

Nucleic Acid Interaction Analysis



159159

Field of the invention

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/or identification of nucleic acid interaction components.

Background of the invention

The study of nucleic acid interactions, in particular interactions in nucleic acid-protein complexes and in chromatin, is important in understanding gene regulation and expression.

Studies of transcription factor binding sites (TFBS) by genome-wide methods such as ChIP-chip (Cawley *et al.*, 2004) and ChIP-PET (Wei *et al.*, 2006), have shown that most TFBS are not located 5' proximal to genes, suggesting extensive remote regulation. While whole genome ChIP methods can identify TFBS, they may be limited in their capability to detect actual interactions between the TFBS, hence other methods are required.

Analysis of these interactions by electron microscopy, light microscopy and fluorescence *in situ* hybridisation are constrained by technical limitations, such as lack of resolution, or difficulty in preparing samples, and may not provide enough information on these nucleic acid interactions.

In addition, several recently reported methods developed to study these interactions, purported to be high-throughput and/or unbiased, also fall short of expectations.

Dekker *et al.*, (2002) described Chromosome Conformation Capture (3C), to study nucleic acid interactions. As this method requires the use of primers for



G00002

analysing interactions, some knowledge of the interacting sites is required, and the method is limited to single point detection.

Carroll *et al.*, (2005) described a method coupling ChIP with 3C. This method indicates the nucleic acid molecules which interact with a particular protein of interest, but otherwise faces the same limitations as 3C.

Simonis *et al.*, (2006) described the chromosome conformation capture-on-chip (4C) technique wherein inverse PCR was first performed using primer sequences targeted to specific sites and subsequently, microarray analysis was used to identify other sites (as available on the microarray) which interact with the targeted sites.

Zhao *et al.*, (2006) described a strategy termed circular chromosome conformation capture (4C) wherein the captured sequences which interact with the specific target sequences are identified by sequencing. This method is limited in its application in that only sequences which interact with specific target sequences can be identified.

Dostie *et al.*, (2007) described a strategy termed 5C which includes performing a multiplex ligation-mediated amplification on a 3C library to generate a 5C library to detect global interactions. The 5C forward and reverse primer sequences are designed for all restriction fragments for the genomic region of interest. As a result, the 5C library is a quantitative "carbon copy" of a part of the 3C library, as determined or limited by the 5C primers used. This method is limited in its application in that only interactions between sites corresponding to 5C primers can be identified.

Ruan *et al* (US 2007-0238101 A1) provides a method utilising ditags capable of high-throughput, global unbiased interrogation of nucleic acid interactions and binding sites. However, in this prior art method, each tag in the ditag is short and occasionally may not be adequate for identification and/or mapping of the

interacting fragments. In addition, this prior art method is not able to separate out interactions between different nucleic acid molecules from self-ligated molecules. Further, ditags may also be formed from long linear composite molecules and these may not reflect actual nucleic acid interactions.

- 5 Solexa sequencing or sequencing by synthesis is an example of a recently developed sequencing technology (WO 2007/091077, WO 2007/010252 and WO 2007/123744). Solexa sequencing has been applied to the analysis of ChIP sequences but only in the context of comparing immuno-enriched chromatin compared to control chromatin to identify binding sites of a protein known as
10 NRSF (Johnson *et al.*, 2007).

There is thus a need for more efficient methods and robust technologies that may effectively address three-dimensional chromosomal interactions.

Summary of the invention

- According to one aspect, the invention provides a method of preparing at least
15 one isolated linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule, wherein the first and second nucleic acid molecules interact in a nucleic acid mixture and wherein the first and second tags are from different nucleic acid molecules; comprising the steps
20 of:

- (i) providing at least one nucleic acid mixture comprising at least one first nucleic acid molecule interacting with at least one second nucleic acid molecule;
25
(ii) performing a ligation reaction to ligate the ends of the nucleic acid molecules to form at least one linear composite nucleic acid molecule; and

(iii) isolating at least one linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule; wherein the first second tags are from different nucleic acid molecules.

5

After ligating the ends of the nucleic acid molecules, several types of nucleic acid molecules may be formed. (See Figure 4 as an example). These include linear composite molecules (linear interactions), circular self-ligated molecules, circular multimers (circular interactions), incompletely ligated and unligated
10 molecules (for example, DNA with no linkers and incomplete linker-containing DNA).

Further, the method according to the invention may separate out and/or provide at least one linear composite nucleic acid molecule or a pool of (more than one)
15 linear composite nucleic acid molecules from the other nucleic acid molecules.

According to another aspect, the method further comprises ligating at least one adapter to at least one end of at least one linear composite nucleic acid molecule to form at least one adapter-ligated linear composite nucleic acid
20 molecule, and isolating the adapter-ligated linear composite nucleic acid molecule(s).

According to another aspect, the present invention provides an isolated linear composite nucleic acid molecule comprising at least one first tag from at least a first nucleic acid molecule and at least a second tag from at least one second
25 nucleic acid molecule, wherein the first and second nucleic acid molecules interact in a nucleic acid mixture; and wherein the first and second tags are from different nucleic acid molecules.

In particular, for each isolated linear composite nucleic acid molecule of the invention, there are at least two tags, with each tag representing a nucleic acid

molecule which interacts with a different nucleic acid molecule represented by at least one other tag.

In particular, the nucleic acid mixture comprising interacting nucleic acid molecules comprises at least one nucleic acid-protein complex and/or
5 chromatin. The nucleic acid mixture may be produced by any method. According to one aspect, the nucleic acid mixture may be prepared by chromatin immunoprecipitation (ChIP).

Further, the linear composite nucleic acid molecule(s) of the invention may be amplified. Any standard amplification method may be used, including but not
10 limited to rolling circle amplification and/or polymerase chain reaction (PCR). Further, the linear composite nucleic acid molecule(s) of the invention may also be isolated or purified for further analysis.

According to another aspect of the invention, the linear composite nucleic acid molecules of the invention may be sequenced. The sequence of the tags
15 comprising the linear composite nucleic acid molecules may be obtained by any sequencing method. In particular, the sequencing method used may obtain sequence information from at least two tags of the linear composite nucleic acid molecules.

According to a further aspect, the sequenced tags representing interacting
20 nucleic acid molecules may be mapped to at least one reference sequence. In particular, the reference sequence comprises at least one sequence database, such as a genome sequence database. Mapping the sequenced tags to the genome may determine which nucleic acid molecules interact in close spatial proximity to which other nucleic acid molecules. The interacting nucleic acid
25 molecules identified may be located at some distance on the same chromosome or they may be on different chromosomes. Importantly, the invention accordingly may provide a global *de novo* picture of these interactions.

