Efficient integration of artificial transposons into plasmid targets *in vitro*: a useful tool for DNA mapping, sequencing and genetic analysis

Scott E.Devine and Jef D.Boeke*

Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Received May 9, 1994; Revised and Accepted 21 July 1994

EMBL accession no. U09476

ABSTRACT

We have developed efficient methods for creating artificial transposons and inserting these transposons into plasmid targets *in vitro*, primarily for the purpose of DNA mapping and sequencing. A novel plasmid has been engineered to convert virtually any DNA sequence, or combination of sequences, into an artificial transposon; hence, custom transposons containing any desired feature can be easily designed and constructed. Such transposons are then efficiently inserted into plasmid targets, *in vitro*, using the integrase activity present in yeast Ty1 virus-like particles. A single *in vitro* integration reaction, which resembles a simple restriction digestion in the complexity of the reaction, gives rise to thousands of recoverable insertion events within DNA target molecules; this frequency approaches one insertion per phosphodiester bond in typical plasmids. Importantly, transposon insertions are recovered from all regions of DNA inserts carried on plasmid targets, indicating that integration is a random or nearly-random process. Because of its versatility, this technology offers a generalized method of generating recombinant DNA molecules of a desired structure. We have adapted this system for DNA sequencing by developing a customized artificial transposon to insert new primer binding sites into internal regions of DNA inserts carried on cloning vectors. Transposon insertions have been generated throughout several different yeast and human DNA inserts carried on plasmids, allowing the efficient recovery of sequence information from these inserts. Our results demonstrate the overall utility of this method for both small and large-scale DNA sequencing, as well as general DNA restructuring, and indicate that it could be adapted for use with a number of additional applications including functional genetic analysis.

INTRODUCTION

DNA sequencing has helped revolutionize the way that genes and genomes are studied, and has led to a greater understanding of most aspects of biology. Nevertheless, with efforts underway to map and sequence the genomes of a variety of organisms, the need to improve the efficiency of DNA sequencing has never been greater (1). One of the major problems associated with sequencing large segments of DNA is obtaining sequence information beyond the limits of a single primer extension event. Several techniques are currently used to acquire sequences within the interior of a DNA insert; these include: i) the synthesis of custom primers to 'walk' along a segment of DNA (2, 3), ii) shotgun subcloning, which requires a high degree of redundancy for complete sequence recovery (4), or iii) the construction of overlapping exonuclease deletion clones (3, 5). Each of these methods is time-consuming, idiosyncratic and therefore difficult to automate, and/or costly.

Transposable elements have also been adapted for DNA mapping and sequencing. Examples include: $\gamma\delta$ (6), Tn5 (7), Tn10 (8,9) as well as derivatives of these and others (10, 11). For DNA sequencing, customized transposons are used as mobile priming sites which are integrated into the interior of a DNA insert carried on a cloning vector. As a result of transposon integration, unique priming sites of a known sequence are introduced adjacent to previously uncharacterized sequences, allowing the recovery of sequence information from these regions. Transpositions are generally carried out *in vivo* in bacteria, using a donor strain carrying the transposon either on a vector or as integrated copies in the genome. A target of interest is introduced into the transposon host where it receives integrations. Targets bearing insertions are then recovered from the host by genetic selection.

The complete DNA integration reaction employed by certain retroviruses and retrotransposons as part of their normal life cycles can be carried out completely *in vitro* (12–16) offering a possible alternative to *in vivo* transposon insertion techniques for DNA sequencing. An *in vitro* system offers several advantages

*To whom correspondence should be addressed

over *in vivo* transposition systems: i) special bacterial strains are not required, ii) potential host effects are avoided, and iii) an *in vitro* reaction is amenable to biochemical alteration and parameter optimization. Hence, we have developed a transposon insertion technique that is carried out entirely *in vitro* that may be applied to a variety of problems, including DNA sequencing. This technique employs artificial transposons which are created using a novel plasmid construct, and yeast Ty1 integrase provided in the form of Ty1 virus-like particles (VLPs), which mediates the insertion of these transposons into plasmid targets *in vitro*.

Artificial transposons are ideal for DNA sequencing: i) a large number of transposon insertions can be easily assembled from a single *in vitro* integration reaction, allowing the recovery of insertions suitably spaced to facilitate sequencing of a DNA segment, ii) the transposon can be engineered to contain desired features useful for DNA mapping or sequencing, and iii) since each transposon carries unique primer sites at both ends, the nucleotide sequence flanking each insertion site can be rapidly and efficiently determined. A set of plasmids bearing artificial transposon insertions is especially useful for sequencing because all the plasmids can be sequenced in parallel using a defined pair of primers. This is in contrast to the inefficient 'series' approach of primer walking, in which each sequence is used to specify the next primer. Hence, artificial transposons are novel, flexible, and extremely efficient for generating DNA sequencing templates useful for both small and large-scale DNA sequencing projects.

