Preface

Several developmental and historical threads are woven and displayed in these two volumes of *Bacterial Artificial Chromosomes*, the first on *Library Construction, Physical Mapping, and Sequencing*, and the second on *Functional Studies*. The use of large-insert clone libraries is the unifying feature, with many diverse contributions. The editors have had quite distinct roles. Shaying Zhao has managed several BAC end-sequencing projects. Marvin Stodolsky during 1970–1980 contributed to the elucidation of the natural bacteriophage/prophage P1 vector system. Later, he became a member of the Genome Task Group of the Department of Energy (DOE), through which support flowed for most clone library resources of the Human Genome Program (HGP). Some important historical contributions are not represented in this volume. This preface in part serves to mention these contributions and also briefly surveys historical developments.

Leon Rosner (deceased) contributed substantially in developing a PAC library for drosophila that utilized a PI virion-based encapsidation and transfection process. This library served prominently in the Drosophila Genome Project collaboration. PACs proved easy to purify so that they substantially replaced the YACs used earlier. Much of the early automation for massive clone picking and processing was developed at the collaborating Lawrence Berkeley National Laboratory. However, the P1 virion encapsidation system itself was too fastidious, and P1 virion-based methods did not gain popularity in other genome projects.

Improving clone libraries was an early core constituent of the DOE genome efforts. Cosmid-based libraries with progressively larger inserts were developed within the DOE National Laboratories Gene Library Program. But quality control tests by P. Youdarian indicated that perhaps 25% of human insert cosmids had some instability, possible owing to the multicopy property of the system. Both for this reason and to provide for larger inserts of cloned DNAs, DOE supported the investigation of several new cloning systems. Of the eukaryotic host systems, the Epstein-Barr virus-based system from Jean-M. Vos (deceased) was quite successful. But the added costs and care needed for use of eukaryotic cells precluded its wide adoption in HGP production efforts.

Among the bacterial host systems, two developed in the lab of Melvin Simon provided pivotal service. Ung-Jin Kim developed fosmids. They are maintained as single copy replicons and utilize the reliable encapsidation pro-

cesses developed for cosmids. Fosmids proved to be highly stable. BACs were developed by Hiroaki Shizuya. They were introduced into *E. coli* by electroporation and stability was generally good, though there is an unstable BAC minority ***(1)***. This BAC resource emerged after the chimeric properties of the large YACs was recognized. BACs were thus initially viewed with appropriate suspicion. But at the nearby Cedar-Sinai Medical Center, J. R. Korenberg and X.-N. Chen implemented a very efficient FISH analysis. They found that chimerism in any of the BACs was at worst around 5% and the BACs were well distributed across all the chromosomes. Overall human genome coverage was estimated in the 98–99% range, with even centromeric and near telomeric regions represented.

Two examples of this good coverage soon emerged. Isolation of the BRAC1 breast cancer gene had failed with all other clone resources. But when Simon's group was provided with a short cDNA probe, they soon returned a BAC clone carrying an intact BRAC1 gene. Pieter de Jong had acquired the technology of cloning long DNA inserts from the Simon lab, initially using a PAC vector and electroporation. After a first successful library, DOE advised de Jong to broadly distribute this new PAC resource. Shortly thereafter, he assembled a 900 kb contig for the candidate region of the BRAC2 gene. The subsequent DNA sequence generated at the Washington University then revealed the BRAC2 gene. These striking easy successes stimulated broad usage of the BAC and PAC resources.

End sequences of clonal inserts have been used to facilitate contig building since the 1980s in small-scale mapping and sequencing projects. Glen Evans for example was piloting with DOE support a "mapping plus sequencing" strategy on chromosome 11, before the BAC resources were available. Once a covering set of cloned DNAs with sequenced ends is generated, clones to efficiently extend existing sequence contigs can be chosen. As the need for high throughput genome sequencing to meet HGP timelines became imminent, only a few human chromosomes had adequate contig coverage. L. Hood, H. Smith, and C. Venter proposed a Sequence Tag Connector (STC) strategy to alleviate this bottleneck. With application to the entire human genome, concurrent BAC contig building and sequencing would be implemented.

The DOE instituted a fast track review of two STC applications in the spring of 1996 ***(2)***. One was from a team comprised of L. Hood, H. Smith, and C. Venter, and the second from a team comprised of G. Evans, P. de Jong, and J. R. Korenberg. A panel with broad international representation reviewed applications from two teams. Interested colleagues from the NIH and NSF were observers. Although the overall STC concept was reviewed favorably,

initial pilot implementations to better define the economics were recommended. A year later, progress was reviewed and a DOE commitment to a full scale implementation was made. At the request of the NIH, the DOE later increased support to accelerate a 20-fold coverage of the genome.

The STC data set has had multiple beneficial roles. Sequence Tag Sites (STSs) were defined within the STC sequences and used to enrich the Radiation Hybrid (RH) maps of the genome, thus providing for an early correspondence of the RH maps and the maturing contig maps. Validity constraints on sequence contigs were provided by the spanning BACs. Most broadly, the STC resource had an indispensable role for both the strategies of Celera Genomics Inc., and the international public sector collaboration, in the rapid generation of draft sequences of the human genome. The STC strategy is now implemented in many current genomic projects, including the NIH sponsored mouse and rat genome programs.

