MINIREVIEW SERIES FOR THE 50TH VOLUME

The development and applications of the bacterial artificial chromosome cloning system

Hiroaki Shizuya and Hosein Kouros-Mehr

Beckman Institute, Division of Biology, California Institute of Technology, Pasadena, CA, USA

Abstract. The development of the Bacterial Artificial Chromosome (BAC) system was driven in part by the Human Genome Project as a means to construct genomic DNA libraries and physical maps for genomic sequencing. The BAC system is based on the well-characterized *Escherichia coli* F-factor, a low copy plasmid that exists in a supercoiled circular form in host cells. The structural features of the F-factor allow stable maintenance of individual human DNA clones as well as easy manipulation of the cloned DNA. BACs are currently used in a wide array of applications from genome sequencing to gene discovery. This paper describes the key elements in the development of the BAC system and its current notable applications. (Keio J Med 50 (1): 26–30, March 2001)

Key words: BAC, YAC, pBeloBAC, pIndigoBAC

The Human Genome Project will undoubtedly have a profound effect on our understanding of the biological sciences and medicine. The knowledge generated by these efforts will change the practice of medicine in terms of the diagnosis, treatment, and prevention of disease. By the time this manuscript is published, most of the human genome and possibly the genomes of other organisms such as the mouse will have been sequenced. Genomic sequencing is only the beginning of an exciting and fruitful advancement in medical genetics. The post-sequencing era of functional genomics, which will require the annotation, deciphering, and interpretation of the genomic sequence, will be more difficult to complete and may take a considerably longer period of time. The annotation will require intimate collaboration between scientists working in biological sciences and in information technology. Physicians will also take more active roles in advancing medical sciences by utilizing the information generated by functional genomics studies. We feel honored to have been a part of these endeavors from the beginning through the Bacterial Artificial Chromosome (BAC) system that we developed. The BAC system is now widely used by many scientists in sequencing efforts as well as in studies of genomics and functional genomics.

The Development of the BAC Cloning System

As the Human Genome Project was underway in the early 1990s, there came the need to create high-resolution physical maps of each human chromosome, which would permit the isolation of short DNA fragments for direct sequencing and other manipulations. In response to this need, the Yeast Artificial Chromosome (YAC) system was developed.[1] Although YACs can carry DNA as large as one mb, subsequent studies indicated that the YAC system presented several difficulties in the creation of a human genome map.[2–4] Additionally, yeast cells were not as familiar to molecular biologists as other organisms such as *Escherichia coli*. To circumvent these difficulties, we developed a bacterial cloning system based on the well-characterized *E. coli* F-factor, a low-copy plasmid that exists in a supercoiled circular form.[5] We had three major goals in mind when we developed the BAC cloning system. First, we wanted to have a stable bacterial cloning system that was easy to manipulate for mapping and genome analysis. The long linear structure of YACs made it difficult to recover the cloned DNA in pure form, as YACs were more susceptible to shearing. Our second goal was to have a large insert that could serve

Reprint requests to: Dr. Hiroaki Shizuya, Beckman Institute, Division of Biology, 139-74, California Institute of Technology, Pasadena, CA 91125, USA, e-mail: shizuya@its.caltech.edu

as a substrate for DNA sequencing. Other bacterial cloning systems such as cosmid and P1 vectors had been developed but these had maximum cloning capacities of 45 kb and 100 kb, respectively.[6] The third goal was to provide a tool that would later be used for functional genomic studies once the human genome and other genomes had been sequenced.

BAC vectors

We based the BAC system on the *Escherichia coli* F-factor because of the wealth of information in its genetics and molecular biology. The F-factor had long ago been used as a tool for studying gene regulation and mutation analysis in a variety of *E. coli* operons. Replication of the F-factor is strictly controlled by the regulatory functions of *E. coli*, and as a result the F-factor is maintained in low copy number (i.e., one or two copies per cell). This allows stable maintenance of large DNA inserts and reduces the potential for recombination between DNA fragments carried by the vector. The latter property is reflected in the fact that genomic inserts in high-copy number cloning systems such as cosmids often undergo extensive and unwanted rearrangements. In addition to stable maintenance, the structural stability of F-factors allows complex genomic DNA inserts to be maintained with a great degree of structural stability in the *E. coli* host.

