(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 September 2007 (13.09.2007)

PCT

(10) International Publication Number
WO 2007/102006 A2

(51) International Patent Classification:
C12N 15/09 (2006.01)

(21) International Application Number:
PCT/GB2007/000808

(22) International Filing Date: 8 March 2007 (08.03.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/780,715 9 March 2006 (09.03.2006) US

(71) Applicant *(for all designated States except US)*: **SOLEXA LIMITED** [GB/GB]; Chesterford Research Park, Little Chesterford, Saffron Walden, Essex CB10 1XL (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **RIGATTI, Roberto** [IT/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden, Essex CB10 1XL (GB). **GORMLEY, Niall, Anthony** [IE/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden, Essex CB10 1XL (GB). **BURGAR, Helen, Rachel** [GB/GB]; SOLEXA LIMITED, Chesterford Research Park, Little Chesterford, Saffron Walden, Essex CB10 1XL (GB).

(74) Agents: **WHITE, Nina, Louise** et al.; BOULT WADE TENNANT, Verulam Gardens, 70 Gray's Inn Road, London WC1X 8BT (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2007/102006 A2

(54) **Title:** NON-CLONING VECTOR METHOD FOR GENERATING GENOMIC TEMPLATES FOR CLUSTER FORMATION AND SBS SEQUENCING

(57) **Abstract:** The invention relates to a method of preparing a 5' and 3' modified library of template polynucleotides for use in solid phase nucleic acid amplification and sequencing by synthesis which does not require cellular transformation.

Non-cloning vector method for generating genomic templates for cluster formation and SBS sequencing.

Field of the invention

This invention relates to the field of nucleic acid amplification and sequencing. More specifically, the invention relates to a method of preparing a 5′ and 3′ modified library of template polynucleotides for use in solid phase nucleic acid amplification and sequencing by synthesis which does not require cellular transformation.

Background to the invention

Several publications and patent documents are referenced in this application in order to more fully describe the state of the art to which this invention pertains. The disclosure of each of these publications and documents is incorporated by reference herein.

The ability to acquire and analyse DNA sequence data has increased phenomenally over the past few years. As a result nucleic acid analysis has become increasingly important in many areas of biology, biotechnology and medicine. Various approaches have been proposed for sequencing large genomes such as whole genome shotgun sequencing.

Whole genome shotgun sequencing has been made feasible by recent advances in both sequencing technology and the computational processes required to assemble contiguous stretches of DNA from large groups of overlapping DNA sequences. Previous sequencing methodologies relied on prior knowledge of restriction map data and collections of

stable clones containing large fragments of DNA. In contrast, shotgun sequencing relies on random cloning of relatively small pieces of genomic DNA and subsequent sequencing of many of these clones to provide 10 to 12 fold sequence coverage of the whole genome.

The procedure for shotgun sequencing involves a number of steps. Briefly, the first of these is usually the isolation and preparation of quantities of genomic DNA from the organism to be sequenced. Next genomic DNA is randomly sheared into smaller fragments which are used to construct plasmid libraries and transformed into bacteria. The bacteria are propagated and used to maintain and replicate the DNA fragments. The DNA fragments can then be purified from individual bacterial clones and sequenced. Sequence data obtained in this way is assembled into contiguous sequence.

Propagation and maintenance of these DNA fragments in bacterial clones suffers from a number of drawbacks however which lead to cloning bias, that is, an under-representation of particular DNA fragments. Such cloning bias leads to gaps in shotgun libraries and incomplete genome sequence.

Cloning bias may occur for a number of different reasons. For example, to differentiate between host cells carrying non-recombinant and recombinant vector, DNA fragments are usually inserted into a chromogenic gene such as Lac Z. On insertion of the DNA fragment the chromogenic gene is inactivated with the result that recombinant colonies are white (Messing et al, Proc. Natl. Acad. Sci. USA (1977) 79, 3642-3646; Yanisch-Perron et al, Gene (1985)

33, 103-119). When the numbers of non-recombinant colonies is high in comparison to recombinant colonies, it may be difficult to successfully differentiate between the two, leading to errors in cloned libraries. To reduce such problems, other vectors have been developed with positive selection capability. Many of these vectors function on the basis of insertional inactivation of lethal genes (U.S. Patent Nos. 5,910,438 and 5,891,687). Both systems of positive and negative selection are susceptible to a tendency for high numbers of false positives due to, for example, exonuclease activity in reagents. Exonuclease activity can delete nucleotides from the cloning site, resulting in re-circularisation of vector with inactivation of chromogenic or lethal genes, which in turn leads to false positive transformants.

Cloning bias may also be caused by high AT content in the genome to be sequenced or the formation of strong secondary structure. Other causes are vector driven expression of deleterious sequences, insert-driven transcription into the vectors interfering with vector stability, cloning of promoter regions or other control elements and also unreliable selection of bacterial clones with plasmids not containing DNA fragments.

Other problems encountered in maintaining large numbers of bacterial clones include the large volume of space required to store and propagate libraries of individual genomic clones representative of a whole genome and also the time involved preparing and manipulating them. In addition, as genome complexity increases, so achieving library diversity becomes more difficult. For example, $6x10^9$ clones

would be required for a diploid human library. A library of this size is routinely not achievable or maintainable due to problems with efficiency of bacterial transformation.

Because the amount of DNA that can be isolated from even large numbers of cells is very small, the only other practical method with which to prepare enough DNA for sequencing applications involves amplification of DNA in vitro.

The first PCR based genome amplification approach was developed by Nelson et al (Proc. Natl. Acad. Sci. USA (1989) 86, 6686-6690) who used primers designed to anneal to repetitive sequences (Alu repeats) within the human genome for the purpose of amplification of sequences. Since Alu repeats are not randomly dispersed throughout the genome this approach leads to bias towards regions containing repeats.

Telenius et al (Genomics (1992) 13, 718-725) used degenerate oligonucleotide primed PCR to amplify whole genome DNA. The main disadvantages of this approach are the generation of non-specific amplification artefacts (Cheung and Nelson, Proc. Natl. Acad. Sci. USA (1996) 93, 14676-14679) and incomplete coverage of loci (Paunio et al, Clin. Chem. (1996) 42, 1382-1390).

In WO03/050242, applicants fragment genomic DNA and ligate a universal sequence to the 3′ end of the random DNA fragments. The method suffers from a number of problems - because only one universal sequence is ligated to the 3′ end, it is likely that non-specific amplification artefacts

(primer-dimers) will accumulate during whole genome PCR. Also, because each fragment is of a different length and comprises the universal sequence at the 3′ end, use of a second primer for the purposes of conventional sequencing potentially leads to background signal and inefficient sequencing for example where the universal sequence overlaps with genomic DNA. In addition, since the library fragment size is variable, PCR amplification will favour shorter fragments, again leading to amplification bias.

Callow et al (Nucleic Acids Research (2004) 32:2, e21) describe an approach for specific selection and amplification of genomic DNA fragments of interest. Application of the method to whole genome amplification suffers from the drawback that large numbers of oligonucleotide adaptors are required.

Other methods of PCR amplification utilising adaptors are also known, however, these methods suffer from the problem that two or more randomly cut DNA fragments may ligate during cloning steps to form DNA sequences not contained within the normal genome.

Thus it is desirable to provide a method for efficiently amplifying genomic DNA for sequencing which does not suffer the problems of cloning bias, amplification bias, non-specific amplification or random ligation of sequence.

Summary of the invention

In a first aspect the invention provides a method of generating a 5′ and 3′ modified library of template

polynucleotide molecules from one or more primary polynucleotide molecules characterised in that said method is carried out entirely *in vitro*, comprising:

(a) fragmenting said one or more primary polynucleotide molecules to produce target polynucleotide duplexes;

(b) ligating a vector polynucleotide sequence to the target polynucleotide duplexes to form combined ligated polynucleotide sequences under conditions biased towards formation of the combined ligated polynucleotide sequences that contain the target duplexes;

(c) preparing an amplification reaction comprising said combined ligated polynucleotide sequences and at least two different primer oligonucleotides wherein said two different primer oligonucleotides comprise sequences complementary to regions of the vector polynucleotide sequence portion of the combined ligated polynucleotide sequences, said regions being located on either side of the inserted target; and

(d) carrying out an amplification reaction in which one of said at least two different primer oligonucleotides is annealed to complementary parts of each strand of the vector polynucleotide sequence portion of the combined ligated polynucleotide sequences and extended by sequential addition of nucleotides to generate amplification products complementary to the combined ligated polynucleotide sequences and wherein said amplification products comprise a common sequence Y at their 5′ ends and common sequence Z at their 3′ ends and collectively provide a 5′ and 3′ modified library of template polynucleotide molecules, wherein the

templates comprise the targets plus additional common sequences Y and Z.

A second aspect of the invention relates to the use of a 5′ and 3′ modified library of template polynucleotide molecules prepared according to the method of the first aspect of the invention for solid-phase nucleic acid amplification. Thus, in a particular embodiment the invention provides a method of solid-phase nucleic acid amplification of template polynucleotide molecules which comprises: preparing a 5′ and 3′ modified library of template polynucleotide molecules which have common sequences at their 5′ and 3′ ends using the method according to the first aspect of the invention and carrying out a solid-phase nucleic acid amplification reaction wherein said template polynucleotide molecules are amplified.

Brief description of the drawings

Figure 1 is a diagrammatic representation of the non-cloning vector method.

Figure 2 is a diagrammatic representation of the non-cloning vector method using TA vector cloning.

Figure 3(a) shows the distribution of fragments on 4-20% TBE gel (Invitrogen, EC62252) of fragmented Phix174 DNA, stained in Vistra green stain according to the manufacturer's instructions (Amerham, RPN5786). Sonicated phiX174 DNA. 5 ug of Phix174 DNA was sonicated at power 1 for 0, 0.5, 1 and 2 mins.

Figure 3(b) shows the distribution of fragments on 4-20% TBE gel (Invitrogen, EC62252) of fragmented Phix174 DNA, stained in Vistra green stain according to the manufacturer's instructions (Amerham, RPN5786). Nebulised Phix174 DNA. 10 ug Phix174 DNA was nebulised at

i) 30 psi (2 bar) for 4 mins on ice in shearing buffer (10 mM Tris, 1 mM EDTA, 10 % glycerol, pH 8.0)

ii) 32 psi (2.2 bar) for 5 mins on ice, in nebuliser buffer (37 mM Tris, 5.5 mM EDTA, 53 % glycerol, pH 7.5).

