Systematic sequencing of cDNA clones using the transposon Tn5

Yuriy Shevchenko[1], Gerard G. Bouffard[1,2], Yaron S. N. Butterfield[3], Robert W. Blakesley[1,2], James L. Hartley[4], Alice C. Young[1], Marco A. Marra[3], Steven J. M. Jones[3], Jeffrey W. Touchman[1,2] and Eric D. Green[1,2,*]

[1]NIH Intramural Sequencing Center, National Institutes of Health, Gaithersburg, MD 20877, USA, [2]Genome Technology Branch, National Human Genome Research Institute, Bethesda, MD 20892, USA, [3]Genome Sciences Centre, BC Cancer Research Centre, Vancouver, BC V5Z 4E6, Canada and [4]Invitrogen Corporation, Rockville, MD 20850, USA

Received January 29, 2002; Revised and Accepted April 8, 2002

ABSTRACT

In parallel with the production of genomic sequence data, attention is being focused on the generation of comprehensive cDNA-sequence resources. Such efforts are increasingly emphasizing the production of high-accuracy sequence corresponding to the entire insert of cDNA clones, especially those presumed to reflect the full-length mRNA. The complete sequencing of cDNA clones on a large scale presents unique challenges because of the generally small, yet heterogeneous, sizes of the cloned inserts. We have developed a strategy for high-throughput sequencing of cDNA clones using the transposon Tn5. This approach has been tailored for implementation within an existing large-scale 'shotgun-style' sequencing program, although it could be readily adapted for use in virtually any sequencing environment. In addition, we have developed a modified version of our strategy that can be applied to cDNA clones with large cloning vectors, thereby overcoming a potential limitation of transposon-based approaches. Here we describe the details of our cDNA-sequencing pipeline, including a summary of the experience in sequencing more than 4200 cDNA clones to produce more than 8 million base pairs of high-accuracy cDNA sequence. These data provide both convincing evidence that the insertion of Tn5 into cDNA clones is sufficiently random for its effective use in large-scale cDNA sequencing as well as interesting insight about the sequence context preferred for insertion by Tn5.

INTRODUCTION

The technologies and strategies for large-scale DNA sequencing have matured dramatically over the past decade (1), resulting in the generation of impressive amounts of sequence data. Most striking has been the production of large collections of single-pass cDNA sequences [expressed-sequence tags (ESTs)] (2–5) and near-complete sequences of eukaryotic genomes, such as that of the nematode (6), *Drosophila* (7,8) and human (9,10). Similarly, efforts to sequence the mouse, rat and pufferfish genomes are reaching a mature stage (1). These data sets provide the substrate for rigorous annotation of genomic sequence, including the identification of all encoded genes. While available ESTs are useful for such efforts, they are imperfect and incomplete. Among the best tools for the accurate annotation of genes are sequences corresponding to full-length mRNAs. As a result, attention is increasingly being paid to the construction and sequencing of full-length cDNA clones. This has resulted in the launching of major cDNA-sequencing initiatives, such as the effort at the RIKEN Genomic Sciences Center (11) and the Mammalian Gene Collection (MGC) program (12) (see mgc.nci.nih.gov).

Full-insert, high-accuracy sequencing of cDNA clones presents unique technical and logistical challenges. This mostly relates to the size characteristics of cDNA inserts, which are both relatively small (typically averaging 1.5–2.0 kb) and heterogeneous in nature (often ranging from <500 bp to >6 kb). The simple application of a random shotgun-sequencing strategy (13), such as that used for sequencing large-insert clones (e.g. bacterial artificial chromosomes) in genomic sequencing projects (9), is relatively inefficient when applied to cDNA clones. There are many reasons for this, one being the effort and expense that would be required to construct a shotgun library for each cDNA clone.

A number of different approaches have been implemented for full-insert sequencing of cDNA clones. A relatively common one involves primer walking, whereby synthetic oligonucleotides are successively designed along a cDNA template following each round of sequence acquisition. This method has been applied to the large-scale, full-insert sequencing of cDNA clones (14); however, it is also associated with several limitations. First, the extensive use of synthetic oligonucleotides adds considerably to the overall costs. Secondly, the sequencing of larger cDNA clones is associated

[*]To whom correspondence should be addressed at: NIH, 50 South Drive, Building 50, Room 5523, Bethesda, MD 20892, USA. Tel: +1 301 402 0201; Fax: +1 301 402 4735; Email: egreen@nhgri.nih.gov
Present address:
James L. Hartley, Protein Expression Laboratory, National Cancer Institute, Frederick, MD 21702, USA

with many iterative walking steps, in some cases requiring a protracted effort. Thirdly, there are logistical demands of ensuring correct primer–template associations, especially when applied on a large scale.

Another approach, concatenated cDNA sequencing (15), represents an adaptation of conventional shotgun sequencing. In this method, multiple cDNA inserts are isolated, pooled and enzymatically concatenated. The entire population of concatenated cDNAs is then subjected to shotgun sequencing (as if it was a single large-insert genomic clone), with the individual cDNA sequences then derived by computer analysis following sequence assembly. This approach is well suited for existing high-throughput sequencing environments. However, it is also associated with a number of technical challenges, including the initial construction of the cDNA concatemers and shotgun libraries, the computational de-convolution of the cDNA sequences, and problems associated with the uneven molar representation of individual cDNAs.

Another option for full-insert sequencing of cDNA clones involves the use of transposons. Transposon-based strategies have been developed for various sequencing applications (16–20). In this case, the transposon simply serves to introduce sites in the cDNA clone for the subsequent annealing of a sequencing primer(s). Typically, each cDNA clone is subjected to a transposon reaction to produce a population of subclones, each harboring a transposon at a distinct location. The subclones are then sequenced using transposon-specific primers, most often from each end of the inserted transposon. Sequencing methods utilizing *in vivo* transposon systems, such as those employing the γδ transposon, have been described (16). More recently, technically simpler *in vitro* transposon systems have become available, including those using the transposons Tn5 and Mu.

As participants in the MGC program, we sought to develop a pipeline for the cost-effective, full-insert sequencing of candidate full-length cDNA clones. The goal was to design a scalable approach for integration with an existing sequencing operation that was mostly geared towards the shotgun sequencing of genomic clones. It also needed to readily produce highly accurate cDNA sequence (less than one error per 50 000 bp, as required by the MGC program). Here we describe our strategy, which involves the use of an *in vitro* transposon system, specifically one based on the transposon Tn5. Our extensive utilization of this strategy in sequencing more than 4200 cDNA clones has provided valuable insight about the transposon Tn5, both with respect to its utility for cDNA sequencing and some of its inherent biological properties.

