Plasmid Tagging for Efficient Large-Scale Sequence Completion of Entire Clone Inserts

BioTechniques 34:604-608 (March 2003)

Alan A. Jackson and Glenn K. Fu
Incyte Genomics
Palo Alto, CA, USA

With the advent of capillary gel electrophoresis, DNA sequencing has become quite robust and reliable (1). Still, there has always been a need to sequence large segments of cloned DNA. Various methods have been devised over the years to accomplish such sequencing. The shotgun approach to DNA sequencing utilizes random fragmentation of cloned DNA that can then be subcloned into a bacteriophage M13 vector (2). Historically, the focus of shotgun cloning has been to sequence-determine large genomic clones, but the technique is also often applied to sequence cDNA clones. With the shotgun method, a library of overlapping clones is created with sufficient redundancy such that—with a vector primer annealing site(s)—random segments of target DNA can be sequence-determined and then aligned by computer analysis to assemble the complete insert DNA sequence (3). A relatively new modification of the shotgun method is concatenated cDNA sequencing. Here, multiple DNA fragments are pooled, enzymatically concatenated, and then sequenced simultaneously as a single shotgun library (4). A major drawback to the shotgun process is that clones representing alternatively spliced transcripts of a given gene may be misassembled because of the dependency on sequence alignments in the assembly algorithm. Indeed, recent reports (5) suggest that as many as 40%-60% of human genes have alternative splice forms and that such variants may explain the functional intricacy of the human genome. Another method, primer walking, relies on sequence information that flanks an unknown DNA sequence gap on a clone insert. With the known sequence information, a gene-specific primer is designed and then used in combination with a vector primer or gene-specific primer at the opposite end to close the gap; if the target DNA is large, then further walking steps can be used until the gap is closed (6). Improvements to the primer walking process include the use of hexamers for economical, large-scale sequencing (7). While the primer walking method does allow accurate assembly of sequence fragments from an individual clone, it can be time consuming to sequentially design new primers. Moreover, the cost of synthetic oligonucleotides and the labor associated with multiple walking steps can be disadvantages for large-scale sequencing operations.

Other technologies have taken a different approach to DNA sequencing. One technique generates nested deletions with exonuclease III (8). The method requires two restriction enzymes, one to create a four-base 3' protrusion that protects the remainder of the given vector and sequencing primer site, and the other to create a 5' protruding end or blunt end that allows unidirectional digestion with exonuclease III. Since exonuclease III digests at a uniform rate, aliquots of the target DNA can be removed and then be treated with S1 nuclease to eliminate the exposed single strands. The last step in the procedure is to recircularize the timed aliquots with T4 DNA ligase and transform competent cells so that sequencing analysis can be performed. A more recent technology employs Tn5 intramolecular transposition to create nested deletions and inversions (9). The required components of this system are the mutant transposase, specific 19bp Tn5 end sequences, and target DNA (10). In this case, transposase mediates the insertion of a transpon containing an ori, a selectable resistance gene, and the two hyperactive end sequences into target DNA (9). The net result is a population of deletions and insertions that can be sequenced from primer sites within the transposon and on the given vector (9).

With access to one of the largest collections of partially sequenced human cDNA clones in the world, we sought to devise improvements to the shotgun sequencing method to sequence them completely in high-throughput fashion so as to identify clones containing full-length open reading frames. Here we describe a method to tag each shotgun fragment uniquely so that sequences can be assembled back to their respective pieces accurately, even in the presence of homologous or identical stretches within a pool of multiple clones. Our method uses plasmid pPDM-1 from EPICENTRE (Madison, WI, USA). A double-stranded nucleotide containing NotI and PiScel sites (5'-CTAGCGG- GCGCATCTATGTCCGCTTGAAGAGTTTAATGAAATGGCA-3' and 5'-GATCTGCGATTATTACC- TTTTCTCGCACCCGAATAGATA- TCGGCGGCG-3') was ligated to pPDM-1 at the Xbal-BamHI sites to form pPDM/NotPisce. Next, a popula-
tion of double-stranded random 20-mer signature tags was prepared by annealing 5'-GGGGTACCNNNNNNNNNNNNNNNNNNNGAATTCTGTACC-ACCTGCTAAC-3' and 5'-GTTAGCAGTGGTACAGA-3'. The tags were treated with Klenow and digested with EcoRI and KpnI before ligating to pPDM/NotPisce to form pUNISIG (Figure 1). Plasmid DNA from approximately 1 x 10^6 carbenicillin-resistant recombinant clones was isolated using plasmid maxiprep columns from Qiagen (Valencia, CA, USA). This cloning vector stock is essentially a library of unique signatures that can distinctly tag inserted DNA. Given the complexity of each signature, it is very unlikely that two distinct clones could share the same tag. PCR on 106 randomly selected clones showed that 99 (93.3%) had the signature insert. To test our vector, we chose to RT-PCR GenBank® accession no. AY166586 from a cDNA pool with

**Figure 1. The multiple cloning site of the pUNISIG cloning vector.** pPDM-1 was modified to contain a PiScel site and to incorporate a 20-mer signature tag. The location of the PvuII cloning site used for TA-cloning, as well as the position of the reverse sequencing primer, is shown as indicated.
Figure 2. Sequence coverage of assembly fragments generated using timed exonuclease deletions of the parent clone. The top line in the figure represents the 4917-bp mRNA transcript (GenBank accession no. AY166586) cloned into our vector using RT-PCR. The location of each signature-tagged shotgun sequence derived from timed exonuclease deletions of this clone is shown as indicated.

ACKNOWLEDGMENTS

We thank Gina Gorgone and Tania Ng for technical assistance.