Bacterial Artificial Chromosomes in its two volumes provides a comprehensive collection of the protocols and resources developed for BACs in recent years. These two volumes collectively cover four topics about BACs: (1) library construction, (2) physical mapping, (3) sequencing, and (4) functional studies. The laboratory protocols follow the successful *Methods in Molecular Biology*™ series format by containing a clear sequence of steps followed by extensive troubleshooting notes. The protocols cover simple techniques such as BAC DNA purification to such complex procedures as BAC transgenic mouse generation. Both routine and novel methodologies are presented. Besides protocols, chapter topics include scientific reviews, software tools, database resources, genome sequencing strategies, and case studies. The books should be useful to those with a wide range of expertise from starting graduate students to senior investigators. We hope our books will provide useful protocols and resources to a wide variety of researchers, including genome sequencers, geneticists, molecular biologists, and biochemists studying the structure and function of the genomes or specific genes.

We would like to thank all those involved in the preparation of this volume, our colleagues, and friends for helpful suggestions, and Professor John Walker, the series editor, for his advice, help and encouragement.

Shaying Zhao
Marvin Stodolsky

References

1. http://www.ornl.gov/meetings/ecr2/index.html
2. http://www.ornl.gov/meetings/bacpac/body.html

2

Exon Trapping for Positional Cloning and Fingerprinting

Scott E. Wenderfer and John J. Monaco

1. Introduction

Positional cloning involves the genetic, physical, and transcript mapping of specific parts of a genome ***(1)***. Linkage analysis can map specific activities, or phenotypes, to a quantitative trait locus (QTL), a genomic region no smaller than 1 centiMorgan (cM) or megabase (Mb) in length. Physical mapping can then provide a map of higher resolution. Physical maps are constructed from clones identified by screening genomic libraries. Genomic clones can be characterized by fingerprinting and ordered to create a contig, a contiguous array of overlapping clones. Transcript identification from the clones in the contig results in a map of genes within the physical map. Finally, expressional and functional studies must be performed to verify gene content.

Bacterial artificial chromosomes (BACs) and P1 artificial chromosomes (PACs), both based on *Escherichia coli* (*E. coli*) and its single-copy plasmid F factor, can maintain inserts of 100–300 kilobases (kb). Their stability and relative ease of isolation have made them the vectors of choice for the development of physical maps. Once BAC clones are obtained, exon trapping can be performed as a method of transcript selection even before characterization of the contig is complete. Trapped exons are useful reagents for expressional and functional studies as well as physical mapping of BAC clones to form the completed contig.

Exon trapping was first used by Apel and Roth ***(2)*** and popularized by Buckler and Housman ***(3)***. A commercially available vector, pSPL3 ***(4)***, has been used in multiple positional cloning endeavors ***(5–8)***. Exon trapping relies on the

From: *Methods in Molecular Biology, vol. 256:*
Bacterial Artificial Chromosomes, Volume 2: Functional Studies
Edited by: S. Zhao and M. Stodolsky © Humana Press Inc., Totowa, NJ

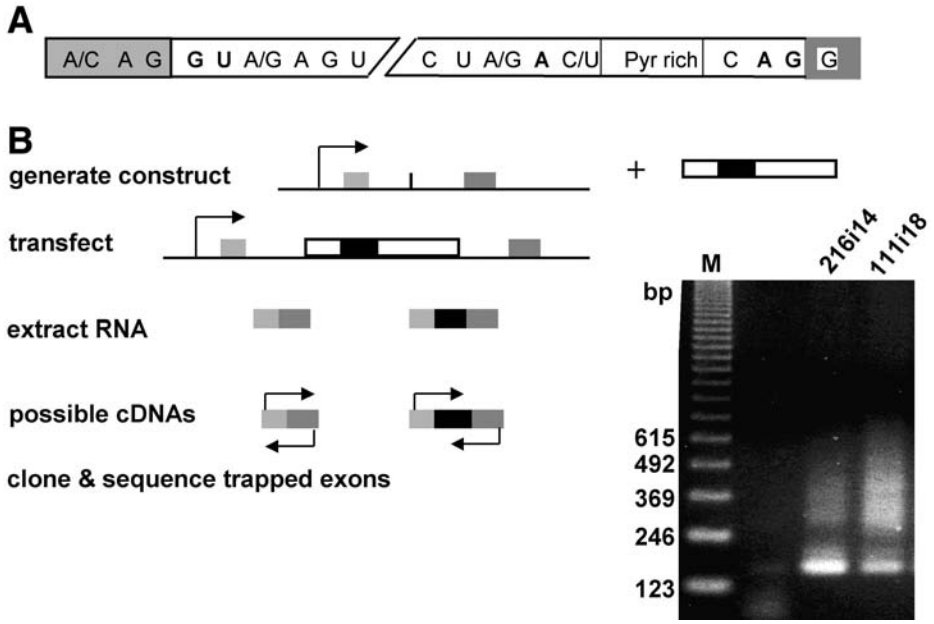

Fig. 1. **(A)** Exon splicing is conserved in eukaryotes. The sequences at the splice junctions are conserved. The gray box represents the 5′ exon and the checkered box represents the 3′ exon. The white box represents the intron. The bold bases indicate the 3′ splice acceptor, the branch point A, and the 5′ splice donor from left to right. **(B)** Because splicing is conserved, a genomic fragment (white bar) containing an exon (black box) from any species can be inserted within the intron of an expression construct for exon trapping. COS7 cells are transfected with the construct and 48 h later RNA is collected. The expressed recombinant mRNA can be isolated by RT-PCR using primers for the upstream and downstream exon of the expression construct. Genomic fragments lacking an exon would allow the upstream and downstream exons of the expression construct to splice together, resulting in a smaller RT-PCR product (the 177 bp band). We screened BAC clones by shotgun cloning small fragments into the intron of the HIV tat gene behind an SV40 early promoter. The RT-PCR products from two exon trapping experiments are shown.

conservation of sequence at intron–exon boundaries in all eukaryotic species (*see* **Note 1**). By cloning a genomic fragment into the intron of an expression vector, exons encoded in the genomic fragment will be spliced into the transcript encoded on the expression vector (*see* **Fig. 1**). Reverse transcriptase polymerase chain reaction (RT-PCR) using primers specific for the transcript on the expression vector will provide a product for analysis by electrophoresis and sequencing.

