Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 6213–6217, August 1990
Biochemistry

Probe mapping to facilitate transposon-based DNA sequencing

(transposon γδ/Tn*1000*/insertion localization/ordered sequence acquisition/genome project)

LINDA D. STRAUSBAUGH*, MICHAEL T. BOURKE, MARTIN T. SOMMER, MICHAEL E. COON, AND CLAIRE M. BERG

Department of Molecular and Cell Biology, The University of Connecticut, Storrs, CT 06269-3125

Communicated by Mary Jane Osborn, June 4, 1990

ABSTRACT **A promising strategy for DNA sequencing exploits transposons to provide mobile sites for the binding of sequencing primers. For such a strategy to be maximally efficient, the location and orientation of the transposon must be readily determined and the insertion sites should be randomly distributed. We demonstrate an efficient probe-based method for the localization and orientation of transposon-borne primer sites, which is adaptable to large-scale sequencing strategies. This approach requires no prior restriction enzyme mapping or knowledge of the cloned sequence and eliminates the inefficiency inherent in totally random sequencing methods. To test the efficiency of probe mapping, 49 insertions of the transposon γδ (Tn*1000*) in a cloned fragment of *Drosophila melanogaster* DNA were mapped and oriented. In addition, oligonucleotide primers specific for unique subterminal γδ segments were used to prime dideoxynucleotide double-stranded sequencing. These data provided an opportunity to rigorously examine γδ insertion sites. The insertions were quite randomly distributed, even though the target DNA fragment had both A+T-rich and G+C-rich regions; in G+C-rich DNA, the insertions were found in A+T-rich "valleys." These data demonstrate that γδ is an excellent choice for supplying mobile primer binding sites to cloned DNA and that transposon-based probe mapping permits the sequences of large cloned segments to be determined without any subcloning.**

The current commitment to determining the sequence of the genomes of humans and other model organisms is stimulating technical advances in molecular genetics and giving new insights into genome structure, gene function, and evolution. Techniques for the rapid determination of nucleotide sequence have been developed (1, 2), and vectors such as cosmids (3) and bacteriophage P1 (4) permit large genomic segments to be cloned in *Escherichia coli*. Typically, large cloned DNAs are fragmented so that relatively short pieces can be brought into juxtaposition with fixed primer binding sites for sequencing, either by subcloning or by the generation of nested deletions (5). Even larger DNA segments may be sequenced directly by oligomer walking (6), by using a library of primers (7), or by the ExoMeth method (8).

A fundamentally different sequencing approach that uses transposons to provide mobile sites for primer binding has several attractive features: subcloning and/or the isolation of nested deletions are not required, only two primers are needed, and large cloned fragments can be sequenced directly. Transposons have been successfully used to provide DNA sequencing primer sites in plasmids, phage λ, and in chromosomal DNA (9–14). However, even with transposons, sequence acquisition is either random and therefore highly repetitive, or it requires the time-consuming construction of a restriction map to locate the insertions. Furthermore, the usefulness of transposon-borne primer sites may be limited if the transposon inserts nonrandomly.

We demonstrate here an efficient method for localizing and orienting mobile primer sites using the transposon γδ (Tn*1000*) (15, 16) in a probe-mapping strategy. This involves digestion of the plasmid with a single restriction enzyme that recognizes sites in the transposon and vector, but not in the cloned fragment. Hybridization patterns with specific probes permit the transposon's location and orientation to be determined. Probe mapping of transposons requires no knowledge of the target DNA sequence or prior restriction enzyme mapping. In addition, we show that the previously reported preference of γδ for A+T-rich 5-base-pair (bp) target sites (11) does not translate into regional insertion specificity, even though the *Drosophila melanogaster* target fragment selected for this study has regions of both high and low G+C content. The use of probe-mapped γδ insertions to provide primers for DNA sequencing avoids the duplication of effort inherent in totally random sequencing methods and shows particular promise for adaptation to the analysis of large cloned fragments.

