Title: THREE DIMENSIONAL MATRIX ANALYSES FOR HIGH THROUGHPUT SEQUENCING

Abstract: Methods and systems for combining a multi-dimensional matrix (with at least three dimensions) and high-throughput sequencing technologies to identify/recover genomic locations of each insert in thousands of transgenic plants simultaneously. In some embodiments, multiplex sequencing is carried out and sequencing data are imported in parallel into sequence data base for displaying in the multi-dimensional matrix.
THREE DIMENSIONAL MATRIX ANALYSES FOR HIGH THROUGHPUT SEQUENCING


FIELD OF THE INVENTION

[0001] This invention is generally related to the field of plant molecular biology, and more specifically to the field of high-throughput sequencing and genomic insertion-site recovery.

BACKGROUND

[0002] Determining the genomic location and the chromosomal flanking sequence adjacent to an inserted transgene is technically challenging. Various methods have been developed to overcome the limitation of identifying the unknown DNA sequences which flank a known DNA sequence. However, these traditional PCR methods for the identification of genomic chromosomal sequences which flank a known transgene such as LM-PCR (also described as Genome Walking) and other methods including inverse PCR (i-PCR), thermal asymmetric interlaced PCR (TAIL-PCR), anchored PCR (a-PCR) and randomly primed PCR (rm-PCR) are hindered by low detection sensitivity (requiring large quantities of template DNA) or low specificity because of losses of DNA during preparation.

[0003] The development of a method which can improve detection sensitivity by purifying chromosomal DNA fragments which contain both the known and unknown DNA sequences can result in a sensitive method for detecting and characterizing unknown DNA regions which are located adjacent to a known DNA sequence. The development of the Linear Amplification Mediated Polymerase Chain Reaction (LAM PCR) method achieves these goals. The LAM PCR method is particularly suited to amplify and analyze DNA fragments, the sequence of which is only known in part.

[0004] Previous methodology for identifying or recovering genomic insertion-site is labor intensive, time-consuming, and not adaptable as high-throughput analysis. Thus, there remains a need for systems and methods for analyzing high throughput sequencing data, especially for genomic insertion-site recovery.

SUMMARY OF THE INVENTION

[0005] Methods and systems for combining a multi-dimensional matrix (with at least three dimensions) and high-throughput sequencing technologies to identify/recover genomic
locations of each insert in thousands of transgenic plants simultaneously. In some embodiments, multiplex sequencing is carried out and sequencing data are imported in parallel into sequence data base for displaying in the multi-dimensional matrix.

[0006] In one aspect, provided is a method for use preferentially in a computerized system for identifying/recovering genomic insertion sites. The method comprises:

(a) generating a three (or N) dimensional matrix with a pre-selected number of coordinates;
(b) generating a sequence database of genomic insertion-sites using a sequencing module;
(c) assigning a coordinate to each of genomic insertion-sites of the sequence database to be screened in the matrix; and
(d) pooling each of genomic insertion-sites into its assigned coordinate vertically, horizontally, and laterally.

[0007] In one embodiment, the three dimensional matrix comprises a plural of vertical columns, a plural of horizontal floors, and a plural of laterally layers. In a further or alternative embodiment, the method further comprises displaying the three dimensional matrix in a color-coded mode, where vertical columns, horizontal floors, and laterally layers are displayed in different colors.

[0008] In another embodiment, the genomic insertion sites are located in genome of a transgenic plant. In a further or alternative embodiment, the transgenic plant is selected from the group consisting of corn, soybean, sunflower, canola, cotton, and wheat.

[0009] In one embodiment, the pre-selected number of coordinates is between 20 and 20,000. In another embodiment, the pre-selected number of coordinates is bigger than 20. In another embodiment, the sequencing module comprises a high throughput sequencing system. In a further or alternative embodiment, the sequencing module comprises a next generation sequencing system. In another embodiment, the sequencing module performs multiplex sequencing. In a further embodiment, sequencing data from the multiplex sequencing are imported in parallel into the sequence database.

[0010] In one embodiment, the sequencing module identifies/recovers genomic insertion-sites using a method comprising:

(i) shearing or cleaving genomic DNA into fragments;
(ii) performing primer extensions;
(ii) capturing primer extension products;
(iv) ligating adaptors with identifiers to the primer extension products;
(v) optionally amplifying the ligated products; and
(vi) sequencing the ligated products.

[0011] In another embodiment, the fragments of DNA comprise between 100 base pairs (bp) and 500 bp DNA. In another embodiment, the fragments of DNA comprise between 50 base pairs (bp) and 500 bp DNA. In another embodiment, the fragments of DNA comprise longer than 50 bp DNA. In another embodiment, the primer extension products are captured using magnetic beads or biotin. In another embodiment, the ligated products are amplified using polymer chain reaction (PCR) where adaptors with identifiers are used as primers for amplification. In another embodiment, the ligated products are sequences using a next generation sequencing system. In a further embodiment, sequencing data of the ligated products are imported in parallel into the sequence database. In some embodiments, each of the columns, floors, and layers is associated with a particular identifier in the adapters.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Figure 1 shows an exemplary embodiment of the three dimensional matrix disclosed in a "cube" form. The samples are pooled vertically (y in floors), horizontally (x in columns), and laterally (z in layers).

[0013] Figure 2 shows an exemplary embodiment of the sequencing module disclosed for obtaining high-throughput sequencing data for genomic insertion-site recovery.

DETAILED DESCRIPTION OF THE INVENTION

[0014] T-DNA and transposon insertion mutagenesis are important tools for the isolation of new genes and study their function. The insertion of elements (randomly) within a genome might up- or down-regulate the transcription of nearby genes and therefore create altered phenotypes. The genes responsible for generating these modified phenotypes can then be associated by genetic analysis. This association is usually achieved by elucidating the genomic regions flanking the T-DNA and transposon insertions. Provided are methods and systems to combine a dimensional (multi-D) matrix pooling system and high-throughput sequencing technologies (for example, next generation sequencing (NGS)) to identify the genomic locations of each insert in thousands of transgenic lines simultaneously. The methods and systems disclosed can be used to associate generated phenotypes to the genes responsible of them.

[0015] In forward and reverse genetics, the function of a gene that is responsible for a particular phenotype within an organism can be identified. For instance, an activation tagging platform in maize is developed using transposon technology that enables a few transformation events containing the enhancers to be amplified into many activation tagged
lines with insertions at different locations. The insertions are predicted to generate phenotypes, and in order to associate these to the gene responsible for them the genomic locations of those insertions need to be known. To date, the main approach used to identify the genomic locations of inserts is TAIL PCR or Ligation-Mediated (LM) PCR followed by sequencing. However, these methods demand significant time and efforts and are gene-by-gene approaches (for example, see Alonso, JM et al. (2003) "Genome-wide Insertional mutagenesis of Arabidopsis thaliana." Science 301:653-657. Although a relatively novel method in a "Sudoku fashion" was recently designed to pool samples for NGS, that approach is more complex and the pooling is limited to two dimensions. Provided are methods and systems using a multi-dimension matrix (with at least three dimensions) for the molecular screening and genomic mapping of transgenic elements/insertions in large populations of plants/organisms (thousands and above). Accordingly, insertion sites in over 100,000 individual plants/organisms can be identified in a single NGS run. The methods and systems disclosed are also useful in cell culture systems for high-throughput mapping of multiple events.

[0016] In some embodiments, insert locations are identified using adapter mediated PCR on pools of DNA from several individuals. To form these pools genomic DNA is isolated from each line and pooled in multiple dimensions as described in Figure 1.

[0017] The methods and systems disclosed provide a very cost efficient and fast screening approach to identify the genomic locations of inserts in a large number of transgenic plants/organisms. The methods and systems disclosed provide the following advantages: (1) enable the identification of gene function by association with phenotypes in thousands of lines that otherwise will require much more time and resources; (2) enable the discovery of optimal genomic locations (in hundreds of thousands of plants) for targeted gene insertion, which may help overcome transgene silencing immediately and over generations for any transgene; (3) provide a new approach to pool indexed samples and/or sequencing libraries to maximize the sequencing capability for NGS; and (4) provide early screening of events in cell systems.

