

## NUCLEIC ACID INTERACTION ANALYSIS

### Field of the Invention

[001] The present invention generally relates to the field of gene expression. More specifically, the present invention relates to nucleic acid interactions. In particular, the invention relates to analysis, detection and identification of nucleic acid-protein interaction events and components

### Background of the Invention

[002] Chromatin interaction is important in gene regulation. The recently completed human genome sequences provide frameworks of genetic information. However, the human genome structure and information is often presented as a one-dimensional linearity, which is short to explain the complexity and coordination of a cellular system. An entirely different perspective is required for understanding how a genome actually functions as an orchestrated system in a three-dimensional nucleus in living cells. Genomic DNA (estimated to be two meters long stretched out) is condensed in chromosomes only a few microns across in a nucleus. It is known that chromosomes are unevenly organized into euchromatins and heterochromatins, which are packaged by chromatin proteins and communicated by transcription factors for transcription and replication.

[003] These activities appear to be ordered. It has been observed that large chromosomal loops contain active genes. Furthermore, it has been suggested that distal regulatory elements such as locus control regions (LCR), enhancers, and insulators act by repositioning specific genetic loci to regions with active or silent transcription. Recent work has demonstrated in  $\beta$ -globin, and most recently, in cytokine gene (IFN- $\gamma$ ) that LCRs may directly interact with promoters in long distances on the same chromosome and even interact with promoters on different chromosomes. It is possible that intra- and inter-chromosomal interactions are general phenomena occurring at multiple genetic loci in coordinating gene regulation of important pathways. Inter-chromosomal interactions have also been

implicated in diseases. For example, dysregulation of myc transcript is achieved by chromosomal translocations that juxtapose the c-myc/pvt-1 locus on chromosome 15 with one of the immunoglobulin loci on chromosome 12. Further analysis of interchromosomal interactions at whole genome level is necessary to identify all interactions, and will shed light on high-order gene regulations in cells.

[004] Technologies used for studies of chromatin interactions - A number of approaches have been used to study the three-dimensional structure and chromatin interactions, all with considerable limitations. Technologies applicable to this question may be roughly classified as visualization tools, such as microscopy, Fluorescence In Situ hybridization (FISH), and RNA-TRAP (RNA Tagging and Recovery of Associated Proteins); and molecular methodologies, such as chromosome conformation capture (3C), and 3C followed by chromatin immunoprecipitation (3C-ChIP).

[005] Microscopy was used in many early studies to investigate chromatin spatial organization in nuclei. However, such cytogenetic approach may only provide rough segment information of chromatins in chromosomes. FISH is a significant improvement in this direction, which localizes specific genetic loci to particular physical locations on chromosomes through fluorescence labelled DNA or RNA probes hybridizing to genomic DNA. However, the resolution was very still limited. Modified from FISH, RNA-TRAP is a method that may show distal enhancers in close physical proximity with gene promoters.

[006] Chromosome Conformation Capture (3C) was originally designed to investigate chromosomal conformation in yeast (Dekker et al, 2002), and has been used to study interactions of genetic elements that are separated in long distance and/or in different chromosomes. In 3C, DNA-protein (chromatin) structures are formaldehyde cross-linked in vivo, and chromatins are fragmented by restriction enzyme digestion. DNA fragments tethered by DNA binding proteins are then joined together by ligation, and the junctions of two suspected known elements are detected by PCR. The detection of chromatin interactions mediated by specific DNA binding protein or transcription factors may be further enhanced by chromatin immunoprecipitation (3C-ChIP), in which the chimerical DNA fragments cross-linked with protein resulted in the 3C procedure are enriched by antibody pull-down.

[007] Though each individual technique and some combinations have been demonstrated to be useful in identifying some specific intra- and inter-chromosomal interactions, these approaches relied on existing knowledge or conjecture as to what possible distal chromatin interactions may be present and primers designed to detect such junctions by PCR one region at a time. Therefore, the current technologies for study of chromatin interactions are extremely limited for identification of novel chromatin interactions and large scale at whole genome level.

[008] Despite considerable interest in the way that chromosomes are spatially organized within the nucleus and how that may regulate transcription of distal genes in concert, only scattered and indirect information is currently available. The lack of information in this aspect is largely due to the lack of robust technologies that may effectively address three-dimensional questions of chromosomal interactions.

[009] There is a need in the art for more efficient methods and robust technologies that may effectively address three-dimensional questions of chromosomal interactions that may overcome the disadvantages and limitations of the existing art.

### Summary of the Invention

The present invention solves the problems mentioned above by providing a new method of detecting, identifying and/or preparing at least one nucleic acid sequence or fragment from a nucleic acid complex, in particular from a nucleic acid-protein complex. In particular, the method according to the invention provides a method of detecting, identifying and/or preparing at least two nucleic acid sequences or fragments from a nucleic acid-protein complex. The present invention also relates to oligonucleotides prepared with a method according to any embodiment of the invention. There is also provided a method to identify chromatin interaction events mediated by specific DNA binding proteins across long distances and between different chromosomes.

In another aspect, the present invention solves the problem mentioned above by providing an isolated oligonucleotide comprising at least one first tag and at least

one second tag, wherein the first and second tags are tags of a nucleic acid-protein complex. In particular, there is provided an isolated oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex. The first and second polynucleotides may be part of the same nucleic acid region or from different nucleic acid regions of the nucleic acid-protein complex.

The isolated oligonucleotide may further comprise at least one restriction enzyme recognition site at least one linker. In particular, the at least one restriction enzyme recognition site may be included in a linker. The linker may be inserted between the tags or the linker may flank at least one tag (that is, positioned upstream and/or downstream to at least one tag). The at least one restriction enzyme recognition site may be asymmetric. The at least one restriction enzyme recognition site may be, for example, for a type IIs restriction enzyme or for a homing restriction enzyme.

[0010] The at least one first tag may comprise a 5' terminus and a 3' terminus from the first polynucleotide and the at least one second tag comprises a 5' terminus and a 3' terminus from the second polynucleotide. The isolated oligonucleotide may further comprise at least one linker. The linker may be inserted between the tags or the linker may positioned upstream and/or downstream to at least one tag. The linker comprises at least one restriction enzyme recognition site. The at least one restriction enzyme recognition site may be asymmetric. The at least one restriction enzyme recognition site may be, for example, for a type IIs restriction enzyme or for a homing restriction enzyme.

The linker may comprise a first restriction recognition site recognized by a restriction enzyme capable of cleaving the first polynucleotide to obtain the first tag, and a second restriction recognition site recognized by a restriction enzyme capable of cleaving the second polynucleotide to obtain the second tag. In particular, the linker may comprise a first restriction recognition site recognized by a restriction enzyme capable of cleaving the first polynucleotide to obtain a 3' terminus of the first polynucleotide, and a second restriction recognition site recognized by a second restriction enzyme capable of cleaving the second

polynucleotide to obtain a 5' terminus of the second polynucleotide. The first and second restriction recognition sites may be recognized by the same or different restriction enzyme.

[0011] According to another aspect, the first polynucleotide may be further cleaved by a third restriction enzyme recognizing a third recognition site to obtain a 5' terminus of the first polynucleotide, and the second polynucleotide is cleaved by a fourth restriction enzyme recognizing a fourth recognition site to obtain a 3' terminus of the second polynucleotide. According to this embodiment, at least one first tag is obtained from ligating the 5' and 3' termini of the first polynucleotide and at least one second tag is obtained from ligating the 5' and 3' termini of the second polynucleotide. The third and fourth recognition sites may be recognized by the same or different restriction enzymes. The further recognition sites may be included in adaptors ligated to the 5' end of the first polynucleotide and to the 3' end of the second polynucleotide, respectively. Alternatively, the third and fourth restriction sites may be present in a vector into which the structure first polynucleotide-linker-second polynucleotide is inserted. In this case, the third restriction site flanks the 5' end of the first polynucleotide and the fourth restriction site flanks the 3' end of the second polynucleotide.

[0012] The nucleic acid-protein complex of the isolated oligonucleotide may be part of a chromatin structure. The nucleic acid fragment to which a protein of interest binds may be any nucleic acid fragment comprising a region to which a protein of interest binds, for example, histone binding site. The polynucleotides may be DNA or RNA.

In another aspect, the present invention provides a concatemer of oligonucleotides comprising at least two isolated oligonucleotides, each isolated oligonucleotide comprising at least one first tag and at least one second tag, wherein the first and second tags are tags of a nucleic acid-protein complex. IN particular, there is provided a concatemer of oligonucleotides comprising at least two isolated oligonucleotides, each isolated oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is a tag of a first polynucleotide and the second tag is a tag of a second polynucleotide, the first and second polynucleotides are from a nucleic acid-protein complex. The concatemer may further comprise at least one linker. The linker may comprise at

least one restriction enzyme recognition site. Further, each isolated oligonucleotide of the concatemer may comprise at least one restriction enzyme recognition site. The at least one restriction enzyme recognition site may be comprised in at least one linker and/or in at least one adapter.

The linker may be inserted between the tags or the linker may be positioned upstream and/or downstream to at least one tag. The linker may comprise at least one restriction enzyme recognition site. The at least one restriction enzyme recognition site may be for a type IIs restriction enzyme or for a homing restriction enzyme. The first tag of each oligonucleotide of the concatemer may comprise a 5' terminus and a 3' terminus from the first polynucleotide, and the second tag may comprise a 5' terminus and a 3' terminus from the second polynucleotide.

[0013] The oligonucleotide or the concatemer according to the invention may be inserted into a vector and/or a cell. The cell may be a bacterial cell.

[0014] The the polynucleotide nucleic acid-protein complex may be part of a chromatin structure. The polynucleotides may be located on the same chromosome or may be located on different chromosomes.

[0015] In another aspect, the present invention provides a library of oligonucleotides or concatemer of oligonucleotides comprising at least one oligonucleotide, the oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex.

[0016] The at least one oligonucleotide of the library may comprise at least one linker. The linker may be inserted between the tags or the linker may be positioned upstream and/or downstream to at least one tag. The first tag may comprise a 5' terminus and a 3' terminus from the first polynucleotide and the second tag further comprises a 5' terminus and a 3' terminus from the second polynucleotide.

In another aspect, the present invention provides a method of preparing at least one isolated oligonucleotide comprising:

- (a) providing a nucleic acid-protein complex;
- (b) preparing an oligonucleotide comprising at least one first tag and at least one second tag, wherein the first and second tags are obtained from a nucleic acid-protein complex; and

(c) isolating the oligonucleotide.

In particular, the present invention provides a method of preparing at least one isolated oligonucleotide comprising:

- (a) providing a nucleic acid-protein complex;
- (b) preparing an oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex; and
- (c) isolating the oligonucleotide.

[0017] In one embodiment, step (b) of this aspect of the invention comprises:

- (i) inserting at least one linker comprising at least one restriction enzyme recognition site, and
- (ii) cleaving the first and second polynucleotides with a least one restriction enzyme recognizing the at least one recognition site in the linker to form an oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide and the linker between the tags.

[0018] In another embodiment, step (b) of this aspect of the invention comprises:

- (i) inserting at least one linker comprising at least one restriction enzyme recognition site between the first polynucleotide and the second polynucleotide of the complex;
- (ii) adding at least one restriction enzyme recognition site to the 5' terminus of the first polynucleotide and to the 3' terminus of the second polynucleotide;
- (iii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing at least one recognition site to obtain cleaved fragments; and
- (iv) ligating cleaved fragments to form the oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide, the tags comprising 5' and 3' termini of each polynucleotide, and the linker inserted between the tags and the at least one restriction enzyme recognition site in step (ii) is part of an adaptor or part of a vector.

[0019] The polynucleotides may be obtained from the nucleic acid-protein complex by incorporating a photoactivatable moiety into the nucleic acid (for example, DNA) and/or the protein of interest and isolation of nucleic acid/protein complex by

antibody-mediated precipitation or by affinity-mediated technique. Examples of such affinity-based techniques include streptavidin/ biotin, Glutathione-S-transferase/ glutathione matrix, maltose-binding protein/ amylose matrix interactions.

[0020] In another aspect, the present invention provides a method of detecting and/or identifying at least two polynucleotides from a nucleic acid-protein complex, the method comprising:

- (a) providing a nucleic acid-protein complex;
  - (b) preparing an oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex;
  - (c) sequencing the oligonucleotide; and
  - (d) mapping the at least two polynucleotides based on the nucleotide sequences of the first and second tags,
- thereby detecting and/or identifying the at least two polynucleotides.

[0021] The oligonucleotide obtained in step (b) may be amplified before being sequenced in step (c). The amplification may be by polymerase chain reaction. The amplified oligonucleotides may be subjected to at least one purification step after amplification but before being sequenced in step (c). The at least one purification step may be gel electrophoresis.

[0022] The oligonucleotide according to the invention may be concatenated with at least one further oligonucleotide obtained by steps (a) to (b) before being sequenced.

[0023] The sequencing may be carried out by the Sanger method or by multiplex sequencing such as pyrosequencing.

[0024] The detecting and/or identifying may be for transfection or translocation of the polynucleotides. Accordingly, the method may be used to detect and/or identifying polynucleotides and/or genes in proximity with each other in nucleic acid-protein complexes such as those in chromatin. In addition, the nucleic acid fragment to which a protein of interest binds may be any nucleic acid fragment comprising a region to which a protein of interest binds, for example, histone binding site. The polynucleotides may be DNA or RNA.

[0025] The oligonucleotide may be transfected into at least one cell. The transfection may be by electroporation. The cell may be a bacterial cell. The polynucleotides may be DNA or RNA.

[0026] In another aspect, the present invention provides a vector comprising the oligonucleotide, the concatemer of oligonucleotides or the library of oligonucleotides or concatemers.

### Brief Description of the Drawings

[0027] FIG.1 illustrates the overview of one method of the present invention for Chromatin Interaction Analysis by paired end (di)tag(s) (CIA-PET) sequencing.