MATERIALS AND METHODS

Construction of pAT-1 and pAT-2

pAT-1 (pSD544) and pAT-2 (pSD545) were constructed as follows. First, the plasmid pRS316 (ref. 17; a derivative of pBLUESCRIPT, Stratagene) was modified to eliminate the ampicillin resistance gene. (We found that donor plasmids lacking a functional *bla* gene were preferable to those containing a functional *bla* gene; donor plasmids containing a function *bla* gene were a source of false positives upon *E. coli* transformation and selection on plates containing ampicillin and trimethoprim.) The *bla* gene was eliminated from pRS316 by ligating together two fragments of pRS316 (a 2.1 kb *Ssp*I fragment and a 2.1 kb *Bsa*I/*Ssp*I fragment), thus creating the plasmid pSD528; this plasmid can be propagated in the pyrimidine-requiring *E. coli* strain MH1066 since the yeast *URA3* gene on this construct complements the bacterial *pyrF* auxotrophy (18). pAT-1 and pAT-2 were constructed from plasmid pSD528 by replacing the pBLUESCRIPT multicloning site (mcs) (from the unique *Kpn*I site to the unique *Sac*I site) with polymerase chain reaction (PCR) adapters containing the appropriate sequences to create the structure indicated in Figs. 2 and 3. These PCR adapters were generated using primers SD112 (JB661) (5'-AAAAGCTGGG-TACCGAACATGTTCTCGAGGTCGACGGTATCG-3') and SD113 (JB662) (5'-GCGAATTGGAGCTCGAACATGTTCA-CCGCGGTGGCGGCCGCTC-3') with plasmids pBLUE-SCRIPT and pSD511 as templates. The resulting PCR products were digested with *Kpn*I and *Sac*I, and ligated to *Kpn*I/*Sac*I-digested pSD528 to generate pAT-1 and pAT-2. The structures of these constructs were confirmed by restriction mapping and sequence analysis.

In vitro reaction conditions

A typical *in vitro* DNA integration was carried out in a 20 μ l reaction volume, and contained the following. 100–500 ng

artificial transposon (0.8 kb), 1 μ g CsCl-purified plasmid target (a 10 to 1 molar ratio of transposon to target), 2 μ l 10 $\times$ reaction buffer (150 mM MgCl₂, 100 mM Tris HCl, pH 7.5, 100 mM KCl, and 10 mM DTT), 5 μ l 20% [w/v] PEG 8000, 2 μ l VLPs, and water to 20 μ l. The reaction was incubated at 30° Celsius for 30 minutes followed by 37° Celsius for 10 minutes, and then was terminated by adding 1.0 ml 0.5 M EDTA and heating to 65° Celsius for 20 minutes. Finally, the nucleic acids were phenol/chloroform extracted, ethanol precipitated, collected by centrifugation, washed with 70% ethanol, and resuspended in 10 μ l TE (10 mM Tris, pH 8.0, 1 mM EDTA). 1 μ l was used to transform 6 μ l DH10B *E. coli* (Gibco/BRL) to drug resistance by electroporation.

PCR, sequencing, primers, plasmid constructions, CsCl preps

The PCR was carried out using reagents obtained from Perkin Elmer, as described (19). DNA sequencing was carried out using Sequenase (USB), and analyzed as described (20). Custom oligonucleotide primers were obtained from Operon Technologies, Inc. (Alameda, California). The primers for sequencing from within the pAT-2 transposon are SD110 (JB532) (5'-GGTGATCCCTGAGCAGGTGG-3'), SD111 (JB563) (5'-G-ACACTCTGTTATTACAAATCG-3'), SD118 (JB781) (5'-C-ATTGGATGCTGAGAATTTCG-3') and SD119 (JB802) 5'-GCCGTCTATCCTGCTTGC. The integration site of each insertion was determined with the aid of the Wisconsin GCG package. Plasmids were constructed using standard DNA cloning methods (21), and were purified from *E. coli* cultures by either STET miniprep (22) or alkaline lysis followed by CsCl banding (23).

Preparation of artificial transposons from pAT-1 and derivatives

20 μ g of CsCl-purified plasmid DNA was digested with 50 units of *Xmn*I (NEB) or *Asp*700 (Boehringer Mannheim) for 1.5–4 hours at 37° Celsius. The resulting fragments were separated on a 1% agarose/TBE gel, and the transposon fragment was electroeluted from the gel using an IBI electroelution device. Some batches of agarose were found to yield inactive transposon preparations, and highly purified agarose (e.g., ultraPURE LMP agarose, Gibco BRL) generally worked better than cruder grades.