Accordingly, the invention may identify and/or discover protein binding site(s) including for example, transcription factor binding sites. In particular, the invention may identify and/or discover sites on different and/or separate nucleic acid molecules which interact with one another and with protein(s) and/or transcription factor(s).

Brief description of the figures

Figure 1 is a diagram which illustrates an example of the method and the isolated oligonucleotide(s) of the invention.

Figure 2 is a diagram which illustrates in general the nucleic acid interaction analysis method of Ruan *et al.*, (US application 2007-0238101 A1). This method available in the prior art generally results in short ditags because the tags are generated following restriction. In addition, these tags also include tags obtained as a result of self-ligation or circularisation of the same nucleic acid molecule.

Figure 3 is a diagram which illustrates an example of the method of the invention. In particular, Figure 3 illustrates how non-linker ligated molecules may be eliminated following purification and immobilisation and how circular molecules may be eliminated following adapter ligation and amplification.

Figure 4 is a diagram that illustrates examples of molecules which are produced following ligation and reverse cross-linking.

Figure 5 is a diagram that illustrates examples of molecules which are retained after immobilisation to streptavidin beads (SA).

Figure 6 is a diagram that illustrates examples of molecules following ligation to adapter and amplification.

Figure 7 is a diagram that illustrates the molecules which represents linear interactions of nucleic acid molecules after size selection.

Figure 8 is a diagram illustrating mapping of the tags from the method of the present invention. The vertical lines indicate the sequenced part of the tags that make up the paired tags of the isolated oligonucleotides of the present invention. The sequenced parts of the tags are used for mapping to the reference genome and the sequence is then extended by about 500 bp according to the sequence information of the reference genome to form virtual DNA fragments (firm horizontal lines). As illustrated, it is expected that most of the mapped tags of the invention will identify interactions from distal parts of the same chromosome (cis-interactions) or interactions from different chromosomes (trans-interactions) The dotted horizontal lines indicate the "span" of the interactions. (See boxed mapped interactions in Figure 8). The mapped tags (not in the box) illustrate that different mapped regions may have different mapped tag densities.

Figure 9 is a diagram illustrating mapping of the tags from the method of Ruan *et al.* (US 2007-0238101) which identifies single binding sites (from self-ligations) as well as interactions from distal part of the same chromosome and interactions from different chromosomes. The vertical lines indicate the sequenced part of the tags that make up the paired tags of the isolated oligonucleotides of this prior art invention. The sequenced parts of the tags are used for mapping to the reference genome and the sequence is then extended by about 500 bp according to the sequence information of the reference genome to form virtual DNA fragments (firm horizontal lines). The dotted horizontal lines indicate the "span" of the interactions.

Definitions

“Adapter” refers to an oligonucleotide molecule which is to be ligated or is ligated to the ends of a nucleic acid molecule. Adapters may be used for amplification, sequencing and/or inserting a nucleic acid fragment into a vector.

- 5 “Circular multimer(s)” refer to multimers formed when different or separate nucleic acid molecules are ligated at their ends to form closed nucleic acid molecules. A circular multimer will include at least two separate nucleic acid molecules ligated at the ends to form a closed molecule but may include more than two separate nucleic acid molecules.
- 10 “Composite nucleic acid molecule” refers to a nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecules, wherein the first and second tags are from different nucleic acid molecules. A composite nucleic acid molecule comprises at least two tags, but may comprise three or
15 more tags, with each tag being from a different nucleic acid molecule. A composite nucleic acid molecule may be linear or circular.

A “linear composite nucleic acid molecule” has two free ends, i.e. the ends are not joined together. A linear composite nucleic acid molecule comprises at least two tags, but may comprise three or more tags (multimer), with at least two tags
20 being from different nucleic acid molecules.

“Isolated composite nucleic acid molecule” means a composite nucleic acid molecule which is separated or isolated, and is substantially free of other nucleic acid molecules which are not composite nucleic acid molecules. Isolated composite nucleic acid molecules may comprise at least one composite nucleic
25 acid molecule or may comprise several or many different composite nucleic acid molecules as long as they are substantially free of non-composite nucleic acid molecules. Substantially free is taken to mean at least about 60% free, in

particular about 75% free, more in particular about 85%, preferably about 90% free and most preferably about 95%..

“Isolated linear composite nucleic acid molecule” means a linear composite nucleic acid molecule which is separated or isolated, and is substantially free of
5 other nucleic acid molecules which are not linear composite nucleic acid molecules. Isolated linear composite nucleic acid molecules may comprise at least one linear composite nucleic acid molecule or may comprise several or many different linear composite nucleic acid molecules as long as they are substantially free of non-linear composite nucleic acid molecules. Substantially
10 free is taken to mean at least about 60% free, in particular about 75% free, more in particular about 85%, preferably about 90% free and most preferably about 95%.

“Linker” refers to an oligonucleotide molecule which is used to link or links the ends of nucleic acid molecules together. A linker may be used for ligating to
15 another linker to join the ends of nucleic acid molecules.

“Nucleic acid material” comprises nucleic acid molecules.

“Nucleic acid-protein complex” refers to an interaction of nucleic acid material and a protein such as that found in chromatin or when a transcription factor binds to a stretch of nucleic acid.

20 A “tag” refers to a nucleic acid molecule which represents a larger nucleic acid molecule, such as a gene. The tag may be sequenced and the sequence of the tag may be used to map the tag to a genome. Tags may also be used to identify genes.

“Tag length”, in general means the length of a nucleic acid tag comprising a
25 nucleic acid fragment from a longer nucleic acid molecule. The tags of the present invention are generally longer than 19-21 bp. The tags lengths of the linear composite oligonucleotides of the present invention obtained depend on

the sonication conditions used and are expected to be at least about 50 bp or longer. However, appropriate sizes of the linear composite suitable for further analysis may be selected for. For example, through size selection of the linear composite nucleic acid molecules, tag lengths about 50-1000 bp may be selected for. In particular, tag lengths of about 50-300 bp long may be selected for. The sizes of the linear composite molecules are longer than the tag lengths. For example, if the tag lengths are 50-300 bp, the linear composite molecules would be about 200-500 bp (note that the tags are expected to be unequal and have different lengths).