[0028] FIG.2 illustrates the overview of another method of the present invention for CIA-diPET method.

[0029] FIG.3 illustrates mapping of CIA-PET. CIA-PETs representing real interacting regions are expected to span two different genomic regions, and multiple CIA-PETs are expected to be clustered.

[0030] FIG. 4 illustrates the relevant restriction enzyme recognition sequences for CIA-PET. Each base indicated with N or X may be any nucleotide (either A, C, G or T). The regions comprising N and X represent portions obtained from different polynucleotides. The sequences which appear in the figure from top to bottom are as follows: SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:13, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:14, SEQ ID NO:24, SEQ ID NO:25 (SEQ ID Nos:24 and 25 repeated thrice).

[0031] FIG.5 illustrates mapping CIA-diPET. CIA-diPETs representing the interacting regions are expected to span two different genomic regions, and multiple CIA-diPETs are expected to be clustered. The additional information provided by the two PETs of the CIA-diPET will be useful in mapping the CIA-diPETs to the genome.

[0032] FIG. 6 illustrates the relevant restriction enzyme recognition sequences for CIA-diPET. The bases N and X represent any nucleotide (either A, C, G or T) and they are obtained from different polynucleotides. The numerals 1 and 2 represent nucleotides from two regions or termini obtained from one polynucleotide while the numerals 3 and 4 represent nucleotides from two regions or termini obtained from another polynucleotide. The numerals (1, 2, 3, and/or 4) may represent any nucleotide (either A, C, G or T). The sequences which appear in the figure from top to bottom are as follows: SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, SEQ ID NO:35 (SEQ ID Nos:24 and 25 repeated thrice).

[0033] Fig.7 illustrates the vector pGIS8 containing multiple unique cloning sites, which is an example of a vector that may be used under the present invention.

[0034] FIG. 8 illustrates how oligonucleotides that do not have a linker inserted will not be cut by the restriction enzymes and hence are too long and will be removed by electrophoresis. The sequences which appear in the figure from top to bottom are as follows: SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44 and SEQ ID NO:45.

[0035] FIG 9 illustrates how if only a part of a linker is inserted into an oligonucleotide, it will result in that oligonucleotide that is incorrectly cut and hence will be removed by electrophoresis as it will be too long. The sequences which appear in the figure from top to bottom are as follows: SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52 and SEQ ID NO:53.

[0036] FIG. 10 illustrates how an incorrectly inserted linker will give rise to oligonucleotides that are too long and variable in length and hence can be removed by electrophoresis. The sequences which appear in the figure from top to

bottom are as follows: SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62 and SEQ ID NO:63.

[0037] FIG. 11 shows the nucleic acid sequence of the vector pGIS8 used in the examples with the various restriction enzyme recognition sites indicated. The sequences appear as SEQ ID NO:18 and SEQ ID NO:19 in the sequence listing.

### Detailed Description of the Invention

#### [0038] Definitions.

[0039] Restriction enzyme - A restriction enzyme (or restriction endonuclease) is an enzyme that cuts double-stranded DNA. The enzyme makes two incisions, one through each of the phosphate backbones of the double helix without damaging the bases. The chemical bonds that the enzymes cleave may be reformed by other enzymes known as ligases so that restriction fragments obtained from different chromosomes or genes may be spliced together, provided their ends are complementary. Type II enzymes recognized specific nucleic sequences and cut DNA at defined positions close to or within their recognition sequences sites. They produce discrete restriction fragments and distinct gel banding patterns. Type IIs enzymes cleave outside of their recognition sequence to one side. Mmel as well as most of the type IIs restriction enzymes produce variable end lengths (Dunn et al, 2002 showed that Mmel can cut 18/20 or 19/21 bases away in a rough proportion of 1:1. Therefore, the sequences given in all figures each represent one common variant of the use of Mmel.

[0040] The PET intermediates also have variable lengths (that is, M and G adapter DNA sequences after cloning into pGIS8 plasmids and M PET intermediate), because the polynucleotide can be of differing lengths. Type III enzymes are also large combination restriction-and-modification enzymes. They cleave outside of their recognition sequences and require two such sequences in opposite orientations within the same DNA molecule to accomplish cleavage. Homing endonucleases are rare double-stranded DNases that have large,

asymmetric recognition sites (12–40 base pairs) and coding sequences that are usually embedded in either introns (DNA) or inteins (proteins). Restriction enzymes may make cuts that leave either blunt end or sticky ends with overhangs. A sticky-end fragment can be ligated not only to the fragment from which it was originally cleaved, but also to any other fragment with a compatible cohesive or sticky end. As such, ends produced by different enzymes may also be compatible. A sticky end can thus be also referred as an end capable of being ligated. Many type II restriction enzymes cut palindromic DNA sequences. If a restriction enzyme has a non-degenerate palindromic cleavage site, all ends that it produces are compatible. A "palindromic" sequence is one where the sequence on one strand reads the same in the opposite direction on the complementary strand. As such, it is possible for nucleic acid strands treated to obtain palindromic cohesive ends to have its ends mate and the nucleic acid strands to self-circularize. The meaning of "palindromic" in this context is different from its linguistic usage. For example, the sequence GTAATG is not a palindromic DNA sequence, while the sequence GTATAC is. Examples of restriction enzymes leaving cohesive or sticky ends include BamH1, EcoR1 and HindIII. Examples of restriction enzymes leaving blunt, non-cohesive or non-sticky ends include BseR1 and AluI.

[0041] Nucleotide – a phosphoric ester of nucleoside; the basic structural unit of nucleic acids (DNA or RNA). Nucleotides form base pairs – one of the pairs of chemical bases joined by hydrogen bonds that connect the complementary strands of a DNA molecule or of an RNA molecule that has two strands; the base pairs are adenine with thymine and guanine with cytosine in DNA and adenine with uracil and guanine with cytosine in RNA. Short strands of nucleotides are referred to as oligonucleotides; longer strands are referred to as polynucleotides. Nucleotides may be joined with or concatenated with other nucleotides. The term nucleotide may be used interchangeably with the term nucleic acid. A stretch of nucleic acids possess a 5' end and a 3' end. The end regions of a stretch of nucleic acid may be referred to as the 5' terminus and the 3' terminus respectively. With 5' or 3' terminus of a polynucleotide, it is understood any region, fragment or whole piece of a polynucleotide that comprises the actual 5' or 3' terminus of the polynucleotide are included.

[0042] Concatemer – It is composed by at least two nucleotide monomers sequences linked end to end, optionally separated by a linker or spacer. For the purpose of the present invention, a concatemer comprises at least two oligonucleotides prepared according to the method of the invention.

[0043] Clone, cloning – To transfer nucleotides, such as a gene from one organism to another and/or to replicated the nucleotide by genetic engineering techniques.

[0044] Library - a collection of cloned nucleic acid sequences, oligonucleotides or polynucleotides, usually comprised in one or more plasmids.

[0045] Vector - A bacteriophage, plasmid, or other agent that transfers genetic material from one cell to another.

[0046] Obtain, derive – to use molecular biology and genetic engineering and manipulation techniques on biological material such as nucleic acids and proteins to confer upon the material certain desired characteristics. The terms obtain and derive may be used interchangeably under the present invention.

[0047] Amplification – increasing the copy number of nucleic acids. One method commonly used is that of polymerase chain reaction (PCR). Other amplification methods known to a skilled person may also be used.

[0048] Transfection or transformation - any method for introducing foreign molecules into a cell. Lipofection, calcium phosphate precipitation, retroviral deliver, electroporation and biolistic transformation are just a few of the teachings that may be used.

[0049] Chromatin - A complex of nucleic acids and proteins, primarily histones, in the cell nucleus that stains readily with basic dyes and condenses to form chromosomes during cell division. Chromatin is an example of a nucleic acid-protein complex. Regions of chromosomes may interact with other regions either on the same or different chromosome. The interaction event may thus be an inter- or intra- chromosomal event and may involve rearrangement of the genetic material at the regions involved.

[0050] Transfusion – rearrangement of genetic information at the RNA processing level to form a new chimerical transcript.

[0051] Translocation - rearrangement of genetic information at the genomic DNA level.

[0052] Nucleic acid-protein complex — an interaction of genetic material and a protein such as that found in chromatin or when a transcription factor binds to stretch of nucleic acid. A DNA-protein-DNA (DPD) complex is a more specific structure wherein a protein binds between two stretches of nucleic acid (DNA) of interest. A stretch of nucleic acid such as DNA may be manipulated to become a tag or an identifiable sequence of DNA.

[0053] Tag, tag-linker structure - A tag or signature is an identifiable sequence of nucleic acids refers to either the 5'- or 3'-most terminal nucleic acid sequence (terminus; usually 18-20 bp) derived from any contiguous DNA region, or a tag may comprise both the 5' and 3' most terminal nucleic acid sequences or termini of any contiguous DNA region. A linker is an artificial sequence of nucleic acids. The tag-linker-tag structure is thus an arrangement of nucleic acids wherein a linker is inserted between two tags. Another possible arrangement is a linker-tag-tag-linker structure where a linker flanks a tag (that is, it is positioned upstream and/or downstream to at least one of the tag). The terms tag and signature may be used interchangeably under the present invention.

[0054] Ditag — A short (usually 12-60 bp) nucleic acid fragment derived terminal tags or signatures of polynucleotides. A ditag may be prepared according to US 20050255501 and/or US 20050059022, the contents of which are herein incorporated by reference.

Sequencing - The methods used to determine the order of constituents in a biopolymer, in this case, a nucleic acid. Sequencing techniques used include Sanger method and modified variations thereof, as well as pyrosequencing or the "454 method" of sequencing.

[0055] In the following description, details and specific quantities and parameters are provided to describe the embodiments of the present invention. It shall be apparent to one skilled in the art, however that the invention may be practiced without such details. Some of the details may not be described at length so as not to obscure the invention.

[0056] For the performance of the methods of the present invention for a particular embodiment, any description disclosed for the purpose of carrying out other embodiments of this invention may also be used and are herein incorporated by reference. In particular, technique(s), reagents, experimental conditions,

restrictions sites, enzymes, vectors, primers, and the like. In particular, it will be evident to any skilled person how to adapt techniques and material disclosed for the other embodiments to the present embodiment of the invention.

[0057] A person skilled in the art will appreciate that techniques not specifically taught herein may be found in standard molecular biology reference books such as *Molecular Cloning: A Laboratory Manual* by Sambrook and Russell, Third Edition, 2001, published by Cold Spring Harbor Laboratory Press.

[0058] Description

[0059] The present invention relates to a new method of detecting, identifying and/or preparing at least one nucleic acid sequence or fragment from a nucleic acid complex, in particular from a nucleic acid-protein complex. In particular, the method according to the invention provides a method of detecting, identifying and/or preparing at least two nucleic acid sequences or fragments from a nucleic acid-protein complex. The invention also provides for oligonucleotides and/or concatamers of oligonucleotides.

[0060] According to one aspect, the present invention provides a method for chromatin interaction analysis (CIA). The CIA is designed to capture novel information about distal control regions and inter-chromosomal interactions de novo. This method is designed to identify chromatin interaction events mediated by specific DNA binding proteins, such as histones, across long distances and between different chromosomes.

[0061] Two embodiments of the CIA method are provided under the present invention to detect such DNA junctions, namely the CIA-PET method that extracts a single tag signature (about 20bp) from each of the two ligated DNA fragments to form a "tag1-linker-tag2" (also referred to as "first tag-linker-second tag) paired end ditag (PET) structure (Figure 1), and the CIA-diPET method that obtains two paired end ditags (PET) to represent the two related DNA fragments in a structure of "PET1-linker-PET2", so called diPET (Figure 2). The tags of CIA-PET and CIA-diPET may be directly sequenced a multiplex sequencing technique such as the "454" pyrosequencing method, or concatenated for cloning and sequencing using conventional sequencing method.

Accordingly, the present invention provides a method of preparing at least one isolated oligonucleotide comprising:

- (a) providing a nucleic acid-protein complex;
- (b) preparing an oligonucleotide comprising at least one first tag and at least one second tag, wherein the first and second tags are obtained from a nucleic acid-protein complex; and
- (c) isolating the oligonucleotide.

The first tag may be obtained from a first polynucleotide and the second tag is obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex.

The step (b) may comprise:

- (i) inserting at least one linker comprising at least one restriction enzyme recognition site, and
- (ii) cleaving the first and second polynucleotides with a least one restriction enzyme recognizing the at least one recognition site in the linker to form an oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide and the linker.

In particular, step (b) may comprise:

- (i) inserting at least one linker comprising at least one restriction enzyme recognition site between the first polynucleotide and the second polynucleotide of the complex;
- (ii) adding at least one restriction enzyme recognition site each to the 5' terminus of the first polynucleotide and the 3' terminus of the second polynucleotide;
- (iii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing at least one recognition site to obtain cleaved fragments; and
- (iv) ligating cleaved fragments to form the oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide, the tags comprising 5' and 3' termini of each polynucleotide, and the linker inserted between the tags.

[0062] The present invention also provides a method of detecting and/or identifying at least two polynucleotides from a nucleic acid-protein complex, the method comprising:

- (a) providing a nucleic acid-protein complex;
- (b) preparing an oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex;
- (c) sequencing the oligonucleotide; and
- (d) mapping the at least two polynucleotides based on the sequence, thereby detecting and/or identifying the at least two polynucleotides.

[0063] As the linker sequence carries two Type II restriction enzyme recognition sites at each end (FIGS 2 and 4), therefore, after a Type II restriction digestion, one tag sequence signature each (about 20bp) may be excised from the two linked DNA fragments, to form a tag-linker-tag structure, in which one tag represents one DNA region of a chromosome while the other tag represents a locus from a far away region on the same or in different chromosome. This paired end ditag structure is referred to as "CIA-PET", which may be efficiently sequenced through concatenation into longer stretch of DNA or directly analyzed by sequencing.

