Title
Transposon mediated multiplex sequencing

International Patent Classification(s)
C12Q 1/68 (2006.01)  C12N 15/10 (2006.01)
C12N 15/09 (2006.01)

Application No: 2001273185  Date of Filing: 2001.07.05

WIPO No: WO02/04674

Priority Data
Number 60/216,381  Date 2000.07.07  Country US

Publication Date: 2002.01.21
Publication Journal Date: 2002.04.18
Accepted Journal Date: 2006.05.18

Applicant(s)
Aventis Pharmaceuticals Inc.

Inventor(s)
August, Paul R.; Wiencis, Anna; Draper, Michael; Keagle, Pamela J.; Call, Katherine; Long, Henry

Agent / Attorney
Watermark Patent & Trademark Attorneys, 290 Burwood Road, Hawthorn, VIC, 3122

Related Art
Krysan et al. (1999) Plant Cell. 11: 2263-2290
Title: TRANSPOSON MEDIATED MULTIPLEX SEQUENCING

Abstract: The present invention relates to an automated method of transposon-mediated multiplex sequencing of DNA fragments inserted into a vector. It relates more particularly to an increased efficiency in such automated methods, where the increased efficiency is obtained by screening out before the sequencing those constructs in which the transposon inserted into the vector sequence. This prevents a waste of time and resources in performing reactions sequencing the vector instead of the DNA fragments of interest.
Transposon Mediated Multiplex Sequencing

FIELD OF THE INVENTION

The present invention relates to an automated method of transposon-mediated multiplex sequencing of DNA fragments inserted into a vector. It relates more particularly to an increased efficiency in such automated methods, where the increased efficiency is obtained by screening out before the sequencing those constructs in which the transposon inserted into the vector sequence. This prevents a waste of time and resources in performing reactions sequencing the vector instead of the DNA fragments of interest.

BACKGROUND OF THE INVENTION

The enormous wealth of information that has been acquired from genomic and expressed sequence tag (EST) sequencing in the last 10 years has contributed significantly to efforts to clone full-length cDNA representatives. Although it is anticipated that genomic sequencing projects from human and mouse will be completed in the near future, the transcriptome of these species will remain ambiguous for some time. The complexities involved in predicting, with complete certainty, the splicing program of mRNAs from genomic sequences have compelled additional genomic research focused on obtaining the sequences of full-length cDNAs. In addition, full-length cDNA sequencing efforts are also required for the confirmation of cDNA sequences after methods that involve amplification of the cDNA have been employed for cloning. This scenario is particularly prevalent in genomic centers that are focused on validating gene targets for drug discovery efforts. Clearly, after great expense and effort has been expended, it would be senseless for a putative target to fail the validation process simply because the coding sequence of the target gene was incorrect. Therefore, approaches are required at genomic centers
to sequence large numbers of full-length clones quickly, inexpensively, and accurately. For this purpose, Applicants have created a new, integrated high-throughput process called transposon expedited multiplex sequencing (TEMS).

In the last 20 years, many methods have been developed for sequencing large inserts to plasmids. However, many of these methods were cumbersome and could not be transferred to high-throughput, automated systems. Sequencing by primer walking is slow, expensive, and often fails since primers are designed to the sequence in a poorly characterized region. Similarly, sequencing by creating a collection of clones by exonuclease digestion from the ends of the target clone is slow, clone specific, and extremely sensitive to the purity and integrity of the template DNA. In addition, the success rate using this approach is quite variable. Shotgun sequencing of clones is a higher-throughput method, however it requires isolating the insert from each clone and then recloning smaller fragments generated by a wide variety of methods. Additionally, shotgun libraries that utilize restriction digests result in a cloning bias and subsequently a non-random distribution of DNA sequence data.


Traditionally, this method was cumbersome since it required moving plasmids through different host strains for the cloning and transposon-insertion steps. Recently however, several commercial molecular biology vendors have developed in vitro transposition systems to take advantage of the random insertion of a modified transposon (i.e. Tn5 etc.). Unfortunately, some of these systems result in a high background of false positives, and are difficult to use with methods to screen positive clones by polymerase chain reaction ("PCR"). Applicants have utilized several of these transposon based sequencing methods and have not experienced any of these difficulties with a modified version of the in vitro GPS-1 transposition system from New England Biolabs.
Nevertheless, transposon insertions cannot be directed exclusively to the target DNA of interest and appear in the vector with a high frequency. In an effort to solve this problem and increase the efficiency of transposon facilitated sequencing, Applicants have developed a unique, high-throughput procedure called transposon expedited multiplex sequencing (TEMS).

Accordingly, it is an object of this invention to provide a high-throughput, efficient, and inexpensive process for the sequencing of DNA fragments.

It is a further object of this invention to provide a high-throughput, efficient and inexpensive process for transposon-mediated sequencing of target DNA fragments which minimizes the amount of non-target DNA sequence generated.

It is yet another object of this invention to provide a PCR-based screen to distinguish between the desired constructs with transposons inserted into the target DNA sequence and the undesired constructs with transposons inserted elsewhere.

