



(51) International Patent Classification:

C12Q 1/6869 (2018.01)

C12Q 1/6876 (2018.01)

C12Q 1/6874 (2018.01)

Published:

— with international search report (Art. 21(3))

(21) International Application Number:

PCT/US2020/047067

(22) International Filing Date:

19 August 2020 (19.08.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/888,714

19 August 2019 (19.08.2019)

US

(71) Applicant: **UNIVERSAL SEQUENCING TECHNOLOGY CORPORATION** [US/US]; Suite 800, 780 Dedham Street, Canton, Massachusetts 02021 (US).

(72) Inventors: **CHEN, Zhoutao**; 6728 Whitesail Street, Carlsbad, California 92011 (US). **WU, Tsai-Chin**; 741 Avenida Leon, San Marcos, California 92069 (US). **PHAM, Long Kim**; 7780 Stylus Drive, San Diego, California 92108 (US). **WANG, Yong**; 5 Glenview Road, Sharon, Massachusetts 02067 (US). **LEI, Ming**; 6 Marie Ave., Sharon, Massachusetts 02067 (US).

(74) Agent: **WU, George T.**; Culhane Meadows, PLLC, 2591 Dallas Parkway Suite 300, Frisco, Texas 75034 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: METHODS AND COMPOSITIONS FOR TRACKING NUCLEIC ACID FRAGMENT ORIGIN FOR NUCLEIC ACID SEQUENCING

(57) Abstract: The present disclosure provides methods and compositions for tracking nucleic acid fragment origin by target-specific barcode tagging when original nucleic acid targets break into small fragments. Nucleic acid targets are captured in vitro on a solid support with clonally localized nucleic acid barcode templates. Many nucleic acid targets can be processed simultaneously in a massively parallel fashion without partition. These nucleic acid target tracking methods can be used for a variety of applications in both whole genome sequencing and targeted sequencing in order to accurately identify genomic variants, haplotype phasing and assembly, for example.

METHODS AND COMPOSITIONS FOR TRACKING NUCLEIC ACID FRAGMENT ORIGIN FOR NUCLEIC ACID SEQUENCING

FIELD

5 **[0001]** The present invention relates in general to methods and compositions for nucleic acid sequencing. In particular, the methods and compositions provided herein are related to preparation of nucleic acid library and generation of sequencing data therefrom.

10 BACKGROUND

[0002] Nucleic acid sequencing can provide information for a wide variety of biomedical applications, including diagnostics, prognostics, pharmacogenomics, and forensic biology. Sequencing may involve basic low throughput methods including Maxam-Gilbert sequencing (chemically modified nucleotide) and Sanger sequencing (chain-termination) methods, or high throughput next-generation methods including massively parallel pyrosequencing, sequencing by synthesis, sequencing by ligation, semiconductor sequencing, and others. For most sequencing methods, a sample, such as a nucleic acid target, needs to be processed into a sequencing library prior to be sequenced on a sequencing instrument. For example, a sample may be fragmented, amplified or
15 attached to an identifier. Unique identifiers are often used to identify the origin of a particular sample.

[0003] Most commercially available sequencing technologies have limited sequencing read length. Second generation sequencing technologies particularly can sequence only several hundred bases and hardly reach a thousand bases. However, nucleic acid sequences of a gene can span from several kilobases to tens and hundreds of
25 kilobases, which means sequencing read length of tens of kilobases is necessary to successfully determine the haplotypes of all genes.

[0004] To overcome the short sequencing read length problem, many methods have been developed to target-specifically label long nucleic acid targets when they are broken into small fragments for sequencing library preparation. Such methods include
30 Complete Genomics's long fragment read, Illumina's synthetic long read, 10x Genomics's linked-read (Zheng et al, 2016), Illumina's single tube method (Zhang et al, 2017) and our own single tube method (WO2017/151828). These target-specific labels are short nucleic acid sequences, called barcodes. The origin of these short nucleic acid fragments can be identified based on their unique, associated barcode. The broader
35 diversity of a barcode population used in the method, the better specificity it provides for identification. 10x Genomics's linked-read method has been used widely. However, it requires a water-in-oil emulsion method to keep the clonality of target-specific barcode labelling reaction. This requirement significantly increases the complexity of its sample

preparation procedure and cost. Several methods, including the method belonging to Complete Genomics (US Patent 9328382), Illumina's single tube method (Zhang et al, 2017) and our method (WO2017/151828) use transposase-based system and remove the need of partition of nucleic acid targets with emulsion droplets in the reaction. These methods enable target-specific barcoding in a single tube reaction format in principle. The present invention provides novel transposase based single tube barcoding methods with significant improvement on reaction efficiency and simplicity of the workflow.

SUMMARY

[0005] In one aspect, described herein are methods of tracking nucleic acid fragment origin by barcode tagging. The methods include providing a plurality of solid support having clonal barcode templates or semi-clonal barcode templates immobilized thereon, wherein the barcode templates comprise at least two different barcode sequences, a majority barcode sequence and a minority barcode sequence; the majority barcode sequence on a barcoded solid support is substantially different from the majority barcode sequence of other barcoded solid supports, and a minority barcode sequence is the same as the minority barcode sequence of other barcoded solid supports, and providing a plurality of transpososomes, each transpososome comprising transposable DNA and transposase, wherein at least one transposable DNA in the transpososome is capable of being captured by the barcode template on the solid support directly or indirectly by hybridization or an affinity moiety. A nucleic acid target contacts the solid support, and the transpososomes in one reaction vessel to attach the barcode information on the solid support to the nucleic acid target by simultaneous strand transfer and capture reactions. The aforementioned reaction occurs at substantially the same time without additional partition of each nucleic acid target from another nucleic acid target within the total plurality of nucleic acid targets. The nucleic acid target is broken into fragments by breaking strand transfer complexes, wherein at least one fragment is attached to a barcode template on the solid support.