“Virtual DNA fragment” refers to a tag wherein the sequence is extended from the sequenced part of the tag following mapping to the reference genome. Following sequencing, the sequence of the tag is used to map the tag to a reference genome. Once mapped, a virtual DNA fragment is generated by including the sequence from the region it is mapped to for about 500 bp, thus extending the sequence. The direction for extending the sequence information of the tag for the virtual DNA fragment is so as to include the sequence of the part of the tag which is not sequenced, and may be deduced from the method of generating the tag and the sequencing method used. For example, if Solexa sequencing is used, the sequence of the virtual DNA fragment is extended in the same direction as the sequencing method. Further, as another example, the direction for extending the sequence information of a first tag may be towards the adjacent tag ligated to it.

Detailed description of the invention

The present invention relates to a method of detecting, identifying and/or preparing at least one isolated nucleic acid molecule from a nucleic acid mixture. In particular, the invention provides a method of preparing at least one isolated composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least

one second nucleic acid molecule, wherein the first and second nucleic acid molecules interact in the nucleic acid mixture and wherein the first and second tags are from different nucleic acid molecules.

According to a first aspect, the invention provides a method of preparing at least one isolated linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule, wherein the first and second nucleic acid molecules interact in a nucleic acid mixture and wherein the first and second tags are from different nucleic acid molecules; comprising the steps of:

(i) providing at least one nucleic acid mixture comprising at least one first nucleic acid molecule interacting with at least one second nucleic acid molecule;

(ii) performing a ligation reaction to ligate the ends of the nucleic acid molecules to form at least one linear composite nucleic acid molecule; and

(iii) isolating at least one linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule; wherein the first and second tags are from different nucleic acid molecules.

According to a second aspect, the present invention also provides an isolated linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule; wherein the first and second nucleic acid molecules interact in a nucleic acid mixture; and wherein the first and second tags are from different nucleic acid molecules.

The nucleic acid mixture may comprise at least one nucleic acid-protein complex and/or chromatin. The nucleic acid mixture may be prepared by any method. In particular, the nucleic acid mixture may be prepared by a modified chromatin immunoprecipitation (ChIP). A standard ChIP protocol (Orlando *et al.*,
5 1997) may be used for obtaining the chromatin but with some modifications. In particular the modifications are that the nucleic acid material is not eluted, reverse cross-linked to remove the formaldehyde cross-links or purified. The chromatin prepared by this method may be bound to sepharose beads or any other suitable type of beads.

10 Sonication may be used to break up long nucleic acid molecules into separate molecules, however, the nucleic acid interactions are retained because of the cross-linking. In this way, interacting nucleic acid sites which are widely separated on the same long nucleic acid molecules are broken into separate molecules and may be identified by the method of the invention.

15 The ends of the nucleic acid molecules in the chromatin are then end-filled, for example by T4 DNA polymerase. A ligation reaction is then performed to ligate the ends of the nucleic acid molecules together.

This may involve a "first ligation" step comprising ligating linker(s) with at least one non-ligatable end to at least one end of the nucleic acid molecule(s). After
20 the first ligation, excess linkers may be removed and the ligated linker(s) may be converted to make the non-ligatable end(s) ligatable. The linker(s) are then ligated, thus linking the ends of the nucleic acid molecules to form at least one linker-ligated linear composite molecule.

The following are non-limiting examples of linkers which may be used for
25 ligating or linking the ends of the nucleic acid molecules.

Linker H (first strand): 5' GCG ATA TCG GAT CCA AC 3' (SEQ ID NO: 1)

Linker H (second strand): 5' GTT GGA TCC GAT ATC GCG GCC 3' (SEQ ID NO: 2)

The first strand and second strand of linker H forms the linker H with an overhang:

```

5  5'          GCGATATCGGATCCAAC      3'
   3'      CCGGCGCTATAGCCTAGGTTG      5'

```

The blunt end is for ligating to the ends of the nucleic acid molecules while the overhang is for ligating with another linker, thus joining the ends of the nucleic acid molecule.

For example, at least one dephosphorylated linker may be ligated to at least one end of the nucleic acid molecules, as shown in Figure 1. The linker(s) may be coupled to biotin for subsequent purification with avidin or streptavidin. It is understood that the linker(s) are not limited to being coupled to biotin molecules and may be coupled to any other component. In particular, the component may be used to purify or isolate the linker-ligated nucleic acid molecules.

In one non-limiting embodiment, the fifth base, T (in bold) of the first strand of Linker H (SEQ ID NO: 1) is biotinylated.

The ligation conditions may be modified or adjusted, as required, so as to maximise the ligation of the linker(s) to the end(s) of the nucleic acid molecules, or to optimize other conditions relevant to particular embodiments. Any ligation condition may be modified or adjusted, including but not limited to increasing or decreasing the time for the ligation reaction and/or concentration of the reagents. Following the first ligation, the nucleic acid molecules may be eluted off the sepharose beads. Alternatively, they may also be left on the beads.

As the linker(s) are dephosphorylated, it is expected that the linker(s) are unlikely to ligate to form dimers or concatemers of linker(s).

In addition, since the linker(s) are dephosphorylated, it is expected that the linker(s) may ligate to the phosphorylated ends of the nucleic acid molecule but cannot ligate to link together the ends of the nucleic acid molecule until they are phosphorylated.

- 5 The dephosphorylated ends of the ligated linker(s) are then phosphorylated. Following this, a second ligation is performed, to allow the ends of the nucleic acid molecules to ligate. The ligation conditions may be modified or adjusted as required, so as to maximise the ligation of the ends of different nucleic acid molecules via the linkers to form linear nucleic acid molecules and to reduce the
- 10 self-ligation or circularisation of the ends of any one nucleic acid molecule and/or formation of circular multimers. Any of the ligation conditions may be modified, including but not limited to increasing or decreasing the time for the ligation reaction and/or concentration of the reagent(s). In other words, the ligation reactions are adjusted or modified to maximise intermolecular ligation of
- 15 separate nucleic acid molecules. In particular, the ligation is performed under very dilute conditions of the nucleic acid molecules to maximise the ligation of the ends of different nucleic acid molecules and to reduce the formation of circular multimers.

Alternatively, at least one dephosphorylated linker comprising at least one

20 restriction enzyme site may be ligated to at least one end of the nucleic acid molecules in the nucleic acid mixture. This is followed by cutting the ligated linker(s) to produce ligatable ends. The ligatable ends of the ligated linker(s) may then be ligated to ligate the ends of the nucleic acid molecules. The ligatable end after restriction may be a blunt or have a cohesive end with a 5' or

25 3' overhang. In particular, a restriction enzyme which cuts rarely may be used. In other words, the restriction site for the enzyme should not occur too frequently. The reason being, to reduce the possibility of cutting the nucleic acid material in the nucleic acid mixture to very short fragments.

