

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
24 September 2009 (24.09.2009)

PCT

(10) International Publication Number
WO 2009/116863 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C12N 15/10* (2006.01)

(21) International Application Number:
PCT/NL2009/050128

(22) International Filing Date:
17 March 2009 (17.03.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
08152859.8 17 March 2008 (17.03.2008) EP

(71) Applicant (for all designated States except US): **EXPRESSIVE RESEARCH B.V.** [NL/NL]; Nieuwe Kanaal 7b, NL-6709 PA Wageningen (NL).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **DE BOER, Anne Douwe** [NL/NL]; Oude Maasdijk 32, NL-6621 AC Dreumel (NL). **EBSKAMP, Michaël Johannes Marcus** [NL/NL]; Sterregaarde 25, NL-3436 ZA Nieuwegein (NL). **LANGEVELD, Simon Albertus** [NL/NL]; Leidsevaart 6, NL-2121 AX Bennebroek (NL). **LAROS, Ivo** [NL/NL]; Hicksweg 19, NL-6871 KS Renkum (NL).

(74) Agent: **HATZMANN, M.J.**; Vereenigde, Johan de Wittlaan 7, NL-2517 JR Den Haag (NL).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, 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, ME, 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, ST, 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, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, 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 (Rule 48.2(g))

(54) Title: EXPRESSION-LINKED GENE DISCOVERY

(57) Abstract: The invention relates to a method for analysing a genomic region of an organism, comprising four major parts. The first part involves the isolation of mRNA from a selected organism that is used for the preparation of small single stranded DNA fragments with one adaptor containing an affinity label. These DNA fragments are used in part three. In the second part genomic DNA from the same or a related organism is isolated. This genomic DNA is fragmented and ligated to adaptor molecules. In the third part these genomic fragments are hybridized with single stranded DNA fragments from part one and the hybrids formed in this process are used for synthesis of DNA fragments. These fragments will be used in part four, which involves sequencing of these fragments using one of the available high throughput sequencing methods.



WO 2009/116863 A2

Title: Expression-Linked Gene Discovery

The invention relates to the field of molecular biology and
5 biotechnology, more specifically to the field of sequencing, detection and
identification of nucleic acid sequences in genomic DNA. More in particular,
the invention relates to the application of a method in the identification of
and/or detection of nucleotide sequences which represent the majority of
transcribed regions and its surroundings in a genome and which are related to
10 a wide variety of genetic traits, genes and combinations thereof. The invention
can be used in the field of high throughput detection and identification of
molecular markers from any origin, be it plant, animal, human, artificial or
otherwise.

15 Breeding technologies have evolved from simple selection of visible
traits into advanced methods for detection of multigenic traits using molecular
markers. In principal, each genetic difference between different lines of a
crossing population can represent an altered trait. However due to the
complexity of most genomes it is not possible to indentify every difference that
20 exists between genomes and to link it to a particular trait. In theory,
sequencing complete genomes would reveal all differences between genomes.
However, this cannot be realized practically, time and cost effectively with
current sequencing technologies. Therefore, methods for detection of genetic
differences have mainly been based on the principle of complexity reduction
25 involving sequencing of a limited but fully defined part of the genomic DNA
from different individuals. With the advancements in sequencing technologies
complexity reduction has become less important for certain applications like
analysis of transcriptomes that represent all expressed gene sequences. The
size of eukaryotic genomes ranging from a few tens to several hundreds of
30 megabases is nevertheless still beyond the capacity of current high throughput
sequencing technologies. Furthermore, the vast majority of genomic DNA in

eukaryotic organisms, especially those with larger genome sizes, provides no valuable information for breeding purposes as it is never expressed and therefore does not seem to contribute to expression of traits.

Therefore, to identify molecular markers, methods that focus on those parts of a genome that are more prone to reveal molecular markers closely linked to traits have an advantage over methods which analyse just random selections from genomes including non-expressed areas. This problem becomes more prominent when genome size increases. The method described makes it possible to determine sequences in a selected part of the genomic DNA representing the coding regions of the majority of expressed genes and their surroundings. Comparison of such selected parts between different individuals allows identification of polymorphic sites that are inside or in close vicinity of expressed genes. Since frequency of polymorphisms is higher in non-coding regions, more polymorphisms can be related to expressed genes as with current technologies. Also large non-coding regions surrounding more conserved genes can be analysed for the presence of polymorphisms. This may ultimately result in the discovery of at least one marker per trait. The method of the invention makes it possible to focus on SNP detection in gene encoding areas and gene regulatory areas by enlightening well defined parts of the genome among different individuals and organisms, even in organisms with complex and large genomes.

Nucleotide sequence polymorphisms, like SNPs are widely applied to construct genome maps. After polymorphisms have been linked to phenotypes in a process called genetic mapping, such polymorphisms can be used as markers in marker assisted breeding technologies to detect a particular phenotype at any stage of development. Nucleotide sequence polymorphisms are generally identified in genomic DNA. As genome sizes of all eukaryotic organisms are far exceeding the number of nucleotides that can be analysed with current high throughput sequencing technologies, reproducible procedures for complexity

reduction are needed to analyse selected parts of whole genomes for finding genetic differences between individuals that can be used for genomic mapping. However, the statistical nature of complexity reduction methods which are currently applied, implies that these methods do not reveal *a priori* those genetic differences that can be linked to single phenotypes or are map closely to genes which contribute to a particular phenotype.

Current technologies strongly focus on discovery of single nucleotide polymorphisms (SNPs) for several reasons: SNPs are more frequently present in genomes than any other type of polymorphism, SNPs allow accurate detection of homozygous and heterozygous alleles, SNPs can be applied in high throughput applications and many industrial platforms are available which make SNP detection at any desired scale of application cost effective. Although SNP discovery would be the method of choice in situations where low levels of polymorphism occur, like conserved genes encoding areas and genomes of closely related individuals, using EST banks for SNP discovery in individuals that are closely related may be less effective because of an inherently low level of polymorphism.

In conclusion, an SNP discovery method should ideally reveal all present SNPs that are physically linked to traits of interest, but should not be impeded by the lower levels of polymorphisms as they occur in gene encoding areas of the genome or be impeded by any requirement for knowledge of genome sequences. Thus there is a need for a method that can reproducibly determine concomitantly sequences in areas of genomic DNA that represent the majority of gene coding regions and their surrounding regions, this without prior knowledge of genome or transcriptome sequences

Summary of the invention

The current inventors now have found a method for analysing a genomic region of an organism, comprising four major parts.

The first part involves the isolation of mRNA from a selected organism that is used for the preparation of small single stranded DNA fragments with one adaptor containing an affinity label. These DNA fragments are used in part three. In the second part genomic DNA from the same or a related organism is isolated. This genomic DNA is fragmented and ligated to adaptor molecules. In the third part these genomic fragments are hybridized with single stranded DNA fragments from part one and the hybrids formed in this process are used for synthesis of DNA fragments. These fragments will be used in part four, which involves sequencing of these fragments using one of the available high throughput sequencing methods.

Said method for the identification of genomic DNA in a sample thus comprises the steps of:

- a) isolation and purification of mRNA from tissue samples of an organism;
- b) synthesis of cDNA using said mRNA as a template;
- 15 c) optionally complexity reduction of said cDNA;
- d) fragmentation of said cDNA;
- e) optionally size selection of said fragments;
- f) optionally removing polyA-containing fragments by binding to streptavidin-coated affinity beads;
- 20 g) polishing of said fragments of cDNA;
- h) ligation of said fragments with one adaptor comprising a recognition site for a rare restriction enzyme and another adaptor containing a biotin label;
- i) optionally size selection of said fragments;
- j) nick repair of said fragments;
- 25 k) selection of said fragments that contain both adaptor sequences;
- l) amplification of said fragments using primers annealing to the adaptor sequences described in step h, wherein one primer is complementary to the adaptor with a rare restriction site and the other primer contains a biotin label.
- 30 m) binding said fragments to streptavidin-coated affinity beads;

- n) removing adaptors containing the rare restriction site using the corresponding restriction enzyme from said fragments;
- o) removing single strands not attached to affinity beads by a biotin-streptavidin interaction from double stranded DNA fragments attached to affinity beads;
- 5 p) isolation and purification of genomic DNA for example from the organism of step a
- q) fragmentation of said genomic DNA;
- r) optionally polishing of said genomic DNA;
- 10 s) ligation of said genomic DNA with one single type of adaptor or with two different type of adaptors;
- t) melting out of said genomic DNA into single stranded DNA;
- u) hybridizing genomic DNA from step t) with cDNA on beads from step o);
- v) remove unbound genomic DNA by washing;
- 15 w) extension of the cDNA-genomic DNA hybrid by a polymerase to create a double-stranded template;
- x) performing PCR on said genomic DNA-cDNA hybrid
- y) selection of fragments larger than about 100 basepairs from said PCR;
- z) optionally purification of said fragments and
- 20 aa) high throughput sequencing of said fragments.

In another embodiment, the method is extended to a method for identifying polymorphisms, comprising all the steps of the method according to claim and additionally comparing sequence data from two or more samples to identify polymorphisms.

25

Definitions

In the following description and examples a number of terms are used. In order to provide a clear and consistent understanding of the specification and claims, including the scope to be given such terms, the following definitions are provided. Unless otherwise defined herein, all technical and scientific terms

30

used have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. The disclosures of all publications, patent applications, patents and other references are incorporated herein in their entirety by reference.

5

Nucleic acid: a nucleic acid according to the present invention may include any polymer or oligomer of pyrimidine and purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively (See Albert L. Lehninger, *Principles of Biochemistry*, at 793-800 (Worth Pub. 1982)). The present invention contemplates any deoxyribonucleotide, ribonucleotide or peptide nucleic acid component, and any chemical variants thereof, such as methylated, hydroxymethylated or glycosylated forms of these bases, and the like. The polymers or oligomers may be heterogeneous or homogenous in composition, and may be isolated from naturally occurring sources or may be artificially or synthetically produced. In addition, the nucleic acids may be DNA or RNA, or a mixture thereof, and may exist permanently or transitionally in single-stranded or double-stranded form, including homoduplex, heteroduplex, and hybrid states.

10

SNP: A single nucleotide polymorphism is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species (or between paired chromosomes in an individual) at a specific locus. SNPs are the most common type of genetic variation. SNPs may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code. A SNP in which both forms lead to the same polypeptide sequence is termed synonymous and if a different polypeptide sequence is produced it is termed non-synonymous. As SNPs are

20
25

evolutionarily conserved, they can be used as markers for quantitative trait loci (QTL) analysis and in association studies.

Intron: Introns are non-coding sections of a gene that are removed from pre-mRNAs in a process named splicing to produce functional mRNAs.

Exon: An exon is any region of DNA within a gene that is transcribed to the final messenger RNA (mRNA) molecule, rather than being spliced out from the transcribed RNA molecule like introns.

10

cDNA: cDNA is an artificial form of DNA synthesized by the reverse transcriptase enzyme using RNA molecules as a template.

Genomic DNA: the term genomic DNA indicates that the DNA is derived from the situation 'as is'. This means that genomic DNA harbors sequences, as they will be found in nature, e.g. including introns and regulatory sequences.

Genomic DNA can be derived from different sources, like chromosomes, but also from extrachromosomal sources such as mitochondria, chloroplasts and plasmids.

20

Cot-1 DNA: The technique that is used to determine the sequence complexity of any genome involves the denaturation and renaturation of DNA. DNA is denatured by heating and this melts the H-bonds and renders the DNA single-stranded. If the DNA is rapidly cooled, the DNA remains single-stranded. But if the DNA is allowed to cool slowly, sequences that are complementary will find each other and eventually base pair again. The rate at which the DNA reanneals (another term for renatures) is a function of the species from which the DNA was isolated, also identified as the "Cot" curve. DNA that has a high Cot value is highly repetitive DNA, while DNA with a low Cot value is only available in low copies or is unique. In the method we use DNA with a Cot

30

value of 1, a fraction of total genomic DNA that is enriched for highly repetitive DNA sequences.

Annotation: annotation of cDNA sequences comprises two steps. The obtained
5 sequences are compared to nucleotide and/or amino acid sequences as
available in (public) databases. Methods for alignment of sequences for
comparison purposes are well known in the art. This comparison is typically
performed with help of a program such as the NCBI Basic Local Alignment
Search Tool (BLAST) describe by Altschul et al., 1990). This program is
10 available from several sources including the National Center for Biological
Information (NCBI, Bethesda, Md.) and on the internet
([HTTP://www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/)). This program compares the identified
cDNA/EST (Expressed Sequence Tags) sequences with the sequences present
in databases and presents the results based on a certain score and a
15 probability parameter. The program can select those cDNA/EST sequences
with a certain predetermined lower limit of said probability parameter. The
selected cDNA/EST sequences are then in a second step provided with an
annotation (i.e. a link to a sequence as present in a database). This kind of
annotation is called an "electronic annotation".

20

Clustering: with the term "clustering" is meant the building of a collection of
sequences with similarity through pairwise comparison of two or more
nucleotide sequences and selection on the presence of short or long stretches of
identical or similar nucleotides. Several methods for alignment of nucleotide
25 sequences are known in the art, as will be further explained below. Sometimes
the terms "assembly" or "alignment" are used as synonyms.

Identifier: a short sequence that can be added to an adaptor or a primer or
included in its sequence or otherwise used as label to provide a unique
30 identifier. Such a sequence identifier can be a unique base sequence of varying

but defined length uniquely used for identifying a specific nucleic acid sample. For instance 4 bp tags allow $4(\text{exp}4) = 256$ different tags. Typical examples are ZIP sequences, known in the art as commonly used tags for unique detection by hybridization (Iannone et al. Cytometry 39:131-140, 2000). Using such an
5 identifier, the origin of a PCR sample can be determined upon further processing. In the case of combining processed products originating from different nucleic acid samples, the different nucleic acid samples are generally identified using different identifiers.

