previously distant portions of the DNA molecule. Subsequent processing can allow for the sequence of the associated segments to be ascertained, yielding read pairs whose separation on the genome extends up to the full length of the input DNA molecules. Since the read pairs are derived from the same molecule, these pairs also contain phase information.

[0006] Many aspects of health and fitness are impacted by the rich microbial communities in gastro-intestinal tracts, on skin, and in other locations. Herein are described simple and powerful approaches to revealing the full genomic complexity of such microbial communities. These techniques can allow quick, accurate, and quantitative assaying of the full genetic repertoire present in locations such the human body (e.g., gut) and other sites where microbial communities are found.

[0007] Such techniques include *in vitro* proximity-ligation methods, e.g. for fecal metagenomics applications. These techniques can provide a powerful and efficient approach to *de novo* metagenomics assembly that will allow research and biomedical analysis to move beyond methods such as single locus molecule counting or statistical inference.

[0008] The techniques of the present disclosure can provide a single, integrated workflow for accurate assembly of all major constituents of complex metagenomics communities. These techniques can enable a comprehensive understanding of the ways the microbiome (e.g., the gut microbiome) influences health and disease in humans, other animals, plants, other life forms, and environments.

[0009] Techniques disclosed herein can provide for efficient capture and representation of the diversity of microbes present in a sample, such as a human fecal sample. Also disclosed are computational approaches to metagenomics assembly that exploits the rich datatype these techniques generate. Such computational approaches can achieve highly contiguous scaffolding and strain deconvolution. Techniques of the present disclosure can provide for robust, fool-proof laboratory protocols and software products that can allow generation of a comprehensive view of a dynamic microbial environment (e.g., human gut) from a small sample (e.g., fecal sample) in a manner of days.

[0009a] In on aspect, the disclosure provides a method of assaying for nucleic acid species diversity in a heterogeneous sample comprising at least two species, comprising

a) obtaining a stabilized nucleic acid sample comprising a diverse plurality of nucleic acids from at least two species stabilized such that, for at least a first member of the plurality, a first nucleic acid segment and a second nucleic acid segment are held together independent of their common phosphodiester backbone, wherein said phosphodiester backbone is cleaved between said first nucleic acid segment and said second nucleic acid segment, and for at least a second member of the plurality, a third nucleic acid segment and a fourth nucleic acid segment are held together independent of their common phosphodiester backbone, wherein said phosphodiester backbone is cleaved between said third nucleic acid segment and said fourth nucleic acid segment;

b) tagging said first nucleic acid segment with a first tag and said second nucleic acid segment with a second tag, such that said first nucleic acid segment and said second nucleic acid segment are identifiable as arising from a common nucleic acid of the diverse plurality of nucleic acids, and tagging said third

nucleic acid segment with a third tag and said fourth nucleic acid segment with a fourth tag, such that said third nucleic acid segment and said fourth nucleic acid segment are identifiable as arising from a common nucleic acid of the diverse plurality of nucleic acids;

c) sequencing at least an identifiable portion of said first nucleic acid segment and said first tag, of said second nucleic acid segment and said second tag, of said third nucleic acid segment and said third tag, and of said fourth nucleic acid segment and said fourth tag;

d) constructing at least a first sequence scaffold comprising said first nucleic acid segment and said second nucleic acid segment and a second sequence scaffold comprising said third nucleic acid segment and said fourth nucleic acid segment; such that a plurality of segments of said diverse plurality of nucleic acids are assigned to at least one of the first or second sequence scaffold; and

e) counting a plurality of sequence scaffolds constructed,

wherein nucleic acid segments tagged such that they are identifiable as arising from a common nucleic acid of the diverse plurality of nucleic acids are assigned to a common scaffold; and

wherein the number of scaffolds generated indicates the species diversity in the heterogeneous sample.

[0010] In some examples, the disclosure provides methods that can produce high quality assemblies with far less data than previously required. For example, the methods disclosed herein provide for genomic assembly from only two lanes of Illumina HiSeq data.

[0011] In other examples, the disclosure provides methods that can generate chromosome-level phasing using a long-distance read pair approach. For example, the methods disclosed herein can phase 90% or more of the heterozygous single nucleotide polymorphisms (SNPs) for that individual to an accuracy of at least 99% or greater. This accuracy is on par with phasing produced by substantially more costly and laborious methods.

[0012] In some examples, methods that can produce fragments of genomic DNA up to megabase scale can be used with the methods disclosed herein. Long DNA fragments can be generated to confirm the ability of the present methods to generate read pairs spanning the longest fragments offered by those extractions. In some cases, DNA fragments beyond 150 kbp in length can be extracted and used to generate XLRP libraries.

[0013] The disclosure provides methods for greatly accelerating and improving *de novo* genome assembly. The methods disclosed herein utilize methods for data analysis that allow for rapid and inexpensive *de novo* assembly of genomes from one or more subjects. The disclosure provides that the methods disclosed herein can be used in a variety of applications, including haplotype phasing, and metagenomics analysis.