MATERIALS AND METHODS

Transposon insertion

Arrayed cDNA clones [most of which were constructed with the vector pOTB7 (see http://image.llnl.gov/image/html/vectors.shtml#pOTB7)] were inoculated into individual wells of 96-square-well titer plates (Beckman Instruments) containing 0.8 ml of TB medium and 25 µg/ml of chloramphenicol, covered with AirPore™ tape sheets (Qiagen), and incubated at 37°C overnight with vigorous shaking (~325 r.p.m.). Plasmid DNA was isolated using a 96-well alkaline lysis-based method (see http://genome.wustl.edu/gsc/Protocols/pucprep.shtml). Purified plasmid DNA was resuspended in 100 µl of TE buffer. The transposon reactions, consisting of 5 µl total volume with 15 fmol of transposon DNA and 100 ng of purified plasmid DNA, were performed in a 96-well format using the EZ::TN™ <KAN-2> Insertion Kit (Epicentre). Completed transposon reactions were diluted to a total volume of 200 µl with TE buffer, then 2 µl were used to transform 15 µl of MultiShot TOP10 chemically competent cells (Invitrogen) in a 96-well format. Each transformation reaction (from a single well of a 96-well plate) was then plated onto an individual LB agar plate containing both 25 µg/ml of chloramphenicol and 50 µg/ml of kanamycin. Note that the chloramphenicol helps to minimize the recovery of subclones with a transposon inserted into the vector backbone by selecting against subclones harboring a transposon within the chloramphenicol-resistance gene. The resulting colonies were picked using a QPix robot (Genetix) into 96-square-well titer plates, with each well containing 0.8 ml of 2× YT medium plus 50 µg/ml of kanamycin. Following incubation at 37°C overnight with vigorous shaking (~325 r.p.m.), plasmid DNA was purified using the Concert96™ Plasmid Purification System (Invitrogen) and resuspended in 40 µl of TE buffer.

DNA sequencing

DNA-sequencing reactions were performed in 384-well plates (Marsh), with each reaction (10 µl) containing 2 µl (~200 ng) of purified plasmid DNA, 2 µl of BigDye terminator premix (Applied Biosystems), 1 µl of buffer [400 mM Tris–HCl (pH 9.0), 10 mM $MgCl_2$], 1 µl (3.2 pmol) of sequencing primer and 4 µl of deionized water. Sequence reads were derived from the starting cDNA clones using universal M13 forward (5′-TGTAAAACGACGGCCAGT-3′) and reverse (5′-CAGGAAACAGCTATGACC-3′) primers as well as oligo$(dT)_{23}$V (V = A + C + G) primer. Sequence reads were derived from the transposon-containing subclones using primers specific for each end of the inserted transposon (5′-ACCTACAACAAAGCTCTCATCAACC-3′ and 5′-GCAATGTAACATCAGAGATTTTGAG-3′). The sequencing reactions were subjected to thermal cycling as follows: 96°C for 1 min followed by 35 cycles of 96°C for 10 s, 55°C [or 50°C in the case of the oligo$(dT)_{23}$V primer] for 10 s, and 60°C for 4 min. Each sample was then precipitated with 15 µl of isopropanol, washed once with 25 µl of 80% ethanol, air dried for 15 min, resuspended in 20 µl of water, and analyzed using a model 3700 sequencing instrument (Applied Biosystems). During the sequence-finishing phase (required for a minority of cDNA clones), sequence reads were generated using custom primers and either dGTP terminator chemistry or a 4:1 mixture of standard BigDye terminator and dGTP terminator chemistries, with the precipitated samples then resuspended in 20 µl of Hi-Di Formamide (Applied Biosystems) and analyzed on a model 3100 sequencing instrument (Applied Biosystems).

Data processing

Sequence reads were analyzed using Phred (21,22) and assembled using Phrap (http://www.phrap.org), with the assemblies then viewed and edited using Consed (23). The sequence assemblies consisted of both the end reads derived from the starting cDNA clones and the transposon-derived reads. A cDNA sequence was considered finished when there was a single contiguous assembly that corresponded to the entire cDNA insert with an estimated error frequency of less than one in 50 000 bp (as computed by Phrap).

Analysis of transposon-insertion sites

The final sequences of 1955 cDNA clones were analyzed to establish the orientation of each transposon-derived sequence read. Depending on the orientation of the transposon read, that of the contig, and the locations where the read starts and stops, it was possible to determine the point where the read was generated from the transposon and, hence, the site where the transposon inserted (24).

To assess the randomness of insertions of the transposon Tn5 into cDNA clones, a binomial test was utilized. Specifically, each transposon-insertion event in each of the 1955 analyzed cDNA clones was assigned to a bin based on its location. For each clone, the number of bins corresponded to the square root of the total number of insertion sites (25,26). The length of each bin varied, equaling the cDNA sequence length divided by the number of bins. The number of transposon-insertion events falling within each bin was then tabulated for each cDNA sequence. A binomial test was then performed for each bin, yielding the probability of an insertion event occurring *q* times in *n* attempts, where the probability of it occurring in a single attempt being *P* (*P* = 1/number of bins, *q* = number of insertions in that bin, *n* = total number of insertions in the cDNA sequence). The results for each bin (7135 bins total) were grouped into ranges of 0.01.

Gateway-mediated transfer of cDNA inserts following transposon insertion

For cDNA clones with large vector backbones (e.g. the ~4.4-kb pCMV-SPORT6.0), transposon insertion was first performed and then the cDNA inserts were transferred to a new vector using the Gateway™ cloning technology (27,28) (Invitrogen). Specifically, immediately after the transposon reaction (performed as above), each sample was diluted with water to a total volume of 40 µl and precipitated by the addition of 4 µl of 3 M sodium acetate (pH 5.5) and 100 µl of ethanol. Following incubation at 4°C for 1 h, the precipitated DNA was collected by centrifugation at 3000 *g* for 45 min, washed twice with 80% ethanol, air dried, and resuspended in 5 µl of TE buffer. BP Clonase enzyme mix and buffer (Invitrogen) and 150 ng of *Eco*RI-linearized pDONR223 vector (Invitrogen) were added to each sample according to the manufacturer's instructions (except the final volume was 10 µl instead of 20 µl). Following incubation at 25°C for 20 h, the samples were treated with proteinase K (as per the manufacturer's instructions), and 2 µl were immediately used to transform 20 µl of MAX Efficiency DH10B™ competent cells (Invitrogen), which were then plated on LB agar medium containing 100 µg/ml of spectinomycin and 50 µg/ml of kanamycin. Sequencing reactions performed with the pDONR223-containing subclones used the following two Tn5-specific primers: 5′-ACCTACAACAAAGCTCTCATCAACC-3′ and 5′-GATTTTGAGACACAATTCATCG-3′.