10 Sequencing: The term sequencing refers to determining the order of nucleotides (base sequences) in a nucleic acid sample, e.g. DNA or RNA. High-throughput, screening, often abbreviated as HTS, is a method for scientific experimentation especially relevant to the fields of biology and chemistry. Through a combination of modern robotics and other specialized
15 laboratory hardware, it allows a researcher to effectively screen large amounts of samples simultaneously, more specifically this is a sequencing technology such as disclosed herein elsewhere (from 454 Life Sciences, www.454.com and Illumina, www.illumina.com). E.g. the Illumina Solexa sequencing approach relies on attachment of randomly fragmented genomic DNA to a planar,
20 optically transparent surface and solid phase amplification to create an ultra-high density sequencing flow cell with > 10 million clusters, each containing ~1,000 copies of template per sq. cm. These templates are sequenced using a robust four-color DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescence. This approach ensures
25 high accuracy and avoidance of artefacts with homopolymeric repeats. High sensitivity fluorescence detection is achieved using laser excitation and total internal reflection optics.

Restriction endonuclease: a restriction endonuclease or restriction enzyme is
30 an enzyme that recognizes a specific nucleotide sequence (target site) in a

double-stranded DNA molecule, and will cleave both strands of the DNA molecule at or near every target site.

Restriction fragments: the DNA molecules produced by digestion with a restriction endonuclease are referred to as restriction fragments. Any given genome (or nucleic acid, regardless of its origin) will be digested by a particular restriction endonuclease into a discrete set of restriction fragments. The DNA fragments that result from restriction endonuclease cleavage can be further used in a variety of techniques and can for instance be detected by gel electrophoresis.

Ligation: the enzymatic reaction catalyzed by a ligase enzyme in which two double-stranded DNA molecules are covalently joined together is referred to as ligation. In general, both DNA strands are covalently joined together, but it is also possible to prevent the ligation of one of the two strands through chemical or enzymatic modification of one of the ends of the strands. In that case the covalent joining will occur in only one of the two DNA strands.

Synthetic oligonucleotide: single-stranded DNA molecules having preferably from about 10 to about 50 bases, which can be synthesized chemically are referred to as synthetic oligonucleotides. In general, these synthetic DNA molecules are designed to have a unique or, desired nucleotide sequence, although it is possible to synthesize families of molecules having related sequences and which have different nucleotide compositions at specific positions within the nucleotide sequence. The term synthetic oligonucleotide will be used to refer to DNA molecules having a designed or desired nucleotide sequence.

Adaptors: short double-stranded DNA molecules with a limited number of base pairs, e.g. about 10 to about 30 base pairs in length, which are

designed such that they can be ligated to the ends of restriction fragments. Adaptors are generally composed of two synthetic oligonucleotides that have nucleotide sequences which are partially complementary to each other. When mixing the two synthetic oligonucleotides in solution under appropriate
5 conditions, they will anneal to each other forming a double-stranded structure. After annealing, one end of the adaptor molecule is designed such that it is compatible with the end of a restriction fragment and can be ligated thereto; the other end of the adaptor can be designed so that it cannot be ligated, but this need not be the case (double-ligated adaptors). Specifically, in the present
10 invention an adaptor is a double-stranded DNA molecule, usually between 15 and 60 base pairs, which can be prepared by annealing of two (partially) complementary oligonucleotides. The adaptors used herein can be blunt-ended, or have a specific overhang for ligation to DNA molecules with a complementary overhang, such as those created by restriction endonuclease
15 digestion. The adaptors can have an additional, non-compatible overhang on the other end, which cannot ligate to a blunt-ended DNA fragment and not to DNA with a specific overhang created by a type II restriction endonuclease, thereby preventing the ligation of multiple adaptors with each other.

The adaptors provide annealing sites for primers during PCR to
20 amplify the different DNA fragments ligated to the adaptors in a single PCR reaction.

One of the adaptors ligated to the cDNA molecules can carry a type
25 II's restriction endonuclease recognition sequence to cleave the adaptor off the cDNA molecule once bound to the solid phase.

25

Adaptor-ligated restriction fragments: restriction fragments that have been capped by adaptors.

Primers: in general, the term primers refer to DNA strands, which
30 can prime the synthesis of DNA. DNA polymerase cannot synthesize DNA *de*

*nov*o without primers: it can only extend an existing DNA strand in a reaction in which the complementary strand is used as a template to direct the order of nucleotides to be assembled. We will refer to the synthetic oligonucleotide molecules that are used in a polymerase chain reaction (PCR) as primers.

5

DNA amplification: the term DNA amplification will be typically used to denote the *in vitro* synthesis of double-stranded DNA molecules using PCR or a comparable amplification system. It is noted that other amplification methods exist and they may be used in the present invention. Methods of the invention can in principle be performed by using any nucleic acid amplification method, such as the Polymerase Chain Reaction (PCR; Mullis 1987, U.S. Pat. No. 4,683,195, 4,683,202, en 4,800,159) or by using amplification reactions such as Ligase Chain Reaction (LCR; Barany 1991, Proc. Natl. Acad. Sci. USA 88:189-193; EP Appl. No., 320,308), Self-Sustained Sequence Replication (3SR; 10 Guatelli et al., 1990, Proc. Natl. Acad. Sci. USA 87:1874-1878), Strand Displacement Amplification (SDA; U.S. Pat. Nos. 5,270,184, en 5,455,166), Transcriptional Amplification System (TAS; Kwoh et al., Proc. Natl. Acad. Sci. USA 86:1173-1177), Q-Beta Replicase (Lizardi et al., 1988, Bio/Technology 6:1197), Rolling Circle Amplification (RCA; U.S. Pat. No. 5,871,921), Nucleic 20 Acid Sequence Based Amplification (NASBA), Cleavase Fragment Length Polymorphism (U.S. Pat. No. 5,719,028), Isothermal and Chimeric Primer-initiated Amplification of Nucleic Acid (ICAN), Ramification-extension Amplification Method (RAM; U.S. Pat. Nos. 5,719,028 and 5,942,391) or other suitable methods for amplification of DNA.

25

In order to amplify DNA with a small number of mismatches to one or more of the amplification primers, an amplification reaction may be performed under conditions of reduced stringency (e.g. a PCR amplification using an annealing temperature of 38°C, or the presence of 3.5 mM MgCl₂). The person

skilled in the art will be able to select conditions of suitable stringency.

Polishing (also called end repair) refers to the conversion of non-blunt ended DNA into blunt ended DNA.

5 Polishing digestion of genomic DNA (gDNA) templates with DNase I in the presence of Mn²⁺ produces fragments of DNA that are either blunt-ended or have protruding termini with one or two nucleotides in length. Similarly, fragmentation of DNA by mechanical means provides a combination of fragments with blunt-ends or overhanging ends. These DNA fragments,
10 whether generated enzymatically or mechanically, may be "polished" using the procedure described below.

In one method, polishing may be performed by treatment of 3'-overhanging fragments with a single strand-specific exonuclease, such as BAL32 nuclease or Mung Bean nuclease. Generally, the nuclease should be
15 calibrated prior to use.

In another method, blunt ends are created with Pfu DNA polymerase or with other DNA polymerases such as T4 DNA polymerase or Klenow DNA polymerase. Pfu "polishing" or blunt ending can be used to increase the amount of blunt-ended species generated following genomic template digestion
20 with DNase I. Pfu DNA polymerase fills in 5' overhangs. Additionally, Pfu DNA polymerase exhibits 3' to 5' exonuclease activity. Thus, the enzyme can be used to remove single and double nucleotide extensions to further increase the amount of blunt-ended DNA fragments available for adaptor ligation (see, e.g., Costa, G. L. and M. P. Weiner, 1994, *Protocols for cloning and analysis of*
25 *blunt-ended PCR-generated DNA fragments. PCR Methods Appl 3 (5): S95;* Costa, G. L., A. Grafsky and M. P. Weiner, 1994, *Cloning and analysis of PCR-generated DNA fragments. PCR Methods Appl 3 (6): 338;* Costa, G. L. and M. P. Weiner, 1994, *Polishing with T4 or Pfu polymerase increases the efficiency of cloning of PCR products. Nucleic Acids Res. 22 (12): 2423*).

The present inventors have found that by providing the genomic DNA and by using adaptor-ligated cDNA derived fragments to act as primers for amplification of genomic DNA fragments, it appears possible to detect genome sequences outside regions that are actually transcribed, i.e. the promoter, 5 intron and terminator sequences. It thus combines the possibility of rapid identification of genomic DNA fragments in expressed regions of the genome, wherein multiple samples can be sequenced in a single run, with the possibility to study surrounding genomic DNA fragments of these regions and the possibility to detect genetic variation in gene encoding fragments, intron 10 fragments and fragments covering regulatory genome sequences.

A further advantage is that the method of the invention is generally applicable, i.e. for all organisms. No information on the genome or genomic organization is needed beforehand. Another advantage is that no cloning step is required in the present method. This enables the sequencing of sequences 15 coding for toxic substances or for regulatory proteins, which otherwise would be impossible because the host organism in which such sequences would be cloned and expressed would not or hardly survive. In the same sense also sequences that would create problems in cloning can now be approached and there is no limit as to the length of the sequences, which also may cause 20 problems in cloning schedules.

A further advantage is that the current method does not require full length cDNA, but may be used with shorter sequences. This is especially useful when analyzing large genomes (e.g. from bulbous plants) since then a focus on the most interesting or relevant regions of the genome can be 25 maintained.

A next advantage is that it is possible with the current method to generate primers for all expressed DNA sequences, which means that it is possible to obtain the genomic data from the expressed sequences.

Brief description of the drawings

Figure 1. Schematic representation of creation of small single stranded DNA molecules from mRNA. C is an adaptor, D is a different adaptor, B is biotin, P is a phosphate group, bp is base pairs, ds is double stranded, ss is single stranded, RE is a restriction enzyme, SA is streptavidin, LD-PCR is long distance PCR, NBA₍₃₀₎ and NVT₍₃₀₎ are single-letter code for nucleotides according the Nomenclature Committee of the International Union of Biochemistry (NC-IUB) syntax rules for nucleotides, A₍₃₀₎ and T₍₃₀₎ indicate stretches of 30 As and 30 Ts respectively.

10 Figure 2. Schematic representation of creation of fragments of genomic DNA and the subsequent hybridization with small single stranded DNA molecules from mRNA (figure 1). After several steps, DNA fragments are obtained that can be sequenced using high throughput sequencing methods. C and D are adaptors also depicted in figure 1. E and F are different adaptors, B is biotin, bp is base pairs, ds is double strand, ss is single strand and SA is streptavidin.

Figure 3. Non-phosphorylated A and B adaptors are ligated to the ends of phosphorylated, polished, double-stranded genomic DNA fragments. The A and B adaptors differ in both nucleotide sequence and the presence of a 5' biotin tag on the B adaptor. Nicks are present at the 3'-junctions of each of the adaptors and the fragments are filled in by the strand-displacement activity of Bst DNA polymerase. Streptavidin- biotin interactions are used to remove fragments flanked by homozygous adaptor sets (A/A and B/B) and to generate single stranded library templates. Fragments are bound to Streptavidin beads; unbound material (composed of homozygous A/A adaptor sets, which lack biotin) is washed away. The immobilized fragments are then denatured; both strands of the B/B fragments remain immobilized through the biotinylated B adaptor, while A/B fragments are washed free and used in subsequent steps.

Figure 4 ELGD schemes cases 1 and 2 and 3. **Panel A. Case 1.**

Schematic representation of the alignment of apple genomic sequence FRA8S6E02IN5QW, apple EST contig cg8984st_846_7 and *Arabidopsis thaliana* gene AT1G70160 (not drawn to scale). Only the relevant 3'-part of AT1G70160 is shown. Introns are represented as white boxes in FRA8S6E02IN5QW and AT1G70160. Spliced introns are shown as thin lines in cg8984st_846_7. The protein coding sequences (CDS) are shown in black. The 3'-untranslated regions are shown in grey. The polyA-tail in the cDNA sequence is indicated with (A)_n. **Panel B. Case 2.** Schematic representation of the alignment of apple genomic sequence FRA8SE02HOH39 and apple EST contig cg12357st_1428_21 (not drawn to scale). The protein coding sequence (CDS) is shown in black. The 5'- and 3'-untranslated regions are shown in grey. The polyA-tail in the cDNA sequence is indicated with (A)_n. **Panel C. Case 3.** Schematic representation of the alignment of apple genomic sequence 02-H03 and apple EST 91044590 (not drawn to scale). The protein coding sequence (CDS) is shown in black. The 5'- and 3'-untranslated regions are shown in grey. The non-transcribed promoter region in the genomic sequence is shown in white.

20 Detailed description of the invention

In one aspect, the invention relates to a method for the identification of genomic DNA in a sample comprising of 4 different parts.

