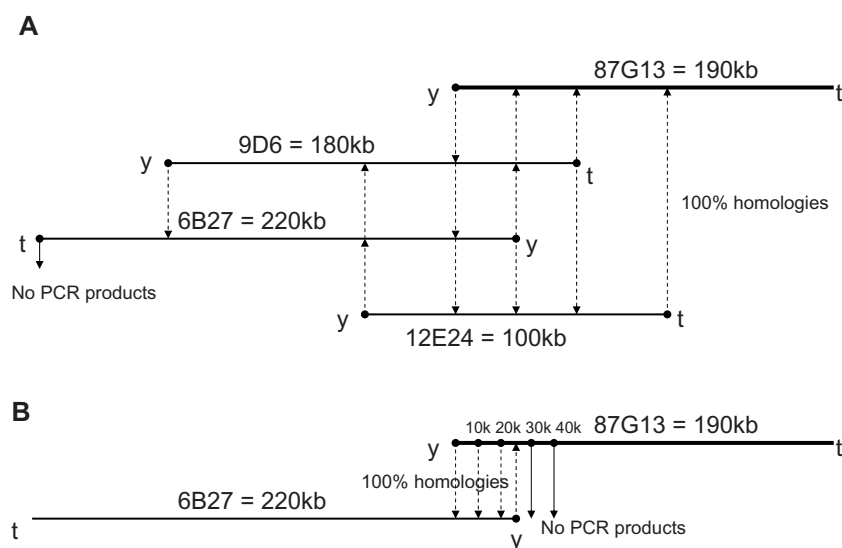

Figure 1. The scheme of contig assembling. BAC names used on the picture are an example

A: Strategy used when the seed BAC (87G13) is not sequenced. Searching for clones and overlaps confirmation is based on BAC end sequences (BESs). Primer pairs are designed for all candidate BAC clones BESs confirmed by PCR (marked by dashed arrows) to have homology with the seed BAC BES, in this case 87G13y. These primers are used for back PCR confirmation on each of potentially overlapping BAC clones, their PCR products are sequenced to verify 100% identity of homologous region. BAC clones orientation (y to t end) is determined

B: Strategy used when the seed BAC is sequenced. Several primer pairs in defined positions can be designed for multiple PCR confirmation of overlap between seed BAC 87G13 and the largest candidate BAC 6B27. PCR products should be sequenced to verify 100% identity. The length of overlap is unambiguously defined by position of homology between BAC 6B27 BES and 87G13 sequence. Analogous strategies are used for searching on the second end of seed BAC and also on the opposite ends of confirmed overlapping BACs

Back confirmation of positive clones

The overlaps in selected BAC clones must be verified by more than one PCR confirmation. Strategy is illustrated on Figure 1. If the BES is the only known seed BAC sequence, BESs of candidate positive BACs should be used for back confirmation of overlap. A list of BESs of all BAC clones in the library may be very helpful. To obtain such quantity of sequencing data characterizing each clone is very expensive and time consuming, but can significantly accelerate research progress. If BES is not available, sequencing of BAC ends must be performed. BAC DNA may be isolated by a common method, but in the case of direct BAC DNA sequencing with BAC vector specific primers, DNA template must be very pure. Quiagen or Promega plasmid isolation midi kits are good choice because they provide high yield (~10 μg) of ultrapure DNA from 50–70 ml of bacteria cultured in LB medium. One sequencing reaction requires 1.5–2.0 μg of BAC DNA. BAC ends are sequenced with T7 and M13 (or SP6) primers, respectively. BES originating in sequencing with T7 primer is called "BAC_NAME_t" and BES sequenced with M13 or SP6 primer is called "BAC_NAME_y", which is useful for determining of BAC clone orientation.

The other method of BES sequencing is BAC end cloning (Chen and Gmitter 1999) when BAC end subclone is used for sequencing instead of large BAC DNA. During this approach, BAC DNA is digested with one restriction enzyme cleaving inside the BAC vector simultaneously with another restrictase digesting inside BAC insert. This method is time consuming and has a high probability that subcloned BAC end sequence is short. Subsequently, primers for obtained BESs are designed and used for back PCR confirmation. Only candidate BACs whose BESs derived primers amplify products of correct size on seed BAC DNA can be considered as overlapping with seed BAC. Amplified PCR products should be then sequenced and aligned to BESs of primers origin (BLAST – align two sequences). Ideally, 100% identity can prove the overlap. In some cases, such double overlap confirmation does not guarantee accuracy of results and more regions used for confirmation should be analyzed. Namely repetitive elements and duplicated regions presence in BESs of both seed BAC and candidate BACs can cause mistakes in contig assembling. If the seed BAC is already sequenced, it can serve as a source of next primers useful for overlap confirmation and also for overlap length estimation. When the complete sequence of seed BAC is not available, more candidate BACs overlapping with seed BAC can be used for contig assembly (Figure 1A).