[0018] In some embodiments, the methods disclosed comprise (a) the process of generating a 3D matrix (can be expanded to N dimensions), (b) assigning a location (coordinate) to the plants/organisms to be screened in the matrix, (c) pooling samples vertically, horizontally and laterally (or in all N dimensions), and (d) Generation of NGS sequencing libraries of T-DNA tagged genomic sequences that will be indexed according to their "pools." In some embodiments, the methods disclosed also include: (i) shear or cleave
the DNA with restriction enzymes in smaller pieces of 200-400bp; (ii) perform primer
extensions with or without modified primers, capture or amplify the said extensions with
beads or other methods; (iii) ligate adaptors with barcodes/indexes or other differentiating
method, amplify the ligated pieces; (iv) NGS sequence the amplified pieces; and (v) use
bioinformatics tools to assign each piece a location in the genome.

[0019] The polymerase chain reaction (PCR) is a commonly employed molecular biology
method. The method is performed by denaturing double-stranded template DNA, annealing
oligonucleotide primers to the DNA template, and extension of a DNA strand via a DNA
polymerase. The oligonucleotide primers are designed to anneal to opposite strands of the
DNA and positioned so that the DNA strand produced by the DNA polymerase serves as a
template strand for the other primer. Each cycle is repeated, resulting in the exponential
amplification of a DNA fragment. (Mullis et al. U.S. Pat. No. 4,683,195, 4,683,202, and
4,800,159). The use of PCR by those skilled in the art is fundamental for amplifying and
isolating DNA fragments for subsequent analysis.

[0020] Isolation and analysis of DNA templates via the polymerase chain reaction (PCR)
requires knowledge of the flanking DNA sequences. Unfortunately, this requirement limits
PCR amplification to regions of known DNA sequence. The use of PCR methodologies to
identify the location of a transgene location within a genome is hindered by the random
insertion of the transgene into an unknown chromosomal location within the genome of an
organism. Methods to identify unknown DNA sequences which are located adjacent to a
known DNA sequence are necessary for the identification of a transgene location within the
chromosome of an organism. In addition such methods can be used to identify novel gene
sequences to identify new traits, to determine the genomic location of a transposon or viral
sequence which has been inserted into the genome of an organism, or to identify the
chromosomal location of polynucleotide sequences inserted into the genome via insertion
mutagenesis.

[0021] Various methods have been developed to overcome the limitation of the unknown
DNA sequences which flank a known DNA sequence. A Ligation Mediated PCR (LM PCR)
method wherein a genomic library is generated and adapters are annealed to DNA fragments
for PCR amplification is marketed as the GENOME WALKER UNIVERSAL KIT™ (see
U.S. Pat. No. 5,565,340, and U.S. Pat. No. 5,759,822). Another method commonly used is
the inverse PCR reaction (see Silver and Keerikatte (1989), J. Virol., 63:1924-1928), wherein
DNA is digested with a restriction enzyme and self ligated resulting in a contiguous circle.
PCR amplification using oligonucleotide primers which bind to known sequences results in
amplification and elucidation of the unknown flanking sequences. Unfortunately these methods are inefficient and time consuming. These and other traditional PCR methods (including thermal asymmetric interlaced PCR [TAIL-PCR], anchored PCR [a-PCR] and randomly primed PCR [rm-PCR]) are hindered by low detection sensitivity (requiring large quantities of template DNA) or low specificity because of losses of DNA during preparation. [0022] The development of a method which can improve detection sensitivity by purifying chromosomal DNA fragments which contain both the known and unknown DNA sequences can result in a sensitive method for detecting and characterizing unknown DNA regions which are located adjacent to a known DNA sequence. The development of the LAM PCR method achieves these goals. The Linear Amplification Mediated Polymerase Chain Reaction (LAM PCR) method is a modified PCR method that is used for analyzing unknown chromosomal flanking sequences located adjacent to a known DNA sequence. The LAM PCR method can be used to identify and/or sequence an unknown DNA or RNA sequence flanking a known DNA or RNA region. [0023] The LAM PCR method consists of the following steps. A primer extension reaction is performed using a chromosomal DNA as a template and an oligonucleotide primer which binds to a known DNA sequence within the chromosomal DNA. The oligonucleotide primer is complementary to a long terminal repeat (LTR) sequence, which is a sequence characteristic of a retrovirus, and labeled with biotin at the end of the oligonucleotide primer. The single-stranded DNA product of the linear PCR is bound to magnetic beads having immobilized streptavidin. This step serves to isolate the single-stranded amplified DNA fragment containing the known LTR sequence and an unknown sequence derived from the chromosome. The single-stranded DNA is converted into a double-stranded DNA by synthesizing the complementary strand. The double-stranded DNA is cleaved with a restriction enzyme that recognizes a sequence and cleaves the double-stranded DNA at the sequence. A double-stranded DNA called a linker cassette is ligated to the terminus. Subsequent PCR reactions are conducted using the thus obtained ligation product as a template as well as a primer complementary to the LTR and a primer complementary to the linker cassette. A DNA fragment that contains the LTR and chromosome DNA flanking sequence adjacent to the LTR is amplified. As a result the previously unknown retrovirus integration site can be determined. [0024] The LAM PCR method is currently considered to be an effective system for analyzing unknown DNA sequences adjacent to a known DNA sequence. However, modifications and improvements to the LAM PCR method have been described in the art.
The LAM PCR method was modified in U.S. Pat. App. US2007/0037139 to improve the detection of a biological sample having a retrovirus integrated at various sites. The reaction conditions of the traditional LAM PCR method produced results that did not reflect the actual state of clones existing in the cell population of the sample. A modification was developed in which more integration fragments were PCR amplified without being biased toward a fragment amplified from a specific clone. The modification to the LAM PCR method allowed researchers to determine the extent of cells having an integrated gene in the population and to determine the ratio of a specific cell in the population.

In addition, Harkey et al., (2007) describe an optimized, multiarm, high throughput modification of the LAM PCR method wherein the detection capacity was improved 90% with exhaustive sampling. The modified protocol facilitated accurate estimates of the total pool size, thus providing a rapid, cost-effective approach for generating large insertion-site data of preferred genomic locations for vector integration.

In some embodiments, the subject invention describes another modification wherein, the traditional LAM-PCR method which requires the steps of completion of a primer extension reaction, generation of a double stranded DNA fragment and, the subsequent digestion of the double stranded DNA fragment with restriction enzymes are not required. Rather the subject invention describes an initial restriction enzyme digestion of gDNA, the ligation of a double stranded adapter to the gDNA fragment and a primer extension reaction. In addition to other modifications of the method.

As used herein, the phrase "comprises," "comprising," "includes," "including," "has," "having," "contains," "containing," or any other variation thereof, is intended to be non-exclusive and open-ended. For example, a composition, a mixture, a process, a method, an article, or an apparatus that comprises a list of elements is not necessarily limited to only those elements but may include other elements not expressly listed or inherent to such composition, mixture, process, method, article, or apparatus. Further, unless expressly stated to the contrary, "or" refers to an inclusive or and not to an exclusive or. For example, a condition A or B is satisfied by any one of the following: A is true (or present) and B is false (or not present), A is false (or not present) and B is true (or present), and both A and B are true (or present).

As used herein, the indefinite articles "a" and "an" preceding an element or component of the invention are intended to be nonrestrictive regarding the number of
instances, i.e., occurrences of the element or component. Therefore, "a" or "an" should be read to include one or at least one, and the singular word form of the element or component also includes the plural unless the number is obviously meant to be singular.

[0030] As used herein, the phrase "plant" includes plants and plant parts including but not limited to plant cells and plant tissues such as leaves, stems, roots, flowers, pollen, and seeds. The class of plants that can be used in the present invention is generally as broad as the class of higher and lower plants amenable to mutagenesis including angiosperms (monocotyledonous and dicotyledonous plants), gymnosperms, ferns and multicellular algae. Thus, "plant" includes dicotyledons plants and monocotyledons plants. Examples of dicotyledons plants include tobacco, Arabidopsis, soybean, tomato, papaya, canola, sunflower, cotton, alfalfa, potato, grapevine, pigeon pea, pea, Brassica, chickpea, sugar beet, rapeseed, watermelon, melon, pepper, peanut, pumpkin, radish, spinach, squash, broccoli, cabbage, carrot, cauliflower, celery, Chinese cabbage, cucumber, eggplant, and lettuce. Examples of monocotyledons plants include corn, rice, wheat, sugarcane, barley, rye, sorghum, orchids, bamboo, banana, cattails, lilies, oat, onion, millet, and triticale.