[0014] In certain examples, the disclosure provides for a method for genome assembly comprising the steps of: generating a plurality of contigs; generating a plurality of read pairs from data produced by probing the physical layout of chromosomes, chromatin, or reconstituted chromatin; mapping or assembling the plurality of read pairs to the plurality of contigs; constructing an adjacency matrix of contigs using the read-mapping or assembly data; and analyzing the adjacency matrix to determine a path through the contigs that represent their order and/or orientation to the genome. In some examples, the disclosure provides that at least about 90% of the read pairs are weighted by taking a function of each read's distance to the edge of the contig so as to incorporate information about which read pairs indicate short-range contacts and which read pairs indicate longer-range contacts. In other examples, the adjacency matrix can be re-scaled to down-weight the high number of contacts on some contigs that represent promiscuous regions of the genome, such as conserved binding sites for one or more agents that regulate the scaffolding interactions of chromatin, like transcriptional repressor CTCF. In other examples, the disclosure provides for a method for the genome assembly of a human subject, whereby the plurality of contigs is generated from the human subject's DNA, and whereby the plurality of read pairs is generated from analyzing the human subject's chromosomes, chromatin, or reconstituted chromatin made from the subject's naked DNA.

[0015] In some examples herein, a benefit is a reduction on the number of steps required to isolate complexes tagged so as to provide phase information. In many techniques in the prior art, complexes comprise tagged nucleic acids or tagged association moieties such as proteins or nanoparticles, for example biotin-tagged, so as to facilitate binding of complexes to a solid surface labeled with, for example, avidin or streptavidin. In some methods and compositions of the present disclosure, solid surfaces are coated with a moiety that binds complexes either directly or mediated through a solvent, such that the complex does not need to be modified with a ligand to facilitate binding to the solid surface. A number of moieties are contemplated herein, such as hydrophilic moieties, hydrophobic moieties, positively charged moieties, negatively charged moieties, PEG, polyamines, amino-moieties, poly-carboxylic acid moieties, or other moieties or combinations of moieties. In some cases the surface is a SPRI surface, such as a SPRI surface that binds the association moiety-nucleic acid complex directly or through a solvent.

[0016] The disclosure provides that a plurality of contigs can be generated by using a shotgun sequencing method comprising: fragmenting long stretches of a subject's DNA into random fragments of indeterminate size; sequencing the fragments using high throughput sequencing methods to generate a plurality of sequencing reads; and assembling the sequencing reads so as to form a plurality of contigs.

[0017] In certain examples, the disclosure provides that a plurality of read pairs can be generated by probing the physical layout of chromosomes, chromatin, or reconstituted chromatin using a chromatin capture based technique. In some examples, the chromatin capture based technique comprises, crosslinking chromosomes, chromatin, or reconstituted chromatin with a fixative agent, such as formaldehyde, to form DNA-protein cross links; cutting the cross-linked DNA-Protein with one or more nuclease enzymes (e.g., restriction enzymes) so as to generate a plurality of DNA-protein complexes comprising sticky ends; filling in the sticky ends with nucleotides containing one or more markers, such as biotin, to create blunt ends that are then ligated together; fragmenting the plurality of DNA-protein complexes into fragments; pulling down junction containing fragments by using the one or more of the markers; and sequencing the junction containing fragments using high throughput sequencing methods to generate a plurality of read pairs. In some examples, the plurality of read pairs for the methods disclosed herein is generated from data produced by probing the physical layout of reconstituted chromatin.

[0018] In some examples, the present disclosure provides methods for generating a tagged sequence, comprising: binding the DNA molecule to an association molecule; cutting the bound DNA-Protein so as to generate a plurality of DNA-protein complexes comprising segment ends; ligating the segment ends to tags; and sequencing the junction containing fragments using high throughput sequencing methods to generate a plurality of read pairs. A number of association molecules that bind DNA are contemplated, including chromatin constituents sensu strictu such as histones, but also chromatin constituents more generally defined, such as DNA binding proteins, transcription factors, nuclear proteins, transposons, or non-polypeptide DNA binding association molecules such as nanoparticles having surfaces comprising DNA- affinity molecules. In some cases, the tags are ligated to segment ends, for example using ligases or using transposases loaded using tag molecules. In some cases, the segment ends comprising a common tag are assigned to a common molecule of origin, which is often indicative of phase. In some examples, the plurality of read pairs for the methods disclosed herein is generated from data produced by probing the physical layout of reconstituted chromatin.

[0019] In various examples, the disclosure provides that a plurality of read pairs can be determined by probing the physical layout of chromosomes or chromatin isolated from cultured cells or primary tissue. In other examples, the plurality of read pairs can be determined by probing the physical layout of reconstituted chromatin formed by complexing naked DNA obtained from a sample of one or more subjects with isolated histones.

[0020] The disclosure provides methods to determine haplotype phasing comprising a step of identifying one or more sites of heterozygosity in the plurality of read pairs, wherein phasing data for allelic variants can be determined by identifying read pairs that comprise a pair of heterozygous sites.

[0021] In various examples, the disclosure provides methods for high-throughput bacterial genome assembly, comprising a step of generating a plurality of read pairs by probing the physical layout of a plurality of microbial chromosomes using a modified chromatin capture based method, comprising the

modified steps of: collecting microbes from an environment; adding a fixative agent, such as formaldehyde, so as to form cross-links within each microbial cell, and wherein read pairs mapping to different contigs indicate which contigs are from the same species.

[0022] In some examples, the disclosure provides methods for genome assembly comprising: (a) generating a plurality of contigs; (b) determining a plurality of read pairs from data generated by probing the physical layout of chromosomes, chromatin, or reconstituted chromatin; (c) mapping the plurality of read pairs to the plurality of contigs; (d) constructing an adjacency matrix of contigs using the read-mapping data; and (e) analyzing the adjacency matrix to determine a path through the contigs that represent their order and/or orientation to the genome.

