METHODS AND COMPOSITIONS FOR PROXIMITY LIGATION

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 62/867,463 filed June 27, 2019, U.S. Provisional Patent Application No. 62/931,069 filed November 5, 2019, U.S. Provisional Patent Application No. 63/011,490 filed April 17, 2020, U.S. Provisional Patent Application No. 62/870,297 filed July 3, 2019, and U.S. Provisional Patent Application No. 63/014,422 filed April 23, 2020, each of which is incorporated herein by reference in its entirety.

BACKGROUND

[0002] Obtaining high-quality, contiguous genome sequences is often difficult, especially in cases when limited source material is available for sequence analysis. While obtaining raw sequence data has become faster and available at a lower cost, suitable methods for analyzing and assembling the data efficiently and accurately remains a challenge.

SUMMARY

[0002a] In one aspect, provided herein is a method comprising: (a) obtaining a stabilized biological sample comprising a nucleic acid molecule complexed to at least one nucleic acid binding protein; (b) contacting the stabilized biological sample to a non-specific endonuclease to cleave the nucleic acid molecule into a plurality of segments; (c) attaching a first segment and a second segment of the plurality of segments at a junction; and (d) subjecting the plurality of segments to size selection to obtain a plurality of selected segments wherein attaching comprises contacting at least the first segment and the second segment to at least one bridge oligonucleotide.

[0003] In one example, there are provided methods of nucleic acid analysis. In some cases, methods may comprise: (a) obtaining a stabilized biological sample comprising a nucleic acid molecule complexed to at least one nucleic acid binding protein; (b) contacting the stabilized biological sample to a non-specific endonuclease to cleave the nucleic acid molecule into a plurality of segments; (c) attaching a first segment and a second segment of the plurality of segments at a junction; and (d) subjecting the plurality of segments to size selection to obtain a plurality of selected segments. In some cases, the plurality of selected segments is about 145 to about 600 bp. In some cases, the plurality of selected segments is about 100 to about 2500 bp. In some cases, the plurality of selected segments is about 100 to about 600 bp. In some cases, the plurality of selected segments is about 600 to about 2500 bp. In some cases, the method further comprises, prior to step (d), preparing a sequencing library from the plurality of segments. In some cases, the method further comprises subjecting the sequencing library to a size selection to obtain a size-selected library. In some cases, the size-selected library is between about 350 bp and 1000 bp in size. In some cases, the size selection is conducted with gel electrophoresis, capillary electrophoresis, size selection beads, or a gel filtration column. In some cases, the method further comprises analyzing the plurality of selected segments to obtain a QC value. In some cases, the QC value is a chromatin digest efficiency (CDE) based on the proportion of segments between 100 and 2500 bp in size prior to step (d). In some cases, the method further comprises selecting a sample for further analysis

when the CDE value is at least 65%. In some cases, the QC value is a chromatin digest index (CDI) based on the ratio of a number of mononucleosome-sized segments to a number of dinucleosome-sized segments prior to step (d). In some cases, the method further comprises selecting a sample for further analysis when the CDI value is greater than -1.5 and less than 1. In some cases, the method further comprises subsequent to the contacting the stabilized biological sample to a non-specific endonuclease, binding the plurality of segments to one or more surfaces. In some cases, the one or more surfaces comprise one or more beads. In some cases, the one or more beads are solid phase reversible immobilization (SPRI) beads. In some

barcode. In some cases, the method does not comprise a shearing step. In some cases, the method further comprises: (e) obtaining at least some sequence on each side of the junction to generate a first read pair. In some cases, the method further comprises (f) mapping the first read pair to a set of contigs; and (g) determining a path through the set of contigs that represents an order and/or orientation to a genome. Alternatively or in combination, the method further comprises (f) mapping the first read pair to a set of contigs; and (g) determining, from the set of contigs, a presence of a structural variant or loss of heterozygosity in the stabilized biological sample. Alternatively or in combination, the method further comprises (f) mapping the first read pair to a set of contigs; and (g) assigning a variant in the set of contigs to a phase. In some cases, the variant is a human leukocyte antigen (HLA) variant. In some cases, the variant is a killer-cell immunoglobulin-like receptor (KIR) variant. Alternatively or in combination, the method further comprises (f) mapping the first read pair to a set of contigs; (g) determining, from the set of contigs, a presence of a variant in the set of contigs; and (h) conducting a step selected from one or more of: (1) identifying a disease stage, a prognosis, or a course of treatment for the stabilized biological sample; (2) selecting a drug based on the presence of the variant; or (3) identifying a drug efficacy for the stabilized biological sample. In some cases, the DNase is coupled or fused to an immunoglobulin binding protein or a fragment thereof. In some cases, the DNase is coupled to two or more immunoglobulin binding proteins or fragments thereof. In some cases, the immunoglobulin binding protein is selected from a Protein A, a Protein G, a Protein A/G, and a Protein L.