Part 1. In this part small sequences are generated from cDNA that will be used as priming sequences in part 3. The part consists of the following steps:

- a) isolation and purification of mRNA from tissue samples of an organism;
- b) synthesis of cDNA using said mRNA as a template;
- c) optionally complexity reduction of said cDNA;
- 30 d) fragmentation of said cDNA;

- e) optionally size selection of said fragments;
- f) optionally removing polyA-containing fragments by binding to (streptavidin-coated) affinity beads;
- g) polishing of said fragments of cDNA;
- 5 h) ligation of said fragments with a first adaptor comprising a recognition site for a rare restriction enzyme and a second adaptor containing a biotin label;
- i) optionally size selection of said fragments;
- j) nick repair of said fragments;
- k) selection of said fragments that contain both adaptors sequences;
- 10 l) amplification of said fragments using primers annealing to the adaptor sequences described in step h; one primer complementary to the adaptor with the rare restriction site and the other primer containing a biotin label.
- m) binding said fragments to streptavidin-coated affinity beads;
- n) removing adaptors containing the rare restriction site from said fragments
- 15 o) removing single strands not attached to affinity beads by a biotin-streptavidin interaction from the double stranded DNA fragments attached to affinity beads.
- 20 This step results in single strands of DNA bound to streptavidin affinity beads;

Part 2:

- p) isolation and purification of genomic DNA for example from the organism of step a);
- 25 q) fragmentation of said genomic DNA;
- r) optionally polishing of said genomic DNA;
- s) ligation of said genomic DNA with one single type of adaptor or with two different type of adaptors (preferred)
- t) melting out of said genomic DNA into single stranded DNA

Part 3:

- u) hybridizing genomic DNA from step t) with cDNA on beads from step o);
- v) remove unbound genomic DNA by washing
- w) extension of the cDNA-genomic DNA hybrid by a polymerase to create a
5 double-stranded template
- x) performing PCR on said genomic DNA-cDNA hybrid
- y) selection of fragments larger than about 100 basepairs from said PCR by
size fractionation
- z) optionally purification of said fragments

10

Part 4

- aa) Use the fragments obtained in step z) for high throughput sequencing
according to manufacturer.

- 15 Additionally, when the sequence of the genomic DNA has been determined a
step can follow to identify changes in the sequences of two or more samples:
- ab) Comparing the data of two or more sample to identify polymorphisms.

20 By treating a sample nucleic acid in this way it is possible to analyse
reproducibly genomic regions of an organism covering gene encoding and
connected regions, without having any information about the structure or
content of the genome of the organism. When the sequencing protocol allows
for adaptors with identifiers, multiple samples can be combined in a single
sequencing run.

25

The method starts with the isolation and purification of a sample of mRNA
from an organism. Obtaining a sample of mRNA in this way is a routine
procedure nowadays.

- 30 Similarly for the next step, wherein – with the aid of the enzyme reverse
transcriptase – a DNA copy, the so called cDNA is made from the RNA in the

sample. This cDNA comprises the transcriptome of the cell from which the RNA has been derived, representing the total of genetic information that has been transcribed at the moment that the sample was taken. Thus, depending on the cell type, the tissue from which the cell was derived, the age of the cell, the developmental phase of the cell and environmental conditions, the transcriptome of each cell of the same organism will differ, and even from the same cell, when sampled at different times and/or different conditions, different transcriptomes would be obtained. The nucleic acid in the original sample will usually be in the form of mRNA. However, also RNA or DNA derived from other sources can be useful, such as RNA or DNA derived from gene libraries. The nucleic acid in the sample may be double stranded, single stranded, and double stranded DNA denatured into single stranded DNA. The sample can be from any organism, whether plant, animal, synthetic or human. It will be understood that if a DNA sample is obtained, no reverse transcriptase reaction is needed.

Since the sample, if it is derived from whole cell mRNA, contains the total transcriptome, sometimes it would be desirable to retrieve only a subset of it. This can be achieved in several ways: one possible way is to discriminate on size of the cDNA, e.g. by ultraspeed centrifugation. The principle underlying this method is known as complexity reduction. Other options for complexity reduction are for example hybridization methods which select for abundant or non-abundant transcripts or for example methods for capturing specific transcripts either to remove them from the pool of cDNA molecules or to select them for further analysis or for example methods like cDNA-AFLP that create a subset of the pool of cDNA molecules by restriction digestion.

Once the desired sample of cDNA has been obtained, the cDNA is fragmented, which can be done enzymatically or mechanically. The nucleic acid sample is digested with at least one restriction endonuclease to provide a set of restriction fragments. In certain embodiments, two or more endonucleases can

be used to obtain restriction fragments. The endonuclease can be a frequent cutter (a recognition sequence of 3-5 bp; such as MseI) or a rare cutter (recognition sequence of >5 bp, such as EcoRI). In certain preferred embodiments, a combination of a rare and a frequent cutter is preferred. In certain embodiments, in particular when the sample contains or is derived from a relatively large genome, it may be preferred to use a third enzyme (rare or frequent cutter) to obtain a larger set of restriction fragments of shorter size.

As a restriction endonuclease, any endonuclease will suffice. Typically, Type II endonucleases are preferred such as EcoRI, MseI, PstI etc. In certain embodiments a type II's endonuclease may be used, i.e. an endonuclease of which the recognition sequence is located distant from the restriction site, i.e. such as AceIII, BbvI, BbvII, BbsI, Bed, Bce83I, BceI, BcgI, BlnI, BsaI, BsgI, BsmAI, BsmFI, BspMI, Esp3I, FaeI, FokI, GsuI, HgaI, MboII, MmeI, MnlII, SapI, SfaNI, TaqI and ZthII.

Mechanical fragmenting can be done by shearing the cDNA, wherein the severity and duration of shearing determine the amount of fragmenting. One such shearing method is nebulization. A nebulizer is a small plastic device that uses compressed air to atomize liquids. They are easily adapted for shearing DNA, extremely efficient and simple to use (Surzycki, S., 2000, Basic Methods in Molecular Biology, NY – Springer Verlag). They are commercially available from different sources (e.g. Invitrogen Corporation).

Again complexity reduction can be optionally performed on this fragmented DNA by size selection. Optionally complexity reduction can be obtained by size selection of the fragments, e.g. by electrophoresis. Additionally, or alternatively, poly-A fragments can be removed by binding to streptavidin coated affinity columns.

In the next step of the method, the cDNA fragments are made blunt ended, a process also indicated as 'polishing'. Insert polishing is used to remove 3 prime overhang nucleotide(s) or fill in 5 prime overhang nucleotides from restriction

enzyme generated, PCR-generated DNA fragments or sheared DNA fragments. Kits for polishing are commercially available (e.g. Quick blunt kit, New England Biolabs Inc.)

In the next step of the method of the invention, the cDNA fragments are provided with adaptors through a ligation reaction. In this reaction two different types of adaptors are ligated to a mixture of cDNA fragments. One adaptor is carrying a biotin label. The other adaptor contains a restriction site for a type II restriction enzyme that cuts outside its recognition sequence. An example of this is SapI that recognises the GCTCTTCNNNN sequence. The enzyme cuts in the NNNN sequence leaving an NNN overhang on the 5' end. In this way the complete adaptor sequence containing the restriction recognition sequence can be removed from the fragment. It is important to use an enzyme that is a rare cutter to prevent shortening of the fragments by a more frequent occurrence of recognition sequences in the fragment DNA.

Again, at this stage, optionally the fragments can be selected dependent upon their size by e.g. electrophoresis.

After ligation of the adaptors to the fragments, the fragments are subjected to nick repair, to fill any gap in the DNA backbone that may have been created by ligation of the adaptors.

Fragments that only contain both adaptors can be selected as described in the Supplementary Figure 1 of the publication by M. Margulies et al. in Nature 437, pages 376-380, 2005 (figure 3). Fragments that contain both adaptors can then be amplified. In this step a PCR amplification reaction is performed with one primer annealing to one of the complementary adaptor sequences in combination with another primer carrying a biotin molecule annealing to the other complementary adaptor sequence. After performing an amplification step using this set of primers, the amplified double stranded DNA fragments are captured on streptavidin affinity beads while other reaction products are removed from the reaction mixture.

In a next step, the bound DNA fragments, having different adaptor sequences at their 5' and 3' termini, are treated with an asymmetric endonuclease cutting the fragment in such a way as to remove a complete adaptor sequence from one end of the fragment, as described earlier, thus providing one fragment terminus that is perfectly complementary to genomic DNA. After this restriction step, free single stranded forms of the fragments are removed by elution from the beads under alkaline conditions, resulting in a set of single stranded nucleic acid fragments bound to streptavidin beads. An alternative method for generating single stranded fragments from double stranded fragments is treatment with lambda exonuclease enzyme. Lambda exonuclease degrades from a double stranded DNA molecule those strands that possess phosphorylated 5' termini, leaving single strands with 5'OH termini intact. Since one of the 5' ends of the fragment is protected by the biotin label that is bound to the streptavidin, treatment with lambda exonuclease also creates single strand DNA fragments bound to the streptavidin beads. These single strand DNA fragments will be used in later steps.

Next to the steps described above, in the second part genomic DNA is isolated from the organism. This organism can be the same as from which the mRNA is isolated (cultured under the same or different conditions) or it can be a different organism (different strain, different species) and it can even be a collection of genomic DNAs (e.g. a BAC clone library). Procedures for isolating genomic DNA are standard in the field and are e.g. described in Ausubel et al. (Preparation of genomic DNA from plant tissue. pp. 2.3.1-2.3.7 in Ausubel et al., eds. Current Protocols in Molecular Biology. John Wiley & Sons, Inc. Budelier. 1993). After isolation, the genomic DNA is fragmented by enzymatic or mechanical fragmentation procedures as described above. If restriction enzyme digestion is used, the resulting fragments are provided with adaptors matching the different restriction site overhangs at the termini of the fragments. The adaptor molecules attached to genomic DNA fragments have

sequences that are different from the adaptor molecules attached to cDNA derived fragments of part 1.

If mechanical fragmentation is used, the different adaptors are ligated to the genomic DNA fragments and fragments having different adaptors at the 5' and 3' termini are selected according to the procedure previously described for cDNA fragments (e.g. a polishing step can be applied).

Alternatively, also one single adaptor can be ligated to the genomic fragments. In that case adaptor sequences with panhandles can be used to prevent non-specific amplification in later steps (D H Jones and S C Winistorfer, PCR Methods Appl. 1993 2: 197-203).

The fragmented genomic DNA fragments can be optionally size-selected as described above.

To obtain single stranded genomic DNA fragments a melting out step is provided.

After thus providing an appropriate sample of fragmented genomic single stranded DNA, the single stranded cDNA fragments from part 1, carrying a single adaptor molecule with a biotin label which may be optionally bound to streptavidin affinity beads, are mixed in preferentially a molar excess of, for hybridization with the fragmented genomic DNA carrying different adaptor molecules (part 3). After a denaturation step, annealing conditions are applied to allow formation of hybrid double stranded genomic DNA – cDNA molecules. The hybridization procedure may include a short period of pre-annealing with non-fragmented cDNA (without adaptors) to normalize the hybridization reaction by reducing the effects of higher concentrations of cDNA fragments derived from abundant transcripts. As an optional modification, *Cot-1*-DNA may be used in a pre-annealing step to reduce possible anomalies caused by sequence repeats in genomic DNA.

In this step the cDNA fragments will anneal to the genomic DNA at places that are homologous or identical to the sampled mRNA.

After extension of the cDNA-genomic DNA hybrid by a polymerase a double-stranded template is made.

The annealed double stranded fragments can now be amplified by a PCR reaction using one primer complementary with the cDNA adaptor and another primer complementary to one of the genomic DNA adaptors. Optionally the annealed material can be separated into two separate fractions to also use the cDNA adaptor and the other primer complementary to the genomic DNA adaptor. The amplification provides PCR fragments that do not only contain copies of the genomic DNA corresponding part of the originally sampled nucleic acid, but also contain non-transcribed sequences, such as regulatory sequences and introns.

After the PCR amplification, fragments of more than 100 basepairs, more preferably more than about 200 basepairs, even more preferably more than about 300 basepairs and most preferably of about 400 basepairs or more are selected through size fractionation. Optionally, these fragments are purified according to meet the requirements for the next part of the method of the invention.

In the next part (part 4) said fragments are sequenced. Sequencing of the amplified adaptor-ligated fragments provides sequence information on at least part of the adaptor-ligated fragment and the 3' flanking genomic sequences. The information contained in the adaptor-derived part contains information about the sample from which the fragment is obtained if adaptors carry sample specific tags, whereas sequence information from the fragment itself (an identifier sequence) provides information about the fragment and allows for identification of the fragment. This sequence information on the fragment is used to identify the fragment with an accuracy that depends on the number of nucleotides that is determined and the number of fragments in the set of amplified adaptor-ligated fragments.

To provide a solution to the problem of sampling variation in transcript frequencies between samples, which affects the accuracy of identifying

molecular markers by sequencing contained in a set of multiple fragments, the present inventors have also found that detection of markers via sequencing is preferably performed with sufficient redundancy (depth) to sample all fragments at least once and accompanied by statistical means which address the issue of sampling variation in relation to the accuracy of the genotypes called. In order to increase the accuracy, preferably an amplification step proceeds the sequencing step. After sufficient cycles of amplification, the redundancy of the amplified adaptor-ligated restriction fragments is at least 6, preferably at least 7, more preferably at least 8 and most preferably at least 9. In preferred embodiments thus the sequence of each adaptor-ligated restriction fragment is determined at least 6, preferably at least 7, more preferably at least 8 and most preferably at least 9 fold. In certain embodiments, the redundancy is selected such, assuming a 50/50 overall chance of identifying the locus correctly as homozygous, that the chance of correct identification of the locus is more than 95%, 96%, 97%, 98%, 99%, 99.5%.