[0023] The disclosure provides methods to generate a plurality of read pairs by probing the physical layout of chromosomes, chromatin, or reconstituted chromatin using a chromatin capture based technique. In some examples, the chromatin capture based technique comprises (a) crosslinking chromosomes, chromatin, or reconstituted chromatin with a fixative agent to form DNA-protein cross links; (b) cutting the crosslinked DNA-Protein with one or more nuclease (e.g., restriction) enzymes so as to generate a plurality of DNA-protein complexes comprising sticky ends; (c) filling in the sticky ends with nucleotides containing one or more markers to create blunt ends that are then ligated together; (d) shearing the plurality of DNA-protein complexes into fragments; (e) pulling down junction containing fragments by using one or more of the markers; and (f) sequencing the junction containing fragments using high throughput sequencing methods to generate a plurality of read pairs.

[0024] In certain examples, the plurality of read pairs is determined by probing the physical layout of chromosomes or chromatin isolated from cultured cells or primary tissue. In other examples, the plurality of read pairs is determined by probing the physical layout of reconstituted chromatin formed by complexing naked DNA obtained from a sample of one or more subjects with isolated histones.

[0025] In some examples, at least about 60%, about 70%, about 80%, about 90%, about 95% or about 99% or more of the plurality of read pairs are weighted by taking a function of the read's distance to the edge of the contig so as to incorporate a higher probability of shorter contacts than longer contacts. In some examples, the adjacency matrix is re-scaled to down-weight the high number of contacts on some contigs that represent promiscuous regions of the genome.

[0026] In certain examples, the promiscuous regions of the genome include one or more conserved binding sites for one or more agents that regulate the scaffolding interactions of chromatin. In some examples, the agent is transcriptional repressor CTCF.

[0027] In some examples, the methods disclosed herein provide for the genome assembly of a human subject, whereby the plurality of contigs is generated from the human subject's DNA, and whereby the plurality of read pairs is generated from analyzing the human subject's chromosomes, chromatin, or reconstituted chromatin made from the subject's naked DNA.

[0028] In other examples, the disclosure provides methods for determining haplotype phasing, comprising identifying one or more sites of heterozygosity in the plurality of read pairs, wherein phasing

data for allelic variants can be determined by identifying read pairs that comprise a pair of heterozygous sites.

[0029] In yet other examples, the disclosure provides methods for meta-genomics assemblies, wherein the plurality of read pairs is generated by probing the physical layout of a plurality of microbial chromosomes using a modified chromatin capture based method, comprising: collecting microbes from an environment; and adding a fixative agent so as to form cross-links within each microbial cell, and wherein read pairs mapping to different contigs indicate which contigs are from the same species. In some examples, the fixative agent is formaldehyde.

[0030] In some examples, the disclosure provides methods of assembling a plurality of contigs originating from a DNA molecule, comprising generating a plurality of read-pairs from the DNA molecule and assembling the contigs using the read-pairs, wherein at least 1% of the read-pairs span greater than 50 kB on the DNA molecule and the read-pairs are generated within 14 days. In some examples, at least 10% of the read-pairs span a distance greater than 50 kB on the DNA molecule. In some examples, at least 1% of the read-pairs span a distance greater than 100 kB on the DNA molecule. In some cases, the read-pairs are generated within 7 days.

[0031] In other examples, the disclosure provides methods of assembling a plurality of contigs originating from a single DNA molecule, comprising generating a plurality of read-pairs from the single DNA molecule *in vitro* and assembling the contigs using the read-pairs, wherein at least 1% of the read-pairs span a distance greater than 30 kB on the single DNA molecule. In some examples, at least 10% of the read-pairs span a distance greater than 30 kB on the single DNA molecule. In other examples, at least 1% of the read-pairs span a distance greater than 50 kB on the single DNA molecule.

[0032] In yet other examples, the disclosure provides methods of haplotype phasing, comprising generating a plurality of read-pairs from a single DNA molecule and assembling a plurality of contigs of the DNA molecule using the read-pairs, wherein at least 1% of the read-pairs spans a distance greater than 50 kB on the single DNA molecule and the haplotype phasing is performed at greater than 70% accuracy. In some examples, at least 10% of the read-pairs span a distance greater than 50 kB on the single DNA molecule. In other examples, wherein at least 1% of the read-pairs span a distance greater than 100 kB on the single DNA molecule. In some examples, the haplotype phasing is performed at greater than 90% accuracy.

[0033] The disclosure provides methods of haplotype phasing, comprising generating a plurality of read-pairs from a single DNA molecule *in vitro* and assembling a plurality of contigs of the DNA molecule using the read-pairs, wherein at least 1% of the read-pairs spans a distance greater than 30 kB on the single DNA molecule and the haplotype phasing is performed at greater than 70% accuracy. In some examples, at least 10% of the read-pairs span a distance greater than 30 kB on the single DNA molecule. In other examples, at least 1% of the read-pairs span a distance greater than 50 kB on the single DNA molecule. In yet other examples, the haplotype phasing is performed at greater than 90% accuracy. In some examples, the haplotype phasing is performed at greater than 70% accuracy.