Recovery of clones carrying transposon insertions using ampicillin/trimethoprim plates

E. coli clones carrying plasmids with transposon insertions were identified by selection on M9 minimal plates (24) containing 1.0 mM thiamine HCl, 50 μ g/ml ampicillin (Amp) and 100 μ g/ml trimethoprim (Tmp; Sigma). After one to two days incubation at 37° Celsius, the majority of colonies growing on M9/Amp/Tmp plates contained plasmids with a transposon insertion. Dilutions of the transformation were routinely plated on LB plates containing 50 μ g/ml Amp (24); this control monitored the number of target plasmids successfully carried through the procedure. When compared to the number of colonies on M9/Amp/Tmp plates, the frequency of transposon insertion could be estimated (frequency of insertion = [# colonies on M9/Amp/Tmp plates] / [# colonies on LB/Amp plates]). A positive control plasmid, pSD511, containing both Amp^R and Tmp^R markers, routinely gave rise to equivalent numbers of colonies on LB/Amp (50 μ g/ml), M9/Tmp (100 μ g/ml), or M9/Amp/Tmp (50/100 μ g/ml) plates under these conditions.

Transformation of *E. coli*

The two strains transformed routinely in this work were DH5 α (25) and DH10B (26). DH5 α was prepared for electroporation as described (27), and electrocompetent DH10B cells were purchased from Gibco/BRL. Transformation by electroporation was accomplished for both strains using a Biorad Genepulser with 1 mm cuvettes and the following settings: capacitance: 25 μ FD; voltage: 1.8 kV; and resistance: 200 ohms. Using pUC19 or pBLUESCRIPT as a test plasmid, freshly-prepared electrocompetent DH5 α generally showed transformation efficiencies of 10^7 – 10^8 colonies/ μ g DNA, whereas electrocompetent DH10B purchased from BRL/Gibco generally showed efficiencies of 5×10^8 to 5×10^9 colonies/ μ g DNA.

VLP preparation

VLPs were prepared from yeast cultures as described (28). Fractions from the sucrose gradient showing integrase activity were aliquoted and frozen at -70° Celsius where they were stable for more than 6 months. Established methods for VLP preparation are very reproducible; however, the ability to purchase pretested VLPs from a commercial source would make this method more accessible, and we are pursuing this possibility.

Figure 1. Overview of artificial transposon insertion into plasmid targets. The basic steps involved in generating artificial transposon insertions in target plasmids are indicated. Note the following: target plasmid (double circle); artificial transposon (rectangle); *dhfr* (trimethoprim resistance) gene (shaded box); Ty1 U3 termini (filled boxes); DNA sequences to be determined (dashed line).

RESULTS

Basic reaction

There are three macromolecular components to the *in vitro* integration reaction: i) an artificial transposon, ii) Ty1 virus-like particles, which are the source of integrase activity, and iii) a plasmid target. These three components are mixed together in a reaction containing the appropriate buffer and cofactors (see Materials and Methods). The reaction is briefly incubated at 30° and 37° Celsius, and terminated by adding EDTA and heating to 65° Celsius. Finally, the nucleic acids are phenol/chloroform extracted and ethanol precipitated. The recovered DNA is used to transform *E. coli* to drug resistance, allowing the identification of target molecules which have received a transposon integration (Fig. 1). A set of transposon-bearing plasmids may then be used directly to obtain the DNA sequences flanking the insertion sites, using subterminal primers adjacent to the transposon termini; a collection of such insertions can be used for the efficient recovery of DNA sequence information from the region of interest.

Production of artificial transposons

We have devised a simple and versatile method of producing artificial transposons (Fig. 2). Any DNA sequence can, in theory, be converted into an artificial transposon, using this approach (although the upper limits of size are unknown). The strategy involves an altered pBLUESCRIPT (Stratagene) mcs, which has been adapted to include Ty1 terminal sequence cassettes flanking the set of unique cloning sites (in pSD528). These Ty1 terminal cassettes, in addition to Ty1 U3 terminal nucleotide sequences, have been designed to also contain the recognition sequences for the restriction endonuclease *Xmn*I. Hence, because there are two

Figure 2. pAT-1 and pAT-2. A. The backbone common to pAT-1 and pAT-2 is shown to contain the yeast *URA3* gene, a bacterial origin of replication (*ori*) and a multicloning site (*mcs*). pAT-2, containing the Ty1 *mcs* and artificial transposon insert, is depicted. B. The artificial transposon, AT-2, which is created upon digestion with *Xmn*I contains the *dhfr* gene (stippled), subterminal segments of the pBLUESCRIPT *mcs* (white boxes), and Ty1 U3 cassettes (filled boxes), which are incorporated into *Xmn*I cleavage sites. The arrow below the *dhfr* gene indicates the direction of transcription. C. The sequence at Ty1 U3/*Xmn*I cassettes. The arrows indicate the *Xmn*I cleavage site. The shaded areas indicate Ty1 U3 sequences (one on either side of the arrows), while the entire sequence encodes a recognition site for *Xmn*I.