[0064] Alternatively, linker sequences may flank the two tags to yield a linker-tag-tag-linker structure.

[0065] In this method, the nucleic acid-protein complexes, such as native DNA-protein-DNA (DPD) complexes, are cross-linked by a suitable fixative such as formaldehyde, glutaldehyde or methanol. The cross-linked DPD complex may then be fragmented by sonication, hydroshearing (Hydroshear, Gene Machines), repeated drawing through a hypodermic syringe needle or by restriction enzyme digestion. The DNA fragments originated from different chromosomes or at long distance are tethered by DNA binding proteins in the DPD complexes. Unlike the 3C technique that requires existing knowledge or conjecture as to what these DNA fragments are so as to generate PCR primers, the ends of DNA fragments with distal relationship bounded in a DPD complex are joined by specific linker through ligation.

[0066] The linker (about 20bp) contains two Type II restriction enzyme recognition sites to join the ends of different DNA fragments tethered by proteins in each DPD complex. The DNA junction of the two related DNA fragments may then be tagged by the paired end ditagging (PET) strategy (US 20050255501). Type IIs restriction enzyme recognitions sites are preferred. Besides Type II restriction enzymes, any other suitable restriction enzyme, including Type III or homing restriction enzymes may be used.

[0067] Thus the present invention provides in one aspect, a method to identify chromatin interaction events mediated by specific DNA binding proteins, such as histones, across long distances and between different chromosomes. In another aspect, the present invention provides an isolated oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex. The tags correspond to regions of chromatin in nucleic acid-protein complexes. These tags may then be sequenced to analyse, identify, and/or detect chromatin interaction events (FIGS 3 and 5).

[0068] The isolated oligonucleotide may further comprises at least one linker. The linker may be inserted between the tags or a linker may be positioned upstream and/or downstream the at least one tag. The linker may comprises at least one restriction enzyme recognition site; the at least one restriction enzyme recognition site may be asymmetric, the at least one restriction enzyme recognition site may be for a type IIs restriction enzyme or for a homing restriction enzyme.

[0069] Alternatively, the at least one first tag may comprise a 5' terminus and a 3' terminus from the first polynucleotide and the at least one second tag comprises a 5' terminus and a 3' terminus from the second polynucleotide. The isolated oligonucleotide may further comprise at least one linker. The linker may be inserted between the tags or a linker may be positioned upstream and/or downstream the at least one tag. The linker comprises at least one restriction enzyme recognition site; the at least one restriction enzyme recognition site may be asymmetric, the at least one restriction enzyme recognition site may be for a type IIs restriction enzyme or for a homing restriction enzyme.

[0070] The linker may comprise a first restriction recognition site recognized by a restriction enzyme capable of cleaving the first polynucleotide to obtain the first tag, and a second restriction recognition site recognized by a restriction enzyme capable of cleaving the second polynucleotide to obtain the second tag (FIGS. 1 and 4).

[0071] The linker may comprise a first restriction recognition site recognized by a restriction enzyme capable of cleaving the first polynucleotide to obtain a 3' terminus of the first polynucleotide, and a second restriction recognition site recognized by a second restriction enzyme capable of cleaving the second polynucleotide to obtain a 5' terminus of the second polynucleotide. Alternatively, the linker may comprise a first restriction recognition site recognized by a restriction enzyme capable of cleaving the first polynucleotide to obtain a 5' terminus of the first polynucleotide, and a second restriction recognition site recognized by a second restriction enzyme capable of cleaving the second polynucleotide to obtain a 3' terminus of the second polynucleotide.

[0072] The first polynucleotide may be further cleaved by a third restriction enzyme recognizing a third recognition site to obtain a 3' terminus of first polynucleotide and the second polynucleotide may be further cleaved by a fourth restriction enzyme recognizing a fourth recognition site to obtain a 5' terminus of the second polynucleotide; the at least one first tag obtained from ligating the 5' and 3' termini of the first polynucleotide and the at least one second tag obtained from ligating the 5' and 3' termini of the second polynucleotide (FIGS. 2 and 6). The first and third recognition sites may possess the same sequence and the second and fourth recognition sites may possess the same sequence.

[0073] The nucleic acid-protein complex of the isolated oligonucleotide may be part of a chromatin structure. The nucleic acid fragment to which a protein of interest binds may be any nucleic acid fragment comprising a region to which a protein of interest binds, for example, histone binding site. The polynucleotides may be DNA or RNA.

[0074] The resulting CIA-PET sequences may be mapped to reference genome sequences to localize the chromatin interaction junction points, which may be either long range intra-chromosome (on the same chromosome) or inter-

chromosome (on different chromosomes). This information may be used for study of chromosomal 3-dimensional organization structures in nuclei, coordinated gene regulation in long distance and across different chromosomes mediated by transcription factors, and other epigenetic related questions (FIGS. 3 and 5).

[0075] With complete genome sequences available, it is possible to develop whole genome approaches to identify all potential chromatin interactions. Whole genome tiling array is an attractive approach for genome interrogation, in which 20-60mer oligonucleotides are tiled to cover the entire genome in microarrays, and DNA probes with biological contents are hybridized to the array, and the profiles of hybridization signal intensity of all the array elements is the readout for interrogation of genetic elements. Tiling arrays have been demonstrated useful for identifying exons, and for localizing transcription factor binding sites when coupled with ChIP (ChIP-chip). Though the array-based approach is efficient due to its highly multiplex and parallel nature, it is inconceivable that the hybridization-based detection will be capable of detecting non-linear relationship of two DNA fragments in chromatin interactions.

#### Preparation of the oligonucleotide according to one embodiment

[0076] The nucleic acid-protein complex may be obtained through Chromatin immunoprecipitation (ChIP).

#### Chromatin immunoprecipitation (ChIP)

[0077] ChIP has been used to enrich and thereby allow the identification of genomic regions associated with specific proteins such as histones and other proteins binding to nucleic acids in nucleic-acid protein complexes (reviewed in Taverner et al., Genome Biol, 2004. 5(3): p. 210). The aim is to cross-link proteins with DNA at their sites of interaction. This is accomplished quickly and efficiently by adding a suitable fixative such as formaldehyde, glutaldehyde or methanol directly to living cells in culture.

[0078] Crude extracts of these fixed cells are then prepared, and the chromatin sheared by sonication, hydroshearing, repeated drawing through a hypodermic syringe needle or by restriction enzyme digestion to an average size of usually about 1 kb, then used in immunoprecipitation reactions with antibodies raised

against the DNA-associated protein of interest (e.g. transcription factors or histones). DNA fragments enriched in each immunoprecipitation are then de-linked and purified to allow their identification by a variety of methods. The advantage of using ChIP is that this approach is able to "freeze" the in vivo gene regulatory network by rapid cross-linking of chromatin and other non-histone proteins, thereby in theory representing a "true" picture of the regulatory system at any point in time, free of potential artifacts imposed by heterologous expression, for instance.

[0079] Recently, ChIP has been combined with whole-genome (Lieb et al., *Nat Genet*, 2001. 28(4): p. 327-34), whole-chromosomal (Euskirchen et al., *Mol Cell Biol*, 2004. 24(9): p. 3804-14) and CpG island (Weinmann et al., *Genes Dev*, 2002. 16(2): p. 235-44) or microarrays in a "ChIP-chip" or "ChIP-on-chip" approach that promises to enable the genome-level localization of protein-binding sites such as transcription factor binding sites (TFBS) (reviewed in Buck and Lieb, 2004). While the usefulness of this approach has been demonstrated for small genomes such as yeast (Lieb et al., *Nat Genet*, 2001. 28(4): p. 327-34), the cost and complexity of producing whole-genome microarrays for more complex organisms still remains a limiting factor.

[0080] CpG island microarrays contain human genomic fragments of high CpG content, and because CpG islands often correspond to promoter regions (Antequera and Bird, *Proc Natl Acad Sci USA*, 1993. 90(24): p. 11995-9), such microarrays represent a possible compromise. However, the location of putative protein-binding sites still has to be indirectly inferred by examining genomic DNA upstream and downstream (usually 1-2 kb, as this is the approximate size of sonicated ChIP fragments) of the CpG-rich probe spotted on the array.

[0081] As an alternative, cloning and sequencing of the ChIP-enriched DNA fragments has previously been attempted but with limited success. The problem is that the targets of ChIP enrichment are obtained against a high background of the entire genome. Even a 100-fold enrichment of specific targets would still represent only a small fraction of clones in a ChIP library, making standard DNA sequencing a very costly solution. Therefore, sequencing ChIP clones under these circumstances is not a good approach for identifying the enriched targets. Serial Analysis of Gene Expression (SAGE) and Massively Parallel Signature

Sequencing (MPSS) (Brenner et al., Nat Biotechnol, 2000. 18(6): p. 630-4) have also been suggested as useful quantitative tools for detecting ChIP enrichment, the underlying principle being that the tags generated from ChIP-enriched DNA fragments would be present in larger numbers compared to the non-specific background.

[0082] These tags could then be mapped to the genome sequence for identification of the general region of interest (i.e. assumed to be 1-2kb, representing the sonicated fragments). Although the 20bp SAGE and MPSS tags should be specific enough in most instances to define the specific genome location, one still has to examine all sequences approximately 1-2kb upstream and downstream of the tag when mapping to the genome. This is the same problem faced by the CpG island microarray approach. Furthermore, complete coverage using these methods depends on the availability of prerequisite restriction enzyme recognition sites (mapping-enzyme sites); if a recognition site is absent from a certain genomic location, that particular tag will be missing from the corresponding ChIP fragment, and hence that location will be a "blind spot" within the genome.

[0083] From the issues described above, it is clear that what is required to facilitate genome level transcriptional regulatory analysis is a method to accurately and rapidly pinpoint the nucleic acid sequences flanking protein binding regions, as an alternative to whole genome arrays. In this regard, the novel approach provided by the present invention, possesses several advantages: (i) the tag sequences generated by one method of the present invention provide higher specificity for mapping, because each tag would already be known to have been derived from a contiguous DNA segment encompassed by the 5' and 3' signatures. This information facilitates precise localization of the genomic region of interest, and obviates the need to repeatedly examine every sequence an arbitrary 1-2kb upstream and downstream of a standard SAGE or MPSS tag; (ii) the method of the present invention is thus independent of any requirement for the presence of mapping-enzyme sites; (iii) the concatenation of tags prior to sequencing means that several tags may be identified within one sequencing read; (iv) the region that is common to (i.e. overlapped by) all mapped tags in that

cluster therefore defines the DNA regions involved in the nucleic acid-protein complex in question.

#### Tags and ditags

[0084] For the purpose of the present application, a tag is a nucleotide sequence or signature obtained from a nucleic acid molecule and represents the polynucleotide from which the tag was obtained or derived from. The polynucleotide which is intended to shrink or represent may be RNA, mRNA, genomic DNA, full-length cDNA, or cDNA.

[0085] Under the present invention, two tags that are present in an oligonucleotide of the present invention may also each be called a ditag. Like tags, a ditag is shorter than the original nucleic acid molecule from which it originates or which it represents. Preferably, the ditag must be much shorter than the original nucleic acid molecule. As consequence of the "shrinking", the ditag may essentially comprises either or both the 5' end region (also indicated as 5' tag) and 3' end region (also indicated as 3' tag) of the original nucleic acid molecule. Hence, the portion of the original nucleic acid molecule that is originally between or inside the 5' tag and 3' tag is not included in the ditag. The ditag according to the invention retains the most informative features of the original nucleic acid molecule, namely the start and the end signatures of the nucleic acid.

[0086] The 5' tag and 3' tag forming the ditag may have the same or different size. Preferably, they have the same number of nucleotides. The ditag may be of any size, but needs to be meaningful and advantageous over the size of the parental sequence from which it is derived. The preferred size of a tag or ditag is determined by genome complexity. For a bacterial genome a tag from about 8 bp to about 16 bp may be sufficient whereas for a complex genome like the human genome, a 16-20 bp tag (or in other words, a 32-40bp ditag) may be considered. In general, the size of the ditag is from about 12-60 bp.

[0087] For the purpose of the present application, the terms 5'-terminus, 5'-end and 5'-tag are equivalent to each other and may be used interchangeably. In the same way, the terms 3'-terminus, 3'-end and 3'-tag are equivalent to each other and may be used interchangeably. In an original nucleic acid molecule or portion inside a nucleic acid molecule that one intends to reduce or represent, each of the

5'-end and 3'-end represents a region or portion closest to the extremity and most far from the middle region of the molecule.

[0088] According to one aspect of the present invention, the 5'-tag and 3'-tag comprised in the ditag are the regions of the molecule cleaved by a restriction enzyme closest to the 5'-end and 3'-end, respectively, of the nucleic acid molecule or portion thereof which is intended to be reduced or represented. Accordingly, the size of the ditag may be determined by the restriction enzyme or enzymes used.

[0089] Accordingly, the invention provides an isolated oligonucleotide comprising an isolated oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex. The oligonucleotide may further comprise at least one linker inserted between the tags. The oligonucleotide may also further comprise at least one linker inserted to flank a tag (that is, the at least one linker may be positioned upstream and/or downstream the at least one tag).

#### Linker

[0090] In particular, each linker may comprise at least: one first restriction site and at least a second adjacent restriction site. Therefore, the number of restriction sites present in each linker may be one or more, preferably two. The restriction site may be an asymmetric restriction site. Examples of asymmetric restriction sites are homing endonuclease asymmetric recognition sites, and some type II (or class II) recognition sites. Type II's restriction enzymes that cut to one side of their recognition sites are preferable.

[0091] However, any recognition site known in the art may be used. Restriction enzyme recognizing at least one recognition site within the nucleic acid molecule and which may be used will be evident to those skilled in the art (see for example, Current Protocols in Molecular Biology, Vol. 2, 1995, Ed. Ausubel, et al., Greene Publish. Assoc. & Wiley Interscience, Unit 3.1.15; New England Biolabs Catalog, 2005). A list of possible restriction sites and corresponding restriction enzymes recognizing is reported below.