Accordingly, the ligated nucleic acid molecule(s) may comprise at least one linker linking the first nucleic acid molecule and the second nucleic acid molecule.

Following the second ligation, the nucleic acid material is then treated to
5 reverse the formaldehyde cross-links and purified. Purification of at least one linker-ligated nucleic acid molecule may be performed using a binding system comprising at least two components, where in at least one first component is coupled to the linker, and at least a second component binds the first component to isolate the linker-ligated nucleic acid molecule(s). The
10 components include but are not limited to streptavidin-biotin, avidin-biotin, protein-antibody and/or magnet/magnetic material.

In particular, biotinylated linker-ligated nucleic acid material is purified using streptavidin beads. Only the nucleic acid material that contains biotinylated linkers will be immobilized on the streptavidin beads. If another component is
15 bound to the linkers used, other systems of purifying the nucleic acid molecules suitable for the component may be used.

Following purification, adapters may then be ligated to the ends of the purified nucleic acid material (See Figure 1). As can be understood from Figure 3, only blunt ended, non-circularised nucleic acid material can be ligated to the
20 adapters. Self-ligated nucleic acid molecules and circular multimers cannot be ligated to the adapters. The linear composite nucleic molecule(s) with both adapters and linkers ligated may then be isolated.

Either the linker or the adapters or both may be coupled to at least one first component(s). The coupled component(s) may be used to isolate ligated
25 nucleic acid material. In particular, the coupled component(s) may be used to isolate linear composite nucleic acid molecule(s). The coupled first component(s) bind to at least one second component(s) which may be used to isolate nucleic acid molecule(s) which are coupled to the first component(s).

The components may include but are not limited to streptavidin-biotin, avidin-biotin, protein-antibody and/or magnet/magnetic material.

Accordingly, the method is able to obtain at least one isolated linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule, wherein the first and second nucleic acid molecules interact in a nucleic acid mixture, and wherein the first and second tags are from different nucleic acid molecules. In addition, the isolated linear composite nucleic acid molecule(s) may further comprise adapters.

- 10 The isolated linear composite molecule(s) may comprise at least one linker linking the tags and/or at least one adapter. Either the linker(s) and/or adapters may be biotinylated or coupled to another component for use in purifying or isolating the linker-ligated and/or adapter ligated linear composite nucleic acid molecule(s).
- 15 The method of the invention does not isolate interacting nucleic acid ligated to form circular multimers (circular interactions in Figure 5). However, it is expected that the number of non-circular ligations is high, and with careful adjustment of the ligation conditions, it may be possible to maximise the ligation of the ends of separate nucleic acid molecules and reduce circularisation of the ends of any one nucleic acid molecule and/or formation of circular multimers.

Further, it is expected that the method should not produce substantial bias as there is no aspect of the method that will make it more likely that DNA species of a particular type will only form circular interactions instead of linear interactions. In addition, while PCR of fragments of different lengths may result in a length bias, the size range of the fragments used in this protocol is small. In addition, sonication tends to produce mostly random breaks in the DNA, hence sonication will tend to have an equal chance of producing short or long fragments.

In addition, the method of the invention generates nucleic acid molecules comprising longer tags than the nucleic acid interaction analysis method disclosed in Ruan *et al.*, (US 2007-0238101 A1). The tag length of the linear composite nucleic acid molecules are expected to be longer than 19-21 bp in length. The tag lengths in the linear composite molecules are dependent on the sonication conditions used. In general, the tag lengths of the linear composite molecules may be at least about 50 bp or longer. Linear composite molecules of suitable sizes for further analysis may be selected for using size selection methods known to the person skilled in the art. For example, through size selection of the linear composite nucleic acid molecules, tag lengths about 50-1000 bp may be selected for. In particular, tag lengths of about 50-300 bp long may be selected for.

According to a further aspect of the invention, PCR may be performed to amplify the linear composite nucleic acid molecules. The linear composite nucleic acid molecules, before or after amplification, may be selected for suitable sizes, by any standard method, including gel electrophoresis and gel excision. The main considerations in selection for the appropriate sizes of the linear composite nucleic acid molecules are that the sizes should be above the size of primer dimers and unannealed adapters and below the sizes of long linear multimers. Long linear multimers are undesirable because they may complicate the data analysis. Long multimers are chimeras and may comprise several linear composite molecules linked together (Figure 6). They represent false interactions and accordingly are regarded as "interaction noise" (Figure 6). Moreover, long linear multimers may result from the ligation of several nucleic acid molecules and accordingly may represent false interactions. In particular, linear composite nucleic acid molecules with sizes of approximately 200-500 bp may be selected. (The tag length of each tag in the linear composite nucleic acid molecules are unequal and about 50-300 bp). Further, if the ligation reactions are optimised such that long linear multimers are formed only with a

low frequency, it may be possible that larger sized fragments of linear composite nucleic acid molecules may be selected. Accordingly, with size selection, an advantage is that long linear multimers may be eliminated as their sizes will be above the size range. Similarly, fragments that are too short, unannealed adapters and primer dimers may also be eliminated.

In addition, the linear composite nucleic acid molecules may be concatenated by standard methods. In particular, the length controlled concatenation method (Ruan *et al.*, US application 11/449,872) may be used.

According to yet another aspect of the invention, the isolated linear composite nucleic acid molecules may be sequenced. In particular, the sequencing method used may obtain sequence information from different regions of the nucleic acid molecule, for example, from at least two tags of the linear composite nucleic acid molecule.

For example, the isolated nucleic acid molecules may be sequenced according to the Solexa method (Illumina) which is capable of pairwise sequencing two different regions of a target nucleic acid molecule. Accordingly, the Solexa method may be adapted for sequencing at least two different tags from each linear composite molecule. However, other known sequencing methods, including but not limited to Sanger sequencing, 454 sequencing, sequencing by ligation (Applied Biosystems SOLiD system) and the true single molecule sequencing (Helicos Biosciences Corporation) may also be used.

The Solexa method is a technology that can provide ultra-high output sequencing of paired tags from the isolated oligonucleotides (WO 2007/010252, WO 2007/091077 and WO 2007/123744), the entire contents of which is herein incorporated by reference. The method, also known as Sequencing by Synthesis, basically involves the following steps: (1) fragmentation of the DNA; (2) ligation of adapters to both ends of the fragments; and (3) attachment of single stranded fragments randomly to the inside surface of a flow cell.