Figure 3. Map and sequence of artificial transposon AT-2. **A.** Map of AT-2. Note the U3 termini (black boxes at ends), subterminal multicloning regions (white boxes), and the *dhfr* gene including the transcription start site and translational ORF. The sequencing primer sites (SD110, 111, 118, and 119) are shown. **B.** AT-2 sequence (Genbank accession # U09476). Note the U3 4 bp termini (bold) and multicloning regions (boxed). The *dhfr* gene sequence is not fully shown but is represented by a box in the center. The sequencing primer sites are shaded.

such sites (one on either side of the mcs), digestion with *XmnI* both cleaves the fragment from the plasmid and precisely creates *Ty1* U3 termini at the ends of the liberated fragment, thus allowing it to act as a substrate for *Ty1* integrase; as a consequence, the fragment can act as a transposon. Moreover, since these *XmnI*-*Ty1* sequence cassettes flank the mcs, *XmnI* digestion potentially converts any sequence carried as an insert in the mcs into an artificial transposon, provided that the insert does not contain *XmnI* sites. New sequences can easily be cloned into the adapted multicloning site, since the open reading frame allowing blue/white selection was maintained. The resulting plasmid, containing this *Ty1* mcs, was named pAT-1 (for Artificial Transposon -1).

Using a strategy similar to that used for the construction of pAT-1, an artificial transposon was created for the purpose of DNA sequencing. The resulting plasmid carrying this transposon is pAT-2 (Figs. 2 and 3). A drug resistance marker (the *dhfr* gene from the bacterial transposon Tn7, kindly provided by N.Craig) was added to the mcs, allowing facile identification of colonies containing a transposon insertion clone (clones bearing an insertion are trimethoprim resistant). Hence, the *XmnI*-processed AT-2 transposon contains: i) the *dhfr* gene, ii) the majority of the mcs (including a unique *NotI* site), iii) unique subterminal primer sites at both ends, and iv) *Ty1* U3 termini (Figs. 2 and 3).

***In vitro* integration of 'primer island' transposons into a cloned segment of yeast chromosome III carried on a plasmid target**

We next generated artificial transposon insertions *in vitro* using the transposon from pAT-2 and various plasmid targets. One of the primary test clones consisted of a pRS200 backbone (a derivative of pBLUESCRIPT) with an 8.0 kb insert that spans

Table 1. Recovery of artificial transposon insertions into clone 76-2

Rxn	EDTA ^a	Total transformants ^b	Total insertion plasmids ^c	Frequency of transposition ^d
-C	-	0	0	N.A.
+C	-	3.1×10^8	4.5×10^8	N.A.
1	-	3.1×10^8	1.3×10^4	4.2×10^{-5}
2	+	5.7×10^8	5.0×10^2	9.1×10^{-7}

Reaction: -C, negative transformation control (no DNA added); +C, positive transformation control (pSD511, which contains both *Amp^R* and *Tri^R* markers); 1, complete integration reaction using p76-2 as the target; 2, same as reaction 1, but EDTA was added (inhibits integrase activity); N.A., not applicable.

^a+, EDTA added to 25 mM.

^bTotal number of *Amp^R* transformants.

^cTotal number of *Amp^R* + *Tmp^R* transformants.

^dNumber of transpositions into target plasmid (*Amp^R* + *Tmp^R* colonies) divided by the total number of transformants (*Amp^R* colonies).

bp 136,155 to 144,333 of yeast chromosome III; this plasmid is called p76-2 (the details of the isolation of this clone will be described elsewhere; Devine and Boeke, in preparation). With a single *in vitro* integration reaction, we recovered approximately 13,000 transposon insertions in p76-2 (Table 1). It is noteworthy that a linear DNA molecule containing two U3 termini can behave as a substrate for *Ty1* integrase, since *Ty1* itself has only a single U3 terminus (and a U5 terminus at the other end). In fact, our data indicate that a configuration of two U3's works at least as well as U3/U5 (not shown). Thus linear DNA molecules with terminal sequences other than the U3/U5 configuration behave as substrates for complete *Ty1* integrase-mediated insertions, making the *XmnI* strategy discussed above generally useful for the generation of artificial transposons.

By measuring the number of colonies transformed to ampicillin resistance vs. combined trimethoprim and ampicillin resistance, we determined that the frequency of transposon insertion recovery was approximately 4.2×10^{-5} (i.e., 1 insertion per 2.4×10^4 target molecules; Table 1). In other experiments, this frequency was as high as 1 insertion per 2.5×10^3 targets). Although such frequencies are not likely to represent the upper limits of optimization, they are sufficiently high that a large number of insertion events are readily recovered, while sufficiently low that a single target is generally limited to a single transposon insertion (two transposon insertions in a single target might be useful for some purposes, but would render the molecule useless as a sequencing template).

