Magnetic bead capture of expressed sequences encoded within large genomic segments

Danilo A. Tagle, Manju Swaroop, Michael Lovett and Francis S. Collins

Magnetic bead capture utilizes biotin–streptavidin magnetic bead technology to isolate cDNAs rapidly from large genomic intervals, giving several thousand-fold enrichment of the selected cDNAs. The technique can allow parallel analysis of several large genomic segments of varying complexities and can be applied to the isolation of expressed sequences from various tissue sources.

RAPID identification and isolation of transcribed sequences from defined chromosomal regions is critical to the success of positionally cloning any gene responsible for an inherited disorder[1]. Conventional methods of finding transcripts, such as using phylogenetically conserved sequences and CpG-rich fragments as hybridization probes, are laborious and cumbersome when the candidate interval spans several hundred thousand to a few million base pairs. Recently, several new approaches aimed at identifying and isolating transcripts from large genomic segments, such as from heteronuclear RNA derived from human-rodent hybrid cell lines[2,3], trapping of exons based on the presence of splice junctions[4,5], direct screening of cDNA libraries with total yeast artificial chromosomes (YACs)[6,7] and enrichment of cDNAs using immobilized YACs[8,9], have been reported. Although each of these methodologies has certain advantages, each has limitations that dictate specific applicability. The isolation of human transcripts from somatic cell hybrids[2,3] has the advantage of scanning entire chromosomes or subchromosomal regions present in the hybrids, but in general, only constitutively-expressed genes can be isolated from these cell lines. As exon trapping[4,5] rescues transcriptional units directly from genomic clones, it does not directly rely on the the availability of tissue- or temporal-specific cDNA libraries. However, thus far the method requires relatively small genomic inserts, such as cosmids, for trapping. In addition, cryptic splice sites in the human genome can cause false positives, and the technique will also miss intronless genes. Direct cDNA screening[6,7] using YACs allows several hundred thousand base pairs to be scanned at once but because of the complexity of the probe, it suffers from a low signal-to-noise ratio, resulting in low reproducibility. In contrast, cDNA enrichment schemes[8,9] are potentially powerful and rather straightforward. It is questionable, however, whether immobilizing YACs on membrane filters, as described in these original enrichment methods[8,9], provides easy access for complex cDNA probes to find their targets.

More recently, our group[10,11] and others[12] have modified the cDNA enrichment strategy by hybridizing biotinylated genomic DNA to cDNA probes in solution and subsequently capturing the hybridized complexes using streptavidin-coated magnetic beads (Fig. 1). We have used this method[10] to isolate cDNAs encoded over a 2.2-megabase (Mb) interval for the Huntington disease (HD) candidate re-