[0031] As used herein, the phrase "plant material" refers to leaves, stems, roots, flowers or flower parts, fruits, pollen, egg cells, zygotes, seeds, cuttings, cell or tissue cultures, or any other part or product of a plant. In some embodiment, plant material includes cotyledon and leaf.

[0032] As used herein, the phrase "plant tissue" refers to a group of plant cells organized into a structural and functional unit. Any tissue of a plant in planta or in culture is included, for example: whole plants, plant organs, plant seeds, tissue culture and any groups of plant cells organized into structural and/or functional units.

[0033] As used herein, the phrases "nucleic acid," "polynucleotide," "polynucleotide sequence," and "nucleotide sequence" refer to a polymer of nucleotides (A,C,T,U,G, etc. or naturally occurring or artificial nucleotide analogues), e.g., DNA or RNA, or a representation thereof, e.g., a character string, etc., depending on the relevant context. The phrases "nucleic acid" and "polynucleotide" are used interchangeably herein; these phrases are used in reference to DNA, RNA, or other novel nucleic acid molecules of the invention, unless otherwise stated or clearly contradicted by context. A given polynucleotide or complementary polynucleotide can be determined from any specified nucleotide sequence. A nucleic acid may be in single- or double-stranded form.

[0034] As used herein, the phrase "isolated," refers to material, such as a nucleic acid or a protein, which is: (1) substantially or essentially free from components which normally
accompany or interact with the material as found in its naturally occurring environment or (2) if the material is in its natural environment, the material has been altered by deliberate human intervention to a composition and/or placed at a locus in the cell other than the locus native to the material.

[0035] As used herein, the phrase "promoter" refers to a DNA sequence which directs the transcription of a structural gene to produce RNA. Typically, a promoter is located in a region 500 base pairs upstream of a gene, proximal to the transcription start site. If a promoter is an inducible promoter, then the rate of transcription increases or decreases in response to an exogenous or endogenous inducing agent. In contrast, the rate of transcription is regulated to a lesser degree by an inducing agent if the promoter is a constitutive promoter.

[0036] As used herein, the phrase "transgenic plant" refers to a plant or progeny thereof derived from a transformed plant cell or protoplast, wherein the plant DNA contains an introduced exogenous DNA molecule not originally present in a native, non-transgenic plant of the same.

[0037] As used herein, the phrase "vector" refers to any recombinant polynucleotide construct that may be used for the purpose of transformation, i.e., the introduction of heterologous DNA into a host cell.

[0038] As used herein, the phrase "complementary strand" refers to nucleic acid sequences or molecules in which each base in one molecule is paired with its complementary base in the other strand, to form a stable helical double strand molecule. The individual strands are termed complementary strands.

[0039] As used herein, the phrase "oligonucleotide primer" refers to a sequence of linear oligonucleotides of about ten to about fifty nucleotides in length that are complementary to nucleotide sequences 5' or 3' to be amplified. A pair of oligonucleotide primers, in which one of the primers is complementary to a nucleotide sequence 5' of the polynucleotide fragment to be amplified while the other primer of the pair is complementary to a nucleotide sequence located 3' polynucleotide fragment to be amplified can be used to amplify a polynucleotide sequence. One skilled in the art understands that a pair of oligonucleotide primers means two oligonucleotides complementary to opposite strands of nucleic acid and flanking the polynucleotide sequence to be amplified.

[0040] As used herein, the phrase "adapter" refers to a short, oligonucleotide polynucleotide segment that can be joined to a polynucleotide molecule at either a blunt end or cohesive end. Adapters may contain restriction enzyme recognition sequences within the polynucleotide fragment. The size of the adapter can vary from about ten to about one-
hundred and fifty nucleotides in length. Adapters can either be single stranded or double stranded.

[0041] As used herein, the phrase "isolated complementary strand" refers to a polynucleotide fragment which comprises an adapter joined to a second DNA fragment that contains a portion or all of the known polynucleotide sequence and an adjacent unknown polynucleotide sequence via a ligation reaction. The "isolated complementary strand" is flanked by an adapter on one end and a known polynucleotide sequence on the other end.

[0042] A method of analyzing, in chromosomal DNA having a transgene incorporated therein, a DNA flanking region derived from the chromosome which is adjacent to the transgene. Wherein, the DNA flanking region is characterized by isolation and digestion of genomic DNA with a restriction enzyme, ligation of a double stranded adapter to the isolated and digested genomic DNA, a primer extension reaction of the adapter ligated genomic DNA, and the isolation of the primer extension reaction product via a streptavidin-biotin interaction. The DNA flanking region is further characterized via subsequent PCR amplification reactions and DNA sequencing.

[0043] Disclosed within this application is a modified LAM PCR method used to identify genomic chromosomal sequences which flank an inserted transgene. The disclosed method contains modifications to the traditional LAM PCR method that are developed to improve the accuracy, sensitivity, and reproducibility of the detection of the unknown chromosomal sequences which flank a known DNA sequence. This modified LAM PCR method can be deployed as a high throughput method to quickly and efficiently identify the genomic chromosomal sequences which flank a transgene. Further analysis of these sequences can be used to characterize the transgene insertion site for the identification of rearrangements, insertions and deletions which result from the integration of the transgene. In addition, analysis of the chromosomal flanking sequences can be used to identify the location of the transgene on the chromosome. Finally, the method can be broadly applied to determine an unknown DNA sequence which is adjacent to any known DNA sequence.

[0044] In another aspect, provided is a method for finding an unknown polynucleotide sequence adjacent to a known polynucleotide sequence in isolated plant DNA. The comprises:

(a) digesting the isolated plant DNA that contains a portion or all of the known polynucleotide sequence and an adjacent unknown polynucleotide sequence with one or more suitable restriction enzymes to produce a plurality of digested polynucleotide restriction fragments;
(b) ligating a double stranded adapter to the digested polynucleotide restriction fragments;
(c) synthesizing a complementary strand of the adapter ligated polynucleotide restriction fragments using an oligonucleotide primer sequence having an attachment chemistry bound to the 5’ end of the oligonucleotide primer sequence;
(d) isolating the complementary strand by binding the attachment chemistry to a suitable isolation matrix;
(e) performing a PCR amplification of the isolated complementary strand using a first PCR primer designed to bind to the known polynucleotide sequence and a second PCR primer designed to bind to the adapter sequence to produce a PCR amplicon; and,
(f) sequencing the PCR amplicon to ascertain the sequence of the unknown polynucleotide sequence.

[0045] In another aspect, provided is a method for the isolation and identification of transgene border sequences. In another aspect, provided is a method which is readily applicable for high throughput analysis to determine the transgenic copy number and the chromosomal location of a genomic insertion site. In addition, the methods disclosed can be used for the simultaneous detection of multiple insertion sites within one reaction. In some embodiments, the methods disclosed have improved sensitivity and specificity for the detection of unknown polynucleotide fragments which flank a known polynucleotide fragment. Moreover, the methods disclosed can be deployed to detect the unknown DNA sequences which are located adjacent to any target sequence, including viral sequences and insertional mutagenesis sites created via transposon mutagenesis or mutagenesis generated via T-strand integration.

[0046] Another aspect relates in part to transgenic event identification using such flanking, junction, and insert sequences. In one embodiment, modified PCR analysis and DNA sequencing analysis methods using amplicons that span across inserted transgene DNA and its borders can be used to detect or identify commercialized transgenic plant varieties or lines derived from the proprietary transgenic plant lines.

[0047] The transgene border and adjacent chromosomal flanking sequences disclosed can be diagnostic for a transgenic event. Based on these sequences, transgenic plant lines can be identified in different plant genotypes by analysis of the chromosomal flanking and transgene sequences. Thus, an embodiment of the subject invention describes a method that can be used to identify transgenic plant lines.

[0048] The chromosomal flanking sequences of the subject invention are especially useful in conjunction with plant breeding, to determine which progeny plants comprise a given
event, after a parent plant comprising an event of interest is crossed with another plant line in an effort to impart one or more additional traits of interest in the progeny. Also provided are methods for the determination of the chromosomal flanking/junction sequences to benefit breeding programs as well as quality control, especially for commercialized transgenic plant lines.