Analysis of 156 randomly chosen *Amp^R*/*Tmp^R* colonies indicated that transposon insertions occurred into all areas of the plasmid target, including both the pRS200 backbone (6.0 kb) and the 8.0 kb chromosome III insert, as determined by restriction mapping and/or sequence analysis (Table 2). More than 86% of these 156 clones (134) had easily-identifiable, single artificial transposon insertions; of these, 78 (50%) were in the cloned 8 kb insert, while 56 (36%) were in the vector. A small percentage of the clones were found to have two superimposed restriction maps/and or sequences. There are several likely explanations for this result, including the possibility that two plasmids transformed a single *E. coli* clone, or that two transposon insertions occurred into a single plasmid target; the available evidence indicates that most of these clones are explained by such mechanisms (Table 2). Hence, a small portion of clones recovered from an *in vitro* integration reaction would not be suitable for direct DNA sequence analysis for this reason (12 % in this example, Table

2). Likewise, vector insertions would not be useful for DNA sequence analysis. Smaller vectors containing only those features essential for plasmid function (6) would be expected to minimize the number of insertions recovered in the vector backbone. Nevertheless, even using a relatively large plasmid vector such as that used here, one of every two Amp^R/Tnp^R colonies analyzed from a single *in vitro* reaction could be used directly to obtain DNA sequence from the cloned insert (Table 2). Furthermore, analysis of only 156 minipreps led to the assembly of 78 useful insertions in an 8 kb insert, corresponding to an expected distribution of roughly one insertion per 100 bp.

Since the entire yeast chromosome III sequence has been previously determined (29), we could easily identify the precise sites of transposon integration by determining the nucleotide sequences at the insertion junctions. Indeed, the 78 insertions were distributed throughout the entire 8 kb insert, and were found in all 6 genes and open reading frames (ORFs) known to occupy this region of chromosome III. Insertions occurred into both RNA polymerase II and III-transcribed genes as well as a Ty1 solo δ (Fig. 4). The *PGK1* gene for example, which is a polymerase II-transcribed gene whose ORF occupies 1.2 kb from bp 137,349

to 138,597, received 14 insertions. The *SUF16* tRNA gene, which is a polymerase III-transcribed gene of approximately 70 bp located from bp 142,307 to 142,377, received an integration within the proposed leader sequence at a position 5 bp upstream of the 5' end of the mature tRNA. A little less than half of the 78 insertions were in the forward orientation (34/78 or 44%), indicating a slight orientation bias for this target. However, since primer extensions can be initiated into the sequences flanking the insertion on both sides irrespective of the transposon orientation, an orientation bias does not affect the utility of an insertion for purposes of DNA sequencing. The interval sizes between adjacent insertions ranged from 1 bp to 377 bp, and the mean interval size was 102.3 $\pm$ 88.1 (Table 3). Only six of the intervals were greater than 250 bp, and the largest of these was 377 bp. Hence, the vast majority of the intervals between adjacent transposon insertions were well below the maximum distance that can be reached with an average primer extension under sequencing conditions. A property of Ty1 integrase is that it creates characteristic 5 bp target sequence duplications flanking the insertion site upon integration (13–15, 30). As expected, 5 bp target site duplications were found at each integration site examined (only a small portion of the insertions were sequenced at both ends in this example). No deletions or rearrangements were observed.

Table 2. Examination of Amp^R+Tnp^R colonies from a single *in vitro* integration reaction

		%
Total number of Amp ^R + Tnp ^R clones examined	156	100
# minipreps recovered	153	98
# easily-identifiable insertions	134	86
In insert	78	50
In vector	56	36
Other	19	12
double insertions/cotransformants ^a	13	8
unknown plasmid map	5	3
no transposon	1	<1

^aThis class contains some plasmids that apparently had two independent insertions in the target as determined by restriction mapping, and others with DNA sequence that was readable to the insertion junction, at which point two superimposed sequences were observed.