[0034] In some examples, the disclosure provides methods of generating a first read-pair from a first DNA molecule, comprising: (a) binding the first DNA molecule to a plurality of association molecules *in vitro*, wherein the first DNA molecule comprises a first DNA segment and a second DNA segment; (b) tagging the first DNA segment and the second DNA segment and thereby forming at least one tagged DNA segment; and (c) sequencing the tagged DNA segment, or at least a recognizable portion of the tagged DNA segment, such as a portion adjacent to the tag or a portion at an opposite end from the tagged end, and thereby obtaining the tagged sequence, wherein the plurality of association molecules are not covalently modified with an affinity label prior to and during steps (a), and (b).

[0035] In certain examples, the present disclosure provides methods of generating a tagged sequence from a first DNA molecule, comprising: (a) crosslinking binding said first DNA molecule to a plurality of association molecules in vitro; (b) immobilizing said first DNA molecule on a solid support; (c) severing said first DNA molecule to generate a first DNA segment and a second DNA segment; (d) tagging said first DNA segment and said second DNA segment and thereby forming at least one tagged DNA segment; and sequencing said tagged DNA segment, or at least a recognizable portion of the tagged DNA segment, such as a portion adjacent to the tag or a portion at an opposite end from the tagged end, or sequencing a recognizable portion of each end of the tagged DNA segment, and thereby obtaining said tagged sequence, wherein said first DNA molecule is directly bound to said solid support. In some examples, the solid support comprises a polymer bead (e.g. SPRI bead) that binds to DNA without further modifications with any affinity label (e.g. biotin, streptavidin, avidin, polyhistidine, digoxigenin, EDTA, or derivatives thereof).

[0036] In some examples, a plurality of association molecules, such as from reconstituted chromatin, are cross-linked to the first DNA molecule. In some examples, the association molecules comprise amino acids. In some cases, the association molecules are peptides or proteins. In certain examples, the association molecules are histone proteins. In some cases, the histone proteins are from a different source than the first DNA molecule. In various examples, the association molecules are transposases. In some cases, the first DNA molecule is non-covalently bound to the association molecules. In other cases, the first DNA molecule is covalently bound to the association molecules. In certain examples, the first DNA molecule is crosslinked to the association molecules. In certain examples, the first DNA molecule is cross-linked with a fixative agent. In some examples, the fixative agent is formaldehyde. In various examples, the method comprises immobilizing the plurality of association molecules on a solid support. In some cases, the solid support is a bead. In some examples, the bead comprises a polymer. In some examples, the polymer is polystyrene. In certain examples, the polymer is polyethylene glycol (PEG). In certain examples, the bead is a magnetic bead. In some examples, the bead is a solid-phase reversible immobilization (SPRI) bead. In certain cases, the solid support comprises a surface, wherein the surface comprises a plurality of carboxyl groups. In various cases, the solid support is not covalently linked to any polypeptide (e.g. streptavidin). In some

cases, the association molecule is not covalently linked to an affinity label (e.g. biotin) prior to immobilization to the solid support.

[0037] In some examples, the first DNA segment and the second DNA segment are generated by severing the first DNA molecule. In some cases, the first DNA molecule is severed after the first DNA molecule is bound to the plurality of association molecules. In certain cases, the first DNA molecule is severed using a restriction enzyme (e.g. MbolI). In some cases, the first DNA molecule is severed using a transposase (e.g. Tn5). In other cases, the first DNA molecule is severed using a physical method (e.g. sonication, mechanical shearing). In certain examples, the first DNA and the second DNA segment are modified with an affinity label. In some examples, the affinity label can comprise biotin, which can be captured with a streptavidin bead, an avidin bead, or derivatives thereof. In certain examples, the affinity label is a biotin-modified nucleoside triphosphate (dNTP). In some examples, the affinity label is a biotin-modified deoxyribocytosine triphosphate (dCTP). In some examples, the affinity label is a biotin-modified deoxyribocytosine triphosphate (dGTP). In some examples, the affinity label is a biotin-modified deoxyribocytosine triphosphate (dATP). In some examples, the affinity label is a biotin-modified deoxyribocytosine triphosphate (dUTP). In certain cases, the first DNA segment is tagged at at least a first end with a first tag and the second DNA segment is tagged at at least a second end with a second tag. In certain examples, the first tag and the second tag are identical. In various examples, the first DNA segment and the second DNA segment are tagged using a transposase (e.g. Tn5). In some cases, the first DNA segment is tagged with the second DNA segment and the second DNA segment is tagged with the first DNA segment. For example, the first DNA segment can be linked to the second DNA segment. In some examples, the first DNA segment is linked to the second DNA segment using a ligase. In some cases, the linked DNA segment is severed prior to the sequencing in step (c). In certain examples, the linked DNA segment is severed using a restriction enzyme (e.g. ExoIII). In other cases, the linked DNA segment is severed using a physical method (e.g. sonication, mechanical shearing).

[0038] In some examples, the first DNA segment is washed for less than about 10 times before the first DNA segment is linked to the second DNA segment. In some examples, the first DNA segment is washed for less than about 6 times before the first DNA segment is linked to the second DNA segment. In some examples, the method comprises connecting the linked DNA segment to sequencing adaptors.

[0039] In certain examples, the method comprises assembling a plurality of contigs using the tagged sequence. In some examples, each of the first and the second DNA segment are connected to at least one affinity label and the linked DNA segment is captured using the affinity label. In various examples, the method comprises phasing the first DNA segment and the second DNA segment using the tagged sequence. In some cases, ‘tagging’ is effectuated by ligating a first DNA segment to a second DNA segment, thereby generating a read pair segment.