Figure 4 shows the sequence of the KH1 vector. The EcoRV restriction site is shown in bold and underlined.

Figure 5 shows the cloning site of KH1.

Figure 6 shows the size distribution of the 5′ and 3′ modified library of template polynucleotides on a 4-20% TBE PAGE gel stained in Vistra green stain and scanned on Typhoon.

Figure 7 shows photographs of two 2 % TAE agarose gel showing the results of a PCR screen on 20 isolated colonies from two ligations

i) Purified EcoRV digested ligation

ii) Purified ligation (no EcoRV digestion)

The PCR screen was used to determine the number of transformants that contain genomic DNA inserts. The sizes of the PCR products were determined from the Invitrogen 1 kb plus DNA ladder (10787-018) that was loaded. Religated KH1 containing no insert gave a PCR product of 322 bp in size.

Any clones containing genomic DNA have a PCR product >322 bp in size.

Detailed description of the invention

The invention relates to a method of generating a 5′ and 3′ modified library of template polynucleotide molecules for use in sequencing by synthesis that does not involve transformation or propagation of the library in a biological cell. As a result, the method of the invention is advantageous because it avoids cloning bias present in traditional library construction.

The method comprises a first step of fragmenting one or more primary polynucleotide molecules to produce target polynucleotide duplexes.

As used herein, the term "polynucleotide" refers to deoxyribonucleic acid (DNA), but where appropriate the skilled artisan will recognise that the method may also be applied to ribonucleic acid (RNA). The terms should be understood to include, as equivalents, analogs of either DNA or RNA made from nucleotide analogs and to be applicable to single stranded (such as sense or antisense) and double stranded polynucleotides. The term as used herein also encompasses cDNA, that is complementary or copy DNA produced from an RNA template, for example by the action of reverse transcriptase.

The primary polynucleotide molecules may originate in double-stranded DNA (dsDNA) form (e.g. genomic DNA fragments, PCR and amplification products and the like) or

may have originated in single-stranded form, as DNA or RNA, and been converted to dsDNA form. By way of example, mRNA molecules may be copied into double-stranded cDNAs suitable for use in the method of the invention using standard techniques well known in the art. The precise sequence of the primary polynucleotide molecules is generally not material to the invention, and may be known or unknown.

In a particular embodiment, the primary polynucleotide molecules are DNA molecules. More particularly, the primary polynucleotide molecules represent the entire genetic complement of an organism, for example plants, bacteria, viruses, mammals, and are genomic DNA molecules which include both intron and exon sequence (coding sequence), as well as non-coding regulatory sequences such as promoter and enhancer sequences. Although it could be envisaged that particular sub-sets of polynucleotide sequences or genomic DNA could also be used, such as particular chromosomes, for example. Yet more particularly, the sequence of the primary polynucleotide molecules is not known. Still yet more particularly, the primary polynucleotide molecules are human genomic DNA molecules.

The sequence of the primary polynucleotide molecules may be the same or different, for example, a mixture of primary polynucleotide molecules of different sequences may be prepared by mixing a plurality, greater than one, of individual primary polynucleotide molecules. For example, DNA from more than one source can be prepared if each DNA sample is first tagged to enable its identification after it has been sequenced. Many different suitable DNA-tag methodologies already exist in the art, for example as

described in WO05068656, which is included herein by reference, and are well within the purview of the skilled person.

Random fragmentation refers to the fragmentation of a polynucleotide molecule in a non-ordered fashion by enzymatic, chemical or mechanical means. Such fragmentation methods are known in the art and utilise standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition). The random fragmentation is designed to produce fragments irrespective of the sequence identity or position of nucleotides comprising and/or surrounding the break. More particularly the random fragmentation is by mechanical means such as nebulisation or sonication to produce fragments of about 50 base pairs in length to about 1500 base pairs in length, still more particularly 50 to 700 base pairs in length yet more particularly 50-400 base pairs in length. Most particularly, the method is used to generate smaller fragments of from 50-150 base pairs in length.

Fragmentation of polynucleotide molecules by mechanical means (nebulization, sonication and Hydroshear for example) results in fragments with a heterogeneous mix of blunt and 3'- and 5'-overhanging ends. It is therefore desirable to repair the fragment ends using methods or kits (such as the Lucigen DNA terminator End Repair Kit) known in the art to generate ends that are optimal for insertion, for example, into blunt sites of cloning vectors. In a particular embodiment, the fragment ends of the population of nucleic acids are blunt ended. More particularly, the fragment ends are blunt ended and phosphorylated.

In a particular embodiment of the invention, the targets may be treated to obtain a 3′ overhanging sequence of one of more nucleotides. This may help to prevent the blunt ended targets for ligating to each other and forming concatamers. The 3′ overhanging sequence may be obtained using a nucleoside triphosphate and an exonuclease deficient DNA polymerase, such as Klenow exo minus DNA polymerase, or Taq polymerase.

The term "target polynucleotide duplexes" refers to nucleic acid molecules that it is desired to sequence. Template polynucleotides are target polynucleotides that have known ends, and can thus be amplified. Templates can originate as duplexes or single strands. For ease of reference, single stranded templates are described herein. When viewed as a single strand, the 5′ ends and the 3′ ends of the templates may comprise different sequences, herein depicted as Y and Z for ease of reference. The other strand will be amplified in any amplification reaction, but would be noted 5′-X-target′-Y′-3′, where X is the complement of Z, and target′ and Y′ are the complementary copies of the target region and sequence Y. This strand may be present in many or all of the processes described herein, but is not further discussed.

The second step of the method comprises ligating a linearised vector polynucleotide sequence to the target polynucleotide duplexes to form combined ligated polynucleotide sequences.

A linearised vector polynucleotide sequence can be obtained from a circular polynucleotide duplex that has been opened to give a linear duplex. The circular vector will generally be opened with a restriction endonuclease to give a blunt ended sequence at a specific location, The linearised vector polynucleotide sequences will generally comprise sequences near to each terminus that will contain sequences X and Y that become part of the template upon amplification. The linearised blunt ended vector may be treated with a nucleotide triphosphate and a nucleotide polymerase as described above to give a 3′ overhanging sequence. The nucleoside added to the linearised vector may be complementary to the nucleoside added to the target fragments. Such a treatment may help to prevent the re-circularisation of the vector without a target fragment during the ligation step. One key objective of the method is to bias the closure of the vector such that all the circularised constructs contain target fragments, and the constructs that do not carry a target sequence remain linear. Thus upon amplification with primers X and Y, where X is complementary to template sequence Z, the only amplified product that can be produced is the desired template.

Combined ligated sequences particularly comprise at least one target polynucleotide duplex ligated, as an insert, to both termini of one linearised vector polynucleotide sequence to form a circular polynucleotide molecule. The at least one target polynucleotide duplex may be in either orientation (sense or antisense). The definition also captures the case where more than one target polynucleotide duplexes are ligated together to form a

concatemer, the concatemer then being ligated, as an insert, to both termini of the linearised vector polynucleotide sequence to form a circular polynucleotide molecule.

It will be clear to one skilled in the art that the combined ligated polynucleotide sequences could also be formed by ligation of multiple linearised vector polynucleotide sequences with one or more target polynucleotide duplexes. However in this case there is a bias against formation of concatemers of linearised vector polynucleotide sequences. Thus if more than one linearised vector polynucleotide sequence is present in a combined ligated sequence it is an object of the invention that it will not be ligated directly to another linearised vector polynucleotide sequence and there will be at least one target polynucleotide duplex between, and separating, them. Therefore it will also be clear to the skilled artisan that there will be variation in the sizes of the combined ligated polynucleotide sequences, that is, within a number of such sequences generally they will not be of a fixed size.

Linearised vector polynucleotide sequence can be produced by a number of methods well known in the art (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition). In a particular embodiment, the linearised vector is produced from circular polynucleotide molecules such as plasmids cut (digested) with one or more restriction endonuclease enzymes for example. In a more particular embodiment, the one or more restriction endonuclease enzymes are 'rare blunt cutters'.

Equally, a linearised vector could be produced by amplification of a portion of a circular polynucleotide molecule such as a plasmid, for example, to produce a linear amplification or PCR product. It is well within the ability of one skilled in the art to produce linearised vector polynucleotide sequences having defined ends. In a particular embodiment, the ends of the linearised vector polynucleotide sequence are blunt-ended. In a further embodiment, the ends of the linearised vector comprise a sequence Y and Z that can form the ends of the template fragments.

Ligation methods are known in the art and utilise standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition). Such methods utilise ligase enzymes such as DNA ligase to effect or catalyse joining of the ends of the two polynucleotide strands of, in this case, the linearised vector polynucleotide sequence and the target polynucleotide duplexes such that covalent linkages are formed. In this context, joining means covalent linkage of two polynucleotide strands which were not previously covalently linked. In a particular aspect of the invention, such joining takes place by formation of a phosphodiester linkage between the two polynucleotide strands, but other means of covalent linkage (e.g. non-phosphodiester backbone linkages) may be used.

It is usual and commonplace in the art for the ends of a linearised vector to be non-phosphorylated to prevent re-ligation of the vector during subsequent ligation reactions. Re-ligation of vector significantly reduces the efficiency of ligation reactions. As would be known by the skilled

practitioner, a ligase enzyme requires the presence of a phosphate molecule at the 5' end of the molecule to be ligated. Such a phosphate moiety may be provided by using insert sequences (target polynucleotide duplexes) with phosphorylated ends. As a consequence ligation of a phosphorylated insert sequence and a non-phosphorylated vector sequence results in a ligated sequence which contains nicks at the site of the ligation. On transformation of the ligated sequence, into for example a bacterial cell, these nicks are repaired by the cell. Surprisingly *in vitro* nick repair methods are inefficient and consequently PCR amplification across nicks repaired by *in vitro* methods will fail.

The present inventors have found a novel solution to this problem which not only significantly increases the efficiency of the ligation reaction, it also negates the requirement for a transformation step whilst also enabling amplification across the 'join'. By carrying out a ligation reaction in the presence of a restriction endonuclease enzyme when using phosphorylated linearised vector polynucleotide sequence, if the vector ligates to itself the restriction endonuclease enzyme re-cleaves the vector. Therefore it is advantageous that if the vector re-ligates, the restriction endonuclease recognition site is reformed such that the vector is susceptible to further digestion by the same enzyme. Thus the ligation efficiency is significantly increased and formation of combined ligated sequences is favoured over, for example, self-ligation of vector or target polynucleotide duplexes. Thus there is a bias against re-ligation of linearised vector without insert (target polynucleotide sequence).