[0006] In some instances, the clonal barcode templates or semi-clonal barcode templates immobilized on the solid support are produced by methods of direct synthesis, clonal amplification, or a combination thereof. The clonal amplification can be emulsion PCR, bridge PCR, isothermal PCR, template walking, nanoball generation, and a combination thereof. In particular instances, the barcoded solid support is prepared by a clonal amplification method without separating amplified and unamplified populations. Further instances, the barcoded solid support is prepared by a clonal amplification with only or predominantly enriched amplified populations.

[0007] In one aspect, the reactions are in a buffer system with a controlled viscosity to decrease diffusion and increase suspension by adding substance selected from the

group of polyethylene glycol, pluronic, cellulose, agarose, and their derivatives, and other polymers, and a combination thereof, preferably with a viscosity at about 2-200 mPa·s.

5 [0008] In one aspect, the transpososomes can be replaced with individual transposable DNAs and transposases without pre-assembling them into transpososomes.

[0009] In another aspect, a second transpososome is added after the initial reaction in the reaction vessel; wherein the previously added transpososome is referred to as the first transpososome, and the first and second transpososomes can be of the same type or different type, or different transposon sequences of the same type. In another aspect, 10 a second transpososome is added after breaking the nucleic acid target into fragments.

[00010] In another aspect, the nucleic acid target can be pre-attached to the solid support by non-specific binding.

15 [00011] In another aspect, the capture reaction in the vessel is by ligation, or by hybridization, or by affinity tag, or by antibody and antigen reaction, or by click chemistry, or a combination thereof.

[00012] In another aspect, the capture reaction comprises first hybridization and then ligation.

20 [00013] In another aspect, the barcoded solid support has transposable DNAs or transpososomes pre-attach to the end of some barcode templates immobilized on the solid support.

[00014] In another aspect, said transposase is selected from the group consisting of Tn, Mu, Ty, and Tc transposases in a wildtype, a mutant or a tagged version thereof, and a combination thereof. In particular, the transposase is a MuA transposase, or a Tn5 transposase in a wildtype or a mutant or a tagged version thereof, or a combination 25 thereof.

30 [00015] In another aspect, said transposable DNA contains a transposon, wherein the transposon is selected from the group consisting of Tn, Mu, Ty, and Tc transposon DNAs in a wildtype or a mutant version thereof, and a combination thereof. In particular, the transposon is a MuA transposon, or a Tn5 transposon, wild type or mutant, or a combination thereof.

[00016] In another aspect, said transposable DNA further comprises an adaptor sequence.

35 [00017] In another aspect, said transposable DNA capable of being captured by said barcode template has no complementary sequences to said barcode template, and the capture of said transposable DNA to said barcode template is facilitated by a linker.

[00018] In another aspect, said transpososome comprises at least one type of transposase, at least one type of transposable DNA or a combination thereof.

[00019] In one aspect, described herein are methods of tracking nucleic acid fragment origin by barcode tagging. The methods include providing a plurality of solid support having clonal barcode templates or semi-clonal barcode templates immobilized thereon, wherein the barcode templates comprise at least two different barcode sequences, a majority barcode sequence and a minority barcode sequence; the majority barcode sequence on a barcoded solid support is substantially different from the majority barcode sequence of other barcoded solid supports, and a minority barcode sequence is the same as the minority barcode sequence of other barcoded solid supports, and capturing a nucleic acid target to the solid support via non-specific binding, and providing a plurality of transpososomes, each transpososome comprising transposable DNA and transposase, wherein at least one transposable DNA in the transpososome is capable of being specifically captured to the barcode template on the solid support directly or indirectly. Contact the non-specifically bound nucleic acid target on the solid support with the transpososomes in one reaction vessel to attach the barcode information on the solid support to the nucleic acid target by simultaneous strand transfer and capture reactions. The nucleic acid target is broken into fragments by breaking strand transfer complexes, wherein at least one fragment is attached to a barcode template on the solid support.

[00020] In one aspect, described herein are methods of tracking nucleic acid fragment origin by barcode tagging. The methods include providing a plurality of solid support having clonal barcode templates or semi-clonal barcode templates immobilized thereon, wherein the barcode templates comprise at least two different barcode sequences, a majority barcode sequence and a minority barcode sequence; the majority barcode sequence on a barcoded solid support is substantially different from the majority barcode sequence of other barcoded solid supports, and a minority barcode sequence is the same as the minority barcode sequence of other barcoded solid supports, and providing a plurality of transpososomes, each transpososome comprising transposable DNA and transposase, wherein at least one transposable DNA in the transpososome is capable of being captured specifically to the barcode template on the solid support directly or indirectly. A nucleic acid target contacts the transpososomes to form stable strand transfer complexes. The strand transfer complexes are captured to the solid support via non-specific binding. The barcode information on the solid support is attached to the nucleic acid target by a specific capture reaction. The nucleic acid target is broken into fragments by breaking the strand transfer complexes, wherein at least one fragment is attached to a barcode template on the solid support. In some embodiments, the captured nucleic acid target with strand transfer complexes on the solid support is broken into fragments first by breaking the strand transfer complexes. The nucleic acid fragments are kept on the solid support by non-specific binding. The barcode information on the solid support is then attached to the nucleic acid fragments by a specific capture

reaction, wherein at least one fragment is attached to a barcode template on the solid support.