[0040] In some examples, the method comprises: (a) providing a plurality of association molecules, such as from reconstituted chromatin, to at least a second DNA molecule; (b) crosslinking the association

molecules to the second DNA molecule and thereby forming a second complex *in vitro*; (c) severing the second complex thereby generating a third DNA segment and a fourth segment; (d) linking the third DNA segment with the fourth DNA segment and thereby forming a second linked DNA segment; and (e) sequencing the second linked DNA segment and thereby obtaining a second read-pair. In some examples, less than 40% of the DNA segments from the DNA molecules are linked with DNA segments from any other DNA molecule. In some examples, less than 20% of the DNA segments from the DNA molecules are linked with DNA segments from any other DNA molecule.

[0041] In some examples, the disclosure provides methods of generating a first read-pair from a first DNA molecule comprising a predetermined sequence, comprising: (a) providing one or more DNA-binding molecules to the first DNA molecule, wherein the one or more DNA-binding molecules bind to the predetermined sequence; (b) crosslinking the first DNA molecule *in vitro*, wherein the first DNA molecule comprises a first DNA segment and a second DNA segment; (c) linking the first DNA segment with the second DNA segment and thereby forming a first linked DNA segment; and (d) sequencing the first linked DNA segment and thereby obtaining the first read-pair; wherein the probability that the predetermined sequence appears in the read-pair is affected by the binding of the DNA-binding molecule to the predetermined sequence.

[0042] In some examples, the DNA-binding molecule is a nucleic acid that can hybridize to the predetermined sequence. In some examples the nucleic acid is RNA. In other examples, the nucleic acid is DNA. In other examples, the DNA-binding molecule is a small molecule. In some examples, the small molecule binds to the predetermined sequence with a binding affinity less than 100 μM. In some examples, the small molecule binds to the predetermined sequence with a binding affinity less than 1 μM. In some examples, the DNA-binding molecule is immobilized on a surface or a solid support.

[0043] In some examples, the probability that the predetermined sequence appears in the read-pair is decreased. In other examples, the probability that the predetermined sequence appears in the read-pair is increased.

[0044] The present disclosure provides methods for generating a plurality of tagged sequences from a plurality of DNA molecules, comprising: (a) binding the plurality of DNA molecules to a plurality of association molecules *in vitro*; (b) severing each of the DNA molecules to generate at least a plurality of DNA segments; (c) tagging at least a portion of the DNA segments to form a plurality of tagged DNA segments; and (d) sequencing the tagged DNA segments, or at least a recognizable portion of the tagged DNA segments, such as a portion adjacent to the tag or a portion at an opposite end from the tagged end, to obtain a plurality of tagged sequences; wherein the plurality of association molecules are not covalently modified with an affinity label prior to and during steps (a) and (b). In some cases, less than 40% of DNA segments from the DNA molecules are linked with DNA segments from any other DNA molecule. In some cases, less than 20% of DNA segments from the DNA molecules are linked with DNA segments from any other DNA molecule.

[0045] In some examples, the association molecules comprise amino acids joined by peptide bonds. In certain examples, the association molecules are polypeptides or proteins. In some examples, the association molecules are histone proteins. In some examples, the histone proteins are from a different source than the DNA molecules. For example, the histone proteins can be isolated from a non-human organism and the DNA molecules can be isolated from humans. In various examples, the association molecules are transposases (e.g. Tn5). In some cases, the first DNA molecule is non-covalently bound to the association molecules. In other cases, the first DNA molecule is covalently bound to the association molecules. In certain examples, the first DNA molecule is crosslinked to the association molecules. In some examples, the DNA molecules are cross-linked with a fixative agent. For example, the fixative agent can be formaldehyde. In some cases, the method comprises immobilizing the plurality of association molecules on a plurality of solid supports. In certain cases, the solid supports are beads. In some examples, the beads comprise a polymer. In some examples, the polymer is polystyrene. In certain examples, the polymer is polyethylene glycol (PEG). In certain examples, the beads are magnetic beads. In some examples, the beads are SPRI beads. In various examples, the solid support comprises a surface, wherein the surface comprises a plurality of carboxyl groups. In various cases, the solid support is not covalently linked to any polypeptide (e.g. streptavidin). In some cases, the association molecule is not covalently linked to an affinity label (e.g. biotin) prior to immobilization to the solid support.

[0046] In some examples, the first DNA molecule is severed after the first DNA molecule is bound to the plurality of association molecules. In some cases, the first DNA molecule is severed using a restriction enzyme (e.g. MboII). In certain cases, the first DNA molecule is severed using a transposase (e.g. Tn5). In certain examples, the portion of the DNA segments are modified with an affinity label. In some cases, the affinity label comprises biotin. In some examples, the affinity label is a biotin-modified nucleoside triphosphate (dNTP). In some examples, the biotin-modified nucleoside triphosphate (dNTP) is a biotin-modified deoxyribocytosine triphosphate (dCTP). In some cases, a portion of the DNA segments are tagged at tat least a first end with a first tag. In some examples, the DNA segments are tagged using a transposase. In various cases, a portion of the DNA segments are tagged by linking each of said DNA segments to at least one other DNA segment. In some examples, the portion of DNA segments are linked to the other DNA segments using a ligase. In certain cases, the linked DNA segment is severed prior to step (c). In various cases, the linked DNA segment is severed using a physical method (e.g. sonication, mechanical shearing). In some examples, the method comprises connecting the linked DNA segments to sequencing adaptors.