MATERIALS AND METHODS

Transposon Mutagenesis. The *E. coli* K-12 strains used were MG1063 (F$^+$ *recA56 thi*) (15), containing monomeric pDmHJ6.7, and CBK861 [F$^-$, *rpsL109 thyA* Δ(*proB-lac*) *recA*]. Plasmid pDmHJ6.7 contains a 6.7-kilobase (kb) fragment of *D. melanogaster* DNA cloned into the *Eco*RI site of pBR325 (Amp^r, Tet^r, Cam^s) (M.T.B. and L.D.S., unpublished data). Transposition of γδ from the F factor to pDmHJ6.7 occurred in MG1063. The transpositional cointegrate was transferred by conjugation to CBK861, where it underwent resolution. Therefore, each plasmid in CBK861 contained a single γδ insertion (see refs. 15–17 for additional information).

To ensure that every insertion was independent, plate matings were used. MG1063 (pDmHJ6.7) was streaked on L agar (LA) plates containing ampicillin (100 μg/ml) and the plates were incubated at 37°C. After overnight growth, these plates had both confluent growth and small isolated colonies. They were replica plated to a lawn of CBK861 (10^7 cells) freshly spread on LA plates containing thymine (0.079 mM), tetracycline (25 μg/ml), and streptomycin (200 μg/ml) and were incubated at 37°C. After 1 or 2 days of incubation, isolated single colonies were picked from the replica plates and purified on the same medium.

Probe Mapping of Transposon Insertions. Plasmid DNA was prepared by a rapid boiling method (18) from 1.5-ml cultures that had been grown overnight in L broth plus thymine, tetracycline, and streptomycin. Samples of DNA were digested with *Eco*RI according to the supplier's (BRL) instructions and the resulting fragments were electrophoresed in 1% agarose, visualized by ethidium bromide staining,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*To whom reprint requests should be addressed.

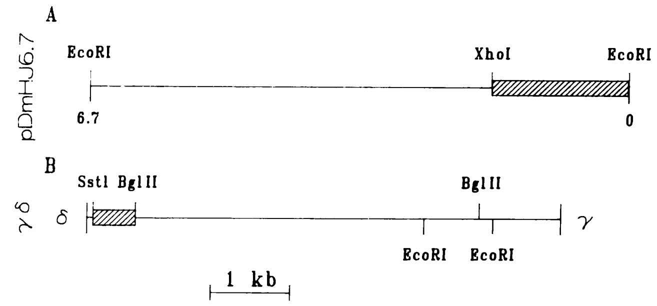

FIG. 1. Partial restriction enzyme maps of the *D. melanogaster* target DNA (*A*) and of γδ (*B*). Restriction enzyme recognition sites used in probe mapping are shown. Probes were synthesized from the restriction fragments marked by hatched boxes.

and transferred to nitrocellulose by standard procedures. Restriction enzyme fragments for probe preparation (a 0.5-kb *Bgl* II/*Sst* I transposon δ end fragment and a 1.4-kb *Xho* I/*Eco*RI terminal pDmHJ6.7 fragment; Fig. 1) were labeled in low melting temperature agarose (BRL) with [^{32}P]dCTP (New England Nuclear) by a random-primer method (19). Procedures for hybridization, rinsing, and autoradiography have been described elsewhere (20). Prior to rehybridization, filters were stripped by alkali treatment (21).

DNA Sequencing. Plasmid DNA was prepared for sequencing by alkaline lysis followed by polyethylene glycol precipitation (22). Oligonucleotide primers for the γ and δ transposon ends were 5′-TCAATAAGTTATACCAT-3′ and 5′-GAATTATCTCCTTAACG-3′, respectively (11). Dideoxynucleotide chain-termination reactions (2) were conducted with deoxyadenosine 5′-[γ-[^{35}S]thio]triphosphate (New England Nuclear) using a commercially available kit (Pharmacia sequencing kit). Reaction products were electrophoresed and visualized by standard DNA sequencing techniques (23).