It will be clear to one skilled in the art that when using a restriction endonuclease enzyme, even a rare cutter, some of the target polynucleotide sequences may be cleaved. Therefore it may be desirable to use two different rare blunt cutters to avoid any possible bias or loss of sequence from the 5′ and 3′ modified library. By way of non-limiting example, in one embodiment the vector could be designed so that the restriction site of one enzyme can fit inside the restriction site of the other enzyme, within the one vector, with both enzymes linearising the vector at the same position. Enzymes such as BoxI and eco721 could be used which have recognition (restriction) sites of GACNN|NNGTC and CAC|GTG respectively wherein '|' represents the point of cleavage and the vector would contain the cloning site GACACGTGTC. In another embodiment, the enzymes could cut at two different sites.

The 5′ and 3′ modified library can therefore be prepared in two halves. Each half of the 5′ and 3′ modified library can be cut with a different blunt-end cutter, so that if a genomic fragment is digested by one of these restriction enzymes the undigested fragment is likely to be present in the 5′ and 3′ modified library prepared using the other restriction enzyme. Both halves of the 5′ and 3′ modified library can then be combined.

It could also be envisaged that restriction enzymes that are blocked by mammalian CpG methylation could also be used. The genomic DNA could be methylated prior to the ligation. This would prevent digestion of the genomic DNA during linearisation of the re-ligated vector.

Since both the vector and insert are phosphorylated, there are no nicks in the combined ligated sequence and rendering it possible to directly amplify across the 'joins'. In a particular aspect, therefore, the ends of the linearised vector polynucleotide sequence of the present invention are phosphorylated. It could be envisaged that non-enzymatic ligation techniques (e.g. chemical ligation) could also be used provided that the non-enzymatic ligation leads to the formation of a covalent linkage which allows read-through of a polymerase, such that the resultant construct can be copied in an amplification reaction.

The desired products of the ligation reaction are combined ligated sequences in which the target polynucleotide duplexes are ligated into the linearised vector polynucleotide sequence to form a continuous, circular polynucleotide molecule. Conditions of the ligation reaction should, therefore, be optimised to maximise formation of this product, in preference to, for example, linear polynucleotide molecules.

In an alternative embodiment it may be envisaged that the restriction endonuclease enzyme may be omitted and added in a later step or a further restriction digest step be added subsequent to the ligation reaction.

In yet another alternative embodiment, the target polynucleotide sequences and vector polynucleotide sequences are prepared with single overhanging nucleotides by, for example, activity of certain types of DNA polymerase such as Taq polymerase or Klenow exo minus polymerase which has a nontemplate-dependent terminal transferase activity that

adds a single deoxyadenosine (A) to the 3′ ends of, for example, PCR products. Such enzymes can be utilised to add a single nucleotide, for example 'A' to the 3′ terminus of each strand of the polynucleotide duplexes. Thus, an A could be added to the 3′ terminus of each duplex strand of polynucleotide duplex by reaction with Taq or Klenow polymerase whilst the vector polynuceotide could be a T-vector with a compatible 'T' present on the 3′ terminus of each duplex strand. This end modification would prevent self-ligation of both vector and target such that there is a bias towards formation of the combined ligated sequences (Figure 2). This method of ligation is the first step of the 'TA cloning' protocol that is known in the art.

The 5′ and 3′ modified library of template polynucleotide molecules is particularly suitable for use in solid phase sequencing methods. Because sequence reads may be short, that is around 25 base pairs in length, unlike conventional methods of cloning, it is of no consequence if multiple different target polynucleotide duplexes are ligated into a single linearised vector polynucleotide. Because the sequence read is shorter than the length of the individual target polynucleotide duplexes, there is no risk of artificial concatamers of sequence data being produced.

Optionally the combined ligated polynucleotide sequences, unligated vector polynucleotide sequence, and unligated target polynucleotide duplexes may be purified from any components of the ligation and/or restriction endonuclease digest(s) such as enzymes, buffers, salts and the like. Suitable purification methods, for example polyacrylamide gels, are known in the art and utilise

standard methods (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition).

In a next step according to the invention an amplification reaction is prepared. The contents of an amplification reaction are known by one skilled in the art and include appropriate substrates (such as dNTPs), enzymes (e.g. Taq polymerase) and buffer components required for an amplification reaction. Generally amplification reactions require at least two amplification primers, often denoted "forward" and "reverse" primers (primer oligonucleotides) that are capable of annealing specifically to a part of the polynucleotide sequence to be amplified under conditions encountered in the primer annealing step of each cycle of an amplification reaction. In certain embodiments the forward and reverse primers may be identical. Thus the primer oligonucleotides must include a "template-specific portion", being a sequence of nucleotides capable of annealing to a part of, that is, a primer-binding sequence, in the polynucleotide molecule to be amplified (or the complement thereof if the template is viewed as a single strand) during the annealing step.

In the context of the present invention, the term "polynucleotide molecule to be amplified" refers to the original or starting template added to the amplification reaction. The "template-specific portion" in the forward and reverse amplification primers refers to a sequence capable of annealing to the original or starting template present at the start of the amplification reaction and reference to the length of the "template-specific portion" relate to the length of the sequence in the primer which anneals to the

starting template. It will be appreciated that if the primers contain any nucleotide sequence which does not anneal to the starting template in the first amplification cycle then this sequence may be copied into the amplification products (assuming the primer does not contain a moiety which prevents read-through of the polymerase). Hence the amplified strands produced in the first and subsequent cycles of amplification cycles may be longer than the starting template. Typically the invention relates to the use of forward and reverse primers of from 20 to 25 nucleotides in length although it is conceivable that primers may be used which are longer than this, such as 26-30, 30-35 or 35 to 45 nucleotides in length. The nucleotide sequences of the template-specific portions of the forward and reverse primers are selected to achieve specific hybridisation to the template to be amplified under the conditions of the annealing steps of the amplification reaction, whilst minimising non-specific hybridisation to any other sequences present. Skilled readers will appreciate that it is not strictly required for the template-specific portion to be 100% complementary to the template, a satisfactory level of specific annealing can be achieved with less than perfectly complementary sequences. In particular, one or two mis-matches in the template-specific portion can usually be tolerated without adversely affecting specificity for the template. Therefore the term "template-specific portion" should not be interpreted as requiring 100% complementarity with the template. However, the requirement that the primers do not anneal non-specifically to regions of the template other than their respective primer-binding sequences must be fulfilled.

Amplification primers are generally single stranded polynucleotide structures. They may also contain a mixture of natural and non-natural bases and also natural and non-natural backbone linkages, provided that any non-natural modifications do not preclude function as a primer - that being defined as the ability to anneal to a template polynucleotide strand during conditions of the amplification reaction and to act as an initiation point for synthesis of a new polynucleotide strand complementary to the template strand. Primers may additionally comprise non-nucleotide chemical modifications, again provided such that modifications do not prevent primer function. Chemical modifications may, for example, facilitate covalent attachment of the primer to a solid support. Certain chemical modifications may themselves improve the function of the molecule as a primer, or may provide some other useful functionality, such as providing a site for cleavage to enable the primer (or an extended polynucleotide strand derived therefrom) to be cleaved from a solid support.

It is desired that the template for amplification will be the combined ligated polynucleotide sequences, namely the target sequences with adapter sequences Y and Z attached to the ends. More particularly the at least two different primer oligonucleotides will comprise sequences which are complementary to a part of the linearised vector polynucleotide sequence portion of the combined ligated polynucleotide sequences. To amplify a single stranded template structured 5′-Y-target-Z-3′, amplification primers comprising sequences X and Y may be used, where X is complementary to Z, and can hybridise to the 3′ end of the template. Primer X can be extended in a first amplification

cycle, and denatured to give a single stranded copy of the original template. Primer Y can hybridise to this single stranded copy, and a further primer X can hybridise to the original Y-target-Z construct. Cycles of amplification and denaturing thus give a universal amplification of all targets. References to 'a part of the linearised vector polynucleotide sequence portion' are intended to mean a part of the linearised vector polynucleotide sequence, denoted Y or Z, of which the sequence is known and which is suitable for use as a primer binding site. In the combined ligated sequences such primer binding sites should flank the target polynucleotide duplex(es). Under conditions suitable for amplification, primers annealed to such primer binding sites will be extended by polymerase activity to produce complementary copies of the target polynucleotide duplex (see below).

In a next step of the method of the invention an amplification reaction is performed in which said at least two different primer oligonucleotides comprising sequences X and Y are annealed to complementary parts of the linearised polynucleotide sequence portion of the combined ligated polynucleotide sequences and extended by sequential addition of nucleotides to generate amplification products complementary to at least one strand of the combined ligated polynucleotide sequences and wherein said amplification products have common sequences at their 5′ ends and common sequences at their 3′ ends and collectively provide a 5′ and 3′ modified library of template polynucleotide molecules.

In this context the term "common" is interpreted as meaning common to all templates in the 5′ and 3′ modified

library. As explained in further detail below, all templates within the 5′ and 3′ modified library will contain regions of common sequence at (or proximal to) their 5' and 3' ends, wherein the common sequence at the 5' end of each individual template in the 5′ and 3′ modified library is not identical and not fully complementary to the common sequence at the 3' end of said template. The term "5′ and 3′ modified library" merely refers to a collection or plurality of template molecules which share common sequences at their 5' ends and common sequences at their 3' ends. Use of the term "5′ and 3′ modified library" to refer to a collection or plurality of template molecules should not be taken to imply that the templates making up the 5′ and 3′ modified library are derived from a particular source, or that the "5′ and 3′ modified library" has a particular composition. By way of example, use of the term "5′ and 3′ modified library" should not be taken to imply that the individual templates within the 5′ and 3′ modified library must be of different nucleotide sequence or that the templates be related in terms of sequence and/or source.

In its various embodiments the invention encompasses use of so-called "mono-template" libraries, which comprise multiple copies of a single type of template molecule, each having common sequences at their 5' ends and their 3' ends, as well as "complex" libraries wherein many, if not all, of the individual template molecules, when viewed as a single strand comprise different target sequences (as defined below), although all share common sequences Y and Z at their 5' ends and 3' ends. Such complex template libraries may be prepared from a complex mixture of target polynucleotides such as (but not limited to) random genomic DNA fragments,

cDNA libraries etc. The invention may also be used to amplify "complex" libraries formed by mixing together several individual "mono-template" libraries, each of which has been prepared separately starting from a single type of target molecule (i.e. a mono-template). In certain embodiments more than 50%, or more than 60%, or more than 70%, or more than 80%, or more than 90%, or more than 95% of the individual polynucleotide templates in a complex 5′ and 3′ modified library may comprise different target sequences, although all templates in a given 5′ and 3′ modified library will share common sequence Y at their 5' ends and common sequence Z at their 3' ends.

