

DOE/ER/60862--2

PROGRESS REPORT - DE-FG02-89ER-60862

DE91 006541

## Transposon Facilitated DNA Sequencing

Douglas E. Berg, Claire M. Berg and Henry V. Huang

02 year: through 27 Aug. 1990 (entire 2 yr. 7 mo. grant period began 1 July 1989).

The purpose of this research is to investigate and develop methods that exploit the power of bacterial transposable elements for large scale DNA sequencing: Our premise is that the use of transposons to put primer binding sites randomly in target DNAs should provide access to all portions of large DNA fragments, without the inefficiencies of methods involving random subcloning and attendant repetitive sequencing, or of sequential synthesis of many oligonucleotide primers that are used to march systematically along a DNA molecule. Two unrelated bacterial transposons, Tn5 and  $\gamma\delta$ , are being used because they have both proven useful for molecular analyses, and because they differ sufficiently in mechanism and specificity of transposition to merit parallel development.

## I - PROGRESS

1. "Deletion factory". Nested sets of deletions extending to sites in cloned DNAs are attractive for DNA sequencing because (i) DNA adjacent to the constant end can serve as a fixed primer binding site for sequencing the DNAs that remain adjacent to the variable end in each deletion mutant; (ii) the portion of target DNA to be accessed in this way can be estimated from plasmid size; and (iii) deletions of specific sizes can be selected by preparative gel electrophoresis. We are therefore examining the use of intramolecular transposition to generate nested deletions.

1-a. Initial tests were carried out using 8 kb plasmids with synthetic Tn5 elements that contain a gene conferring sucrose sensitivity (*sacB*). Deletions, selected by sucrose resistance, were obtained at frequencies of about  $10^{-6}$ , and constituted about one-fifth of all sucrose-resistant variants. Restriction mapping and DNA sequencing of deletion plasmids showed that the endpoints were widely distributed in the several kb long DNA targets, and that the deletions started precisely at the transposon ends. Inversions were also obtained, but were only about one-sixth as frequent as deletions. A manuscript describing these analyses has been accepted for publication (Tomcsanyi, C. Berg, Phadnis and D. Berg, 1990, *J. Bacteriol.*). We are continuing this work with the aims of (i) developing methods to preferentially stimulate intramolecular transposition, so that deletions will greatly outnumber point mutations in *sacB*; (ii) developing the use of two different contraselectable markers (e.g. *sacB* and *thyA* (see below)) to facilitate the selection of deletions without complication from other types of mutations; and (iii) extending this deletion factory approach to larger DNA segments.

**MASTER**

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

1-b.  $\gamma\delta$ , which is unrelated to Tn5, inserts with a different specificity, and preferentially into plasmids; the transposase that mediates  $\gamma\delta$  movement operates more effectively in trans than does the transposase of Tn5. We have constructed a test plasmid with a mini- $\gamma\delta$  element near *thyA*, and selected trimethoprim resistant derivatives, which result from loss of *thyA* function (see below). Preliminary tests indicated that more than half of the *Suc<sup>R</sup>* derivatives contain deletions. Experiments underway will test the expectation that  $\gamma\delta$  is generating sets of nested deletions with one endpoint fixed at a  $\gamma\delta$  end, and will compare the Tn5 and  $\gamma\delta$  intramolecular transposition systems for specificity, ease and efficiency, and suitability for analyzing very large cloned DNAs (e.g. in cosmid, P1 and F factor vectors).

## 2. Probe mapping to facilitate transposon-based DNA sequencing.

We have shown that transposon insertions in cloned DNAs can be easily located by a fast probe mapping/Southern blotting strategy, without prior knowledge of the restriction map of the target region. Our tests involved intermolecular  $\gamma\delta$  transposition to a cloned fragment of *Drosophila melanogaster* DNA. The particular clone used as a target was chosen because prior partial sequence determinations had revealed regions of atypical base composition that would help test the randomness of  $\gamma\delta$  insertion. The probe mapping results showed that, at the level of resolution needed for sequencing (a few hundred bp), the transposon insertions were quite randomly distributed. These findings, which have just been published (Strausbaugh, Bourke, Sommer, Coon and C. Berg. 1990. PNAS 87:6213-6217), served as the basis of the application for DOE grant DE-FGO2-90-ER61000 to L. Strausbaugh (PI) and C.M. Berg (co-PI).