[0047] In some cases, the DNA segments are washed for less than about 10 times before the DNA segments are linked to form the linked DNA segments. In certain cases, the DNA segments are washed for less than about 6 times before the DNA segments are linked to form the linked DNA segments. In various cases, the method comprises assembling a plurality of contigs of the DNA molecules using the tagged segments. In some cases, the method comprises phasing the DNA segments using the tagged segments.

[0048] The disclosure provides an *in vitro* library comprising a plurality of read-pairs each comprising at least a first sequence element and a second sequence element, wherein the first and the second sequence elements originate from a single DNA molecule and wherein at least 1% of the read-pairs comprise first and second sequence elements that are at least 50 kB apart on the single DNA molecule. In some examples, at least 10% of the read-pairs comprise first and second sequence elements that are at least 50 kB apart on the single DNA molecule. In other examples, at least 1% of the read-pairs comprise first and second sequence elements that are at least 100 kB apart on the single DNA molecule. In some examples, less than 20% of the read-pairs comprise one or more predetermined sequences. In some examples, less than 10% of the read-pairs comprise one or more predetermined sequences. In some examples, less than 5% of the read-pairs comprise one or more predetermined sequences.

[0049] In some examples, the predetermined sequences are determined by one or more nucleic acids that can hybridize to the predetermined sequences. In some examples, the one or more nucleic acids is RNA. In other examples, the one or more nucleic acids is DNA. In some examples, the one or more nucleic acids is immobilized to a surface or a solid support.

[0050] In some examples, the predetermined sequences are determined by one or more small molecule. In some examples, the one or more small molecule binds to the predetermined sequences with a binding affinity less than 100 μM. In some examples, the one or more small molecule binds to the predetermined sequences with a binding affinity less than 1 μM.

[0051] The disclosure provides a composition comprising a DNA fragment and a plurality of association molecules, such as from reconstituted chromatin, wherein: (a) the association molecules are cross-linked to the DNA fragment in an *in vitro* complex; and (b) the *in vitro* complex is immobilized on a solid support.

[0052] The disclosure provides a composition comprising a DNA fragment, a plurality of association molecules, and a DNA-binding molecule, wherein: (a) the DNA-binding molecule is bound to a predetermined sequence of the DNA fragment; and (b) the association molecules are cross-linked to the DNA fragment. The DNA-binding molecule is a nucleic acid that can hybridize to the predetermined sequence in some cases. In some examples, the nucleic acid is RNA. In other examples, the nucleic acid is DNA. In some examples, the nucleic acid is immobilized to a surface or a solid support. In other examples, the DNA-binding molecule is a small molecule. In some examples, the small molecule binds to the predetermined sequence with a binding affinity less than 100 μM. In other examples, the small molecule binds to the predetermined sequence with a binding affinity less than 1 μM.

[0053] The present disclosure provides a composition comprising a plurality of association molecules bound to a DNA fragment in an in vitro complex, wherein said in vitro complex is immobilized on a solid support, and wherein said solid support is not covalently linked to any polypeptides. In some cases, the solid support is not covalently linked to streptavidin. In some cases, the solid support is a bead. In some examples, the bead comprises a polymer. In some examples, the polymer is polystyrene. In certain examples, the polymer is polyethylene glycol (PEG). In certain examples, the bead is a magnetic bead. In

some examples, the bead is a solid-phase reversible immobilization (SPRI) bead. In certain cases, the solid support comprises a surface, wherein the surface comprises a plurality of carboxyl groups. In various cases, the solid support is not covalently linked to any polypeptide (e.g. streptavidin).

[0054] In some examples, the association molecules comprise amino acids bound by peptide bonds. In some examples, the association molecules are peptides or proteins. In certain examples, the association molecules are histone proteins. In some cases, the histone proteins are from a different source than the first DNA molecule. In certain examples, the association molecules are transposases. In some cases, the first DNA molecule is non-covalently bound to the association molecules. In other cases, the first DNA molecule is non-covalently bound to the association molecules. In some examples, the first DNA molecule is crosslinked to the association molecules. In certain examples, the first DNA molecule is cross-linked with a fixative agent. In some examples, the fixative agent is formaldehyde.

[0055] In certain examples, the DNA fragment is modified with an affinity label. In some examples, the affinity label can comprise biotin, which can be captured with a streptavidin bead, an avidin bead, or derivatives thereof. In certain examples, the affinity label is a biotin-modified nucleoside triphosphate (dNTP). In some examples, the affinity label is a biotin-modified deoxyribocytosine triphosphate (dCTP). In some cases, the linked DNA segment is further severed prior to the sequencing in step (c). In certain examples, the linked DNA segment is severed using a restriction enzyme (e.g. ExoIII). In other cases, the linked DNA segment is severed using a physical method (e.g. sonication, mechanical shearing).

[0056] Methods and compositions disclosed herein are useful for the assembly of genome information into scaffolds up to and including phased whole chromosomes. In some cases the information generated herein guides assembly of previously generated sequence information into scaffolds up to and including phased whole chromosomes. In some cases the methods and compositions herein are used to assemble de novo generated nucleic acid information into phased scaffolds up to and including whole chromosomes.

[0057] Tag information does not in all cases strictly correspond to phase, but is informative as to phase information. Generally referring to the disclosure herein, the presence of a common tag pattern on a pair of sequence reads indicates that the reads either 1) originated from a common molecule, or 2) are shared in common by chance.