The conditions encountered during an amplification reaction will generally be known to one skilled in the art (see Sambrook et al., 2001, Molecular Cloning A laboratory Manual, 3rd Ed, Cold Spring Harbor Laboratory Press). During the amplification reaction the two different primers act as starting points for initiation of polymerization mediated by an enzyme with polymerase activity (e.g. Taq). In a particular embodiment, the primers anneal to two different sites of the combined ligated polynucleotide sequence. Yet more particularly, the primers anneal to primer binding sites (Z and Y′) on either side of the target polynucleotide duplex portion of the combined ligated polynucleotide sequences such that the products of the amplification reaction (5′ and 3′ modified library of template polynucleotide molecules) will be of the structure: 5′[common sequence I]-[target polynucleotide duplex sequence]-[common sequence II]-3′.

Typically "common sequence I" (containing sequence Y) and "common sequence II" (containing sequence Z) will consist of more than 20, or more than 40, or more than 50, or more than 100, or more than 300 consecutive nucleotides. The precise length of the two sequences may or may not be identical. The nucleotide sequences of "common sequence I" and "common sequence II" in the 5′ and 3′ modified library of template polynucleotide molecules will be determined in part by the sequences of the linearised vector polynucleotide sequence Y and Z and in part by the sequence(s) of the at least two different primer oligonucleotides used in the amplification reaction. The primers comprising sequences X and Y may contain additional bases than just those with hybridise to sequences Y and Z, and therefore extra bases may be copied into the template in addition to just bases Y and Z from the linearised vector.

Following amplification, the 5′ and 3′ modified library of template polynucleotide molecules can be purified using methods well known in the art. Such methods utilise by way of non-limiting example, size inclusion chromatography such as agarose gel electrophoresis or commercial kits comprising silica-gel-membrane assemblies (Qiagen purification kits).

Use of the template 5′ and 3′ modified library

Template libraries prepared according to the method of the invention may be used in essentially any method of nucleic acid analysis which requires further amplification of the templates and/or sequencing of the templates or amplification products thereof. Exemplary uses of the template libraries include, but are not limited to, providing templates for bridging amplification or other

forms of solid-phase amplification. In a particular embodiment, the use is in solid-phase amplification carried out on either a planar solid support or on a pool of beads or microparticles.

Whole-genome amplification

Template libraries prepared according to the method of the invention starting from a complex mixture of genomic DNA fragments representing a whole or substantially whole genome provide suitable templates for so-called "whole-genome" amplification. The term "whole-genome amplification" refers to a nucleic acid amplification reaction (e.g. PCR) in which the template to be amplified comprises a complex mixture of nucleic acid fragments representative of a whole (or substantially whole genome)

Solid-phase amplification

Once formed, the 5′ and 3′ modified library of templates prepared according to the methods described above can be used for solid-phase nucleic acid amplification.

Thus, in further aspects the invention provides a method of solid-phase nucleic acid amplification of template polynucleotide molecules which comprises preparing a 5′ and 3′ modified library of template polynucleotide molecules which have common sequences at their 5' and 3' ends using a method according to the first aspect of the invention described herein and carrying out a solid-phase nucleic acid amplification reaction wherein said template polynucleotide molecules are amplified.

Sequencing can be performed as an array of single molecules, or can be amplified prior to sequencing. The amplification can be carried out using one or more immobilised primers. The immobilised primer(s) can be a lawn on a planar surface, or on a pool of beads. The pool of beads can be isolated into an emulsion with a single bead in each 'compartment' of the emulsion. At a concentration of only one template per 'compartment', only a single template is amplified on each bead.

The term 'solid-phase amplification' as used herein refers to any nucleic acid amplification reaction carried out on or in association with a solid support such that all or a portion of the amplified products are immobilised on the solid support as they are formed. In particular, the term encompasses solid-phase polymerase chain reaction (solid-phase PCR) and solid phase isothermal amplification which are reactions analogous to standard solution phase amplification, except that one or both of the forward and reverse amplification primers is/are immobilised on the solid support. Solid phase PCR covers systems such as emulsions, wherein one primer is anchored to a bead and the other is in free solution, and colony formation in solid phase gel matrices wherein one primer is anchored to the surface, and one is in free solution.

Although the invention encompasses 'solid-phase' amplification methods in which only one amplification primer is immobilised (the other primer usually being present in free solution), the solid support may also be provided with both the forward and the reverse primers immobilised. In practice, there will be a 'plurality' of identical forward

primers and/or a 'plurality' of identical reverse primers immobilised on the solid support, since the amplification process requires an excess of primers to sustain amplification. References herein to forward and reverse primers are to be interpreted accordingly as encompassing a 'plurality' of such primers unless the context indicates otherwise.

As will be appreciated by the skilled reader, any given amplification reaction requires at least one type of forward primer and at least one type of reverse primer specific for the template to be amplified. However, in certain embodiments the forward and reverse primers may comprise template-specific portions of identical sequence, and may have entirely identical nucleotide sequence and structure (including any non-nucleotide modifications). In other words, it is possible to carry out solid-phase amplification using only one type of primer, and such single-primer methods are encompassed within the scope of the invention. Other embodiments may use forward and reverse primers which contain identical template-specific sequences but which differ in some other structural features. For example one type of primer may contain a non-nucleotide modification which is not present in the other.

In all embodiments of the invention, primers for solid-phase amplification may be immobilised by single point covalent attachment to the solid support at or near the 5' end of the primer, leaving the template-specific portion of the primer free to anneal to its cognate template and the 3' hydroxyl group free for primer extension. Any suitable covalent attachment means known in the art may be used for this purpose. The chosen attachment chemistry will depend on

the nature of the solid support, and any derivatisation or functionalisation applied to it. The primer itself may include a moiety, which may be a non-nucleotide chemical modification, to facilitate attachment. In a particular embodiment, the primer may include a sulphur-containing nucleophile, such as phosphorothioate or thiophosphate, at the 5' end. In the case of solid-supported polyacrylamide hydrogels (as described below), this nucleophile will bind to a bromoacetamide group present in the hydrogel. A more particular means of attaching primers and templates to a solid support is via 5' phosphorothioate attachment to a hydrogel comprised of polymerised acrylamide and *N*-(5-bromoacetamidylpentyl) acrylamide (BRAPA), as described fully in WO05065814, whose contents are incorporated herein by reference.

Certain embodiments of the invention may make use of solid supports comprised of an inert substrate or matrix (e.g. glass slides, polymer beads, etc) which has been "functionalised", for example by application of a layer or coating of an intermediate material comprising reactive groups which permit covalent attachment to biomolecules, such as polynucleotides. Examples of such supports include, but are not limited to, polyacrylamide hydrogels supported on an inert substrate such as glass. In such embodiments, the biomolecules (e.g. polynucleotides) may be directly covalently attached to the intermediate material (e.g. the hydrogel), but the intermediate material may itself be non-covalently attached to the substrate or matrix (e.g. the glass substrate). The term "covalent attachment to a solid support" is to be interpreted accordingly as encompassing this type of arrangement.

The 5′ and 3′ modified library may be amplified on beads wherein each bead contains a forward and reverse amplification primer. In a particular embodiment, the library of templates is used to prepare clustered arrays of nucleic acid colonies, analogous to those described in WO 00/18957 and WO 98/44151, by solid-phase amplification and more particularly solid phase isothermal amplification. The terms 'cluster' and 'colony' are used interchangeably herein to refer to a discrete site on a solid support comprised of a plurality of identical immobilised nucleic acid strands and a plurality of identical immobilised complementary nucleic acid strands. The term 'clustered array' refers to an array formed from such clusters or colonies. In this context the term 'array' is not to be understood as requiring an ordered arrangement of clusters.

The term solid phase, or surface, is used to mean either a planar array wherein primers are attached to a flat surface, for example, glass, silica or plastic microscope slides or similar flow cell devices; beads, wherein either one or two primers are attached to the beads and the beads are amplified; or an array of beads on a surface.

Use in sequencing/methods of sequencing

The invention also encompasses methods of sequencing amplified nucleic acids generated by whole genome or solid-phase amplification. Thus, the invention provides a method of nucleic acid sequencing comprising amplifying a 5′ and 3′ modified library of nucleic acid templates using whole genome or solid-phase amplification as described above and

carrying out a nucleic acid sequencing reaction to determine the sequence of the whole or a part of at least one amplified nucleic acid strand produced in the whole genome or solid-phase amplification reaction.

Sequencing can be carried out using any suitable sequencing technique, wherein nucleotides are added successively to a free 3' hydroxyl group, resulting in synthesis of a polynucleotide chain in the 5' to 3' direction. The nature of the nucleotide added may be determined after each nucleotide addition. Sequencing techniques using sequencing by ligation, wherein not every contiguous base is sequenced, and techniques such as massively parallel signature sequencing (MPSS) where bases are removed from, rather than added to the strands on the surface are also within the scope of the invention, as are techniques using detection of pyrophosphate release (pyrosequencing). Such pyrosequencing based techniques are particularly applicable to sequencing arrays of beads where the beads have been amplified in an emulsion such that a single template from the library molecule is amplified on each bead.

The initiation point for the sequencing reaction may be provided by annealing of a sequencing primer to a product of the whole genome or solid-phase amplification reaction. In this connection, one or both of the adapters added during formation of the template 5′ and 3′ modified library may include a nucleotide sequence which permits annealing of a sequencing primer to amplified products derived by whole genome or solid-phase amplification of the template 5′ and 3′ modified library.

The products of solid-phase amplification reactions wherein both forward and reverse amplification primers are covalently immobilised on the solid surface are so-called "bridged" structures formed by annealing of pairs of immobilised polynucleotide strands and immobilised complementary strands, both strands being attached to the solid support at the 5' end. Arrays comprised of such bridged structures may provide inefficient templates for nucleic acid sequencing, since hybridisation of a conventional sequencing primer to one of the immobilised strands is not favoured compared to annealing of this strand to its immobilised complementary strand under standard conditions for hybridisation.

In order to provide more suitable templates for nucleic acid sequencing, substantially all, or at least a portion of, one of the immobilised strands in the "bridged" structure may be removed in order to generate a template which is at least partially single-stranded. The portion of the template which is single-stranded will thus be available for hybridisation to a sequencing primer. The process of removing all or a portion of one immobilised strand in a "bridged" double-stranded nucleic acid structure may be referred to herein as "linearisation".