[0004] In another example, there are provided methods comprising: (a) obtaining a stabilized biological sample comprising a nucleic acid molecule complexed to at least one nucleic acid binding protein; (b) contacting the stabilized biological sample to a micrococcal nuclease (MNase) to cleave the nucleic acid molecule into a plurality of segments; and (c) attaching a first segment and a second segment of the plurality of segments at a junction. In some cases, the method further comprises (d) subjecting the plurality of segments to size selection to obtain a plurality of selected segments. In some cases, the plurality of selected segments is about 145 to about 600 bp. In some cases, the plurality of selected segments is about 100 to about 2500 bp. In some cases, the plurality of selected segments is about 100 to about 600 bp. In some cases, the plurality of selected segments is about 600 to about 2500 bp. In some cases, the method further comprises prior to step (d), preparing a sequencing library from the plurality of segments. In some cases, the method further comprises subjecting the sequencing library to a size selection to obtain a size-selected library. In some cases, the size-selected library is between about 350 bp and 1000 bp in size. In some cases, the size selection is conducted with gel electrophoresis, capillary electrophoresis, size selection beads, or a gel filtration column. In some cases, the method further comprises analyzing the plurality of selected segments to obtain a QC value. In some cases, the QC value is a chromatin digest efficiency (CDE) based on the proportion of segments between 100 and 2500 bp in size prior to step (d). In some cases, the method further comprises selecting a sample for further analysis when the CDE value is at least 65%. In some cases, the QC value is a chromatin digest index (CDI) based on the ratio of a number of mononucleosome-sized segments to a number of dinucleosome-sized segments prior to step (d). In some cases, the method further comprises selecting a sample for further analysis when

bridge oligonucleotide is coupled to two or more immunoglobulin binding proteins or fragments thereof. In some cases, the immunoglobulin binding protein is selected from a Protein A, a Protein G, a Protein A/G, and a Protein L. In some cases, attaching comprises contacting at least the first segment and the second segment to a barcode. In some cases, the method does not comprise a shearing step. In some cases, the method further comprises (e) obtaining at least some sequence on each side of the junction to generate a first read pair. In some cases, the method further comprises (f) mapping the first read pair to a set of contigs; and (g) determining a path through the set of contigs that represents an order and/or orientation to a genome. Alternatively or in combination, the method further comprises (f) mapping the first read pair to a set of contigs; and (g) determining, from the set of contigs, a presence of a structural variant or loss of heterozygosity in the stabilized biological sample. Alternatively or in combination, the method further comprises (f) mapping the first read pair to a set of contigs; and (g) assigning a variant in the set of contigs to a phase. In some cases, the variant is a human leukocyte antigen (HLA) variant. In some cases, the variant is a killer-cell immunoglobulin-like receptor (KIR) variant. Alternatively or in combination, the method further comprises (f) mapping the first read pair to a set of contigs; (g) determining, from the set of contigs, a presence of a variant in the set of contigs; and (h) conducting a step selected from one or more of: (1) identifying a disease stage, a prognosis, or a course of treatment for the stabilized biological sample; (2) selecting a drug based on the presence of the variant; or (3) identifying a drug efficacy for the stabilized biological sample. In some cases, the MNase is coupled or fused to an immunoglobulin binding protein. In some cases, the MNase is coupled or fused to two or more immunoglobulin binding proteins or fragments thereof. In some cases, the immunoglobulin binding protein is selected from a Protein A, a Protein G, a Protein A/G, and a Protein L.