RESULTS

Pipeline for transposon-based cDNA sequencing

The pipeline we developed for the high-throughput, full-insert sequencing of cDNA clones is outlined in Figure 1. The process begins with the generation of sequence reads from both ends of each cDNA insert using a universal forward, a universal reverse, and an oligo(dT)$_{23}$V primer. These end reads both serve as unique identifiers for each cDNA clone and participate in the subsequent assembly. Since one of the universal reads must go through a stretch of adenines contained in the poly(A) tail (in some cases more than 100 As), the resulting read can often be compromised. Thus, the sequence read generated with the oligo(dT)$_{23}$V primer often provides valuable sequence data immediately proximal to the poly(A) tail. All end reads are, in turn, subjected to an initial sequence assembly. Each nascent assembly is then evaluated to assess the number of transposon-containing subclones (if any) that should be used for sequencing that cDNA clone.

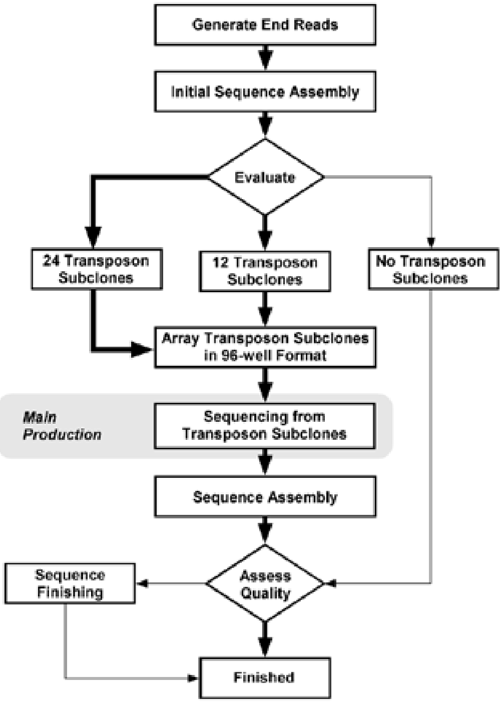

Figure 1. Pipeline for transposon-based sequencing of cDNA clones. The general pipeline for the systematic sequencing of cDNA clones using the transposon Tn5 is depicted, with additional details provided in the text. Note that 'transposon subclones' correspond to subclones derived from the starting cDNA clone by the insertion of a transposon. For some steps, the thickness of the arrows is intended to reflect the relative number of cDNA clones traversing that portion of the pipeline (see Table 1). The gray box designates the portion of the pipeline that can be readily performed as part of the 'main production' component of a DNA-sequencing facility.

For assemblies where there is no evidence of an overlap among the reads from each end of the clone, 24 transposon-containing subclones are then used to generate 48 additional sequence reads (bidirectionally from each transposon insertion). For assemblies where there is evidence of an overlap among the end reads, 12 transposon-containing subclones are then used to generate 24 additional sequence reads. Note that in a small minority of cases (see below and Table 1), the initial end reads yield a single contiguous assembly of notably high quality; in these cases, no transposon-containing subclones are used and that cDNA sequence is then shunted to the quality-assessment step of the pipeline (see Fig. 1).

Table 1. Experience in sequencing 1186 cDNA clones using a transposon Tn5-based strategy

No. of clones (% of total)	Transposon reads	Finishing reads	Average size (bp)
13 (1.1)	No	No	678
25 (2.1)	No	Yes	926
1026 (86.5)	Yes	No	1867
122 (10.3)	Yes	Yes	1988

The data summarized in this table reflect a representative subset of 1186 cDNA clones sequenced to date. Specifically, these clones were sequenced after the transposon-based pipeline described here (see Fig. 1) was fully optimized. In addition, cDNA clones were included in this table only if they originated from a 96-well plate for which the sequencing was complete for at least 94 of those clones (so as not to introduce biases due to the incomplete processing of defined groups of clones).

The choice of using either 12 or 24 transposon-containing subclones allows the convenient batch processing of samples in a 96-well format. Specifically, the transposon-containing subclones from a given cDNA clone are arrayed into either one or two rows of an 8 × 12 microtiter plate. Thus, each cDNA clone is treated independently but, at the same time, the transposon-containing subclones from multiple cDNA clones can be processed together in groups of 96. Such an organizational step facilitates the early and direct integration of the cDNA-sequencing effort with a high-throughput sequencing pipeline, which invariably is geared towards the processing of samples arrayed in 96- and/or 384-well formats.

Sequence reads are generated from the transposon-containing subclones, in each case deriving sequence bidirectionally from each transposon-insertion site. All available sequence reads produced for each cDNA clone (including the initial end reads and the transposon-derived reads) are then assembled, and the quality of the assembled sequence is computationally assessed. In the great majority of cases, the cDNA sequence is finished at this stage, with an estimated accuracy rate of less than one error in 50 000 bp (see below and Table 1). In the remaining, small minority of cases, additional sequence reads must be generated to achieve the requisite quality level, which almost exclusively involves the generation of directed sequence reads using custom-designed primers. The latter is quite similar to the 'sequence-finishing phase' used in the sequencing of genomic clones (13). In very rare circumstances (almost always corresponding to particularly large cDNA inserts), additional transposon-containing subclones need to be selected and used to generate additional sequence reads prior to completing the sequence of that clone.

Note that in generating the transposon-containing subclones, a double-antibiotic selection scheme is used to minimize the recovery of subclones harboring a transposon in the vector backbone (which subsequently might not contribute cDNA sequence). Specifically, only subclones with both chloramphenicol resistance provided by the vector and kanamycin resistance provided by the transposon are selected. Compared with selection with kanamycin alone, we found in our initial studies that such double-antibiotic selection results in the recovery of 4-fold fewer subclones with a transposon inserted in the vector. Also note that by employing a mutated Tn5 transposase (29), which is ~1000-fold more active than the wild-type enzyme, and by balancing the concentrations of the transposon and purified cDNA-containing plasmid, we are able to routinely generate the requisite number of subclones using small-reaction volumes in 96-well microplates. In addition, we have found that only a very small fraction of the recovered subclones (~3%) contain more than one transposon.