[00021] In one aspect, described herein are methods of tracking nucleic acid fragment origin by barcode tagging. The methods include providing a plurality of solid support
5 having clonal barcode templates or semi-clonal barcode templates immobilized thereon, wherein the barcode templates comprise at least two different barcode sequences, a majority barcode sequence and a minority barcode sequence; the majority barcode sequence on a barcoded solid support is substantially different from the majority barcode sequence of other barcoded solid supports, and a minority barcode sequence is the
10 same as the minority barcode sequence of other barcoded solid supports, and providing a plurality of transpososomes, each transpososome comprising transposable DNA and transposase, wherein at least one transposable DNA in the transpososome is capable of being captured specifically to the barcode template on the solid support directly or indirectly. A nucleic acid target is captured to the solid support via non-specific binding.
15 The transpososomes contact the non-specifically bound nucleic acid target to form stable strand transfer complexes on the solid support. The barcode information on the solid support is attached to the nucleic acid target by a specific capture reaction. The nucleic acid target is broken into fragments by breaking the strand transfer complexes, wherein at least one fragment is attached to a barcode template on the solid support. In
20 some embodiments, the captured nucleic acid target with strand transfer complexes on the solid support is broken into fragments first by breaking said strand transfer complexes. The nucleic acid fragments are kept on the solid support by non-specific binding. The barcode information on the solid support is then attached to the nucleic acid fragments by ligation, wherein at least one fragment is attached to a barcode template
25 on the solid support.

[00022] In one aspect, described herein are methods for determining linkage information of a nucleic acid target. The methods comprise generating barcode tagged fragments of a nucleic acid target according to any one of methods described in this invention, determining the sequence of the nucleic acid fragments and the barcodes, and
30 determining the linkage information of the nucleic acid target based on the barcode sequences when at least two fragments from the same nucleic acid target receiving identical barcode information.

[00023] In one aspect, described herein are methods for generating a soluble library of barcode tagged fragments of a nucleic acid target. In some embodiments, the soluble
35 library comprises sequence information for a whole genome. In some embodiments, the soluble library comprises sequence information for a targeted region. In some embodiments, the soluble library is used for sequencing to determine phasing

information of the nucleic acid target. In some embodiments, the soluble library is used for sequencing to determine the identity of duplicated reads.

BRIEF DESCRIPTION OF THE DRAWINGS

- 5 **[00024]** Fig. 1 illustrates a method of generating clonal barcode tagged nucleic acid fragments with simultaneous strand transfer and ligation reaction onto barcoded solid support in an open bulk reaction without partition of nucleic acid target.
- 10 **[00025]** Fig. 2 shows different transposable DNA designs, (A) transposon complementary strand with 3' over hang in one piece, (B) transposon complementary strand with a separated complementary linker oligo, (C) transposon joining strand with a 5' overhang, (D) transposon with a blunt end at the non-joining end.
- [00026]** Fig. 3 shows examples of different free linker design. (A) single stranded linker, (B) double stranded linker, (C) partially double stranded linker.
- 15 **[00027]** Fig. 4 illustrates a method of generating clonal barcode tagged nucleic acid fragments with simultaneous strand transfer and ligation reaction onto barcoded solid support using different transpososomes simultaneously in an open bulk reaction without partition of nucleic acid target.
- [00028]** Fig. 5 shows removal of single stranded polynucleotides on the solid support using exonuclease I.
- 20 **[00029]** Fig. 6 illustrates a method of generating clonal barcode tagged nucleic acid fragments with simultaneous strand transfer and ligation reaction onto barcoded solid support using different transpososomes sequentially in an open bulk reaction without partition of nucleic acid target.
- 25 **[00030]** Fig. 7 illustrates a method of generating clonal barcode tagged nucleic acid fragments with simultaneous strand transfer and ligation reaction onto barcoded solid support using different transpososomes sequentially in an open bulk reaction without partition of nucleic acid target. The workflow order is different from that shown in Fig. 6.
- 30 **[00031]** Fig. 8 illustrates a method with alternative workflow to generate clonal barcode tagged nucleic acid fragments with simultaneous strand transfer and ligation reaction onto barcoded solid support.
- [00032]** Fig. 9 illustrates a method with alternative workflow to generate clonal barcode tagged nucleic acid fragments with simultaneous strand transfer and ligation reaction onto barcoded solid support.
- 35 **[00033]** Fig. 10 illustrates a method of generating clonal barcode tagged nucleic acid fragments with non-specific binding and ligation onto barcoded solid support.
- [00034]** Fig. 11 illustrates a method with alternative workflow to generate clonal barcode tagged nucleic acid fragments with non-specific binding and ligation onto barcoded solid support.

[00035] Fig. 12 shows a method of introducing adaptor onto immobilized barcode tagged fragments using transpososome.

[00036] Fig. 13 shows a method of introducing adaptor onto immobilized barcode tagged fragments with fragmentation and ligation reaction.

5 [00037] Fig. 14 shows a method of releasing a copy or copies of immobilized barcode tagged fragments by primer extension and/or PCR amplification.