Furthermore, the identification of chromosomal flanking sequences can be used to specifically identify the genomic location of each transgenic insert. This information can be used to develop molecular marker systems specific for each event. These molecular marker systems can be used for accelerated breeding strategies and to establish linkage data. Also provided are methods for the development of molecular marker systems for marker assisted breeding.

Still further, the chromosomal flanking sequence information can be used to study and characterize transgene integration processes, genomic integration site characteristics, event sorting, stability of transgenes and their flanking sequences, and gene expression (especially related to gene silencing, transgene methylation patterns, position effects, and potential expression-related elements such as MARS [matrix attachment regions], and the like).

The methods disclosed can be used to obtain and ascertain the sequence of the unknown polynucleotide from a transgenic organism. In some embodiments, the sample can be genomic DNA and the transgenic organism can be a transgenic plant. Transgenic plants analyzed by any of the methods of this invention can be selected from plants consisting of barley, corn, oat, sorghum, turf grass, sugarcane, wheat, alfalfa, banana, broccoli, bean, cabbage, canola, carrot, cassava, cauliflower, celery, citrus, cotton, a cucurbit, eucalyptus, flax, garlic, grape, onion, lettuce, pea, peanut, pepper, potato, poplar, pine, rye, rice, sunflower, safflower, soybean, strawberry, sugar beet, sweet potato, tobacco, tomato, ornamental, shrub, nut, millet, and pasture grass.

The methods disclosed can be used to obtain and ascertain the sequence of the unknown polynucleotide from a non-transgenic organism. In some embodiments, the sample can be genomic DNA and the non-transgenic organism can be a plant. Plants analyzed by any of the methods of this invention can be selected from plants consisting of barley, corn, oat, sorghum, turf grass, sugarcane, wheat, alfalfa, banana, broccoli, bean, cabbage, canola, carrot, cassava, cauliflower, celery, citrus, cotton, a cucurbit, eucalyptus, flax, garlic, grape, onion, lettuce, pea, peanut, pepper, potato, poplar, pine, rye, rice, sunflower, safflower, soybean, strawberry, sugar beet, sweet potato, tobacco, tomato, ornamental, shrub, nut,
millet, and pasture grass. In some embodiments, the unknown polynucleotide sequence adjacent to a known polynucleotide sequence can be a native polynucleotide of agronomic interest.

[0053] A ligation reaction can be completed by an enzyme, generally referred to as a ligase that catalyzes the formation of a phosphodiester bond between adjacent 3’-OH and 5’-P termini in DNA.

[0054] Isolation of a plant DNA can be accomplished by methods known in the art. Generally, the isolation of a plant DNA results in obtaining purified plant DNA which is free of lipids, proteins and other cellular debris. One example for plant DNA isolation methods include: lysis, heating, alcohol precipitation, salt precipitation, organic extraction, solid phase extraction, silica gel membrane extraction, CsCl gradient purification, and any combinations thereof. Another example for plant DNA isolation method is the silica-gel-membrane technology marketed as the DNeasy kit (Qiagen, Valencia, CA) or the Cetyltrimethylammonium Bromide (CTAB) DNA isolation protocol.

[0055] Restriction enzyme digestions, also referenced as restriction endonuclease digestions, are performed when a nuclease enzyme is used to cleave the polynucleotide sequences. There are numerous restriction enzymes available to those skilled in the art. As described at World Wide Web: neb.com/nebecomm/tech_reference/restriction_enzymes/overview.asp, four classifications are used to characterize restriction enzymes. These classifications are made on the basis of subunit composition, cleavage position, sequence specificity and cofactor requirements.

[0056] Type I enzymes randomly cut DNA at locations which are a distance from the recognition/binding sequence (> 1,000 bp away). The recognition sites which are bound by a Type I enzyme are asymmetrical. As a result these enzymes are not used for gene cloning because these enzymes do not produce discrete restriction fragments or distinct gel-banding patterns. Type I enzymes are multifunctional and the different subunits which comprise a Type I restriction enzyme are responsible for different activities (i.e. subunit HsdR encodes restriction, subunit HsdM encodes methylation of DNA, and subunit HsdS encodes specificity of the recognition sequence).

[0057] Type II enzymes digest DNA at positions located within close proximity of the recognition sequences. These enzymes function as a dimer, wherein a subunit binds to the sense strand and a second copy of the subunit binds to the antisense strand at a palindromic sequence which is typically between 4-8 nucleotides in length. The Type II dimer that binds to the DNA can be either a homodimer which bind to symmetric DNA sequences, or a
heterodimer which binds to asymmetric DNA sequences. The enzymes can recognize either continuous sequences or discontinuous sequences. Type II enzymes are commercially available and commonly used for DNA analysis and gene cloning. Widespread usage of these enzymes is a result of distinct restriction fragments which are produced and can be resolved on an agarose gel.

[0058] Type II enzymes are a collection of unrelated proteins which are highly divergent in amino acid sequence similarity. Type II enzymes have been divided into subcategories which are labeled using a letter suffix. Type IIB restriction enzymes are multimers that contain more than one subunit. These enzymes cut both sides of the recognition sequence, thereby resulting in removal of the recognition sequence. Type HE and Type IIF restriction enzymes cleave DNA following interaction with two copies of their recognition sequence. Type IIG restriction enzymes are comprised of a single subunit. The N-terminal portion of the enzyme possesses a DNA cleavage domain and DNA modification domain. The C-terminal portion of the enzyme possesses a DNA sequence binding domain. These enzymes cleave outside of their recognition sequence. Type IIM restriction enzymes recognize and cut methylated DNA. Type IIS restriction enzymes function as a dimer and cleave DNA at a location which is outside of the non-palindromic asymmetric recognition sites. These enzymes are comprised of two distinct domains, one for DNA binding and the other for DNA cleavage.

[0059] Type III enzymes are combination restriction-and-modification enzymes. These enzymes recognize two separate non-palindromic sequences and cleave outside of their recognition sequences. Type III enzymes require two recognition sequences in opposite orientations within the same DNA molecule to accomplish cleavage.

[0060] Type IV enzymes recognize methylated DNA. Examples include the McrBC and Mrr systems of E. coli.

[0061] Other methods are known in the art for cleaving polynucleotides and can be used in place of digesting the polynucleotide with a restriction enzyme, any of the group consisting of: lysis, a sequence-specific cleavage agent, non-sequence specific cleavage agent, sonication, shear-stress, French press, UV radiation, ionizing radiation, and DNase. In addition, to the restriction enzymes described above, homing endonucleases or Flap endonucleases or any combination of these enzymes could be used to digest the isolated DNA. A preferred method for digesting isolated plant DNA is the use of a Typell restriction enzyme which is known to cut outside of the transgene sequence being transformed into the plant. Another preferred method for digesting isolated plant DNA is the use of a Typell
restriction enzyme which is known to cut at a site which is in close proximity of the end of the transgene sequence.

[0062] Primer extension reactions are used to produce a DNA or RNA strand which contains a known polynucleotide sequence and an unknown adjacent polynucleotide sequence. Primer extension methodologies result in the production of a complementary strand of DNA or RNA which contains the unknown polynucleotide sequence. The complementary strand of DNA or RNA is produced by a polymerase which extends along a template strand of DNA or RNA after complexing with an oligonucleotide primer which has bound to the known template strand of DNA or RNA. The oligonucleotide primer is designed to specifically bind to the known DNA or RNA sequence within the template strand of DNA or RNA. Numerous types of polymerase are commercially available for the extension reaction; T4 polymerase, TAQ polymerase, PFU polymerase, or Reverse Transcriptase are a few non-limiting examples of commonly used polymerases. Each polymerase has special buffer requirements and function at a specific temperature for optimal reaction conditions. For example a primer extension reaction is the use of the TAQ polymerase marketed as the Platinum Taq kit.