Efficacy of transposon-based cDNA sequencing

The strategy detailed above has been extensively utilized to sequence the complete inserts of candidate full-length cDNA clones for the MGC program (12). Specifically, we have now generated the complete sequence for more than 4200 cDNA clones, which together corresponds to >8 Mb of cDNA sequence; all of this sequence data has been submitted to GenBank (also see http://mgc.nci.nih.gov). Importantly, this extensive data set allows the efficacy of our sequencing strategy to be assessed.

Table 1 provides a representative breakdown of the cDNA clones traversing each of the different paths within the pipeline depicted in Figure 1. For the set of clones analyzed, note that transposon-containing subclones were generated and used in ~97% of cases. In more recent experience and with the acquisition of longer end reads from the starting cDNA clones, we are finding that this figure is dropping slightly. Particularly remarkable is our finding that the sequences of the great majority of cDNA clones (~87% for the clones analyzed in Table 1) are finished with an accuracy of less than one error in 50 000 bp following the generation of the initial set of transposon-derived sequence reads, in these cases requiring no additional sequence-finishing effort. Note that only a small fraction (~10%) of the cDNA clones requiring transposon-derived sequence reads need to be subjected to any sequence finishing. To date, we have not encountered a cDNA clone whose sequence could not be finished by the approach described here. In addition, we have found that such finishing is considerably easier than that required for genomic sequence. Indeed, in our experience to date, this has simply involved the generation of an average of 2.3 custom sequence reads per cDNA clone.

The transposon-based pipeline described here thus yields high-quality, finished sequence for ~90% of cDNA clones after the acquisition of insert-end and transposon-derived sequence reads. For these clones, this is associated with the generation of an average of just over 25 reads per kb of cDNA sequence (based on producing ~4.3 Mb of sequence from more than 2250 cDNA clones). Interestingly, this is roughly the number of sequence reads per kb required for generating high-quality genomic sequence by a shotgun-sequencing strategy (1). While it might be possible to implement a variant of the described pipeline that would require slightly fewer sequence reads on a per-clone basis, it is worth pointing out the remarkably high-quality sequence data generated by the described routine. Specifically, the accuracy of the cDNA sequence we have generated to date (>8 Mb from more than 4200 clones) is estimated to be less than one error in 1 000 000 bp (indeed, >20-fold better than the goal of less than one error in 50 000 bp). Thus, a valuable by-product of the described pipeline is the generation of extremely accurate sequence, which is particularly important for cDNA clones.

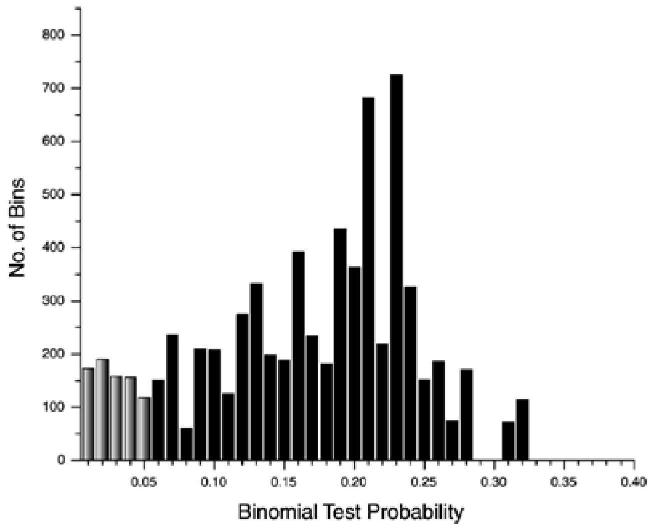

Figure 2. Assessing randomness of transposon Tn5 insertions. The binomial test was used to assess the distribution of transposon-insertion events. The insertions of Tn5 into 1955 cDNA clones were analyzed and assigned to bins (see Materials and Methods for details). The resulting *P*-values reflect the likelihood that the observed insertion events were not random. Plotted are the numbers of bins grouped into *P*-value ranges of 0.01. *P*-values >0.05 correspond to bins for which the observed insertion events are likely to be random. *P*-values ≤0.05 (indicated by gray bars) correspond to bins for which the observed insertion events cannot be confidently described as random occurrences.

Analysis of Tn5 insertions in cDNA clones

The efficient sequencing of cDNA clones by a transposon-based strategy requires insertion of transposons in a relatively random fashion within the cDNA, so that sequence data can be acquired in a generally uniform fashion across the cloned insert. To investigate the randomness of Tn5 insertion into cloned cDNA, we examined 1955 of our finished cDNA sequences (3.59 Mb) to establish the sites of transposon insertion. The sequence length of each cDNA was divided into equal-sized bins, where the number of bins was equal to the square root of the number of insertion sites (25,26). A total of 7135 bins were tested, with the median bin size being 480 bp. Each transposon-insertion event (27 493 total) was assigned to a bin based on its insertion site, and a binomial test was then performed on each bin to assess the randomness of the Tn5 insertion events.

Figure 2 shows the results of this analysis. The depicted plot, which graphs the number of bins as a function of the probability of an insertion event in a particular bin, approximates a Gaussian distribution in the range of *P*-values from 0.06 to 0.40; thus, for this data range, the distribution is nearly random. However, for *P*-values of ≤0.05, the number of bins is unexpectedly high, accounting for 795 bins (or ~11% of the total). These bins contained insertions into 577 cDNA clones. There was not a distinguishable difference between any of the groups of bins with respect to size or GC content. A Chi-square test (26) for the randomness of transposon insertions into the same bins gave similar results. In this test, 89% of the clones yielded a *P*-value of >0.05. Thus, two statistical tests revealed evidence of very slight non-randomness of Tn5 insertion into cDNA clones; however, this was small relative to the average length of the sequence reads, making its impact negligible.