[00038] Fig. 15 is an example of Illumina's sequencing library generated from barcode tagged fragments.

10 [00039] Fig. 16 shows an electropherogram of a barcode tagged Illumina sequencing library ran on a TapeStation (A) and sequencing read count histogram based on read distance to the next alignment for reads with the same barcode (B).

[00040] Fig. 17 shows another electropherogram of a barcode tagged Illumina sequencing library ran on a TapeStation (A) and sequencing read count histogram based on read distance to the next alignment for reads with the same barcode (B).

15 [00041] Fig. 18 illustrates a method of generating clonal barcode tagged nucleic acid fragments with simultaneous strand transfer and hybridization reaction onto barcoded solid support in an open bulk reaction without partition of nucleic acid target.

[00042] Fig. 19 shows three different transposase-based methods to generate clonal barcode tagged fragments for sequencing library construction.

20 [00043] Fig. 20 shows a 2% agarose E-gel EX picture of three amplified sequencing libraries using three different transposase-based methods. M1, Method 1; M2, Method 2; M3, Method 3. The fragment sizes of the 100bp DNA ladder from top to bottom are 3000bp, 2000bp, 1500bp, 1000bp, 900bp, 800bp, 700bp, 600bp, 500bp, 400bp, 300bp, 200bp, and 100bp, respectively.

25 [00044] Fig. 21 shows sequencing read count histograms based on read distance to the next alignment for reads with the same barcode from sequencing data generated from three different preps of barcoded beads.

[00045] Transposases in all the figures are illustrated as a tetramer in the transpososome based on the MuA transposition system. However, other transposases can be also used.

30

DETAILED DESCRIPTION

[00046] As used herein and in the appended claims, a barcode template and a solid support with clonal barcode templates or semi-clonal barcode templates immobilized thereon, i.e. barcoded solid support, are described in patent application

35 WO2017/151828, which is hereby incorporated by reference in its entirety. In some embodiments, all the solid support has barcode templates attached. In some embodiments, only a fraction of solid support has barcode templates attached. The

fraction of solid support with barcodes can be ranged from 1% to 99%. When a solid support is physically separable, such as a bead or a microparticle, barcoded solid support can be prepared by a clonal amplification method with or without enriching amplified solid support from unamplified solid support. The barcode sequences have significant diversity among different barcode templates. There are at least 1000 unique barcode sequences used in a reaction. The more unique barcodes used in the reaction, the higher identification power for detection or tracking.

[00047] The term “adaptor” as used herein refers to a nucleic acid sequence that can comprise a primer binding sequence, a barcode, a linker sequence, a sequence complementary to a linker sequence, a capture sequence, a sequence complementary to a capture sequence, a restriction site, an affinity moiety, unique molecular identifier, and a combination thereof.

[00048] The term “transposase” as used herein refers to a protein that is a component of a functional nucleic acid protein complex capable of transposition and which is mediating transposition, including but not limited to Tn, Mu, Ty, and Tc transposases. The term “transposase” also refers to integrases from retrotransposons or of retroviral origin. It also refers to wild type protein, mutant protein and fusion protein with tag, such as, GST tag, His-tag, etc. and a combination thereof.

[00049] The term “transposon”, as used herein, refers to a nucleic acid segment that is recognized by a transposase or an integrase and is an essential component of a functional nucleic acid-protein complex capable of transposition. Together with transposase they form a transpososome and perform a transposition reaction. It refers to both wild type and mutant transposon.

[00050] A “transposable DNA” as used herein refers to a nucleic acid segment that contains at least one transposon unit. It can also comprise an affinity moiety, un-natural nucleotides and other modifications. The sequences besides the transposon sequence in the transposable DNA can contain adaptor sequences.

[00051] The term “transpososome” as used herein refers to a stable nucleic acid and protein complex formed by a transposase non-covalently bound to a transposon. It can comprise multimeric units of the same or different monomeric unit.

[00052] A “transposition reaction” as used herein refers to a reaction where a transposon inserts into a target nucleic acid. Primary components in a transposition reaction are a transposon, a transposase or an integrase, and its target nucleic acid.

[00053] A “strand transfer reaction” as used herein refers to a reaction between a nucleic acid and a transpososome, in which stable strand transfer complexes form.

[00054] The term “strand transfer complex (STC)” as used herein refers to a nucleic acid-protein complex of transpososome and its target nucleic acid into which transposons insert, wherein the 3' ends of transposon joining strand are covalently connected to the

two strands of its target nucleic acid. It is a very stable form of nucleic acid and protein complex and resists extreme heat and high salt *in vitro* (Burton and Baker, 2003).

[00055] A “transposase binding region” as used herein refers to the nucleotide sequences that are always within the transposon end sequence where a transposase specifically binds when mediating transposition. The transposase binding region may comprise more than one site for binding transposase subunits.

[00056] A “transposon joining strand” as used herein means the strand of a double stranded transposon DNA that is joined by the transposase to the target nucleic acid at the insertion site.

[00057] A “transposon complementary strand” as used herein means the complementary strand of the transposon joining strand in the double stranded transposon DNA.

[00058] A “solid support” as used herein is selected from the group consisting of a bead, a microparticle, a well, a tube, a slide, a plate, a flow cell, and a combination thereof, and wherein when the solid support is physically separable, such as a bead or a microparticle, the barcode template is clonally or semi-clonally immobilized onto the entire surface, and when the solid support is a contiguous flat surface, such as a well, a tube, a slide, a plate or flow cell, the barcode template is immobilized onto the surface as separable clonal clusters or semi-clonal clusters.