The amplification of the adaptor-ligated restriction fragments leads to a set of amplified adaptor-ligated restriction fragments, sometimes referred to as amplicons. The amplicons (or at least part thereof) are subjected to a step that comprises at least the determination of the sequence of the sample specific identifier to determine the origin of the fragment and of part of the sequence of the restriction fragment. In practice this amounts also to the determination of the sections located in between such as the remains of the recognition sequence of the restriction endonuclease. By sequencing the sample specific identifier in combination with part of the fragment located adjacent to the adaptor derived sequence, it is possible to uniquely identify restriction fragments and their 3' flanking genomic sequences. From this information it is possible to restore the genomic genetic information of a complete gene.

The high-throughput sequencing used in the present invention is a method for

scientific experimentation especially relevant to the fields of biology and chemistry.

It is preferred that the sequencing is performed using high-throughput sequencing methods, such as the methods disclosed in WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, and WO 2005/003375 (all in the name of 454 Life Sciences), by Seo et al. (2004) Proc. Natl. Acad. Sci. USA 101:5488-93, and technologies of Helios, Solexa, US Genomics, etcetera, which are herein incorporated by reference.

The technology described allows sequencing of 40 million bases in a single run and is 100 times faster and cheaper than competing technology. The sequencing technology roughly consists of 5 steps: 1) fragmentation of DNA and ligation of specific adaptors to create a library of single-stranded DNA (ssDNA), 2) annealing of ssDNA to beads, emulsification of the beads in water-in-oil microreactors and performing emulsion PCR to amplify the individual ssDNA molecules on beads, 3) selection or enrichment for beads containing amplified ssDNA molecules on their surface, 4) deposition of DNA carrying beads in a PicoTiter(TM) Plate; and 5) simultaneous sequencing in 100,000 wells by generation of a pyrophosphate light signal. The method will be explained in more detail below.

In this respect the following calculation may be illustrative: The sequencing technology of Illumina Solexa as described herein elsewhere, provides for 40,000,000 reads of about 25 bp each, totaling a staggering 1 billion bp in one single run. Assuming a redundancy in sampling of 10 times, 4,000,000 unique fragments can be assessed in one run. Combining 100 samples allows for 40,000 fragments to be sequenced for each sample.

In a preferred embodiment, the sequencing comprises the steps of: (a) annealing adapted fragments to beads, each bead being annealed with a single adapted fragment;

(b) emulsifying the beads in water-in-oil microreactors, each water-in-oil microreactor comprising a single bead;

(c) loading the beads in wells, each well comprising a - single bead; and generating a pyrophosphate signal.

In the first step (a), sequencing adaptors are ligated to fragments within the combination library. Said sequencing adaptor includes at least a "key" region for annealing to a bead, a sequencing primer region and a PCR primer region. Thus, adapted fragments are obtained. In a first step, adapted fragments are annealed to beads, each bead annealing with a single adapted fragment. To the pool of adapted fragments, beads are added in excess as to ensure annealing of one single adapted fragment per bead for the majority of the beads (Poisson distribution). In a next step, the beads are emulsified in water-in-oil microreactors, each water-in-oil microreactor comprising a single bead. PCR reagents are present in the water-in-oil microreactors allowing a PCR reaction to take place within the microreactors. Subsequently, the microreactors are broken, and the beads comprising DNA (DNA positive beads) are enriched.

In a following step, the beads are loaded in wells, each well comprising a single bead. The wells are preferably part of a PicoTiter(TM)Plate allowing for simultaneous sequencing of a large amount of fragments. After addition of enzyme-carrying beads, the sequence of the fragments is determined using pyrosequencing. In successive steps, the PicoTiter(TM)Plate and the beads as well as the enzyme beads therein are subjected to different deoxyribonucleotides in the presence of conventional sequencing reagents, and upon incorporation of a deoxyribonucleotide a light signal is generated which is recorded. Incorporation of the correct nucleotide will generate a pyrosequencing signal which can be detected.

Pyrosequencing itself is known in the art and described inter alia on www.biotagebio.com; www.pyrosequencing.com / section technology. The technology is further applied in e.g. WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, and WO 2005/003375 (all in the name of 454 Life Sciences). In the present invention, the beads are

preferably equipped with primer (binding) sequences or parts thereof that are capable of binding the amplicons, as the case may be. In other embodiments, the primers used in the amplification are equipped with sequences, for instance at their 5' -end, that allow binding of the amplicons to the beads in order to allow subsequent emulsion polymerization followed by sequencing. Alternatively the amplicons may be ligated with sequencing adaptors prior to ligation to the beads or the surface. The sequenced amplicons will reveal the identity of the identifier and thus of the presence or absence of the restriction fragment in the sample .

10

Illumina-Solexa technologies

One of the methods for high throughput sequencing is available from Illumina, United Kingdom (www.illumina.co.uk) and described inter alia in WO0006770, WO0027521, WO0058507, WO0123610, WO0157248, WO0157249, WO02061127, WO03016565, WO03048387, WO2004018497, WO2004018493, WO2004050915 , WO2004076692 , WO2005021786 , WO2005047301, WO2005065814, WO2005068656, WO2005068089 and WO2005078130. In essence, the method starts with adaptor-ligated fragments of genomic DNA. The adaptor-ligated DNA is randomly attached to a dense lawn of primers that are attached to a solid surface, typically in a flow cell. The other end of the adaptor ligated fragment hybridizes to a complementary primer on the surface. The primers are extended in the presence of nucleotides and polymerases in a so-called solid-phase bridge amplification to provide double stranded fragments. This solid phase bridge amplification may be a selective amplification.

25

Denaturation and repetition of the solid-phase bridge amplification results in dense clusters of amplified fragments distributed over the surface. The sequencing is initiated by adding four differently labeled reversible terminator nucleotides, primers and polymerase to the flow cell. After the first round of primer extension, the labels are detected, the identity of the first incorporated

30

bases is recorded and the blocked 3' terminus and the fluorophore are removed from the incorporated base. Then the identity of the second base is determined in the same way and so sequencing continues.

5 In the present invention, the adaptor ligated restriction fragments or the amplicons are bound to the surface via the primer binding sequence or the primer sequence. The sequence is determined as outlined, including the identifier sequence and the restriction fragment. Currently available Solexa technology allows for the sequencing of fragments of about 30 base pairs. By
10 smart design of the adaptors and the surface bound primers, the sequencing step reads through the sample identifier and the remains of the recognition sequence of the restriction endonuclease used. For example when an 3 bp sample identifier is used and the remains of the rare cutter EcoRI (GAACCT) are present, an internal sequence of the restriction fragment of 7 bp that can
15 be used to uniquely identify the restriction fragment in the sample. In a preferred embodiment based on the Illumina-Solexa sequencing technology above, the amplification of the adaptor ligated restriction fragments is performed with a primer that contains at most one selective nucleotide at its 3' end, preferably no selective nucleotides at its 3' end, i.e. the
20 primer is only complementary to the adaptor (a +0 primer) .

In alternative embodiments directed to the sequencing methods described herein, the primers used in the amplification may contain specific sections (as alternative to the herein described primer or primer binding sequences) that are used in the subsequent sequencing step to bind the adaptor-capped
25 restriction fragments or amplicons to the surface. These are generally depicted as the key region or the 5'-primer compatible sequence.

In one embodiment of the invention, the nucleic acid sample is digested with at least one restriction enzyme and at least one adaptor is ligated that comprises a recognition sequence for a type II restriction endonuclease. The subsequent
30 digestion of the adaptor-ligated restriction fragment with a type II's restriction

endonuclease yields, as the distance between the recognition and restriction site of a type II enzyme is relatively short (up to about 30 nucleotides), a shorter and a longer restriction fragment, to which a II restriction site compatible adaptor can be ligated. Typically, the overhang of the II-restricted site is unknown such that a set of adaptors may be used that are degenerated in the overhang. After (selective) amplification, the amplicons can be sequenced. The adaptor sequence in this embodiment generally can be described as: 5' - primer binding site - sample identifier sequence - degenerate type II cohesive end sequence - 3'. The associated PCR primer generally follows: primer sequence - sample identifier sequence - degenerate type II cohesive end sequence - selective nucleotides - 3'. The primer used to initiate the sequencing-by-synthesis then generally has the structure: 5' - primer binding site - 3'. A size selection step may be preferred after digesting with the II enzyme to remove the smaller fragments. As in this embodiment the remains of the restriction site are for this type of enzyme typically in the order of 2-4 bp, this results in combination with a 6 bp sample identifier in the sequencing of 15-17 bp of a restriction fragment.

Thus the method of the invention is perfectly suited to identify the regulatory genomic sequences of genes belonging to the transcriptome of a cell or organism, without any initial sequence information or prior genetic knowledge of the cell and/or of the organism from which the cell is derived. Accordingly, the promoter region, leader sequences and other 5' UTR regions, introns and exons, the 3' UTR sequences and terminator of an expressed gene can be identified according to the present method. Since no cloning step is involved, it is also feasible to determine the genomic sequences of genes that cause problems in cloning steps, e.g. genes that are toxic to the host organism, genes that code for regulatory proteins and/or genes that cause otherwise problems in cloning.

Further, it is possible on basis of this information to directly analyze all polymorphisms (including SNPs) associated with the alleles of an expressed gene, whether these polymorphisms occur in the coding sequence or in the non-coding sequence of the gene. Thus, it will be possible to detect aberrations in promoter sequences that cause a regulation of gene expression, it will be possible to detect mutants with polymorphisms in the introns that can cause different splice variants, and so on.

In order to increase a correct interpretation of the sequenced nucleic acid sequences and the differences found therein, automatic annotation can be performed on the sequenced fragments or contigs.

Similarly, the obtained sequence information can be used to compare the sequences with sequences from an EST library. In that way intron sequences or gene internal non-coding sequences, as well as promoter sequences and 3' and 5' UTRs can be identified. The EST library can be taken from the same organism or from a related species.

In a further aspect, the invention relates to kits with which the method of the invention can be performed. Such kits would comprise one or more adaptors and optionally one or more primers complementary to said adaptors, ligase, and/or restriction enzymes that are specific for cutting the adaptors, aside from conventional components for amplification kits per se, like dNTPs, polymerase etc. Further, the kit should provide instructions for use, wherein an instruction scheme for performing the method of the invention.

Furthermore, the present invention finds application in, amongst others, use of the method for the identification of molecular markers, for genotyping, bulk segregant analysis, genetic mapping, marker-assisted back-crossing, mapping of quantitative trait loci, linkage disequilibrium mapping, and determination of methylation patterns.

EXAMPLE

cDNA procedure**5 RNA isolation and cDNA synthesis**

Total RNA was isolated from skin of apple fruit, *Malus x domestica*, cultivar Kanzi, following the method of Chang *et al.* (1993). Apples were derived from 4 different orchards and picked at 5 different time-points in 2007 from early August until late September. Equal amounts of total RNA from
10 these 20 samples were pooled into one sample and purified with the RNeasy Plus Micro Kit (QIAGEN, Hilden, Germany, 74034) to remove genomic DNA contamination according to manufacturer's instructions.

First-strand cDNA synthesis was performed with 2 µg total RNA as input using the Mint cDNA synthesis kit (Evrogen, Moscow, Russia, SK001)
15 according to manufacturer's instructions. Double-stranded (ds) cDNA synthesis by PCR amplification was performed using the Mint cDNA synthesis kit for an optimal of 18 cycles. The obtained ds cDNA was purified using QIAquick® PCR Purification columns (QIAGEN, 28104) and the concentration was measured spectrophotometrically. The cDNA was analyzed on a 1%
20 agarose gel. The ds cDNA ranged from 200 to 2000 bp.

cDNA blunting, phosphorylation, concatenation and nebulization

The ds cDNA was blunted and phosphorylated using the Quick Blunting™ Kit (New England Biolabs, Ipswich, MA, USA, E1201S). 38 µl
25 cDNA (8 µg) was mixed with 5 µl 10 x Blunting Buffer, 5 µl 1 mM Deoxynucleotide Solution Mix and 2 µl Blunting Enzyme Mix and incubated at room temperature for 30 minutes, followed by 10 minutes at 70°C.

Subsequently, 48 µl of this cDNA blunting mixture was concatenated by mixing it with 10 µl 10 x T4 DNA Ligase Reaction Buffer, 5 µl T4 DNA Ligase
30 (both from New England Biolabs, M0202S, 400,000 U/ml), 25 µl 40% (w/v) polyethyleneglycol 8000 and 12 µl water. The ligation mixture was incubated for 2 hours at room temperature and concatenation was confirmed by agarose gel analysis. The T4 ligase was inactivated by incubation at 65°C for 10 minutes.

35 The concatenated cDNA was subjected to shearing by nebulization. The cDNA ligation mixture of 100 µl was mixed with 650 µl nebulization buffer (10 mM Tris-HCl, 1 mM EDTA, 50 % glycerol, pH 8.0) and pipetted into a nebulizer (Invitrogen, Paisley, UK, K7025-05). Nebulization was carried out following manufacturer's instructions for 15 minutes at 48 psi using nitrogen
40 gas 5.0 (Praxair, Danbury, CT, USA). After short centrifugation of the nebulizer, the collected nebulized cDNA was transferred to a microcentrifuge tube and precipitated by addition of 2 µl glycogen (Sigma-Aldrich, St. Louis, MO, USA, 20 mg/ml, G 1767), 0.1 volume of 3M sodium acetate pH5.2 and 1

volume of isopropanol, and incubation for 10 minutes at -80 °C. The cDNA was pelleted by centrifugation for 15 minutes at 20,800 g, washed with 70% ethanol, dried and dissolved in 50 µl 10 mM Tris-HCl, 1 mM EDTA, pH 8.0.