The wealth of data provided by the sequencing of a large number of cDNA clones also provided the opportunity to investigate the preferred sequence context for Tn5 transposition. For this analysis, we again examined all of the Tn5-insertion sites for the same set of 1955 sequenced cDNA clones, in this case cataloging the 20 bases on either side of each insertion site. The resulting data are shown in Figure 3 and Table 2. In both cases, the insertion site corresponds to position 1, while positions 1–9 represent the target sequence that becomes duplicated during Tn5 transposition (30). Striking symmetry with respect to GC content is observed across these nine bases (Fig. 3). This symmetry appears to extend for several bases on either side, but beyond that the base composition of the sequence appears random. Earlier studies involving substantially smaller data sets (31,32) showed a somewhat symmetrical target sequence for Tn5 transposition. For example, Reznikoff and co-workers (32) predicted a consensus sequence for Tn5 insertion of 5′-A^GNT(CT)(AT)(AG)ANC^T-3′, where the caret (^) indicates the boundaries of the 9-bp duplicated target sequence. Examination of our data, which includes 24 493 insertion events, confirms a strong symmetry across the target sequence (Fig. 3). Based on the nucleotide frequencies at each position in the area of symmetry (Table 1), we propose a refined consensus sequence for Tn5 insertion of 5′-G(CT)(CT)(CT)(AT)(AG)(AG)(AG)C-3′. Note that we also calculated the frequency of all 9mers found in the 1955 generated cDNA sequences. From this, it is apparent that there were many other 9mer sites for potential Tn5 transposition than were actually utilized, lending additional support for the above consensus sequence for Tn5 insertion.

Modified transposon-based sequencing strategy

The studies described above in developing, implementing and characterizing a Tn5-based pipeline for full-insert cDNA sequencing were performed using clones constructed with the vector pOTB7. This vector is relatively compact, and thus a typical cDNA insert is roughly the same size as (or even larger than) the vector backbone (~1.8 kb). Such small-vector clones are well suited for transposon-based sequencing since the occurrence of transposon insertions into the vector that fail to contribute any cDNA sequence is negligible (Fig. 4).

cDNA clones constructed with larger vectors (e.g. pCMV-SPORT6.0, whose backbone is ~4.4 kb in size) present an additional challenge to transposon-based sequencing strategies since a larger fraction of the transposon insertions will occur within the vector. As a remedy, one could simply increase the total sequence reads to compensate for a larger fraction of the subclones harboring transposons within the vector backbone. Alternatively, one could analyze each subclone to map the transposon-insertion site (e.g. by PCR), and then only generate sequence reads from the subset of suitable subclones with transposons residing within the cDNA insert. Both of these options are associated with significant additional costs, wasted vector-specific sequence reads in the former case and an expensive up-front mapping effort in the latter case.

We have developed a different, more efficient solution to this problem by a simple modification of our general Tn5-based sequencing strategy (Fig. 4). This modified pipeline utilizes the Gateway cloning technology (27,28) and can be readily applied to any cDNA clone containing suitable Gateway-recombination sequences (e.g. *att*B1 and *att*B2 sites) flanking

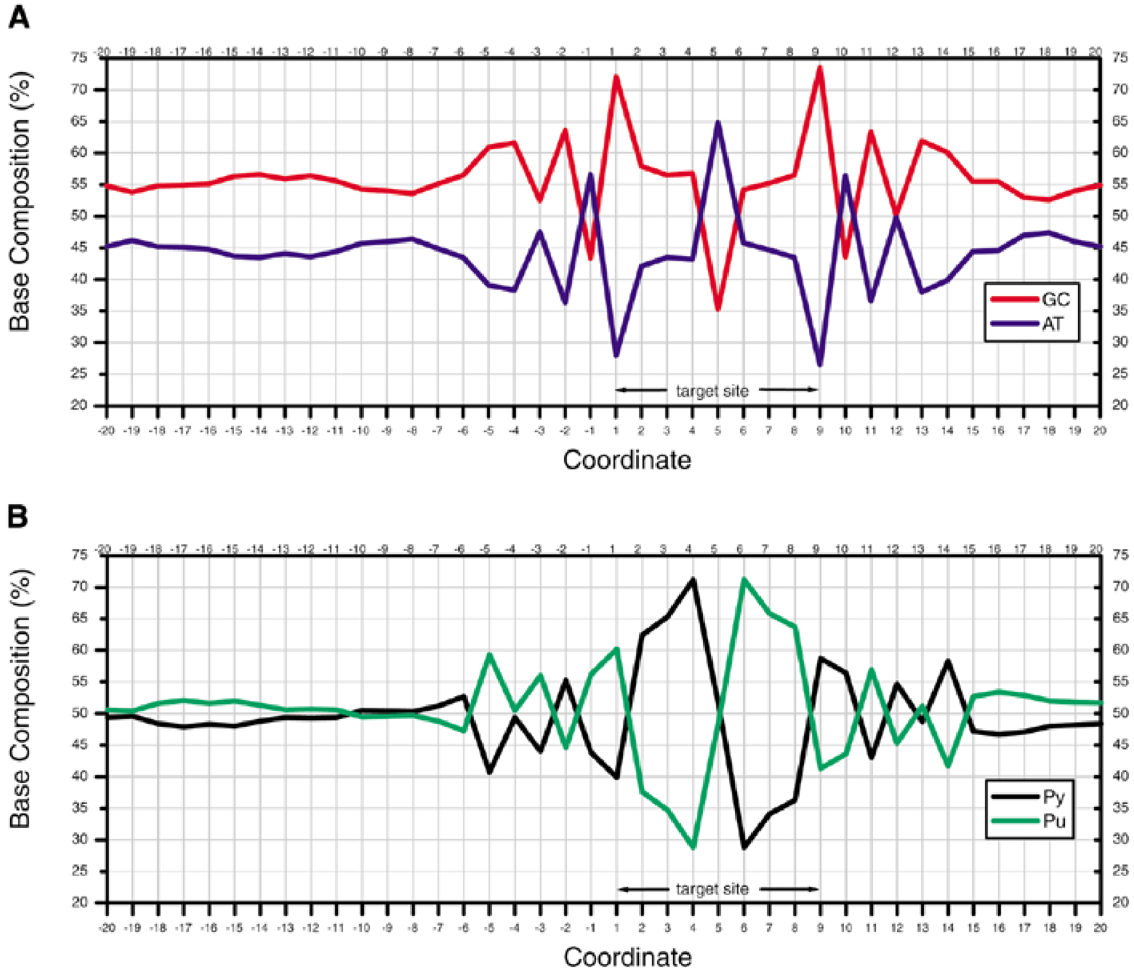

Figure 3. Base composition at the Tn5-insertion site and immediately flanking it. The frequency of each base flanking 24 493 Tn5-insertion events was cataloged (see Table 2). From those data, the relative compositions of GC and AT (**A**) and pyrimidine (Py) and purine (Pu) nucleotides (**B**) were determined and then plotted relative to the 9-bp target site, where position 1 is the 5′ end of the site.

