Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 14680–14685, December 1996
Genetics

Transposon Tc1-derived, sequence-tagged sites in *Caenorhabditis elegans* as markers for gene mapping

(genome project/forward genetics/reverse genetics)

HENDRIK C. KORSWAGEN*, RICHARD M. DURBIN†, MIRIAM T. SMITS*, AND RONALD H. A. PLASTERK*‡

*Division of Molecular Biology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands; and
†The Sanger Centre, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, United Kingdom

William B. Wood, University of Colorado, Boulder, CO, September 23, 1996 (received for review June 10, 1996)

ABSTRACT We present an approach to map large numbers of Tc1 transposon insertions in the genome of *Caenorhabditis elegans*. Strains have been described that contain up to 500 polymorphic Tc1 insertions. From these we have cloned and shotgun sequenced over 2000 Tc1 flanks, resulting in an estimated set of 400 or more distinct Tc1 insertion alleles. Alignment of these sequences revealed a weak Tc1 insertion site consensus sequence that was symmetric around the invariant TA target site and reads CAYATATRTG. The Tc1 flanking sequences were compared with 40 Mbp of a *C. elegans* genome sequence. We found 151 insertions within the sequenced area, a density of ≈1 Tc1 insertion in every 265 kb. As the rest of the *C. elegans* genome sequence is obtained, remaining Tc1 alleles will fall into place. These mapped Tc1 insertions can serve two functions: (*i*) insertions in or near genes can be used to isolate deletion derivatives that have that gene mutated; and (*ii*) they represent a dense collection of polymorphic sequence-tagged sites. We demonstrate a strategy to use these Tc1 sequence-tagged sites in fine-mapping mutations.

Within the next few years the complete genomic sequence of several organisms, including the nematode *Caenorhabditis elegans*, will be available (1–3). This will drastically alter molecular biology; all genes will be cloned and sequenced. The next challenge for biology will be to relate these genes to phenotypes and function (4). This requires efficient methods for targeted gene disruption and efficient strategies for mapping mutations of known phenotype to the sequence map.

The *C. elegans* transposon Tc1 is an effective tool in both these genetic approaches: Tc1 insertions can be used to inactivate genes, and they can also serve as convenient genetic markers for mapping mutations. Tc1 is a member of the Tc1/mariner family of transposons and is present in all *C. elegans* strains analyzed (5–8). The copy number of Tc1 varies among different strains (9–11). The commonly used wild-type strain Bristol N2 has about 30 copies of Tc1; this number is stable, since germ-line transposition of Tc1 is absent in this strain (12, 13). Mutator strains show active germ-line transposition of Tc1 (12, 14–17) and are used for transposon tagging (15) and targeted gene disruption in *C. elegans* (18). Insertion of Tc1 can directly inactivate a gene. If this is not the case, the Tc1 insertion can be used to generate deletion derivatives (18). As well as using mutator strains for direct mutation, they can also be used as a source of genetic markers for mapping purposes. Mutator strains can contain large numbers of polymorphic Tc1 insertions, as high as 500 Tc1 insertions in strain Bergerac BO (11). One approach has been to genetically identify a Tc1 insertion that maps close to a mutation of interest and to use this insertion to locate the mutation on the physical map (19). A more powerful application of polymorphic Tc1 insertions is as sequence-tagged sites (STSs) (20). Individual Tc1 insertions can be visualized by PCR using primers directed to the transposon and the flanking genomic sequence. Williams *et al.* (21, 22) cloned and sequenced a set of 40 strategically located Bergerac BO Tc1 insertions. Combinations of these insertions can be detected by multiplex PCR, and linkage of a mutation to any of the Tc1 STSs can be assessed by PCR on single progeny of crosses to Bergerac BO. Tc1 STSs have a number of advantages over conventional genetic markers (21, 22). (*i*) STS analysis enables genome-wide mapping of mutations. Depending on the set of PCR primers used, a mutation can be mapped to any of the six linkage groups or mapped further to a specific region. This eliminates laborious two- and three-factor crosses using conventional visible genetic markers. (*ii*) Multiple Tc1 STS markers can be analyzed in a single cross. Apart from allowing efficient mapping strategies, this enables mapping of multiple genes involved in complex phenotypes such as aging (23). (*iii*) Tc1 STSs have no associated phenotypes, which is useful when mapping mutations with subtle phenotypes. (*iv*) Tc1 STS markers can be used to efficiently map lethal mutations because dead embryos and larvae can serve as substrates for PCR. (*v*) Tc1 STSs provide a direct link from genetic data to the physical map. A current limitation of Tc1 STSs, however, is the restricted number of Tc1 insertion site sequences that are available.

In this paper, we describe a method to shotgun sequence large numbers of polymorphic Tc1 insertions present in high Tc1 copy number strains and to map these insertions to the genomic sequence of *C. elegans*. Our approach makes optimal use of the available genome sequence; instead of using laborious genetic or physical methods, we mapped polymorphic Tc1 insertions by comparing short flanking sequences to the genomic sequence. We have obtained a Tc1 STS map with a density of about one Tc1 insertion in every 265 kb, and we show an approach to using these STSs in fine-mapping mutations.