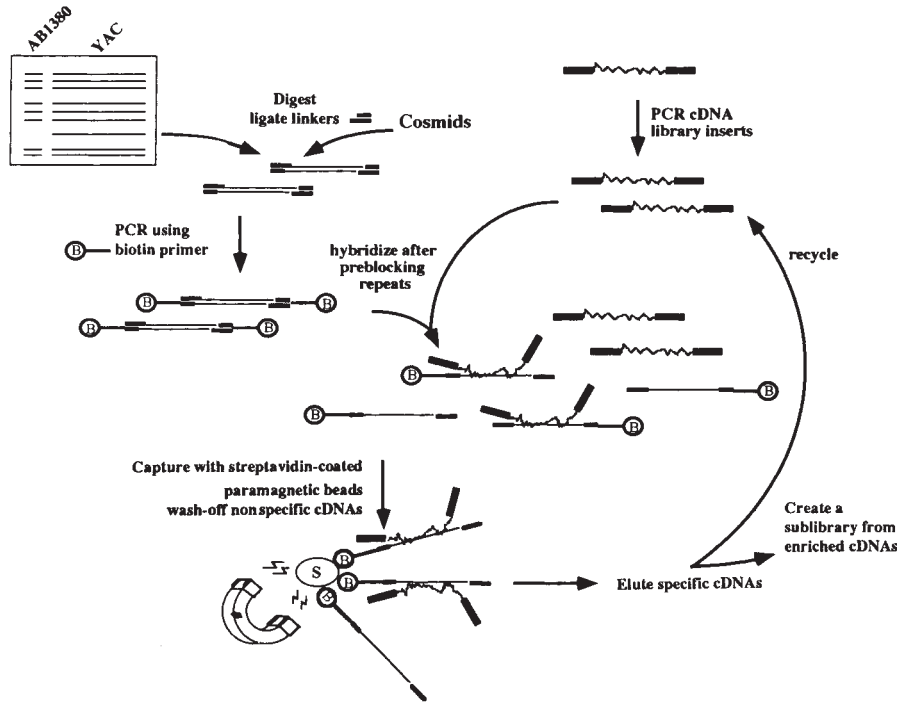

FIG. 1 General scheme for magnetic bead capture of expressed cDNAs from cosmids and YACs. Pools of miniprep cosmid DNA or of pulsed-field gel purified YAC genomic DNA are digested with four-base cutters and ligated to linkers. In place of restriction enzyme digestion, genomic DNA can also be sonicated and blunt-ended. The linkered genomic segments are PCR-amplified using a 5′-biotinylated primer whose sequence matches one of the linker arms. cDNA library inserts are PCR-amplified using vector primers and hybridized at high stringency to the biotinylated genomic fragments after suppression of repetitive sequences. The biotinylated genomic-cDNA complexes are captured using streptavidin-coated paramagnetic beads and the unbound, nonspecific cDNAs are washed off. The captured cDNAs are eluted and PCR-amplified. The PCR products are digested with *Eco*RI and subcloned as a region-specific sublibrary. To increase selectivity, the captured cDNA material is recycled through a second enrichment. In a representative study, the minimal set of 8 overlapping YACs that spans the 2.2-Mb HD candidate interval from D4S126 to D4S98[13,14] was used either individually or combined together as a pool. The YACs ranged in size from 200 to 600 kb and were deemed to contain only chromosome 4 sequences[14]. In addition, a pool of 17 arrayed cosmids (kindly provided by Gillian Bates of the Imperial Cancer Research Fund) that spans the region from D4S43 to D4S98 (approximately 550 kb) was used to capture cDNAs. Inserts from a commercially available random-primed, fetal brain cDNA library in λ-ZAP were amplified using T3 and T7 vector primers. A more detailed protocol is available from the authors upon request.

gion[13,14], which may contain as many as 100 genes in this very transcript-rich region of chromosome 4p16.3. Our comparisons using (1) pooled cosmid DNA, (2) individual YACs and (3) pooled YACs emphasize a key aspect of the methodology, that is, the technique is amenable to the parallel analysis of several large genomic segments of varying complexities in isolating expressed sequences from various tissue sources. As both genomic clones and cDNA probes are amplified by the polymerase chain reaction (PCR), the method provides another substantial advantage in that only a small quantity of starting materials is required. Specific 4p16.3 transcripts found in low abundance can be enriched several thousand fold with an associated decrease in nonspecific, ubiquitous transcripts, such as β-actin.

Technical considerations

Figure 1 shows schematically the general procedure involved in magnetic bead capture of cDNAs encoded from large genomic insert clones. The modification of biotinylating genomic DNA coupled with very high-affinity binding with streptavidin-coated paramagnetic beads provides better control of solution hybridization conditions compared to the kinetics of filter hybridization[15–17]. The strength and stability of the biotin–streptavidin coupling allows DNA manipulations, such as thermal denaturation and elution of annealed cDNAs or relative ease in changing buffers and wash solutions. This ensures a flexible system where preblocking, hybridization and washing can be performed easily. In addition, the beads are of equal size (monodispersed) and thus they follow uniform kinetics when subjected to a magnetic field[18].

The biotin can be introduced into the genomic clones by photobiotinylation[12] or nick-translation[12] where, in both cases, biotin is incorporated randomly along the length of the DNA. An alternative way[10,11] is to use a biotinylated primer during an *in vitro* PCR amplification reaction. The latter method has the advantage of tagging each amplified DNA strand at the 5′ end with only one biotin molecule, which can prevent any possible steric hindrance with *Taq* polymerase during PCR or with cDNA probes during hybridization.

The genomic DNA used for capturing cDNAs can be from any genomic source. However in our experience, enrichment with cosmid DNA seems to work more efficiently than with YACs as during the gel-purification of the latter, YACs often co-migrate with degraded, higher molecular weight yeast chromosomes that harbour tandem ribosomal sequences. This has led to interesting enrichment artefacts where we[10,11] and others[8] often find that 30–60 per cent of the clones from the enriched sublibrary from YACs contain ribosomal sequences. This is most likely due to cross-species hybridization of highly conserved ribosomal sequences between human and yeast and not from yeast contamination of the cDNA libraries, as the same phenomenon is observed when cDNA is extracted and used directly from cell lines[11].