BACs have several additional advantages over YACs. It was observed that a large percentage of YACs carried chimeric inserts, making mapping efforts confusing and difficult. BACs, in contrast, are virtually free from chimerism. Another problem with YACs is that multiple YAC chromosomes may coexist in a single yeast cell. In the BAC system, the F-factor-encoded parA and parB genes are involved in exclusion of multiple F-factors, and as a result multiple BACs cannot coexist in a single cell. We have thus far analyzed more than 10,000 BAC clones and have not yet found cells carrying multiple BACs.

BAC vector cloning site

The BAC vector incorporates several sites and markers. The cloning segment includes (1) two bacteriophage markers, lambda cosN and P1 loxP; (2) three restriction enzyme sites (*Eco*RI, *Hind*III and *Bam*HI) for cloning; and (3) a GC-rich *Not*I restriction enzyme site for potential excision of the inserts. The cosN site provides a fixed position for cleavage by bacteriophage lambda enzyme terminase, which allows the convenient generation of a linear form of the BAC DNA. The cosN site is also used to package approximately 50 kb DNA into the bacteriophage lambda head as a particle. The method, known as Fosmid for *F*-based cosmid system, is extremely efficient and thus very useful when DNA is precious or available in limited amounts. The P1 loxP site allows the retrofitting of additional components to the BAC vector at a later stage; *e.g.*, the addition of the G418 resistant marker for selection in eukaryotic cells. The loxP site can also be used to linearize BACs through the P1 phage protein Cre, which catalyzes strand exchange between two DNA strands at the loxP sites.

The BAC vector pBeloBAC11 contains an additional component in the cloning site – β-galactosidase (*lacZ*), which allows α complementation. This allows clones with inserts to be readily identified as an X-gal color change. A more recent BAC vector known as pIndigoBAC displays a much faster and deeper X-gal color change as a result of a point mutation in the 3′ end of the *lacZ* gene. This frameshift causes premature termination of the β-galactosidase polypeptide. All BAC vectors, including pBeloBAC11 and pIndigoBAC, contain cloning sites that are flanked by T7 and SP6 promoters, which can be utilized for DNA sequencing of the insert at the vector-insert junction (BAC end sequencing). Additionally, the BAC vectors carry a chloramphenicol resistance gene in order to screen for host cells that carry BACs.

BAC cloning

Two important enzymes were key in the creation of the BAC system – agarase and temperature-sensitive alkaline phosphatase. Although we knew how to separate large DNA through Pulsed-Field Gel Electrophoresis, we did not know how to purify DNA from gel matrices without damaging the DNA. After trying several procedures, we found that the enzyme agarase could be used to recover the DNA by breaking down the agarose matrices. The main problem we faced is that there was no pure agarase that could be used without damaging DNA larger than 100 kb; fortunately, an enzyme supply company aided us by providing DNAse-free agarase. The second enzyme, temperature-sensitive alkaline phosphatase, was greatly needed to prepare BAC vector DNA. Since the BAC vector is a single-copy plasmid, it was extremely laborious to obtain sufficient amounts for use in cloning, especially since the DNA must first be dephosphorylated. At the time, the available enzyme for dephosphorylation was *E. coli* alkaline phosphatase, which is extremely difficult to remove from the system subsequent to the reaction and can ultimately result in a considerable loss of DNA. We needed a phosphatase that could be destroyed by raising the temperature to 65°C. Fortunately, we had purified such an enzyme about 10 years ago from a bacterial species isolated in Antarctica, and in collaboration with the same company we obtained the

temperature-sensitive alkaline phosphatase that we needed.