Because the expression vector utilizes its own exogenous promotor, exon trapping is independent of transcript abundance and tissue expression. Moreover, exon trapping provides rapid sequence availability. It has proven to be a very sensitive method for transcript identification ***(9,10)*** (*see* **Note 2**). By pooling subclones via shotgun cloning of cosmids, BACs, or yeast artificial chromosomes (YACs) into the pSPL3 vector, 30 kb–3 Mb can be screened in a single experiment.

Disadvantages include dependence on introns, splice donor and acceptor sites. False negatives are caused by missing genes with only one or two exons, interrupting exons by cloning into the expression vector, and possibly by not meeting unidentified splicing requirements. False positives are caused by cryptic splice sites ***(11)***, exon skipping ***(12)***, and pseudogenes.

No one method for transcript identification has become the stand-alone method for positional cloning. Genomic sequence analysis, when sequence is available, should be the primary tool for identification of genes within a genomic region of interest. Bulk sequencing provides a template for computer selection of gene candidates via long open reading frames (ORFs), sequence homology, or motif identification. Gene Recognition and Assembly Internet Link (GRAIL) analysis can be performed manually at a rate of 100,000 kb per person-hour ***(13)***. PCR primer pairs can be made for each set of GRAIL exon clusters. Alternatively, predicted GRAIL exons may be represented in the expressed sequence tag (EST) database, a collection of sequences obtained from clones randomly selected from cDNA libraries encompassing a wide range of tissues or cell types. If an EST exists, corresponding cDNA clones can be purchased from the IMAGE consortium ***(14)***. Motif and ORF searching does suffer from a lack of specificity and sensitivity and tend to be both time consuming and software/hardware dependent. Exon trapping is an excellent tool for verification of genes predicted in the sequence, as well as for identification of genes missed by computational techniques. A cluster of trapped exons likely encodes a functional gene product if several correspond to exons also predicted by GRAIL and together they encode a long ORF.

When no genomic sequence is available, exon trapping is the method of choice for initially identifying genes. Not only are new genes identified and known genes mapped, but also trapped exons, bona fide or false positives, become markers for the generation of a physical map. Southern or colony blots made from BAC clones can be hybridized with exon probes to map them to specific locations on individual BACs, or to BACs in a contig. Trapped exon probes an also be used to screen further genomic BAC libraries. In our experience, more than 100 markers were generated for every 1 Mb region, resulting in a marker density of one per 10 kb. Therefore, the number of markers generated during a completed exon trapping study will be sufficient for genome

sequencing centers to begin obtaining and aligning sequence information in this contig ***(15)***.

Most other strategies for positional cloning use "expression-dependent" techniques. Direct selection is the selection of transcribed sequences from a library of expressed cDNAs using solution hybridization with labeled genomic clones ***(16,17)***. A similar technique, cDNA selection, selects transcribed sequences by hybridization screening of blotted genomic clones with labeled cDNA libraries ***(18–20)***. Transcript selection techniques depend on the knowledge of mRNA distribution and abundance in different tissues. They are difficult to perform with BAC clones, as most will contain regions of repetitive sequence that must be blocked with competing unlabeled DNA. Performed together with exon trapping, they have been proven complimentary.

Exon trapping is not intended for extremely high-throughput gene identification or mapping. Whole genome sequencing and large-scale sequencing of cDNA library clones together have been the most efficient high-throughput gene identification methodology. EST databases contain a large number of gene markers that can be used for expressional profiling by RT-PCR or DNA chip technology. Radiation hybrid mapping of these EST clones has become a high-throughput technology for gene mapping ***(21)***. However, EST databases tend to be overrepresented with genes expressed in high abundance. Researchers interested in a genomic region in a species that has been the subject of high-throughput analyses, such as *Homo sapiens*, may wish to obtain BAC clones and use exon trapping as a complimentary method.

Once trapped, exon clones can be used for expression analysis. Querying sequences of candidate exons against Genbank's EST dataset can be used to identify multiple tissues where the gene has been previously identified by sequencing of cDNA libraries. Hybridization to northern blots with total RNA from brain, heart, kidney, liver, lung, skeletal muscle, spleen, and thymus will give a general screen for expression appropriate for all candidate exons. Hybridization to blots with total RNA from cell lines can provide information on constitutive and inducible expression in different cell types. Alternatively, exon sequences can be used to generate a DNA chip for expressional profiling, allowing all exons to be tested in a single experiment.

2. Materials

2.1. Subclone BAC DNA into pSPL3 Exon Trapping Vector

1. Appropriate BAC or PAC clones may be purchased (Incyte Genomics, St. Louis, MO; Roswell Park Cancer Institute, Buffalo, NY).
2. BAC DNA should be isolated from 500 mL bacterial cultures by alkaline lysis. Lysates are passed through Nucleobond filters onto AX-500 columns (Clontech,

Palo Alto, CA), eluted, then precipitated with isopropanol, washed with ethanol, and reconstituted in 100 µL distilled H_2O. Aliquots of 5 µL of separate *EcoRI* and *NotI* digests can be analyzed by electrophoresis on agarose gels. Contamination of preps with bacterial DNA does not preclude their use, but may increase the false-positive rate.