A significant number of cDNA libraries contain repetitive sequences[19] in the form of expressed repeats or unspliced messages and some cDNA clones contain repeats, particularly in their 3′ untranslated region. Artefactually enriching these repeat-containing cDNAs can be partly avoided by suppressing repetitive sequences in the genomic target DNA with sheared human placental DNA. However, low-copy repeats are never adequately quenched, and can contribute up to 10 per cent of the enriched sublibrary. This is especially important as some chromosomal regions have relatively higher frequency of low-copy repeat sequences than others.

Specificity and enrichment

The specificity of the enrichment process can be assessed by examining the nature of the captured cDNAs. In the example of the HD region, PCR-amplified inserts from the starting fetal brain cDNA library were dot blotted along with amplified inserts from magnetic capture with each YAC from the HD candidate interval (Fig. 2). A cDNA probe, α-adducin, known to map telomeric to the D4S95 locus[20] which is spanned by two overlapping YACs (70D11 and YGA2) hybridized only to enriched cDNAs from these two YACs (Fig. 2*a*). Similarly, the YC70 cDNA probe only hybridizes to captured cDNAs using YACs 177B7 and YGA10 (Fig. 2*b*) which span across the D4S166 locus where YC70 has been mapped[21]. These demonstrate the high specificity of the captured cDNAs despite the complexity of the target DNA.

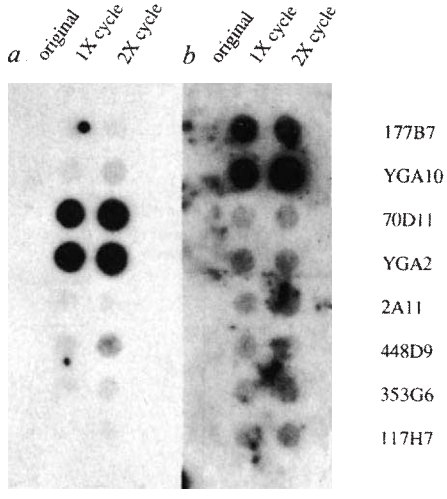

FIG. 2 Hybridization of control cDNA probes to captured cDNAs from YACs. The 8 YAC clones from the HD candidate interval[14] used in the magnetic bead capture experiments were 177B7 (580 kb), YGA10 (450 kb), 70D11 (500 kb), YGA2 (600 kb), 2A11 (400 kb), 448D9 (350 kb), 353G6 (350 kb) and 117H7 (200 kb). Equal amounts (by weight) of PCR-amplified cDNA inserts from the starting library and from the once- and twice-enriched sublibraries for each of the YACs were dot-blotted onto membrane filter. Panels *a* and *b* show hybridization results when the blot was hybridized with the cDNA probes α-adducin[20] (encoded in the region spanned by YACs 70D11 and YGA2) and YC70[21] (encoded in the region spanned by YACs 177B7 and YGA10), respectively.

A quantitative assessment of the degree of enrichment can be determined by hybridizing cDNA probes to plaque lifts from original and twice-enriched cDNA libraries and counting the number of plaques detected by each probe. Approximately $1–2 \times 10^6$ and 1,000 clones from the original and twice-cycled cDNA libraries, respectively, were surveyed (data not shown). A screen of α-adducin[20] in the starting fetal brain cDNA library resulted in the detection of only one clone per 500,000 plaques. In comparison, α-adducin was represented in the twice-cycled cDNA library from pooled 4p16.3 YACs at a frequency of 1 in 250 clones, indicating a 2,000-fold enrichment of this cDNA. A cDNA clone, MBC6, isolated using pooled 4p16.3 cosmids was detectable in the primary cDNA library at 1 per 2×10^6 clones. MBC6 was detected in the twice-cycled cDNA sublibrary using the pooled 4p16.3 cosmids at 1 per 1000 plaques giving also a 2,000-fold enrichment. Similar levels of enrichment of cDNAs have been reported from a 425-kb YAC from chromosome 5 (ref. 11) and from two cosmid contigs (550 kb and 350 kb) from the X chromosome[12]. As a corollary, we tested the level of nonspecific transcripts, such as β-actin, in the enriched material. β-actin was depleted at least 50-fold from the twice-cycled pooled YAC cDNA library.