Following this, (4) bridge amplification is performed. The bridge amplification involves (a) allowing single stranded fragments to bend to come into contact with primers randomly bound on the inside surface of the flow cell, (b) single stranded fragments becoming double stranded fragments, which are then (c) denatured. Cycles of amplification are repeated such that millions of clonal clusters of double stranded DNA are generated in each channel of the flow cell, (5) Sequencing is then performed, using reversible, fluorescent terminators and polymerase.

Solexa sequencing is ultra-high output in the sense that it can sequence many different molecules concurrently and hence generate a large quantity of sequence information and is suitable for use with the method of the present invention.

The method of the invention provides linear composite nucleic acid molecules comprising longer tags (in particular, selected to be about 50-300 bp) than the nucleic acid interaction analysis method disclosed in Ruan *et al.*, (US 2007-0238101 A1). The tag length of the linear composite nucleic acid molecules are expected to be longer than 19-21 bp in length. In particular, the tag lengths about 50-300 bp are selected for.

The linear composite nucleic acid molecules comprising the longer tags may accordingly be sequenced directly by the high-throughput Solexa method according to the manufacturer's instructions.

The Solexa method has several important advantages: (1) the Solexa method can sequence "tags" of 20-36 base pairs per read presently. This tag length may also increase with further technology improvements (2) Solexa may be capable of sequencing tags from both sides of the DNA fragment, providing paired end tag information (3) Solexa is capable of performing reads on long DNA fragments of variable length.

Solexa sequencing may be carried out directly on the nucleic acid molecules of the invention by ligating adapters which are also Solexa adapters or primers to the nucleic acid molecules. The invention therefore provides a streamlined method wherein the isolated linear composite nucleic acid molecules may
5 directly be used for the high-throughput Solexa sequencing method.

The whole sequencing process is computerised, and many sequencing reactions occurring in parallel on the surface may be detected.
(WO 2007/123744).

It is understood that other sequencing methods may also be used for
10 sequencing the isolated linear composite nucleic acid molecules. In particular, sequencing methods capable of sequencing from at least two different regions of the nucleic acid molecules, for example, from two tags of the linear composite nucleic acid molecules may be used. As alternative examples, multiplex polony sequencing (Shendure *et al.*, 2005), mate-paired sequencing
15 with the Applied Biosystems SOLiD system and paired end sequencing using the Heliscope Single Molecule Sequencer (Helicos Biosciences Corporation) may also be used to sequence the linear composite nucleic acid molecules. Mate-paired sequencing may be carried out using the SOLiD Mate-Paired Library Sequencing Kit (Applied Biosystems).

20 Following sequencing, each tag of the linear composite nucleic acid molecules can then be mapped to the genome. The method produces nucleic acid molecules comprising longer tags (in particular, selected to be about 50-300 bp in length) which may be mapped accurately to the genome. Since the tags are longer, the rate at which the tags may map uniquely to the genome is expected
25 to be high. Each single tag may be used to identify binding sites. Paired tags from each isolated nucleic acid of the invention may be used to identify nucleic acid interactions, particularly interactions between binding sites of proteins in nucleic acid-protein complexes.

The invention provides isolated linear composite nucleic acid molecules comprising tags from at least two different or separate interacting nucleic acid molecules, and therefore eliminates nucleic acid molecules comprising tags from the ends of the same nucleic acid molecule ligating together. Such self-ligated nucleic acid molecules being circular are not present in the material for sequencing. Accordingly, following sequencing and subsequent mapping, only tags representing different or separate interacting nucleic acid molecules are obtained. (See Figure 7). The “noise” from sequences contributed by self-ligated nucleic acid molecules and long linear multimers is reduced.

10 This can be better understood as discussed below with reference to Figures 3-7. Following ligation and reverse cross-linking steps shown in Figure 3, linear interactions, self-ligations (true site), self-ligations (ChIP noise), DNA with no linkers, circular multimers (circular interactions), long linear multimers and incomplete linker-containing DNA are obtained (See Figure 4).

15 DNA with no linkers may be eliminated during the purification and immobilisation step (see Figure 3 and Figure 5).

Self-ligation (true site) of a nucleic acid molecule comprises a site that interacts with a particular protein of interest but it only represents a single site interaction. Self-ligation (ChIP noise) on the other hand, is a false negative piece of nucleic acid molecule which does not interact with the particular protein of interest at all and can be regarded as a “contamination” of the ChIP preparation. As the method of the invention provides for studying interactions of sites located at great distances or on different molecules, these self-ligated molecules may be eliminated because adapters may not be ligated to them and accordingly, they may also not be amplified by PCR. (See Figure 6).

In addition, although incomplete linker-containing DNA are also obtained after immobilisation to streptavidin beads (See Figure 5), it is likely that the presence of the streptavidin bead provides a hindrance to the ligation of adapters to one

end of these molecules and these molecules may not be amplified by PCR during the next step. Further, a possible reason why there is incomplete ligation of the linker with another linker on these molecules is because of an incomplete reversal of the reaction to make the ends ligatable. It follows that these molecules may also not ligate to the adapters and hence may not be amplified (See Figure 6).

After amplification by PCR (Figure 3), both linear interactions and long linear multimers may be obtained (Figure 6). However, long linear multimers may be eliminated by size selection (Figure 7). Size selection of nucleic acid molecules may be carried out by any standard method, including gel electrophoresis and gel excision.

The present invention allows the study of global nucleic acid interactions *de novo* without limitation to the primers or microarrays used. However, the prior art methods utilising primers may require initially an assumption of interactions (pre-conceived surmise) and may not be truly global, since the construction of primer pairs representing a global perspective of the reference genome is impractical. The present invention accordingly provides a method capable of studying global nucleic acid interactions, with longer tags representing nucleic acid interactions and a reduction in "noise".

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

EXAMPLES

Standard molecular biology techniques known in the art and not specifically described were generally followed as described in Sambrook and Russel, Molecular Cloning: A Laboratory Manual, Cold Springs Harbor Laboratory, New York (2001).

Example 1 Chromatin immunoprecipitation

Cells were grown harvested, formaldehyde cross-linked and chromatin immunoprecipitated according to standard procedures with slight modifications (see, for example Orlando *et al.*, 1997). However, following the last wash with TE buffer, the material is not eluted from the beads, reverse cross-linked or purified. Instead the material is diluted 1 in 2 and stored at 4 °C until use.

As a control, an aliquot was taken and subjected to elution and reverse cross-linking according to the standard protocol (Orlando *et al.*, 1997). The amount of DNA may be determined according to any standard protocol. The aliquot may also be checked for proper ChIP enrichment according to standard protocols (Haring *et al.*, 2007).