[0058] In most cases, common tagging will not arise by chance, and thus most commonly tagged sequences, particularly commonly tagged sequences that are independently mapped to a common contig, are safely inferred to map to a common phase of that contig, that is, to the same haploid molecule of a diploid organism. Groups of reads that map together to a single or a few contigs suspected of being adjacent and that share a tag sequence are likely to be in phase on a single molecule. Groups of reads that share a common tag sequence but that map to contigs suspected to be on separate chromosomes, for example, are more likely to have obtained their common tag sequences by chance. Multiple instances of sequence clusters sharing the exact tag sequence but mapping to two separate contigs or suspected chromosomes, however, may indicate that a translocation has occurred by which a fragment of one

corresponding to how many scaffolds are generated; wherein the number of scaffolds generated corresponds to the nucleic acid molecular diversity of the heterogeneous sample. In some examples, tagging said first nucleic acid segment and said second nucleic acid segment comprises adding a first oligo to the first nucleic acid segment and adding a second oligo to the second segment, said first oligo and said second oligo sharing a common sequence. In some examples, nucleic acid segments having said common oligo sequence are assigned to a common scaffold. In some examples, the method further comprises mapping said identifiable portion of said first nucleic acid segment to a contig dataset, and including any matching contig of said contig data set into said common scaffold. In some examples, the contig data set is concurrently generated. In some examples, the contig dataset is obtained from a database. In some examples,tagging said first nucleic acid segment and said second nucleic acid segment comprises ligating said first nucleic acid segment to said second nucleic acid segment, and wherein said first nucleic acid segment and said second nucleic acid segment are assigned to a common scaffold. In some examples, the method further comprises mapping said identifiable portion of said first nucleic acid segment to a contig dataset, and including any matching contig of said contig data set into said common scaffold. In some examples, the contig data set is concurrently generated. In some examples, the contig dataset is obtained from a database. In some examples, the heterogeneous sample comprises a plurality of allelic variants. In some examples,the number of allelic variants is greater than the number of scaffolds. In some examples, the number of allelic variants is equal to the number of number of scaffolds generated. In some examples, said phosphodiester backbone is cleaved subsequent to said obtaining a stabilized sample. In some examples, said stabilized sample is contacted to a crosslinking agent. In some examples, said stabilized sample is an FFPE sample. In some examples, the method further comprises contacting said heterogeneous sample to a reverse transcriptase. In some examples, the method further comprises searching at least one of said scaffold against a nucleic acid sequence database. In some examples, the method further comprises categorizing said scaffold as novel if nucleic acid sequence uniquely mapping to said scaffold is absent from said database. In some examples, the method further comprises categorizing said scaffold as corresponding to a sample condition when a plurality of samples correlating to said condition have said scaffold and if a plurality of samples lacking said condition lack said sample In some examples, the heterogeneous sample comprises nucleic acids mapping to at least two individuals of a common species. In some examples, the heterogeneous sample comprises nucleic acids mapping to at least three individuals of a common species. In some examples, the heterogeneous sample comprises nucleic acids mapping to at least two species In some examples, the heterogeneous sample comprises nucleic acids mapping to at least three species. In some examples, the heterogeneous sample comprises nucleic acids mapping to at least four species. In some examples, the sequence reads assemble into at least two nucleic acid scaffolds without reference to exogenous sequence information. In some examples, the sequence reads assemble into at least three nucleic acid scaffolds without reference to exogenous sequence information. In some examples, the sequence reads assemble into at least two nucleic acid scaffolds, such that at least 50% of a first genome and at least 50% of a second genome are represented in said at least two nucleic acid scaffolds. In some

examples, the sequence reads assemble into at least two nucleic acid scaffolds, such that at least 60% of a first genome and at least 60% of a second genome are represented in said at least two nucleic acid scaffolds. In some examples, the sequence reads assemble into at least two nucleic acid scaffolds, such that at least 70% of a first genome and at least 70% of a second genome are represented in said at least two nucleic acid scaffolds. In some examples, the sequence reads assemble into at least two nucleic acid scaffolds, such that at least 80% of a first genome and at least 80% of a second genome are represented in said at least two nucleic acid scaffolds. In some examples, the method comprises using SPRI beads. In some examples, the stabilized sample comprises no greater than about 5 micrograms of DNA.

INCORPORATION BY REFERENCE

[0079] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference. All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference in its entirety as well as any references cited therein.

BRIEF DESCRIPTION OF THE DRAWINGS

[0080] The novel features of the disclosure are set forth with particularity in the appended claims. A better understanding of the features and advantages of the disclosure will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the disclosure are utilized, and the accompanying drawings of which:

[0081] **Figure 1** presents an illustration of genome assembly using high-throughput sequencing reads. The genome to be assembled is shown (top). Typically, genomes have many repeat sequences that are difficult to assemble. Random, high-throughput sequence data from genomes (middle) are collected and assembled into “contigs” in regions that are unique in the genome (bottom). Contig assembly generally stops at the many repeat sequences. The final output is a set of thousands of contigs whose order and orientation relative to one another are not known. In the figure, they are arbitrarily numbered from longest to shortest.

[0082] **Figures 2A-D** illustrates a chromatin capture based protocol of the disclosure: **(A)** demonstrates where DNA is cross-linked and processed to created biotinylated junction fragments for sequencing; and **(B-D)** provide contact map data on human chr14 for a variety of restriction enzymes. As shown, most contacts are local along the chromosome.