MATERIALS AND METHODS

Nematode Culture. Nematodes were cultured as described by Sulston and Hodgkin (24). High Tc1 copy number strains used in this study were RW7000, which is a derivative of Bergerac BO (12), CB4000 [*sma-1*(*e30*)V] (J. Hodgkin, unpublished result in ref. 25), and KR1787 [*unc-13*(*e51*)I] (17). Strains used in validation of the vectorette amplification approach were NL233 [*prk-2*::Tc1(*pk26*)III] and NL300 [*gpa-2*::Tc1(*pk2*)V]. Strains used in mapping experiments were CB1489 [*him-8*(*e1489*)IV], CB164 [*dpy-17*(*e164*)III], and PB49 [*egl-5*(*n486*)*unc-36*(*e251*)III; *him-5*(*e1490*)V].

DNA analysis. Genomic DNA was isolated as described (18). The DNA was further purified by phenol/chloroform extraction and ethanol precipitation. Genomic DNA (100 ng) was digested with *Sau*3A as recommended by the supplier (New England Biolabs). After heat inactivation of *Sau*3A (15 min at 65°C), 15 pmol of annealed vectorette oligonucleotides (26) (top strand, pGATCCAAGGAGAGGACGCTGTCTGTCGAAGGTAAGGAACGGACGAGAGAAGGAGA; bottom strand,

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.

Abbreviation: STS, sequence-tagged site.
‡To whom reprint requests should be addressed.

TCTCCCTTCTCGAATCGTAACCGTTCGTACGAGAATCGCTGTCCTCTCCTTG) was ligated to the digested DNA in a 100-μl reaction containing ligation buffer (Boehringer Mannheim), 1 mM ATP, and 10 units of T4 DNA ligase. Ligation was overnight at 16°C. Three microliters of ligation reaction was used for PCR with the Tc1 primers Right 2 or Left 2 (18) and the vectorette primer N505 (CGAATCGTAACCGTTCGTACGAGAATCGCT). PCR was carried out for 38 cycles as described (18). Twenty microliters of PCR on vectorette-ligated DNA of strains NL233 and NL300 was separated on a 1% agarose gel, blotted onto a nitrocellulose filter, and hybridized with a genomic *prk-2* fragment spanning the site of the Tc1 insertion (27). For cloning and sequencing amplified fragments, the PCR on vectorette-ligated DNA of RW7000, CB4000, or KR1787 was diluted 10 times after 30 cycles of PCR, and 1 μl was used for nested PCR using the Tc1-inverted repeat primer N412 (GCAGTGGAATTCTTTTTGGCCAGCACTG) and the vectorette primer C337 (AAAGGGGCATGCCGTACGAGAATCGCTGTCCTC). These primers contain restriction sites for *Eco*RI and *Sph*I, respectively. The resulting product after 35 cycles of PCR was digested with these enzymes and purified over a Wizard DNA purification column (Promega). One-tenth of the purified PCR product was ligated into M13 *mp19* (New England Biolabs) digested with *Eco*RI and *Sph*I. Blue/white selection enabled identification of recombinant plaques. M13 was cultured for 6 h at 37°C with vigorous shaking in 600 μl of YT medium (27) containing *Escherichia coli* JM101 inoculated with a single plaque. Single-stranded M13 DNA was isolated from the culture supernatant by precipitation with 120 μl of 20% PEG-6000: 2 M NaCl for 15 min at 4 °C. The DNA was extracted with 0.1 M Tris·HCl (pH 8.0)-saturated phenol, precipitated, and dissolved in 15 μl of water. One microliter of DNA was used for PCR sequencing with dye-labeled M13 forward primer as recommended by the supplier (Applied Biosystems).

Computer Analysis and Statistics. Raw sequence data were transferred to a UNIX workstation (Sun Microsystems, Mountain View, CA). They were then edited using the program TRACE EDITOR (28) to remove Tc1 termini and vector sequences and to correct obvious base calling errors. Sequence data were compared with each other using the Smith–Waterman algorithm (29) and clustered at the 12.5% identity level. Multiple alignments of each cluster were made with CLUSTAL V (30), and consensus sequences were made using the HMMER package [S. Eddy, The HMMER package (http://genome.wustl.edu/eddy/hmm.html)]. Alignments were checked by hand and some sequences were reprocessed after further vector removal and editing. Consensus sequences were compared to genomic sequence using BLASTN (31), and significant matches were examined by eye. Standard contingency table χ^2 tests were used for analyzing the insertion site sequence distribution (32).

Genetic Mapping. Heterozygous *egl-5* (*n486*)*unc-36*(*e251*)/++; *him-8*(*e1489*)/+ and *dpy-17*(*e164*)/+; *him-8*(*e1489*)/+ males were mated with RW7000 hermaphrodites. F_1 progeny was selfed, and 20 Unc or Dpy F_2 animals were pooled in 40 μl of single worm lysis buffer (21, 22). Lysis was 60 min at 65°C, and denaturation was 15 min at 95°C. Two microliters was used for PCR with Tc1-specific primers Right 1 or Right 2 (18) and primers directed to the flanking genomic sequence (pkP417, TTCGCATATCTTTCTGAGAG; pkP409, TAGAGTGTGGAGAAATAGAC; pkP406, ATCGTCTGCAGAATTGCGCG; pkP410, TCTTTCAGGAACACAAGCCC; pkP402, AGAATCCGAAATAGAACGGC; pkP415, TCTGCGTCGCGACGGGAGGC; pkP403, TGAATTGATTCCAACGCCTC; pkP400, TTGCAAATGCTCCTGTAACC; pkP645, CTTCTGTGTTGGACCTCAGGC; pkP503, GTTGAAATGTACGCCACACTGC; pkP411, AATTAGTTGGTCCAAAATGG). Tc1 insertions were visualized in a single round of PCR or using a second round of PCR with nested primers (pkP417.2, TCACTTGCTAACAGAGTGAG; pkP409.2, CTGAGCAATTACGATGTGACG; pkP406.2, CAGTACTTCCCACGTCGTCATC; pkP410.2, TATTTGGCCACGTGTCCGTC; pkP402.2, CGTCCCACAAGATCAACAAG; pkP415.2, GATTCTCGAGGGATAGATCAG; pP403.2, GTTCCCTACTGTAAACATGC; pkP400.2, GAAAGGTCCATCGCCCTAACG) (as in refs. 18, 21, and 22).