Other targets tested

In addition to clone 76-2 containing a DNA insert from yeast chromosome III, we have tested other plasmid targets. These plasmids had a variety of backbone structures and carried various cloned inserts. The backbones included pUC19 and pBLUESCRIPT as well as others, and the DNA inserts originated from different species including yeast and human. In each case, results similar to those shown for clone 76-2 were obtained: i) insertions were mapped to all regions of these targets, ii) a large number of insertions was readily recovered from reactions using each target, and iii) recovered insertions consistently served as successful sequencing templates. Moreover, in two cases other than p76-2 (pCAR143 and pWAFP), this system was used to recover 90–100% of the nucleotide sequence from clones with previously unknown sequences (data not shown; C.Baker

Figure 4. Artificial transposon insertions in clone p76-2. The 8 kb insert of clone p76-2, containing a segment of yeast chromosome III, is shown along with the sites of 78 independent AT-2 insertions (vertical arrows). The orientation of insertion is indicated: (↓) Forward (the *dhfr* gene in the artificial transposon is transcribed left to right, or (↑) Reverse. The region of chromosome III contained on the insert contains: 5 ORFs of significant size (29), including the *PGK1* gene (black boxes); the *SUF16* locus, encoding a glycine tRNA (black circle with arrowhead); and a Ty1 solo δ (or LTR; stippled box). The direction of each open reading frame or sequence of interest is indicated by a horizontal arrow over the map. The insertion locations were determined by sequencing one or both insertion junctions.

Table 3. Statistics on insertions

Region of target	Number of insertions per kb target DNA
136,155 to 137,000	13
137,000 to 138,000	9
138,000 to 139,000	17
139,000 to 140,000	14
140,000 to 141,000	6
141,000 to 142,000	10
142,000 to 143,000	9
143,000 to 144,000	6
144,000 to 144,333	6

Mean interval distance = 102.3 $\pm$ 88.1

Mean number of insertions per kb target DNA = 10.2 $\pm$ 3.7

Number of insertions in genes/ORFs = 45

Number of genes/ORFs receiving insertions = 6

Orientation

Forward 34 (44%)

Reverse 44 (56%)

n = 78

Brachmann; T.Waldman, W.El-Deiry, and B.Vogelstein, personal communications). Hence, *in vitro* integration of artificial transposons is expected to work well with most or all plasmid targets, making it both a generally useful sequencing tool and a general method of integrating new DNA sequences into plasmid targets for a variety of applications.

DISCUSSION

We have developed new methods for creating artificial transposons and efficiently inserting these transposons into plasmid targets *in vitro*. There are three key aspects of the process: i) the *in vitro* integration reaction is highly efficient, giving rise to thousands of integrations per reaction; with most plasmid targets, this efficiency approaches one insertion per phosphodiester bond, ii) the insertion process is sufficiently random that transposon integrations occur throughout target plasmid sequences, and iii) virtually any DNA sequence or combination of sequences can, in principle, serve as an artificial transposon. These three features combine to make this an extremely versatile method of generating recombinant DNA molecules.

The mechanism of moving the artificial transposon from donor to target plasmid differs from the *in vivo* retrotransposition mechanism employed by Ty1, and resembles more closely the 'cut and paste' mechanism used by Tn7 (31) and Tn10 (32). The first step involves the cleavage of the artificial transposon from a plasmid backbone. Although this step is accomplished with a restriction enzyme in our system, it resembles the first step of 'cut and paste' transposition. The second step is the integration of the excised transposon into a target site. This is accomplished with the machinery that mediates the integration of Ty1 during retrotransposition, namely Ty1 integrase in the form of VLPs. Thus, the mechanism of transposition used here incorporates features of both 'cut and paste' and retrotransposition mechanisms, and is carried out entirely *in vitro*.

We have focused our initial efforts on developing a specific application of this technology, i.e., *in vitro* insertion of 'primer island' artificial transposons into plasmid targets for the purpose of DNA mapping and sequencing. In addition to the features

mentioned above (efficiency of integration, randomness of insertion and flexibility of transposon), this system has other advantages compared with existing methods, including: i) the *in vitro* protocol is simple and highly reliable, ii) artificial transposons do not contain large terminal repeats which, in Tn5 and Tn10-based systems, hinder efficient access to sequences flanking the insertion junctions, and iii) the reaction is carried out completely *in vitro* and is therefore amenable to biochemical alteration and parameter optimization; this may be especially useful with unusual DNA templates such as those containing tandem sequence repeats, high GC content, or unusual template topology which might, in principle, represent difficult targets.

Importantly, transposon integration within targets was sufficiently random that insertions were recovered from all regions of target plasmids, and into all known or proposed functional regions. On the chromosome III insert, these included at least 6 genes or ORFs representing both polymerase II and III-transcription units (29). Unlike Ty1 integration *in vivo*, where ORFs are avoided (30), more than half (44/78) of the *in vitro* insertions were found in ORFs. Moreover, insertions were suitably spaced within ORFs to completely recover the corresponding nucleotide sequences with conventional dideoxy methods. Since transposon insertion within the ORF would be expected to disrupt gene function in most cases, these insertions would be predicted to be useful for mapping the functional regions of the 8 kb insert as well. Since the original linkage would be maintained throughout the analysis, correlating functional regions with map positions would be simplified. Thus using a set of such insertions, and an appropriate assay, combined physical and functional genetic analysis could be directly performed on a region of interest.