Example 2 End filling nucleic acid fragments with T4 DNA polymerase

Following ChIP, the beads were resuspended by pipetting. A volume of beads suspension corresponding to ~ 100ng of ChIP DNA, as determined from the DNA quantitation was used for end filling. This volume of beads were collected by centrifugation at 100 g for 5 minutes and resuspended in 150 µl of sterile distilled water.

The reaction mix was prepared as follows:

	Beads	150 μ l
	10 $\times$ Buffer for T4 DNA Polymerase (Promega)	50 μ l
	10mM dNTPs (Eppendorf)	5 μ l
5	T4 DNA Polymerase (Promega, 7.9u/ μ l)	6.3 μ l
	<u>Sterile distilled H₂O</u>	<u>438.7 μl</u>
		500 μ l

The reaction was then incubated at 37 °C for 15 minutes with gentle mixing on a rotator. (Intelli-Mixer RM-2L, Skyline). In particular, the manufacturer's pre-program F8 (30 rpm) may be used.

Following the reaction, the beads were washed thrice with 1 ml Wash buffer (10 mM Tris-Cl, pH 7.5; 1 mM EDTA; 500 mM NaCl, all reagents from Ambion). The beads were mixed well by inverting the microfuge tubes and collected by centrifugation at 100 g for 5 minutes at 4 °C.

15 Following the last wash, the beads were collected for use in ligation of biotinylated linkers.

Example 3 Ligation of dephosphorylated biotinylated linkers

The reaction mix was prepared as follows.

	Biotinylated linkers (200 ng/ μ l) (IDT)	10 μ l
	5 $\times$ T4 DNA Ligase Buffer with PEG (Invitrogen)	200 μ l
5	T4 DNA Ligase (30u/ μ l, Fermentas)	4 μ l
	H ₂ O	86 μ l
		<hr/>
		1000 μ l

The final concentration of T4 DNA ligase was at ≥ 0.1 u/ μ l. Following the final wash from above, the beads were suspended in this reaction mix and incubated at 16 °C with gentle mixing on a rotator (Intelli-Mixer RM-2L, Skyline) for 16 hours. The manufacturer's pre-program F8 (30 rpm) may be used.

Following the ligation, the beads were washed three times as described in Example 2 to remove the excess linkers.

Example 4 Phosphorylation of linkers

15 The reaction mix was prepared as follows:

	10 $\times$ T4 DNA Ligase Buffer (NEB)	100 μ l
	T4 DNA Polynucleotide Kinase (10u/ μ l, NEB)	20 μ l
	Sterile distilled H ₂ O	880 μ l
		<hr/>
		1000 μ l

20 The washed beads from the ligation reaction were resuspended in the above reaction mix and incubated at 37 °C with gentle mixing on a rotator (Intelli-Mixer

RM-2L, Skyline) for 30 minutes. In particular, the manufacturer's pre-program F8 (30 rpm) may be used.

Example 5 Second ligation reaction

5 The second ligation reaction was performed to ligate the phosphorylated linkers by adding 6 μ l of T4 DNA Ligase (30 U/ μ l, Fermentas). The contents were mixed immediately and incubated at 22 °C with gentle mixing on a rotator (Intelli-Mixer RM-2L, Skyline) for 16 hours. In particular, the manufacturer's pre-program F8 (30 rpm) may be used.

Example 6 Elution and reverse-crosslinking

10 Following ligation, the ligation reaction was separated into 575 μ l aliquots, with adjustment to the required volume using sterile distilled water if necessary. 150 μ l of Elution buffer (1 % SDS, 0.1M NaHCO₃) and 5 μ l of proteinase K (Ambion, 20 mg/ml) were added to each aliquot and the beads were incubated at 65 °C overnight to remove the cross-links and elute the nucleic acid
15 molecules.

Example 7 DNA purification

The solution was then centrifuged at 16,100 g for 2 minutes and the supernatant was transferred into new microfuge tubes. The DNA was extracted using phenol: chloroform:isoamyl alcohol (25:24:1), pH7.9 (Ambion).
20 Alternatively, a phase-gel tube (Eppendorf) may be used to extract the DNA.

The DNA was further purified using sodium acetate and isopropanol precipitation as follows:

	DNA (after phenol extraction)	~500 μ l
	3M NaOAc, pH5.2 (Ambion)	50 μ l
5	GlycoBlue (15mg/ml) (Ambion)	2 μ l
	Isopropanol (-20 °C) (Sigma)	500 μ l

The reaction was incubated at -80 °C for 1 hour and the DNA precipitate was obtained by centrifugation at 16,100 g at 4 °C for 30 minutes. The DNA pellet obtained was washed twice with 500 μ l of 75 % ethanol and resuspended in 30 μ l of EB buffer (Qiagen).

Example 8 Preparation of Dynabeads

50 μ l of resuspended Dynabeads M-280 Streptavidin (10 mg/ml), (Invitrogen) were transferred to a 1.5 ml microfuge tube using a magnetic particle collector (Invitrogen) according to the manufacturer's instructions. The beads were washed twice with 2XB&W buffer (10 mM Tris-HCL pH 7.5, 1 mM EDTA, 2 M NaCl) and resuspended in 30 μ l of 2XB&W buffer.

Example 9 Immobilization of the nucleic acid molecules comprising the paired tags

30 μ l of the DNA (from Example 7) was added to the resuspended Dynabeads prepared from above and mixed well. The mixture was then incubated at 22 °C with gentle mixing on a rotator (Intelli-Mixer RM-2L, Skyline) for 30 minutes. In particular, the manufacturer's pre-program F8 (30 rpm) may be used. The beads were washed twice with 100 μ l of 1XB&W buffer (5 mM Tris-HCL pH 7.5,

0.5 mM EDTA, 1 M NaCl). The Dynabeads were collected during each wash with a Magnetic Particle Collector (Invitrogen).

Example 10 Ligation of Solexa adapters to the immobilised nucleic acid molecules

- 5 The A-tailing of the DNA as well as the ligation of Solexa adapters (Illumina) to the immobilised nucleic acid molecules was performed using the Solexa/Illumina Paired End Sample Prep Kit item #1001809 according to the manufacturer's instructions except that the beads were washed twice after each enzymatic reaction with 100 µl of 1XB&W buffer (5 mM Tris-HCL pH 7.5, 0.5 mM EDTA, 1 M NaCl). The Dynabeads were collected during each wash with a
10 Magnetic Particle Collector (Invitrogen).