[00059] A “ligase” as used herein is selected from the group consisting of DNA ligase, or RNA ligase in a wildtype, a mutant or a tagged version thereof, and a combination thereof; it is used for a ligation reaction.

[00060] A “capture reaction” as used herein means specific capture via ligation, hybridization, affinity binding with an affinity moiety, such as, biotin and streptavidin, antibody and antigen, click chemistry, or a combination of any of these, etc.

[00061] A “reaction vessel” as used herein means a substance with a contiguous open space to hold liquid; it is selected from the group consisting a tube, a well, a plate, a well in a multi-well plate, a slide, a spot on a slide, a droplet, a tubing, a channel, a bottle, a chamber and a flow-cell.

[00062] The methods and materials in this invention are exemplified by employing *in vitro* MuA transposition (Haapa et al, 1999 and Savilahti et al, 1995). Other transposition systems or combination of these different transposition systems can be used, e.g. Ty1 (Devine and Boeke, 1994), Tn7 (Craig, 1996), Tn10 and IS10 (Kleckner et al, 1996), Mariner transposase (Lampe et al, 1996), Tc1 (Vos et al, 1996), Tn5 (Park et al, 1992), P element (Kaufman and Rio, 1992) and Tn3 (Ichikawa and Ohtsubo, 1990), bacterial insertion sequences (Ohtsubo and Sekine, 1996), retroviruses (Varmus and Brown, 1989), and retrotransposon of yeast (Boeke, 1989).

[00063] The present invention relates in general methods and compositions for nucleic acid sequencing. In particular, the methods and compositions provided herein related in preparation of nucleic acid library and generation of sequencing data therefrom.

[00064] In one aspect, the methods and compositions relate to haplotype phasing the target nucleic acid. In some embodiments, the nucleic acid target is DNA. In some embodiments, the nucleic acid target is genomic DNA. In some embodiments, the nucleic acid target is amplified DNA. In some embodiments, the DNA is modified DNA. The modifications include un-natural nucleotide, affinity moiety, chemical treatment (e.g. bisulfite treated or formalin fixed paraffin embedded), and protein attachment (e.g. histone, transcription factor). In some embodiments, the nucleic acid target is synthesized DNA. In some embodiments, the nucleic acid target is RNA. In some embodiments, the nucleic acid target is mRNA. In some embodiments, the nucleic acid target is complementary DNA (cDNA). In some embodiments, the nucleic acid target is first strand cDNA and RNA hybrid. In some embodiments, the nucleic acid target is a DNA and RNA hybrid. In some embodiments, the target nucleic acid is from single cell. In some embodiments, the target nucleic acid is cell free DNA. The length of the nucleic acid target can be varied a lot. It can range from about 50bp to 1Mb, or more. The longer the length of the nucleic acid targets, the better the result for phasing application. The number of nucleic acid targets in a reaction can be from one to billions, or even more. In some embodiments, a reaction vessel is a tube, a well, a plate, a well in a multi-well plate, a slide, a spot on a slide, a droplet, a tubing, a channel, a bottle, a chamber or a flow-cell. The reaction happens in a bulk format without partition of each nucleic acid target from another nucleic acid target within the total plurality of nucleic acid targets. Examples of such as partition are emulsion, wells, droplets, dilution, etc. The present invention dramatically simplifies the workflow and make it easy to scale and automate without the need of partition.

Strand transfer reaction onto nucleic acid targets with simultaneous specific capture of barcode template

[00065] The present invention provides methods and compositions that capture nucleic acid targets by both strand transfer reaction with a transposition system and specific capture reaction, such as, with a ligase and/or hybridization, simultaneously to a clonally barcoded solid support. The captured nucleic acid target can be fragmented by breaking strand transfer complex, which generates small fragments from the nucleic acid target with a target-specific barcode attached (Fig. 1).

[00066] In one embodiment, a transposable DNA may comprise only one transposon sequence. The transposon sequence in the transposable DNA is thus not linked to another transposon sequence by a nucleotide sequence, i.e., the transposable DNA

contains only one transposase binding region (Fig. 2). In addition, the 5' end of joining strand of the transposable DNA has a phosphate, which can ligate to a 3' end of any DNA strand with -OH group through single stranded end to end ligation, double stranded end to end ligation, or via a linker molecule. Fig. 2 shows some examples of ligatable transposable DNA. The 5' end of the transposon joining strand can ligate to the polynucleotide on the solid support with or without the presence of the transposases on the transposable DNA. In some cases, the 3' end of the transposon complementary strand and the 5' end of joining strand of the transposable DNA are in different length. In some cases, the 3' end of the transposon complementary strand and the 5' end of joining strand of the transposable DNA are in the same length. In some cases, the 3' end of the transposon complementary strand is modified with, such as, a dideoxy nucleotide, C3-spacer, a phosphate group, thiophosphate group, an azido group or amino linker to block self-ligation. In some cases, the 3' end of the transposon complementary strand may have a single nucleotide overhang or single nucleotide recess or mismatch nucleotide(s) to block self-ligation. In some cases, a single nucleotide overhang on both double stranded barcode template and a single nucleotide overhang double stranded transposon are complementary and used to facilitate ligation. In some cases, more than one nucleotide overhang on both double stranded barcode template and double stranded transposon may be used to facilitate ligation. In some cases, the barcode templates on the solid support are double stranded. In some cases, the barcode templates on the solid support are single stranded. In some cases, the barcode templates on the solid support are partially single stranded and partially double stranded. In some cases, some sequence variation in the transposon sequence is used as an additional sample identifier. In some cases, transposable DNA comprises an adaptor. In some cases, sequence variation in the adaptor is used as an additional sample identifier. Samples reacted with different transposons and/or transposable DNA sequences can be pooled together after the reaction to simplify downstream process when needed.