Table 2. Frequency of bases immediately flanking the site of transposon Tn5 insertion

A	18.8	22.5	21.9	27.6	20.2	31.0	16.5	15.6	16.1	13.7	**31.4**	**31.6**	**29.2**	**29.2**	12.7	26.4	17.4	21.6	17.4	17.1	25.3
G	28.5	36.8	28.6	28.4	24.4	25.2	**43.7**	22.0	18.6	15.1	17.3	**39.6**	**36.6**	**34.5**	28.6	17.2	39.5	23.7	33.8	24.6	27.4
C	28.0	24.1	33.0	24.1	39.2	18.2	28.4	**35.9**	**37.9**	**41.7**	18.0	14.6	18.6	22.0	**44.9**	26.4	23.8	26.5	28.1	35.5	28.1
T	24.7	16.6	16.4	19.9	16.1	25.6	11.5	**26.5**	**27.4**	**29.5**	**33.4**	14.2	15.5	14.3	13.8	30.0	19.2	28.2	20.6	22.8	19.1
	–6	–5	–4	–3	–2	–1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Consensus							**G**	**C/T**	**C/T**	**C/T**	**A/T**	**A/G**	**A/G**	**A/G**	**C**						

Depicted here is the relative frequency of each base immediately surrounding Tn5 insertions, based on the analysis of 24 493 insertion events into cDNA clones. Positions 1–9 represent the 9-bp target site that is duplicated upon insertion of Tn5. The numbers in bold indicate the predominant base(s), when present, at each position. Along the bottom is depicted the deduced consensus sequence preferred by Tn5 for insertion.

the cloned insert. Specifically, cDNA-containing plasmids are subjected to a standard transposon reaction (as above), purified, and then incubated with BP Clonase enzyme mix in the presence of a new recipient vector (e.g. pDONR223, which harbors *att*P1 and *att*P2 sites as well as a *spn*^{r} gene). Following bacterial transformation, two antibiotics are used to select for subclones that consist of the transposon-containing cDNA insert shuttled to the recipient vector (Fig. 4). Such a scheme selects against transposons that were inserted into the parent vector backbone and, in principle, should produce an even greater fraction of sequence reads that contribute cDNA sequence.

As a proof of principle, we subjected more than 400 cDNA clones in pCMV-SPORT6.0 (with a 4.4-kb vector backbone) to the Gateway-mediated transfer step following the transposon reaction. The resulting subclones were subjected to the same sequencing pipeline as above (see Fig. 1). The sequencing of

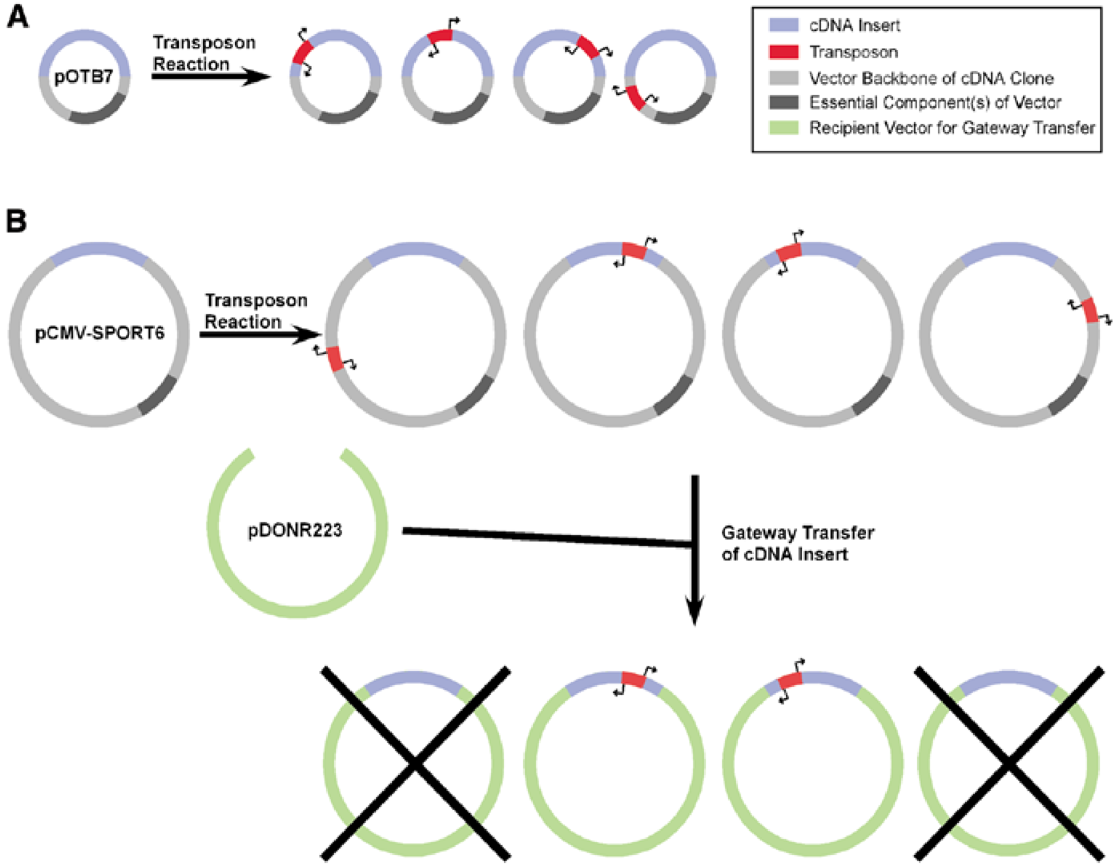
A
pOTB7
Transposon Reaction
cDNA Insert
Transposon
Vector Backbone of cDNA Clone
Essential Component(s) of Vector
Recipient Vector for Gateway Transfer
B
pCMV-SPORT6
Transposon Reaction
pDONR223
Gateway Transfer of cDNA Insert

Figure 4. Modified strategy for transposon-based sequencing of cDNA clones involving Gateway cloning technology. The transposon-based approach for sequencing cDNA clones described here can be implemented in the most straightforward fashion with clones containing relatively small vectors, such as pOTB7 (**A**). In these cases, most of the resulting transposon-containing subclones harbor a transposon within the cDNA insert. While insertions within the vector backbone occur, those inserting within the essential components of the vector (e.g. antibiotic resistance gene, origin of replication) yield non-viable subclones; thus, only a small minority of the recovered subclones harbor a transposon in the vector. For cDNA clones with larger vectors, such as pCMV-SPORT6.0 (**B**), a much larger proportion of transposon-insertion events occur within the vector backbone, with only a small fraction occurring within the essential components of the vector. Undesirable 'background' subclones (i.e. those with an inserted transposon in the vector) can be eliminated by using the Gateway-transfer system (27,28) to shuttle the cDNA inserts into a suitable recipient vector (e.g. pDONR223). By then selecting for the recipient vector backbone and the presence of a transposon, virtually all of the resulting subclones should harbor a transposon within the transferred cDNA insert. Subclones containing a cDNA insert devoid of a transposon would be non-viable (indicated by crosses). Note that the vectors and cDNA inserts are not drawn to scale. At both ends of each inserted transposon are annealing sites for sequencing primers (arrows).

this set of cDNA clones was virtually identical to our extensive experience with pOTB7-derived clones. Based on these data, we are convinced that our modified pipeline is robust and can be used to sequence cDNA clones regardless of the starting vector size, as long as suitable Gateway-recombination signals are present in that vector.