[0005] In additional examples, there are provided nucleic acid libraries comprising: (a) a first cell genome library component comprising a plurality of first cell genome fragment pairs, wherein at least one of said first cell genome fragment pairs comprises two first cell genome segments tethered via a nucleic acid segment comprising a first cell genome library-indicative tag; and (b) a second cell genome library component comprising a plurality of second cell genome fragment pairs, wherein at least one of said second cell genome fragment pairs comprises two second cell genome segments tethered via a nucleic acid segment comprising a second cell genome library-indicative tag. In some cases, the two first cell genome segments tethered via a nucleic acid segment comprising a first cell genome library-indicative tag indicate a first cell genome configuration in a first cell. In some cases, the two second cell genome segments tethered via a nucleic acid segment comprising a second cell genome library-indicative tag indicate a second cell genome configuration in a second cell, wherein the second cell genome configuration is different from the first cell genome configuration. In some cases, the first cell genome library component is obtained from an isolated eukaryotic nucleus. In some cases, the plurality of first cell genome fragment pairs are bookended by recombinase sites. In some cases, the recombinase site is an integrase integration site. In some cases, the recombinase site is a transposase mosaic end. In some cases, at least one recombinase site of the recombinase sites comprises an exonuclease resistant moiety. In some cases, the exonuclease resistant moiety comprises a phosphorothioate. In some cases, the nucleic

acid segment comprising a first cell genome library-indicative tag further comprises a recombinase left border and a recombinase right border. In some cases, the recombinase is an integrase. In some cases, the recombinase is a transposase. In some cases, the nucleic acid segment comprising a first cell genome library-indicative tag comprises an affinity tag. In some cases, the affinity tag comprises biotin. In some cases, at least some library members are clonal copies. In some cases, co-occurrence of read pairs that map to comparable regions of a nucleic acid reference indicate distance of the regions to one another in the cell. In some cases, the distance is a relative distance.

[0006] In additional examples, there are provided systems comprising: a plurality of cell genome aliquots, wherein at least some of the cell genome aliquots comprise genome binding moieties that preserve genome component positional information; and a plurality of recombinase nucleic acid aliquots, wherein at least some of the recombinase nucleic acid aliquots comprise distinguishable sequence relative to at least one other aliquot. In some cases, the recombinase is an integrase. In some cases, the recombinase is a transposase. In some cases, at least some of the cell genome aliquots comprise fragmented genome molecules. In some cases, at least some of the fragmented genome molecules comprise integration site ends. In some cases, the cell genome aliquots comprise eukaryotic cell genome aliquots. In some cases, the genome binding moieties that preserve genome component positional information comprise chromatin constituents. In some cases, the genome binding moieties that preserve genome component positional information comprise nucleosomes. In some cases, the plurality of cell genome aliquots comprise integrase enzymes. In some cases, the plurality of recombinase nucleic acid aliquots comprise integrase nucleic acid molecules having integrase integration sites. In some cases, the plurality of cell genome aliquots comprise transposase enzymes. In some cases, the plurality of recombinase nucleic acid aliquots comprise transposase nucleic acid molecules having transposase mosaic ends. In some cases, at least one integration site of the integration sites comprises an exonuclease resistant moiety. In some cases, at least one mosaic end of the mosaic ends comprises an exonuclease resistant moiety. In some cases, the exonuclease resistant moiety comprises a phosphorothioate. In some cases, at least some of the recombinase nucleic acid molecules comprise a recombinase left border and an recombinase right border. In some cases, the nucleic acid segment comprising a first cell genome library-indicative tag comprises an affinity tag. In some cases, the affinity tag comprises biotin. In some cases, the distinguishable sequence relative to at least one other aliquot comprises a plurality of identical nucleic acid sequences in an aliquot. In some cases, a first aliquot comprises a plurality of nucleic acid molecules having a common distinguishable sequence relative to at least one other aliquot. In some cases, the plurality of cell genome aliquots comprise single-cell genome aliquots.