To analyse the enriched cDNAs further, 50 clones from the twice-enriched cDNA sublibrary from the pooled 4p16.3 cosmids were randomly picked and mapped to a panel containing *Eco*RI digests of these cosmids (data not shown). Of the 50 clones, 35 mapped to at least one cosmid in the contig while the remaining 15 hybridized to cosmid vector bands, which indicates a possible contamination of the cDNA library with vector sequences.

Conclusions and prospects

Magnetic bead capture constitutes a powerful tool for the isolation of cDNAs encoded from large genomic regions. The technique is highly efficient and can be applied to many sources of cDNA and genomic clone pools in parallel. Furthermore, the method tends to normalize transcript levels, with a greater enrichment of rare messages than abundant ones. Several potential limitations, as well as possible solutions, are associated with this method. First, the resulting small insert sublibrary necessitates screening a full-length cDNA library to obtain full-length transcripts. The time involved in rescreening, however, can be cut substantially by screening random-primed cDNA

libraries. Alternatively, the insert size of the enriched cDNAs can be preserved by direct cloning of the PCR products. Second, genomic clones that contain region-specific low-copy repeats can cause a significant number of enriched cDNAs to contain these repeat sequences. Once identified, however, a running catalogue of these repeats can be used to pre-screen the sublibrary or can be incorporated as blocking agents in future magnetic bead capture experiments. Likewise, artefactually selected ribosomal cDNA clones can be pre-screened with ribosomal DNA probes. Third, since the methodology is based on hybridization by sequence homology, pseudogenes and members of multigene families may also be enriched. However, these can be sorted through by further mapping or sequencing of the enriched cDNAs. Last, in order to capture all encoded cDNAs from a given region, it would be ideal to use a cDNA library representing all possible transcribed human sequences. In this regard, it is possible to use pooled and normalized libraries from multiple tissue sources and from various developmental stages to approach such a complete cDNA library.

Magnetic bead capture of cDNAs provides an effective and straightforward alternative for isolation of expressed sequences from large genomic tracts. This method should be especially useful for isolating candidate genes where the defined candidate interval has been delineated by linkage analysis. It is anticipated that this technique will expedite the process of finding the proverbial needle in a haystack for many disease loci. As the methodology is PCR-based, it may be also possible to extend the application of this technique to chromosome microdissected material, either as whole chromosome or as region-specific chromosomal scrapes, or to flow-sorted chromosomes, in order to generate a catalogue of chromosome-specific transcribed sequences that can be integrated with the genetic and physical maps of the genome. As such gene-hunting techniques become more efficient, isolating candidate genes will no longer be the rate-limiting step in positional cloning. □

Danilo A. Tagle[1] and Francis S. Collins[1,2] are at the [1]Departments of Human Genetics and [2]Internal Medicine, University of Michigan Medical Center, Ann Arbor, Michigan 48109–0650, USA. Manju Swaroop and FSC are with the Howard Hughes Medical Institute. Michael Lovett is at the Department of Biochemistry, University of Texas - Southwestern Medical Center, Dallas, Texas 75235, USA. *DAT is supported by a National Institutes of Health - National Research Service Award and a Postdoctoral Fellowship Grant from The Wills Foundation. We thank our colleagues from the Hereditary Disease Foundation Collaborative Research Group, especially Gillian Bates and Hans Lehrach who have provided the ordered cosmids used in this study, Marcy MacDonald, Russell Snell and John Valdes who provided the control cDNA probes. We also thank Iris Gonzales for providing the human 45S rDNA probes, and Yossi Shiloh and Lawrence Brody for helpful discussions.* *For more information on the protocol, fill in reader service number 100.*