RESULTS

Rationale of Probe Mapping. Probe mapping permits the efficient localization of insertions and bypasses the time-consuming steps involved in the prior generation of a restriction enzyme map of the target DNA and the subsequent mapping of the transposon-mutagenized plasmids to determine the insertion sites. Thus, it is expected to compete very favorably with random sequence acquisition methods. To test the effectiveness of probe mapping, we exploited existing features of wild-type γδ and of a 6.7-kb *D. melanogaster* DNA *Eco*RI fragment cloned into pBR325. γδ, a member of the Tn*3* transposon family, is 6 kb long and contains two internal *Eco*RI sites (see refs. 15 and 16). Radioactive probes were prepared from a unique subterminal restriction enzyme fragment of γδ and from a terminal restriction fragment of the cloned *D. melanogaster* target DNA (Fig. 1). This particular eukaryotic fragment was chosen to test probe mapping because prior partial sequence determination had revealed regions of atypical base composition that would be useful for testing the randomness of γδ insertion. Furthermore, since the restriction map of this fragment was known, it could be used to design the probe.

To determine the location and orientation of transposon insertions, plasmid DNA samples were digested with *Eco*RI and the four resulting fragments were separated in agarose gels. Plasmids with an insertion in the cloned fragment yielded two fragments of fixed sizes, one of 6.0 kb derived from the vector and the other of 0.8 kb derived from within the transposon, and two variably sized fragments, each derived from one end of the transposon plus adjacent cloned DNA (Fig. 2*B*; lanes 2, 5, 7, and 9). Plasmids with insertions in the vector were identified by the presence of the 6.7-kb intact cloned DNA (Fig. 2*B*; lanes 1, 3, 4, 6, and 8). Of 85 plasmids examined, 49 (58%) contained the γδ insertion in the cloned fragment, consistent with its relative target size. After transfer of the digestion products to nitrocellulose, the inserts were mapped by hybridization to the probes. One of the two variably sized fragments in each lane hybridized to the probe specific for the δ end of the transposon (Fig. 2*A*), and the other contained the γ end of the transposon. Similarly, for plasmids with inserts in the cloned region (lanes 2, 5, 7, and 9), one of the variably sized fragments hybridized to a probe specific for one end of the cloned fragment (Fig. 2*C*). In two cases (lanes 5 and 7), both probes hybridized to the same fragment. From the hybridization patterns, the location and orientation of each insertion were determined (Fig. 3*A*).

γδ Insertion Specificity. A central issue in the effectiveness of any transposon in facilitating DNA sequencing is its randomness of insertion. Many analyses have shown that γδ insertions tend to be random along the length of a target fragment (see ref. 16), although the 5-bp target sites duplicated upon insertion are very A+T-rich (11). The target DNA chosen for probe mapping provided a substrate for detecting possible transposon insertion bias because it contains regions of distinctly different base compositions, ranging from 23% to 65% G+C (Fig. 3*B*). Nonetheless, the distribution of γδ insertions in this nonuniform target DNA sequence was not strongly biased (Fig. 3*A*).