[00067] In one embodiment, there are no complementary capture sequences between barcode templates on the solid support and transposable DNA which will be captured. A linker-based capture method may be used to facilitate the capture reaction. Fig.3 shows some examples of linker-based ligation and/or capture. In one embodiment, the linker molecule is single stranded. In one embodiment, the linker molecule is double stranded. In one embodiment, the linker molecule is partially double stranded. In some cases, the linker oligonucleotides may be pre-bound with transposable DNA. In some cases, the linker oligonucleotides may be pre-bound with the immobilized polynucleotides on a solid support. In some cases, the linker oligonucleotides may be added only when capture reaction happens to join the 5' end of joining strand of the transposable DNA to 3' end of the immobilized polynucleotide on a solid support. The free linker method tends to

generate fewer PCR by products than pre-bound linker method. The length of linker molecule can be varied, for example, 5b(p), 10b(p), 20b(p), 30b(p), 40b(p), 50b(p), 100b(p), 200b(p) or more.

[00068] A method for clonally fragmenting and barcoding nucleic acid targets is described as following (Fig. 1). In one reaction vessel, a double stranded nucleic acid target, an assembled transpososome, a ligase and a clonally or semi-clonally barcoded solid support are mixed together without compartmentation by emulsions or dilution. The transposable DNA in the transpososome is ligatable to the barcode template on the solid support. Strand transfer reaction between the transpososomes and the nucleic acid target, and ligation reaction between transposable DNA and barcode template happen simultaneously in the same solution. Stable STCs formed during strand transfer reaction will keep the nucleic acid target in one piece. The barcode sequence will be clonally attached to a nucleic acid target with STCs through ligatable transposable DNA in the STCs. After the reactions, the nucleic acid target is broken into small fragments by breaking STCs with a SDS solution. Many small fragments contain the barcode sequences and the fragments from the same nucleic acid target will have the same barcode. The simultaneous strand transfer reaction and barcode ligation reaction are critical for the high efficiency of this clonal barcode tagging method. The yield of barcode tagged fragments generated with present invention is much higher than that of method in the patent application WO2017/151828, wherein a nucleic acid target forms STCs with transpososomes in the solution first, transposable DNA in the STCs then ligate to a barcode template on a solid support. The reaction efficiency of present invention is also much better than the method in the US patent 9,328,382B2 and the single tube article (Zhang et al, 2017), wherein the transpososomes are immobilized on a barcoded bead first, which then capture free nucleic acid targets in the solution by strand transfer reaction only. In the present invention, both ligation and strand transfer reactions are used to capture the nucleic acid targets. An explanation for low yield from method using ligation only for capture is that a nucleic acid target which is full of STCs may create steric hindrance which limits the ligation efficiency. An explanation for low efficiency from method using strand transfer only for capture is that transposable DNA are immobilized on the bead surface with fixed spatial arrangement and location, which may restrain the efficiency of transpososome formation and/or strand transfer reaction with the free nucleic acid targets. To fully take advantage of the simultaneous reactions of strand transfer and ligation, reaction conditions including buffer composition, pH and temperature are optimized.

[00069] A plurality of nucleic acid targets can be used in one reaction vessel. The reaction happens in a bulk format without partition of each nucleic acid target from another nucleic acid target within the total plurality of nucleic acid targets. The present

invention dramatically simplifies the workflow and make it easy to scale and automate without the need of partition. The plurality of nucleic acid targets is dissolved in a solution homogeneously in order to be uniformly captured on a solid support after the reaction. In some embodiments, limiting the diffusion rate in the reaction solution is used to facility uniformity capture on the solid support. The solid support can be a continuous surface as in a well, a tube, a slide, a plate or a flow cell with isolated clonally or semi-clonally immobilized barcode template clusters. It can also be physically separated as individual bead or microparticle. The bead and microparticle can have sizes ranging from 50nm to 100µm, preferably 1µm to 15µm. Each bead or microparticle has a plurality of barcode templates with unique sequence. The major advantage in the present disclosure is that target specific barcode tagging can occur in an open bulk reaction without partition of nucleic acid targets with wells, microwells, spots, nanochannels, droplets, emulsion droplets, capsules, or dilution, etc. For better results, the bead or microparticle size should be controlled between 50nm to 100µm (diameter), preferably 1µm to 15µm, though it can be smaller than 50nm or larger than 100µm. For uniform reaction, beads or microparticles should keep suspended during reaction by controlling the viscosity of the solution using polyethylene glycol, pluronic, cellulose, agarose, or their derivatives, or other polymers, or a combination thereof, with a final viscosity ranging from 1 to 100 mPa·s at 20°C, most preferably 1.5-30 mPa·s at 20°C. For the barcode clusters on a solid surface, such as a flow cell surface, the cluster size should be controlled between 50nm to 200µm (diameter), preferably 100nm to 10µm. The larger the cluster separation distance is, the smaller the chance of one target nucleic acid molecule being tagged by two or more barcodes.