RESULTS

Shotgun Sequencing of Tc1 Insertions. To sequence random Tc1 insertion sites, we generated strain- and orientation-specific libraries of Tc1 flanks. These libraries were constructed by cloning amplified left or right Tc1 flanking sequences directly into sequencing vectors. Flanking genomic sequences of Tc1 insertions present in the high Tc1 copy number strains RW7000, CB4000 and KR1787 was amplified using an anchored PCR-based method (Fig. 1*A*). RW7000 is a derivative of the natural isolate Bergerac BO (12) whereas CB4000 and KR1787 independently acquired mutator activity and high Tc1 copy numbers in a Bristol N2 background (17, 25). Genomic DNA was digested with the frequently cutting restriction enzyme *Sau*3A. This enzyme cuts the genomic DNA in fragments ≈0.2 kb in length. In addition, *Sau*3A cuts at known positions in Tc1, leaving part of the transposon sequence attached to the flanking DNA. A double-stranded oligonucleotide containing the appropriate 5′ overhang (termed a vectorette) was ligated to the digested DNA (26). The vectorette serves as an anchor to amplify Tc1 flanks, using one primer in the transposon terminus and one in the vectorette. The restriction sites of *Sau*3A in Tc1 are outside the terminal inverted repeats of Tc1, so flanks of the right and the left side of the transposon could be amplified separately. The specificity of this amplification method was assessed by Southern analysis of the total vectorette PCR product of a strain with a Tc1 insertion in the gene *prk-2* and an equivalent strain without this insertion. Hybridization with a genomic *prk-2* probe shows an amplified *prk-2* fragment in the *prk-2*::Tc1 strain but not in the strain without this insertion (Fig. 1*B*), demonstrating that this method can be used to specifically amplify the flanks of a complex mixture of Tc1 insertions.

To clone the amplified Tc1 flanks, a second round of PCR with nested primers containing unique restriction sites was performed. The PCR product was cloned directly into M13-sequencing vectors, and clones were sequenced using an automatic sequenator. Over 90% of these sequence tracks contained the Tc1 terminus and flanking genomic sequence. Sequence data were edited to remove Tc1 and vector sequences and obvious sequencing errors. A total of 2478 Tc1 flanking sequences were obtained from six different libraries: left and right flanks of strains RW7000, CB4000, and KR1787. Sequencing of random clones resulted in individual Tc1 flanks being represented by multiple sequence tracks. Therefore, we clustered homologous Tc1 flanking sequences into distinct alleles. Fig. 2 shows the distribution of the set of sequence tracks over the different Tc1 flanks they represent. Approximately one-half of the sequenced Tc1 flanks are represented by multiple sequence tracks. The other half are represented by single sequence tracks only. The distribution is clearly not random, reflecting an inherent bias of the amplification approach; some Tc1 flanks are amplified more efficiently than others. The distribution also shows that we have not reached saturation in sequencing all Tc1 flanks represented in the different libraries. We obtained 378 alleles of left flanks (283 of RW7000, 65 of CB4000, and 19 of KR1787; 11 insertions are present in more than one strain) and 340 alleles of right flanks (195 of RW7000, 84 of CB4000, 51 of KR1787, and 10 common insertions). The 21 Tc1 flanking alleles that are present in more than one strain are presumably Tc1 insertions that are shared by different *C. elegans* strains.

Tc1 Insertion Consensus Sequence. The genomic sequence surrounding the canonical TA target site of Tc1 was analyzed. To eliminate the noise from sequencing errors, only consensus sequences of Tc1 flanks with multiple sequence reads were aligned. We separately aligned the sequences of left and right Tc1 flanks and focused on the first seven positions from the TA target site of Tc1. The base distributions of the left and right Tc1 flanking sequences were not statistically different ($\chi^2_{21} = 30.3$; $P \approx 0.1$), suggesting a symmetric insertion site preference. For further analysis, we combined the data of left and right flanking sequences and asked if there were significant differences in base composition at each position (Table 1). There was a weak but significant preference for a T at position 1, an A or a G at position 2, and a G at position 4 and a highly significant preference for a T at position 3. Positions 5–7 do not show a significant bias in base composition. These results suggest that the consensus sequence for Tc1 insertion is symmetric; in simplified form, it can be written as CAYA**TA**TRTG. This symmetry is confirmed by the analysis of 83 Tc1 insertion sites mapped to the genomic sequence. Of these, 57 have a T at position +3, with 47/57 having a corresponding A at position −3, showing, at least for this position, a symmetric target site preference.