[0007] In additional examples, there are provided methods of assaying for a chromosomal conformation variation between at least two cells. In some cases, methods comprise obtaining genomic nucleic acids from the two cells wherein chromosomal conformation variation is preserved; introducing internal breaks into genomic nucleic acids from the two cells; and linking two exposed ends adjacent to the internal breaks to one another via one of a plurality of tagged segments, wherein tags of a first cell are distinguishable from tags of a second cell. In some cases, the genomic nucleic acids from the two cells

are isolated prior to said linking. In some cases, amplifying nucleic acid molecules resultant from said linking to produce amplicons comprising chromosomal break junction ends linked by an internal segment comprising distinguishable sequence such that chromosomal break junctions from the two cells are distinguishable. In some cases, isolating fragments from a first cell of the two cells comprising two formerly exposed ends linked by a first tagged segment and fragments from a second cell of the at least two cells comprising two formerly exposed ends linked by a second tagged segment. In some cases, obtained paired end sequence information comprising at least some first tag information and at least some second tag information. In some cases, assigning paired ends to a common proximity in a cell. In some cases, assigning paired ends to a common proximity comprises assessing a number of occurrences of paired ends mapping to two common clusters, and correlating the relative proximity to the number of occurrences. In some cases, said fragments from a first cell and said fragments from a second cell are isolated in a common volume. In some cases, said fragments from a first cell and said fragments from a second cell are sequenced in a common volume. In some cases, linking two exposed ends adjacent to the internal breaks to one another comprises linking two exposed ends that were not immediately adjacent prior to said introducing internal breaks. In some cases, linking two exposed ends adjacent to the internal breaks to one another comprises linking two exposed ends that were remote to one another on a common nucleic acid molecule prior to said introducing internal breaks. In some cases, linking two exposed ends adjacent to the internal breaks to one another comprises linking two exposed ends that were in physical proximity to one another prior to said introducing internal breaks. In some cases, the at least two cells comprise at least two cell populations.

[0008] In further examples, there are provided methods comprising obtaining a stabilized sample comprising a nucleic acid molecule complexed to at least one nucleic acid binding protein; cleaving the nucleic acid molecule into a plurality of segments comprising at least a first segment and a second segment; attaching adapters comprising first recombinase sites to the first segment and to the second segment; and contacting the first segment and the second segment with a linker comprising second recombinase sites in the presence of a recombinase, thereby generating a linked nucleic acid comprising a first sequence from the first segment, a linker sequence from the linker, and a second sequence from the second segment. In some cases, the recombinase is an integrase. In some cases, the recombinase is a transposase. In some cases, the method further comprises sequencing at least a portion of the linked nucleic acid. In some cases, the sequencing comprises sequencing at least a portion of the first sequence and at least a portion of the second sequence. In some cases, the method further comprises mapping at least a portion of the first sequence and at least a portion of the second sequence to a genome. In some cases, the method further comprises conducting three-dimensional genomic analysis using information from the sequencing. In some cases, the stabilized sample is a cross-linked sample. In some cases, obtaining the stabilized sample comprises obtaining a sample and stabilizing the sample. In some cases, obtaining the stabilized sample comprises obtaining a sample that was previously stabilized. In some cases, the nucleic acid binding protein comprises chromatin or a constituent thereof. In some cases, cleaving comprises enzymatic digestion. In some cases, the enzymatic digestion comprises digestion with

one or more restriction enzymes. In some cases, the enzymatic digestion comprises digestion with one or more non-specific nucleases. In some cases, the one or more non-specific nucleases comprise DNase or MNase. In some cases, the attaching the first recombinase sites comprises ligation. In some cases, the first recombinase sites and the second recombinase sites comprise integrase sites attP and attB. In some cases, the adapters further comprise sequencing adapter regions. In some cases, the sequencing adapter regions comprise Y adapters. In some cases, the sequencing adapter regions comprise P5 and/or P7 adapters. In some cases, the first recombinase sites and the second recombinase sites comprise transposase mosaic ends. In some cases, the linker sequence comprises an affinity tag. In some cases, the affinity tag is biotin. In some cases, the linker sequence comprises a barcode sequence. In some cases, the barcode sequence is indicative of a partition of origin. In some cases, the barcode sequence is indicative of a cell of origin. In some cases, the barcode sequence is indicative of a cell population of origin. In some cases, the barcode sequence is indicative of an organism of origin. In some cases, the barcode sequence is indicative of a species of origin.