Bridged template structures may be linearised by cleavage of one or both strands with a restriction endonuclease or by cleavage of one strand with a nicking endonuclease. Other methods of cleavage can be used as an alternative to restriction enzymes or nicking enzymes, including *inter alia* chemical cleavage (e.g. cleavage of a

diol linkage with periodate), cleavage of abasic sites by cleavage with endonuclease, or by exposure to heat or alkali, cleavage of ribonucleotides incorporated into amplification products otherwise comprised of deoxyribonucleotides, photochemical cleavage or cleavage of a peptide linker. Methods of linearization are detailed in co-pending application WO07010251, whose contents are included herein by reference.

It will be appreciated that a linearization step may not be essential if the solid-phase amplification reaction is performed with only one primer covalently immobilised and the other in free solution.

In order to generate a linearised template suitable for sequencing it is necessary to remove the cleaved complementary strands in the bridged structure that remain hybridised to the uncleaved strand. This denaturing step is a part of the 'linearisation process', and can be carried out by standard techniques such as heat or chemical treatment with hydroxide or formamide solution. In a particular embodiment, one strand of the bridged structure is substantially or completely removed by the process of chemical cleavage and denaturation. Denaturation results in the production of a sequencing template which is partially or substantially single-stranded. A sequencing reaction may then be initiated by hybridisation of a sequencing primer to the single-stranded portion of the template.

Thus, the invention encompasses methods wherein the nucleic acid sequencing reaction comprises hybridising a sequencing primer to a single-stranded region of a linearised amplification product, sequentially incorporating

one or more nucleotides into a polynucleotide strand complementary to the region of amplified template strand to be sequenced, identifying the base present in one or more of the incorporated nucleotide(s), or one or more of the bases present in the oligonucleotides, and thereby determining the sequence of a region of the template strand.

One particular sequencing method which can be used in accordance with the invention relies on the use of modified nucleotides having removable 3′ blocks, for example as described in WO04018497 and US7057026. Once the modified nucleotide has been incorporated into the growing polynucleotide chain complementary to the region of the template being sequenced there is no free 3'-OH group available to direct further sequence extension and therefore the polymerase can not add further nucleotides. Once the nature of the base incorporated into the growing chain has been determined, the 3' block may be removed to allow addition of the next successive nucleotide. By ordering the products derived using these modified nucleotides it is possible to deduce the DNA sequence of the DNA template. Such reactions can be done in a single experiment if each of the modified nucleotides has attached thereto a different label, known to correspond to the particular base, to facilitate discrimination among the bases added at each incorporation step. Alternatively, a separate reaction may be carried out containing each of the modified nucleotides separately.

The modified nucleotides may carry a label to facilitate their detection. Particularly this is a fluorescent label. Each nucleotide type may carry a

different fluorescent label. However the detectable label need not be a fluorescent label. Any label can be used which allows the detection of an incorporated nucleotide.

One method for detecting fluorescently labelled nucleotides comprises using laser light of a wavelength specific for the labelled nucleotides, or the use of other suitable sources of illumination. The fluorescence from the label on the nucleotide may be detected by a CCD camera or other suitable detection means.

The invention is not intended to be limited to use of the sequencing method outlined above, as essentially any sequencing methodology which relies on successive incorporation of nucleotides into a polynucleotide chain can be used. Suitable alternative techniques include, for example, Pyrosequencing™, FISSEQ (fluorescent in situ sequencing), MPSS (massively parallel signature sequencing) and sequencing by ligation-based methods, for example as described in US6306597.

The target polynucleotide to be sequenced using the method of the invention may be any polynucleotide that it is desired to sequence. Using the template 5′ and 3′ modified library preparation method described in detail herein it is possible to prepare template libraries starting from essentially any double or single-stranded target polynucleotide of known, unknown or partially known sequence. With the use of clustered arrays prepared by solid-phase amplification it is possible to sequence multiple targets of the same or different sequence in parallel. Sequencing may result in determination of the sequence of a whole or a part of the target molecule.

Example of the Preparation of a 5′ and 3′ modified library using restriction enzymes to reduce sequences without the target present

Experimental overview

The following experimental details describe the exposition of one embodiment of the invention as described above.

1) **Preparation of blunt ended genomic DNA fragments**

(a) Sonication of genomic DNA (Figure 3a)

100 µl of 50 ng/µl PhiX174 RF1 DNA (NEB, N3021L) diluted in 10 mM Tris, 0.1 mM EDTA pH 8.0 was sonicated using a VirTis Virsonic 600 sonicator at power 1.0 for 30 secs on ice.

(b)Nebulisation of genomic DNA (Figure 3b)

Two alternate methods to generate different sized fragments:

i) <1500 bp fragments. 10 µl (10 µg) of PhiX174 RF1 DNA (NEB, N3021L) was diluted in 740 µl of shearing buffer (10 mM Tris, 1 mM EDTA, 10 % glycerol, pH 8.0) and fragmented using a disposable nebuliser (Invitrogen, 45-0072) using an argon cylinder set to 30 psi (2 bar) for 4 mins on ice, in the fume hood. The nebuliser was centrifuged at 1000 rpm for 1 min to recover the fragmented DNA.

ii) <1000bp fragments. 10 µl (10 µg) of PhiX174 RF1 DNA (NEB, N3021L) was diluted in 740 µl of nebuliser buffer (37 mM Tris, 5.5 mM EDTA, 53 % glycerol, pH 7.5) and fragmented using a disposable nebuliser (Invitrogen, 45-0072) using an argon cylinder set to 32 psi (2.2 bar) for 5 mins on ice, in the fume

hood. The nebuliser was centrifuged at 1000 rpm for 1 min to recover the fragmented DNA.

(c) End repair and purification of nebulised/sonicated genomic DNA

Sonicated/nebulised DNA was precipitated using Novagen pellet paint NF co-precipitant (70748-4) according to the manufacturer's instructions (general precipitation of nucleic acids method). The pellet was resuspended in 38 µl of 10 mM Tris HCl pH 8.5, per 10 µg of fragmented DNA. The DNA was end repaired using the Lucigen DNA terminator end repair kit (40035-2) using 10 µg of fragmented DNA, 2 µl of end repair enzymes in 1X end repair buffer (50 ul total volume) for 30 mins at RT. The reaction was heat inactivated at 70°C for 15 mins. The end repaired DNA was purified using a Qiagen PCR purification kit column (28106) to purify all the end repaired fragments, or by gel extraction on a 2 % TAE agarose gel using a Qiagen gel extraction kit (28706), to purify products of specific size (usually 400-700 bp in size). The Qiagen columns were used according to manufacturers instructions and the DNA eluted using 30 µl of 10 mM Tris HCl pH 8.5. The samples were stored at -20°C.

2) **Preparation of linearised vector DNA**

10 µg of KH1 vector (figure 4) was digested with 40 U of EcoRV in 1X NEB buffer 3, 100 µg/ml BSA (NEB, R0195L) (100 µl total volume) for 2 hours at 37°C. The digest was heat inactivated at 80 °C for 20 mins. The linearised fragment (3.1 kb) was gel purified on a 1 % TAE agarose gel using a Qiagen gel extraction kit (28706) according to the manufacturer's instructions. The purified DNA was eluted

from the Qiagen column with 50 µl of 10 mM Tris HCl pH 8.5 and stored at -20°C.

3) **Ligation** (figure 5)

Ligation can be done in the presence or absence of EcoRV

a) Ligation without EcoRV

15 ng of linearised KH1 (from step 2), 120 ng of blunt ended genomic DNA (from step 1), 0.5 µl of Quick ligase in 1X quick ligase buffer (NEB, M2200L) (total volume of 10 µl) was incubated for 15 mins at RT.

b) Ligation in the presence of EcoRV

15 ng of linearised KH1 (from step 2), 120 ng of blunt ended genomic DNA (from step 1), 2000 U of T4 DNA ligase (NEB, M0202M), 10 U EcoRV (NEB, R0195L) in 1X ligase buffer (NEB, M0202M) (total volume of 20 µl) was incubated overnight at RT.

4) **Purification of ligation**

The ligation (from step 3) was purified using a Qiagen PCR purification kit (28106) according to the manufacturer's instructions. The DNA was eluted from the column with 30 µl of 10 mM Tris HCl pH 8.5.

5) **EcoRV digest of ligation**

The purified ligation (from step 4) was digested with 50 U EcoRV in 1X NEB buffer 3, 100 µg/ml BSA (NEB, R0195L) for 2.5 hours at 37°C, to linearise any religated vector,

containing no insert. The digest was heat inactivated at 80°C for 20 mins.

6) **Purification of EcoRV digest**

The digest (from step 5) was purified using a Qiagen PCR purification kit (28106), according to the manufacturer's instructions. The DNA was eluted from the column with 30 µl of 10 mM Tris HCl pH 8.5.

7) **Amplification of Genomic Fragments from the Ligation** (Figure 6)

The genomic fragments were amplified from the purified digested ligation, with the appropriate P5/sequencing primer and P7 adaptors on each end. The ligation was added to REDTaq Ready mix (Sigma, R2523) (1x final concentration), the P5 primer (AATGATACGGCGACCACCGA SEQ ID NO:1) and P7 primer (CAAGCAGAAGACGGCATACGA SEQ ID NO:2) (both at 1 µM final concentration).

PCR Programme

94°C 30 secs
65°C 30 secs
72°C 1min/kb
} 20 cycles

72°C 10 mins

Reactions were cooled to 4°C until required.

8) **Gel purification of genomic fragments**

PCR products from any religated KH1 (that were not linearised by the EcoRV digest) were 88 bp in size. Correct sized fragments (>200 bp) were gel purified from a 2 % TAE

agarose gel using a Qiagen gel extraction kit (28706), according to manufacturers instructions. The DNA was eluted from each Qiagen column with 30 µl of 10 mM Tris HCl pH 8.5.

9) **Validation of the Genomic 5′ and 3′ modified library by Conventional Dideoxy Sequencing**

3.6 µl of the purified fragments from step 8 were removed in order to verify the DNA fragments by conventional dideoxy sequencing (by Lark). The fragments were cloned into pGEM-T Easy vector from Promega (A1360), according to the manufacturer's instructions. The pGEM-T Easy ligations were transformed into XL10 Gold ultracompetent cells (Stratagene, 200315), according to the manufacturer's instructions using blue-white selection. 22 white colonies from each 5′ and 3′ modified library were screened by PCR using M13F and M13R primers. The colonies were picked and each added directly to 25 µl of PCR solution (2 µM M13 forward primer (GTAAAACGACGGCCAG SEQ ID NO:3), 2 µM M13 reverse primer (CAGGAAACAGCTATGAC SEQ ID NO:4) in 1X REDTaq Ready mix (Sigma, R2523)).