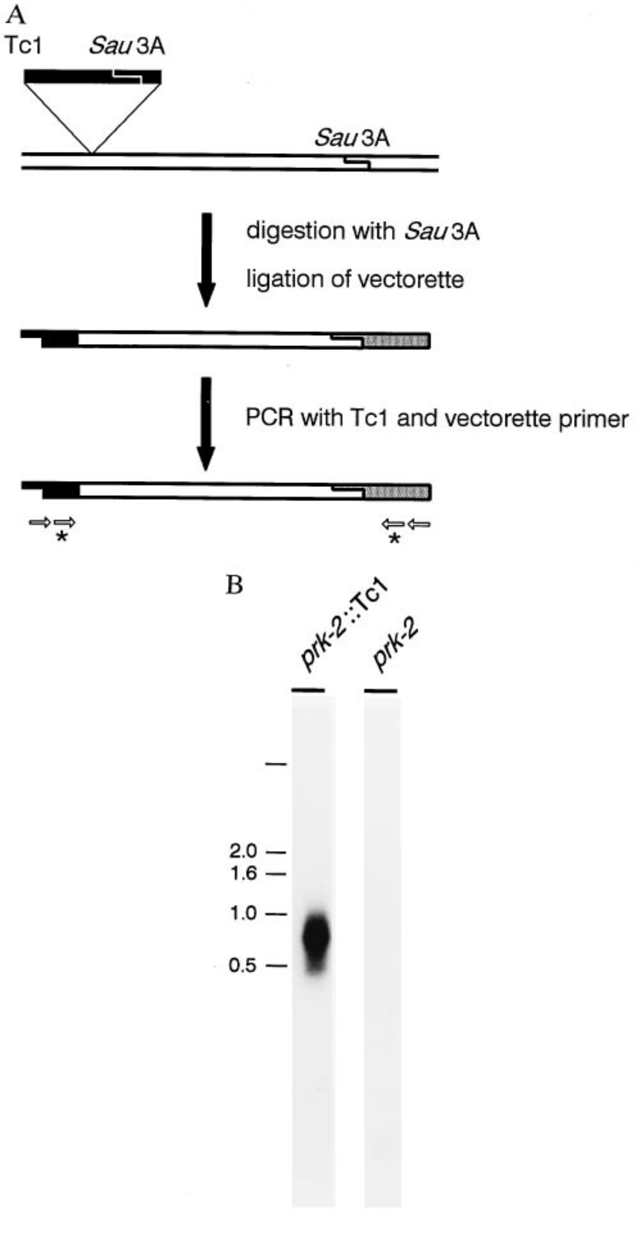

FIG. 1. Amplification of Tc1 flanking genomic sequence using an anchored PCR based method. (*A*) Schematic representation of the anchored PCR based approach. Genomic DNA was digested with *Sau*3A, resulting in small fragments consisting of Tc1 sequence and flanking genomic sequence. To amplify these fragments, a vectorette oligonucleotide anchor was ligated to the digested DNA. The Tc1 flanking fragments were amplified using primers that anneal to Tc1 and the vectorette anchor. Note that vectorette anchors can ligate at both ends of the fragments. The vectorette is, however, constructed in such a way that the vectorette PCR primers can only anneal after a complementary strand has been generated in the first round of PCR by synthesis from the Tc1-specific primer. Therefore, fragments containing only vectorette anchors, but no Tc1 sequence are not amplified. *, Internal Tc1 and vectorette primers containing restriction sites for *Eco*RI and *Sph*I, respectively. (*B*) Southern blot analysis of the total vectorette PCR product using genomic DNA of a strain with a Tc1 insertion in the gene *prk-2*, compared with a similar strain without this insertion. Hybridization with a ^{32}P-labeled *prk-2* genomic probe shows specific amplification of a *prk-2* fragment in the *prk-2*::Tc1 strain but not in the strain without this insertion.

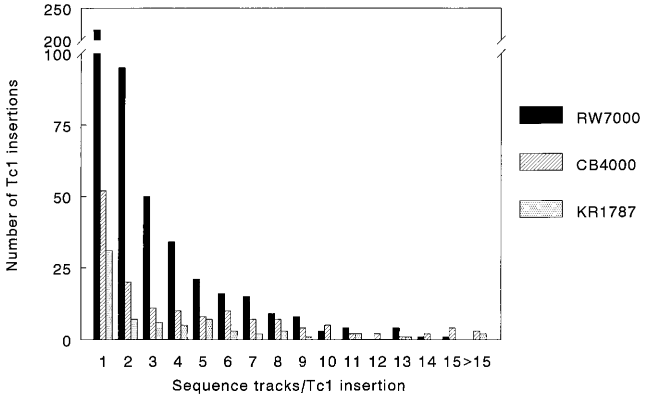

FIG. 2. Distribution of independent Tc1 flanking sequence reads over the different Tc1 insertions they represent. Sequence reads were clustered according to sequence similarity into distinct Tc1 insertion alleles. Approximately one-half of the sequenced Tc1 insertion sites is represented by multiple sequence reads.

Mapping Tc1 Insertion Alleles to the Sequenced Area of the Genome. The alleles of left and right Tc1 flanks were aligned with the 40 Mbp of *C. elegans* genome sequence available in June 1996. Mismatches between alignments were reexamined by recalling the original sequence data, and alignments were ignored when multiple positions remained ambiguous. Thus far, we have found 151 Tc1 insertions within the sequenced area of the genome. These matches are on the sequenced regions of the five autosomes and the X chromosome. As shown in Fig. 3, the mapped Tc1 insertions are distributed uniformly over the sequenced areas of the chromosomes, and, as expected from the independent origin of RW7000, CB4000, and KR1787, the pattern of Tc1 insertion sites does not overlap between the three strains. Twenty-seven Tc1 alleles showed matches to multiple regions in the genome. These insertions are located within repeats (like the rDNA cluster) or duplicated regions and consequently can not be mapped. As more genomic sequence becomes available, the remaining sequenced Tc1 insertions will fall in place. Detailed information on the location of these Tc1 insertions is available through the *C. elegans* data base ACeDB (S. Jones, personal communication) (2, 3) and at http://www.sanger.ac.uk/~rd/tc1.polyinfo.html.