[0009] In further examples there are provided methods comprising: obtaining a stabilized sample comprising a nucleic acid molecule complexed to at least one nucleic acid binding protein; cleaving the nucleic acid molecule into a plurality of segments comprising at least a first segment and a second segment; attaching the first segment to the second segment, thereby creating proximity ligated segments; recovering the proximity ligated segments; and sequencing at least a portion of the proximity ligated segments, wherein sequencing adapters are not attached to the proximity ligated segments after the recovering. In some cases, attaching is conducted by ligating the first segment to the second segment. In some cases, attaching is conducted using recombinase. In some cases, attaching is conducted via a linker. In some cases, the linker comprises an affinity tag. In some cases, the affinity tag is biotin. In some cases, the method further comprises prior to the attaching in step (c), attaching recombination adapters comprising recombinase sites to the first segment and the second segment. In some cases, the recombination adapters comprise sequencing adapters. In some cases, the sequencing adapters comprise Y adapters. In some cases, the sequencing adapters comprise P5 and/or P7 adapters.

[0010] Provided herein are methods comprising: (a) obtaining a stabilized biological sample comprising a nucleic acid molecule complexed to at least one nucleic acid binding protein; (b) contacting the stabilized biological sample to a DNase to cleave the nucleic acid molecule into a plurality of segments; (c) attaching a first segment and a second segment of the plurality of segments at a junction; and (d) subjecting the plurality of segments to size selection to obtain a plurality of selected segments. In some cases, the plurality of selected segments is about 145 to about 600 bp. In some cases, the plurality of selected segments is about 100 to about 2500 bp. In some cases, the plurality of selected segments is about 100 to about 600 bp. In some cases, the plurality of selected segments is about 600 to about 2500 bp. In some cases, the method further comprises, prior to step (d), preparing a sequencing library from the plurality of segments. In some cases, the method further comprises subjecting the sequencing library to a size selection to obtain a size-selected library. In some cases, the size-selected library is between about 350 bp and 1000 bp in size. In some cases, the size selection is conducted with gel electrophoresis,

a presence of a structural variant or loss of heterozygosity in the stabilized biological sample. In some cases, the method further comprises: (f) mapping the first read pair to a set of contigs; and (g) assigning a variant in the set of contigs to a phase. In some cases, the variant is a human leukocyte antigen (HLA) variant. In some cases, the variant is a killer-cell immunoglobulin-like receptor (KIR) variant. In some cases, the method further comprises: (f) mapping the first read pair to a set of contigs; (g) determining, from the set of contigs, a presence of a variant in the set of contigs; and (h) conducting a step selected from one or more of: (1) identifying a disease stage, a prognosis, or a course of treatment for the stabilized biological sample; (2) selecting a drug based on the presence of the variant; or (3) identifying a drug efficacy for the stabilized biological sample.