3. *BamHI*, *BglII*, *DraI*, *EcoRV*, *EcoRI*, *NotI*, *HincI*, *NotI*, *PvuII*, and *T4* DNA ligase.
4. pSPL3 plasmid may be purchased as part of the exon amplification kit (Gibco-BRL, Gaithersburg, MD). Plasmid preps can be performed using alkaline lysis kits from Qiagen (Valencia, CA).
5. *E.coli* strain DH10b electromax cells can be purchased from Gibco BRL.
6. GenePulser bacterial cell electroporator and cuvets (Bio-Rad, Richmond, CA).
7. Luria Bertani broth with 100 µg/mL ampicillin (LB-amp).
8. Routine gels can be prepared from electrophoresis grade agarose (Bio-Rad).
9. DNA can be purified from low-melt agarose gel slices using the MP kit from U.S. Bioclean (Cleveland, OH).

2.2. Transient Transfections

1. COS-7 green monkey kidney cells may be obtained from ATCC (Rockville, MD) and maintained in 10 mL Dulbecco's modified Eagle's media (DMEM) with 10% fetal bovine serum (FBS) and 2 m*M* sodium pyruvate (GibcoBRL) at 37°C, 5–10% CO2. All manipulation should be performed in a hood under sterile conditions.
2. Phosphate buffered saline (GibcoBRL), stored at 4°C.
3. GenePulser mammalian cell electroporator and cuvets (Bio-Rad).

2.3. Exon Trapping

1. Superscript II RT, *Bst*XI, RNAse H, Taq DNA polymerase, Trizol reagent for total RNA isolation, uracil DNA glycosylase (UDG), prelinearized pAMP10 vector, and DH10b max efficiency competent cells.
2. Oligo SA2 sequence: ATC TCA GTG GTA TTT GTG AGC.
3. First strand buffer contains a final concentration of 50 m*M* Tris-HCl pH 8.3, 75 m*M* KCl, 3 m*M* $MgCl_2$, 10 m*M* dithiothreitol (DTT), and 0.5 m*M* dNTP mix.
4. PCR buffer contains a final concentration of 10 m*M* Tris pH 9.0, 50 m*M* KCl, 1.5 m*M* $MgCl_2$, and 0.2 m*M* dNTP mix.
5. Oligo SD6 sequence: TCT GAG TCA CCT GGA CAA CC.
6. Oligo dUSD2 sequence: ATA GAA TTC GTG AAC TGC ACT GTG ACA AGC TGC.
7. Oligo dUSA4 sequence: ATA GAA TTC CAC CTG AGG AGT GAA TTG GTC G.
8. RT reaction and PCR can be performed in a DNA thermocycler 480 (Perkin Elmer–Applied Biosystems, Norwalk, CT).
9. Water for manipulation and storage of RNA should be treated with 0.1% diethyl pyrocarbonate to remove RNAses and then autoclaved. When working with RNA, change gloves often and use only reagents prepared with RNAse-free water.

2.4. Screening Trapped Exons to Exclude False Positives and Previously Sequenced Exon Clones

1. LB-amp broth.
2. Sterile 96-well microtiter plates with lids (Fisher).
3. 96-pin replicator may be purchased from Fisher (Pittsburgh, PA), should be stored in 95% ethanol bath, and can be flame sterilized before and after each bacterial colony transfer.
4. Appropriately sized rectangular agar plates can be made by pouring molten LB agar into the lid of a standard 96-well microarray plate and solidifying overnight at 4°C.
5. Magnabond 0.45-µm nylon filters (Micron Separations Inc., Westborough, MA).
6. Prehyb solution contains a final concentration of 1 *M* NaCl, 1% sodium dodecyl sulfate (SDS), 10% dextran sulfate, and 100 µg/mL denatured salmon sperm DNA.
7. *Acc*I, *Ava*I, *Bgl*II, *Sal*I, T4 DNA kinase and exonuclease-free Klenow fragment.
8. T4 forward reaction buffer contains a final concentration of 70 m*M* Tris-HCL pH 7.6, 10 m*M* MgCl2, 100 m*M* KCl, and 1 m*M* 2-mercaptoethanol.
9. DNA replication buffer contains a final concentration of 0.2 *M* HEPES, 50 m*M* Tris-HCL pH 6.8, 5 m*M* $MgCl_2$, 10 m*M* 2-mercaptoethanol, 0.4 mg/mL bovine serum albumin (BSA), 10 µ*M* dATP, 10 µ*M* dGTP, 10 µ*M* dTTP, and 5 OD_{260} U/mL random hexamers mix.
10. [γ-^{32}P]dATP and [α-^{32}P]dATP. Proper shielding should be used when handling all solutions containing ^{32}P.
11. $pSPL3_{VV}$ oligo sequence: CGA CCC AGC A|AC CTG GAG AT.
12. $pSPL3_{1021}$ oligo sequence: AGC TCG AGC GGC CGC TGC AG.
13. $pSPL3_{1171}$ oligo sequence: AGA CCC CAA CCC ACA AGA AG.
14. $pSPL3_{1056}$ oligo sequence: GTG ATC CCG TAC CTG TGT GG.
15. pPSL3 intron probe can be prepared in bulk by double digest of pSPL3 vector with *Ava*I and *Sal*I. The 335 bp and 2086 bp bands can be isolated by agarose gel electrophoresis and purified using the U.S. Bioclean MP kit. It can be stored at –20°C, thawed on ice, and refrozen multiple times.
16. Previously sequenced exon clone (PSEC) probes can be prepared from double digests of trapped exons in pAMP10 using 5 U each of *Acc*I and *Bgl*II. Vector bands of 4 kb and either 50 or 109 bp (depending on direction in which trapped exon is cloned into pAMP10) should be avoided when probes are isolated from gel slices. PSEC probes can be stored at –20°C, thawed on ice, and refrozen multiple times.
17. Probe purification columns can be made by filling disposable chromatography columns with either Sephadex G-25 (for oligos) or G-50 (for longer single-stranded DNA probes) and spinning out buffer into a microfuge tube.
18. 2X SSC/SDS contains a final concentration of 0.3 *M* NaCl, 30 m*M* sodium citrate, and 0.5% SDS. 0.2X SSC/SDS contains 0.03 *M* NaCl, 3 m*M* sodium citrate, and 0.5% SDS.
19. X-OMAT AR film (Eastman Kodak Company, Rochester, NY).
20. Phosphor screen and phosphorimager (Molecular Dynamics (Amersham Pharmacia Biotech, Piscataway, NJ).