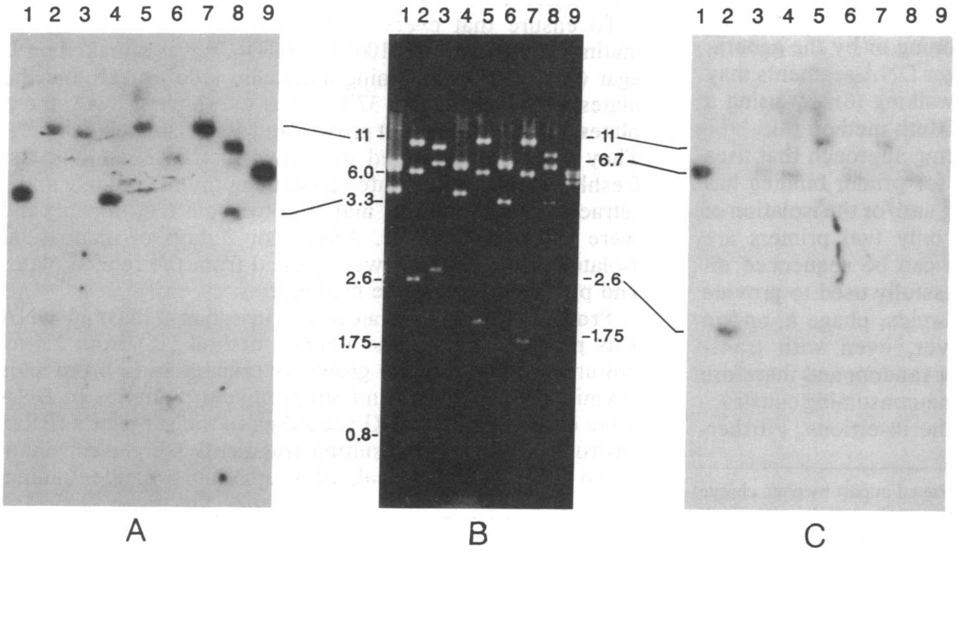

FIG. 2. Sample data for probe mapping. Mutagenized plasmids were digested with *Eco*RI, and the resulting fragments were separated in 1% agarose gels and stained with ethidium bromide (*B*). After transfer to nitrocellulose, fragments were visualized by hybridization to a probe specific for the δ end of the transposon (*A*). After autoradiography, the hybridized probe was removed by alkali treatment, and the filter was rehybridized with a probe specific for one end of the target DNA (*C*). Conditions for hybridization, washing, and autoradiography have been described (20). The 4.2-kb band in lane 8 is due to partial digestion.

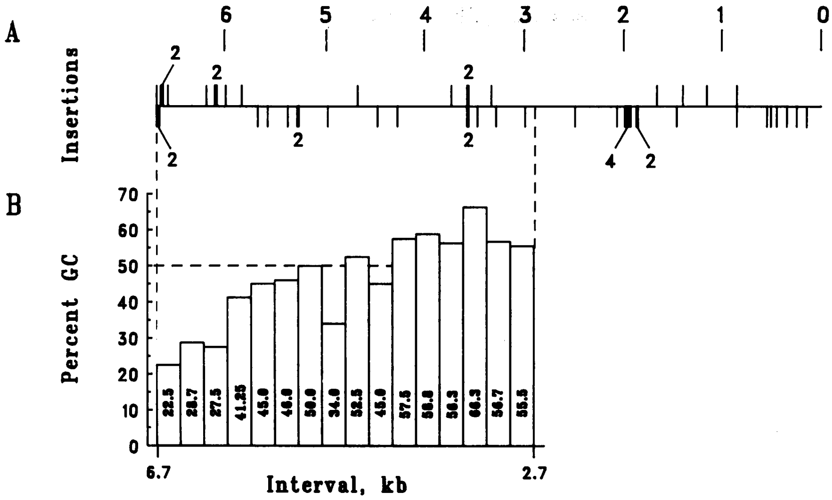

FIG. 3. Distribution of $\gamma\delta$ insertions in the target DNA. (*A*) Location of $\gamma\delta$ insertions. Lines above the bar indicate that the δ end of the transposon is on the left; lines below indicate that the γ end of the transposon is on the left. Heavy lines with numbers indicate multiple insertions within 100 bp of each other. (*B*) Histogram of G+C content (calculated in 250-bp intervals) of that portion of the target DNA previously sequenced (M.T.B. and L.D.S., unpublished data).

An examination of three insertions with landing sites located in regions of low (25%), medium (45%), and high (65%) G+C composition was made (Table 1). No obvious shared features of the DNA sequences adjacent to the landing sites were revealed. The DNA sequence of the 5-bp landing sites that were duplicated by $\gamma\delta$ insertions at 17 sites are shown in Table 2. The duplications were A+T-rich, although less so than reported previously (11). We have also compared the G+C composition of the landing sites with those of their immediate environments. There is no simple relationship between the two, although in 12/14 cases the 5-bp landing sites have lower G+C content than their immediate surroundings (Table 2). Another way to examine landing sites is by considering average G+C composition. In G+C-rich regions, insertions were found in A+T-rich valleys (Fig. 4). These probe-mapping and DNA-sequencing studies show that $\gamma\delta$ insertions are quite randomly distributed. In other words, although $\gamma\delta$ target *sites* tend to be A+T-rich, there is no obligate requirement for an A+T-rich *region*.

Statistics of the Approach. Probe mapping of transposon-borne primer sites will be advantageous only if this additional step is practical in terms of materials, time, and labor. A previously developed statistical formula (24) may be used to calculate mapping requirements, assuming random insertions. Given a circular plasmid of length t, the probability that n arcs of length a will cover the entire circle is

$$P(N_a \leq n) = \sum_{j=0}^{n}(-1)^j\binom{n}{j}\left(1-\frac{ja}{t}\right)_{(+)}^{n-1}.$$

[The ambiguous symbol $(x)_{(+)}^{y}$ indicates that the function vanishes for $x \leq 0$ and equals $(x)^y$ when $x \geq 0$. Therefore, whenever $(ja/t) \geq 1$, the last term in the equation becomes 0. The conditional statement, $P(N_a \leq n)$ is the probability that the total number of arcs of size a randomly distributed on a circle sufficient to cover the entire circle is less than or equal to n.]