An additional obstacle was to choose the correct bacterial strain for cloning. Although most *E. coli* strains can be transfected with DNA of a maximum size of 100 kb, we found that certain strains of *E. coli* such as DH10B or DH5 can take up DNA larger than 100 kb. There may be a genetic disposition or specific genetic markers in the DH10B or DH5 strains that may be responsible for this differential uptake.

Construction of BAC libraries

Having constructed the BAC cloning system, we subsequently made four human BAC libraries from three individuals.[7] Each library contains BAC clones of multiple coverage and an average insert size of 125–175 kb. A fraction of Library A contained the original BAC vector pBAC108L, which lacked the lacZ gene for Xgal color selection. As a result, we needed to perform colony hybridization in order to identify clones carrying human DNA inserts. Libraries A and B were constructed from a cell line obtained from the American Type Culture Collection. The cell line was derived from a deceased male individual whose body was donated for research and whose name was held in confidence. Library C was made from a donor who signed informed consent; however, the donor's identity was not anonymous. The D library was constructed from an anonymous donor in accordance with the NIH/DOE Guidelines. For all four library construction, we received approval from the Caltech Institutional Review Board (IRB).

Donor confidentiality

When large scale sequencing began at various centers, there was growing concern about maintaining confidentiality of the DNA donors from which BAC DNA libraries were constructed, and also the sequences were obtained. As a result the NHGRI-DOE Guidance was issued in 1996. In response to the Guidance, we obtained approvals from both the Caltech IRB and NIH-DOE with regard to the informed consent form and the sperm collection protocol. The approved process was utilized to construct library D.

Before donors were asked to provide samples of blood and semen for the construction of BAC libraries, they were fully informed about the nature of the experiment and were subsequently asked to sign the informed consent form. The form fully describes the benefits and potential problems of participating in the program. It starts by describing how donor confidentiality will be maintained, what will be done with the samples, why we need samples, and who will use samples. The form goes on to explain the benefits of participating in the program; *e.g.*, the donor's contribution to the advancement of human health and human genetics. At the same time, it stresses that the donor will not receive financial gain and also describes risks of the program; for example, if the donor's DNA sequence becomes known, the donors and their families may inadvertently discover the risk of an illness that had not been known previously. There would also be the added problem that if the information is retrieved by employers or insurance companies, it may become difficult to procure a job or health insurance.

In order to reduce the possibility that the identity of the donor would become revealed, we implemented the following protocol. To receive samples, we contacted a sperm bank and designated one person to contact donors. Donors who expressed interest in participating in the project came to the contact person without providing their names, addresses, phone numbers, and any other identifiable information. The contact person explained the nature of the human genome project and the potential benefits and risks associated with the project. After donating blood and semen samples, the donors were thereafter represented by a randomly-chosen four-digit number (even during future semen donations). Twenty persons served as donors. The semen samples were shipped to Caltech and recorded by alphabetization. The blood samples were shipped to MIT and used to establish permanent lymphocyte cell lines. From the twenty semen samples, we randomly chose one sperm sample for BAC library construction. In this way, neither the contact person nor we knew the identity of the donor since the contact person did not know the identity of the alphabetized samples and since we did not know who participated in the program. Thus, no person, including the donors, knew who had been chosen for the construction of the BAC library.

Current Applications of the BAC System

BACs are currently used in a wide array of applications, from genome sequencing to gene discovery. Below are several of the most notable applications to date.

Genome sequencing

BACs are the major driving force in the effort to sequence the human genome. The collaborative effort to sequence human chromosome 22 was aided in part by chromosome 22 BAC sub-libraries, which were made by screening human BAC libraries with sequence tagged site (STS) markers known to be located on chromosome 22.[8–9] By using chromosome 22 BAC sub-libraries in addition to cosmid, Fosmid, and P1-derived artificial chromosome (PAC) libraries, the consortium

was able to cover chromosome 22q in 11 clone contigs with 10 gaps, spanning nearly 33.4 megabases of sequence.[10] Recently, the human chromosome 21 has also been sequenced using BAC maps.[11] The relatively short DNA insert sizes (100–300 kb) and structural stability make BACs the ideal substrates for direct sequencing. Additionally, the use of primers based on pBeloBAC and pBAC108 vectors can be used to quickly and efficiently sequence BAC ends, a fact that greatly facilitates sequencing efforts. For these reasons, BACs are currently being used in large-scale projects to sequence the human and mouse genomes.