[0063] Attachment chemistries attached to an isolation matrix such, as magnetic bead-based systems, are used to isolate the DNA produce by the primer extension reaction. The DNA strand which is produced by the primer amplification reaction can be purified from genomic DNA via a streptavidin - biotin interaction. Biotinylation is widely used to enable isolation, separation, concentration and further downstream processing and analysis of biomolecules (for example, methods described in U.S. Patent No. 5,948,624, U.S. Patent No. 5,972,693, and U.S. Patent No. 5,512,439). There are a variety of commercially available biotinylation reagents that target different functional groups like primary amines, sulphydryls, carboxyls, carbohydrates, tyrosine and histidine side chains and cyanidine and cytosine bases. The use of short, sequence-specific oligonucleotide primers functionalized with biotin (or the equivalent, e.g., digoxigenin) and magnetic beads to separate specific DNA sequences from the genome for subsequent analysis have multiple uses. Isolation using the bead-based method allows for enrichment of a population of DNA for a particular sequence, allowing subsequent analysis to be carried out that could not be done in the presence of the entire genomic complement of DNA. Such bead-based methods are suited for high throughput automation.

[0064] Although the biotin - streptavidin interaction is the best described binding pair, other molecules which have a strong affinity for one another are known. Attachment
chemistries that can be included into a oligonucleotide primer include: ACRYDITE™ an attachment chemistry based on an acrylic phosphoramidite that can be added to oligonucleotides as a 5’-modification, and covalently reacts with thiol-modified surfaces; Alkyne modifications which react with azide labeled functional groups to form stable bonds through the azide alkyne Huisgen cycloaddition reaction (also referenced as the Click reaction); and, Thiol modifications which can couple and interact with high affinity to a corresponding ligand or surface (such as a gold surface). These molecules can be used for purification or enrichment of DNA sequences. Wherein, a primer is labeled with a first molecule and the second molecule is bound to a matrix which can immobilize the first molecule (e.g. magnetic beads). A DNA strand produced from the primer labeled with the first molecule can be isolated by running the DNA over a column containing the immobilized matrix (e.g., magnetic beads) labeled with the second molecule. As a result of the affinity for the second molecule, the amplified DNA sequences containing the primer labeled with the first molecule are isolated. Preferred attachment chemistries include acrylic - thiol interactions, alkyne - azide interactions, and thiol - ligand interactions. Another example of attachment chemistry is the streptavidin - biotin interaction.

[0065] As used herein, the phrase ”isolation matrix” refers to a surface to which a molecule of any sort may be attached. For example, an isolation matrix is an insoluble material to which a molecule may be attached so that said molecule may be readily separated from other components in a reaction. For example, isolation matrices may include, but is not limited to, a filter, a chromatography resin, a bead, a magnetic particle, or compositions that comprise glass, plastic, metal, one or more polymers and combinations thereof. A more preferred isolation matrix is the magnetic bead-based system.

[0066] Adapters can be ligated to double stranded genomic DNA via a ligase. Commercially supplied ligases are widely available for joining double stranded DNA fragments. For example double stranded ligases are commercially available and marketed as T4 Ligase (New England Biolabs; Ipswich, MA or Roche Biosciences; Indianapolis, IN), Taq Ligase (New England Biolabs; Ipswich, MA), and E. coli DNA Ligase (New England Biolabs; Ipswich, MA). For another example double stranded DNA ligase is the Quick Ligation kit from New England Biolabs (Ipswich, MA).

[0067] DNA sequence analysis can be used to determine the nucleotide sequence of the isolated and amplified fragment. The amplified fragments can be isolated and sub-cloned into a vector and sequenced using chain-terminator method (also referred to as Sanger sequencing) or Dye-terminator sequencing. In addition, the amplicon can be sequenced with
Next Generation Sequencing. NGS technologies do not require the sub-cloning step, and multiple sequencing reads can be completed in a single reaction. Three NGS platforms are commercially available, the Genome Sequencer FLX from 454 Life Sciences/Roche, the Illumina Genome Analyser from Solexa and Applied Biosystems' SOLiD (acronym for: 'Sequencing by Oligo Ligation and Detection'). In addition, there are two single molecule sequencing methods that are currently being developed. These include the true Single Molecule Sequencing (tSMS) from Helicos Bioscience and the Single Molecule Real Time sequencing (SMRT) from Pacific Biosciences.

[0068] The Genome Sequencer FLX which is marketed by 454 Life Sciences/Roche is a long read NGS, which uses emulsion PCR and pyrosequencing to generate sequencing reads. DNA fragments of 300 - 800 bp or libraries containing fragments of 3-20 kbp can be used. The reactions can produce over a million reads of about 250 to 400 bases per run for a total yield of 250 to 400 megabases. This technology produces the longest reads but the total sequence output per run is low compared to other NGS technologies.

[0069] The Illumina Genome Analyser which is marketed by Solexa is a short read NGS which uses sequencing by synthesis approach with fluorescent dye-labeled reversible terminator nucleotides and is based on solid-phase bridge PCR. Construction of paired end sequencing libraries containing DNA fragments of up to 10 kb can be used. The reactions produce over 100 million short reads that are 35 - 76 bases in length. This data can produce from 3 - 6 gigabases per run.

[0070] The Sequencing by Oligo Ligation and Detection (SOLiD) system marketed by Applied Biosystems is a short read technology. This NGS technology uses fragmented double stranded DNA that can be up to 10 kbp in length. The system uses sequencing by ligation of dye-labeled oligonucleotide primers and emulsion PCR to generate one billion short reads that result in a total sequence output of up to 30 gigabases per run.

[0071] tSMS of Helicos Bioscience and SMRT of Pacific Biosciences apply a different approach which uses single DNA molecules for the sequence reactions. The tSMS Helicos system produces up to 800 million short reads that result in 21 gigabases per run. These reactions are completed using fluorescent dye-labeled virtual terminator nucleotides that are described as a "sequencing-by-synthesis" approach.

[0072] The SMRT Next Generation Sequencing system marketed by Pacific Biosciences uses a real time sequencing by synthesis. This technology can produce reads of up to 1000 bp in length as a result of not being limited by reversible terminators. Raw read throughput
that is equivalent to one-fold coverage of a diploid human genome can be produced per day using this technology.

[0073] Embodiments of the present invention are further defined in the following Examples. It should be understood that these Examples are given by way of illustration only. From the above discussion and these Examples, one skilled in the art can ascertain the essential characteristics of this invention, and without departing from the spirit and scope thereof, can make various changes and modifications of the embodiments of the invention to adapt it to various usages and conditions. Thus, various modifications of the embodiments of the invention, in addition to those shown and described herein, will be apparent to those skilled in the art from the foregoing description. Such modifications are also intended to fall within the scope of the appended claims.

[0074] The disclosure of each reference set forth herein is incorporated herein by reference in its entirety.

[0075] In one aspect, provided is a method for finding an unknown polynucleotide sequence adjacent to a known polynucleotide sequence in isolated plant DNA. The embodiment comprises:

(a) digesting the isolated plant DNA that contains a portion or all of the known polynucleotide sequence and an adjacent unknown polynucleotide sequence with one or more suitable restriction enzymes to produce a plurality of digested polynucleotide restriction fragments;

(b) ligating a double stranded adapter to the digested polynucleotide restriction fragments;

(c) synthesizing a complementary strand of the adapter ligated polynucleotide restriction fragments using an oligonucleotide primer sequence having an attachment chemistry bound to the 5' end of the oligonucleotide primer sequence;

(d) isolating the complementary strand by binding the attachment chemistry to a suitable isolation matrix;

(e) performing a PCR amplification of the isolated complementary strand using a first PCR primer designed to bind to the known polynucleotide sequence and a second PCR primer designed to bind to the adapter sequence to produce a PCR amplicon; and,

(f) sequencing the PCR amplicon to ascertain the sequence of the unknown polynucleotide sequence.

[0076] In one embodiment, the sample analyzed is a plant genomic DNA. In another embodiment, the unknown polynucleotide sequence is a transgene border. In another embodiment, the unknown polynucleotide sequence is a chromosomal sequence which flanks
a known polynucleotide sequence. In another embodiment, the unknown polynucleotide sequence is an endogenous gene sequence which encodes a trait. In another embodiment, the known polynucleotide sequence is a known polynucleotide viral sequence.

[0077] In one embodiment, the known polynucleotide sequence is a known polynucleotide transgene sequence. In another embodiment, the known polynucleotide sequence is a known polynucleotide transposon sequence. In another embodiment, the known polynucleotide sequence is a known polynucleotide gene sequence that encodes a trait. In another embodiment, the method is used to identify the chromosomal location of a known polynucleotide sequence inserted into the isolated plant DNA via insertion mutagenesis.