Factors affecting method efficiency

Several factors affect efficiency of this method. The most important factor is the BAC library quality and size (insert size, coverage of genome). When the library is too small and does not cover complete genome, hybridization screening might result in no positive clones. Even in a large library, some genomic regions may be missing. It can be caused by the absence of used restriction enzymes target sites surrounding a studied sequence. Repetitive sequences and duplications can cause big problems as well. They can lead to incorrect contig assembly, namely when repetitive regions are highly conserved or young and not much diverse. The use of a large probe could help in these situations. Supporting methods such as restriction fragment analysis, long range PCR (Barnes 1994, Cheng et al. 1994) or fiber FISH (Heiskanen et al. 1995, Weier et al. 1995) can also be used.

Acknowledgments

REFERENCES

Aerts J., Crooijmans R., Cornelissen S., Hemmatian K., Veenendaal T., Jaadar A., van der Poel J., Fillon V., Vignal A., Groenen M. (2003): Integration of chicken genomic resources to enable whole-genome sequencing. Cytogenet. Genome Res., *102*: 297–303.

Barnes W. (1994): PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc. Natl. Acad. Sci. USA, *91*: 2216–2220.

Bender W., Spierer P., Hogness D.S. (1983): Chromosomal walking and jumping to isolate DNA from the *ACE* and *rosy* loci and the bithorax complex in *Drosophila melanogaster*. J. Mol. Biol., *168*: 17–33.

Broggini G.A.L., Le Cam B., Parisi L., Wu C., Zhang H.B., Gessler C., Patocchi A. (2007): Construction of a contig of BAC clones spanning the region of the apple scab avirulence gene AvrVg. Fungal Genet. Biol., *44*: 44–51.

Chen C., Gmitter F.G. Jr. (1999): Direct cloning and sequencing of bacterial artificial chromosome (BAC) insert ends based on double digestion. Plant Mol. Biol. Rep., *17*: 231–238.

Cheng S., Fockler C., Barnes W.M., Higuchi R. (1994): Effective amplification of long targets from cloned inserts and human genomic DNA. Proc. Natl. Acad. Sci. USA, *91*: 5695–5699.

Heiskanen M., Hellsten E., Kallioniemi O.P., Makela T.P., Alitalo K., Peltonen L., Palotie A. (1995): Visual mapping by fiber-FISH. Genomics, *30*: 31–36.

Kejnovsky E. (2000): Construction of plant BAC libraries. Biol. Listy, *65*: 181.

Kim U.J., Birren B.W., Slepak T., Mancino V., Boysen C., Kang H.L., Simon M.I., Shizuya H. (1996): Construction and characterization of a human bacterial artificial chromosome library. Genomics, *34*: 213–218.

McCubbin A.G., Wang Y., Wang X., Kao T.H., Tsukamoto T., Huang S.S., Yi K.W. (2004): Chromosome walking in the *Petunia inflata* self-incompatibility (*S-*) locus and gene identification in an 881-kb contig containing S$_2$-RNase. Plant Mol. Biol., *54*: 727–742.

Riordan J.R., Rommens J.M., Kerem B., Alon N., Rozmahel R., Grzelczak Z., Zielenski J., Lok S., Plavsic N., Chou J.L., Drumm M.L., Iannuzzi M.C., Collins F.S., Tsui L.C. (1989): Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science, *245*: 1066–1073.

Shizuya H., Birren B., Kim U.J., Mancino V., Slepak T., Tachiiry 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*: 8794–8797.

Song J., Dong F., Lilly J.W., Stupar R.M., Jiang J. (2001): Instability of bacterial artificial chromosome (BAC) clones containing tandemly repeated DNA sequences. Genome, *44*: 463–469.

Stein N., Feuillet C., Wicker T., Schlagenhauf E., Keller B. (2000): Subgenome chromosome walking in wheat: A 450-kb physical contig in *Triticum monococcum* L. spans the *Lr10* resistance locus in hexaploid wheat (*Triticum aestivum* L.). Proc. Natl. Acad. Sci. USA, *97*: 13436–13441.

Weier H.U.G., Wang M., Mullikin J.C., Zhu Y., Cheng J.F., Greulich K.M., Bensimon A., Gray J.W. (1995): Quantitative DNA fiber mapping. Hum. Mol. Genet., *4*: 1903–1910.

Corresponding author:

Mgr. Zdeněk Kubát, Akademie věd České republiky, Biofyzikální ústav, v. v. i., Královopolská 135, 612 65 Brno, Česká republika
phone: + 420 541 517 203, e-mail: kubat@ibp.cz