5 *cDNA sizing and blunting*

The sheared cDNA was incubated at 65°C for 10 minutes, gel loading buffer was added and the cDNA was divided over 5 slots of a 2% agarose gel in Tris-acetate (TAE) buffer (Sambrook *et al.*, 1989). After electrophoresis, cDNA fragments of 100-400 bp were isolated from the gel using the GenElute Gel
10 Extraction kit (Sigma-Aldrich, NA1111). A small sample of the purified cDNA was checked on gel and the concentration was found to be low. Therefore, the procedures above of cDNA blunting, phosphorylating, concatenation, nebulization and gel purification was repeated several times with a total of 24 µg ds cDNA obtained with the Mint cDNA synthesis kit. The cDNA fragments
15 of 100-400 bp were concentrated by ethanol precipitation and dissolved in 19 µl molecular biology-grade water.

The frayed ends of the sheared cDNA were blunted and phosphorylated by mixing with 2.5 µl 10 x Blunting Buffer, 2.5 µl 1 mM Deoxynucleotide Solution Mix and 1 µl Blunting Enzyme Mix from the Quick Blunting™ Kit
20 (New England Biolabs, E1201S) and incubation at room temperature for 30 minutes, followed by 10 minutes at 70°C. Subsequently, the cDNA was purified using the MinElute® PCR Purification Kit (QIAGEN, 28004).

Adaptor ligation and nick-repair

25 Adaptor ELTD-AdC was prepared by annealing of the partially complementary oligonucleotides ELTD-Primer-C (5'-AGTCCGTCGCATCGCTCTTC-3') and ELTD-AdC2 (5'-GAAGAGCGATGCGACG-3'). This adaptor is blunt on one side and has a 4 nt (AGTC) 5'-overhang on the other side to achieve directionality of ligation to the
30 cDNA and to prevent ligation of multiple adaptors to the cDNA. The ELTD-AdC adaptor also contains the rare cutting SspI restriction site:

GCTCTTCN/NNN
CGAGAAGNNNN/

This restriction site enables removal of ELTD-AdC from the cDNA during a
35 later step in the protocol. Adaptor ELTD-AdD was prepared by annealing of the partially complementary oligonucleotides ELTD-Primer-D (5'-Biotin-TEG-AGTGGGTGTCCTGGGTCAAC-3') and ELTD-AdD2 (5'-GTTGACCCAGGACACC-3'). This adaptor also has a 4 nt (AGTG) 5'-overhang on one side, which is labeled with biotin via a tetra-ethyleneglycol (TEG)
40 spacer arm. The biotin label will enable immobilization of the cDNA to streptavidin-coated beads during later steps in the protocol. All oligonucleotides were ordered HPLC-purified from Sigma-Aldrich and dissolved in 1 mM Tris-HCl, 0.1 mM EDTA, pH 8.0. The adaptors were prepared by mixing 50 µl of each of the appropriate oligonucleotide (400 µM)

with 100 μ l of 2 x annealing buffer (20 mM Tris-HCl, 100 mM NaCl, 2 mM EDTA, pH 7.6), incubating the mixture for 5 minutes at 95°C in a thermoblock (Thermomixer Compact, Eppendorf, Hamburg, Germany), then switching off the thermoblock allowing the samples inside to slowly cool down to less than 30°C (taking 3 hours). This yielded the double-stranded adaptors ELTD-AdC and ELTD-AdD at 100 μ M concentrations.

Both adaptors were ligated to the cDNA in the following reaction: 9.2 μ l cDNA from the MinElute purification column, 1.25 μ l water, 0.4 μ l adaptor ELTD-AdC (100 μ M), 0.4 μ l adaptor ELTD-AdD (100 μ M), 12.5 μ l 2 x Quick Ligation Reaction Buffer and 1.25 μ l Quick T4 DNA Ligase (Quick Ligation™ Kit, New England Biolabs, M2200S). The ligation mixture was incubated at 25°C for 20 minutes and purified using the GenElute PCR Clean-Up Kit (Sigma-Aldrich, NA1020).

The adaptor-ligated cDNA was nick-repaired in the following reaction: 47 μ l cDNA from the GenElute purification column, 8 μ l 10 x ThermoPol Reaction Buffer (New England Biolabs), 8 μ l 1 mg/ml BSA, 2 μ l 10 mM dNTPs, 1 μ l 8U/ μ l Bst DNA Polymerase, Large Fragment (New England Biolabs, M0275) and 14 μ l water. The nick-repair reaction was incubated for 30 minutes at 65°C and purified using a QIAquick PCR purification column yielding 50 μ l 100-400 bp adaptor-ligated cDNA.

Amplification of adaptor-ligated cDNA

The cDNA was amplified with a high-fidelity DNA polymerase in a PCR reaction containing: 10 μ l cDNA from the QIAquick PCR purification column, 10 μ l 5 x Phusion™ HF buffer, 1 μ l 10 mM dNTPs, 2.5 μ l 10 μ M ELTD-Primer-C, 2.5 μ l 10 μ M ELTD-Primer-D, 0.5 μ l 2U/ μ l Phusion Hot Start DNA Polymerase (Finnzymes, Espoo, Finland, F-540) and 23.5 μ l water. First, a test was carried out to determine the optimal number of PCR cycles for cDNA amplification. The reaction mixture was placed in a thermal cycler, denatured at 98°C for 30 sec and subsequently subjected to 5 cycles of denaturation-annealing-elongation: 5 sec at 98°C, 10 sec at 60°C, 15 sec at 72°C. After this, 5 μ l was removed from the reaction mixture and kept on ice (sample after 5 cycles). The remaining reaction mixture was subjected to three more PCR cycles as above and 5 μ l was removed and kept on ice (sample after 8 cycles). The above was repeated 5 more times until a total of 23 cycles had been reached. The 5 μ l samples of 5, 8, 11, 14, 17, 20 and 23 cycles were analyzed on a 1.5% agarose gel. The optimum number of cycles was determined to be 17 cycles, after that a plateau was reached as more cycles led to the appearance of a smear above the expected size of the cDNA. To make more cDNA, two PCR reaction mixtures each with 10 μ l cDNA were prepared as described above. The reaction mixtures were placed in a thermal cycler, denatured at 98°C for 30 sec and subsequently subjected to 17 cycles of denaturation-annealing-elongation: 5 sec at 98°C, 10 sec at 60°C, 15 sec at 72°C. A final extension step of 5 min at 72°C followed. The amplified cDNA was purified using a QIAquick

PCR purification column, followed by a GenElute PCR Clean-Up column (Sigma-Aldrich) to remove primers and possible primer-dimers.

Isolation of single-stranded CD-adapted cDNA

5 Next, cDNA obtained in the previous step was enriched for molecules carrying ELTD-Ad-C on one end and ELTD-Ad-D on the other end by binding to streptavidin-coated beads, washing and alkaline elution. cDNA molecules carrying ELTD-Ad-C on both ends (from now on called CC molecules) cannot bind to streptavidin and are washed away from the beads. cDNA molecules
10 carrying ELTD-Ad-D on both ends (from now on called DD molecules) will stay bound to the beads during alkaline elution, since both strands are biotinylated. cDNA molecules carrying ELTD-Ad-C on one end and ELTD-Ad-D on the other end (from now on called CD molecules) will stay bound to the beads with one biotinylated strand, the other non-biotinylated strand will be eluted by
15 treatment with NaOH.

Streptavidin-coated paramagnetic Dynabeads® M-270 (Invitrogen, 653.05) were thoroughly resuspended and 50 µl (corresponding to 0.5 mg) beads was transferred to a siliconized microcentrifuge tube (Sigma-Aldrich, T4816). Beads were washed three times with 100 µl 1 x B&W buffer (5 mM
20 Tris-HCl, 0.5 mM EDTA, 1 M NaCl, pH 7.5) using a Dynal magnetic stand (MPC®-E-1, Invitrogen) for separation of the beads following manufacturer's instructions. Beads were resuspended in 100 µl 2 x B&W buffer containing 0.02% Tween-20 to decrease aspecific binding. Next, 45 µl of PCR-amplified and purified cDNA mixed with 55 µl of water was added to the bead
25 suspension. cDNA was allowed to bind for 15 min at room temperature with gentle rotation of the tube. The tube was placed in the magnetic stand to separate the beads from the supernatant, which was transferred to a fresh tube. This fraction is called AB (after binding) and contains cDNA which was not bound to the beads. Subsequently, the bead pellet was washed as follows:
30 once with 200 µl 2 x B&W buffer containing 0.02% Tween-20, once with 500 µl 2 x B&W buffer containing 0.02% Tween-20 and twice with 500 µl water. After resuspension of the beads in water for the second time, beads were transferred to a fresh siliconized tube before transfer to the magnet. Finally, the beads were resuspended in 250 µl freshly prepared 0.1 M NaOH and the tube was
35 gently rotated for 2-3 minutes. The supernatant represents the first elution. The beads were again resuspended in 250 µl 0.1 M NaOH and gently rotated for 2-3 minutes, this supernatant represents the second elution. Elutions were each mixed with 1250 µl PBI buffer (QIAquick PCR purification kit) and 7.2 µl 20% acetic acid and purified over QIAquick PCR purification columns. Also,
40 the AB fraction was purified over a QIAquick PCR purification column. Remaining beads were washed once with 200 µl water, once with 200 µl 10 mM Tris-HCl, 1 mM EDTA, pH 8.0, once again with 200 µl water and finally resuspended in 50 µl water and stored at 4 °C.

Five microliter of the AB fraction and both alkaline elutions were checked on a 1.5% agarose gel. cDNA was found in the AB fraction and the first alkaline elution, but not in the second alkaline elution, which was discarded. Control PCR reactions were carried out on 1 μ l of the AB fraction, first alkaline elution and the beads (50 μ l total volume each). Each was mixed with 12.5 μ l REDTaq[®] ReadyMix[™] (Sigma-Aldrich, R2523), 1 μ l 10 μ M ELTD-Primer-C or 1 μ l 10 μ M ELTD-Primer-D or [1 μ l 10 μ M ELTD-Primer-C and 1 μ l 10 μ M ELTD-Primer-D] and water to a total volume of 25 μ l. PCR conditions were: 1 min@94°C, (30 sec@94°C, 30 sec@50°C, 30 sec@72°C) for 6, 9, 12 and 15 cycles, 5 min@72°C. Of each reaction 5 μ l was loaded on a 1.5% agarose gel. The results showed that there were more CC than DD and CD-molecules in the AB fraction as expected, because CC molecules cannot bind to the beads. In the alkaline elution and the bead fraction the occurrence was: CC < DD < CD. The conclusion was that there was enrichment for CD molecules in the alkaline eluate as expected, but that CC and especially DD molecules were also still present, probably due to aspecific binding of CC and incomplete binding of DD molecules to the beads. The alkaline elution (QIAquick-purified) is called the single-stranded CD-enriched cDNA.

20 **Amplification of the CD-enriched cDNA**

After testing the optimal PCR conditions, the single-stranded CD-enriched cDNA was amplified as follows. Sixteen PCR reactions were set up each containing: 0.5 μ l of the QIAquick column-purified first alkaline elution of the Dynabeads described above, 10 μ l 5 x Phusion[™] HF buffer, 1 μ l 10 mM dNTPs, 2.5 μ l 10 μ M ELTD-Primer-C, 2.5 μ l 10 μ M ELTD-Primer-D, 0.5 μ l 2U/ μ l Phusion Hot Start DNA Polymerase (Finnzymes, F-540) and 33 μ l water. The reaction mixtures were placed in a thermal cycler, denatured at 98°C for 30 sec and subsequently subjected to 11 cycles of denaturation-annealing-elongation: 5 sec at 98°C, 10 sec at 60°C, 15 sec at 72°C. A final extension step of 5 min at 72°C followed. The amplified cDNA was purified using three parallel QIAquick PCR purification columns. The purified cDNA was analyzed on a 1.2% agarose gel and the concentration was measured spectrophotometrically. A total of 27.5 μ g of double-stranded CD-enriched cDNA was obtained.

35

Binding of the CD-enriched cDNA to Dynabeads

Five microgram of the double-stranded CD-enriched cDNA from the previous step was used to bind to Dynabeads M-270. The procedure is described above under 'Isolation of single-stranded CD-adapted cDNA' with the following modifications. An amount of 27.32 μ l corresponding to 5 μ g of the CD-enriched cDNA was mixed with water to a volume of 100 μ l in total, and this mixture was added to the beads in 100 μ l 2 x B&W buffer containing 0.02% Tween-20. After binding of the cDNA and washing with 2 x B&W buffer containing 0.02% Tween-20 and washing with water, the

40

beads were washed 2 times with 200 µl 1 x NEBuffer 4 (New England Biolabs). Finally, the beads with the bound cDNA were resuspended in 100 µl 1 x NEBuffer 4 and transferred to a new siliconized microcentrifuge tube.

5 **Digestion of the CD-enriched cDNA-bead preparation with SapI**

The CD-enriched cDNA on the beads was digested with SapI to remove adaptor ELTD-AdC from the cDNA molecules, while the cDNA remained attached to the beads via biotinylated adaptor ELTD-AdD. Five microliters
10 of Sap I (2 U/µl, New England Biolabs, R0569) was added to the cDNA-bead suspension and incubated for 1.5 hour at 37°C. The beads were vortexed at 1400 rpm every 10 minutes to keep them in suspension during this step. Next, the beads were placed in the magnetic stand for 1 min to separate the beads, the supernatant was discarded and the beads were washed two times
15 with 500 µl 2 x B&W buffer containing 0.02% Tween-20 and two times with 500 µl water.