[0092] As an example, a restriction enzyme recognizing a restriction site may be used for the purpose of the preparation of the ditag according to the invention. In particular a type IIs enzyme, for example Mmel. When Mmel is used, this enzyme recognizes a sequence inside each of the two adaptors that flank the nucleic acid molecule that one intends to reduce, but cuts inside the nucleic acid molecule forming a tag comprising 17-21 nucleotides (see Mmel cuts indicated in FIGS. 1 and 4). Two such tags may be additionally processed by blunting and ligation to form a ditag comprising 34-38 nucleotides. The ditag is hence obtained by splicing together or ligating the 5' terminus and the 3' terminus of the same nucleic acid molecule.

[0093] As an example, asymmetric sites may be introduced. Asymmetric site sequences useful for the purpose of the present invention are: i) two homing endonuclease asymmetric recognition site sequences or ii) restriction endonuclease asymmetric cleavage sites sequences recognizable by type II restriction enzymes.

[0094] Homing endonucleases are sold and described by New England Biolabs, Inc.; a description of the asymmetric site sequences is also available in the New England Biolabs Catalog. These homing endonuclease asymmetric recognition site sequences are from 18 to 39 bp. However, in the present invention the recognition site sequences are not limited to those sequences nor to these sizes. Preferably, the restriction homing endonucleases capable of cutting the asymmetric site sequences are selected from the group consisting of: I-CeuI, PI-SceI, PI-PspI and I-SceI. The list mentioned above however is not exhaustive. Other homing endonucleases known in the art and those which may be later discovered are included in the scope of the present invention.

[0095] Examples of type II restriction enzymes include: AarI, AclII, Alol, BaeI, Bbr7I, BbvI, BbvII, BclI, Bce83I, BceAI, Bcefl, Bcgl, BciVI, Bfil, BinI, BpII, BsaXI, BscAI, BseMII, BseRI, BsgI, BsmI, BsmAI, BsmFI, Bsp24I, BspCNI, BspMI, Bsrl, BsrDI, BstF5I, BtgZI, BtsI, CjeI, CjePI, EclI, Eco31I, Eco57I, Eco57MI, Esp3I, Fall, Faul, FokI, GsuI, HaeIV, HgaI, Hin4I, HphI, HpyAV, Ksp632I, MboII, MlyI, Mmel, MnlI, PleI, Ppil, PspI, RleAI, SapI, SfaNI, SspD5I, Sth132I, StsI, TaqII, TspDTI, TspGWI, TspRI and Tth111II (the list in the web site of Rebase Enzymes®: <http://rebase.neb.com/cgi-bin/outsidelist>; see also Szybalski, W., 1985, Gene,

40:169). The list mentioned above however is not exhaustive. Other type II enzymes known in the art and those which may be later discovered are included in the scope of the present invention.

[0096] Examples of recognition sites and cleavage sites of several class II restriction enzymes are (into parenthesis are the recognition site and the cleavage site): BbvI (GCAGC 8/12), HgaI (GACGC 5/10), BsmFI (GGGAC 10/14) SfaNI (GCATC 5/9), and Bsp I (ACCTGC 4/8).

[0097] Artificial restriction endonucleases may also be used. These endonucleases may be prepared by protein engineering. For example, the endonuclease FokI has been engineered by insertions so that it cleaves one nucleotide further away from its recognition site on both strands of the DNA substrates. See Li and Chandrasegaran, Proc. Nat. Acad. Sciences USA 90:2764-8, 1993. Such techniques may be applied to prepare restriction endonucleases with desirable recognition sequences and desirable distances from recognition site to cleavage site.

[0098] Under the present invention, the isolated oligonucleotide of the present invention may be joined or concatenated with other isolated oligonucleotide to form a concatemer of oligonucleotides. Any number of oligonucleotides may be joined together for the purposes of sequencing or for cloning into a suitable plasmid or vector.

[0099] Accordingly, in another aspect, the present invention is a concatemer of oligonucleotides comprising at least two isolated oligonucleotides, each isolated oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex.

[00100] The isolated oligonucleotides of the concatemer of oligonucleotides may further comprise at least one linker. The linker may be inserted between the tags. Alternatively, a linker may flank a tag. The linker may comprise at least one restriction enzyme recognition site; the at least one restriction enzyme recognition site may be for a type II restriction enzyme.

[00101] In one embodiment, the first tag of the concatemer of oligonucleotides may comprise a 3' terminus from the first polynucleotide and the second tag further

comprises a 5' terminus from the second polynucleotide. The isolated oligonucleotides of the concatenated polynucleotide may each further comprise at least one linker inserted between the tags. In another embodiment, the first tag of the concatemer of oligonucleotides may comprise a 5' terminus from the first polynucleotide and the second tag further comprises a 3' terminus from the second polynucleotide.

[00102] In another embodiment, the first tag of the concatemer of oligonucleotides may comprise a 5' terminus and a 3' terminus from the first polynucleotide and the second tag further comprises a 5' terminus and a 3' terminus from the second polynucleotide. The isolated oligonucleotides of the concatemer of oligonucleotides may each further comprise at least one linker inserted between the tags.

[00103] The linkers of either of these embodiments may comprise at least one restriction enzyme recognition site; the at least one restriction enzyme recognition site may be for a type IIs restriction enzyme.

[00104] The concatemer of oligonucleotides may be inserted into a vector or a cell; the cell may be a bacterial cell.

[00105] The concatemer of oligonucleotides may be from a chromatin structure. The polynucleotides are located on the same chromosome or the polynucleotides are located on different chromosomes.

[00106] While these are preferred concatemer(s), it will be apparent that the number of oligonucleotides of the present invention that may be concatenated depends on the length of the oligonucleotides and may be readily determined by those of skill in the art without undue experimentation. After formation of concatemers, multiple tags may be cloned into a vector for sequence analysis, or ditags or concatemers may be directly sequenced without cloning by methods known to those of skill in the art. Hence, the concatenation of ditags allows an efficient analysis of the nucleic acid molecules, like full-length cDNAs, in a serial manner by sequencing multiple ditags within a single vector or clone.

[00107] While term vector or recombinant vector it is intended, a plasmid, virus or other vehicle known in the art that has been manipulated by insertion or incorporation of the ditag genetic sequences may also be used. Such vectors

contain a promoter sequence that facilitates the efficient transcription. The vector typically contains an origin of replication, a promoter, as well as specific genes that allow phenotypic selection of the transformed cells. Vectors suitable for use in the present invention include for example, pBlueScript (Stratagene, La Jolla, CA); pBC, pZErO-1 (Invitrogen, Carlsbad, CA) and pGEM3z (Promega, Madison, WI) or modified vectors thereof as well as other similar vectors known to those of skill in the art. As a particular realization, the pGEM3z vector has been modified, and will be referred to as pGIS8 (FIGS. 7 and 11). pGEM vectors have also been disclosed in US 4,766,072, herein incorporated by reference.

[00108] For the production of the parental polynucleotide or nucleic acid molecule from which the tags or ditags were derived, as full-length libraries, suitable vectors may be used. Accordingly, suitable vectors, which are within the scope of the present invention, are those wherein the backbone of the vector does not comprise the same restriction site comprised in the adaptors flanking the polynucleotide or the ditag, after insertion of the polynucleotide. Preferably, the invention provides a vector wherein the vector backbone (other than within the stuffer region containing multiple cloning sites that is removed during insertion of the polynucleotide) does not comprise the first restriction site and the second or further restriction sites that are comprised into the adaptors. In particular, the vector does not comprise the at least II restriction site (for example type IIs restriction site) and the at least second or further restriction site comprised in the adaptors. More preferably, the vector backbone (other than within the stuffer region that is removed during insertion of the polynucleotide) does not comprise Mmel and BamHI.

[00109] An example of such a vector not comprising Mmel in any region outside of the stuffer is the vector pGIS8 shown in FIGS. 7 and 11. In pGIS8 the Mmel recognition sites were deleted by mutagenesis.

[00110] Accordingly, the present invention provides a library of oligonucleotides comprising at least one oligonucleotide, the oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex.

There is also provided a library of concatemer(s) of oligonucleotides according to the invention.

[00111] The at least one oligonucleotide of the library of oligonucleotides may comprise at least one linker is inserted between the tags. Alternatively, a linker may instead flank a tag. The first tag may comprise a 5' terminus and a 3' terminus from the first polynucleotide and the second tag further comprises a 5' terminus and a 3' terminus from the second polynucleotide.

[00112] The polynucleotide nucleic acid-protein complex may be part of a chromatin structure. The polynucleotides may be located on the same chromosome or the polynucleotides may be located on different chromosomes.

[00113] According to one aspect, the oligonucleotide is amplified. For example, by using PCR or any other known amplification methods. The PCR primers and the probes sequences may be prepared based on the information of the sequence of the ditags. Accordingly, suitable PCR primers corresponding to specific regions inside the vector are used. Such regions flank the oligonucleotide comprising the ditag and adaptors. PCR may be performed directly on the ligation (self-circularization) reaction to obtain short (for example 200bp) PCR products.

[00114] These PCR products that contain the required ditags may then be cut with an enzyme recognizing the at least second restriction site (inside the adaptors) to generate the required short cohesive ditags. As restriction enzyme recognizing the second or further restriction site, BamHI may for example be used, and cohesive ditags of 50bp are generated. The advantage of this amplification step is that of generating ditags circumvents the need to produce a ditag library amplification, which may be avoided by not transforming the self-circularized tagged plasmids. The amplified oligonucleotide may then subsequently be excised from the vector (in this example, by digestion with BamHI) and concatenated in long stretches of DNA or RNA for subsequent cloning and sequencing analysis (FIGS 1 and 2).

[00115] As a particular aspect, the invention discloses a cDNA library wherein the oligonucleotide(s) comprises at least one ditag, and wherein the ditag comprises about 34-38 nucleotides and is obtained by splicing nucleotides from the 5' terminus and nucleotides from the 3' terminus of a full-length cDNA or fragment thereof.

[00116] The ditag library according to the invention is representative of the library comprising the original nucleic acid molecules. For example, when the library comprising the nucleic acid molecules is a full-length polynucleotide library, the ditag library is representative of the full-length ditag library. Each ditag clone comprises sufficient information characterizing the specific full-length clone. More important, the ditag of the invention comprises the 5'-end and 3'-end of an original full-length polynucleotide derived from the nucleic acid-protein complex. Hence, the ditag is representative of the structure of the polynucleotide.

[00117] Accordingly, it is sufficient to sequence and analyze the ditag clones of the ditag library. In case a ditag of interest is found, the corresponding full-length polynucleotide may be selected and prepared from the full-length polynucleotide library, for example by PCR or directly from target RNA samples by reverse transcription-polymerase chain reaction (RT-PCR).

[00118] Sequencing the oligonucleotides of the invention may be performed by the "454" sequencing (pyrosequencing) method (Margulies et al, 2005), or concatemerized into longer DNA stretches for cloning to make CIA-PET library and followed by sequencing using the Sanger capillary method.

[00119] Accordingly, the present invention provides methods to prepare the oligonucleotide of the invention, to detecting and/or identifying at least two polynucleotides from a nucleic acid-protein complex, and to prepare vectors comprising the oligonucleotides and concatenated oligonucleotides of the present invention.

Thus, in another aspect, the present invention provides a method of preparing at least one isolated oligonucleotide comprising:

- (a) providing a nucleic acid-protein complex;
- (b) preparing an oligonucleotide comprising at least one first tag and at least one second tag, wherein the first and second tags are obtained from a nucleic acid-protein complex; and
- (c) isolating the oligonucleotide.

[00120] In particular, there is provided a method of preparing at least one isolated oligonucleotide comprising:

- (a) providing a nucleic acid-protein complex;

- (b) preparing an oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex; and
- (c) isolating the oligonucleotide.

[00121] In one embodiment, step (b) of this aspect of the invention comprises:

- (i) inserting at least one linker comprising at least one restriction enzyme recognition site, and
- (ii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing the at least one recognition site in the linker to form the oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide and the linker between the tags.

[00122] In another embodiment, step (b) of this aspect of the invention comprises:

- (i) inserting at least one linker comprising at least one restriction enzyme recognition site between the first polynucleotide and the second polynucleotide of the complex;
- (ii) adding at least one restriction enzyme recognition site each to the 5' terminus of the first polynucleotide and the 3' terminus of the second polynucleotide;
- (iii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing at least one recognition site to obtain cleaved fragments; and
- (iv) ligating cleaved fragments to form the oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide, the tags comprising 5' and 3' termini of each polynucleotide, and the linker inserted between the tags and the at least one restriction enzyme recognition site in step (ii) is part of a vector.

[00123] The nucleic acid-protein complex may be obtained by chromatin immunoprecipitation.

[00124] In another aspect, the present invention provides a method of detecting and/or identifying at least two polynucleotides from a nucleic acid-protein complex, the method comprising:

- (a) providing a nucleic acid-protein complex;

- (b) preparing an oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex;
- (c) sequencing the oligonucleotide; and
- (d) mapping the at least two polynucleotides based on the nucleotide sequences of the first and second tags, thereby detecting and/or identifying the at least two polynucleotides.

[00125] In one embodiment, step (b) of this aspect of the invention comprises:

- (i) inserting at least one linker comprising at least one restriction recognition site, and
- (ii) cleaving the first and second polynucleotides with a least one restriction enzyme recognizing the at least one recognition site in the linker to form the oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide and the linker between the tags.

[00126] In another embodiment, step (b) of this aspect of the invention comprises:

- (i) inserting at least one linker comprising at least one restriction enzyme recognition site between the first polynucleotide and the second polynucleotide of the complex;
- (ii) adding at least one restriction enzyme recognition site each to the 5' terminus of the first polynucleotide and the 3' terminus of the second polynucleotide;
- (iii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing at least one recognition site to obtain cleaved fragments; and
- (iv) ligating cleaved fragments to form the oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide, the tags comprising 5' and 3' termini of each polynucleotide, and the linker inserted between the tags; the at least one restriction enzyme recognition site in step (ii) may be part of an adaptor or a vector.