2.5. Size Selection of Trapped Exons for Sequencing of Unique Clones

1. LB-amp broth.
2. Sterile 96-well microtiter plates with lids.
3. PCR can be performed for sets of 96 samples using Gene Amp PCR system 9700. (Perkin Elmer–Applied Biosystems).
4. PCR buffer.
5. Individual bacterial clones may be transferred from 96-well plate via toothpicks, sterilized by autoclaving in tin foil, or by flame sterilized 96-pin replicator.
6. *Hind*III and *Pst*I.
7. Sequencing primers dUSA4, dUSD2.

3. Methods

3.1. Subclone BAC DNA into pSPL3 Exon Trapping Vector

1. Isolate genomic BAC clone (*see* **Note 3**).
2. Set up *Dra*I, *Eco*RV, and *Hinc*II digests for each BAC clone individually in three separate tubes (*see* **Note 4**). A total of 10 U restriction enzyme will digest 5 µg in 8 h.
3. Linearize pSPL3 exon trapping vector by digesting with the appropriate restriction enzyme and gel-purify.
4. Subclone each digest individually into linearized pSPL3 with 20,000 U T4 DNA ligase for 1 h at 42°C and transform DH10b bacterial cells by electroporation at 1.8 kV, 25 µF, 200 Ω (*see* **Note 5**).
5. Grow transformants overnight in 50 mL LB-amp broth, isolate DNA from shotgun subclones and test heterogeneity by running a *Pvu*II digest on a 1% agarose gel.

3.2. Transient Transfections

1. Plate 2×10^6 COS7 cells / 75 mm^2 dish and preincubate 24 h.
2. Harvest cells by centrifugation and wash twice in 5 mL ice cold PBS.
3. Resuspend to 4×10^6 cells/mL in ice-cold PBS and transfer 0.7 mL aliquots into labeled electroporation cuvets.
4. Add 15 µg supercoiled plasmid DNA, mix, and incubate on ice for 5 min.
5. Electroporate at a voltage of 350 V and a capacitance of 50 µF.
6. Incubate on ice 5–10 min then dilute cells 20-fold in 14 mL DMEM/FBS.
7. Plate transfected cells in T25 flasks and incubate 48 h (2 generation times).

3.3. Exon Trapping

1. Isolate total RNA using Chomczynski-based method. Resuspend total RNA yield from each T25 flask of cells in 100 µL RNAse-free H_2O and store RNA at –80°C. Run 3 µg RNA on a 1% agarose gel at 50 V to check purity (*see* **Note 6**).
2. Perform reverse transcription reaction on 3 µg total RNA (final concentration = 0.15 µg/mL) with 200 U Superscript II RT and 1 µ*M* SA2 oligo in 20 µL 1st strand buffer for 30 min at 42°C.

3. Preincubate cDNA 5 min at 55°C, then treat with 2 U RNAse H for 10 min, store at 4°C.
4. Perform PCR on 5 µL cDNA (approx 1.2µg) with 2.5 U Taq DNA polymerase and 1 µ*M* each oligos SA2 and SD6 in 40 µL PCR buffer for a total of six cycles (each cycle: 1 min denaturation at 94°C, 1 min annealing at 60°C, and 5 min extension at 72°C).
5. Continue final extension an additional 10 min at 72°C.
6. Treat PCR product with 20 U *Bst*XI restriction endonuclease at least 16 h at 55°C (*see* **Note 7**).
7. Add an additional 4 U *Bst*XI enzyme and treat for another 2 h at 55°C.
8. Perform secondary PCR on 5 µL *Bst*XI digest with 2.5 U Taq DNA polymerase and 0.8 µ*M* each oligo dUSA4 and dUSD2 in 40 µL PCR buffer for a total of 30 cycles (each cycle: 1 min denaturation at 94°C, 1 min annealing at 60°C, and 3 min extension at 72°C).
9. Run 9 µL secondary PCR product on >2% agarose gel to check heterogeneity. See **Fig. 1** for the appearance of a satisfactory exon trapping experiment.
10. Clone 2µL (approx 100 ng) heterogeneous exon mixture into pAMP10 vector using 1 U UDG in 10 µL.
11. Transform 3 µL UDG shotgun subclones into 50 µL DH10b max efficiency competent cells by heat shock, 42°C for 40 s, plate 20% of cells on each of two LB amp plates and grow >16h.