Alkaline elution to prepare the single-stranded CD-enriched cDNA-bead library

20 The bead pellet was resuspended in 250 µl 0.1 M NaOH (freshly prepared) and the tube was gently rotated for 2-3 minutes. The tube was put into the magnetic stand for 1min and the supernatant (= alkaline eluate) was transferred to a fresh tube. The alkaline eluate was mixed with 1250 µl PBI buffer (QIAquick PCR purification kit) and 7.2 µl 20% acetic acid and purified
25 over a QIAquick PCR purification column. The remaining beads were washed once with 200 µl water, once with 200 µl 10 mM Tris-HCl, 1 mM EDTA, pH 8.0, once again with 200 µl water and finally resuspended in 50 µl water and stored at 4 °C. This is the single-stranded CD-enriched cDNA-bead library, ready for hybridization with genomic DNA.

30 The purified alkaline eluate was analyzed on a 1.2% agarose gel together with the after binding fraction from 'Binding of the CD-enriched cDNA to Dynabeads' and a dilution series of double-stranded cDNA of known concentration from 'Amplification of adaptor-ligated cDNA'. Approximately 2 µg of the cDNA was found in the AB fraction and was not bound to the
35 Dynabeads. The alkaline eluate showed a smear of the expected size. The concentration of the alkaline eluate was measured spectrophotometrically and it was found that 410 ng of single-stranded cDNA had been eluted from the beads. In theory, an equal amount of the complementary cDNA strands should be in the single-stranded CD-enriched cDNA-bead library, corresponding to
40 approximately 4 pmol assuming an average size of 300 nt.

Genomic DNA

Isolation of genomic DNA

5 Genomic DNA (gDNA) was isolated from Kanzi leaves following the protocol of Kobayashi *et al.* (1998). After RNase treatment according to Kobayashi *et al.* (1998), the gDNA was precipitated with high-salt to remove impurities by addition of two-third volume 5M NaCl and two volumes ethanol (p.a.), followed by centrifugation for 15 minutes at 20.000 g, washing the pellet
10 with 70% ethanol, drying and dissolving the pellet in 10 mM Tris-HCl, 1 mM EDTA, pH 8.0.

Restriction enzyme digestion of the gDNA

The gDNA was fragmented by restriction enzyme digestion to create
15 non-overlapping fragments. Overlapping fragments may interfere with the hybridization step later in the protocol, leading to networks of hybridizing fragments. Restriction digests were chosen which yielded mostly 1-3 kb fragments. The gDNA was digested with HindIII/BstYI and with EcoRI/BstYI to create two sets of different fragments. Twenty micrograms of gDNA was
20 digested by addition of 10 µl NEBuffer 2, 3 µl EcoRI (New England Biolabs, 20U/µl, R0101) or 1 µl HindIII (New England Biolabs, 100U/µl, R0104) and water until a total volume of 100 µl, followed by incubation at 37°C for 1 hour. Subsequently, 6 µl BstYI (New England Biolabs, 10U/µl, R0523) was added to each tube, followed by incubation at 60°C for 1 hour. The digested DNA's were
25 each loaded in 4 lanes of a 1% agarose gel and separated. Fragments between 1 and 3 kb were cut from gel and purified using the GenElute Gel Extraction Kit (Sigma-Aldrich, NA1111). The above procedure was repeated one time to generate sufficient DNA fragments. The EcoRI/BstYI (EB) and HindIII/BstYI
30 (HB) genomic fragments were ligated to adaptors ELTD-AdE-Eco, ELTD-AdE-Hind and ELTD-AdF-Bst.

Adaptor ligation to the gDNA

Adaptor ELTD-AdE-Eco was prepared by annealing of the partially
complementary oligonucleotides ELTD-AdE-Eco1 (5'-
35 CTTGTAGGGCACGGGTTCGAGAG-3') and ELTD-AdE-Eco2 (5'-
AATTCTCTCGACCCGTGCCCTA-3'). This adaptor has a 5'-AATT overhang on one side, which is compatible to the EcoRI-overhang of the gDNA fragments and a 5'-CTTG overhang on the other side. These overhangs achieve
40 directionality of ligation to the gDNA and prevent ligation of multiple adaptors to the gDNA. Adaptor ELTD-AdE-Hind was prepared by annealing of the partially complementary oligonucleotides ELTD-AdE-Hind1 (5'-
CTTGTAGGGCACGGGTTCGGAGA-3') and ELTD-AdE-Hind2 (5'-
AGCTTCTCCGACCCGTGCCCTA-3'). Analogous to ELTD-AdE-Eco, the ELTD-AdE-Hind adaptor has a HindIII-compatible 5'-AGCT overhang on one

side and a 5'-CTTG overhang on the other side. Adaptor ELTD-AdF-Bst was prepared by annealing of the partially complementary oligonucleotides ELTD-AdF-Bst1 (5'-GAATGGCTGGGAGAGTGCTGAG-3') and ELTD-AdF-Bst2 (5'-GATCCTCAGCACTCTCCCAGCC-3'). Analogous to ELTD-AdE-Eco, the
5 ELTD-AdF-Bst adaptor has a BstYI-compatible 5'-GATC overhang on one side and a 5'-GAAT overhang on the other side. All oligonucleotides were ordered HPLC-purified from Sigma-Aldrich and dissolved in 1 mM Tris-HCl, 0.1 mM EDTA, pH 8.0.

The adaptors were prepared by mixing 15 µl of each of the appropriate
10 oligonucleotide (800 µM) with 60 µl of 2 x annealing buffer (20 mM Tris-HCl, 100 mM NaCl, 2 mM EDTA, pH 7.6) and 30 µl water and incubating the mixture for 5 minutes at 95°C in a thermoblock, then switching off the thermoblock allowing the samples inside to slowly cool down to less than 30°C (taking 3 hours). This yielded the double-stranded adaptors ELTD-AdE-Eco,
15 ELTD-AdE-Hind and ELTD-AdF-Bst at 100 µM concentrations.

Adaptors ELTD-AdE-Eco and ELTD-AdF-Bst were ligated to the EcoRI/BstYI (EB) 1-3 kb gDNA fragments in the following reaction: 1.3 µg EB fragments, 0.4 µl adaptor ELTD-AdE-Eco (100 µM), 0.4 µl adaptor ELTD-AdF-Bst (100 µM), 40 µl 2 x Quick Ligation Reaction Buffer, 4 µl Quick T4 DNA
20 Ligase (Quick Ligation™ Kit, New England Biolabs, M2200S) and water to a total volume of 80 µl. Adaptors ELTD-AdE-Hind and ELTD-AdF-Bst were ligated to the HindIII/BstYI (HB) 1-3 kb gDNA fragments in the following reaction: 1.0 µg HB fragments, 0.4 µl adaptor ELTD-AdE-Hind (100 µM), 0.4 µl adaptor ELTD-AdF-Bst (100 µM), 40 µl 2 x Quick Ligation Reaction Buffer,
25 4 µl Quick T4 DNA Ligase (Quick Ligation™ Kit, New England Biolabs, M2200S) and water to a total volume of 80 µl. The ligation mixtures were incubated at 25°C for 20 minutes and purified using the GenElute PCR Clean-Up Kit (Sigma-Aldrich, NA1020).

30 Nick-repair and purification of EB and HB gDNA

The adaptor-ligated gDNA's were nick-repaired in the following reaction: 40 µl gDNA from the GenElute purification column, 8 µl 10 x ThermoPol Reaction Buffer, 8 µl 1 mg/ml BSA, 2 µl 10 mM dNTPs, 3 µl 8U/µl Bst DNA Polymerase, Large Fragment (New England Biolabs, M0275) and 19
35 µl water. The nick-repair reaction was incubated for 30 minutes at 65°C and purified using the GenElute PCR Clean-Up Kit (Sigma-Aldrich, NA1020). This yielded 50 µl adaptor-ligated gDNA of 0.02 µg/µl (EB) and 0.014 µg/µl (HB), ready for hybridization with the single-stranded CD-enriched cDNA-bead library.

40 The adaptor-ligation step was checked by PCR. One nanogram of the adaptor-ligated and nick-repaired EB preparation was used as template in PCR reactions using either primer ELTD-AdE-Eco1 or ELTD-AdF-Bst1 or a combination of both (E, F, EF respectively). Analogously, the adaptor-ligated HB preparation was used as template in PCR reactions using either primer

ELTD-AdE-Hind1 or ELTD-AdF-Bst1 or a combination of both (E, F, EF respectively). For EB and HB, the PCR reactions gave a smear in the region of 1-3 kb as expected.

5 Hybridization and amplification of selected genomic DNA fragments

Hybridization of cDNA-bead library with gDNA fragments

The hybridization conditions were tested first on firefly luciferase (Luc) gene fragments. Briefly, a single-stranded 200 nt Luc fragment (Luc200) with ELTD-AdC and ELTD-AdD was bound to Dynabeads M-270 via the biotin label of ELTD-AdD. This Luc-bead preparation was hybridized with a 1600 nt Luc fragment ligated with ELTD-AdE and -AdF mixed with an aspecific 1400 nt control DNA fragment. After hybridization and washing, fragments bound to the Luc200 probe were eluted by alkaline treatment and amplified by PCR. It was found that the 1600 nt Luc fragment was eluted at much higher concentration than the aspecific, non-hybridizing control fragment. There was a 24 cycle difference between the appearance of these fragments during PCR ($2^{24} = 1.7 \times 10^7$ fold enrichment of Luc1600 assuming 100% PCR efficiency). The same hybridization and washing conditions were used for the cDNA-bead library and gDNA.

The EB and HB gDNA preparations described above, were denatured by heating. First, 13 μ l of EB and 12 μ l of HB were reduced in a centrifugal vacuum concentrator to 5 μ l. The amount of EB corresponds to 0.26 μ g or approximately 0.2 pmol of on average 2 kb fragments. The amount of HB corresponds to 0.17 μ g or approximately 0.13 pmol of on average 2 kb fragments. The samples were then denatured in a thermal block for 5 min at 95°C and directly placed on ice.

The single-stranded CD-enriched cDNA-bead library described above was divided over two siliconized tubes (25 μ l beads each). The beads were washed three times with 200 μ l 6 x SSC/0.1% SDS (pre-warmed at 60°C) and then resuspended in 100 μ l 6 x SSC/0.1% SDS and kept at 60°C. To one tube the EB gDNA was added, to the other tube the HB gDNA was added. The tubes were incubated for 4 hours at 60°C with gentle rotation. The beads were washed two times quickly with 500 μ l pre-warmed 6 x SSC/0.1% SDS (60°C), three times 15 min with 500 μ l 6 x SSC/0.1% SDS at 60°C with rotation, two times 5 min with 500 μ l 6 x SSC at room temperature with rotation and once quickly with 500 μ l water.

40 Elongation of hybrids and PCR amplification of selected gDNA fragments

The formed cDNA/gDNA hybrids on the beads were used for elongation of the 3'cDNA ends using hybridized gDNA as template. After the final water

5 wash step, the beads were directly resuspended in a mixture of 77 μ l water and 20 μ l 5 x PhusionTM HF buffer and transferred to a new siliconized tube. Next were added: 2 μ l 10 mM dNTPs and 1 μ l 2U/ μ l Phusion Hot Start DNA Polymerase (Finnzymes, F-540). Elongation was at 72°C for 2 min. The beads were kept overnight at 0°C.

The beads were used for PCR amplification with different primer combinations. The beads hybridized with EB gDNA were amplified with the following primer sets:

- 1) ELTD-AdE-Eco1 and ELTD-AdF-Bst1
- 10 2) ELTD-AdE-Eco1
- 3) ELTD-AdF-Bst1.

The beads hybridized with HB gDNA were amplified with the following primer sets:

- 15 4) ELTD-AdE-Hind3 and ELTD-AdF-Bst1
- 5) ELTD-AdE-Hind3
- 6) ELTD-AdF-Bst1.

ELTD-Primer E-Hind3 has the following sequence:

5'-GTAGGGCACGGGTCTGGAGAAGC-3'. It is identical to most part of ELTD-AdE-Hind1 and contains 3 extra nt at the 3'end (AGC) to fit on the HindIII site between the ligated adaptor and the gDNA and to confer stronger 3'end binding to the target during annealing.

The PCR reactions included 1 μ l of EB or HB beads, 10 μ l 5 x PhusionTM HF buffer, 1 μ l 10 mM dNTPs, 2.5 μ l of each primer (10 μ M), 0.5 μ l 2U/ μ l Phusion Hot Start DNA Polymerase (Finnzymes, F-540) and water to a final volume of 50 μ l. To determine the optimal cycle number, the reaction mixtures were divided over five 0.2 ml PCR tubes (10 μ l each), overlaid with mineral oil, placed in a thermal cycler, denatured at 98°C for 30 sec and subsequently subjected to 11, 14, 17, 20 and 23 cycles of denaturation-annealing-elongation: 5 sec at 98°C, 10 sec at 60°C, 2 min at 72°C. A final extension step of 5 min at 72°C followed. Of each reaction, 2.5 μ l was analyzed on a 1.2% agarose gel.

In the PCR reactions with ELTD-AdE and/or ELTD-AdF primers, smears were found of fragments ranging from 1-3 kb as expected for gDNA fragments. The combination of both primers gave PCR products for EB and HB samples already at 17 cycles. With ELTD-AdE alone, products were found after 23 cycles, whereas the same intensity of products was found with ELTD-AdF primers after 20 cycles. This is due to the presence of more BstYI-restriction sites (recognition site: RGATCY) in the genome than EcoRI- or HindIII-sites (GAATTC or AAGCTT).