SUMMARY OF THE INVENTION

The present invention meets the above objects by providing the following method, which may be automated for further convenience. Multiple DNA target sequences, each cloned into a vector, are pooled and selectable transposons with sequencing primers on each end are inserted randomly into the DNA target-containing vectors. Selected, transposon- and DNA target sequence-containing vectors are then individually screened using a PCR reaction to identify those vectors which have transposons located in the DNA target sequence. The PCR reaction uses primers located at each end of the vector sequence and is optimized to provide sufficient extension time for the PCR polymerase to efficiently produce a product the length of the vector, but insufficient extension time for the PCR polymerase to efficiently produce a product the length of the vector plus the transposon. Therefore, significant amounts of PCR product will only be produced for vectors containing the transposon in the DNA target sequence and not the vector. The presence or absence of a significant amount of PCR product for each PCR reaction can be quickly determined by the use of a quantitative fluorescent dye selective for double stranded
DNA ("dsDNA"). Vectors identified by a significant PCR product in the prescreen are then individually sequenced using the sequencing primers on each end of the transposon to read out into each DNA target sequence. The raw DNA target sequences of individual sequencing reactions can be combined to determine the full sequence of each of the DNA targets in the pool.

The parallel processing of clones greatly increases the speed in which reactions can be set up and sequenced. Additionally, the process does not rely on any particular vector and does not require any recloning steps. Most important for overall efficiency, sequencing of the vector backbone is minimized or eliminated with the screening step. Automation of each of the steps of the instant process can be readily achieved using equipment available to those in the art of automated DNA sequencing.

**BRIEF DESCRIPTION OF THE DRAWINGS**

Figure 1 is a diagram of the steps of TEMS.

Figure 2 shows the sequence of GPS-Apra transposon vector.

Figure 3 shows the sequence of GPS-Apra-2µm-URA3 transposon vector.

Figure 4 is a picture of an electrophoresis gel stained to show the products of PCR reactions using various polymerases to multiply the vector portions of the transposon- and target-DNA- containing vectors.

Figure 5 is a picture of an electrophoresis gel stained to show the products of PCR reactions having too great an extension time to eliminate an approximately 9 kb product.

Figure 6 is a picture of an electrophoresis gel stained to show the products of PCR reactions at a variety of extension temperatures and having different numbers of cycles.

Figure 7 is a picture of an electrophoresis gel stained to show the products of PCR reactions and a table of the PICOGREEN® results for the same PCR reactions.

**DEFINITIONS**
As used herein:

A "vector" is any DNA construct which replicates in a cell and has an insertion site at which unknown DNA sequences can be inserted.

"Selectable", in reference to a vector or transposon, means that the vector or transposon carries a phenotype that can be used to identify cells transformed with the vector or transposon. Preferably, the phenotype allows the transformed cells to survive in conditions in which non-transformed cells cannot.

A "target DNA" or "target DNA sequence" is a DNA molecule of known or unknown sequence which the operator of the method desires to sequence.

A "representative number of individual transformants" is a minimum number of individual transformants which ensures that for each target DNA in the pool there will be a sufficient number of transformants containing insertions of the transposon into that target DNA to provide a desired number of independent sequencing determinations of each base in that target DNA.

"Substantial amounts of dsDNA" in regard to PCR products is an amount of dsDNA that exceeds a predetermined threshold. The predetermined threshold can be identified by evaluating the variability of PCR products found within PCR reactions with several constructs which are known to contain the transposon inside or outside the vector sequence.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention provides a method that can be efficiently automated for sequencing multiple target DNAs in parallel. In the method, each target DNA sequence is inserted into a
selectable vector, preferably a plasmid, at identical insertion points in the vector sequence. It is important that all target DNAs are inserted into the same vector or, if different vectors are used, that each vector is substantially the same size and has identical sequences around the insertion point so that one set of primers can be used to amplify all the vectors by PCR reaction.

The vectors containing each of the multiple target DNA sequences are pooled in ratios based on the length of each target DNA sequence, so the pool contains substantially equal amounts of each kb of each target DNA sequence. The ratios for each vector to be included in the pool can be determined by determining the lengths of each target DNA sequence (for instance, by electrophoresis against standards size markers on an agarose gel), and adding to the pool a set amount of each vector for every kb in length of the target DNA sequence in that vector.

A transposon-insertion reaction is then performed on the pool to randomly insert a selectable transposon into the target-DNA containing vectors. Although this step could be performed by traditional, \textit{in vivo}, transposon insertion, that would require moving a plasmid through more than one cell type. \textit{In vitro} transposon-insertion reactions are much simpler and faster. The transposon must carry a selectable marker and must be of sufficient size to enable PCR conditions to be established that efficiently multiply a dsDNA fragment the size of the vector, but do not significantly multiply a dsDNA fragment the size of the vector plus the transposon. Preferably, the transposon is about equal to or greater than the vector in length. Most preferably, the transposon is equal to or greater than $\frac{5}{4}$ the length of the vector. Particularly preferably, the transposon is equal to or greater than $\frac{5}{3}$ the length of the vector. Where the vector is a bacterial plasmid, the transposon is preferably at least about 3 kb, most preferably at least about 4 kb and particularly preferably greater than or equal to about 5 kb.

The pool of vectors from the transposon-insertion reaction is then used to transform cells and the transformants are grown under conditions which select for the transposon. Selected transformants are isolated and grown into cultures. The number of isolated transformants grown into cultures depends on the number and size of target DNA sequences in the pool. Enough transformants must be isolated and grown to ensure that for each target DNA sequence, the isolated transformants will include enough individual insertions of the transposon into that particular target DNA sequence to provide several fold independent sequence determinations of each and every base in that target DNA sequence. To ensure against errors, those of skill in the
art often seek to obtain a 6-fold coverage of sequence data, although they may opt for more or less, as the situation merits. When attempting to obtain 6-fold coverage, as a general rule of thumb, 18 clones will need to be actually sequenced per 1 kb of nucleotides to obtain this coverage. However, to actually sequence 18 clones, a greater number of transformants will have to be isolated and grown, because the transposon will not have integrated into the target DNA in every case, but will have randomly inserted over the whole target DNA-containing vector. Therefore, the number 18 is multiplied by the likelihood that the transposon inserted into the target DNA sequence. So, to obtain 6-fold sequencing coverage, for each target DNA sequence in the pool, a number of transformants to be isolated is equal to: the length of the target DNA sequence in kb times 18, times the ratio (size of vector)/(size of DNA target sequence).