[0011] In additional examples, there are provided methods comprising: (a) obtaining a stabilized biological sample comprising a nucleic acid molecule complexed to at least one nucleic acid binding protein; (b) contacting the stabilized biological sample to a micrococcal nuclease (MNase) to cleave the nucleic acid molecule into a plurality of segments; and (c) attaching a first segment and a second segment of the plurality of segments at a junction. In some cases, methods herein further comprise (d) subjecting the plurality of segments to size selection to obtain a plurality of selected segments. In some cases, the plurality of selected segments is about 145 to about 600 bp. In some cases, the plurality of selected segments is about 100 to about 2500 bp. In some cases, the plurality of selected segments is about 100 to about 600 bp. In some cases, the plurality of selected segments is about 600 to about 2500 bp. In some cases, methods herein further comprise, prior to step (d), preparing a sequencing library from the plurality of segments. In some cases, methods herein further comprise subjecting the sequencing library to a size selection to obtain a size-selected library. In some cases, the size-selected library is between about 350 bp and 1000 bp in size. In some cases, the size selection is conducted with gel electrophoresis, capillary electrophoresis, size selection beads, or a gel filtration column. In some cases, the method further comprises analyzing the plurality of selected segments to obtain a QC value. In some cases, the QC value is a chromatin digest efficiency (CDE) based on the proportion of segments between 100 and 2500 bp in size prior to step (d). In some cases, the method further comprises selecting a sample for further analysis when the CDE value is at least 65%. In some cases, the QC value is a chromatin digest index (CDI) based on the ratio of a number of mononucleosome-sized segments to a number of dinucleosome-sized segments prior to step (d). In some cases, the method further comprises selecting a sample for further analysis when the CDI value is greater than -1.5 and less than 1. In some cases, the stabilized biological sample comprises a stabilized cell lysate. In some cases, the stabilized biological sample comprises a stabilized intact cell. In some cases, the stabilized biological sample comprises a stabilized intact nucleus. In some cases, step (b) is conducted prior to lysis of the intact cell or the intact nucleus. In some cases, methods herein further comprise, prior to step (c), lysing cells and/or nuclei in the stabilized biological sample. In some cases, the stabilized biological sample comprises fewer than 3,000,000 cells. In some cases, the stabilized biological sample comprises fewer than 1,000,000 cells. In some cases, the stabilized biological sample comprises fewer than 100,000 cells. In some cases, the stabilized biological sample comprises less than 10 µg DNA. In some cases, the stabilized biological sample comprises less than 1 µg DNA. In some

cases, the stabilized biological sample is further treated with a DNase. In some cases, the DNase is DNase I. In some cases, the DNase is DNase II. In some cases, the DNase is selected from one or more of DNase I and DNase II. In some cases, the stabilized biological sample has been treated with a crosslinking agent. In some cases, the crosslinking agent is a chemical fixative. In some cases, the chemical fixative comprises formaldehyde. In some cases, the chemical fixative comprises psoralen. In some cases, the chemical fixative comprises disuccinimidyl glutarate (DSG). In some cases, the chemical fixative comprises ethylene glycol bis(succinimidyl succinate) (EGS). In some cases, the crosslinking agent is ultraviolet light. In some cases, the stabilized biological sample is a crosslinked paraffin-embedded tissue sample. In some cases, methods herein comprise contacting the plurality of selected segments to an antibody. In some cases, attaching comprises filling in sticky ends using biotin tagged nucleotides and ligating the blunt ends. In some cases, attaching comprises contacting at least the first segment and the second segment to at least one bridge oligonucleotide. In some cases, the bridge oligonucleotide comprises a barcode sequence. In some cases, attaching comprises contacting at least the first segment and the second segment to multiple bridge oligonucleotides in series. In some cases, the attaching results in cells, nuclei, chromosomes, or nucleic acid molecules of the stabilized biological sample receiving a unique sequence of bridge oligonucleotides. In some cases, attaching comprises contacting at least the first segment and the second segment to a barcode. In some cases, the method does not comprise a shearing step. In some cases, methods herein further comprise: (e) obtaining at least some sequence on each side of the junction to generate a first read pair. In some cases, methods herein further comprise: (f) mapping the first read pair to a set of contigs; and (g) determining a path through the set of contigs that represents an order and/or orientation to a genome. In some cases, methods herein further comprise: (f) mapping the first read pair to a set of contigs; and (g) determining, from the set of contigs, a presence of a structural variant or loss of heterozygosity in the stabilized biological sample. In some cases, methods herein further comprise: (f) mapping the first read pair to a set of contigs; and (g) assigning a variant in the set of contigs to a phase. In some cases, the variant is a human leukocyte antigen (HLA) variant. In some cases, the variant is a killer-cell immunoglobulin-like receptor (KIR) variant. In some cases, methods herein further comprise: (f) mapping the first read pair to a set of contigs; (g) determining, from the set of contigs, a presence of a variant in the set of contigs; and (h) conducting a step selected from one or more of: (1) identifying a disease stage, a prognosis, or a course of treatment for the stabilized biological sample; (2) selecting a drug based on the presence of the variant; or (3) identifying a drug efficacy for the stabilized biological sample.

INCORPORATION BY REFERENCE

[0012] 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.