Table 3 presents the theoretical numbers of random insertions that must be mapped in plasmids of different sizes to achieve a 90% probability of obtaining complete single-strand sequence information, assuming 375-, 500-, or 1000-bp sequence acquisition from each primer. Although double-strand sequencing has some technical difficulties, obtaining 375 bp from each transposon end was accomplished, even from this relatively large plasmid of almost 13 kb. The mapping requirements are quite reasonable for current technologies and will become even more favorable as sequencing efficiency improves (see columns 2 and 3 in Table 3). In practice, only a fraction of the insertions are chosen for sequencing and they can be selected to provide double-strand coverage.

Efficacy of Sequencing with Probe-Mapped Primer Sites. Compared with the use of unmapped transposons to provide primer binding sites for DNA sequencing, the localization and orientation of transposon insertions by probe mapping requires the additional step of a very limited restriction enzyme analysis. The advantages of bidirectional, nonrandom (and therefore nonredundant) sequencing makes this additional step well worth the effort. For example, to sequence a 6.7-kb fragment of DNA by random methods, at least a 6-fold excess, which corresponds to 40 kb, is needed. This also involves additional efforts in computer data entry and sequence alignment. However, with probe-mapped insertions (Fig. 3), only 20 insertions would be required to efficiently generate 14 kb of sequence information (estimating 350 bp from each transposon end). It is important to note that although 49 insertions were mapped to the cloned fragment, only a fraction of these would be actually used for sequencing purposes. In addition to the significantly lowered amount of sequence data that must be collected and entered, very little computer time is required for alignment since the linkage information has already been acquired by probe mapping. Moreover, in random methods, the location and extent of sequence gaps due to nonrandom subcloning or to repetitive DNA are difficult to determine and thus introduce additional uncertainties that are not encountered in probe mapping.

DISCUSSION

Utilizing the transposon $\gamma\delta$ and a cloned *D. melanogaster* fragment in a test system, we demonstrate the efficiency and utility of probe mapping for the localization and orientation of transposon-borne primer sites. The probe mapping strategy permits nonrandom sequence acquisition with no *in vitro*

Table 1. Sequences flanking selected $\gamma\delta$ landing sites

Insert	Sequence
1	TTTTCGTCAC TTTCCATTTT TAAAAGGATG **TAATA** TGGGTTTTGA GACTTTCAAA TGTTGATTGA
37	ATAAGCTTTC CATTGACCAC GGCCGATTTT **AAGTA** CTTCTTGATG AATGGCGCTA ACTTTTGGGC
36	CTGAGCATGG AGAAGCGTGG TCAGGTGCCA **AAGAT** CTTCCACGTC AACTGGTTCC GCAAGAGTG

Duplicated target sequence is in boldface type. See Fig. 4 for details.

Table 2. Transposon landing sites

Insert	Mapped position	Target sequence	% G+C
32	6.6	TAATG	43
31	6.6	TAAAC	57
1	6.6	TAATA	29
8	6.6	ATAAA	—*
80	6.2	GATTA	—*
41	5.7	AGATA	—*
16	5.4	ACAAA	48
37	5.4	AAGTA	43
29	5.3	AAGAT	55
43	3.6	AAGTT	54
60	2.5	AACTC	54
12	2.1	TTAAA	32
14	2.0	GCAGT	42
36	2.0	AAGAT	54
25	1.5	GTATA	49
58	1.4	AACAC	45
62	0.3	GTGAC	49

The target sequences are presented in the 5′–3′ γ end of the insert. The percentage G+C was calculated for 65 bases (the duplicated sequence plus 30 bases on each side).

*Sequence information derived from only one primer.

cloning steps, using only two oligonucleotide primers that are specific for unique subterminal segments at each end of $\gamma\delta$. This approach, which requires no prior restriction enzyme mapping or knowledge of the cloned sequence, eliminates the inefficiency inherent in totally random sequencing methods.