[00127] The nucleic acid-protein complex under the present invention may be obtained by chromatin immunoprecipitation. The nucleic acid-protein complex may be obtained by incorporating a photoactivatable moiety into the DNA and/or the

protein of interest and isolation of DNA/protein complex by antibody-mediated precipitation or by affinity-mediated technique. Examples of such affinity-based techniques include streptavidin/ biotin, glutathione-S-transferase/glutathione matrix, and maltose-binding protein/amylose matrix interactions.

[00128] The oligonucleotide obtained in step (b) may be amplified before being sequenced in step (c); the amplification may be by polymerase chain reaction. The amplified oligonucleotides may be subjected to at least one purification step after amplification but before being sequenced in step (c). The at least one purification step may be gel electrophoresis.

[00129] The oligonucleotide according to the invention may be concatenated with at least one other oligonucleotide obtained by steps (a) to (b) before being sequenced to make a concatemer of oligonucleotides.

[00130] The sequencing may be by the Sanger method or by multiplex sequencing. The multiplex sequencing may be pyrosequencing. Any suitable method of sequencing such as those described by Bonetta (2006) may be employed. The nucleic acid fragment to which a protein of interest binds may be any nucleic acid fragment comprising a region to which a protein of interest binds, for example, a histone binding site. The polynucleotides may be DNA or RNA.

[00131] The polynucleotides may be located on the same chromosome or the polynucleotides may be located on different chromosomes.

[00132] The detecting and/or identifying may be for transfection or translocation of the polynucleotides.

[00133] The oligonucleotide may be transfected into a cell. The transfection may be by electroporation. The cell may be a bacterial cell.

[00134] The embodiments of this aspect of the invention are described in more detail below as the CIA-PET and CIA-diPET techniques below.

[00135] In another aspect, the present invention is a vector comprising the oligonucleotide, the concatemer of oligonucleotides or the library of oligonucleotides of the present invention.

#### CIA-PET

[00136] In the CIA-PET method (FIGS. 1 and 4), DNA fragments tethered by protein in DPD complexes will be joined by linker sequences through ligation. The

linker sequence contains two Mmel sites. After reversal of the cross-linking, the ligated DNA will be digested by Mmel to release the paired end ditag (CIA-PET). Each of the CIA-PETs contains a linker with two flanked tags (about 20bp each) (Step 2 labeled in FIG. 1). The two tags contained in a CIA-PET thus represent two distal genetic regions that are far away from each other in linear genome sequence, but interacting together and mediated by a specific protein such as a histone protein.

[00137] After the CIA-PETs are gel-purified, sequence specific adaptors are added to each side of the CIA-PETs, which are then amplified by PCR (Step 3 in FIG. 1). The amplified CIA-PETs may be directly sequenced by a multiplex sequencing method such as the "454" pyrosequencing method or any other suitable sequencing methods, or concatemerized into longer DNA stretches for cloning to make CIA-PET library and followed by sequencing using the Sanger capillary method. The CIA-PET sequences will be mapped to the reference genome sequences. True chromatin interaction sites may be identified based on frequent occurrence of CIA-PET clusters in specific loci, which will be distinguished from background noise randomly scattered as singletons (FIG. 3).

#### CIA-diPET

[00138] In the CIA-diPET method (FIGS. 2 and 6), paired end ditag sequences representing each of the two related DNA fragments are extracted and sequenced compared to only one tag per sequence. As such, the resulting CIA-diPET will contain longer tag sequences than the tags of the CIA-PET method. It will be appreciated by a person skilled in the art that the CIA detection method of the present invention is thus more specific in mapping chromatin interactions than other methods of the prior art. The DNA fragments in DPD complexes are joined by a linker sequence that contains a Mmel site and a Gsu1 site. After reverse cross-linking, the ligated DNA will be randomly broken by sonication.

[00139] The DNA will then be size fractionated and cloned into a pGIS8 vector (Step 2 of FIG. 2) that contains an immediate Mmel site and a Gsu site at its cloning site (FIG. 7). After transformation and propagation in bacteria cells, the library clones are subjected by sequential manipulations of digestion, self-ligation, and transformation to create the single diPET library. The plasmid DNA of the

single diPET library will be digested by BamH1 to release the diPET structure (Step 4 of FIG. 2), which may be directly sequenced using the MS-PET sequencing method, or further concatenated for cloning and sequencing. The CIA-diPET sequences will be mapped to the reference genome sequences. True chromatin interaction sites may be identified based on frequent occurrence of CIA-diPET clusters in specific loci, which will be distinguished from background noise randomly scattered as singletons (FIG. 5).

[00140] Having now generally described the invention, the same will be more readily understood through reference to the following examples that are provided by way of illustration and are not intended to be limiting of the present invention.

[00141] For quality control and to ensure correct insertion of linkers or adaptors, gel electrophoresis is performed to remove unwanted fragments (FIGS. 8 to 10) during the process.

### Examples

#### Methodology of CIA

[00142] The Nanog transcription factor in mouse ES cell (E14) and ChIP-PET data for Nanog and other key ES cell transcription factors such as Oct4 and Sox2 were used as the biological system to practice the present invention. The ChIP-PET data provides a linear map of where these transcription factors bind to, and will be used as reference to validate the chromatin interaction data. The interactions between chromatins were detected by two methods, CIA-PET and CIA-diPET.

[00143] The sequences used in the examples are as below, and in the sequence listing. Where two strands are shown, the top strand is the sense strand and the bottom strand is the antisense strand. All sequences are shown in the 5' to 3' direction. Where a nucleotide is denoted by N or n (or X or a numeral in the Figures), it means that any nucleotide (A, C, G or T) may be represented in that position.

#### M&M linker

5' GTCGGAGGCCAAGGCGGCCGTACGTCCAACT 3' (31 nt) (SEQ ID NO:1)  
 5' GTTGGACGTACGGCCGCCTTGGCCTCCGAC 3' (31 nt) (SEQ ID NO:2)  
 5' ends are phosphorylated

M&G linker

5' GTCGGAGGCCAAGGCGGCCGTACGCTGGAGT 3' (31 nt) (SEQ ID NO:3)  
5' CTCCAGCGTACGGCCGCCTTGGCCTCCGACT 3' (31 nt) (SEQ ID NO:4)  
5' ends are phosphorylated

P1 concatenating adapter (PMR 011)

5' GGATCCCTTAATCGCCTTGCAGCACATC 3' (28 nt) (SEQ ID NO:5)  
5' GATGTGCTGCAAGGCGATTAAGGGATCCNN 3' (30 nt) (SEQ ID NO:6)  
5' end of top adapter is phosphorylated  
5' end of bottom adapter is not phosphorylated

P2 concatenating adapter (PMR 012)

5' GGATCCCCTGTGTGAAATTGTTATCCGCT 3' (29 nt) (SEQ ID NO:7)  
5' AGCGGATAACAATTTACACAGGGGATCCNN 3' (31 nt) (SEQ ID NO:8)  
5' end of top adapter is phosphorylated  
5' end of bottom adapter is not phosphorylated

D1 diPETing adapter (PMR 011)

5' GGATCCCTTAATCGCCTTGCAGCACATC 3' (28 nt) (SEQ ID NO:9)  
5' GATGTGCTGCAAGGCGATTAAGGGATCCNN 3' (30 nt) (SEQ ID NO:10)  
5' end of top adapter is phosphorylated  
5' end of bottom adapter is not phosphorylated

D2 diPETing adapter (PMR 012)

5' GGATCCAATGCTCCTCCCTGTGTGAAATTGTTATCCGCT 3' (39 nt) (SEQ ID NO:11)  
5' AGCGGATAACAATTTACACAGGGAGGAGCATTGGATCCNN 3' (41 nt) (SEQ ID NO:12)  
5' end of top adapter is phosphorylated  
5' end of bottom adapter is not phosphorylated

PMR011 primer

5' GATGTGCTGCAAGGCGATTAAG 3' (22 nt) (SEQ ID NO:13)  
5' ends are phosphorylated

PMR012 primer

5' AGCGGATAACAATTTACACAGG 3' (23 nt) (SEQ ID NO:14)  
5' ends are phosphorylated

RecA selection oligo

5' AGTCGGAGGCCAAGGCGGCCGTACGCTGGAGT 3' (SEQ ID NO:15)  
(Biotinylated)

pGIS8 vector

This vector was derived from a pGEM vector (Promega) and contains the multiple unique cloning sites for the following restriction enzymes in this order:

BamH1 → Mme1 ; Cleavable region ; Gsu1 → BamH1 → BseR1

The following oligonucleotides are used:

5' AATTGGATCCGACTCGAGGATGAATTCTCCAGGATCCCTCCTC 3' (43 nt)  
(SEQ ID NO:16)

5' TCGAGAGGAGGGATCCTGGAGAATTCATCCTCGAGTCGGATCC 3' (43 nt)  
(SEQ ID NO:17)

5' ends are phosphorylated.

The rest of the vector do not contain any BseR1, BamH1, Mme1 or Gsu1 sites.

The pGIS8 plasmid is one example of a vector that may be used (FIG. 7). The sequence of the sense strand is given in SEQ ID NO:18 and the sequence of the antisense strand is SEQ ID NO:19. A multiple cloning site containing restriction enzyme recognition sites (represented by SEQ ID NOS: 16 and 17) is inserted into the vector. The sequence listing for pGIS8 showing the various restriction enzyme recognition sites are shown in FIG. 11. Any other vector fulfilling the requirements may be prepared and used by any skilled person in the art.

M&M adapter PET

5'NNNNNNNNNNNNNNNNNNNAGTCGGAGGCCAAGGCGGCCGTACGTCCAACNTN  
NNNNNNNNNNNNNNNNNNNN 3' (69 nt) (SEQ ID NO:20)

5'NNNNNNNNNNNNNNNNNNNNNAGTTGGACGTACGGCCGCCTTGGCCTCCGACT  
NNNNNNNNNNNNNNNNNNNNNNN 3' (69 nt) (SEQ ID NO:21)

M&M adapter PET with adapter sequences

5'GATGTGCTGCAAGGCGATTAAGGGATCCNNNNNNNNNNNNNNNNNNNNNNNNNNAGT  
CGGAGGCCAAGGCGGCCGTACGTCCAACNNNNNNNNNNNNNNNNNNNNNNNNNGGA  
TCCCCTGTGTGAAATTGTTATCCGCT 3'(128 nt) (SEQ ID NO:22)

5' AGCGGATAAC AATTTACAC AGGGGATCCNNNNNNNNNNNNNNNNNNNNNNNA  
GTTGGACGTACGGCCGCCTTGGCCTCCGACTNNNNNNNNNN NNNNNNNNNNN  
GGATCCCTTA ATCGCCTTGCAGCACATC 3' (128 nt) (SEQ ID NO:23)

M&M final PET (after cleavage of adapter sequences)

5'GATCCNNNNNNNNNNNNNNNNNNNNNNNNNAGTCGGAGGCCAAGGCGGCCGTACG  
TCCAACNNNNNNNNNNNNNNNNNNNNNNNNNG (77 nt) (SEQ ID NO:24)

5' GATCCNNNNNNNNNNNNNNNNNNNNNNNNNAGTTGGACGTACGGCCGCCTTGG  
CCTCCGACTNNNNNNNNNNNNNNNNNNNNNNNG 3' (77 nt) (SEQ ID NO:25)

M&G adapter DNA sequences after sonication

5'NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGTCGGAGGCCAAGGCGGCCGTACGC  
TGGAGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 3' (SEQ ID NO:26)

5' NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
CCTTGGCCTCCGACTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 3' (SEQ ID NO:27)  
(variable nt)

M&G adapter DNA sequences after cloning into pGIS8 plasmids

(note: rest of the plasmid is not shown)

5'GGATCCGACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNNAGTCGGAGGCCAAGGCGGCCGTACGCTGGAGTNNNNNNNNNN  
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCTCCAGGATCC 3' (140  
nt) (SEQ ID NO:28)

5' GGATCCTGGAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNNACTCCAGCGTACGGCCGCCTTGGCCTCCGACTNNNN  
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGTCG  
GATCC 3' (140 nt) (SEQ ID NO:29)

M PET intermediate

5' GGATCCGACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGT  
CGGAGGCCAAGGCGGCCGTACGCTGGAGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCTCCAGGATCC 3' (132 nt) (SEQ ID NO:30)

5' GGATCCTGGAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACTCCAGCGTACGGCCGCCTTGGCCTCCGACTNNNN  
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGTCGATCC 3' (132 nt) (SEQ ID NO:31)

M and G diPET

5' ATCCGACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGT  
CGGAGGCCAAGGCGGCCGTACGCTGGAGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNCTCCAGGATCC 3' (116 nt) (SEQ ID NO:32)

5' GGATCCTGGAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
CGGCCGCCTTGGCCTCCGACTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NN NNNNNNNNNNGTCGATC 3' (116 nt) (SEQ ID NO:33)





Oligonucleotide of FIG. 9 with partial M&G linker inserted after restriction enzyme digest (rest of plasmid not given)



(c) Cell lysis and chromatin preparation:

- c1. Cells are lysed in lysis buffer (50 mM HEPES, 1 mM EDTA, 0.15 M NaCl, 1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate, all from Ambion).
- c2. The chromatin is solubilized by sonication using a Branson 450 ultrasonic cell disruptor (20% duty power output, 30 second, 5 to 8 times;).
- c3. The chromatin is diluted 10 times to lower the SDS to 0.1%.
- c4. The extract is then cleared by centrifugation at 14,000 rpm for 10 minutes at 4°C.
- c5. This extract is stored at -80°C until use.