Each culture from a selected transformant then has a PCR reaction performed on its DNA. The PCR reaction uses primers complementary to the 3' ends of the vector sequence at the insertion point, which allow PCR multiplication of a fragment the size of the entire vector sequence. The reaction conditions of the PCR reaction are optimized to efficiently produce a fragment the length of the vector, but not to efficiently produce a fragment the length of the vector plus the transposon. The optimization of PCR conditions for any particular vector and transposon combination according to this invention can be determined by one of skill in the art through routine experimentation, as guided by such reference works as Kimmel, B., M.J. Palazzola, C. Martin, J.D. Boeke, and S.E. Devine, 1997, Optimizing PCR Assays, Methods for Improving PCR, Detecting and Characterizing PCR products, Protocols for Detecting and Characterizing PCR Products. In Genomic Analysis: A laboratory manual (ed. E Green, B. Birren, R. Myers, and P. Hieter), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; and Cha, R.S. and Thilly, W.G. (1995) in: PCR Primer, Dieffenbach, C.W. and Dveksler, G.S. (eds.), CSH Press, New York. The particular DNA polymerase used for PCR should not be limiting for the method, although the choice of polymerase may help to optimize the reaction. Generally, the elongation temperature and/or time of the PCR are varied to achieve the desired exclusion of the vector plus transposon sized product. For a demonstration of optimization, see Example 2.

The amount of product produced in each PCR reaction is then measured, preferably optically. Optical measurement could be made by any method which distinguishes dsDNA from
nucleotides, but is preferably made using a fluorescent dye specific for dsDNA. dsDNA specific
dyes include the bis-benzimide dye Hoechst 33342 and the cyanine dyes of Molecular Probes,
Inc, Eugene, Oregon, including PICOGREEN®. PICOGREEN® is preferred, due to its
sensitivity and selectivity for dsDNA. The measured value of fluorescence is compared to a
predetermined threshold level for determining whether a significant amount of dsDNA PCR
product is present. The threshold level of fluorescence is set experimentally using positive and
negative control PCR templates of the size of the vector alone and the vector plus the transposon,
under the same PCR conditions that will be used to screen the isolated transformants.

Those transformants corresponding to PCR reactions which did not produce substantial
amounts of dsDNA are not sequenced, because they contain the transposon in the vector
sequence. Transformants corresponding to PCR reactions which produce substantial amounts of
dsDNA contain the transposon in the target DNA sequence and are sequenced. Two sequencing
reactions are performed for each transformant, using primers from each end of the transposon to
read the sequence of the target DNA sequence surrounding the transposon. The sequencing
reactions can be performed by any of the known methods of the art.

Once sequence data has been collected from all the transformants which passed the PCR
screen, the sequences must be assembled into all the individual members of the pool of target
DNA sequences. This can be accomplished by methods of computer sequence analysis known in
the art, and the exact manner is not limiting on this invention.

The advantages of the instant method include speed, efficiency and an ability to automate
the process. Particularly, the instant method provides a quick PCR screen that allows for
elimination of those unproductive transformants which have the transposon inserted into the
vector sequence and would yield mainly vector sequences if sequenced. The PCR screen is fast,
can be automated and does not require any isolation of the PCR product, such as running an
agarose gel to determine its size. The use of a pool of target DNAs allows a single transposon
reaction and a single transformation to be performed, instead of requiring one for each target
DNA sequence.

EXAMPLES
Example 1 - Choosing a transposon to fit a high-throughput screening process

Original Transposon from New England Biolabs

The Genome Priming System Kit from New England Biolabs, 32 Tozer Road, Beverly, MA 01915 ("NEB") was used to initially prove that a transposon inserted randomly into DNA targets could be an efficient way to completely sequence the targets with a great degree of accuracy. Supplied with the kit was one of two transposons that contained the kanamycin resistance gene. However, kanamycin is a fairly common resistance marker, and if a target DNA sequence was already cloned into a kanamycin-selectable vector, then either a different transposon or recloning of the target DNA sequence would be required. The kit provided another transposon that contained the chloramphenicol resistance marker, but that, too, is fairly common. Therefore, the transposon should be modified to contain an uncommon resistance marker, such as apramycin, to avoid additional cloning steps and reduce human error in choosing the correct transposon for each DNA target. The GPS-Apra transposon (Figure 2) was constructed by modifying the NEB pGPS1 transposon.

Screening method using the modified transposon, GPS-Apra

Initial experiments showed that a screening process could identify from transformants selected for the transposon those transformants containing the transposon in the target DNA sequence and not the vector. This reduces costs by eliminating unproductive sequencing of the vector sequence. The original NEB pGPS1 transposon was 2.614 kb with ~1.7 kb of that, with priming sites at each end, being randomly inserted into the target DNA. By PCR amplification of the 3.0 kb vector using complement M13F and complement M13R primers, two distinct PCR products could be visualized on a slab gel. A 3.0 kb PCR product indicates that the transposon must be in the target DNA sequence, since the vector is 3.0 kb. A 4.7 kb PCR product indicates a vector (3.0 kb) containing the transposon (1.7 kb). Therefore, clones containing the transposon
in the target DNA sequence could be identified by the presence of 3.0 kb PCR products on the gel.