1. Collins, F. S. *Nature Genet.* **1**, 3–6 (1992).
2. Liu, P., Legerski, R. & Siciliano, M. J. *Science* **246**, 813–815 (1989).
3. Corbo, L., Maley, J. A., Nelson, D. L. & Caskey, C. T. *Science* **249**, 652–655 (1990).
4. Duyk, G. M., Kim, S. W., Myers, R. M. & Cox, D. R. *Proc. natn. Acad. Sci. U.S.A.* **87**, 8995–8999 (1990).
5. Buckler, A. J. *et al. Proc. natn. Acad. Sci. U.S.A.* **88**, 4005–4009 (1991).
6. Wallace, M. R. *et al. Science* **249**, 181–186 (1990).
7. Elvin, P. *et al. Nucleic Acids Res.* **18**, 3913–3917 (1990).
8. Lovett, M., Kere, J. & Hinton, L. M. *Proc. natn. Acad. Sci. U.S.A.* **88**, 9628–9632 (1991).
9. Parimoo, S., Patanjali, S. R., Shukla, H., Chaplin, D. D. & Weissman, S. M. *Proc. natn. Acad. Sci. U.S.A.* **88**, 9623–9627 (1991).
10. Tagle, D. A. *et al. Am. J. hum. Genet.* **51**, A22 (1992).
11. Morgan, J. G., Dolganov, G. M., Robbins, S. E., Hinton, L. M. & Lovett, M. *Nucleic Acids Res.* **20**, 5173–5179 (1992).
12. Korn, B. *et al. Hum. molec. Genet.* **1**, 235–242 (1992).
13. Bates, G. P. *et al. Am. J. hum. Genet.* **49**, 7–16 (1991).
14. Bates, G. P. *et al. Nature Genet.* **1**, 180–187 (1992).
15. Britten, R. J., Graham, D. E. & Neufeld, B. R. *Meth. Enzym.* **29**, 363–418 (1974).
16. Britten, R. J. & Davidson, E. H. in *Nucleic Acid Hybridization — A Practical Approach* (eds Hames, B. D. & Higgins, S. J.) 3–15 (IRL, Oxford, 1985).
17. Young, B. D. & Anderson, M. L. M. in *Nucleic Acid Hybridization — A Practical Approach* (eds Hames, B. D. & Higgins, S. J.) 47–71 (IRL, Oxford, 1985).
18. Haukanes, B. -I. & Kvam, C. *Bio/technology* **11**, 60–63 (1993).
19. Adams, M. D. *et al. Science* **355**, 632–634 (1992).
20. Taylor, S. A. M. *et al. Nature Genet.* **2**, 223–227 (1992).
21. Snell, R. G. *et al. Hum. molec. Genet.* (in the press).

On the biotech beat

Featured this week — an immunoglobulin binding protein, a chromosome enumeration system, a capillary electrophoresis system and columns, software, CD-ROMs and new product catalogues.

Two of the most thoroughly characterized members of the 'chaperonin' subgroup of molecular chaperones are the *Escherichia coli* proteins GroEL and GroES. Incubation of the inactive, denatured forms of several proteins, including carboxylase, citrate synthase, rhodanase, and a sigma subunit of bacterial RNA polymerase, with GroEL and GroES in the presence of Mg-ATP, results in *in vitro* reconstitution of their active forms (*Nature* **355**, 33–45; 1992). Epicentre Technologies now offers the **chaperonin proteins**, GroEL and GroES, together with another molecular chaperonin, DnaK protein, as tools for studying protein folding and assembly *(Reader Service No. 101)*. The company says that the preparations of these proteins are rigorously tested for quality, both for the absence of contaminants and for functional activity. GroEL must demonstrate ATPase activity, and GroES, in 2.5-fold molar excess over GroEL, must suppress greater than 90 per cent of the ATPase activity of GroEL. In addition, Epicentre Technologies says that its preparations of GroEL and GroES are tested for their ability to renature inactive citrate synthase in the presence of Mg-ATP to an enzymatically active form.

KappaLock is a new **immunoglobulin binding protein** from Aaston that binds to the kappa light chains of antibody types and, as such, should be useful for the purification of monoclonal antibodies and Fab fragments of antibodies *(Reader Service No. 102)*. As binding is to the light chain at the Fab region of the antibody, the binding activity of the antibody is minimally affected. KappaLock is a 32-kD protein that is derived from a larger natural variant of a *Streptococcal* protein L. Natural protein L is a complex protein that binds albumin, cell wall components and lipids, in addition to kappa light chains. Aaston says that the specificity of KappaLock has been improved over that of the naturally occurring protein L through the removal of the albumin binding and cell wall binding regions. It has four light chain binding regions and has been shown to interact with all subgroups of polyclonal antibodies from human, mouse, rabbit, pig and goat sources. As different antibody groups such as IgG, IgD and IgM have different numbers of kappa and lambda light chains, these different antibody groups can be expected to bind KappaLock with different strengths. This should allow for simple fractionation of antibody types with immobilized KappaLock.

The SpectrumCEP **direct-labelled chromosome enumeration system** from Imagenetics can be used for a wide range of applications in molecular cytogenetic research that uses fluorescence *in situ* hy-