To confirm that the Tc1 insertions mapped in this study were present in the germ line of the strains examined and were not the result of cloning and sequencing somatic insertions, we tested 12 of the mapped Tc1 insertions by PCR or Southern blot analysis. Five of these insertions were identified by both left- and right-sequenced Tc1 flanks, whereas the other seven insertions were identified by one sequenced Tc1 flank only. Eleven insertions were tested by PCR using a primer in Tc1 and a primer in the flanking genomic sequence. Each resulted in the expected PCR fragment (Fig. 4) only in the strain in which the insertion was identified. One insertion was confirmed by Southern blot analysis (data not shown). In addition, six Bristol

Table 1. Consensus sequence for Tc1 insertion

Position	T	A	1	2	3	4	5	6	7
G	0	0	66	**102**	24	**102**	73	68	53
A	0	344	27	**142**	55	93	113	131	117
T	344	0	**172**	56	**248**	72	99	95	104
C	0	0	79	44	17	77	60	50	70
χ^2_3*			102.3*	65.6*	261.2*	42.2*	3.0	9.3	3.2

Flanking genomic sequence of left- and right-sequenced Tc1 flanks were aligned around the TA target site for Tc1 insertion.
*χ^2_3 values greater than 11.3 are significant at the 1% level; values greater than 16.3 are significant at the 0.1% level.

N2 insertions (in cosmids C28F5, ZK1251, T22F3, ZK856, C50H2, and R173) were identified that had also been sequenced by the *C. elegans* genome consortium. Taken together, these data show that the Tc1 alleles mapped by sequence comparison with the genome sequence are indeed present in the germ line of the strains examined.

Strategy for Mapping Mutations Using Tc1 STSs. Tc1 insertions can serve as STSs: polymorphic sequences that can be visualized by PCR and can be used as genetic markers (20–22). We developed a strategy to use Tc1 STSs in mapping mutations. To locate a mutation to a specific region of the genome, the mutation has to be genetically linked to markers of known position. This involves scoring crossover frequencies between such markers and the mutation of interest. The closer a marker is located to the mutation, the lower the crossover frequency between the two. This poses a problem when fine mapping mutations; large numbers of animals have to be analyzed to observe such rare informative crossovers. Using Tc1 STSs it is, however, possible to do these analyses on multiple animals simultaneously. A strain containing the (recessive visible) mutation is crossed with one of the high copy number strains, and homozygous mutant F_2 progeny are analyzed for crossovers of different Tc1 STSs. To score crossovers of closely linked Tc1 markers, mutant F_2 progeny are pooled, and lysates are used for PCR analysis. Linkage of a Tc1 STS to the mutation should result in an underrepresentation of that Tc1 allele in homozygous mutant animals. An example is given in Fig. 5; shown are seven RW7000 Tc1 insertions to position two mutations whose positions are already known, *unc-36* (*e251*) and *dpy-17* (*e164*), to the physical map. Marker 7 is located close to *unc-36*, and markers 3 and 4 are close to *dpy-17*. Analysis of five pools of 20 animals for each of the two genes showed crossover of all markers except 6 and 7 in the case of *unc-36* and 2 and 3 for *dpy-17*, as was anticipated from the location of these mutations on the physical map. As expected, when a Tc1 marker close to one of the mutations did crossover onto the mutant chromosome, so do more distal markers. Using this mapping strategy, a mutation can be mapped to the resolution of the Tc1 STS map using only a single cross and analyzing only a limited number of pools for informative crossovers.

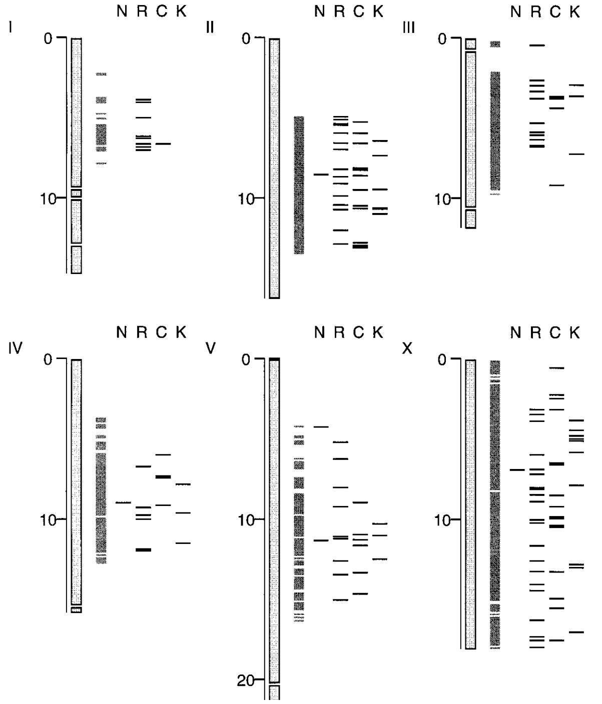

FIG. 3. Distribution of sequenced Tc1 insertion sites mapped to the genomic sequence. The physical maps of chromosomes I, II, III, IV, V, and X are represented by lightly shaded bars; the 40 Mbp of genomic sequence used in this study is indicated by darkly shaded bars. The scale is in approximate megabase pairs. Horizontal lines indicate the location of sequenced Tc1 insertion sites mapped in strains Bristol N2 (N), RW7000 (R), CB4000 (C) and KR1787 (K).