(d) Immuno-precipitation

- d1. Two microgrammes of monoclonal antibody (F7, Santa Cruz) are bound to protein G sepharose (Pharmacia).
- d2. The antibody coated beads are incubated with the chromatin extract at 4°C for 16 hours.
- d3. The beads are then washed with the following buffers (reagents from Sigma Chemical Company):
  - Wash buffer 1 (50 mM HEPES, 1 mM EDTA, 0.15 M NaCl, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate) 2 times
  - Wash buffer 2 (50 mM HEPES, 1 mM EDTA, 0.5 M NaCl, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate) 1 time
  - Wash buffer 3 (20 mM Tris.HCl pH 8.0, 1 mM EDTA, 0.25 M LiCl, 0.5% NP40, 0.5% sodium deoxycholate) 1 time
  - Wash buffer 4 (20 mM Tris.HCl pH 8.0, 1 mM EDTA) 1 time
- d4. The protein-DNA complexes are then eluted from the beads with elution buffer (50 mM Tris.HCl pH 8.0, 1 mM EDTA, 1 % SDS) for 20 min at 65°C.
- d5. The eluant is then dialyzed in PBS (Ambion) to remove SDS for 3 hours at 4°C.

[00146] (2) Prepare CIA-PET

(a) End repair

Perform end-repair with the Epicentre End-It kit and chemicals:

Chromatin (up to 5 ug)

2.5 ul 10x

|                       |          |
|-----------------------|----------|
| End Repair Buffer     | 5 ul     |
| 2.5 mM dNTP Mix       | 5 ul     |
| 10 mM ATP             | 5 ul     |
| End-Repair Enzyme Mix | 1 ul     |
| Nuclease free water   | to 50 ul |

Vortex briefly to mix and then incubate at room temperature for 45 minutes, stop reaction by heating at 70°C for 10 minutes. Adjust the concentration to 65ng/ul by adding 27ul of nuclease free water.

(b) A-tailing

|                           |          |
|---------------------------|----------|
| DNA                       | 20 ul    |
| 10 mM dATP (Roche)        | 0.5 ul   |
| 10x ExTaq buffer (Takara) | 2.5 ul   |
| ExTaq polymerase (Takara) | 0.5 ul   |
| Nuclease free water       | to 25 ul |

Incubate in a PCR machine at 72°C for 30 min, then 4°C. It is best to take out the tubes and perform the ligation immediate once incubation at 72°C ends.

(c) Ligate DNA with M&M linker (SEQ ID NOS: 1 and 2)

|                                        |        |
|----------------------------------------|--------|
| Chromatin DNA (200ug)                  | 3.1 ul |
| M&M adapter – T tailed (38ng)          | 3.1 ul |
| 5X ligase buffer with PEG (Invitrogen) | 6 ul   |
| T4 ligase (5U/ul) (Invitrogen)         | 1 ul   |
| Nuclease free water                    | 12.3ul |

Ligate overnight at 16°C to obtain oligonucleotides (SEQ ID NOS: 20 and 21) inserted with M&M linker.

(d) Reverse cross-links with Proteinase K

Divide DNA sample into 20 ul aliquots, and reverse cross-links by overnight incubation at 65°Celsius in the presence of 15 ul of 20 mg/ml proteinase K (Ambion). The next day, add 1 ul of 10 mg/ml RNase A (Qiagen) to degrade RNA is degraded 45 min at 37°Celsius, followed by phenol extraction and ethanol

precipitation of DNA. Resuspend in 20 ul of Elution Buffer and store at -20°C. The concentration obtained is usually 500 ng/ul.

Quantitate DNA and perform quality control by running 0.5 ul of the DNA obtained on a 1% gel with Takara Wide-Range Ladders and Invitrogen Low-Mass ladders and 0.5 ul of material from (a). The material from (d) should be less smeared on the gel and should not show any bright band at around the 50-100 bp mark.

(e) Mme1 cut

|                                      |        |
|--------------------------------------|--------|
| 10 ug DNA                            | 20 ul  |
| 10x NEBuffer 4 (New England Biolabs) | 20 ul  |
| 10x SAM (New England Biolabs)        | 20 ul  |
| Mme1 (2U/ul) (New England Biolabs)   | 20 ul  |
| Nuclease-free water                  | 120 ul |

Split into two tubes and incubate overnight at 37°C. SAM should be prepared fresh.

Phenol chloroform and ethanol precipitation with glycoblu.

Quantitate DNA and perform quality control by running 3 ul on a 2% gel or PAGE gel together with Takara Wide-Range Ladders and Invitrogen Low-Mass ladders and with 0.5 ul of material from step (d).

(f) Gel purification

Load DNA obtained onto a 2% agarose gel in a Scie-Plas medium-sized unit (60 ul per well) together with appropriate ladders, eg 20 ul of Takara Wide Range ladder. Run at 80V, approximately 1.5 hours, and visualize at 365 nm UV. Oligonucleotides with incorrectly inserted linkers will be removed by the gel electrophoresis (for examples of incorrectly inserted linkers, see FIGS. 8 to 10).

Excise ditag band and electroelute using disposable Fermentas ElutaTubes according to manufacturer's recommendations. Electroelution is performed for 1-1.5 hours at 90 V and the harvested ditags are ethanol precipitated thus:

For every 200 ul eluate, add

|                                |        |
|--------------------------------|--------|
| 3 M NaOAc pH 5.2 (Amresco)     | 20 ul  |
| 1 M MgCl <sub>2</sub> (Ambion) | 4.5 ul |
| Glycoblu (Ambion)              | 2 ul   |

100% ethanol

800 ul

Resuspend precipitated ditags in 12 ul Elution Buffer (Qiagen) and run 2-5 ul on a 4-20% PAGE minigel together with Low Mass Ladders (Invitrogen) for purity check and visual quantification. Correctly obtained ditags have the sequences given in SEQ ID NOS: 22 and 23).

### [00147] (3) Amplify PET

#### (a) Adapter ligation (Concatenating adaptors SEQ ID NOS: 5 to 8)

|                                     |        |
|-------------------------------------|--------|
| DNA (100 ng)                        | 6 ul   |
| Adapter (10 ug)                     | 6 ul   |
| 10x ligase buffer (with spermidine) | 1.5 ul |
| T4 DNA ligase (5U/ul, Invitrogen)   | 1 ul   |
| Nuclease Free Water                 | 0.5 ul |

The total volume is 15 ul. Incubate at 16°C for 16 hours.

10x ligation buffer with Spermidine is made up of:

- 60mM Tris-HCl pH7.5 (Ambion)
- 60mM MgCl<sub>2</sub> (Ambion)
- 50mM NaCl (Ambion)
- 1mg/ml BSA (New England Biolabs)
- 70mM Beta-mercaptoethanol (Sigma)
- 1mM ATP (Invitrogen)
- 20mM DTT (Invitrogen)
- 10mM spermidine (Sigma)

#### (b) PCR amplification

Amplify with primers PMRs 11 and 12 (SEQ ID NOS: 13 and 14 respectively) and the Hotstartaq kit from Qiagen:

|                                       |               |
|---------------------------------------|---------------|
| DNA                                   | 1 ul          |
| 10x PCR buffer (Qiagen)               | 10 ul         |
| dNTP mix (10 mM of each) (Invitrogen) | 2 ul          |
| PMR11                                 | 1 ul (0.2 uM) |

|                                    |               |
|------------------------------------|---------------|
| PMR12                              | 1 ul (0.2 uM) |
| HotStarTaq DNA Polymerase (Qiagen) | 0.5 ul        |
| Nuclease Free Water                | to 100 ul     |

Mix well. Incubate in PCR machine:

1. 15 minutes, 95°C
2. 0.5 min 94°C
3. 0.5 min 55°C
4. 1 min 72°C
5. Repeat from step 2 25 times.
6. 10 min 72°C.

Purify using PCR purification kit (Qiagen).

**[00148] (4) Sequence the CIA-PETs**

The CIA-PETs may be directly sequenced according to the protocols for the 454 multiplex sequencing machine (454 life sciences). The technique is taught in Margulies et al (2005) and US Application No. 20030068629. These references are hereby incorporated in their entirety by reference.

**[00149] (5) Map the CIA-PETs**

Mapping may be performed using the Compressed Suffix Array. Multiple joins across two different DNA fragments ( $n > 3$ ) shall be taken to represent real distal control regions (FIG. 3).

**Example 3 - Method of CIA-diPET**

**[00150]** The CIA-diPET method (FIGS. 2 and 6) comprises the following sections: (1) Generate ChIP DNA-protein-DNA complex, (2) Prepare CIA library (3) Prepare CIA-PET library (4) Sequence and (5) Map the CIA-PET sequences (FIG. 5).

**[00151] (1) Generate ChIP DNA-protein-DNA complex (as above)**

(a) Mouse embryonic stem (ES) cells are cultured under feeder-free conditions in the presence of leukemia inhibitory factor (Chemicon).

(b) About 1 to 2 x 10<sup>8</sup> cells are collected and crosslinked with formaldehyde (final concentration of 1%; Sigma) for 10 minutes at room temperature.

(c) Cell lysis and chromatin preparation:

- c1. Cells are lysed in lysis buffer (50 mM HEPES, 1 mM EDTA, 0.15 M NaCl, 1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate, all from Ambion).
- c2. The chromatin is solubilized by sonication using a Branson 450 ultrasonic cell disruptor (20% duty power output, 30 second, 5 to 8 times;).
- c3. The chromatin is diluted 10 times to lower the SDS to 0.1%.
- c4. The extract is then cleared by centrifugation at 14,000 rpm for 10 minutes at 4°C.
- c5. This extract is stored at -80°C until use.

(d) Immuno-precipitation

- d1. Two micrograms of monoclonal antibody (F7, Santa Cruz) are bound to protein G sepharose (Pharmacia).
- d2. The antibody coated beads are incubated with the chromatin extract at 4°C for 16 hours.
- d3. The beads are then washed with the following buffers (reagents from Sigma Chemical Company):

Wash buffer 1 (50 mM HEPES, 1 mM EDTA, 0.15 M NaCl, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate) 2 times

Wash buffer 2 (50 mM HEPES, 1 mM EDTA, 0.5 M NaCl, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate) 1 time

Wash buffer 3 (20 mM Tris.HCl pH 8.0, 1 mM EDTA, 0.25 M LiCl, 0.5% NP40, 0.5% sodium deoxycholate) 1 time

Wash buffer 4 (20 mM Tris.HCl pH 8.0, 1 mM EDTA) 1 time

d4. The protein-DNA complexes are then eluted from the beads with elution buffer (50 mM Tris.HCl pH 8.0, 1 mM EDTA, 1 % SDS) for 20 min at 65°C.

d5. The eluant is then dialyzed in PBS (Ambion) to remove SDS for 3 hours at 4°C.

[00152] (2) Prepare CIA-diPET library

[00153](a) End repair (as for preparation of CIA-PET), perform end-repair with the Epicentre End-It kit:

|                                   |         |
|-----------------------------------|---------|
| Chromatin (up to 5 ug)            | 2.5 ul  |
| 10x End Repair Buffer (Epicentre) | 5 ul    |
| 2.5 mM dNTP Mix (Epicentre)       | 5 ul    |
| 10 mM ATP (Epicentre)             | 5 ul    |
| End-Repair Enzyme Mix (Epicentre) | 1 ul    |
| Nuclease Free Water               | 31.5 ul |

Briefly vortex to mix, then incubate at room temperature for 45 minutes, stop by heating at 70°C for 10 minutes. Adjust the concentration to 65 ng/ul by adding 27 ul of Nuclease Free Water.

(b) A-tailing, as above for CIA-PET

|                           |        |
|---------------------------|--------|
| DNA                       | 20 ul  |
| 10 mM dATP (Roche)        | 0.5 ul |
| 10x ExTaq buffer (Takara) | 2.5 ul |
| ExTaq polymerase (Takara) | 0.5 ul |
| Nuclease Free Water       | 1.5 ul |

Incubate in a PCR machine at 72°C for 30 minutes, then 4°C forever. It is best to take out the tubes and perform the ligation immediately once the incubation at 72°C ends.

(c) Ligate DNA with M&G linker (SEQ ID NOS: 3 and 4)

|                                        |         |
|----------------------------------------|---------|
| Chromatin DNA (200 ng)                 | 3.1 ul  |
| M&G adapter- T tailed (38 ng)          | 7.6 ul  |
| 5x ligase buffer with PEG (Invitrogen) | 6 ul    |
| T4 ligase (5U/ul) (Invitrogen)         | 1 ul    |
| Final volume                           | 12.3 ul |

Ligate overnight at 16°C to obtain oligonucleotides (SEQ ID NOS: 26 and 27) inserted with M&G linker.

(d) Reverse cross-links with proteinase K, as above for CIA-PET.

Divide DNA sample into 20  $\mu$ l aliquots, and reverse cross-links by overnight incubation at 65°C in the presence of 15  $\mu$ l of 20 mg/ml proteinase K (Ambion). The next day, add 1  $\mu$ l of 10 mg/ml RNase A (Qiagen) to degrade RNA. RNA is degraded 45 min at 37°C, followed by phenol extraction and ethanol precipitation of DNA. Resuspend in 20  $\mu$ l of Elution Buffer and store at -20°C. The concentration is usually 500 ng/ $\mu$ l.

Quantitate and perform quality control check by running 0.5  $\mu$ l on a 1% gel together with Takara Wide-Range Ladders and Invitrogen Low-Mass ladders and 0.5  $\mu$ l of material from (a). The material from (d) should be less smeared out on the gel as well as not show any bright band at around the 50-100 bp mark.

(e) Digest with Nla III

Store Nla III at -80°C (half-life is 6 months at -80°C). Place on ice just before use.

|                                      |              |
|--------------------------------------|--------------|
| DNA (approximately 1 $\mu$ g)        | 2 $\mu$ l    |
| 10x NEBuffer 4 (New England Biolabs) | 5 $\mu$ l    |
| NlaIII (New England Biolabs)         | 1 $\mu$ l    |
| 100x BSA (New England Biolabs)       | 0.5 $\mu$ l  |
| Nuclease-Free Water                  | 41.5 $\mu$ l |

Prepare five tubes of the above reaction. Incubate for 1 hour at 37°C. Phenol chloroform ethanol precipitate with glycoblue, resuspend in 10  $\mu$ l of Elution Buffer (Qiagen).