PCR Programme

95°C 2 mins

95°C 50 secs
50°C 30 secs
72°C 1 min/kb
} 30 cycles

72°C 5 mins

Reactions were cooled to 4°C until required.

5 µl of each PCR were separated on a 2 % TAE agarose gel. The PCR product from the religated pGEM-T Easy vector was

220bp in size. The PCR product from pGEM-T Easy containing only the P5/P7 adaptors (PCR product from religated KH1) was 308 bp in size. PCR products from pGEM-T Easy containing genomic DNA flanked by adaptors (correct products) were >308 bp. 22/22 had PCR products of correct size (>308 bp). Eight correctly sized PCR products were purified from the remaining PCR using a Qiagen PCR purification kit (28106), according to the manufacturer's instructions. The DNA was eluted from the column using 30 µl of 10 mM Tris HCl pH 8.5 and sent for conventional dideoxy sequencing at Lark Technologies using their T7 promoter primer. 7/8 clones contained Phix174 DNA flanked by correct P5/P7 adaptors. 3 clones had point mutations in the Phix174 DNA but Taq probably introduced these during the PCR. By using a high fidelity polymerase, this problem should be resolved. 1/8 contained E. coli DNA flanked by the correct adaptors. This problem is caused by contamination of the NEB Phix174 DNA with E. coli DNA and is not a problem with the sample prep method.

10) **Validation of the Efficiency of the Ligation and EcoRV Digest**

1 µl ($1/30^{th}$) of the purified, digested ligation (from step 6) was transformed into 50 µl of XL10 Gold ultracompetent cells (Stratagene, 200315), according to manufacturers instructions. $1/10^{th}$ of the transformation was plated onto LB agar plates supplemented with 100 µg/ml carbenicillin and incubated overnight at 37 °C. The following morning the number of transformants were counted (92 colonies) and the number of transformants were calculated for the total ligation ($2.76x10^4$ transformants) by multiplying the number of colonies by 300 (30 x 10). This number will be the

minimum size of the library because the transformation efficiency of the cells will not be 100 %. A PCR screen was used to determine the number of transformants that contain genomic DNA inserts.

20 isolated colonies were picked and each added directly to 25 µl of PCR solution (2 µM M13 forward primer (GTAAAACGACGGCCAG SEQ ID NO:3), 2 µM M13 reverse primer (CAGGAAACAGCTATGAC SEQ ID NO:4) in 1X Sigma ReadyTaq or JumpStart). Control: 1 µl of 20 ng/ul KH1 template.

PCR Programme (30 cycles)

95 °C 2 mins

95 °C 50 secs
50 °C 30 secs
72 °C 1 min/kb } 30 cycles
72 °C 5 mins

Reactions were cooled to 4°C until required.

5 µl of each reaction was run on a 2 % TAE agarose gel and the size of the PCR products were determined from the Invitrogen 1 kb plus DNA ladder (10787-018) that was loaded. Religated KH1 containing no insert gave a PCR product of 322 bp in size. Any clones containing genomic DNA have a PCR product >322 bp in size. The proportion of clones containing genomic inserts was determined (9/20) and from this the number of clones containing genomic inserts in the whole library was calculated (1.2×10^4 correct clones). In the case where no EcoRV step was performed (steps 3a without step 5),

none of the clones picked show a target insert, every single one had only the 322 base pair KH1 vector (Figure 7).

Samples prepared according to the methods described herein can be used for surface amplification and sequencing according to the protocols detailed in copending applications WO06064199 or WO07010251, the protocols and contents of which are incorporated herein by reference.

While certain of the embodiments of the present invention have been described and specifically exemplified above, it is not intended that the invention be limited to such embodiments. Various modifications may be made thereto without departing from the scope and spirit of the present invention, as set forth in the following claims.

Claims:

1. A method of generating a 5′ and 3′ modified library of template polynucleotide molecules from one or more primary polynucleotide molecules characterised in that said method is carried out entirely *in vitro*, comprising:

(a) fragmenting said one or more primary polynucleotide molecules to produce target polynucleotide duplexes;

(b) ligating a vector polynucleotide sequence to the target polynucleotide duplexes to form combined ligated polynucleotide sequences under conditions biased towards formation of the combined ligated polynucleotide sequences that contain the target duplexes;

(c) preparing an amplification reaction comprising said combined ligated polynucleotide sequences and at least two different primer oligonucleotides wherein said two different primer oligonucleotides comprise sequences complementary to regions of the vector polynucleotide sequence portion of the combined ligated polynucleotide sequences, said regions being located on either side of the inserted target; and

(d) carrying out an amplification reaction in which one of said at least two different primer oligonucleotides is annealed to complementary parts of each strand of the vector polynucleotide sequence portion of the combined ligated polynucleotide sequences and extended by sequential addition of nucleotides to generate amplification products complementary to the combined ligated polynucleotide sequences and wherein said amplification products comprise a

common sequence Y at their 5′ ends and common sequence Z at their 3′ ends and collectively provide a 5′ and 3′ modified library of template polynucleotide molecules, wherein the templates comprise the targets plus additional common sequences Y and Z.

2. The method of claim 1, wherein fragmentation of the primary polynucleotide molecules is by mechanical fragmentation.

3. The method of claim 2, wherein said mechanical fragmentation of the primary polynucleotide molecules is by sonication, nebulisation or hydrodynamic shearing.

4. The method of claim 1, wherein fragmentation of the primary polynucleotide molecules is by chemical or enzymatic fragmentation.

5. The method of claims 1 to 4, wherein the one or more primary polynucleotide molecules are DNA molecules.

6. The method of claim 5, wherein the one or more primary polynucleotide molecules are genomic DNA molecules.

7. The method of claim 6, wherein the one or more primary polynucleotide molecules are human genomic DNA molecules.

8. The method of claim 1, wherein the vector polynucleotide sequence is prepared from a circular DNA molecule.

9. The method of claim 8, wherein said vector polynucleotide sequence is prepared by cleavage of the

circular DNA molecule with at least one restriction endonuclease enzyme.

10. The method of claim 9, wherein said at least one restriction endonuclease enzyme cleaves the circular DNA molecule to generate a vector polynucleotide sequence with blunt ends.

11. The method of claim 1, wherein said vector polynucleotide sequence is prepared by amplification.

12. The method of any preceding claim, wherein step (b) is performed in the presence of at least one restriction endonuclease enzyme.

13. The method of any preceding claim, wherein two restriction endonuclease enzymes are used.

14. The method of any preceding claim further comprising the steps of:

(b)(ii) optionally purifying the combined ligated sequences, vector polynucleotide sequence and target polynucleotide duplexes from the ligation reaction of step (b)(i); and

(b)(iii) optionally performing a restriction digest reaction on the products of step (b)(ii) in the presence of at least one restriction endonuclease enzyme and optionally purifying the combined ligated sequences, vector polynucleotide sequence and target polynucleotide duplexes from the restriction digest reaction.

15. The method of claims 9, 10, 12, 13 or 14, wherein the at least one restriction endonuclease enzyme used in each step is the same enzyme.

16. The method of claim 15, wherein said at least one restriction endonuclease enzyme used in each step is a rare blunt cutter.

17. The method of claim 1, wherein the vector polynucleotide sequence has a single nucleotide overhang at either the 5′ or 3′ terminus of each duplex strand and the target polynucleotide duplexes have complementary single nucleotide overhangs at the 5′ or 3′ terminus of each duplex strand.

18. The method of claim 17, wherein the vector polynucleotide sequence has a single nucleotide T overhang at the 3′ terminus of each duplex strand and the target polynucleotide duplexes have complementary single nucleotide 'A' overhangs at the 3′ terminus of each duplex strand.

19. The method of claim 17 or 18, wherein the single nucleotide overhang is prepared using Taq or Klenow exo minus polymerase.

20. The method of any preceding claim further comprising a step of obtaining a sequence read from one or more template polynucleotides.

21. A 5′ and 3′ modified library of template polynucleotide molecules prepared by the method of any one of claims 1 to 19.