Our data are in agreement with previous data that showed a preference for A+T-rich 5-bp target sites (11). This bias is not, however, reflected in regional specificity; $\gamma\delta$ insertions were distributed randomly in segments of nonuniform DNA composition (Fig. 3). Nonetheless, an examination of the landing sites shows that in G+C-rich regions, they consistently occur in A+T-rich "valleys" (Fig. 4; data not shown). Thus, it is likely that $\gamma\delta$ insertions occur by a relatively nonspecific interaction with target DNA followed by local searching for an A+T-rich landing site. The essentially random distribution of $\gamma\delta$ insertions makes this transposon a superior candidate for the development of sequencing strategies.

Table 3. Number of mapped insertions required to achieve a 90% probability of sequencing entire plasmid

Size of plasmid, kb	Sequence obtainable from single insertion, kb		
	0.75	1	2
5	37	25	10
10	87	61	25
15	141	100	43

These values were obtained by applying the equation presented in *Results*. These calculations assume that sequencing primers complementary to each transposon end would be used.

The efficiency of a probe-mapping strategy could be significantly improved with several minor modifications. Small vectors and minitransposons containing different subterminal recognition sites for restriction enzymes that cut target DNA infrequently would reduce probe mapping to a single hybridization step, since mapping such asymmetric sites in the vector would eliminate the need for plasmid-specific probes. In addition, the significantly smaller size of a minitransposon provides a positive trade-off, permitting the mapping of corresponding larger cloned fragments. Multiplex sequencing (25) could also be adapted for use with transposons that are engineered to contain different oligonucleotide "tags."

Transposon-based probe mapping can also be used to increase the power of other mapping, amplification, and

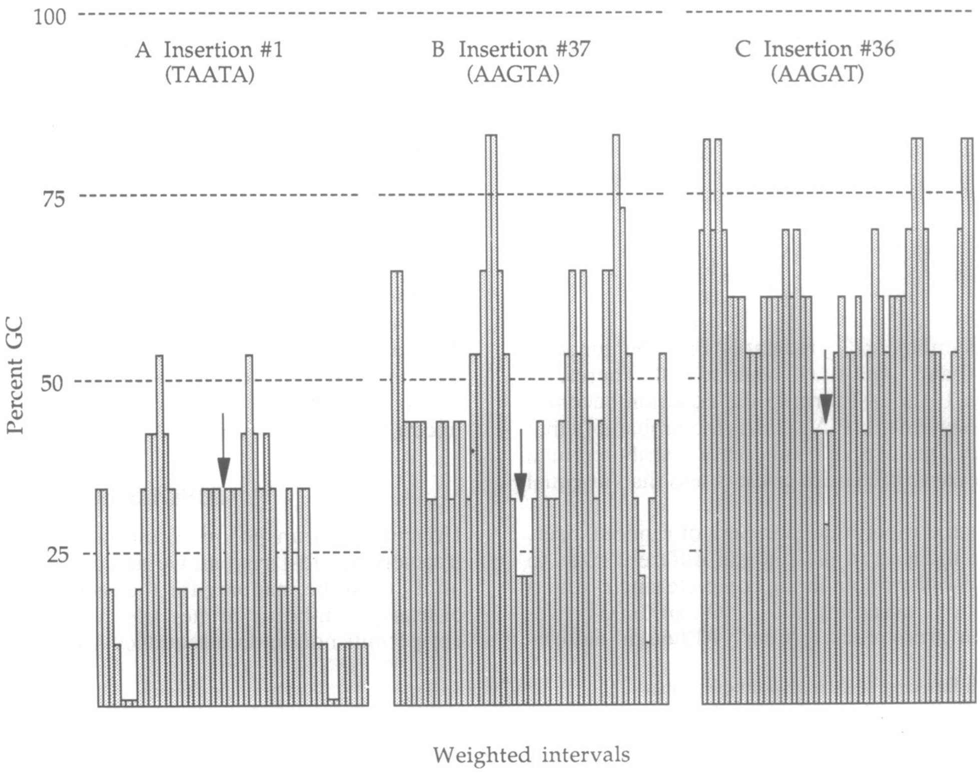

FIG. 4. Graphic representation of base composition of DNA surrounding selected $\gamma\delta$ landing sites. The IBI-Pustell DNA sequence analysis program was used to generate a graph of G+C base composition (using a weighted 10-bp interval) over the 100 bp surrounding selected $\gamma\delta$ landing sites (arrows).