## 3. Efficiency of $\gamma\delta$ insertion (intermolecular).

The standard means of delivering  $\gamma\delta$  to plasmid DNA targets relies on the fact that most plasmid cloning vectors are not self-transmissible in bacterial conjugation. Transpositions from the *E. coli* F factor (a conjugative plasmid containing  $\gamma\delta$ ) to the target are selected in a mating out assay which depends on the transient formation of F-target plasmid cointegrates (the first products of  $\gamma\delta$  transposition). We addressed the problem raised by several recent reports (and confirmed in C. Berg's lab) that the F factor can sometimes cause transfer of pBR322-related plasmids seemingly without  $\gamma\delta$  insertion. We showed that this occurs only when target plasmids are dimers or higher oligomers, and when the F<sup>-</sup> recipient cell in the mating out assay is *recA<sup>+</sup>*. Our analyses showed that seeming failures to generate  $\gamma\delta$  insertions reflects the presence of plasmid dimers (and higher oligomers), transfer of oligomers containing just single inserts of  $\gamma\delta$ , recombination and segregation in the recipient, often leading to the preferential recovery of the smaller  $\gamma\delta$ -free component. This complication is avoided by use of monomeric plasmids in *recA<sup>-</sup>* donor cells (Liu and C. Berg, J. Bact. 172:2814-2816, 1990).

#### 4. Mini- $\gamma\delta$ for more efficient insertion mutagenesis and DNA sequencing.

Wild type  $\gamma\delta$  is 6 kb in size, and is not marked with selectable (e.g. resistance) genes. We developed a 1.8 kb  $\gamma\delta$ -kan<sup>R</sup> element called m $\gamma\delta$ -1 that is more useful than wild type because of its size, distribution of restriction sites, and selectable marker. This element was found to insert quite randomly in a target plasmid containing cloned *Streptomyces* DNA (which is GC-rich), and also, to insert more randomly than wild type  $\gamma\delta$  in a pBR325 test plasmid.

#### 5. Transposon insertion specificity.

We have examined factors controlling the choice of insertion sites during Tn5 and  $\gamma\delta$  transposition.

**Tn5.** Mutational analyses of pBR322-derived plasmid targets showed that the relatively frequent insertion of Tn5 into its major hotspot (the promoter for the *tet* gene) depends on a constellation of factors: the 9 bp sequence of the target site that are duplicated by insertion, upstream sequences, and also extremely high negative supercoiling. Insertion at other sites seems to depend on another subset of sequences, and not to require such high negative supercoiling. (Lodge and D. Berg, 1990. *J. Bact.* in press). Additional analyses showed that the end sequences from Tn5 or its component IS50 elements (the substrates for transposition proteins) do not create insertion hotspots (Lodge, Weston-Hafer and D. Berg, submitted for publication). This outcome may prove important in the design of two-transposon PCR strategies to be developed (see below). More generally, the considerable variety of signals that each contribute to Tn5 insertion helps explain the tens or hundreds of sites in any typical gene that can be used for Tn5 insertion; the near-randomness of Tn5 insertion underlies much of the utility of Tn5 for DNA sequencing and other molecular genetic applications.

$\gamma\delta$ . The  $\gamma\delta$  transposon inserts preferentially into AT-rich target DNAs, and generates 5 bp target duplications that often contain only A and T. These observations led to the expectation that insertions would be clustered in AT-rich regions in plasmids that do not have uniform DNA composition, and that  $\gamma\delta$  would not be useful for analyses of GC-rich DNAs. When this prediction was tested with a cloned *D. melanogaster* fragment that had been found to contain regions of both high and low AT content, however, no clustering of insertions was found (Strausbaugh, Bourke, Sommer, Coon and C. Berg. 1990. *PNAS* 87:6213-6217). We have also isolated insertions of m $\gamma\delta$ -1 (see above) in a plasmid containing a cloned *Streptomyces* DNA fragment that, like other *Streptomyces* DNA, has a very low AT content. In both cases, the anticipated bias against insertion into AT-poor DNA was not observed. Instead, the insertion sites were quite randomly distributed.

An examination of  $\gamma\delta$  landing sites explains this paradox: insertions in GC-rich regions occur preferentially in AT-rich "valleys" within them. The related element Tn3 gives a similar pattern. This suggests that  $\gamma\delta$  and Tn3 find their insertion sites by a two-step process: a relatively non-specific interaction with target DNA, and then local searching for AT-rich segments as small as 5 bp. Analyses of landing sites of several unrelated prokaryotic transposons does not reveal a similar pattern, suggesting that this two-step mechanism may be unique to members of the Tn3 family. In any case, the near random choice of regions for insertion, regardless of overall base composition indicates that  $\gamma\delta$  will be generally useful in transposon-based DNA sequencing strategies.

**6. Genes for use as dual purpose markers: selection and contra-selection.** There are many situations, exemplified by the deletion factory approach just described (1, above) in which it would be advantageous to have a gene that can be selected both for and against. In general, antibiotic resistance determinants are used as selectable markers, e.g. in plasmid vectors and transposons. Only a few genes have been developed as contraselectable markers, most notably *sacB* from *B. subtilis* (sucrose-sensitivity in *E. coli*; 1, above), and none of these have alternative selectable phenotypes.