22. An array comprising the 5′ and 3′ modified library of template polynucleotide molecules of claim 21.

23. Use of the 5′ and 3′ modified library of claim 21 or the array of claim 22 in sequencing.

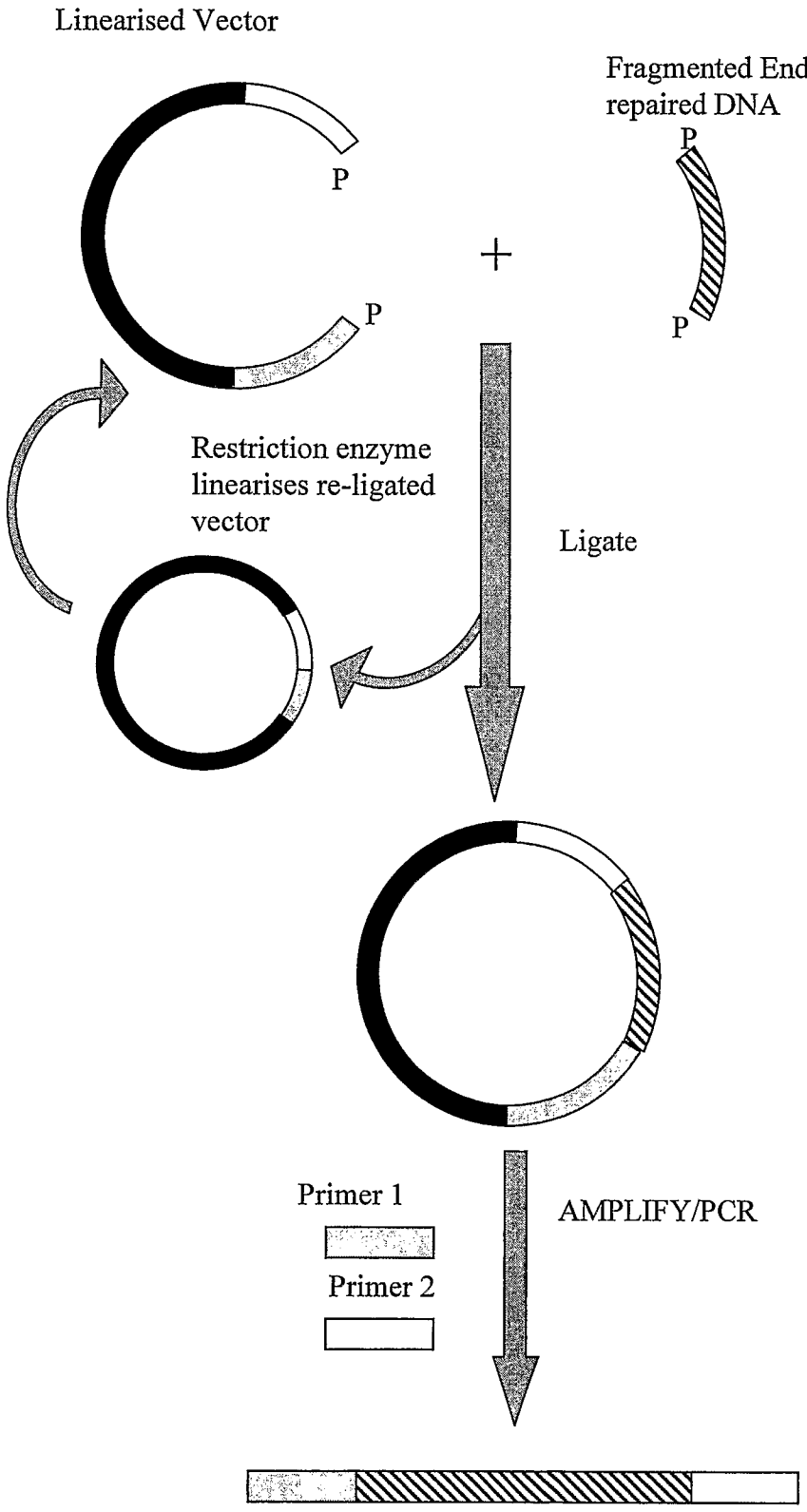

Linearised Vector
Fragmented End-repaired DNA
P
P
P
P
+
Restriction enzyme linearises re-ligated vector
Ligate
Primer 1
Primer 2
AMPLIFY/PCR

Figure 1

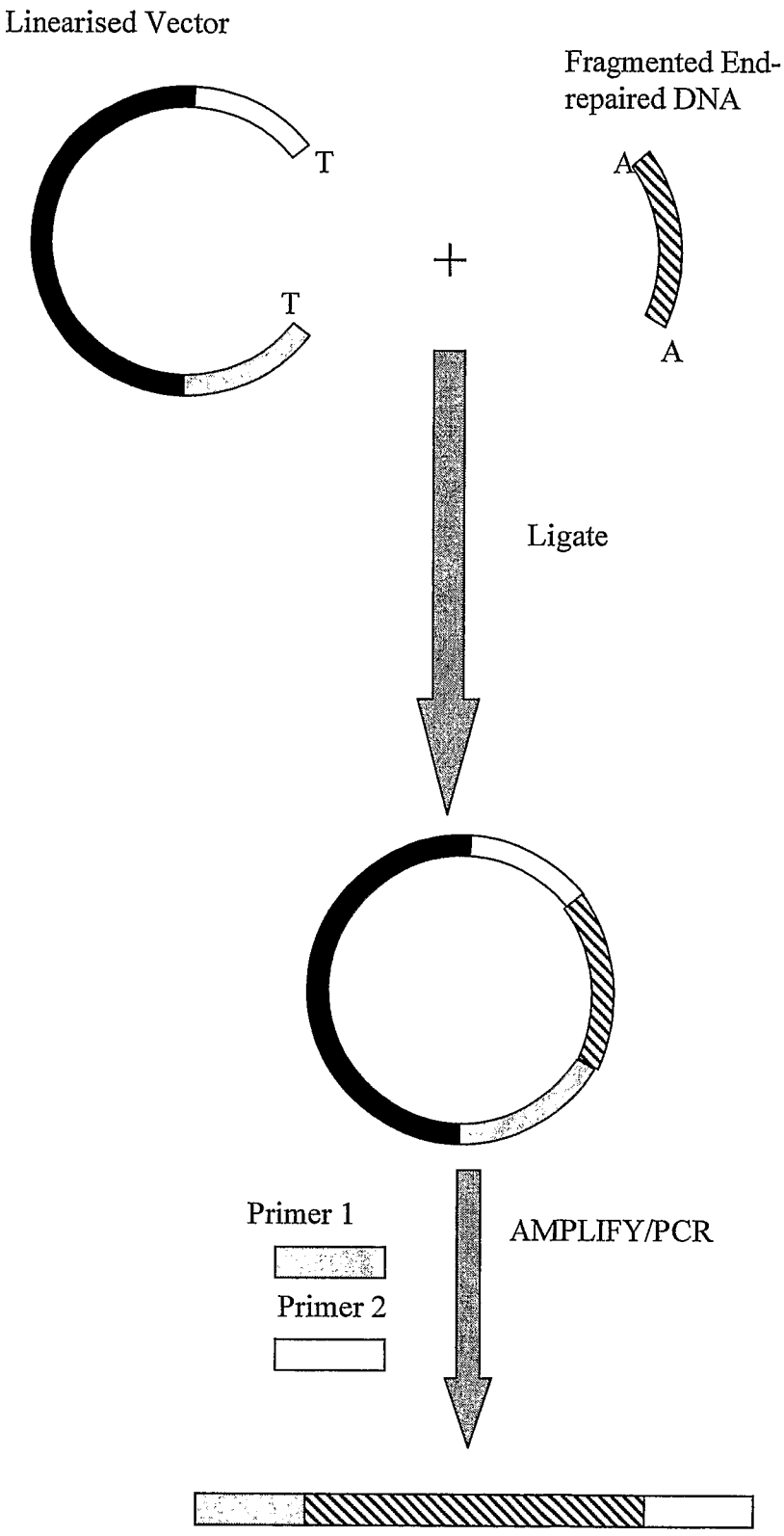
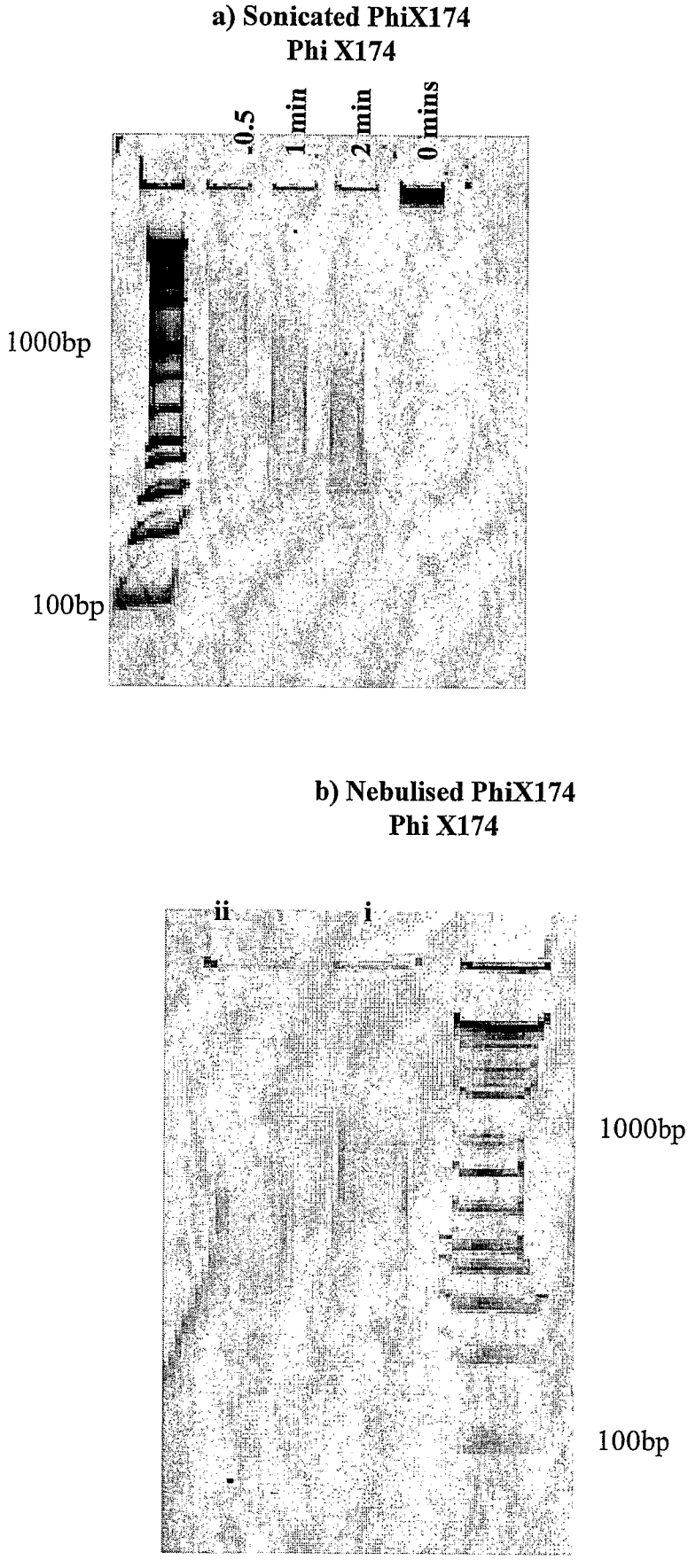

Linearised Vector
Fragmented End-repaired DNA
T
T
A
A
Ligate
Primer 1
Primer 2
AMPLIFY/PCR

Figure 2

Figure 3


a) Sonicated PhiX174
Phi X174
0.5
1 min
2 min
0 mins
1000bp
100bp
b) Nebulised PhiX174
Phi X174
ii
i
1000bp
100bp