[0078] In one embodiment, the insertion mutagenesis is selected from the group consisting of transposon mutagenesis, or T-strand integration mutagenesis. In another embodiment, the method is used to characterize an unknown polynucleotide sequence consisting of a chromosomal sequence which flanks a known polynucleotide sequence. In another embodiment, the characterization of a transgene insertion site identifies polynucleotide sequence consisting of rearrangements, insertions, deletions, or inversions within the unknown polynucleotide sequence consisting of a chromosomal sequence. In another embodiment, the method is used in a high throughput protocol. In another embodiment, the method is used to determine transgene copy number. In another embodiment, the method is used to identify transgenic plant lines. In another embodiment, the method is used to develop molecular marker systems. In a further embodiment, the molecular maker systems are used to accelerate breeding strategies.

EXAMPLES

Example 1

3D Matrix with 27 Coordinates for Genomic Insertion Site Analysis

[0079] Provided is a novel strategy and detailed approach for the cheap and fast elucidation of transgene location in the genome and it is especially useful in large (more than thousands) populations of transgenic plants or other organisms. The process is described in Figure 1 in which a representative 3D matrix of 3 dimensions is shown as an example, although the method is of course designed for larger matrixes. As shown in Figure 1, the matrix can be pooled vertically, horizontally and laterally. Each pool contains 3x3=9 samples in this example. Samples in a pool will share the same adapter (identifier) that will correspond to that pool. Since each sample will be pooled three times (vertically, horizontally and laterally), each sample will have three coordinates each of which will be dictated by the identifier that corresponds to each pool. For instance, if the face in front is considered,
sample 7 (front left corner) will be the only sample having an identifier for the first floor (pool on Y), first layer (pool on Z) and left column (pool on X) altogether. This 3D pooling system in association with NGS will enable screening large numbers of samples at very low costs and very rapidly.

[0080] In some embodiments, the 3D matrix is basically a pooling system that instead of being done in a classical way (2D) array or plate, is done in a "cubic" fashion. In other embodiments, this can be done in n-dimensions as well but the process can be more complicated. The advantage of a 3D is that more samples can be pooled and then deciphered according to their coordinates in the matrix, which saves time and money especially if combined with high-throughput sequencing (for example, NGS). To prepare the matrix of this example, first is to "create" a hypothetical cube and each sample is assigned a place (coordinate) in the matrix. Then, each sample in a certain "floor" is labeled with a certain adaptor or sequence (different for each floor), each sample in a certain "column" is labeled with a certain adaptor or sequence (different for each column) and finally, each "layer" is also labeled with a different sequence or label (different for each layer). In that way, every sample is basically covered three times with a sequence that corresponds to a floor, column and layer and hence, it is possible to pool the samples since the location of the sample in the matrix can be deduced.

[0081] As shown in Figure 2, plant genomic DNA can be extracted and subject to cleavage or shearing. Primer extension can be performed (not mandatory) and primer extension products can be captured using biotin or other molecules (not mandatory). Adapters with bar-codes or other labeling can then be ligated to the primer extension products. The ligated products are optionally subject to PCR amplification, and then submitted for high-throughput sequencing.

**Example 2**

[0082] A plasmid containing a polynucleotide of interest expression cassette and a selectable marker gene expression cassette is used to transform *Zea mays* cv B104 plant tissue via *Agrobacterium*-mediated transformation method. Immature embryos of approximately 1.8 to 2.4 mm in length are isolated from *Zea mays* genotype B104. Isolated embryos are then incubated with an *Agrobacterium* suspension at an Optical Density of 1.0 at 600 nm for 20-30 minutes and placed on co-cultivation medium, oriented scutellum-up for 3-4 days. After that, embryos are transferred onto a selection-free medium containing antibiotics to suppress *Agrobacterium* growth and initiate callus formation. The calli are then placed on medium containing appropriate selection agents to suppress growth of non-
transformed tissue. Following the selection step, calli are placed on selection medium containing plant growth hormone for somatic embryo germination and plant regeneration. Isolated plantlets are moved to plant growth medium without selection agents. Eighty-six (86) transgenic events are produced and used to exemplify the detection method.

[0083] Genomic DNA (gDNA) is isolated from different maize events and untransformed maize controls. Several methods are employed to isolate the gDNA, such as the DNeasy kit (Qiagen, Valencia, CA) or the traditional Cetyl trimethylammonium bromide (CTAB) DNA isolation protocol. The DNA concentrations are determined using a Nanodrop (Thermo Scientific, Wilmington, DE). A total of 250ng of gDNA is digested with a restriction enzyme. Restriction enzymes that bind to degenerate recognition sequences, such as *AflIll* (recognizes the sequence ACRYGT), *BanI* (recognizes the sequence GGYRCC), *BsrYI* (recognizes the sequence RGATCY), *Styl* (recognizes the sequence CCWWGG), and *Stnll* (recognize the sequence CTYRAG), are used to achieve higher levels of cutting frequency which is comparable to four or six base pair cutting restriction enzymes. The resulting digestion fragments are of medium sized lengths, -1,000 bp. The digestion reaction can be further purified using the MinElute Reaction Cleanup Kit (Qiagen, Valencia, CA).

**Example 3**

[0084] Double stranded adapters with twelve (12) different index sequences (see Table 1 below; SEQ ID NO:14 Adapter 5': 5'- Phosphorylation-TYRAACGTGA TAGATCGGAA GAGCGGT-T-Inverted dT-3' which is annealed with its reverse complimentary sequence of the same index sequence SEQ ID NO:26 Adapter 3': 5'- CTCGGCATTC CTGCTGAACC GCTCTTCCGA TCTATCAGT3') are ligated to the digested gDNA pools. The ligation reaction is completed using the Quick Ligation kit (New England Biolabs, Ipswich, MA) per manufacturer's instructions. The ligation reaction is incubated at 25 °C for 30 minutes and then the reaction is stopped by heating the cocktail to 80 °C for 15 minutes. The reaction can be stored indefinitely at 4 °C.

<table>
<thead>
<tr>
<th>Table 1. list of index sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SEQ ID</strong></td>
</tr>
<tr>
<td>SEQ ID NO:14</td>
</tr>
<tr>
<td>SEQ ID NO:15</td>
</tr>
<tr>
<td>SEQ ID NO:16</td>
</tr>
<tr>
<td>SEQ ID NO:17</td>
</tr>
<tr>
<td>SEQ ID</td>
</tr>
</tbody>
</table>

21
| NO : | SEQ ID NO : | dT-3 ' | Inverted
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>5'- Phosphate-TYRAAGCCTAAAGATCGGAAGAGCGGTT-</td>
<td>Inverted</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>5'- Phosphate-TYRAATGGTCAAGATCGGAAGAGCGGTT-</td>
<td>Inverted</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>5'- Phosphate-TYRAACACTGTAGATCGGAAGAGCGGTT-</td>
<td>Inverted</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>5'- Phosphate-TYRAAATTGGCAGATCGGAAGAGCGGTT-</td>
<td>Inverted</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>5'- Phosphate-TYRAATTGGCAGATCGGAAGAGCGGTT-</td>
<td>Inverted</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>5'- Phosphate-TYRAATCAAGTAGATCGGAAGAGCGGTT-</td>
<td>Inverted</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>5'- Phosphate-TYRAACTGATCAGATCGGAAGAGCGGTT-</td>
<td>Inverted</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>5'- Phosphate-TYRAACTGATCAGATCGGAAGAGCGGTT-</td>
<td>Inverted</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTATCACGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTTAGCTTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTGGCTACT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTTTAGGCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTTGACCAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTACAGTGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTACAGTGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTACAGTGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTACAGTGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTACAGTGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTACAGTGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>CTCGGCATTCCTGCTGAACCGCTCTTCGATCTACAGTGT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 4

[0085] Primer extension reactions using the adapter ligated gDNA were completed. A gene specific primer is synthesized by Integrated DNA Technologies Inc. (Coralville, IA) and used for the reaction (SEQ ID NO:38 4468-3PA01-2Btm: 5'-\lambdaDual BiotinV ACACTCTTTCCCTACACGAC GCTCTTCCGA TCTCATATTAA AACGTCGCGCA -3'); The Platinum Taq kit (Invitrogen, Carlsbad, CA) is used to synthesize a DNA strand via primer extension. The following reagents: 1 µL, 10X platinum TAQ buffer; 0.2 µL, 10 mM dNTP; 0.1 µL, 10 µM 4468-3PA01-2Btm; 0.05 µL Platinum TAQ; 6.7 µL H₂O ; 2 µL of adaptor-ligated product are mixed in a tube. Amplification is completed using the following reaction conditions: 1) 94 °C 2.5 minutes; 2) 94 °C 30 seconds; 3) 69 °C 6 minutes; 4) repeat steps 2-3 five times; 5) 4 °C hold.