We have begun developing *thyA* as an efficient dual purpose marker: A *thyA*<sup>+</sup> allele makes *E. coli* highly sensitive to trimethoprim (Tmp) and related dihydrofolate reductase inhibitors. *E. coli* mutants isolated on the basis of Tmp resistance on thymine-containing medium are usually thymine auxotrophs. We have selected plasmid clones carrying the *thyA* gene of either *E. coli* or *Lactobacillus casei* next to  $\gamma\delta$  by transformation of *thyA*<sup>-</sup> mutant *E. coli* to prototrophy, and then selected for  $\gamma\delta$ -mediated deletions by Tmp<sup>R</sup> (see 1-b, above). The *thyA* gene of *L. casei* was used because it is distantly related in sequence to *thyA* of *E. coli*, and this sequence divergence eliminates recombination (gene conversion) from the host chromosomal *thyA*<sup>-</sup> allele as a means of forming Tmp<sup>R</sup> derivatives of *thyA*<sup>+</sup> plasmids. We will continue to exploit *thyA* genes as dual-purpose selectable-contraselectable markers. In addition, we will investigate the utility of the *tetR* determinant of Tn10 as a complementary dual purpose marker. We believe that the *tet* gene is promising because it confers sensitivity to fusaric acid as well as resistance to tetracycline, without the need to construct special (i.e. *thyA*<sup>-</sup>) host strains.

## **II. PLANS for 03 year of grant:**

Most plans for the remaining 16 months of this grant period remain essentially as initially proposed, or as noted at the ends of descriptions, above. However, several additional procedures, based on the polymerase chain reaction (PCR) in combination with transposon insertion will also be developed.

For sequencing DNAs cloned in  $\lambda$  phage DNAs, we will use PCR to amplify cloned DNA segments between sites of transposon insertion and vector DNA adjacent to the cloning site. Because DNA from phage plaques or small colonies can be used for PCR amplification, this approach will eliminate time-consuming procedures of  $\lambda$  phage stock preparation and then DNA extraction. PCR fragments will be sequenced directly using primers specific for transposon ends. No difficulty is expected in obtaining sufficient yields of PCR fragments for insertions within about 5 kb of the vector cloning sites. However, cloned DNAs are typically in the 15-20 kb size range in good  $\lambda$  libraries, and transposon insertions in the middle of such segments (i.e. 8-10 kb from each junction with the vector) may be beyond the reach of normal amplification. We will thus use conditions, some of which will have to be developed, that maximize the lengths of PCR amplifiable segments: much of this will entail use of "next generation" highly processive thermostable polymerases and associated proteins that are being developed by various commercial suppliers, and also temperature cycling conditions that maximize the processivity of the polymerization reactions.

Many transposon insertions in large cloned DNAs will be too distant from the cloning site to allow efficient amplification, even after maximizing PCR processivity as above. We will, accordingly, develop alternative methods for amplification of such DNAs.

We will test the feasibility of a two-transposon system in collaboration with L. Strausbaugh: For example, DNA containing an insertion of Tn5 that is not close enough to the vector site for amplification will be used as the target for  $\gamma\delta$  transposition, and then segments between  $\gamma\delta$  and Tn5 will be amplified.  $\gamma\delta$  insertions that give large PCR fragments with this Tn5 insert will, in many cases, be close enough to a vector site for amplification. Thus, the segment from the vector site to the distant Tn5 transposon insertion will be accessed in two parts: from the vector to the nearby  $\gamma\delta$  insertion; and then from  $\gamma\delta$  to Tn5. Given additional engineered transposons with distinguishable selectable markers and PCR primer tags at their ends, one could, in principle, march systematically through much longer DNA segments; we will consider this extension of the method after evaluating the results of the two-transposon PCR approach.

As an alternative to two transposon PCR, we will test the ease, fidelity and efficiency of "crossover PCR". Two different DNAs, each with a single transposon insertion will be mixed, and used in PCR reactions. Published accounts of crossover PCR (in which it was generally considered an artifact, to be avoided) point to template DNA concentration, primer concentration, and polymerase processivity as important factors (see, e.g. Meyerhans et al., NAR 18:1687-

1691). We will investigate these and other variables using transposon insertions at known sites. Transposon-based crossover PCR is appealing because of its flexibility: many single insertions in any target DNA can be made, and tested pair-wise, without need for prior mapping. It would avoid the need to choose any single insertion-containing DNA for the isolation of additional transposon insertions. Thus, if transposon-based crossover PCR can be made operational with the small amounts of DNAs obtainable from phage plaques it should be more powerful than two-transposon PCR.

## **DISCLAIMER**

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

**END**

**DATE FILMED**

02/07/91