BAC contig assembly

The assembly of BAC contigs remains a critical step in the formation of physical maps. Classically, building contigs has involved utilizing restriction endonuclease digestion to determine pairwise overlap between BAC clones. Recently, we have described a novel method for BAC contig assembly which involves multiplexed fluorescence-labeled fingerprinting.[12] In this method, each BAC clone is digested with three pairs of restriction endonucleases and labeled with three distinct fluorescent dyes, then analyzed on a DNA sequencing machine. By analyzing the fingerprinting patterns of two or more BACs, it is possible to determine overlap among the BACs and thus develop contigs.

Positional cloning

BACs have been widely used to identify disease loci as well as individual genes. One method of identifying genes involves using positional cloning to identify genetic loci responsible for a disorder. The loci can then be analyzed for individual genes by creating physical maps of the region via BAC contig assembly (see above). This method was the basis for finding the breast cancer susceptibility gene BRCA1.[13]

BAC library screening

Another method of using BACs to identify genes entails screening BAC libraries for known sequences. Such a screen can be carried out by PCR analysis of BAC library pools, such as row and column pools for all the plates in a library.[7] This method was employed to search for human pheromone receptor genes using as template sources human ESTs with homology to mouse pheromone receptors (data not published). Once a gene is isolated onto an individual BAC, FISH or Radiation Hybrid mapping can be used to map the cytogenetic localization of the BAC and gene.[14] The short insert sizes and purification efficiency of BACs make them ideal for such cytogenetic analysis.

BAC transgenic mice

Traditionally, transgenic mice have been created by microinjecting copies of a gene as linear DNA into mice fertilized eggs, and then introducing embryos into a mouse. The disadvantage of using single gene copies is that oftentimes upstream and downstream elements of the gene are not included in the transgene. An approach to this problem is to introduce transgenes that incorporate. BACs rather than coding segments; in this way, distant regulatory elements such as enhancers are included with the gene of interest. For example, an 80 kb P1 bacteriophage spanning the human or mouse apoB gene conferred expression of apoB in the liver but not the intestines of transgenic mice. The use of 150–200 kb BACs spanning the apoB gene did confer apoB expression in both the liver and intestines of transgenic mice, suggesting that intestinal expression of apoB is controlled by a distant element that was not present on the apoB-carrying P1 bacteriophage.[15] An alternative approach to BAC transgenic mice is the *in vivo* complementation test in which BAC transgenes are incorporated into mutant mice that lack a gene of interest. This ''cloning by rescue'' approach was used as a complement to positional cloning in order to identify the circadian Clock gene in mice. In the experiment, a 140 kb BAC transgene was shown to completely rescue the loss of rhythm phenotype found in Clock mutant mice, suggesting that the clock gene was located on the BAC.[16]

BAC microarrays

As the genome project nears completion, the nature of genomics research will shift from the ''one gene, one hypothesis'' paradigm to the simultaneous analysis of thousands of genes in a given experimental system. The driving force for this breakthrough paradigm is the microarray, a glass microscope slide that carries thousands of discrete, covalently linked DNA sequences. By using fluorescently linked RNA or cDNA samples, it is possible to profile the gene expression patterns from a cell line or tissue of interest. The creation of microarrays carrying BACs will be particularly useful, especially if the sequence and cytogenetic properties of the BACs are known. In this way it will be possible to link gene expression profiles with chromosomal data and DNA sequences in order to achieve a global understanding of a disorder or genetic abnormality.[17]

Acknowledgements: We are grateful to Dr. Mel Simon and many people who worked on this project, and to the US Department of Energy and National Institutes of Health for supporting this work.