DISCUSSION

We describe a method to identify large numbers of transposon insertions present in high Tc1 copy number strains by shotgun sequencing. Flanking genomic sequence of Tc1 insertions was amplified using an anchored, PCR-based method, cloned in sequencing vectors, and sequenced. The total number of Tc1 insertions obtained depends on the efficiency with which different Tc1 flanks are amplified and the representation of these amplified Tc1 flanks in the collection of sequence tracks. The anchored PCR to amplify Tc1 flanks is biased. Depending on the location of the *Sau*3A site with respect to the Tc1 insertion, Tc1 flanking fragments will have different sizes. Consequently, large fragments will be amplified with lower efficiency, and very small fragments will be lost during cloning procedures. To reduce the effect of this bias, we separately

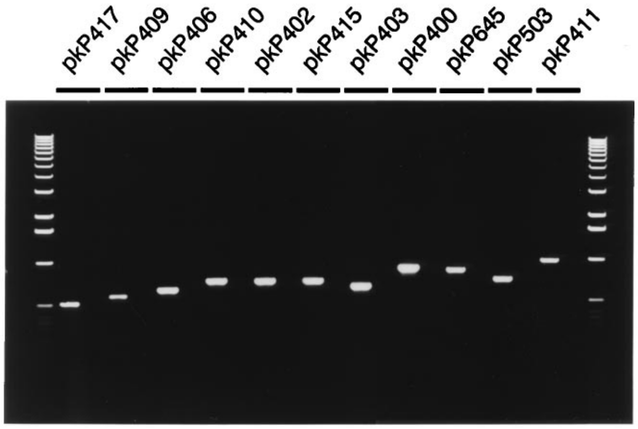

FIG. 4. PCR amplification of polymorphic Tc1 insertions mapped in high copy number strains. Tc1 flanking fragments were amplified using a primer in Tc1 and a primer in the flanking genomic sequence. In each case, the first lane shows the PCR product using template DNA of the high copy number strain in which the insertion was identified, and the second lane shows the PCR product using Bristol N2 DNA. Markers pkP417 to pkP400 are in RW7000, markers pkP645 and pkP503 are in CB4000, and marker pkP411 is in KR1787. A 1-kb DNA ladder (GIBCO/BRL) was used as a DNA fragment size marker.

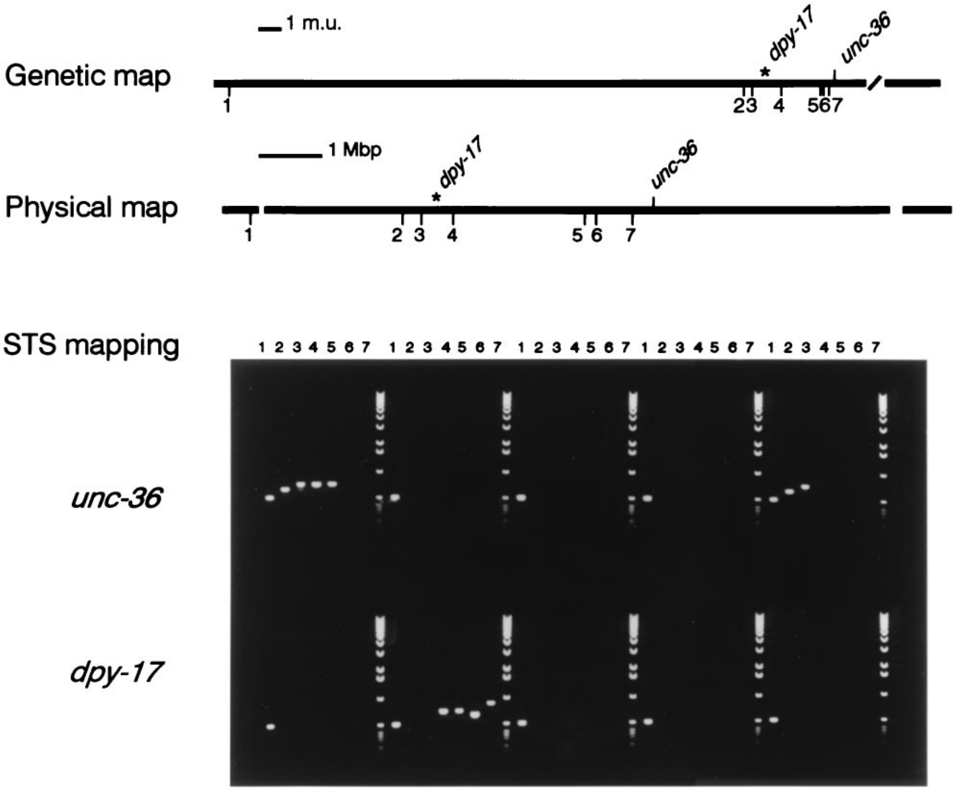

FIG. 5. Genetic mapping of *dpy-17* and *unc-36* using Tc1 STSs. The positions of *dpy-17*, *unc-36*, and RW7000 Tc1 STS markers 1–7 on the genetic and physical maps are indicated. The positions of markers 1–7 on the genetic map are extrapolated from the positions of these markers on the physical map. Note that *unc-36* has been mapped both to the genetic and physical maps whereas *dpy-17* has only been placed on the genetic map. Five pools of 20 Dpy or Unc F_2 animals from crosses to RW7000 were analyzed for crossovers of markers 1–7 by PCR using a primer in Tc1 and unique primers in the flanking genomic sequence of each Tc1 marker. Markers are as follows: 1, pkP417; 2, pkP406; 3, pkP410; 4, pkP402; 5, pkP415; 6, pkP403; and 7, pkP400.

amplified the left and right flanks of Tc1 insertions. In addition, some of the flanking sequences of strain RW7000 were derived from genomic DNA digested with *Nla*III, instead of *Sau*3A. Clustering of identical sequence tracks resulted in 378 left and 340 right Tc1 flanks. Approximately one-half of these were sequenced more than once. The other half was defined by single sequence tracks only, indicating that we probably did not sequence all Tc1 flanks represented in the different libraries.