To demonstrate identification of clones containing the transposon in the insert, colonies were selected from the transformation following the transposon reaction and inoculated into media containing both antibiotics for selection, apramycin and the plasmid antibiotic. The culture was grown up overnight and PCR was done the following day by kerplunking approximately 1 ul of culture into a PCR cocktail in a 96 well plate. The PCR cocktail included the Amplitaq enzyme and reagents from PE Corporation, PE Corporation, 761 Main Avenue, Norwalk, CT 06859. The conditions for PCR were standard {95 C for 5 min, (95 C for 30 sec, 54 C for 1 min, 72 C for 6 min, 30 cycles), 72 C 1 min, 4 C forever}. After completion of the cycle (approximately 2.5 hr), gel electrophoresis was done using 8 ul of the 50 ul PCR reaction on a standard size gel. There was a clearly detectable size distinction between the PCR products from vector sequence without the transposon and vector sequence including the transposon.

To maintain a high-throughput for the sequencing reactions, however, electrophoresis would have to occur on large gels that could hold hundreds of samples at one time. On these large gels, the difference between a 3 kb and a 4.7 kb fragment would be difficult to reproducibly detect with an automated system. Also, gel electrophoresis was very time consuming and would interfere with performing the screen in a high-throughput manner.

In a search for other, high-throughput methods, fluorescent detection of dsDNA was settled on as reproducible and less time consuming. The PICOGREEN® assay seemed particularly convenient. However, fluorescent detection of dsDNA cannot distinguish between different size fragments, so the PCR reaction had to be altered so that the presence or absence of any PCR product indicates the position of the transposon in the target DNA containing vector. If only clones with the transposon in the target DNA sequence produced a dsDNA PCR product, then the PICOGREEN® dsDNA dye would fluoresce at a high level for only the desired clones. The wells that contain the transposon in the vector would not amplify a significant dsDNA product and would result in a low level of fluorescence. A cutoff level of fluorescence between desired and undesired clones can be determined by the use of known samples.

A variety of PCR conditions were tested by lowering the elongation temperature, but the 4.7 kb band would not consistently drop out, without having an increase of false positives. It was
decided that if the transposon could be modified again to greatly increase its size, then PCR conditions could be developed that would consistently drop out the upper, transposon-containing band, thereby lowering the false positive rate. A large "stuffer" gene that would not effect the transposon reaction in a negative way was sought. A yeast 2um plasmid with the URA3 gene met the size criteria and could also be used in yeast experiments in the future. Therefore, a GPS-Apra-2um-URA3 transposon plasmid (Figure 3) was constructed having a total size of 6.1 kb, including the transposon of ~5.0 kb. The PCR conditions were altered successfully to drop out the clones with the transposon in the vector and the transposition reaction was not affected. See Example 2.

Example 2 - Optimization of PCR conditions to reduce false negative and false positive results

BASIC PCR CONCEPT

This screen is based on PCR cycling conditions that have been specifically optimized to amplify a 3.9 kbp fragment of the entire vector DNA. In this case, the vector was PCR4Blunt-TOPO from Invitrogen, 1600 Faraday Ave, Carlsbad, CA 92008. If the vector has the 5 kbp transposon inserted into it, its size will be increased to 8.9 kbp, which is too large to be amplified under the optimized PCR conditions.

The primer sequences selected to amplify the vector portion of the clones are the complement sequences of the universal primers M13F(-20) and M13R. The sequences are as follows: 5'-ACTGGCCGTCGTTTTAC-3' and 5'-CATGGTCATAGCTGTT-3'. This amplification is illustrated in Figure 1.

BACKGROUND ON THE DEVELOPMENT OF THE 384-WELL FORMAT

In order to screen transposon reactions from pools containing multiple clones it was necessary to develop a screen that was both quick and cost efficient. Initial development attempts were based on the 96-well protocol. To change to a 384-well plate, all reagent volumes were halved to accommodate the smaller well volume. However, these conditions proved to be
less than ideal due to the high number of PCR failures that were being observed. A search for PCR reagents and conditions that were more robust for this format was undertaken. The results of experiments that tested a wide range of enzymes and thermocycling conditions will be described in the following sections.

384-WELL PCR OPTIMIZATION PROCESS

Polymerase Selection

A quick screening of a variety of PCR kits from several vendors was done to determine a 384-well replacement for the PE Corporation AmpliTaq polymerase that had been used in the 96-well format. The following PCR kits were tested: Clontech’s Advantage cDNA PCR kit, Epicentre’s Failsafe PCR system, and three kits from Takara; LA Taq, Ex Taq, and Z Taq. (Clontech Laboratories Inc. is located at 1020 East Meadow Circle, Palo Alto, CA 94303-4230; Epicentre Technologies is located at 1402 Emil Street, Madison, WI 53713; and Takara Shuzo Co., LTD has a business address of Biomedical Group Seta 3-4-1, Otsu, Shiga, 520-2193, Japan.) Samples tested represented pooled DNA targets. They were cloned into pBluescript SK- (3.0 kb) from Stratagene, 11011 North Torrey Pines Road, La Jolla, CA 92037. The growth time was 20 hr in 384 well format. Based on preliminary data the Z-taq system was chosen as the enzyme to pursue due to properties it possessed which made it ideal for use in a high-throughput system like TEMS. See Figure 4, which shows a stained electrophoresis gel on which the PCR products of the samples have been run.

Reagent Optimization

It was also desirable to decrease the cost per reaction of the PCR screen. Table 2, shown below, summarizes the various experiments that were done to test the amount by which each of the various reagents could be lowered without compromising the results.
Variable Tested | Condition Tested | Optimal Result
--- | --- | ---
Z- Z-Taq polymerase | 0.625, 0.313, 0.156 Units | 0.313 Units
10X Z-Taq Buffer | 1x, 0.5x | 1x
Primers | 2.5, 5, 10 pmol | 5 pmol
Primers | cM13f & cM13r | cM13f & cM13r
cT3 & cT7
dNTP's | 100,150, 200 μM | 150 μM

Table 2. Table summarizing the various PCR cocktail ingredients that were tested during the optimization process.