sequencing techniques. For DNAs of unknown sequence cloned in large-capacity vectors, probe-mapped transposons can be used (*i*) to provide mobile anchors for polymerase chain reaction amplification (26); (*ii*) as mobile substrates for limited DNA sequencing to obtain new sequence tagged sites (27); and (*iii*) to supply mobile pairs of appropriate rare restriction sites for ExoMeth sequencing (8). Thus, in conjunction with improved vectors and engineered minitransposons, probe-based mapping of transposon insertions has significant potential for improving the efficiency of diverse methods for DNA analysis and characterization.

We thank Douglas Berg and Lin Liu for many stimulating discussions and critical reading of the manuscript, George Church and Gan Wang for suggesting the asymmetric mapping strategy, Joseph Glaz for help with the statistical analysis, Jeffrey Gorra for isolating the insertions, and Kelley Rochefort for typing the manuscript. We also acknowledge Paul Betts and the Biotechnology Center of the University of Connecticut for providing oligonucleotide primers and facilities for computer analysis of DNA sequence data. This work was supported by National Institutes of Health Grant GM28680, National Science Foundation Grant DMB8802310, Department of Energy Grant DE-FG02-89ER-60862, and Grants from the University of Connecticut Research Foundation.

1. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
2. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
3. Collins, J. & Hohn, B. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4242–4246.
4. Sternberg, N. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 103–107.
5. Henikoff, S. (1984) *Gene* **28**, 351–359.
6. Strauss, E. C., Kobori, J. A., Siu, G. & Hood, L. E. (1986) *Anal. Biochem.* **154**, 353–360.
7. Studier, F. W. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6917–6921.
8. Sorge, J. A. & Blinderman, L. A. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9208–9212.
9. Ahmed, A. (1987) *Methods Enzymol.* **155**, 177–204.
10. Adachi, T., Mizuuchi, M., Robinson, E. A., Appella, E., O'Dea, M. H., Gellert, M. & Mizuuchi, K. (1987) *Nucleic Acids Res.* **15**, 771–784.
11. Liu, L., Whalen, W., Das, A. & Berg, C. M. (1987) *Nucleic Acids Res.* **15**, 9461–9469.
12. Chow, W.-Y. & Berg, D. E. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6468–6472.
13. Nag, D. K., Huang, H. V. & Berg, D. E. (1988) *Gene* **64**, 135–145.
14. Phadnis, S. H., Huang, H. V. & Berg, D. E. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 5908–5912.
15. Guyer, M. S. (1983) *Methods Enzymol.* **101**, 362–363.
16. Berg, C. M., Berg, D. E. & Groisman, E. A. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (*Am. Soc. Microbiol.*, Washington, DC), pp. 879–925.
17. Liu, L. & Berg, C. M. (1990) *J. Bacteriol.* **172**, 2814–2816.
18. Holmes, D. S. & Quigley, M. (1981) *Anal. Biochem.* **114**, 193–197.
19. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **137**, 266–267.
20. Fitch, D. H. A., Strausbaugh, L. D. & Barrett, V. (1990) *Chromosoma Berlin* **99**, 118–124.
21. Bucheton, A., Paro, R., Sang, H. M., Pelisson, A. & Finnegan, D. J. (1984) *Cell* **38**, 153–163.
22. Kraft, R., Tardiff, J., Krauter, K. S. & Leinwand, L. A. (1988) *BioTechniques* **6**, 544–546.
23. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY), Vol. 2, pp. 13.3–13.104.
24. Glaz, J. & Naus, J. (1979) *Ann. Probab.* **7**, 900–906.
25. Church, G. M. & Kieffer-Higgins, S. (1988) *Science* **240**, 185–188.
26. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239**, 487–491.
27. Olson, M., Hood, L., Cantor, C. & Botstein, D. (1989) *Science* **245**, 1434–1435.