3.4. Screening Trapped Exons to Exclude False Positives and Previously Sequenced Exon Clones

1. Inoculate 200µL LB-amp broth per well with 286 CFU from each exon-trapping reaction in 96 well plates (three 96-well plates/BAC clone).
2. For each 96-well plate, inoculate one well with a bacterial clone transformed with pSPL3 vector alone (positive control) and a second well with a UDG clone from an exon trapping experiment where no genomic DNA was subcloned (negative control), and grow transformants >16 h.
3. Make three sets of colony dot blots by transferring 96 UDG clones *en mass* with 96-pin replicator to a nylon filter sterilely placed over a rectangular agar plate. Grow colonies >16 h, denature and wash away bacterial debris, and crosslink DNA to nylon at 120,000 µJ/cm^2.
4. Prehybridize for >1 h at 50°C in hybridization bottle.
5. Label 100 ng each of $pSPL3_{VV}$, $pSPL3_{1021}$, $pSPL3_{1171}$, and $pSPL3_{1056}$ oligos together with 75 µCi [γ-^{32}P]dATP and 10 U T4 kinase in 20 µL forward reaction buffer and purify with Sephadex G-25 column (*see* **Note 8**).
6. Add 1×10^7 CPM of labeled four pSPL3 oligo mixture for each milliliter prehybridized solution and hybridize 1 set of colony blots >8 h at 50°C.
7. Washing unbound oligos from blot with 2X SSC/SDS buffer twice at room temperature then four times at 60°C routinely results in appearance of specific signal on film within 16 h or on phosphor screen within 1 h.

8. Hybridize the second set of colony blots with pSPL3 intron, labeled with 75 µCi [α-^{32}P]dATP and 3 U exonuclease-free Klenow fragment in 50 µL DNA replication buffer and purify with Sephadex G-50 column.
9. Hybridize the third set of blots with previously sequenced exon clone (PSEC) mix, labeled with 75 µCi [α-^{32}P]dATP and 3 U exonuclease-free Klenow fragment in 25 µL DNA replication buffer and purify with Sephadex G-50 column (*see* **Note 9**).
10. Washing unbound single stranded DNA probe from blot twice with 2X SSC/SDS buffer, then twice with 0.2X SSC/SDS buffer at 65°C routinely results in appearance of specific signal on film within 16 h, or on phosphor screen within 1 h.

3.5. Size Selection of Trapped Exons for Analysis of Unique Clones

1. Grow bacterial clones transformed with "unsequenced, true positive" candidate exons in LB-amp broth in 96-well plates >16 h.
2. Using a 96-pin replicator, transfer bacterial clones to thin walled PCR tubes containing 40 µL PCR buffer. Colony PCR performed with 2.5 U Taq DNA polymerase and 0.8 µ*M* each of oligos dUSA4 and dUSD2 for a total of 30 cycles (each cycle: 1 min denaturation at 94°C, 1 min annealing at 60°C, and 3 min extension at 72°C).
3. Size select candidate exons by running on a 3% agarose gel (*see* **Note 10**).
4. Grow bacteria transformed with unique clones in LB-amp broth >16 h, and isolate DNA by alkaline lysis.
5. Test size selection by running *Hind*III/*Pst*I double digest on 3% agarose gel.
6. Sequence unique exons from plasmid preps using either oligo dUSA4 or dUSD2. If sequence obtained does not overlap, design additional primers from deduced sequence and repeat until full-length sequence is obtained (*see* **Note 11**).

4. Notes

1. Exon trapping detects exons encoded within the genome. The definition of an exon is well understood. Consensus sequences are present at both splice acceptor and splice donor sites ***(22)***. Small nuclear RNA molecules hybridize to these consensus sequences in the messenger RNA, targeting the splicing machinery to excise the intervening sequence, or introns. Cryptic splice sites exist in the genome, defined as random sequence that mimics either a splice acceptor site or a splice donor site. The chance that a cryptic splice donor and a cryptic splice acceptor would be located close enough together in the genome to cause a false positive exon to be trapped is presumably rare, but the actual number is not known. Our data suggest that the specificity of exon trapping is high. At least 84% of clones have sequences with open reading frames and are expressed in vivo ***(8)***. To help determine the specificity of exon trapping, one can analyze the flanking intron sequence to identify consensus splice sites. Because the sequences at the ends of exons are less conserved, we were unable to analyze the validity of

trapped exons by their sequence alone. Sequencing flanking intron sequence off the BAC clone for every trapped exon is a laborious task, not recommended routinely. However, one BAC clone used in our exon trapping experiments was also sequenced ***(23)***. We did check for the presence of consensus splice sites in introns flanking 22 exons trapped from this BAC clone. Sixteen were exons from genes with published sequence. All 16 are flanked in the genome by consensus splice sites, but two used different splice sites from those published. Five trapped exon clones have open reading frames encoding previously unpublished sequence, and four of the five are flanked by consensus splice sites. The fifth is flanked only by a 5′ splice donor. Only one exon was trapped that lacked an open reading frame in any of the three reading frames, but it too is flanked by consensus splice sites. Therefore, the specificity of the splicing mechanism in our exon trapping experiments appears to be identical to the specificity of the endogenous splice machinery.