DISCUSSION

In light of the increasing interest in the generation of high-quality and complete cDNA-sequence resources, we developed a robust and scalable transposon-based strategy for the full-insert sequencing of cDNA clones. The specific pipeline described here was tailored to a high-throughput environment that emphasizes the shotgun sequencing of genomic clones and that is associated with a relatively low per-sequence-read cost. As such, the goal was to minimize the effort required upstream and downstream of the main sequence production (see Fig. 1), even at the cost of generating a small proportion of extra sequence reads. The upstream transposon-insertion step is simply used to generate suitable plasmid templates with common primer sequences inserted at random positions within the cloned DNA, which can then traverse the main sequence-production pipeline in an efficient fashion.

While each cDNA clone is handled individually, it is batch processed in one of three groups (i.e. those to be sequenced using 0, 12 or 24 transposon subclones). Our approach has proven highly effective, routinely yielding sequence data with error rates of less than one in 50 000 bp. Indeed, ~90% of cDNA clones are finished to this accuracy level after generation of the transposon-derived sequence reads, requiring no additional custom sequence reads. It is interesting to note that the overall effort (as measured in total reads per kb) for sequencing cDNA clones using the described pipeline is quite similar to that required for sequencing large-insert genomic clones by a shotgun-sequencing strategy (1,13).

It is important to emphasize that the transposon-based strategy described can be readily adapted for use in other sequencing environments, including smaller facilities where the per-sequence-read costs might be considerably higher. In the latter situations, one might consider initially sizing each cDNA insert and then titrating the number of transposon subclones sequenced based on each insert's size. Similarly, a smaller number of random transposon-derived sequence reads might be generated, which would then likely require a larger proportion of custom finishing reads. Regardless of the precise implementation scheme, our general strategy provides a robust and efficient path for sequencing cDNA clones in a variety of environments.

Traditionally, an inherent limitation of transposon-based sequencing has been the insertion of transposons into irrelevant regions of the target DNA, such as the cloning vector. Various solutions to this problem have been devised, such as mapping the insertions and then selecting those subclones harboring transposons at desirable positions within the cloned insert (16). The pipeline described here was initially optimized using cDNA clones containing the small pOTB7 vector and employing double-antibiotic selection. In generalizing the strategy to include cDNA clones containing larger vectors, we sought to avoid the addition of labor-intensive steps, such as those involving mapping of transposon-insertion sites. Fortunately, we were able to integrate a single, simple step in the pipeline to address the problem (see Fig. 4). Specifically, following transposon insertion, the cDNA inserts are subjected to Gateway-mediated transfer (27,28) to a new recipient vector, which yields collections of subclones with transposons exclusively within the cDNA inserts. Of course, this adaptation requires the presence of suitable Gateway-recombination signals in the cDNA clones; fortunately, these are present in most clones being generated and sequenced by the MGC program (12).

Our sequencing of a large set of cDNA clones provided the opportunity to rigorously assess the randomness of Tn5 insertions. Quantitative analysis of the Tn5-insertion sites across 3.59 Mb of finished cDNA sequence revealed nearly random behavior. There was little to no evidence for any hot spots or cold spots or the favoring of certain local GC content. Although some rare inter-insertion intervals were occasionally too far or too close to be deemed truly random, the overall spacing was sufficient to produce adequate sequence coverage in most cases. This generally random behavior is consistent with our finding that the sequence for the bulk of cDNA clones is completed after generation of the transposon-based sequence reads. Note that such randomness is not always encountered with genomic DNA. In fact, we have found a handful of clear cold spots for Tn5 insertion during the finishing of large-insert genomic clones.

In the accompanying paper by Butterfield *et al.* (24), an analogous transposon-based strategy for cDNA sequencing is described. Several notable differences with the current study deserve special mention. First, their approach employs the transposon Mu; interestingly, this transposon was also found to provide sufficiently random insertions to support the routine sequencing of cDNA clones. Secondly, these investigators sequence the cDNA clones in large pools rather than individually. Such an adaptation has the potential to be highly efficient by more accurately fine-tuning the number of sequence reads generated per clone based on insert size. However, this requires each cDNA insert to be sized in advance, an available high-quality end read(s) from each clone for identification purposes, and the need to disambiguate the sequence data emanating from the pooled clones. Nonetheless, their approach has also proven effective for the high-throughput sequencing of cDNA clones.

While the general behavior of the transposon Tn5 is sufficiently random to facilitate the sequencing of cDNA clones, its insertion is associated with certain sequence preferences. Indeed, previous studies have sought to define such consensus sequences for various transposons [e.g. Tn3 (33), Tn5 (30,32), Tn7 (34), Tn9 (30) and $\gamma\delta$ (30)]. The data generated here for the transposon Tn5 and that reported in the accompanying paper for the transposon Mu (24) provide almost unprecedented large data sets for examining the preferred sequence for transposon insertion. Based on examining more than 24 000 Tn5 insertions, a reasonable consensus sequence emerged (Table 2 and Fig. 3). Careful scrutiny of this data suggests the consensus could even be extended further than previously recognized, possibly to an 11- or 13-bp sequence. In addition, the symmetry observed in Figure 3 spans further, suggesting the enzyme–transposon complex recognition of target sequence may span two complete helix turns. Finally, there is a pyrimidine–purine symmetry at this site that seems to strengthen the microfilament model proposed by Goryshin *et al.* (32).

In summary, the studies reported here provide convincing evidence that a transposon Tn5-based strategy can be used for the systematic sequencing of cDNA clones. By deploying this approach for sequencing more than 4200 mammalian cDNA clones, important insight has been gained about the effectiveness of our general paradigm for cDNA sequencing and some of the fundamental biological characteristics of Tn5.

ACKNOWLEDGEMENTS

We thank Keith Wetherby, Russell Pearson, Nicole Dietrich, Stephen Granite and the rest of the NIH Intramural Sequencing Center production staff for superb technical and computational assistance.