Based on the above experiments the following optimized reaction cocktail mix (per sample) was determined:

ddH2O: 17.875 μL
d10x Z-Taq buffer: 2.5 μL
dNTP mixture: 1.5 μL
template DNA: 1 μL (deposited using a Kerplunker #96/384, manufactured by Nalge Nunc International, 75 Panorama Creek Drive, Rochester, NY 14625, and Pin Replicator #250520/250393)
Z-Taq DNA polymerase: 0.125 μL (0.3125 Units)
cM13f (5 pmol/μL): 1 μL
cM13r (5 pmol/μL): 1 μL

Optimization of Thermocycler Conditions

Initial thermocycling conditions used were based on recommendations found in the Takara literature. These conditions were:

1. denature: 98 °C for 5 seconds
2. anneal: 55 °C for 10 seconds
3. extension: 72 °C for 60 seconds (15.4 seconds/kb)
4. go to 1, 30 times
5. 4 °C until the PCR reaction is evaluated.

The results of using these conditions are shown in Figure 2. From that data, it was decided that the annealing temperature should be lowered so as to try and decrease the reaction specificity and thus increase the number of positives that are amplified. In the next experiment the annealing temperature was lowered to 52 °C. Although this helped with the overall reaction results, it was discovered that using an extension time of 15.4 seconds per kbp resulted in the amplification of vector containing the transposon insertion. See Figure 5, which shows a stained electrophoresis gel on which PCR products from 47 transformants selected after transposon insertion in target DNA-containing pBluescript SK- and grown for 20 hours in wells of a 384 well plate were run. The appearance of a 9 kbp band, believed to be the amplified product of vector containing the transposon, is the result of an excessive extension time.

Additionally, an annealing temperature gradient between 48 °C and 68 °C was tested to confirm that 52 °C was indeed the optimal temperature to use for this PCR. In the same experiment the number of cycles was also tested. 30, 35, 40, and 45 cycles were tested. From the results of that experiment it was determined that the number of cycles need to remain low to avoid background noise. From the 48 °C and 68 °C gradient the 52 °C annealing temperature was reconfirmed as being optimal. The temperature is not too low that it becomes non-specific and not too high that it risks approaching the 56 °C temperature at which the PCR begins to fail. These results are shown in Figure 6, a picture of a stained electrophoresis gel, where each depicted product represents amplification of the control pBluescript SK- vector at 2.5 ng per well, in a 384 well format with 25 ul PCR cocktail volume. Based on the above data, the thermocycling conditions were determined to perform optimally using the following cycle.

1. denature: 98 °C for 5 seconds
2. anneal: 52 °C for 10 seconds
3. extension: 72 °C for 30 seconds (7.7 seconds/kb)
4. go to 1, 35 times
5. 4 °C until evaluation of the PCR reaction.

Since evaporation was found to be a problem in the 384-well plates, rubber 'P' seals from MJ Research, 590 Lincoln Street, Waltham, MA 02451, are placed over sealed plates before thermocycling begins. Along with this, and as recommended by the manufacturer, the lid temperature is set to 85 °C.

Table 3, shown below, provides a summary of the various PCR thermocycling conditions tested during the optimization process.

<table>
<thead>
<tr>
<th>Variable Tested</th>
<th>Condition Tested</th>
<th>Optimal Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annealing Temperature</td>
<td>38° C – 68° C</td>
<td>52° C</td>
</tr>
<tr>
<td>PCR cycle</td>
<td>30, 35, 40, 45</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 3. Summary of the PCR thermocycling conditions that were tested during the optimization.

SCREENING FOR POSITIVE CLONES USING THE dsDNA QUANTITATION REAGENT PICOGREEN®

To bypass the labor-intensive task of gel electrophoresis, the TEMS PCR screen utilizes the dsDNA quantitation reagent, PICOGREEN®. The PICOGREEN® is diluted to a working dilution of 1:150 in 1x TE buffer. This dilution was found to be optimal, although it is slightly more concentrated than the 1:200 dilution recommended in the manufacturer’s protocol.

25 μL per well of 1:150 PICOGREEN® is aliquoted into black 384-well plates using a Robbins Hydra 96 Dispenser from Robbins Scientific, 1250 Elko Drive, Sunnyvale, CA 94089-2213. 5 μL per well of each PCR reaction is then added to the PICOGREEN®, again using the Hydra Dispenser. The plate is then immediately placed into a Molecular Dynamics Gemini Fluorescence Plate Reader and the plate is then mixed for 10 seconds before its fluorescence signal is measured.
For the assay to function as a screen, there must be a clear difference in the fluorescence signal between positive and negative wells. PCR product amounts of 1.3, 2.5, 5, and 10 μL were tested to determine the greatest assay range of positive signal to background. Results showing that 5 μL of PCR product into 25 μL of PICOGREEN® is the optimal condition to use is shown in Table 4.