2. Our data suggest that exon trapping is 73% sensitive for transcript identification, when several hundred trapped exons are characterized per PAC or BAC clone ***(8)***.
3. Sixfold redundant libraries will result in approximately 50 clones per one Mb. Up to six previously mapped genes or EST clones can be used as probes to screen a genomic BAC library in a single hybridization. A minimum contig of 10 clones should then be shotgun cloned into pSPL3 for exon trapping. With sequence information to aid in development of a contig, this can all be performed in less than a month. Screening 200 exons from each BAC or PAC clone tested should take two weeks, and up to 1000 additional clones can be characterized by PSEC screens in another two weeks.
4. Use of three separate restriction enzyme digests combined prior to ligation to vector minimizes the chance of missing an exon that happens to contain a restriction site within its sequence. An alternative method is to use a *Bam*HI and *Bgl*II double digest along with a *Sau*3AI partial digest in two separate tubes.
5. Transformation of competent cells by electroporation is much more effective than heat shock transformation for bacteria. In our experience, without electroporation of the BAC subclones, the sensitivity of identifying known genes using exon trapping decreased 10-fold.
6. Protocol for using Trizol reagent available from GibcoBRL. Yield of RNA prep is 5–7 µg per T25 flask (approx 10^6 cells). Using a spectrophotometer, the $A_{260/280}$ should be between 1.6–1.8 (less suggests phenol contamination or incomplete dissolution). Gel should show sharp ribosomal bands with the intensity of the 28S twice that of 18S. If the 5S band is as intense as the band at 18S, there is too much degradation to efficiently continue this protocol.
7. The success of the *Bst*XI digestion is critical for the elimination of false negatives. A short 177bp cDNA composed of only pSPS3 vector sequence will predominate unless *Bst*XI digestion is complete. Fresh GibcoBRL enzyme was the only formulation potent enough to approach 100% digestion using this protocol.
8. Cryptic splice sites within the pSPL3 intron were responsible for several false positives, from 10 to 50% of all products of an exon trapping experiment. Screening of trapped exons with four oligos and the entire pSPL3 intron removed 95%

of these false positives from further consideration. Three oligos are named by the location of the complimentary sequence on the pSPL3 vector. The pSPL3 intron sequence runs from 699 to 3094. The fourth oligo ($pSPL3_{v\text{-}v}$) contains sequence complimentary to the exons of the pSPL3 vector after being spliced together (splice junction indicated by a vertical bar in the sequence in the methods section). If the *Bst*XI digestion is incomplete and some pAMP10 clones without trapped exons remain, this fourth oligo will identify them.

9. A difficulty encountered with exon trapping was differential representation of trapped exons within the total pool. Some exons were present at proportions of 1:10 or even 1:4 when hundreds of exons were analyzed from a 100-kb BAC clone. Other exons required characterization of several hundred trapped exons from a particular BAC clone before a single copy was identified. The selection of smaller clones during PCR amplification or cloning does not explain the differences in abundance. Trapped exons from each BAC should be characterized hundreds at a time, first by size selection and sequencing, then by PSEC (spell out) screens. PSECs were isolated as probes, labeled individually and pooled in order to screen additional batches of cloned exons by hybridization. Hundreds of trapped exon clones could be easily screened with all PSECs after generating duplicate colony blots by transfer of bacterial clones from microtiter plates using a 96-pin replicator. Screening 200–300 exons from each exon trapping experiment is recommended. However, if known genes are not identified after characterizing 300, chances are very low that it will be identified in that experiment.

   Exon trapping yield varies between different species and between different regions on the same chromosomes, depending on the gene density. Yield is measured by the following equation:

$$\text{Yield} = \frac{\text{kb DNA screened}}{\text{exons trapped}}$$

   Each exon trapping experiment involves shotgun cloning multiple digests of the same BAC or PAC clone into the pSPL3 trapping vector. Additional experiments may be performed using different restriction endonucleases to generate inserts for shotgun cloning. Running a second experiment for the same BAC clone often doubles the number of exons trapped, but in our hands a third experiment does not result in many new exon clones. Exon trapping of a BAC was considered complete when >95% of trapped exons in a screen were positive for a PSEC. At that point, identification of missed genes by a complimentary "transcript identification" method (sequence analysis, zoo analysis, or expression analysis) would be warranted over screening more trapped exons.

10. Trappable exons have ranged in size from 49 to 465 bp, similar to the range observed for all exons in the genome. Electrophoresis of DNA in this size range is best visualized on 3% agarose gels. Estimating sizes then rerunning samples in order from smallest to largest can verify sizes and is often helpful. Isolation of DNA from 3% agarose gel slices to obtain PSEC probes is possible using the U.S. Bioclean MP kit.

11. Double-stranded sequence was not routinely obtained. Because neither 5′ nor 3′ exons can be trapped by this method, open reading frames are usually a property of true positives identified by exon trapping. An additional method for screening exon trapping products for true positives is zoo blotting. Zoo blotting involves the hybridization of DNA or cDNA from one species with genomic DNA or RNA from various related or divergent species. In one study, 85% of exon trapping products from human DNA demonstrated cross-hybridization to primate sequences, and 56% cross-hybridized to other mammalian sequences ***(9)***. Finally, true positives can be verified by identifying transcripts by Northern blot or by screening cDNA libraries.

    Unfortunately, one drawback of transcript identification is that not all transcripts encode functional gene products. EST databases exemplify this pitfall of transcript identification. An enormous number of cDNA clones represented in the EST database encode repetitive sequence. Sometimes this is owing to isolation of a pre-mRNA in which an intron containing a repeat element has not been spliced out. In other cases, the repetitive element is presumably expressed because of its own LTR, a *cis*-acting factor that drives transcription of the repeat sequence. The importance of repetitive transcripts in health and disease is debatable, but removal of EST sequences containing repeats is straightforward for transcript mapping. A simple algorithm called Repeatmasker is available over the Internet ***(24)***. Entries in the EST database corresponding to novel single-copy sequences that lack ORFs present more of a problem during positional cloning. EST entries by definition are single pass single stranded sequences, and are therefore error-prone. However, there are some transcripts identified numerous times in several tissues, and multiple sequence alignments give a reliable sequence that still lacks an ORF. Moreover, as high-quality bulk genomic sequence becomes available, the presence of stop codons in all frames of EST sequences is often being confirmed. These transcripts have introns, and the resulting exons can be identified by exon trapping. Seeking the function of nontranslated RNAs has been laborious without the aid of sequence similarities. The continuing analysis of quantitative trait loci from spontaneous mutation and large scale induced mutagenesis projects will eventually result in the endorsement of transcribed sequences to convert transcript maps into gene maps.