Computer searches against 40 Mbp of genome sequence resulted in matches for 176 of the flanking sequences, defining 151 Tc1 insertions (in 25 cases, both left and right flanks were sequenced). Extrapolation from these numbers suggests that approximately 616 (718 × 151/176) different insertions are represented in this study. Assuming a genome size of 100 Mbp (1), this would predict an average density of one insertion every 160 kb whereas the observed frequency in the 40 Mbp compared directly was about one every 265 kb. A factor that may have contributed to this difference was the stringency used in examining the alignments between Tc1 flanking sequences and the genome sequence. Alignments that contained multiple mismatches were discarded. Therefore, Tc1 insertions may have been missed. To minimize the risk of mislocating STSs, we also excluded matches to repetitive sequences. This will result in an underestimation of both the total number of Tc1 insertions and the duplication between the sets of sequenced left and right Tc1 flanks. It is also possible that the density of Tc1 insertions is lower in the part of the genome sequenced so far, which concentrates on the central parts of the autosomes and on the X chromosome. However, the distribution of identified sites within the sequenced regions appeared to be uniform (Fig. 3).

The relatively low frequency of matching left and right flanks of any particular insertion site (25/151) again confirms that we did not identify all insertion sites in the strains studied. There appears to be a discrepancy between the degree of coverage estimated from this approach and that obtained by comparing the estimated number of sites sequenced (616) with the number of sites estimated experimentally: 700 total, made from ~500 for RW7000 (11), ~150 for CB4000 (J. Hodgkin, unpublished result in ref. 25) (data not shown), and ~60 for KR1787 (17). This may reflect either an underestimate in the previous experimental results or incompleteness in finding all matching sequences as described above.

Six of the eight Bristol N2 insertions present in the 40 Mbp of genome sequence were identified as well. Screening of the sets of Tc1 flanks against the genomic sequence of *C. elegans* resulted in only 27 insertions that mapped to multiple regions within the genome. This is a reflection of the relatively low abundance of repeated sequences within the *C. elegans* genome (1).

Apart from germ-line transposition, Tc1 is also active in somatic tissues (33, 34). This results in a background noise of somatic Tc1 insertions. The PCR approach used to amplify Tc1 flanks is biased toward germ-line insertions; in the mixture of digested genomic DNA, germ-line Tc1 flanks are present in a much higher template concentration than somatic insertions. Therefore, we did not expect to clone and sequence somatic insertions. Indeed, all 12 Tc1 alleles tested proved to be germ-line insertions. Nevertheless, before a Tc1 insertion mapped in this study is used for further experiments, it is advisable to check first that it is indeed an insertion that is present in the germ line of the strain in which it was identified.

We analyzed the genomic sequence surrounding the canonical TA target site of Tc1 insertion. Previous studies based on small numbers of insertion sites suggested a variety of related consensus sequences that were approximately palindromic (25, 35). Alignment of the 344 consensus flanks confirmed by multiple reads revealed no statistically significant difference between the left and right flanks. When all flanking sequences were combined, a significant bias was seen in the four bases directly flanking the TA target site (Table 1). The resulting consensus sequence is consistent with previously reported results but is now based on the largest set of random germ-line insertions analyzed so far. The symmetry of the Tc1 insertion consensus sequence is a reflection of the orientation independence of Tc1 insertion (36). This is not surprising because Tc1 ends have perfect inverted repeats that are sufficient for insertion when transposase is provided in trans (37).

As a result of the high gene density of the *C. elegans* genome (1), most of the sequenced Tc1 insertions will be located in or close to genes. These Tc1 insertions can be used to obtain mutations in these genes (18). Deletions of flanking genomic sequence occur as a side product of Tc1 transposition; excision of Tc1 results in a double strand break in the chromosome, and repair of this break can result in loss of flanking genomic sequence. Consequently, the ability to induce deletions depends on an intact Tc1 element combined with a genetic background that allows germ-line Tc1 transposition. Tc1 elements are structurally invariant (9), so most Tc1 elements should be competent for excision. Also, the strains used in this study show germ-line Tc1 transposition (12, 17, 25). Therefore, it is, in principle, possible to use the set of mapped Tc1 insertions for deletion mutagenesis. Scaling up the sequencing of polymorphic Tc1 insertions could provide Tc1 insertion alleles of all genes in the *C. elegans* genome. Such Tc1 alleles could be used directly to delete any gene of interest. The current limitation lies in the isolation of more strains with a high Tc1 copy number.