Based on the distribution of insertions (Fig. 4 and Table 3), Ty1 integrase-mediated integration *in vitro* appears, at a minimum, to be a nearly-random process. It may, in fact, be totally random. In contrast, this feature is not generally observed of other transposon systems adapted for DNA sequencing; instead, hotspots and coldspots of insertion frequently lead to a non-random distribution of insertions rendering these systems incapable of accessing large segments of DNA sequence, or to high levels of wasteful redundancy in other regions. These problems have been circumvented in some systems with mutant transposases which display altered target specificity (9). However, this approach provides only a limited relaxation of transposase-specified target specificity. It is known that host cell factors contribute to target specificity for both Tn10 (33) and Ty1 (30); this is most conveniently eliminated by the use of *in vitro* systems. Since artificial transposon integration by Ty1 integrase inherently displays random-like behavior *in vitro* (Fig. 4 and Table 3), it is directly adaptable for DNA sequencing.

Because our method of constructing artificial transposons is very versatile, transposons containing a variety of sequences can be constructed for a number of specific applications. For example, other markers could be inserted into the mcs site of pAT-1, including yeast and mammalian drug-selectable genes, generating marker cassettes that could act as transposons. One possible application would be 'marker addition', i.e., the insertion of a useful auxotrophic marker into an acceptable region of a plasmid of interest. For use in bacteria or yeast, for example, pAT-1 derivatives containing a variety of selectable markers in the mcs could be constructed, and the marker of choice could be added to a target plasmid with a simple *in vitro* integration reaction. Indeed, the products of a single integration reaction could be

viewed as an 'integration library' containing a collection of insertions, each clone containing a single insertion at a particular phosphodiester bond. Should it be necessary, an insertion at any specific phosphodiester bond could be identified with conventional library screening methods, using a junction oligonucleotide as a probe. Hence, using a custom artificial transposon, and applying the appropriate screening method, recombinant molecules of a desired structure could be recovered.

Other applications are also envisioned. For example, with the advent of 'long PCR' (34), artificial transposons could be very useful for mapping, analyzing, and restructuring large segments of genomes. Primers optimized for such applications could be engineered into the artificial transposon, as could rare restriction endonuclease sites (in addition to the *NorI* and other sites already present in the subterminal multicloning sites). Artificial transposons are also likely to be useful for analyzing the functional domains of genes and proteins by 'exon trapping', or mapping by insertional inactivation. They are also likely to be useful tools for making deletions and chimeric constructs containing drug resistance markers for transgenic and other studies.

In addition to the artificial transposon, the other two components of the system, i.e., the integrase and the target, are likely to be equally versatile. For example, it is possible that other integrases or transposases could effect an equivalent or nearly equivalent *in vitro* integration reaction. In addition, mutant integrases might also be useful. The specific properties of such integrases might together provide a wider range of integration preferences or frequencies. Also, rather than providing the integrase in the form of VLPs, purified integrases might display more activity or may be more stable, relative to VLP-associated integrases. In fact, Ty1 integrase has recently been purified (35). On the other hand, while purified integrases can efficiently carry out half-reactions *in vitro* (joining of one terminus), full reactions are inefficiently completed (36); VLPs may, therefore, provide a requisite environment for the efficient completion of full integrations.

It is envisioned that the reaction could also be extended to include other DNA targets. Lambda DNA has been used successfully as a target in similar reactions using Moloney murine leukemia virus (12) and Ty1 integrases (13,14) provided in the form of viral particles. Hence, it is very likely that targets such as cosmids, yeast and bacterial artificial chromosomes as well as P1 vectors and genomic DNA will work equally well as targets, leading to a variety of further applications.

The artificial transposon-based system for generating DNA sequencing templates can be readily applied to the development of high throughput, massively parallel DNA sequencing strategies. The high degree of randomness of insertion and the large fraction of clones generating useful sequence data mean that a shotgun approach to sequencing of large recombinant plasmids, including cosmids as well as P1 and bacterial artificial chromosomes, is feasible and highly suited to automation. Random doubly drug resistant colonies could be picked, their DNA extracted, and fed directly into an automated sequencing apparatus. All of these steps are amenable to automation. Because a single set of optimized primers can be used to sequence an entire set of plasmid derivatives, all of the steps can be done in parallel without operator intervention with regard to primer design and selection, etc. Hence, although artificial transposon-facilitated DNA sequencing is predicted to be very useful for small-scale sequencing projects, it may be even more useful for massive

projects such as the effort underway to map and sequence the human genome.