Example 11 Polymerase Chain Reaction (PCR)

- Following the ligation of the adapters, the dynabeads were washed twice with 100 µl of 1XB&W buffer, again using the Magnetic Particle Collector (Invitrogen),
15 for collecting the Dynabeads for each wash. The Dynabeads were then resuspended in 50 µl of EB buffer (Qiagen) and transferred to a fresh 1.5 ml microfuge tube. PCR was performing using Accuprime according to the Solexa preparation protocol (Illumina).

Example 12 Purification of PCR products.

- 20 The PCR products from, typically 1-25 reactions, may be pooled. The Dynabeads were then removed using the magnetic particle collector leaving the supernatant. The DNA was precipitated from the supernatant with isopropanol precipitation in the presence of GlycoBlue (Ambion) as co-precipitant as described in Example 7. After the final wash, the supernatant was discarded
25 and the DNA resuspended in 40 µl of TE buffer (Qiagen)

The DNA was then electrophoresed in Novex 6 % TBE PAGE gel, 1.0 mm, 5 well (Invitrogen cat. no. DC6264BOX) at 200 V for 30 minutes. The DNA standard size markers used were the 25 bp DNA Ladder and the Low DNA Mass Ladder (both from Invitrogen).

- 5 The DNA from the sample lanes of approximately 200-500 bp in sizes were excised and purified using the gel crush method.

Gel crush method

The excised gel pieces were collected into 0.6 ml microfuge tubes that have been pierced at the bottom with a 21G needle. The pierced 0.6 microfuge tube
10 was placed into a 1.5 ml screw-cap microfuge tube and centrifuged at 16,100 g at 4 °C for 5 minutes. The centrifugation process shreds the gel slices which are then collected in the bottom of each 1.5 ml screw-cap microfuge tube.

200 µl of TE buffer (Qiagen) was added to each 1.5 ml screw-cap microfuge tube and resuspended well with the gel pieces using a pipette tip. The gel
15 pieces should be completely immersed in the buffer.

The 1.5 screw-cap microfuge tubes containing the shredded gel pieces were frozen at -80 °C for 1-2 h, and then transferred directly to 37 °C for 16 h (maceration).

Following the maceration step, the contents of each 1.5 ml tube were
20 transferred to a SpinX centrifuge filter (Corning) and centrifuged at 16, 100 g, 4 °C for 10 minutes. The 1.5 ml tubes used for the maceration were rinsed by adding 200 µl of TE buffer (Qiagen) and subjected to a brief spin in a centrifuge to collect the TE buffer to the bottom of the tube.

The TE buffer used for the rinsing was then transferred into the corresponding
25 SpinX centrifuge containing the gel pieces, resuspending the gel pieces by

stirring briefly with the pipette tip. The TE buffer was collected by centrifuging again at 16,100 g, 4 °C for 20 minutes.

The filtrate was pooled and the DNA was precipitated with isopropanol and resuspended in 21 µl of TE buffer (Qiagen).

- 5 Alternatively the DNA may be purified with QIAEX II (Qiagen) according to the manufacturer's instructions, and eluted in 21 µl of TE buffer (Qiagen).

The quality and quantity of the recovered DNA may be determined by analysing a 1 µl of the recovered DNA with the 1000 DNA Chip kit (Agilent) according to the manufacturer's instructions. Alternatively, the quality and the quantity of the recovered DNA may be determined using picogreen fluorimetry (Quant-iT DNA assay kit, Molecular Probes) according to the manufacturer's instructions.

10

Example 13 Solexa sequencing

Solexa sequencing (Illumina) was performed on the purified PCR products according to the manufacturer's instructions.

- 15 Example 14 Bioinformatics analysis of the sequenced nucleic acid molecules comprising the paired tags

Following sequencing, the paired tags may be analysed using the software described in Chiu *et al.*, (2006). The tags were mapped to the relevant reference genome using an alignment approach based on a compressed suffix array algorithm, in which the entire relevant genome was indexed as a reference database (Hon *et al.*, 2007) and the tags were matched to the database. Alternatively, the tags may also be mapped to the relevant reference genome using the ELAND program provided by Solexa.

20

- 25 Following mapping, the sequenced information from each tag is extended by about 500 bp in the 5' to 3' direction to form a virtual DNA fragment. The virtual

DNA fragment may be used for the analysis, by determining the frequency of tag mapped to a particular region (tag density). The unboxed tags in Figure 8 illustrates that different regions have different map densities. Peaks can be identified as local maxima in the tag density, and binding sites can be identified as peaks with a high number of tags. High confidence interactions can be identified by finding multiple paired tags that map between two binding sites, and medium confidence interactions can be identified by finding multiple paired tags that map between a binding site and a non-binding site, as well as single paired tags that map between two binding sites. In addition, data obtained from the method of Ruan *et al.* (US 2007-0238101) may also be integrated with the data from the present invention to identify binding sites and interactions.

In comparison, the method described in Ruan *et al.* (US 2007-0238101 A1), the types of interactions identified include single binding sites (from self ligations) as well as interactions from distal parts of the same chromosome and interactions from different chromosomes (See Figure 9).

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.

References

- Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W. Meyer, C. A., Szary, A. J., Eeckhoute, J., Shao, W., Hestermann, E. V., Geistlinger, T. R., Fox, E. A., Silver, P. A. and Brown, M. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 122:33-43.
- Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammanna, H., Helt, G., Struhl, K. and Gingeras, T. R. (2004) Unbiased mapping of transcription factor binding sites along human chromosome 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116:499-509.
- Chiu, K. P., Wong, C-H., Chen, Q., Ariyaratne, P., Ooi, H.S., Wei, C-L., Sung, W-K. K. and Ruan, Y. (2006) PET-Tool, a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. *BMC Bioinformatics* 7:390.
- Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science* 295:1306-1311.
- Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., Dekker, J. (2007) Chromosome conformation capture carbon copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research* 16:1299-1309.
- Haring, M., Offermann, S., Danker, T., Horst, I., Peterhansel, C. and Stam, M. (2007) Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. *Plant Methods* 3: 11.
- Hon, W-K., Lam. T-K., Sadakane, K., Sung, W-K. and Yiu, S-M. (2007) A space and time efficient algorithm for constructing compressed suffix arrays. *Algorithmica* 48: 23-26.
- Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316:1497-1502.
- Orlando, V., Strutt, H. and Paro, R. (1997) Analysis of chromatin structure by *in vivo* formaldehyde cross-linking. *Methods: A companion to methods in Enzymology* 11: 205-214.

Ruan, Y., Fullwood, M. J. and Wei, C. L. US application 2007-0238101 A1.