[0086] A capture reaction is completed with 2.0 µL of DYNABEADS® M-270 streptavidin magnetic beads (Invitrogen, Carlsbad, CA). The beads are washed on a magnet with PBST buffer (phosphate buffered saline and tween 20) three times. The beads are resuspended in 10 µL of PBS (phosphate buffered saline) containing 0.1% Tween 20 and 0.5 µg of 10 mg/ml of sheared salmon sperm DNA, the beads are mixed vigorously in the PBS solution. The PBS solution is added to the primer extension reaction at a 1:1 concentration, 20µE of beads are mixed with 20µE of primer extension reaction. The resulting solution is incubated for thirty minutes with gentle pipetting at room temperature. The beads containing the primer extension reaction are then washed over a magnet three times with 50 µL of washing buffer (10mM Tris (pH 8.0) - 1mM EDTA containing 0.1% Tween 20) and then one time with 50 µL of dd-H₂O. All of the wash solutions are removed from the beads.

Example 5

[0087] NGS sequencing libraries are constructed following conventional protocols. PCR reactions are completed using the Takara EX TAQ HS PCR kit (Millipore, Billerica, Ma). The following primers are used to amplify the event and flanking sequence prior to high-throughput sequencing in Mi-Seq (Illumina): Paired End (PE) PCR Primer 1.0 (Illumina), (SEQ ID NO:39 5'- AATGATACGG CGACCACCAGA GTCTACACT CTTTCCCTAC ACGACGCTCT -3') and PE PCR Primer 2.0 (Illumina), (SEQ ID NO:40 5'- CAAGCAGAAG ACGGCATACT CGATCGGTC TCCGATTCCT GCTGAACCGC -3')
Table 2: PCR amplification conditions

<table>
<thead>
<tr>
<th>Conditions</th>
<th>1 cycle</th>
<th>2 min</th>
<th>1 cycle</th>
<th>10 sec</th>
<th>68 °C, 6 min</th>
<th>30 cycles</th>
<th>98 °C, 10 sec</th>
<th>68 °C, 4 min</th>
<th>1 cycle</th>
<th>72 °C, 4 min</th>
<th>1 cycle</th>
<th>4 °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cycle</td>
<td>94 °C</td>
<td></td>
<td>98 °C</td>
<td>10 sec</td>
<td></td>
<td></td>
<td>68 °C</td>
<td></td>
<td></td>
<td>72 °C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 cycle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>68 °C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[0088] The following reagents are used in the PCR reaction: 5 µL, 10X EX-TAQ HS buffer; 4 µL, 2.5 mM dNTP, 1 µL, 10 µM transgene specific primer; 1 µL, 10 µM adapter specific primer; 0.25 µL, EX-Taq HS polymerase; and, 38.75 µL, H₂O. The cocktail is added to the washed beads from the ligation reaction, resuspended well and amplified using the following conditions in Table 2.

Example 6

[0089] The resulting PCR products are column purified and sequenced directly in MiSeq using PE Read 1 Sequencing Primer (SEQ ID NO:41 5' - ACACTCTTTC CCTACACGAC GCTCTTCCGA TCT-3') and PE Read 2 Sequencing Primer (SEQ ID NO:42 5' - CGGTCTCGGC ATTCCTGCTG AACCGCTCTT CCGATCT-3'). The high quality NGS sequence reads containing appropriate index and T-DNA Left Border sequences are trimmed and mapped in the maize genome. The 3' transgene insert and maize genomic flanking sequences from eighty-six (86) events are isolated and identified using the technique described above.

[0090] The characterization of the genomic insertions indicates that event 106685[1]-001 contains one copy of the transgene. The sequencing data of the chromosomal flanking region (SEQ ID NO:6) indicate that one copy of the transgene inserted into a unique location of the Zea mays genome. The site of insertion of the 3' end is mapped to maize chromosome 10: 140567138..140568099, (SEQ ID NO:7).

[0091] The characterization of the genomic insertions indicated that event 106685[1]-010 contains one copy of the transgene. The sequencing data of the chromosomal flanking region (SEQ ID NO:8) indicate that one copy of the transgene inserted into a unique location of the
**Zea mays** genome. The site of insertion of the 3’ end is mapped to maize chromosome 8:161488669–161489160 (SEQ ID NO:9).

[0092] The characterization of the genomic insertions indicated that event 106685[1]-013 contains one copy of the transgene. The sequencing data of the chromosomal flanking region (SEQ ID NO:10) indicate that one copy of the transgene inserted into a unique location of the **Zea mays** genome. The site of insertion of the 3’ end is mapped to maize chromosome 1:43455257–43456230 (SEQ ID NO:11).

[0093] The characterization of the genomic insertions indicated that event 106685[1]-035 contains one copy of the transgene. The sequencing data of the chromosomal flanking region (SEQ ID NO:12) indicate that one copy of the transgene inserted into a unique location of the **Zea mays** genome. The site of insertion of the 3’ end is mapped to maize chromosome 1:2075090–2076004 (SEQ ID NO:13).

**Example 7**

[0094] A plasmid containing a polynucleotide of interest expression cassette and a selectable marker gene expression cassette is used to transform **Zea mays** cv Hi-II plant tissue via the Biorad gene gun. Frame, *et al.*, Production of transgenic maize from bombarded Type II callus: effect of gold particle size and callus morphology on transformation efficiency. *In Vitro Cell. Dev. Biol-Plant.* 36:21-29). The protocol is modified: media components, selection agents and timing are optimized to improve the efficiency of the transformation process. An *Fsp I* linearized fragment of the plasmid is used for the transformation. The resulting transformations produced transgenic maize plants which contains a polynucleotide of interest expression cassette which is linked to the plant selectable marker gene expression cassette. The following transgenic events are produced; 106685[1]-001, 106685[1]-010, 106685[1]-013, and 106685[1]-035. These events are used to exemplify the detection method.

[0095] Genomic DNA (gDNA) is isolated from different maize events and untransformed maize controls. Several methods are employed to isolate the gDNA, such as the DNeasy kit (Qiagen, Valencia, CA) or the traditional Cetyl trimethylammonium bromide (CTAB) DNA isolation protocol. The DNA concentrations are determined using a Nanodrop (Thermo Scientific, Wilmington, DE). A total of 250ng of gDNA is digested with a restriction enzyme. Restriction enzymes that bind to degenerate recognition sequences, such as *AfIIlll* (recognizes the sequence ACRYGT), *Banl* (recognizes the sequence GGYRCC), *BstYI* (recognizes the sequence RGATCY), and *Styl* (recognizes the sequence CCWWGG), are used to achieve higher levels of cutting frequency which is comparable to four or six base
pair cutting restriction enzymes. The resulting digestion fragments are of medium sized lengths, -1,000 bp. The digestion reaction can be further purified using the MinElute Reaction Cleanup Kit (Qiagen, Valencia, CA).

**Example 8**

A double stranded adapter (SEQ ID NO: 1 Adapter 5’: 5’-Phosphorylation-GYRCAGCGGATCGTCT-Inverted dT-3’ which was annealed with SEQ ID NO: 2 Adapter 3’: 5’-GTCCGACCGTCAGAGAATCCAATAGACGATCCGCT-3’) is ligated to the digested gDNA. The ligation reaction is completed using the Quick Ligation kit (New England Biolabs, Ipswich, MA) per manufacturer’s instructions. The ligation reaction is incubated at 25 °C for 30 minutes and then the reaction is stopped by heating the cocktail to 80 °C for 15 minutes. The reaction is stored indefinitely at 4 °C.