ACKNOWLEDGEMENTS

We thank Marc Vidal, Hanlee Ji, Shari Corin and the members of the Boeke lab for helpful discussions, and Lita Braiterman for help with VLP preparation. We also thank Elise Rose of Applied Biosystems for advice and encouragement, as well as Phil Hieter and Hanlee Ji for critical review of the manuscript. This work was supported by American Cancer Society postdoctoral fellowship PF-4040 to SD, NIH grant GM36481 to JB, and American Cancer Society Faculty Research Award FRA-366 to JB.

REFERENCES

- Smith, L.M. (1993) *Science* 262, 530–531.
- Itakura, K., Rossi, J.J., and Wallace, R.B. (1984) *Ann. Rev. Biochem.* 53, 323–356.
- Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular Cloning A Laboratory Manual, Second Edition*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. pp 13.2–13.104.
- Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., et al. (1992) *Nature* 356, 37–41.
- Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. (1989) *Current Protocols in Molecular Biology* 1, 7.2.1–7.2.20.
- Strathman, M., Hamilton, B.A., Mayeda, C.A., Simon, M.I., Meyerowitz, E.M., and Palazzolo, M.J. (1991) *Proc. Natl. Acad. Sci. USA* 88, 1247–1250.
- Phadnis, S.H., Huang, H.V., and Berg, D.E. (1989) *Proc. Natl. Acad. Sci. USA* 86, 5908–5912.
- Way, J.C., Davis, M.A., Morisato, D., Roberts, D.E., and Kleckner, N. (1984) *Gene* 32, 369–379.
- Kleckner, N., Bender, J., and Gottesman, S. (1991) *Methods Enzymol.* 204, 139–180.
- Adachi, T., Mizuuchi, M., Robinson, E., Appella, E., O'Dea, M., Gellert, M., and Mizuuchi, K. (1987) *Nucleic Acids Res.* 15, 771–784.
- Ahmed, A. (1985) *Gene* 39, 305–310.
- Brown, P.O., Bowerman, B., Varnus, H.E., and Bishop, J.M. (1987) *Cell* 49, 347–356.
- Eichinger, D.J. and Boeke, J.D. (1988) *Cell* 54, 955–966.
- Eichinger, D.J. and Boeke, J.D. (1990) *Genes Dev.* 4, 324–330.
- Braiterman, L. and Boeke, J.D. (1994) *Mol. Cell. Biol.*, in press.
- Braiterman, L. and Boeke, J.D. (1994) *Mol. Cell. Biol.*, in press.
- Sikorski, R.S., and Hieter, P. (1989) *Genetics* 122, 19–27.
- Sikorski, R.S., and Boeke, J.D. (1991) *Methods Enzymol.* 194, 302–318.
- Innis, M.A., and Gelfand, D.H. (1990) In: *PCR Protocols A Guide to Methods and Applications*. Academic Press, Inc., San Diego, CA pp 3–12.
- Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.
- Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular Cloning A Laboratory Manual, Second Edition*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. pp 1.53–1.110.
- Holmes, D.S., and Quigley, M. (1981) *Anal. Biochem.* 114, 193–197.
- Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. (1989) *Current Protocols in Molecular Biology* 1, 1.7.1–1.7.11.
- Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) *Molecular Cloning A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. pp68–69.
- Hanahan, D. (1983) *J. Mol. Biol.* 166, 557–580.
- Calvin, N.M., and Hanawalt, P.C. (1988) *J. Bacteriol.* 170, 2796–2801.
- Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. (1989) *Current Protocols in Molecular Biology* 1, 1.8.4–1.8.8.
- Braiterman, L.T., Monakian, G.M., Eichinger, D.J., Merbs, S.L., Gabriel, A., and Boeke, J.D. (1994) *Gene* 139, 19–26.
- Oliver, S.G., van der Aart, Q.J.M., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., and etc. (1992) *Nature* 357, 38–46.

30. Ji, H., Moore, D.P., Blomberg, M.A., Braiterman, L.T., Voytas, D.F., Natsoulis, G., and Boeke, J.D. (1993) *Cell* 73, 1007–1018.
31. Bainton, R.J., Kubo, K.M., Feng, J.N., and Craig, N.L. (1993) *Cell* 72, 931–943.
32. Kleckner, N. (1989) In: *Mobile DNA*, D.E. Berg and M.M. Howe, Eds. American Society for Microbiology, Washington, D.C. pp. 227–268.
33. Lee, S.Y., Butler, D., and Kleckner, N. (1987) *Proc. Natl. Acad. Sci. USA* 84, 7876–7880.
34. Barnes, W.M. (1994) *Proc. Natl. Acad. Sci. USA* 91, 2216–2220.
35. Moore, S.P., and Garfinkel, D.J. (1994) *Proc. Natl. Acad. Sci. USA* 91, 1843–1847.
36. Bushman, F.D., and Craigie, R. (1991) *Proc. Natl. Acad. Sci. USA* 88, 1339–1343.