Ruan, Y., Ng, P. Fullwood, M. J. and Lee, Y. US application 11/449,872.

5 Sambrook and Russel, *Molecular Cloning: A Laboratory Manual*, Cold Springs Harbor Laboratory, New York (2001).

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M. (2005) Accurate multiplex polony sequencing of an evolved genome. *Science* 309:1728-1732.

10 Simonis, M., Kous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture on-chip (4C). *Nature Genetics* 38(11):1348-1354.

WO 2007/010252

15 WO 2007/091077

WO 2007/123744.

Talbert, P. B. and Henikoff, S. (2006) Spreading of silent chromatin: inaction at a distance. *Nature Reviews* 7:793-803.

20 Wei, C. L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Liu, J. J., Zhao, X. D., Chew, J-L., Lee, Y. L., Kuznetsov, V. A., Sung, W-K., Miller, L. D., Lim, B., Liu, E. T., Yu, Q., Ng, H. H. and Ruan, Y. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124: 207-219.

25 Zhao, Z., Tavoosidana, G., Sjölander, M., Göndör, A., Mariano, P., Wang, S. Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S. and Ohlsson, R. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics* 38(11):1341-1347.

30

Claims

1. A method of preparing at least one isolated linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule, wherein the first and second nucleic acid molecules interact in a nucleic acid mixture and wherein the first and second tags are from different nucleic acid molecules; comprising the steps of:

(i) providing at least one nucleic acid mixture comprising at least one first nucleic acid molecule interacting with at least one second nucleic acid molecule;

(ii) performing a ligation reaction to ligate the ends of the nucleic acid molecules to form at least one linear composite nucleic acid molecule; and

(iii) isolating at least one linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule; wherein the first second tags are from different nucleic acid molecules.

2. The method according to claim 1, wherein step (iii) comprises ligating at least one adapter to at least one end of the linear composite nucleic acid molecule to form at least one adapter-ligated linear composite nucleic acid molecule and isolating the adapter-ligated linear composite nucleic acid molecule(s).

3. The method according to claim 1, wherein step (ii) comprises:

(a) ligating at least one linker with at least one non-ligatable end to at least one end of the nucleic acid molecule(s) in the nucleic acid mixture;

(b) converting the non-ligatable end(s) of the ligated linker(s) to make them ligatable; and

(c) ligating the linker(s) to ligate the end(s) of the nucleic acid molecules to form at least one linker-ligated linear composite nucleic acid molecule.

4. The method according to claim 3, wherein step (iii) comprises:

(a) isolating at least one linker-ligated linear composite nucleic acid molecule;

(b) ligating at least one adapter to at least one end of at least one linear composite nucleic acid molecule; and

(c) isolating at least one linear composite nucleic acid molecule with both adapter(s) and linker(s) ligated.

5. The method according to claim 2, wherein isolating the adapter-ligated linear composite nucleic acid molecule comprises using a binding system comprising at least two components, wherein at least one first component is coupled to the adapter, and at least one second component binds the first component to isolate the adapter-ligated linear composite nucleic acid molecule.

6. The method according to claim 4, wherein step (a) comprises using a binding system comprising at least two components, wherein at least one first component is coupled to the linker, and at least one second component binds the first component to isolate the linker-ligated linear composite nucleic acid molecule.

7. The method according to claim 5, wherein the binding system comprises the components of streptavidin-biotin, avidin-biotin, protein-antibody and/or magnet-magnetic material.

8. The method according to claim 6, wherein the binding system comprises the components of streptavidin-biotin, avidin-biotin, protein-antibody and/or magnet-magnetic material.

9. The method according to claim 1, wherein step (ii) comprises:

(a) ligating at least one dephosphorylated linker to at least one end of the nucleic acid molecules in the nucleic acid mixture;

(b) phosphorylating the dephosphorylated end of the ligated linker(s); and

(c) ligating the phosphorylated linker(s) to ligate the ends of the nucleic acid molecules to form linear composite nucleic acid molecules.

10. The method according to claim 1, wherein step (ii) comprises:

(a) ligating at least one dephosphorylated linker comprising at least one restriction enzyme site to at least one end of the nucleic acid molecules in the nucleic acid mixture;

(b) cutting the ligated linker(s) with the restriction enzyme(s) for the site(s) to produce cohesive ends; and

5 (c) ligating the cohesive ends of the linker(s) to ligate the ends of the nucleic acid molecules to form linear composite nucleic acid molecules.

11. The method according to claim 1, further comprising the step of selecting for linear composite nucleic acid molecules of 200-500 bp.

10 12. The method according to claim 1, further comprising the step of amplifying the linear composite nucleic acid molecules.

13. The method according to claim 1, further comprising sequencing at least one linear composite nucleic acid molecules.

15

14. The method according to claim 13, wherein sequencing the linear composite nucleic acid molecules comprises sequencing at least two tags from the linear composite nucleic acid molecules to obtain sequence information for at least two tags.

20

15. The method according to claim 13, further comprising mapping the sequenced tags of the linear composite nucleic acid molecule to at least one reference sequence.

25 16. The method according to claim 15, wherein the reference sequence includes at least one sequence database.

17. The method according to claim 16, wherein the sequence database is at least one genome sequence database.

30

18. The method according to claim 1, wherein the nucleic acid mixture comprises at least one nucleic acid-protein complex and/or chromatin.
19. The method according to claim 1, wherein the nucleic acid mixture is obtained by chromatin immunoprecipitation.
20. An isolated linear composite nucleic acid molecule comprising at least one first tag from at least one first nucleic acid molecule and at least one second tag from at least one second nucleic acid molecule, wherein the first and the second nucleic acid molecules interact in a nucleic acid mixture; and wherein the first and second tags are from different nucleic acid molecules.
21. The isolated linear composite nucleic acid molecule according to claim 20, wherein the nucleic acid molecule further comprises at least one linker linking the tags.
22. The isolated linear composite nucleic acid molecule according to claim 20, further comprising at least one linker linking the tags and/or at least one adapter; wherein the linker and/or adapter is biotinylated and/or coupled to another component for use in purifying or isolating the linker-ligated and/or adapter-ligated linear composite nucleic acid molecule.
23. The isolated linear composite nucleic acid molecule according to claim 20 wherein the nucleic acid molecule further comprises at least one adapter.
24. The isolated linear composite nucleic acid molecule according to claim 22, wherein the adapters are for sequencing the nucleic acid molecule.

25. The isolated linear composite nucleic acid molecule according to claim 20, wherein the nucleic acid mixture comprises at least one nucleic acid-protein complex and/or chromatin.