[00070] Because the very stable nature of STC structure (Surette et al 1987, Mizuuchi et al 1992, Savilahti et al 1995, Burton and Baker 2003, Au et al 2004, Amini et al 2014) and clonal barcode template on a solid support, the barcode tagged fragments generated from this invention keep the identification of their origin nucleic acid target in the barcode sequence. Fragments from the same nucleic acid target share the same barcode sequence. This type of barcode tagged fragments is well known to be used for haplotype phasing, *de novo* assembly and other applications (Zheng et al, 2016, Zhang et al, 2017).

[00071] In one aspect, a nucleic acid target can be bound to a barcoded solid support non-specifically first. It then mixes with transpososome and ligase to attach the barcode information to the nucleic acid target covalently via simultaneous strand transfer reaction and ligation reaction.

[00072] In some embodiments, transpososomes are not pre-assembled before the reaction. A transposase and a transposable DNA are used directly in the reaction with a nucleic acid target, a ligase and a solid support. In some embodiments, the transposable

DNA can be directly ligated to the barcode template on the solid support via single stranded ligation or double stranded ligation (Fig. 2). In some embodiments, a linker unit can be added in the simultaneous strand transfer and ligation reaction to facilitate ligation (Fig. 3). In some embodiments, all the transpososomes contain the same transposable DNA sequence. In some embodiments, the transpososomes contain different transposable DNA sequences (Fig. 4). In some embodiments, only one transposable DNA in the transpososomes can be ligated to barcode template (Fig. 4). In some embodiments, all transposable DNA in the transpososomes can be ligated to barcode template. In some embodiment, all the monomeric unit sequences of transposable DNA in the same transpososome are the same. In some embodiment, the monomeric unit sequences of transposable DNA in the same transpososome are different. In some embodiment, different transposases are used in the reaction. Different methods can be used to break strand transfer complex, such as, protease treatment, high temperature treatment, or a protein denaturing agent, e.g. SDS solution, guanidine hydrochloride, urea, etc., or a combination thereof. In some embodiments, a single stranded exonuclease may be used to remove unwanted single stranded polynucleotides on the solid support after the barcode tagging (Fig. 5).

[00073] In one aspect, transpososomes can be used sequentially in the reaction. In some embodiments, these transpososomes are the same. In other embodiments, these transpososomes are different. In some embodiments (Fig. 6), a first transpososome mixes with a nucleic acid target, a ligase and a barcoded solid support to generate immobilized nucleic acid STC I on the solid support. A second transpososome is then added to attack the immobilized STC I and forms STC II. In some embodiments, the second transpososome may have a different type of transposon and transposase. In some embodiments, the transposable DNA in the second transpososome may have a different transposon sequence of the same type of transposon in the first transpososome. In some embodiments, the second transpososome may have a different transposable DNA sequence but with the same transposon sequence as the first transpososome. In some embodiments, the capture reactions, such as ligation and/or hybridization occur simultaneously again with the second strand transfer reaction. In some embodiment, a reaction buffer is optimized to be used for both first and second simultaneous strand transfer reaction with the transpososomes and capture reaction with hybridization and/or ligation. Without the need of changing any reaction buffer among different steps can significantly simplify the workflow. Target specific barcode can be attached to the fragments after breaking the STCs. In some embodiments, the second transpososome may added only after breaking the first STCs (Fig. 7). This method has better strand transfer efficiency because steric hindrance effect from the first STCs is removed before the second strand transfer reaction. In some embodiments, the second

strand transfer reaction is used to generate shorter fragment size. In some embodiments, the second transpososome is used to introduce a different adaptor or primer sequences from the first transpososome to facilitate downstream amplification and sequencing.

5 [00074] In one aspect, a nucleic acid target is reacted with a first transpososome to form a stable STC I. The nucleic acid with STC I then react with a second transpososome, a ligase and a clonal barcoded solid support to generate target-specific barcode tagged fragments (Fig. 8).

[00075] In one aspect, a transpososome can be attached to barcoded solid support first.

10 To generate target specific barcode tagged fragments, a nucleic acid target, an in-solution transpososome, a ligase and the transpososome attached barcoded solid support are then mixed together in one reaction vessel as Fig. 9. In some embodiments, the in-solution transpososome is the same as the pre-attached transpososome on the solid support. In some embodiments, the in-solution transpososome is different from the pre-attached transpososome on the solid support.

15 [00076] In one aspect, a transposable DNA can be attached to barcode solid support first.

To generate target specific barcode tagged fragments, a nucleic acid target, an in-solution transpososome, a ligase and the transposable DNA attached barcoded solid support are then mixed together in one reaction vessel. In some embodiments, the in-solution transpososome has the same transposon as the pre-attached transposable DNA on the solid support. In some embodiments, the in-solution transpososome has different transposon from the pre-attached transposable DNA on the solid support. In some embodiments, the in-solution transpososome is replaced with individual transposable DNA and transposase.

25 [00077] In one aspect, the ligation reaction in the simultaneous strand transfer reaction and ligation reaction described in Figs. 1, 4, 6, 7, 8 and 9, can be replaced with a hybridization reaction as Fig. 18. Ligase can be added later in the workflow, either before the STC breaking or after the STC breaking, to ligate hybridized transposable DNA covalently onto the barcoded template on the solid support.