Acknowledgments

This work was supported by the Howard Hughes Medical Institute and the John Wulsin foundation. The authors would like to thank Dr. Megan Hersh for critically reviewing this manuscript.

References

1. Menon, A. G., Klanke, C. A., and Su, Y. R. (1994) Identification of disease genes by positional cloning. *Trends Clin. Med.* **4,** 97–102.

2. Apel, T. W., Scherer, A., Adachi, T., Auch, D., Ayane, M., and Reth, M. (1995) The ribose 5-phosphate isomerase-encoding gene is located immediately downstream from that encoding murine immunoglobulin kappa. *Gene* **156,** 191–197.
3. Buckler, A. J., Chang, D. D., Graw, S. L., et al.: (1991) Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci. USA* **88,** 4005–4009.
4. Church, D. M., Stotler, C. J., Rutter, J. L., Murrell, J. R., Trofatter, J. A., and Buckler, A. J. (1994) Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nat. Genet.* **6,** 98–105.
5. Haber, D. A., Sohn, R. L., Buckler, A. J., Pelletier, J., Call, K. M., and Housman, D. E. (1991) Alternative splicing and genomic structure of the Wilms tumor gene WT1. *Proc. Natl. Acad. Sci. USA* **88,** 9618–9622.
6. Taylor, S. A., Snell, R. G., Buckler, A., et al. (1992) Cloning of the alpha-adducin gene from the Huntington's disease candidate region of chromosome 4 by exon amplification. *Nat. Genet.* **2,** 223–227.
7. Lucente, D., Chen, H. M., Shea, D., et al. (1995) Localization of 102 exons to a 2.5 Mb region involved in Down syndrome. *Hum. Mol. Genet.* **4,** 1305–1311.
8. Wenderfer, S. E., Slack, J. P., McCluskey, T. S., and Monaco, J. J. (2000) Identification of 40 genes on a 1-Mb contig around the IL-4 cytokine family gene cluster on mouse chromosome 11. *Genomics* **63,** 354–373.
9. Church, D. M., Banks, L. T., Rogers, A. C., et al. (1993) Identification of human chromosome 9 specific genes using exon amplification. *Hum. Mol. Genet.* **2,** 1915–1920.
10. Trofatter, J. A., Long, K. R., Murrell, J. R., Stotler, C. J., Gusella, J. F., and Buckler, A. J. (1995) An expression-independent catalog of genes from human chromosome 22. *Genome Res.* **5,** 214–224.
11. Wieringa, B., Meyer, F., Reiser, J., and Weissmann, C. (1983) Unusual splice sites revealed by mutagenic inactivation of an authentic splice site of the rabbit beta-globin gene. *Nature* **301,** 38–43.
12. Andreadis, A., Gallego, M. E., and Nadal-Ginard, B. (1987) Generation of protein isoform diversity by alternative splicing: mechanistic and biological implications. *Annu. Rev. Cell Biol.* **3,** 207–242.
13. Xu, Y., Mural, R., Shah, M., and Uberbacher, E. (1994) Recognizing exons in genomic sequence using GRAIL II. *Genet. Eng.* **16,** 241–253.
14. http://image.llnl.gov/, webmaster@image.llnl.gov, Lawrence Livermore National Laboratory. The Image Consortium.
15. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998) New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282,** 682–689.
16. Lovett, M. (1994) Fishing for complements: finding genes by direct selection. *Trends Genet.* **10,** 352–357.
17. Simmons, A. D., Goodart, S. A., Gallardo, T. D., Overhauser, J., and Lovett, M. (1995) Five novel genes from the cri-du-chat critical region isolated by direct selection. *Hum. Mol. Genet.* **4,** 295–302.

18. Parimoo, S., Patanjali, S. R., Shukla, H., Chaplin, D. D., and Weissman, S. M. (1991) cDna selection: efficient Pcr approach for the selection of cDnas encoded in large chromosomal Dna fragments. *Proc. Natl. Acad. Sci. USA* **88,** 9623–9627.
19. Fan, W. F., Wei, X., Shukla, H., et al. (1993) Application of cDNA selection techniques to regions of the human MHC. *Genomics* **17,** 575–581.
20. Goei, V. L., Parimoo, S., Capossela, A., Chu, T. W., and Gruen, J. R. (1994) Isolation of novel non-HLA gene fragments from the hemochromatosis region (6p21.3) by cDNA hybridization selection. *Amer. J. Hum. Genet.* **54,** 244–251.
21. Schuler, G. D., Boguski, M. S., Stewart, E. A., et al. (1996) A gene map of the human genome. *Science* **274,** 540–546.
22. Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., and Sharp, P. A. (1986) Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* **55,** 1119–1150.
23. http://www-hgc.lbl.gov/human-p1s.html, Lawrence Berkeley National Laboratory, Human P1 sequence information.
24. http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/, Smit, A. F. A. and Green, P., Univ. Washington Genome Center. (4/21/99) REPEATMASKER WEB SERVER.