40

Amplification of hybrid-selected gDNA for high-throughput sequencing

The PCR reaction with HB gDNA in combination with primers ELTD-AdE-Hind3 and ELTD-AdF-Bst1 and 20 cycles gave the best result: a

homogeneous smear ranging from 1-3 kb. The following PCR reaction mixture was prepared twenty-fold to raise sufficient DNA for high-throughput sequencing: 1 µl of HB beads, 10 µl 5 x Phusion™ HF buffer, 1 µl 10 mM dNTPs, 2.5 µl ELTD-AdE-Hind3 (10 µM), 2.5 µl ELTD-AdF-Bst1 (10 µM), 0.5 µl 2U/µl Phusion Hot Start DNA Polymerase and water to a final volume of 50 µl. The reaction mixture was placed in a thermal cycler, denatured at 98°C for 30 sec and subsequently subjected to 20 cycles of denaturation-annealing-elongation: 5 sec at 98°C, 10 sec at 60°C, 2 min at 72°C. A final extension step of 5 min at 72°C followed. This so-called hybrid-selected and amplified HB gDNA was purified using two parallel QIAquick PCR purification columns. The yield, which was measured spectrophotometrically, was 14.8 µg. Part of the hybrid-selected HB gDNA was subjected to high-throughput sequencing. Another part was used for cloning and Sanger sequencing.

Amplification of cDNA for high-throughput sequencing

To raise sufficient CD-enriched cDNA for high-throughput sequencing, the following PCR reaction mixture was prepared twenty-fold: 0.5 µl of the QIAquick column-purified first alkaline elution of the Dynabeads described under 'Isolation of single-stranded CD-adapted cDNA', 10 µl 5 x Phusion™ HF buffer, 1 µl 10 mM dNTPs, 2.5 µl 10 µM ELTD-Primer-C, 2.5 µl 10 µM ELTD-Primer-D-NB, 0.5 µl 2U/µl Phusion Hot Start DNA Polymerase (Finnzymes, F-540) and 33 µl water. ELTD-Primer D-NB has the same sequence as ELTD-Primer D, but does not contain a biotin label at the 5' end. The reaction mixtures were placed in a thermal cycler, denatured at 98°C for 30 sec and subsequently subjected to 11 cycles of denaturation-annealing-elongation: 5 sec at 98°C, 10 sec at 60°C, 15 sec at 72°C. A final extension step of 5 min at 72°C followed. The amplified cDNA was purified using four parallel QIAquick PCR purification columns. The purified cDNA was analyzed on a 1.2% agarose gel and the concentration was measured spectrophotometrically. A total of 29 µg of double-stranded CD-enriched cDNA was obtained. The cDNA was subjected to high-throughput sequencing.

Sequence analysis of captured genomic DNA molecules.

After capturing genomic DNA with cDNA beads, the beads with cDNA-gDNA hybrids were used for PCR amplification using the genomic DNA specific primer set ELTD-AdE-Hind3 and ELTD-AdF-Bst1. One part of the genomic DNA was cloned and 800 clones were submitted to Sanger sequencing (mean read length 781 bp). Another part of the genomic DNA was subjected to a high-throughput sequencing technology producing sequence reads having a mean length of 337 bp.

Both sequence datasets consisting of 800 individual Sanger sequences and 1370 high throughput sequence reads were analyzed in detail and compared with all available apple ESTs in a public data bank (NCBI, 262.411 entries on March 9, 2009). The program BLASTN (S. F. Altschul et al. NAR 25:3389-

3402, 1997) was used for global alignment of the genomic DNA sequences with the apple EST sequences.

From 800 Sanger sequences 488 showed significant similarity (e value 10^{-10}) with sequences from the apple EST collection in the public database and from
5 the random set of 1370 genomic DNA sequences, 765 exhibited significant similarity (e value 10^{-10}) showing that specific capturing of cDNA-related genomic DNA was successful. In addition, local alignment of these genomic DNA sequences revealed that these sequences frequently extend from the 5' or
10 3' cDNA boundaries or intron regions, showing that the technique enables identification of non-coding areas. If one takes into account that fragments have only been sequenced from one side, that the size of fragments varies between 1 Kb and 3 Kb and that more than 50% of the sequences have identifiable similarity with EST sequences, it seems obvious that a much
15 higher percentage of sequences matching would have been found if fragments would have been sequenced completely and built into contigs before alignment with public EST sequence data.

Since the size of the Apple genome is around 750 Mb and the number of genes in a plant genome is around 30000 with an average length of 1500 bp yielding
20 45 Mb of coding sequences on the genome, this means that by random sequencing only 6 percent of the genomic fragments could be annotated, whereas with the described method we find over 50 percent annotation.

Three cases of genomic DNA fragments spanning EST boundaries are described from the above example.
25

Description three Cases for Example.

Three cases of apple genomic sequences which were found with the method of
30 the invention are shown in Figures 4 and the sequence alignments below.

Case 1.

Apple genomic sequence FRA8S6E02IN5QW (sequence, see below) showed 98.9% identity with part of apple contig cg8984st_846_7, built from 7 EST
35 sequences (GenBank Acc: CO899363, CO419003, CO052855, CO752637, CO901846, CN927506 and CO066317).

The best BLASTX (S. F. Altschul et al. NAR 25:3389-3402, 1997) hit with contig cg8984st_846_7 of *Arabidopsis thaliana* was gene AT1G70160 (genome locus tag, The Arabidopsis Information Resource, <http://www.arabidopsis.org>)
40 with e-value $2e^{-104}$. Alignment of the three sequences showed that FRA8S6E02IN5QW contains part of the protein coding sequence (CDS) of a protein with unknown function, which partially overlaps with contig cg8984st_846_7. This is shown schematically in Figure 4, panel A and the sequence alignment of relevant parts of the sequences are shown below. The
45 protein coding sequence of FRA8S6E02IN5QW is also 77.8% identical with

AT1G70160, but is interrupted at the 5'-end of the sequence by an intron at the same position as an intron in AT1G70160 (Figures 4, panel A and sequence alignment). The identity between FRA8S6E02IN5QW and AT1G70160 in this region is low, 37.6%. In AT1G70160 the intron is smaller than in apple (the 5'-boundary of this intron is indicated in bold face in the sequence alignment). In FRA8S6E02IN5QW, there are stop codons in all three reading frames of the intron sequence and BLAST searches indicate that the CDS of FRA8S6E02IN5QW is homologous to similar proteins of other plant species, but that homology is lost in the intron region. This indicates that FRA8S6E02IN5QW is indeed a genomic sequence, containing an interrupting intron sequence. Also here the much lower conservation of intron sequences indicates that intron sequences contain more SNP's than exon sequences.

Case 2.

Apple genomic clone FRA8SE02HOH39 (sequence, see below) showed homology to apple contig cg12357st_1428_21, built from 21 EST sequences (GenBank Acc: CN930585, CV525017, CN873920, EB149394, EB121634, EB116211, CN909797, EB115871, EB154300, CN877800, EB121026, CN932122, CN860924, EB110988, CO865849, CN488473, CN497072, CN903918, CN903403, DR996731 and CN894330). Contig cg12357st_1428_21 contains the complete coding sequence for a protein with similarity to ATP-dependent Clp protease proteolytic subunit (best BLASTX hit EEF49880, *Ricinus communis*, score $3e^{-128}$). FRA8SE02HOH39 overlaps with the 3'-untranslated region of cg12357st_1428_21, and extends beyond the polyA tail into the non-transcribed genomic DNA as is shown in Figure 4. panel B and the sequence alignment below. Therefore, FRA8SE02HOH39 is an example of a genomic sequence containing the 3'-flanking sequences of a transcribed region.

Case 3.

Apple genomic sequence 02-H03 (sequence, see below) overlapped with the 5'-part of 40 apple ESTs, with similarity to an ethylene response factor (best BLASTX hit AAV66332, *Cucumis sativus*, score $6e^{-34}$). This is schematically shown in Figure 4, panel C, for EST 91044950 (GenBank Acc: EB155368). The other ESTs had similar 5'ends. The alignment of a relevant part of the sequences is shown below. The genomic sequence extended 744 bp upstream of the translation start codon (ATG) of the gene. No sequence similarity was found to any gene outside the coding region in 02-H03, confirming that this is a genomic sequence. A conserved sequence, upstream of the transcription start is the TATA box. A putative TATA-box (TATAAA) was found 39 bp upstream of the start of the EST.

References

- Chang, S., Puryear, J. and Cairney J. 1993. A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* 11: 113-116.
- 5 Kobayashi, N., Horikoshi, T., Katsuyama, H., Handa, T. and Takayanagi, K. 1998. A simple and efficient DNA extraction method for plants, especially woody plants. *Plant Tissue Culture and Biotechnology* 4: 76-80.
- 10 Sambrook, J., Fritsch, E.F. and Maniatis T. 1989. *Molecular cloning. A laboratory manual*. Second edition. Cold Spring Harbor Laboratory Press. USA.

Case 2.

```

5  FRA856E02HOH39A  CAGCGACCTGTTTACGTTGCAAGGTTTGGATGGAAGAGATTGAACAGTGAATGGCCAGCAAAATCGAATTCAGAGGGTCCAAAACAAATGCT
   CG12357ST_1428_21  CAGCGACCTGTTTACGTTGCAAGGTTTGGATGGAAGAGATTGAACAGTGAATGGCCAGCAAAATCGAATTCAGAGGGTCCAAAACAAATGCT

10 FRA856E02HOH39A  ACTTTTGGACAGTTGAAGAATGACCGACACTTCTTATATTAATCTTGGTTAAGTTTTCAGGACGGAAATCCTCGTCTTTTTCCT
   CG12357ST_1428_21  ACTTTTGGACAGTTGAAGAATGACCGACACTTCTTATATTAATCTTGGTTAAGTTTTCAGGACGGAAATCCTCGTCTTTTTCCT

   ATCAAAAAGAAAGAAAGCGTGAAGTTTATGTTTGTGATGCTTGTGATGGATGACCTAGTGTCAAGTGAAAAAATTCGACGGACAAAACCGTTGGCAATCCA
   ATGAAAAAGAAAGAAAGCGTGAAGTTTATGTTTGTGATGCTTGTGATGGATGACCTAGTGTCAAGTGAAAAA
   polyA

15 FRA856E02HOH39A  ATTTGTGCCCTGTATAAATGTCCTCACGTCCAAACGAGTTTCACATCGAAGAAA
   CG12357ST_1428_21

```

Legend (alignment case 2).

Alignment of apple genomic sequence FRASSE02HOH39 with apple EST contig cg12357st_1428_21. Only relevant parts of the sequences are shown. Matches between the two sequences are shown in grey. The polyA-tail of the cDNA sequence is indicated. The genomic sequence extends beyond the 3'-untranslated region of the cDNA.

Claims

1. A method for the identification of genomic DNA in a sample, comprising
- 5 - isolation of mRNA from a selected organism and preparing from said mRNA small single stranded cDNA fragments with one adaptor containing an affinity label;
- isolation of genomic DNA from the same or a related organism and preparing from said genomic DNA single stranded genomic DNA fragments ligated to
- 10 adaptor molecules
- hybridization of said single stranded genomic DNA fragments with said single stranded cDNA fragments and amplification of said hybrids; and
- high throughput sequencing of said hybrids.
- 15 2. Method according to claim 1, comprising the steps of:
- a) isolation and purification of mRNA from tissue samples of an organism;
- b) synthesis of cDNA using said mRNA as a template;
- c) optionally complexity reduction of said cDNA;
- d) fragmentation of said cDNA;
- 20 e) optionally size selection of said fragments;
- f) optionally removing polyA-containing fragments by binding to streptavidin-coated affinity beads;
- g) polishing of said fragments of cDNA;
- h) ligation of said fragments with one adaptor comprising a recognition site for
- 25 a rare restriction enzyme and another adaptor containing a biotin label;
- i) optionally size selection of said fragments;
- j) nick repair of said fragments;
- k) selection of said fragments that contain both adaptors sequences;
- l) amplification of said fragments using primers annealing to the adaptor
- 30 sequences described in step h, wherein one primer is complementary to the

- adaptor with a rare restriction site and the other primer contains a biotin label.
- m) binding said fragments to streptavidin-coated affinity beads;
 - n) removing adaptors containing the rare restriction site using the
 - 5 corresponding restriction enzyme from said fragments;
 - o) removing single strands not attached to affinity beads by a biotin-streptavidin interaction from double stranded DNA fragments attached to affinity beads via a biotin-streptavidin interaction resulting in single strands of DNA bound to streptavidine affinity beads;
 - 10 p) isolation and purification of genomicDNA for example from the organism of step a
 - q) fragmentation of said genomic DNA;
 - r) optionally polishing of said genomic DNA;
 - s) ligation of said genomic DNA with one single type of adaptor or with two
 - 15 different type of adaptors. (preferred)
 - t) melting out of said genomic DNA into single stranded DNA
 - u) hybridizing genomic DNA from step t) with cDNA on beads from step o);
 - v) remove unbound genomic DNA by washing
 - w) extension of the cDNA-genomic DNA hybrid by a polymerase to create a
 - 20 double-stranded template
 - x) performing PCR on said genomic DNA-cDNA hybrid
 - y) selection of fragments larger than about 100 basepairs from said PCR by size fractionation
 - z) optionally purification of said fragments
 - 25 aa) high throughput sequencing of said fragments .

3. Method for identifying polymorphisms, comprising all the steps of the method according to claim and additionally

- ab) comparing sequence data from two or more samples to identify

polymorphisms.

4. Method according to claim 2 or 3, wherein the sequences
from step aa) are combined into contigs of overlapping individual
5 sequences.

5. Method according to any of claim 2-4, wherein the
sequences from step ab) or the contigs from claim 3 are annotated by
automatic annotation.