**Example 9**

A gene specific primer is synthesized by Integrated DNA Technologies Inc. (Coralville, IA) and used for the reaction (SEQ ID NO: 3 4468-3PA01-2Btn: 5’-\Dual BiotinV GGACAGAGCCACAAACACCACAAGA-3’). The Platinum Taq kit (Invitrogen, Carlsbad, CA) is used to synthesize a DNA strand via primer extension. The following reagents: 1 µE, 10X platinum TAQ buffer; 0.2 µE, 10 mM dNTP; 0.1 µE, 10 µM 4468-3PA01-2Btn; 0.05 µE Platinum TAQ; 6.7 µE H₂O ; 2 µE of adaptor-ligated product are mixed in a tube. Amplification is completed using the following reaction conditions: 1) 94 °C 2.5 minutes; 2) 94 °C 30 seconds; 3) 69 °C 6 minutes; 4) repeat steps 2-3 five times; 5) 4 °C hold.

**Example 10**

A capture reaction is completed with 2.0 µE of DYNABEADS® M-270 streptavidin magnetic beads (Invitrogen, Carlsbad, CA). The beads are washed on a magnet with PBST buffer (phosphate buffered saline and tween 20) three times. The beads are resuspended in 10 µE of PBS (phosphate buffered saline) containing 0.1% Tween 20 and 0.5 µ/l of 10 mg/ml of sheared salmon sperm DNA, the beads are mixed vigorously in the PBS solution. The PBS solution is added to the primer extension reaction at a 1:1 concentration, 20µE of beads are mixed with 20µE of primer extension reaction. The resulting solution is incubated for thirty minutes with gentle pipetting at room temperature. The beads containing the primer extension reaction are then washed over a magnet three times with 50 µl of washing buffer (10mM Tris (pH 8.0) - 1mM EDTA containing 0.1% Tween 20) and then one time with 50 µl of dd-H₂O. All of the wash solutions are removed from the beads.
Example 11

[0099] PCR reactions are completed using the Takara EX TAQ HS PCR kit (Millipore, Billerica, Ma). The following primers are used to amplify the event and flanking sequence: Transgene specific primer, (SEQ ID NO:4 PAT-InvPriF: 5’- CGCTTACGATTGGACAGTTGAGATCTG-3’) and Adaptor primer, (SEQ ID NO:5 lmAdp-Pri: 5’-GTCCGACCGTCAGAGAATCCAAT-3’).

<table>
<thead>
<tr>
<th>Table 3: PCR amplification conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cycle</td>
</tr>
<tr>
<td>1 cycle</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>30 cycles</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1 cycle</td>
</tr>
<tr>
<td>1 cycle</td>
</tr>
</tbody>
</table>

[00100] The following reagents are used in the PCR reaction: 5 µE, 10X EX-TAQ HS buffer; 4 µE, 2.5 mM dNTP, 1 µE, 10 µM transgene specific primer; 1 µE, 10 µM adapter specific primer; 0.25 µE EX-Taq HS polymerase; and, 38.75 µE, H₂O. The cocktail is added to the washed beads from the ligation reaction, resuspended well and amplified using the following conditions in Table 3.

Example 12

[00101] The resulting PCR products are cloned into plasmid pCR2.1 (Invitrogen, Carlsbad, CA). Colonies are isolated and the pCR2.1 plasmid is confirmed to contain a PCR amplicon. The vectors are sequenced using M13 Forward and M13 Reverse primers. The sequencing results are expected to contain the nucleotide sequence of the maize 3’ genomic flanking sequence in addition to the genetic elements present from the plasmid. The 3’ transgene insert and maize genomic flanking sequences from events 106685[1]-001, 106685[1]-010, 106685[1]-013, and 106685[1]-035 are isolated and identified using the technique described above.

[00102] The characterization of the genomic insertions indicated that event 106685[1]-001 contains one copy of the transgene. The sequencing data of the chromosomal flanking region (SEQ ID NO:6) indicate that one copy of the transgene inserted into a unique location of the
Zea mays genome. The site of insertion of the 3’ end is mapped to maize chromosome 10: 140567138..140568099, (SEQ ID NO:7).

[00103] The characterization of the genomic insertions indicated that event 106685[1]-010 contains one copy of the transgene. The sequencing data of the chromosomal flanking region (SEQ ID NO:8) indicate that one copy of the transgene inserted into a unique location of the Zea mays genome. The site of insertion of the 3’ end is mapped to maize chromosome 8:161488669..161489160 (SEQ ID NO:9).

[00104] The characterization of the genomic insertions indicated that event 106685[1]-013 contains one copy of the transgene. The sequencing data of the chromosomal flanking region (SEQ ID NO:10) indicate that one copy of the transgene inserted into a unique location of the Zea mays genome. The site of insertion of the 3’ end is mapped to maize chromosome 1:43455257..43456230 (SEQ ID NO:11).

[00105] The characterization of the genomic insertions indicated that event 106685[1]-035 contains one copy of the transgene. The sequencing data of the chromosomal flanking region (SEQ ID NO:12) indicate that one copy of the transgene inserted into a unique location of the Zea mays genome. The site of insertion of the 3’ end is mapped to maize chromosome 1:2075090.-2076004 (SEQ ID NO:13).
We claim:

1. A method for use in a computerized system for identifying/recovering genomic insertion sites, comprising:
   (a) generating a three dimensional matrix with a pre-selected number of coordinates; 
   (b) generating a sequence database of genomic insertion-sites using a sequencing module; 
   (c) assigning a coordinate to each of genomic insertion-sites of the sequence database to be screened in the matrix; and 
   (d) pooling each of genomic insertion-sites into its assigned coordinate vertically, horizontally, and laterally.

2. The method of claim 1, wherein the three dimensional matrix comprises a plural of vertical columns, a plural of horizontal floors, and a plural of laterally layers.

3. The method of claim 1, further comprising displaying the three dimensional matrix in a color-coded mode, where vertical columns, horizontal floors, and laterally layers are displayed in different colors.

4. The method of claim 1, wherein the pre-selected number of coordinates is bigger than 20.

5. The method of claim 1, wherein the sequencing module comprises a high throughput sequencing system.

6. The method of claim 1, wherein the sequencing module comprises a next generation sequencing system.

7. The method of claim 1, wherein the sequencing module performs multiplex sequencing.

8. The method of claim 7, wherein sequencing data from the multiplex sequencing are imported in parallel into the sequence database.

9. The method of claim 2, wherein the sequencing module identifies/recovers genomic insertion-sites using a method comprising:
(i) shearing or cleaving genomic DNA into fragments;
(ii) performing primer extensions;
(ii) capturing primer extension products;
(iv) ligating adaptors with identifiers to the primer extension products;
(v) optionally amplifying the ligated products; and
(vi) sequencing the ligated products.

10. The method of claim 9, wherein the fragments of DNA comprise longer than 50 bp DNA.

11. The method of claim 9, wherein the primer extension products are captured using magnetic beads or biotin.

12. The method of claim 9, wherein the ligated products are amplified using polymer chain reaction (PCR) where adaptors with identifiers are used as primers for amplification.

13. The method of claim 9, wherein the ligated products are sequences using a next generation sequencing system.

14. The method of claim 13, wherein sequencing data of the ligated products are imported in parallel into the sequence database.

15. The method of claim 9, wherein each of the columns, floors, and layers is associated with a particular identifier in the adapters.
Disrupted/enhanced gene to be found
Shear to 300-400 bp
Extend using biotinylated primers
Capture biotinylated extensions
Ligate adapters with barcodes (for each pool)
PCR
QC, NGS
Combine X pools per channel

Figure 2
INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

INV. C12Q1/68
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data, BIOSIS, EMBASE, Sequence Search

C. DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>wo 2007/055568 AI (KEYGENE NV [NL]; VAN EJK MICHAEL JOSEPHUS THER [NL]; GERATS ANTONIUS) 18 May 2007 (2007-05-18) pages 11,27 pages 28,32 pages 30,33; claim 1; figure 11 pages 18,19</td>
<td>1-15</td>
</tr>
</tbody>
</table>

Further documents are listed in the continuation of Box C. See patent family annex.

Date of the actual completion of the international search
8 February 2013

Date of mailing of the international search report
22/02/2013

Name and mailing address of the ISA/
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040; Fax: (+31-70) 340-3016

Authorized officer
Reuter, Uwe
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>

A


A

<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CN 101310024 A</td>
<td>19-11-2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EP 1960541 Al</td>
<td>27-08-2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2009515518 A</td>
<td>16-04-2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2009208943 Al</td>
<td>20-08-2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 2007055568 Al</td>
<td>18-05-2007</td>
</tr>
</tbody>
</table>