| ul PCR | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|--------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1.3    | 563 | 287 | 206 | 987 | 180 | 345 | 4734 | 351 | 171 | 312 | 306 | 237 | 1102 | 305 | 202 | 246 | 843 | 134 | 126 | 318 | 223 | 308 | 635 | 223 |
| 2.5    | 639 | 627 | 495 | 1624 | 412 | 751 | 2541 | 694 | 397 | 710 | 711 | 413 | 2129 | 541 | 456 | 549 | 1713 | 313 | 217 | 634 | 467 | 690 | 1434 | 552 |
| 5      | 1490 | 1003 | 835 | 2768 | 635 | 1450 | 3703 | 1540 | 813 | 1220 | 1447 | 802 | 2981 | 1190 | 934 | 1003 | 1250 | 674 | 302 | 1259 | 1112 | 873 | 2341 | 1116 |
| 10     | 5990 | 3224 | 7750 | 2231 | 1354 | 2209 | 2857 | 6221 | 2112 | 1376 | 2095 | 1206 | 1746 | 340 | 1862 | 1681 | 1674 | 2002 | 167 | 7383 | 1837 | 1642 | 1560 | 2567 | 1333 |

Table 4. Fluorescence signal detected using the PICOGREEN® assay with various amounts of PCR product. The 24 PCR products are from samples representing a pool of 11 DNA targets with a sizes up to 4.5 kb cloned into the vector PCR4Blunt-TOPO. The transposon used was pGPS-ApaI-2uM-URA3.

The signal output generated by the fluorescence reader is then exported to a histogram display, and a cutoff value for positive clones is assigned from the histogram. Using the current conditions, the cutoff generally falls between 1000-1500.

A Genesis RSP 150 TECAN robotics station (TECAN U.S. INC., P.O. Box 13953, Research Triangle Park, NC 27709) was used to generate 96-deep-well plates containing cells from the clones identified by the screen as having the transposon in the target DNA sequence. Cells from each positive clone are placed in Superbroth and the appropriate antibiotics (dual selection) in one of the wells. These plates are grown and then the cultures are submitted for automated sequencing.

EXAMPLE OF THE OPTIMIZED PCR SCREEN TECHNIQUE FOR 384-WELL PLATES

Use Q-Bot to pick colonies of the transposon containing clones into Genetix 384-well shallow-well plates containing 65 μL per well of Superbroth plus appropriate.
antibiotic. Place plates into a humidified 37°C incubator for 19-22 hours. Aliquot 25 μL per well of sterile 50% glycerol into the growth plates using a Multidrop 384 115V liquid dispenser (Labsystems Inc., 8 East Forge Parkway, Franklin MA 02038). Using the glycerol plate as the template, set-up PCR using the following recipe:

ddH₂O: 17.875 μL
10x Z-Taq buffer: 2.5 μL
dNTP mixture: 1.5 μL
template DNA*: 1 μL (using the Kerplunker, pin-tool)
Z-Taq DNA polymerase: 0.125 μL (0.3125 Units)
cM13f (5 pmol/μL): 1 μL
cM13r (5 pmol/μL): 1 μL

*To ensure that the Kerplunker is sterile it should be stored in 70% ethanol and flamed before use.

Placed the PCR reaction plate into a thermocycler and start the following cycle:
1. denature: 98 °C for 5 seconds
2. anneal: 52 °C for 10 seconds
3. extension: 72 °C for 30 seconds (7.7 seconds/kb)
4. go to 1, 35 times
5. 4 °C until the PCR reaction is evaluated.
Total cycling time: 1 hour, 18 minutes, 30 seconds.

While the reaction is underway, thaw the PICOGREEN® reagent at room temperature.

Prepare a 1:150 dilution of PICOGREEN® in 1x TE Buffer. Prepare 10 mL per 384-well plate. Use the Robbins Hydra to aliquot 25 μL per well to an entire 384-well black plate. Use the Robbins Hydra to dispense 5 μL per well of the PCR reaction into the PICOGREEN® plate. Immediately place the plate into the fluorescence plate reader and mix by shaking for 10 seconds. Export the fluorescence signal data to a floppy disk. Import data into an Excel macro and generate a rearray list of positive clones using 1500
as the threshold. See Figure 7 for an example of final PICOGREEN® data obtained from a PCR screen. Open the rearray list in the TECAN robot and rearray clones from the glycerol stock plate into a 96-well deep-well plate containing Superbroth and appropriate antibiotics. After the plates have been incubated, the cultures in them are sequenced in an automated process.

Example 3 - Transposon Expedited Multiplex Sequencing of unknown sequences in vector PCR4Blunt-TOPO

TRANSPOSON REACTION

The transposon reaction was performed using the Genome Priming System (GPS) kit from New England BioLabs. All of the components provided in the kit were used with the exception of the transposon GPS1. The transposon utilized was a modified version of GPS1. Two variations were generated. In one variant, pGPS-Apra, the existing resistance marker has been replaced by the apramycin resistant gene. This was performed to ensure that the vector could be used to sequence any vector regardless of resistance marker. In another variant, pGPS-Apra-Y2um a "stuffer" region containing the yeast 2um replication origin was cloned into the transposon thus increasing its size to 5 kb (Figure 6). This alteration was created in order to improve the efficiency of the process by screening out transpositions into the vector backbone. The reaction was transformed by electroporation into DH10B E.coli competent cells and plated on Luria-Bertani (1.0% bacto-tryptone, 0.5% bacto-yeast extract, 1.0% NaCl, pH 7.0) agar plates containing the appropriate antibiotics (apramycin (100 μg/ml)/kanamycin (50 μg/ml)) for selection of clones harboring the transposon. The growth period is 20-22 hours.

TEMS (Transposon Expedited Multiplex Sequencing)

The size of all target DNA sequences subjected to TEMS was evaluated by restriction digest of the target DNA-containing vectors with NotI. Plasmid sets containing target DNA sequences of various sizes contained in the same vector backbone were pooled. Equimolar concentrations were calculated for each target DNA sequence prior to pooling. The transposon
reaction was performed treating the pooled DNA template as a single template. For six-fold coverage of each base of each target DNA sequence, the total size of each target DNA sequence was multiplied by our standard factor of 18 clones per 1 kb of sequence to determine the number of randomly selected clones to sequence. Based on the total size of the vector and the total size of the target DNA sequence, a probability of the location of the transposon insertion was calculated. Multiplying this factor with the number of clones for six-fold coverage determines how many clones to randomly select and subject to the PCR screen.