(f) Polish ends with the Epicentre End-It kit:

|                                   |            |
|-----------------------------------|------------|
| DNA (up to 5 $\mu$ g)             | 34 $\mu$ l |
| 10x End Repair Buffer (Epicentre) | 5 $\mu$ l  |
| 2.5 mM dNTP Mix (Epicentre)       | 5 $\mu$ l  |
| 10 mM ATP (Epicentre)             | 5 $\mu$ l  |
| End-Repair Enzyme Mix (Epicentre) | 1 $\mu$ l  |

Incubate at room temperature for 45 minutes; stop by heating at 70°C for 10 minutes.

(i) Clone into a pGIS8 vector (SEQ ID NOS: 18 and 19; FIGS. 7 and 11) with Mme1 and Gsu1 flanking sites

Set up on ice, using 1.7 ml microfuge tubes (use multiple tubes):

|                                        |      |
|----------------------------------------|------|
| 40 ng/ul precut pGIS                   | 1 ul |
| DNA                                    | 6 ul |
| 5x ligase buffer with PEG (Invitrogen) | 2 ul |
| T4 DNA ligase (5U/ul) (Invitrogen)     | 1 ul |

Incubate overnight (12-16 hours) at 16°C to obtain oligonucleotides represented by SEQ ID NOS: 28 and 29).

Also set up a vector self-ligation control.

Transform 1µl of the ligation reaction per 50µl of electrocompetent TOP10 cells (Invitrogen) by electroporation. Recover each aliquot in 1ml LB media at 37°C for 1hr, then plate out in a series of several dilutions (on LB agar + ampicillin) for quality control and titering.

Then scale up this process by plating the remaining culture on large agar plates (Q-trays), and performing maxipreps using the Qiagen HiSpeed Plasmid Maxi kit.

#### [000153] (3) CIA-diPET library

##### (a) Mme1 cut

|                                      |        |
|--------------------------------------|--------|
| 10 ug DNA                            | 100 ul |
| 10x NEBuffer 4 (New England Biolabs) | 20 ul  |
| 10x SAM (New England Biolabs)        | 20 ul  |
| Mme1 (2U/ul) (New England Biolabs)   | 12 ul  |
| Nuclease-free water                  | 48 ul  |

Incubate overnight at 37°C to obtain oligonucleotides represented by SEQ ID NOS: 30 and 31. SAM should be prepared fresh.

Phenol chloroform and ethanol precipitation with glycoblue, resuspend in 12 ul of Elution Buffer.

Quantitate DNA and perform quality control check by running 1 ul on a 2% gel or PAGE gel together with Takara Wide-Range Ladders and Invitrogen Low-Mass ladders and with 1 ul of material from step (e).

## (b) Circularization

Set up a 96 well plate with each well containing the following (MJ Research):

100 ng DNA 50 ul

Ligation Solution 1 (Takara Ligation Kit ver 2) 50 ul

Seal plate tightly to prevent evaporation.

Incubate at 16°C, overnight.

Perform PCR purification (Qiagen) with three columns, resuspending in 40 ul of Elution Buffer each, giving rise to a total of approximately 120 ul.

## (c) Gsu1 cut

Set up 9 tubes of:

Circularized DNA 12 ul

10x buffer TANGO (Fermentas) 8.6 ul

10x SAM (New England Biolabs) 8.6 ul

GsuI (5U/ul) (Fermentas) 1 ul

Nuclease Free Water 55.8 ul

Digest at 30°C for at least 2 h, but do not cut overnight. Oligonucleotides represented by SEQ ID NOS: 32 and 33 are obtained,

## (d) Circularization

Set up 9 tubes containing the following:

Approx 100 ng DNA 50 ul

Ligation Solution 1 (Takara Ligation Kit ver 2) 50 ul

Incubate at 16°C, overnight.

Phenol chloroform ethanol precipitate, resuspend in 12 ul of Elution Buffer.

## (e) Amplification with Rolling Circle Amplification (Templiphi kit, Amersham)

Thaw solutions on ice. Prepare 3-4 tubes of the following:

2.5 ng of DNA 1 ul

Templiphi kit denature buffer (Amersham) 10 ul

Heat at 95°C for 3 minutes, then briefly cool on ice.

To the reaction buffer, add:

Templiphi kit premix (Amersham) 10 ul

Mix well by tapping or gentle vortexing.

Incubate at 30°C, no shaking, 16 – 18 hours.

Examine the material, for example, by micropipetting 1 ul. It should be viscous. Quantitate double-stranded DNA by picogreen fluorimetry (Quant-iT DNA assay kit, Molecular Probes)

(f) BamHI cut

|                                                      |       |
|------------------------------------------------------|-------|
| 100 ug DNA                                           | 1 ul  |
| 10x unique BamHI buffer (New England Biolabs)        | 10 ul |
| 100x Bovine Serum Albumin (New England Biolabs)      | 1 ul  |
| BamHI (20 U/ul; 2-fold excess) (New England Biolabs) | 10 ul |
| Nuclease Free Water                                  | 78 ul |

Prepare more tubes as needed to digest the DNA; incubate at 37°C overnight to obtain oligonucleotides represented by SEQ ID NOS: 34 and 35.

(g) Gel purification

Load onto a 2% agarose gel in a Scie-Plas medium-sized unit (60 ul per well) together with appropriate ladders, eg 20 ul of Takara Wide Range ladder. Run at 80V, approximately 1.5 hours, and visualize at 365 nm UV. Oligonucleotides with incorrectly inserted linkers will be removed by the gel electrophoresis (for examples of incorrectly inserted linkers, see FIGS. 8 to 10).

Excise ditag band and electroelute using disposable Fermentas ElutaTubes according to manufacturer's recommendations. Electroelution is performed for 1-1.5 hours at 90 V and the harvested ditags are ethanol precipitated thus:

For every 200 ul eluate, add

|                       |        |
|-----------------------|--------|
| 3 M NaOAc pH 5.2      | 20 ul  |
| 1 M MgCl <sub>2</sub> | 4.5 ul |
| Glycoblu              | 2 ul   |
| 100% ethanol          | 800 ul |

Resuspend precipitated diPETs in 12 ul Elution Buffer and run 2-5 ul on a 4-20% PAGE minigel together with Low Mass Ladders (Invitrogen) for purity check and visual quantification.

## [00154] (4) Sequence the CIA-diPET

## (a) Concatenation of gel-purified BamHI-cohesive diPETs:

|                                     |      |
|-------------------------------------|------|
| CIA-diPETs 200-1000ng               | 6 ul |
| 10x ligase buffer (with spermidine) | 1ul  |
| T4 DNA ligase (5U/ul) (Invitrogen)  | 1ul  |
| Nuclease Free Water                 | 2 ul |

Incubate at 16°C for 2 hrs to overnight. Then heat-inactivate at 65°C for 10 min.

## (b) Partial BamHI redigestion of CIA-diPET concatemers

Purify the concatemer DNA using the Qiagen PCR purification QuickSpin kit. Then quantitate the DNA by using 1 ul for Nanodrop (Nanodrop technologies), and do a short BamHI re-digest:

|                                                |        |
|------------------------------------------------|--------|
| Concatemers                                    | 20ul   |
| 10x BamHI buffer (New England Biolabs)         | 3ul    |
| BamHI (diluted to 1U/ul) (New England Biolabs) | 0.2 ul |
| 100x BSA (New England Biolabs)                 | 0.5ul  |
| Nuclease Free Water                            | 6.3 ul |

Incubate at 37°C for 30 min, not any longer.

Quickly add 6ul of loading dye, heat at 65°C for 15 min and chill on ice before loading on PAGE gel.

## (c) PAGE purification of concatenated CIA-diPETs

Load the entire sample preferably into 1 well of a 4-20% gradient PAGE minigel, flanked by Takara Wide-Range and Invitrogen Low Mass ladders to allow sizing. Electrophorese at 200V for about 1hr. Stain for 15-30min in SYBR Green I, and visualize on the Dark Reader transilluminator (Clare Chemical) for gel excision.

## (d). Excision of concatemers

Excise the concatenated DNA in 3 separate fractions; low (400-1000bp); medium (1000-2000bp) and high (>2000bp). Place the gel slice of each excised size-fraction into a 0.6ml microfuge tube that has been pierced at the bottom with a 21G needle. This pierced tube is placed inside a 1.7ml microfuge tube, and

centrifuged at 16110 g, 4°C for 5min. The gel pieces are thus conveniently shredded and collected in the bottom of each 1.7ml tube. Add 200µl of LoTE:NH<sub>4</sub>OAc (167:33) (LoTE according to the recipe below, NH<sub>4</sub>OAc from Ambion) to each tube and elute by heating at 65°C for 2hrs.

LoTE buffer:

3 mM Tris-HCl pH 7.5 (Ambion)

0.2 mM EDTA (Ambion)

Separate the supernatant (containing the eluted concatenated DNA) away from the gel pieces with the aid of microspin filter units as before, by spinning at 16110 g, 10min, 4°C. Perform phenol/chloroform extraction on each eluted size-fraction, then ethanol precipitate:

|                          |       |
|--------------------------|-------|
| Eluted DNA fraction      | 200ul |
| 3M Sodium Acetate pH 5.2 | 20ul  |
| GlycoBlue                | 2.2ul |
| 100% Ethanol             | 800ul |

Keep at -80°C, 30min, then spin 16110 g, 4°C, 30min; wash 1x with 75% Ethanol. Resuspend the pellet in 6ul of LoTE buffer.

(e) Ligation to pZErO-1 vector

Prior to use, the pZErO-1 cloning vector is prepared by digesting 2ug of pZErO-1 plasmid DNA (Invitrogen) with 10 units of BamHI (New England Biolabs) for 2 hours at 37°C. The digested plasmid DNA is phenol chloroform extracted and ethanol precipitated, then resuspended in 60 ul of LoTE at a concentration of 33ng/µl. The plasmid may be validated by setting up a vector self-ligation as a control (there should be few colonies), as well as by running on an agarose gel.

Set up the ligation as follows:

|                                          |      |
|------------------------------------------|------|
| Concatemer DNA fraction                  | 6 ul |
| BamHI/pZErO-1                            | 1 ul |
| 5x ligase buffer (with PEG) (Invitrogen) | 2ul  |
| T4 DNA ligase (5U/uL) (Invitrogen)       | 1ul  |

Incubate at 16°C overnight. Do not heat inactivate.

Also set up self-ligation of the vector in parallel as a control. When preparing the vector for self-ligation, replace the concatemer DNA fraction with nuclease free water.

Purify each ligation reaction to remove salts before electroporation: adjust volume to 200ul with nuclease free water, perform phenol/chloroform extraction (pH 7.9) and ethanol precipitate (with GlycoBlue) at -80°C for 30min. Spin, wash the pellet at least twice with 70% ethanol, and resuspend in 12ul LoTE.

Add 1ul of the purified ligation reaction into 25ul electrocompetent cells (e.g. E. cloni from Lucigen; Top10 from Invitrogen), in pre-chilled 1.7ml microfuge tubes. Do NOT pipette up and down to mix; instead, gently stir with the pipette tip. Stand on ice for 5min, then transfer to pre-chilled Biorad electroporation cuvettes (0.1cm gap).

Stand on ice for another 5min. Electroporate using Biorad Micropulser unit, single pulse, program EC1. The time constant is usually between 4.5 to 5 ms.

Add 1ml room temperature plain LB media within 10sec of pulsing; transfer to 15ml Falcon tube, and recover at 37°C, 1hr, 200rpm shaking.

Plate between 20-50ul (out of 1ml) on a small agar plate containing Low Salt LB agar plus Zeocin (Immedia Zeocin Agar, Invitrogen) and incubate overnight at 37°C.

#### (f) Library quality control

Count the numbers of colonies and determine library efficiency after eliminating the self-ligation background. Pick 24-48 colonies for PCR screening using primers PMR011 and PMR012. Include a control PCR on the pZErO-1 vector itself.

Based on the PCR results, pick 1 to 4x 96-well plates of colonies for overnight culture (in Low Salt LB + Zeocin, Immedia, Invitrogen), plasmid purification and sequencing to determine the average number of ditags per insert.

At this stage, the library can be stored in the form of purified ligation mix at -20°C, until one wishes to perform large scale transformations, plasmid extractions and sequencing.

#### (g) Sequencing library plating and colony picking

The transformed TOP10 (Invitrogen) bacteria cells were plated out on 22x22cm agar plates (Q-trays, Genetix) with colony density of less than 2,000 per plate to facilitate robotic picking. Individual colonies were picked and cultured in 384-well plates with LB plus Zeocin (see above) at 37°C overnight. Multiple copies of 384-well plates are replicated and stored in -80°C in the presence of 15% glycerol (Sigma).

(h) Template preparation

Plasmid DNA from the pZERO-1- derived clones is prepared using the Sprintprep solid-phase kit (Agencourt).

(i) DNA sequencing

Plasmids are sequenced using the sequencing primers PMR011 and PMR012 (SEQ ID NOS: 13 and 14 respectively) to sequence in both directions.

[00155] (5) Map the CIA-diPETs

Mapping may be performed using the Compressed Suffix Array. Multiple joins across two different DNA fragments ( $n > 3$ ) shall be taken to represent real distal control regions (FIG. 5). More than 3 PETs have to show rearrangement of the same stretch of DNA – that is, tag 1 and tag 2 separated by a distance of more than 10 kB or located on different chromosomes, before any chimera consisting of tags that map to different locations of the genome, is taken to be representative of genomic rearrangements.

Example 4 - Preparation of M and G vector pGIS8 for use

(a) Obtain pGIS8 vector.