[0083] **Figures 3A-C** provides methods of the disclosure using chromatin capture sequence data to assist genome assembly: **(A)** illustrates where DNA is cross-linked and processed using a chromatin capture based protocol; **(B)** demonstrates where read-pair data is mapped to assembled contigs, generated from random shotgun sequencing and assembly; and **(C)** illustrates that after filtering and weighting, an adjacency matrix summarizing all inter-contig read pair data can be constructed. This matrix can be re-ordered to indicate the correct assembly path. As shown, most of the read pairs will map within a contig.

CLAIMS

WHAT IS CLAIMED IS:

1. A method of assaying for nucleic acid species diversity in a heterogeneous sample comprising at least two species, comprising

a) obtaining a stabilized nucleic acid sample comprising a diverse plurality of nucleic acids from at least two species stabilized such that, for at least a first member of the plurality, a first nucleic acid segment and a second nucleic acid segment are held together independent of their common phosphodiester backbone, wherein said phosphodiester backbone is cleaved between said first nucleic acid segment and said second nucleic acid segment, and for at least a second member of the plurality, a third nucleic acid segment and a fourth nucleic acid segment are held together independent of their common phosphodiester backbone, wherein said phosphodiester backbone is cleaved between said third nucleic acid segment and said fourth nucleic acid segment;

b) tagging said first nucleic acid segment with a first tag and said second nucleic acid segment with a second tag, such that said first nucleic acid segment and said second nucleic acid segment are identifiable as arising from a common nucleic acid of the diverse plurality of nucleic acids, and tagging said third nucleic acid segment with a third tag and said fourth nucleic acid segment with a fourth tag, such that said third nucleic acid segment and said fourth nucleic acid segment are identifiable as arising from a common nucleic acid of the diverse plurality of nucleic acids;

c) sequencing at least an identifiable portion of said first nucleic acid segment and said first tag, of said second nucleic acid segment and said second tag, of said third nucleic acid segment and said third tag, and of said fourth nucleic acid segment and said fourth tag;

d) constructing at least a first sequence scaffold comprising said first nucleic acid segment and said second nucleic acid segment and a second sequence scaffold comprising said third nucleic acid segment and said fourth nucleic acid segment; such that a plurality of segments of said diverse plurality of nucleic acids are assigned to at least one of the first or second sequence scaffold; and

e) counting a plurality of sequence scaffolds constructed, wherein nucleic acid segments tagged such that they are identifiable as arising from a common nucleic acid of the diverse plurality of nucleic acids are assigned to a common scaffold; and

wherein the number of scaffolds generated indicates the species diversity in the heterogeneous sample.

2. The method of claim 1, wherein tagging said first nucleic acid segment and said second nucleic acid segment comprises adding a first oligo to the first nucleic acid segment and adding a second oligo to the second segment, said first oligo and said second oligo sharing a common oligo sequence.

3. The method of claim 2, wherein nucleic acid segments having said common oligo sequence are assigned to a common scaffold.

4. The method of claim 3, comprising mapping said identifiable portion of said first nucleic acid segment to a contig dataset, and including any matching contig of said contig dataset into said common scaffold.

5. The method of claim 4, wherein the contig dataset is concurrently generated.

6. The method of claim 4, wherein the contig dataset is obtained from a database.

7. The method of claim 1, wherein tagging said first nucleic acid segment and said second nucleic acid segment comprises ligating said first nucleic acid segment to said second nucleic acid segment, and wherein said first nucleic acid segment and said second nucleic acid segment are assigned to a common scaffold.

8. The method of claim 7, comprising mapping said identifiable portion of said first nucleic acid segment to a contig dataset, and including any matching contig of said contig dataset into said common scaffold.

9. The method of claim 8, wherein the contig dataset is concurrently generated.

10. The method of claim 8, wherein the contig dataset is obtained from a database.

11. The method of claim 1, wherein said phosphodiester backbone is cleaved subsequent to said obtaining said stabilized nucleic acid sample.

12. The method of claim 1, wherein said stabilized nucleic acid sample is contacted to a crosslinking agent.

13. The method of claim 1, wherein said stabilized nucleic acid sample is an FFPE sample.

14. The method of claim 1, comprising contacting said heterogeneous sample to a reverse transcriptase.

15. The method of claim 1, comprising categorizing said common scaffold as corresponding to a sample condition when a plurality of samples correlating to said condition have said common scaffold and if a plurality of samples lacking said condition lack said sample.

16. The method of claim 1, wherein sequence reads from the nucleic acid segments assemble into at least two nucleic acid scaffolds without reference to exogenous sequence information.

17. The method of claim 1, wherein sequence reads from the nucleic acid segments assemble into at least two nucleic acid scaffolds, such that at least 50% of a first genome and at least 50% of a second genome are represented in said at least two nucleic acid scaffolds.

18. The method of claim 1, wherein the method comprises using SPRI beads.

19. The method of claim 1, wherein the stabilized nucleic acid sample comprises no greater than about 5 micrograms of DNA.

20. The method of claim 1, wherein said first tag and said second tag are the same.

21. The method of claim 1, wherein said third tag and said fourth tag are the same.

22. The method of claim 1, wherein said first tag comprises a covalently-linked DNA segment from said second nucleic acid segment and said second tag comprises a covalently-linked DNA segment from said first nucleic acid segment.

23. The method of claim 1, wherein said third tag comprises a covalently-linked DNA segment from said fourth nucleic acid segment and said fourth tag comprises a covalently-linked DNA segment from said third nucleic acid segment.