The calculated number of colonies were selected and inoculated into Superbroth (bacto-tryptone, yeast extract, NaCl, NaOH) media containing the appropriate antibiotics on 96 or 384 well culture plates. These culture plates were grown overnight without shaking for 18-22 hours depending on the format of the growth plate (96/384 well). Glycerol stocks were made for temporary storage of the plate. 50% glycerol was added directly to the plate and mixed. This allowed indefinite storage in \(-80\) C.

**SCREENING OUT VECTOR INSERTIONS BY PCR**

The primer sequence selected to amplify the vector portion of the clones for screening were the complement sequences of the universal primers M13F(-20) and M13R. The sequence is as follows: 5'-ACTGGCCGTCGTTTTAC-3' and 5'-CATGGTCATAGCTGTT-3'. Culture PCR was assembled with the following conditions: the template was deposited into the PCR reaction reagent preparation using a pin replicator tool which aliquoted approximately 1 µL. For the 96-well format, the 10x PCR buffer composition was 100mM Tris-HCl pH 8.3; 500mM KCl; 15mM MgCl2; 0.01% w/v gelatin. Five picomoles of the forward primer and five pmoles of the reverse primer were used in each reaction. The dNTP concentration in the reaction was 2.5 mM final concentration. The quantity of Taq polymerase added to each reaction was 0.1 units. All reactions were carried out in a total volume of 50 µL. The 384-well format PCR was done in a total volume of 25 µL using the TaKaRa Z-Taq kit obtained from Panvera Corporation, 545 Science Drive, Madison, WI 53711. This kit consists of Z-taq polymerase (2.5 units/ µL), 10x Z-Taq buffer containing a 30 mM concentration of Mg2+, and a dNTP mixture (2.5 mM each). Thermocyclers were run with a heated lid (MJ Research). This PCR cycle was designed to
amplify only the vector without the transposon due to the short elongation time. Black PCR plates were used so that the fluorescent analysis could later be performed directly in the plates.

**ANALYSIS CONDITIONS USING PICOGREEN®**

Upon completion of the PCR, 25 ul of PICOGREEN® was added the PCR plate. A dilution of PICOGREEN® stock 1:200 in 1x TE (tris-acetate, EDTA) buffer at 50 ul volume was mixed with 10 ul of PCR reaction. The plate was read in a spectrofluorometer where the molecules are excited at 480 nm and the signal is emitted at 540 nm. The PCR products fluorescing above a predetermined threshold were the positive clones to be sequenced. The threshold was experimentally determined using known constructs containing the transposon either in the vector or in an insert to the vector. The positive clones were sorted and compressed into a format suitable for array by the TECAN robotic arm. Glycerol stocks were inoculated into the 96 deep well plate containing 1.5 ml of Superbroth (bacto-tryptone, yeast extract, NaCl, 5N NaOH) and appropriate antibiotics. These cultures were grown for a 24 hour time period and then the DNA targets from the cultures are sequenced using the primers at either end of the transposon.

**SEQUENCE ASSEMBLY**

After vector trimming and base calling of the raw sequences using Phred software, the sequences are assembled using Phrap software (see Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. 1998. Genome Research 8:175-185. and Brent Ewing and Phil Green Base-calling of automated sequencer traces using phred. II. Error probabilities. 1998. Genome Research 8:186-194). The assembly is reviewed using consed software. (See Consed: A Graphical Tool for Sequence Finishing Gordon et al 1998.) Criteria for completion are high quality, six-fold coverage of the entire full-length gene. The expected error rate per 10 kb of sequence should be < 0.1. If these criteria are not met, primers are selected along with templates for primer walking sequencing of the error-prone section of target DNA sequence. The primer walking sequencing is
done on either the ABI 310 sequencer or through high-throughput sequencing using the ABI 3700 depending on volume of sequencing needed, and those sequences are subsequently added to the assembly. BLAST analysis is done for comparison especially to known genes. (See:

It is to be understood that the foregoing examples are exemplary and explanatory only and are not restrictive of the invention. Various changes may be made to the embodiments described above by one of skill in the art without departing from the scope of the invention, as defined by the following claims.

Comprises/comprising and grammatical variations thereof when used in this specification are to be taken to specify the presence of stated features, integers, steps or components or groups thereof, but do not preclude the presence or addition of one or more other features, integers, steps, components or groups thereof.
Applicants claim:

1. A process for parallel, transposon-mediated sequencing of DNA, the process comprising:
   i) providing one or more target DNA sequences, each inserted into a vector at an insertion site;
   ii) amplifying each target DNA-containing vector and pooling them, wherein each target DNA sequence is represented in the pool in an substantially equal amount per kb;
   iii) exposing the pool of target DNA-containing vectors to a selectable transposon, wherein the selectable transposon integrates into the target DNA-containing vectors at substantially random sites to form a pool of target DNA- and transposon-containing vectors;
   iv) transforming cells with the pool of target DNA- and transposon-containing vectors and isolating and growing a representative number of individual transformants into cultures under selection conditions;
   v) performing a polymerase chain reaction on DNA from each culture, wherein the polymerase chain reaction uses a pair of primers complementary to the 3' ends of the vector sequence at the insertion site and has an extension time during each reaction cycle sufficient to efficiently produce a full-length copy of the vector sequence but too short to efficiently produce a full-length copy of the vector sequence with the transposon inserted;
   vi) measuring the amount of DNA produced in each polymerase chain reaction; and
   vii) sequencing the transposon flanking regions of the target DNA-containing vectors from those cultures corresponding to polymerase chain reactions which produced substantial amounts of DNA.