1>>GGGCGAATTGGGCCCGACGTCGCATGCTCCCGGCCGCCATGGCGGCCGCGGGAATTCGATTCA
AGCAGAAGACGGCATACGAGATATCCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGAT
CTCGGTGGTCGCCGTATCATTAATCACTAGTGAATTCGCGGCCGCCTGCAGGTCGACCATATGGGA
GAGCTCCCAACGCGTTGGATGCATAGCTTGAGTATTCTATAGTGTCACCTAAATAGCTTGGCGTAA
TCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCG
GAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCT
CACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGG
GGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGT
TCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGG
ATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGC
GTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCA
GAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGC
GCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGC
GCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGT
GTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACC
CGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTAT
GTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTT
GGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAA
CAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGA
TCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAA
GGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGT
TTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAG
GCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAA
CTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCAC
CGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCA
ACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTA
ATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGC
TTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGC
GGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTT
ATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGT
ACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATAC
GGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGC
GAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACT
GATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCG
CAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATT
GAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAAC
AAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGATGCGGTGTGAAATACCGCACAG
ATGCGTAAGGAGAAAATACCGCATCAGGAAATTGTAAGCGTTAATATTTTGTTAAAATTCGCGTTA
AATTTTTGTTAAATCAGCTCATTTTTTAACCAATAGGCCGAAATCGGCAAAATCCCTTATAAATCAA
AAGAATAGACCGAGATAGGGTTGAGTGTTGTTCCAGTTTGGAACAAGAGTCCACTATTAAAGAAC
GTGGACTCCAACGTCAAAGGGCGAAAAACCGTCTATCAGGGCGATGGCCCACTACGTGAACCATC
ACCCTAATCAAGTTTTTTGGGGTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCC
CCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAAAGCGAAA
GGAGCGGGCGCTAGGGCGCTGGCAAGTGTAGCGGTCACGCTGCGCGTAACCACCACACCCGCCGC
GCTTAATGCGCCGCTACAGGGCGCGTCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGA
TCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGT
TGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATTGTAATACGAC
TCACTATA>>3105>

Figure 4

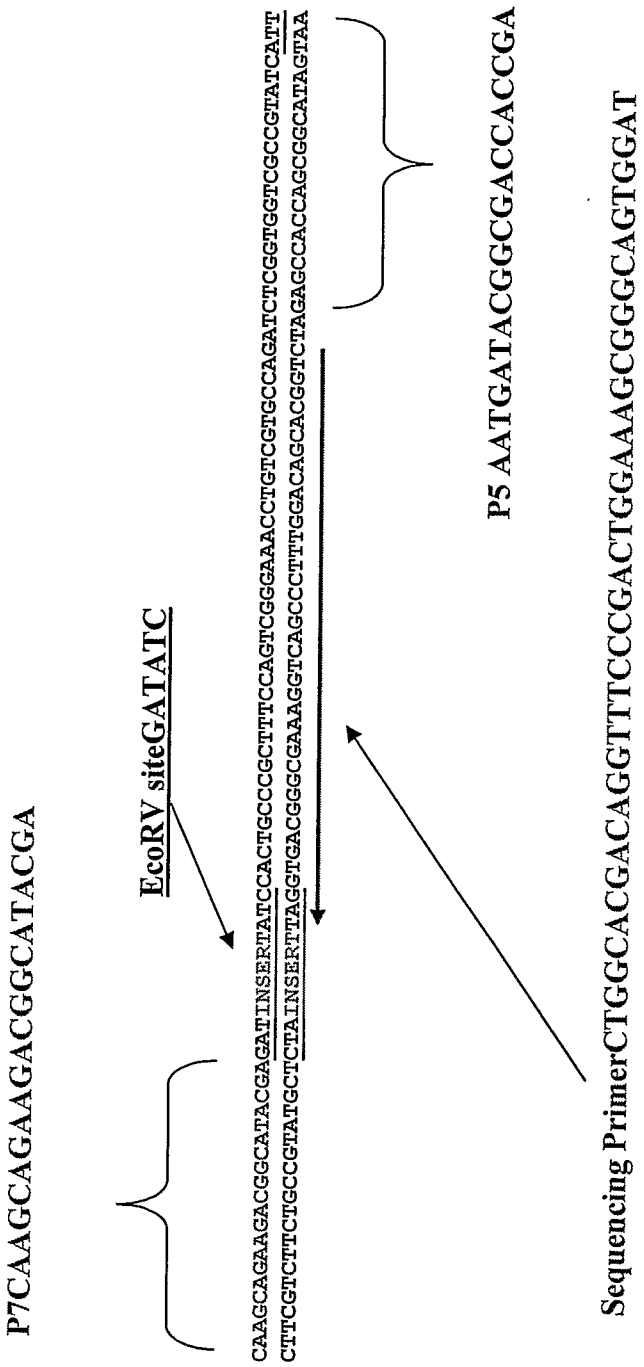

P7CAAGCAGAAGACGGCATACGA
EcoRV siteGATATC
CAAGCAGAAGACGGCATACGAGATINSERTATCCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGATCTCGGTGGTCGCCGTATCATT
CTTCGTCTTCTGCCGTATGCTCTAINSERTTAGGTGACGGGCGAAAGGTCAGCCCTTTGGACAGCACGGTCTAGAGCCACCAGCGGCATAGTAA
P5 AATGATACGGCGACCACCGA
Sequencing PrimerCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGGAT

Figure 5

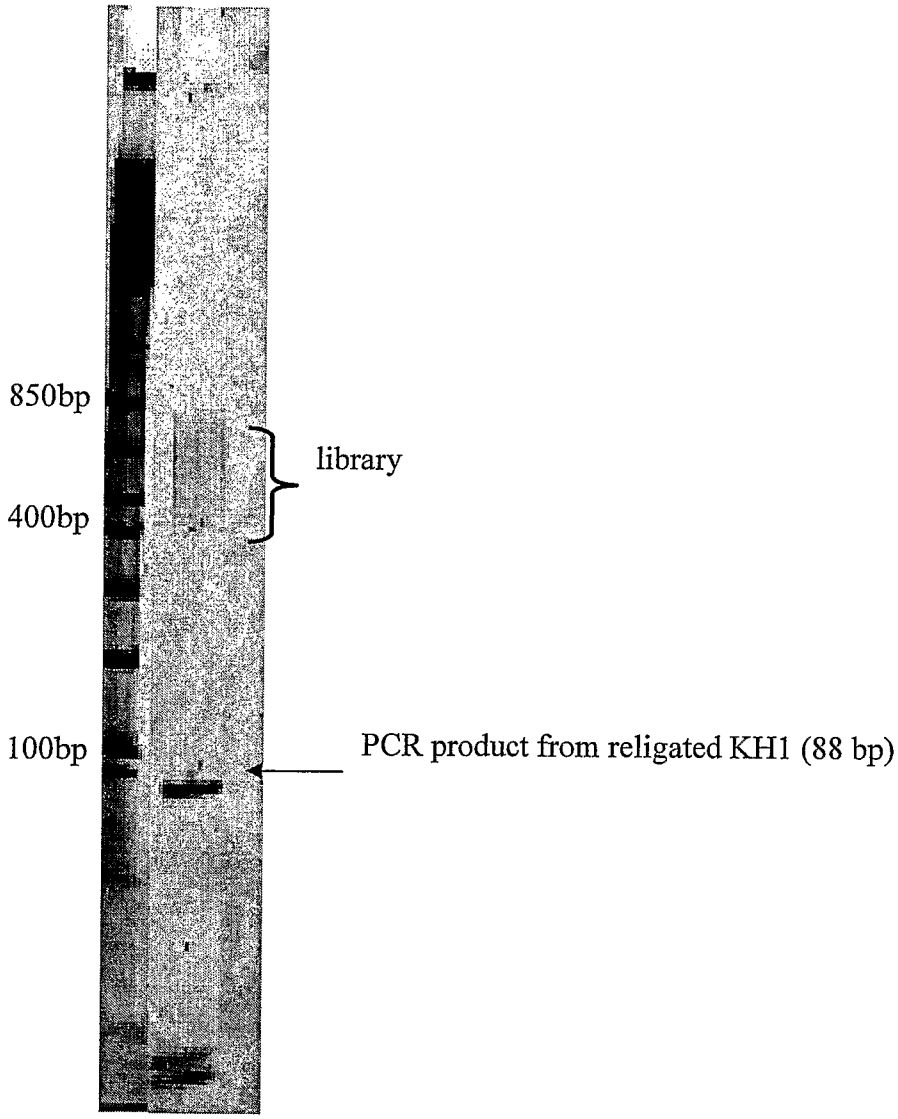

850bp
400bp
100bp
library
PCR product from religated KH1 (88 bp)

Figure 6

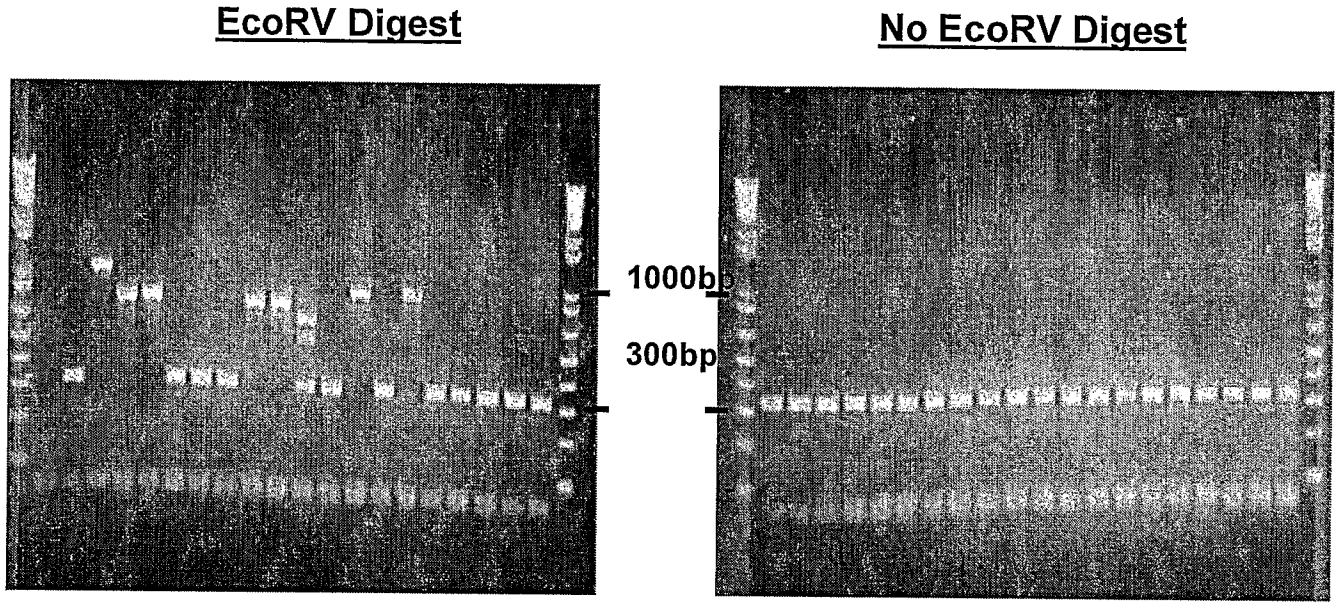


Figure 7: Validation of the Efficiency of the Ligation and EcoRV Digest