10 6. Method according to any of the previous claims, in which
sequences are obtained from individuals belonging to a species and
compared with available EST data to reveal non-coding sequences such as
intron sequences and gene internal non-coding sequences.

15 7. Method according to any of the previous claims, in which
sequences are obtained from one or more individuals belonging to related
species and compared with available EST data reveal non-coding sequences
such as intron sequences and gene internal non-coding sequences.

20 8. Method according to any of the previous claims, in which
sequences are obtained from two or more individuals belonging to the same
species species and compared to reveal polymorphic sites.

25 9. Method according to any of the previous claims, in which
sequences are obtained from one or more individuals from different species
and compared to reveal polymorphic sites.

30 10. Method according to any of the previous claims, in which
sequences are obtained from one or more individuals from different species

and compared to reveal conserved areas in genomic DNA.

11. Method according to any of the previous claims, wherein
the adaptor comprising a recognition site for a rare restriction enzyme from
5 step h) comprises the recognition site for the enzyme SapI.

12. Method according to any of the previous claims, wherein
the fragmentation of the nucleic acids is achieved by nebulisation.

10 13. Kit for performing the method according to claim 1 or 2,
comprising one or more adaptors and instructions for use, and optionally
one or more primers complementary to said adaptors, ligase, and/or
restriction enzymes that are specific for cutting the adaptors, conventional
components for amplification kits per se, like dNTPs, and polymerase.

15

14. Kit according to claim 13, wherein the adaptors are
obtained by annealing of oligonucleotides selected from the group consisting
of:

5'-AGTCCGTCGCATCGCTCTTC-3'

20 5'-GAAGAGCGATGCGACG-3'

5'-Biotin-TEG-AGTGGGTGTCCTGGGTCAAC-3'

5'-GTTGACCCAGGACACC-3'

5'-CTTGTAGGGCACGGGTCGAGAG-3'

5'-AATTCTCTCGACCCGTGCCCTA-3'

25 5'-CTTGTAGGGCACGGGTCGGAGA-3'

5'-AGCTTCTCCGACCCGTGCCCTA-3'

5'-GAATGGCTGGGAGAGTGCTGAG-3'

5'-GATCCTCAGCACTCTCCCAGCC-3' and

5'-GTAGGGCACGGGTCGGAGAAGC-3'.

30

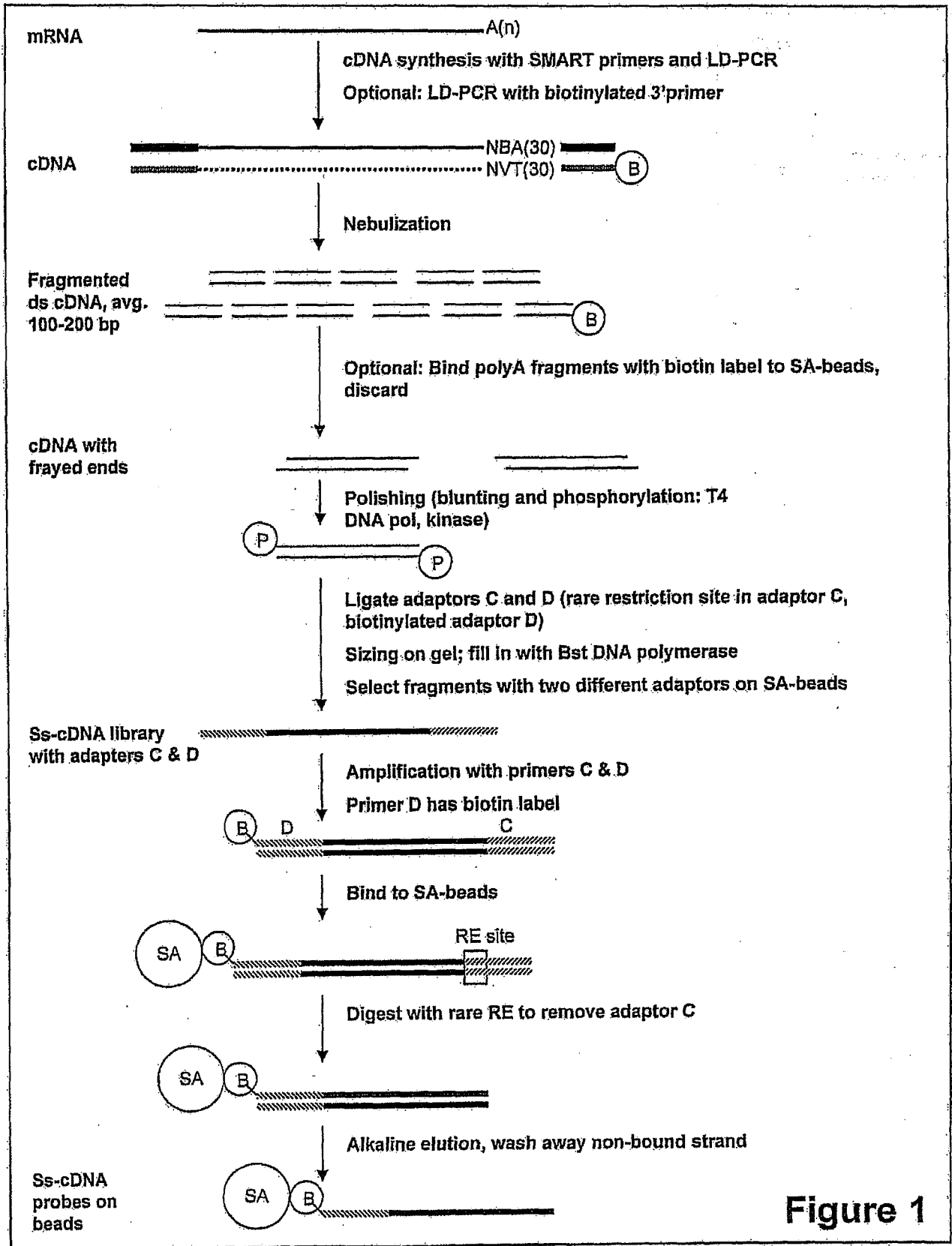


Figure 1

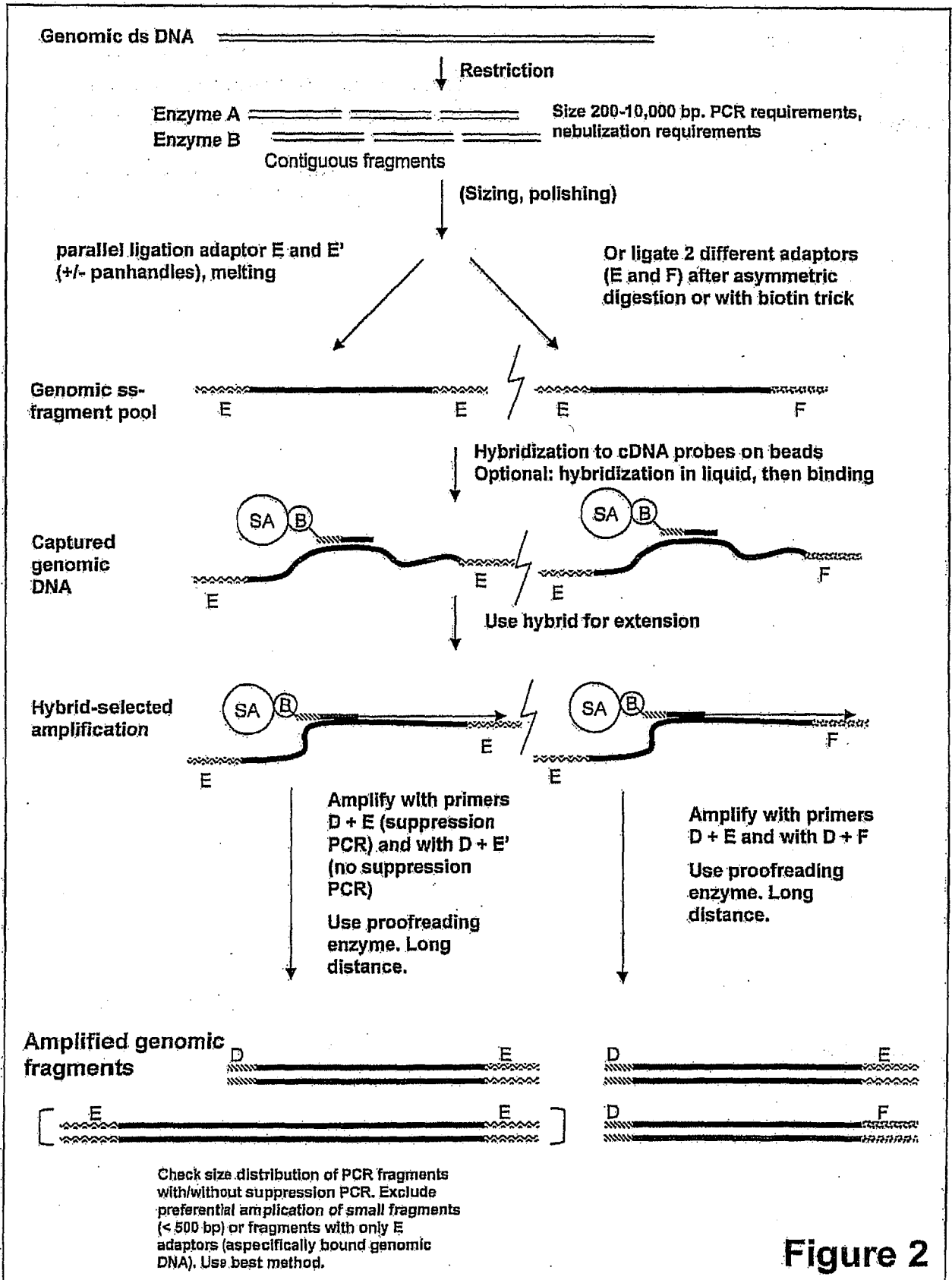


Figure 2

FIGURE 3

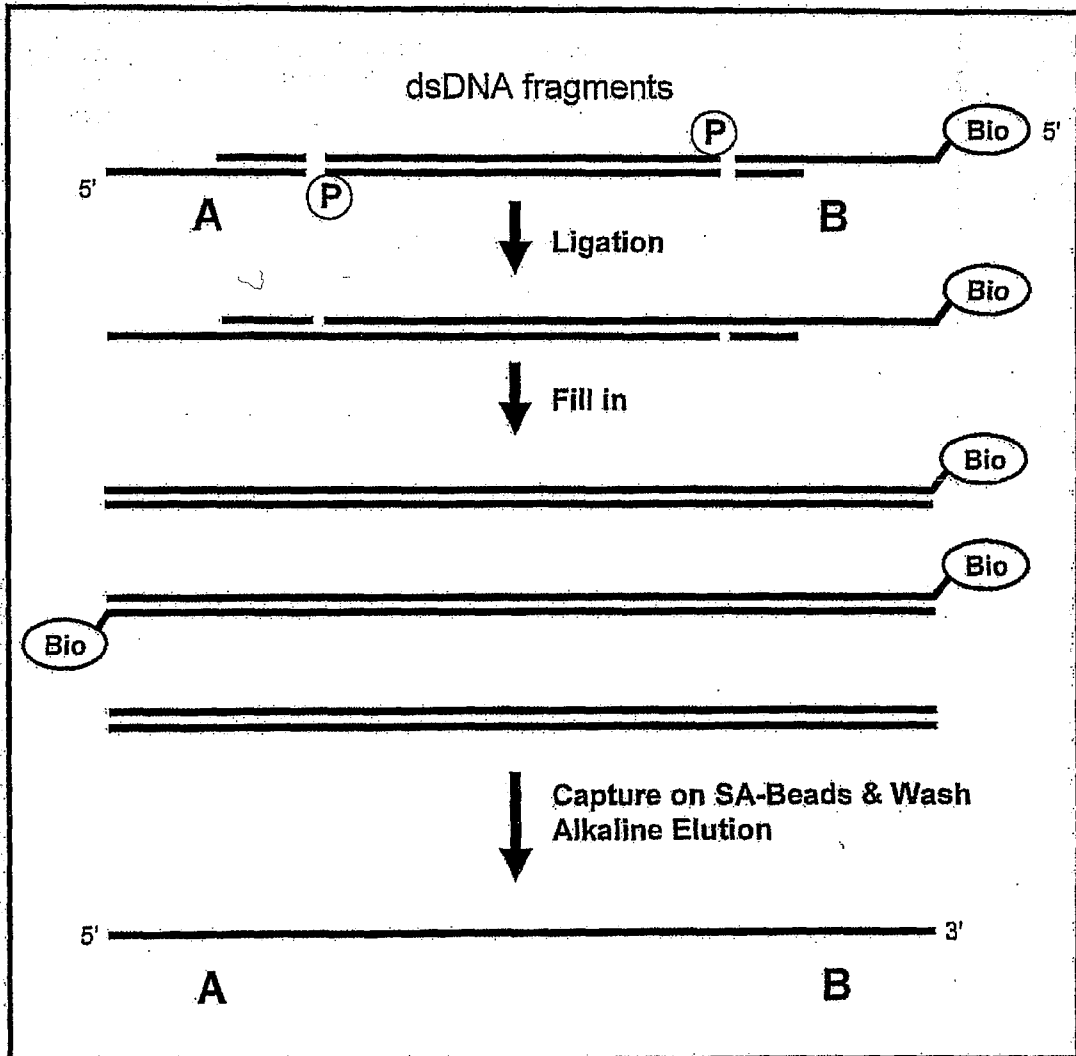


Figure 4