2. The process of claim 1, wherein one or more of steps i) to vii) are automated.

3. The process of claim 1, wherein the step of measuring the amount of DNA produced in each polymerase chain reaction comprises adding a fluorescent stain selective for dsDNA to the finished polymerase chain reaction and measuring the resulting amount of fluorescence.

4. The process of claim 3, wherein one or more of steps i) to vii) are automated.

5. The process of claim 3, wherein the fluorescent stain selective for dsDNA is selected from the group consisting of PICOGREEN® and bisbenzimide dyes.

6. The process of claim 3, wherein the fluorescent stain selective for dsDNA is PICOGREEN®.
7. The process of claim 1, wherein the vector and transposon carry different selectable markers.

8. The process of claim 1, wherein the transposon is at least 3 kb.

9. The process of claim 8, wherein the transposon is at least 4 kb.

10. The process of claim 9, wherein the transposon is GPS-Apra-2um-URA3.

11. The process of claim 8, wherein the transposon is at least 5 kb.

12. The process of claim 1, wherein the transposon is equal to or greater than the vector in length.

13. The process of claim 12, wherein the transposon is at least about 5/4 the length of the vector.

14. The process of claim 13, wherein the transposon is about 5 kb and the vector is about 4 kb.

15. The process of claim 12, wherein the transposon is at least about 5/3 the length of the vector.

16. The process of claim 15, wherein the transposon is about 5 kb and the vector is about 3 kb.

17. A process for parallel, transposon-mediated sequencing of DNA, the process being substantially as hereinbefore disclosed with reference to the Examples.

DATED this 22nd day of December 2005
AVENTIS PHARMACEUTICALS INC

WATERMARK PATENT & TRADE MARK ATTORNEYS
290 BURWOOD ROAD
HAWTHORN VICTORIA 3122
AUSTRALIA

KJS/JMN
Figure 1: TEMS Process

transposon

vector

target DNA

Insertion

Colony Picking

PCR

8 kb

3 kb

Screen

Gel Electrophoresis

dsDNA Fluorescence

Submission to DNA Sequencing

Analyze Data Rearray Positives
Figure 2
Figure 4. Results of a quick screen of various PCR kits tested during the 384-well optimization process.
Figure 5. PCR results showing the appearance of a 9 kbp band believed to be the amplified product of vector containing the transposon. This upper band is the result of an excessive extension time.
Figure 6. Results of 48 °C and 68 °C gradient PCR tested with 30, 35, 40, and 45 cycles.
Figure 7: Correlation between electrophoresis results and PICOGREEN® staining for determining significant amounts of dsDNA

|      | 1    | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    | 11    | 12    | 13    | 14    | 15    | 16    | 17    | 18    |
|------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| A    | 1060.36 | 895.05 | 698.65 | 2406.91 | 2318.39 | 429.20 | 582.09 | 315.05 | 446.50 | 481.50 | 419.26 | 653.24 | 562.34 | 705.15 | 2564.36 | 886.19 | 674.13 | 400.86 | 713.32 | 445.49 |
| B    | 3536.48 | 3421.30 | 8623.91 | 9256.58 | 945.80 | 96.28 | 702.40 | 972.30 | 1720.87 | 1273.87 | 1678.33 | 7705.93 | 7250.67 | 7250.67 | 5205.10 | 7250.67 | 578.50 | 938.92 | 783.60 | 823.28 |
| C    | 281.43 | 3658.90 | 1054.79 | 665.98 | 1470.44 | 375.70 | 954.16 | 1881.78 | 1792.83 | 480.64 | 384.83 | 543.32 | 502.94 | 456.07 | 435.44 | 1564.98 | 2302.68 | 2727.47 | 3352.99 |
| D    | 4857.46 | 17.5.75 | 997.07 | 224.77 | 876.52 | 364.63 | 761.26 | 1713.71 | 1390.96 | 865.11 | 1648.74 | 2141.41 | 1064.34 | 914.84 | 421.47 | 895.27 | 875.91 | 425.30 |
| E    | 912.34 | 967.77 | 3343.23 | 560.33 | 470.20 | 433.29 | 322.81 | 1000.31 | 489.60 | 7690.39 | 430.19 | 752.68 | 531.95 | 3743.89 | 419.24 | 930.36 | 427.54 | 441.30 |
| F    | 340.00 | 918.55 | 917.61 | 2652.18 | 7301.56 | 862.84 | 2603.82 | 899.94 | 745.92 | 717.38 | 875.11 | 859.67 | 855.43 | 745.50 | 526.02 | 527.76 | 627.98 | 360.93 |
| G    | 1518.73 | 2163.87 | 1846.69 | 1029.84 | 473.82 | 934.26 | 843.75 | 2901.56 | 597.96 | 2416.00 | 1491.90 | 629.90 | 449.26 | 335.01 | 786.23 | 824.78 | 1482.40 | 710.79 |
| H    | 3232.52 | 3529.29 | 667.50 | 842.86 | 2194.03 | 758.15 | 746.00 | 563.31 | 2184.25 | 806.75 | 722.38 | 2551.48 | 665.50 | 909.50 | 1560.90 | 886.63 | 1540.00 | 1284.08 |

Legend:
- false positive
- true positive
- match gel
cutoff=1500

Rows A and B
Rows C and D
Rows E and F
Rows G and H