(b) Amplify by Rolling Circle Amplification as before (Templiphi kit, Amersham)

Thaw solutions on ice. Prepare 3-4 tubes of the following:

|                               |       |
|-------------------------------|-------|
| 1 ng of DNA (from Maxiprep)   | 1 ul  |
| Templiphi kit denature buffer | 10 ul |

Heat at 95°C for 3 minutes, then briefly cool on ice.

To the reaction buffer, add:

|                      |       |
|----------------------|-------|
| Templiphi kit premix | 10 ul |
|----------------------|-------|

Mix well by tapping or gentle vortexing.

Incubate at 30°C, no shaking, 16 – 18 hours.

Examine the material. It should be viscous. Quantitate DNA by picogreen fluorimetry (Quant-It kit, Molecular Probes).

(c) Perform restriction enzyme digest

|                                        |       |
|----------------------------------------|-------|
| DNA                                    | 25 ul |
| XhoI (20U/ul) (New England Biolabs)    | 1 ul  |
| EcoRI (20U/ul) (New England Biolabs)   | 1 ul  |
| 10x Buffer EcoRI (New England Biolabs) | 5 ul  |
| 100x BSA (New England Biolabs)         | 1 ul  |
| Nuclease Free Water                    | 17 ul |

Incubate at 37°C for 16 hours.

Purify with PCR purification kit (Qiagen).

Test 5 ul on a 1% agarose gel.

(d) End-it blunt

Prepare two tubes of:

|                                   |         |
|-----------------------------------|---------|
| DNA                               | 22.5 ul |
| 10x End Repair Buffer (Epicentre) | 5 ul    |
| 2.5 mM dNTP Mix (Epicentre)       | 5 ul    |
| 10 mM ATP (Epicentre)             | 5 ul    |
| End-Repair Enzyme Mix             | 1 ul    |
| Nuclease Free Water               | 11.5 ul |

Incubate at room temperature for 45 minutes, stop by heating at 70°C for 10 minutes.

[00156] It will be appreciated that various modifications and improvements may be made by a person skilled in the art without departing from the spirit and scope of the present invention.

[00157] For example, the CIA method of the present invention may be used in other kinds of cells such as yeast cells instead of mammalian cells. Also, instead of using ChIP, the method may be performed by cross-linking with a suitable fixative without need for immunoprecipitation. In this variation, the protocol is:

1. Harvest cells.
2. Cross-link with formaldehyde for 10 min at 36°C.
3. Cell lysis with bead-beater followed by centrifugation to obtain a supernatant.
4. Shear DNA by sonication, hydroshearing, repeated drawing through a hypodermic syringe needle or restriction enzyme digestion.
5. Remove unwanted proteins with SDS and Triton-X treatment.

Thereafter, the DNA may be further processed (end blunting, ligation) as described above.

[00158] Further, in another variation, rolling circle amplification may be used instead of PCR. In such a variation, the protocol following removal of unwanted proteins is:

1. End-blunting of DNA.
2. A-tailing of DNA.
3. Ligation.
4. Rolling circle amplification; may be performed with a suitable commercial kit such as the Templiphi kit from Amersham Biosciences.
5. Quantification of DNA; may be performed with the Invitrogen/Molecular Probes' PicoGreen fluorimetry kit.
6. Digestion with Mmel restriction enzyme to obtain the isolated oligonucleotides.

[00159] Further, in another variation, the A-tailing step may be omitted and a suitable blunt-ended adaptor may be used.

References

- Antequera and Bird (1993), Proc Natl Acad Sci U S A. 90(24): 11995-9
- Ausubel (1995) Current Protocols in Molecular Biology, Vol. 2, 1995, Ed. Ausubel, et al., Greene Publish. Assoc. & Wiley Interscience, Unit 3.1.15
- Bonetta (2006) Nature Methods 3(2): 141-147.
- Brenner et al (2000)., Nat Biotechnol, 2000. 18(6): p. 630-4
- Buck and Lieb (2004) Genomics 83(3):349-60
- Dekker J, Rippe K, Dekker M and Kleckner N (2002) Science 295 (5558):1306-11.
- Dunn et al (2002) Genome Research 12(11):1756-1765
- Euskirchen et al., Mol Cell Biol, 2004. 24(9): p. 3804-14
- Li and Chandrasegaran (1993) Proc. Nat. Acad. Sciences USA 90:2764-8
- Lieb et al., Nat Genet, 2001. 28(4): p. 327-34
- Margulies et al, 2005 Nature 437, 376-380 (15 Sep 2005)
- New England Biolabs Catalog 2005. New England Biolabs (Ipswich, MA)
- Szybalski, W., 1985, Gene, 40:169
- Taverner et al., Genome Biol, 2004. 5(3):210.
- US 20050255501
- US 20050059022
- US 20030068629
- US 4,766,072
- Weinmann et al., Genes Dev, 2002. 16(2):235-44)

## Claims

1. An isolated oligonucleotide comprising at least one first tag, at least one second tag and at least one linker, wherein the first and second tags are tags of a nucleic acid-protein complex.
2. The isolated oligonucleotide according to claim 1, wherein the first tag is of a first polynucleotide and the second tag is of a second polynucleotide, and wherein the first and second polynucleotides are polynucleotides of a nucleic acid-protein complex.
3. The isolated oligonucleotide according to claim 1, wherein the linker comprises at least one restriction enzyme recognition site.
4. The isolated oligonucleotide according to any one of the preceding claims, wherein the at least one linker is inserted between the tags or positioned upstream to the first tag or downstream to the second tag.
5. The isolated oligonucleotide according to claims 3 or 4, wherein the at least one restriction enzyme recognition site is for a type II restriction enzyme.
6. The isolated oligonucleotide according to claims 3 or 4, wherein the at least one restriction enzyme recognition site is for a homing restriction enzyme.
7. The isolated oligonucleotide according to any one of the preceding claims 1 to 6, wherein the at least one first tag comprises a 5' terminus and a 3' terminus of a first polynucleotide and the at least one second tag comprises a 5' terminus and a 3' terminus of a second polynucleotide, and wherein the first and second polynucleotides are polynucleotides of a nucleic acid-protein complex.
8. The isolated oligonucleotide according to any one of the preceding claims, wherein the linker comprises a first restriction recognition site recognized by

a restriction enzyme capable of cleaving a first polynucleotide to obtain the first tag, and a second restriction recognition site recognized by a restriction enzyme capable of cleaving a second polynucleotide to obtain the second tag, the first and second polynucleotides being polynucleotides of a nucleic acid-protein complex.

9. The isolated oligonucleotide according to any one of the preceding claims, wherein the linker comprises a first restriction recognition site recognized by a first restriction enzyme capable of cleaving a first polynucleotide to obtain a 3' terminus of the first polynucleotide, and a second restriction recognition site recognized by a second restriction enzyme capable of cleaving a second polynucleotide to obtain a 5' terminus of the second polynucleotide, the first and second polynucleotides being polynucleotides of a nucleic acid-protein complex.
10. The isolated oligonucleotide according to any one of the preceding claims, wherein the linker comprises a first restriction recognition site recognized by a first restriction enzyme capable of cleaving a first polynucleotide to obtain a 5' terminus of the first polynucleotide, and a second restriction recognition site recognized by a second restriction enzyme capable of cleaving a second polynucleotide to obtain a 3' terminus of the second polynucleotide, the first and second polynucleotides being polynucleotides of a nucleic acid-protein complex.
11. The isolated oligonucleotide according to any one of claims 3, 4 or 9, wherein the linker comprises a first restriction recognition site recognized by a first restriction enzyme capable of cleaving a first polynucleotide to obtain a 3' terminus of the first polynucleotide, and a second restriction recognition site recognized by a second restriction enzyme capable of cleaving a second polynucleotide to obtain a 5' terminus of the second polynucleotide, the first polynucleotide further comprises a third recognition site recognised by a third restriction enzyme capable of cleaving the first polynucleotide to obtain a 5' terminus of first polynucleotide, and the second polynucleotide

comprises a fourth recognition site recognised by a fourth restriction enzyme capable of cleaving the second polynucleotide to obtain a 3' terminus of the second polynucleotide; the first and second polynucleotides being polynucleotides of a nucleic acid-protein complex; the at least one first tag obtained from ligating the 5' and 3' termini of the first polynucleotide and the at least one second tag obtained from ligating the 5' and 3' termini of the second polynucleotide.

12. The isolated oligonucleotide according to any one of the preceding claims, wherein the nucleic acid-protein complex is part of a chromatin structure.
13. The isolated oligonucleotide according to any one of the preceding claims, wherein the polynucleotides are DNA or RNA.
14. A vector comprising the oligonucleotide of any one of the preceding claims.
15. A concatemer of oligonucleotides comprising at least two isolated oligonucleotides, each isolated oligonucleotide comprising at least one first tag, at least one second tag and at least one linker, wherein the first and second tags are tags of a nucleic acid-protein complex.
16. The concatemer of oligonucleotides according to claim 15, wherein the first tag is a tag of a first polynucleotide and the second tag is a tag of a second polynucleotide, the first and second polynucleotides being from a nucleic acid-protein complex, and wherein each isolated oligonucleotide comprises at least one restriction enzyme recognition site.
17. The concatemer of oligonucleotides according to claim 16, wherein the at least one recognition site is included in the at least one linker.
18. The concatemer of oligonucleotides according to any one of claims 16 to 18, wherein the at least one linker is inserted between the tags or positioned upstream to the first tag and/or downstream to the second tag.

19. The concatemer of oligonucleotides according to any one of claims 16 to 18, wherein the at least one restriction enzyme recognition site is for a type II restriction enzyme.
20. The concatemer of oligonucleotides according to any one of claims 15 to 19, wherein the first tag comprises a 5' terminus and a 3' terminus from the first polynucleotide and the second tag further comprises a 5' terminus and a 3' terminus from the second polynucleotide.
21. An oligonucleotide library comprising at least one oligonucleotide according to any one of claims 1 to 13.
22. A method of preparing at least one isolated oligonucleotide comprising:
  - (a) providing a nucleic acid-protein complex;
  - (b)(i) inserting at least one linker comprising at least one restriction enzyme recognition site between a first polynucleotide and a second polynucleotide of the complex;
  - (ii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing the at least one recognition site in the linker to form an oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide and the linker; and
  - (c) isolating the oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag is obtained from a second polynucleotide, the first and second polynucleotides obtained from the nucleic acid-protein complex.

23. A method of preparing at least one isolated oligonucleotide comprising:
- (a) providing a nucleic acid-protein complex;
  - (b)(i) inserting at least one linker comprising at least one restriction enzyme recognition site between a first polynucleotide and a second polynucleotide of the complex;
  - (ii) adding at least one restriction enzyme recognition site each to the 5' terminus of the first polynucleotide and the 3' terminus of the second polynucleotide;
  - (iii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing at least one recognition site to obtain cleaved fragments; and
  - (iv) ligating the cleaved fragments to form an oligonucleotide comprising a first tag comprising the 5' and 3' termini from the first polynucleotide, a second tag comprising the 5' and 3' termini from the second polynucleotide, and the linker inserted between the tags; and
  - (c) isolating the oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag is obtained from a second polynucleotide, the first and second polynucleotides obtained from the nucleic acid-protein complex.
24. The method according to claim 23, wherein the at least one restriction enzyme recognition site in step b(ii) is part of an adaptor or of a vector.
25. The method according to any one of claims 22 to 24, wherein the nucleic acid-protein complex is obtained by chromatin immunoprecipitation.

26. A method of detecting and/or identifying at least two polynucleotides from a nucleic acid-protein complex, the method comprising:
- (a) providing a nucleic acid-protein complex;
  - (b)(i) inserting at least one linker comprising at least one restriction recognition site between a first polynucleotide and a second polynucleotide of the complex;
  - (ii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing the at least one recognition site in the linker to form an oligonucleotide comprising a first tag obtained from the first polynucleotide, a second tag obtained from the second polynucleotide and the linker between the tags;
  - (c) sequencing the oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from the nucleic acid-protein complex; and
  - (d) mapping the at least two polynucleotides based on the nucleotide sequences of the first and second tags, thereby detecting and/or identifying the at least two polynucleotides.
27. A method of detecting and/or identifying at least two polynucleotides from a nucleic acid-protein complex, the method comprising:
- (a) providing a nucleic acid-protein complex;

- (b)(i) inserting at least one linker comprising at least one restriction enzyme recognition site between a first polynucleotide and a second polynucleotide of the complex;
  - (ii) adding at least one restriction enzyme recognition site each to the 5' terminus of the first polynucleotide and the 3' terminus of the second polynucleotide;
  - (iii) cleaving the first and second polynucleotides with at least one restriction enzyme recognizing at least one recognition site to obtain cleaved fragments; and
  - (iv) ligating cleaved fragments to form an oligonucleotide comprising a first tag comprising the 5' and 3' termini from the first polynucleotide, a second tag comprising the 5' and 3' termini from the second polynucleotide, and the linker inserted between the tags;
  - (c) sequencing the oligonucleotide comprising at least one first tag and at least one second tag, wherein the first tag is obtained from a first polynucleotide and the second tag obtained from a second polynucleotide, the first and second polynucleotides obtained from a nucleic acid-protein complex; and
  - (d) mapping the at least two polynucleotides based on the nucleotide sequences of the first and second tags, thereby detecting and/or identifying the at least two polynucleotides.
28. The method according to claim 27, wherein the at least one restriction enzyme recognition site in step b(iii) is part of an adaptor or of a vector.
29. The method according to any one of claims 26 to 28, wherein the nucleic acid-protein complex is obtained by chromatin immunoprecipitation.

30. The method according to any one of claims 26 to 29, wherein the oligonucleotide is concatenated with at least one further oligonucleotide obtained by steps (a) to (b) before being sequenced.
31. The method according to any one of claims 26 to 30, wherein the polynucleotides are located on the same chromosome or the polynucleotides are located on different chromosomes.