REFERENCES

1. Green,E.D. (2001) Strategies for the systematic sequencing of complex genomes. *Nature Rev. Genet.*, **2**, 573–583.
2. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
3. Hillier,L., Lennon,G., Becker,M., Bonaldo,M.F., Chiapelli,B., Chissoe,S., Dietrich,N., DuBuque,T., Favello,A., Gish,W. *et al.* (1996) Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.*, **6**, 807–828.
4. Marra,M., Hillier,L., Kucaba,T., Allen,M., Barstead,R., Beck,C., Blistain,A., Bonaldo,M., Bowers,Y., Bowles,L. *et al.* (1999) An encyclopedia of mouse genes. *Nature Genet.*, **21**, 191–194.
5. Marra,M.A., Hillier,L. and Waterston,R.H. (1998) Expressed sequence tags—ESTablishing bridges between genomes. *Trends Genet.*, **14**, 4–7.
6. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
7. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
8. Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanigan,M.J., Kravitz,S.A., Mobarry,C.M., Reinert,K.H.J., Remington,K.A. *et al.* (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.
9. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

10. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
11. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
12. Strausberg,R.L., Feingold,E.A., Klausner,R.D. and Collins,F.S. (1999) The Mammalian Gene Collection. *Science*, **286**, 455–457.
13. Wilson,R.K. and Mardis,E.R. (1997) Shotgun sequencing. In Birren,B., Green,E.D., Klapholz,S., Myers,R.M. and Roskams,J. (eds), *Genome Analysis: A Laboratory Manual. Analyzing DNA*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, Vol. 1, pp. 397–454.
14. Wiemann,S., Weil,B., Wellenreuther,R., Gassenhuber,J., Glassl,S., Ansorge,W., Bocher,M., Blocker,H., Bauersachs,S., Blum,H. *et al.* (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.*, **11**, 422–435.
15. Yu,W., Andersson,B., Worley,K.C., Muzny,D.M., Ding,Y., Liu,W., Ricafrente,J.Y., Wentland,M.A., Lennon,G. and Gibbs,R.A. (1997) Large-scale concatenation cDNA sequencing. *Genome Res.*, **7**, 353–358.
16. Kimmel,B.E., Palazzolo,M.J., Martin,C.H., Boeke,J.D. and Devine,S.E. (1997) Transposon-mediated DNA sequencing. In Birren,B., Green,E.D., Klapholz,S., Myers,R.M. and Roskams,J. (eds), *Genome Analysis: A Laboratory Manual. Analyzing DNA*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, Vol. 1, pp. 455–532.
17. Davies,C.J. and Hutchison,C.A.,III, (1991) A directed DNA sequencing strategy based upon Tn3 transposon mutagenesis: application to the *ADE1* locus on *Saccharomyces cerevisiae* chromosome I. *Nucleic Acids Res.*, **19**, 5731–5738.
18. Devine,S.E. and Boeke,J.D. (1994) Efficient integration of artificial transposons into plasmid target *in vitro*: a useful tool for DNA mapping, sequencing and genetic analysis. *Nucleic Acids Res.*, **22**, 3765–3772.
19. Devine,S.E., Chissoe,S.L., Eby,Y., Wilson,R.K. and Boeke,J.D. (1997) A transposon-based strategy for sequencing repetitive DNA in eukaryotic genomes. *Genome Res.*, **7**, 551–563.
20. Haapa,S., Suomalainen,S., Eerikainen,S., Airaksinen,M., Paulin,L. and Savilahti,H. (1999) An efficient DNA sequencing strategy based on the bacteriophage mu *in vitro* DNA transposition reaction. *Genome Res.*, **9**, 308–315.
21. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using *Phred*. I. accuracy assessment. *Genome Res.*, **8**, 175–185.
22. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using *Phred*. II. error probabilities. *Genome Res.*, **8**, 186–194.
23. Gordon,D., Abajian,C. and Green,P. (1998) *Consed*: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
24. Butterfield,Y.S.N., Marra,M.A., Chan,S.Y., Guin,R., Kryzwinski,M.I., Lee,S.S., MacDonald,K.W.K., Mathewson,C.A., Olson,T.E., Pandoh,P.K. *et al.* (2002) An efficient strategy for large-scale high-throughput transposon-mediated sequencing of cDNA clones. *Nucleic Acid Res.*, **30**, 2460–2468.
25. Sokal,R.R. and Rohlf,F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research.* Freeman, New York.
26. Chissoe,S.L., Marra,M.A., Hillier,L., Brinkman,R., Wilson,R.K. and Waterston,R.H. (1997) Representation of cloned genomic sequences in two sequencing vectors: correlation of DNA sequence and subclone distribution. *Nucleic Acids Res.*, **25**, 2960–2966.
27. Walhout,A.J., Temple,G.F., Brasch,M.A., Hartley,J.L., Lorson,M.A., van den Heuvel,S. and Vidal,M. (2000) GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.*, **328**, 575–592.
28. Hartley,J.L., Temple,G.F. and Brasch,M.A. (2000) DNA cloning using *in vitro* site-specific recombination. *Genome Res.*, **10**, 1788–1795.
29. Goryshin,I.Y. and Reznikoff,W.S. (1998) Tn5 *in vitro* transposition. *J. Biol. Chem.*, **273**, 7367–7374.
30. Berg,C.M., Berg,D.E. and Groisman,E.A. (1989) Transposable elements and the genetic engineering of bacteria. In Berg,D.E. and Howe,M.M. (eds), *Mobile DNA*. American Society for Microbiology, Washington, DC, pp. 879–925.
31. Berg,D.E., Schmandt,M.A. and Lowe,J.B. (1983) Specificity of transposon Tn5 insertion. *Genetics*, **105**, 813–828.
32. Goryshin,I.Y., Miller,J.A., Kil,Y.V., Lanzov,V.A. and Reznikoff,W.S. (1998) Tn5/IS50 target recognition. *Proc. Natl Acad. Sci. USA*, **95**, 10716–10721.
33. Davies,C.J. and Hutchison,C.A.,III (1995) Insertion site specificity of the transposon Tn3. *Nucleic Acids Res.*, **23**, 507–514.
34. Biery,M.C., Stewart,F.J., Stellwagen,A.E., Raleigh,E.A. and Craig,N.L. (2000) A simple *in vitro* Tn7-based transposition system with low target site selectivity for genome and gene analysis. *Nucleic Acids Res.*, **28**, 1067–1077.