The other application of polymorphic Tc1 insertions is gene mapping. The set of Tc1 insertions forms a dense collection of polymorphic sequence tagged sites. Each Tc1 insertion can be visualized by PCR using a primer in Tc1 and a unique primer in the flanking genomic sequence. Extending the work of Williams *et al.* (21, 22), we have demonstrated an efficient method to use Tc1 STSs in fine mapping mutations. The mutation in a Bristol N2 background is crossed with one of the high Tc1 copy number strains and mutant F_2 progeny are analyzed for linkage to any of the Tc1 STSs. To fine-map mutations, we pooled independently segregated homozygous mutant progeny instead of analyzing single animals. Pooling of single animals allows rare, informative crossovers of Tc1 STSs to be readily detectable. Instead of analyzing many single animals, using PCR on a limited number of pools is sufficient. Depending on the mapping resolution required, the complexity of these pools can be varied. To fine map mutations to the resolution of the STS map, in principle, pools of as many as 100 or more animals can be analyzed. Given the density of the Tc1 STS distribution over the genome, a mutation can now be located with a resolution of ≈265 kb to the physical map, a region corresponding to about 10 cosmids, which is small enough to directly attempt to identify the cosmid containing the mutant gene by transgenesis experiments.

We thank Drs. Jonathan Pettitt and Henri van Luenen for critically reading the manuscript. The *Caenorhabditis* Genetics Center provided some of the strains used in this study. This work was supported by a National Institutes of Health-National Center for Research Resources Grant 5 RO1 RR10082-02 and a Human Frontier Science Program grant to R.H.A.P. Work at the Sanger Centre was supported by grants from the Wellcome Trust and the Medical Research Council.

1. Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., *et al* (1994) *Nature (London)* **368,** 32–38.
2. Hodgkin, J., Plasterk, R. H. A. & Waterston, R. H. (1995) *Science* **270,** 410–414.
3. Waterston, R. & Sulston, J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10836–10840.
4. Plasterk, R. H. A. (1996) *Genome Res.* **6,** 169–175.
5. Abad, P., Quiles, C., Tares, S., Piotte, C., Castagnone-Sereno, P., Abadon, M. & Dalmasso, A. (1991) *J. Mol. Evol.* **33,** 251–258.
6. Henikoff, S. (1992) *New Biol.* **4,** 382–388.
7. Robertson, H. M. (1995) *J. Insect Physiol.* **2,** 99–105.
8. Plasterk, R. H. A. (1996) in *Current Topics in Microbiology and Immunology*, eds. Saedler, H. & Gierl, A. (Springer, Heidelberg), pp. 125–143.
9. Emmons, S. W., Yesner, L., Ruan, K. & Katzenberg, D. (1983) *Cell* **32,** 55–65.
10. Liao, L. W., Rosenzweig, B. & Hirsh, D. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3585–3589.
11. Egilmez, N. K., Ebert, R. H. & Shmookler Reis, R. J. (1995) *J. Mol. Evol.* **40,** 372–381.
12. Moerman, D. G. & Waterston, R. H. (1984) *Genetics* **108,** 859–877.
13. Eide, D. & Anderson, P. (1985) *Genetics* **109,** 67–79.
14. Eide, D. & Anderson, P. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1756–1760.
15. Moerman, D. G., Benian, G. M. & Waterston, R. H. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 2579–2583.
16. Collins, J., Saari, B. & Anderson, P. (1987) *Nature (London)* **328,** 726–728.
17. Babity, J. M., Starr, T. V. B. & Rose, A. M. (1990) *Mol. Gen. Genet.* **222,** 65–70.
18. Zwaal, R. R., Broeks, A., van Meurs, J., Groenen, J. T. M. & Plasterk, R. H. A. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7431–7435.
19. Ruvkun, G., Ambros, V., Coulson, A., Waterston, R., Sulston, J. & Horvitz, H. R. (1989) *Genetics* **121,** 501–516.
20. Olson, M., Hood, L., Cantor, C. & Botstein, B. (1989) *Science* **245,** 1434–1435.
21. Williams, B. D., Schrank, B., Huynh, C., Shownkeen, R. & Waterston, R. H. (1992) *Genetics* **131,** 609–624.
22. Williams, B. D. (1995) in *Caenorhabditis elegans: Modern Biological Analysis of an Organism*, eds. Epstein, H. F. & Shakes, D. C. (Academic, San Diego), pp. 81–96.
23. Ebert, R. H., Cherkasova, V. A., Dennis, R. A., Wu, J. H., Ruggles, S., Perrin, T. E. & Shmookler Reis, R. J. (1993) *Genetics* **135,** 1003–1010.
24. Sulston, J. & Hodgkin, J. (1988) in *The Nematode Caenorhabditis elegans*, ed. Wood, W. B. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 587–606.
25. Moerman, D. G. & Waterston, R. H. (1988) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. for Microbiol., Washington, DC), pp. 537–556.
26. Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J. C. & Markham, A. F. (1990) *Nucleic Acids Res.* **18,** 2887–2890.
27. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
28. Gleeson, T. & Hillier, L. (1991) *Nucleic Acids Res.* **19,** 6481–6483.
29. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
30. Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992) *Comput. Appl. Biosci.* **8,** 189–191.
31. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–407.
32. Armitage, P. & Berry, G. (1987) *Statistical Methods in Medical Research* (Blackwell, Oxford).
33. Emmons, S. W. & Yesner, L. (1984) *Cell* **36,** 599–605.
34. Emmons, S. W., Roberts, S. & Ruan, K. (1986) *Mol. Gen. Genet.* **202,** 410–415.
35. Mori, I., Benian, G. B., Moerman, D. G. & Waterston, R. H. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 861–864.
36. van Luenen, H. G. A. M. & Plasterk, R. H. A. (1994) *Nucleic Acids Res.* **22,** 262–269.
37. Vos, J. C., De Baere, I. & Plasterk, R. H. A. (1996) *Genes Dev.* **10,** 755–761.