References

1. Burke DT, Carle GF, Olson MV: Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 1987; 236: 806–812
2. Green ED, Riethman HC, Dutchik JE, Olson MV: Detection characterization of chimeric yeast artificial-chromosome clones. Genomics 1991; 11: 658–669
3. Bellis M, Gerard A, Charlieu JP, Marcais B, Brun ME, Viegas-Pequignot E, Carter DA, Roizes G: Construction and characterization of a partial library of yeast artificial chromosomes from human chromosome 21. DNA Cell Biol 1991; 10: 301–310
4. Wada M, Little RD, Abidi F, Porta G, Labella T, Cooper T, Della Valle G, D'Urso M, Schlessinger D: Human Xq24–Xq28: approaches to mapping with yeast artificial chromosomes. Am J Hum Genet 1990; 46: 95–106
5. Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M: Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factory-based vector. Proc Natl Acad Sci USA 1992; 89: 8794–8797
6. Sternberg N: Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. Proc Natl Acad Sci USA 1990; 87: 103–107
7. Kim UJ, Birren BW, Slepak T, Mancino V, Boysen C, Kang HL, Simon MI, Shizuya H: Construction and characterization of a human bacterial artificial chromosome library. Genomics 1996; 34: 213–218
8. Kim UJ, Shizuya H, Kang HL, Choi SS, Garrett CL, Smink LJ, Birren BW, Korenberg JR, Dunham I, Simon MI: A bacterial artificial chromosome-based framework contig map of human chromosome 22q. Proc Natl Acad Sci USA 1996; 93: 6297–6301
9. Kim UJ, Shizuya H, Chen XN, Deaven L, Speicher S, Solomon J, Korenberg J, Simon MI: Characterization of a human chromosome 22 enriched bacterial artificial chromosome sublibrary. Gen Anal 1995; 12: 73–79
10. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, *et al.*: The DNA sequence of human chromosome 22. Nature 1999; 402: 489–495
11. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, *et al.*: The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. Nature 2000; 405: 311–319
12. Ding Y, Johnson MD, Colayco R, Chen YJ, Melnyk J, Schmitt H, Shizuya H: Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting. Genomics 1999; 56: 237–246
13. Neuhausen SL, Swensen J, Miki Y, Liu Q, Tavtigian S, Shattuck-Eidens D, Kamb A, Hobbs MR, Gingrich J, Shizuya H, *et al.*: A P1-based physical map of the region from D17S776 to D17S78 containing the breast cancer susceptibility gene BRCA1. Hum Mol Genet 1994; 3: 1919–1926
14. Trask BJ, Massa H, Brand-Arpon V, Chan K, Friedman C, Nguyen OT, Eichler E, van den Engh G, Rouquier S, Shizuya H, Giorgi D: Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. Hum Mol Genet 1998; 7: 2007–2020
15. Nielsen LB, McCormick SP, Pierotti V, Tam C, Gunn MD, Shizuya H, Young SG: Human apolipoprotein B transgenic mice generated with 207- and 145-kilobase pair bacterial artificial chromosomes. Evidence that a distant 5′-element confers appropriate transgene expression in the intestine. J Biol Chem 1997; 272: 29752–29758
16. Antoch MP, Song EJ, Chang AM, Vitaterna MH, Zhao Y, Wilsbacher LD, Sangoram AM, King DP, Pinto LH, Takahashi JS: Functional identification of the mouse circadian Clock gene by transgenic BAC rescue. Cell 1997; 89: 655–667
17. Korenberg JR, Chen XN, Sun Z, Shi ZY, Ma S, Vataru E, Yimlamai D, Weissenbach JS, Shizuya H, Simon MI, *et al.*: Human genome anatomy: BACs integrating the genetic and cytogenetic maps for bridging genome and biomedicine. Genome Res 1999; 9: 994–1001