30 [00078] In one aspect, the hybridization reaction in the simultaneous strand transfer reaction and hybridization reaction described in Figs. 18, can be replaced with other capture reactions, such as, affinity tags (e.g. biotin and streptavidin), antibody to antigen, click chemistry, or a combination thereof.

35 **Clonally capture nucleic acid target by non-specific binding on a solid support for barcode tagging**

[00079] The present invention provides methods and compositions that capture nucleic acid targets by non-specific binding on a clonally barcoded solid support. The captured

nucleic acid target can be covalently attached to the barcode templates on the solid support and generate small fragments from the nucleic acid target with a target-specific barcode attached.

[00080] In one aspect, a nucleic acid target reacts with transpososomes and forms strand transfer complexes. The nucleic acid target with the STCs bind non-specifically to a solid support with clonally or semi-clonally barcode templates immobilized on the surface (Fig. 10). In some embodiments, a nucleic acid target can bind non-specifically to a barcoded solid support first. The bound nucleic acid then reacts with transpososomes in the solution to form STCs on the solid support. In one aspect, the STCs are broken by a SDS treatment and the nucleic acid target breaks into small fragments which are still attached to the solid support by non-specific binding under the condition (Fig. 10). A ligase is added to covalently attach the small fragments to the barcode templates on the solid support, and generate small fragments attached with target specific barcode sequences (Fig. 10). In another aspect, the nucleic acid target with the STCs bound non-specifically on the solid support ligates to the barcode templates on the solid support via the ligatable transposable DNA in the STCs first. The STCs are then broken by a SDS treatment and generate small fragments attached with target specific barcode sequences (Fig. 11).

[00081] In one aspect, a nucleic acid target can bind non-specifically to a barcoded solid support first. The bound nucleic acid then reacts with transpososomes and ligase in the solution to form STCs and ligate the transposable DNA to the barcode templates simultaneously on the solid support.

[00082] Many conditions can make nucleic acid and nucleic acid & protein complex bind to a solid support non-specifically. Most notably, polyethylene glycol with salt (Lis and Scheif, 1975), polyamines and cobalthexamine (Pelta et al, 1996), and alcohols (Crouse and Amorese, 1987) are widely used to precipitate and/or condense nucleic acid.

[00083] In one aspect, the ligation reaction described in the invention, can be replaced with other capture reactions, such as, hybridization, affinity tags (e.g. biotin and streptavidin), antibody to antigen, click chemistry, or a combination thereof.

Tracking clonal or semi-clonal barcoded solid support

[00084] Clonal barcoded solid support comprises a plurality of barcode templates with an identical barcode sequence on its surface. Semi-clonal barcoded solid support comprises a plurality of barcode templates with more than one identical barcode sequence. In most cases, the barcode sequences among different clonal or semi-clonal barcoded solid support are different. In order to track different batches of clonal or semi-clonal barcoded solid support, a plurality of barcode templates with an identical barcode is attached to the clonal or semi-clonal barcoded solid support during the preparation so

that all these clonal or semi-clonal barcoded solid supports in the same batch of preparation comprise an additional barcode with the same sequences among them. This additional shared barcode template can comprise up to 50% barcode population on each clonal or semi-clonal barcoded surface or cluster and is defined as minority barcode group in order to differentiate it from the barcode templates on the solid support used for tracking nucleic acid fragment origin which is now defined as majority barcode group. Preferably this shared minority barcode template comprises less than 10% barcode population per clonal or semi-clonal barcoded surface/cluster. The amount of the minority barcode on the barcode solid support will not affect the ability of majority barcode used for tracking nucleic acid fragment origin and related applications. In addition, the minority barcode sequence is predefined and can be filtered out informatically when needed. In some embodiment, more than one minority barcode template with different barcode sequence can be used. This minority barcode on the clonal or semi-clonal barcode solid support can serve as an identifier for barcode solid support. In some embodiment, it can be used to monitor the production and track the usage of the barcoded solid support; in some embodiments, it can be used to detect any potential cross sample contamination and sequencing system contamination, such as, index hopping identified on Illumina sequencing system. This kind of beads, i.e. beads with different barcode templates on the surface, can also be used for nucleic acid barcoding reaction in a compartmentalized reactor, such as, aliquots or droplets.

Releasing clonally barcode tagged nucleic acid fragments to generate sequencing library

[00085] The barcode tagged fragments are immobilized on the solid support. They can be used to make sequencing library. In some embodiments, it can be further manipulated for other applications, such as, treatment with bisulfite for methylation study. In some embodiments, additional sequencing adaptor can be attached to the barcode tagged fragments using transposase-based tagging method (Fig. 12). In some embodiments, barcode tagged fragments can be further fragmented with physical shearing methods and/or enzymatic fragmentation methods, then additional sequencing adaptor can be ligated on (Fig. 13). Immobilized barcode tagged fragments can be released from the solid support in many ways. In one embodiment, a cleavable link or a rare restriction site may be included in the oligonucleotide sequence which is attached to the solid support. With a cleavage reaction or a restriction enzyme digestion, the barcode tagged fragments can be released from the solid support. In some cases, a primer extension may be performed to make a copy or copies of the barcode tagged fragments (Fig. 14A). In some embodiments, the primer is random primer. In some embodiments, the primer is target specific primer (Fig. 14A, 14C). The target of the specific primer can be exon, intron, gene, exome, etc. More detail application for targeted sequencing is described in

patent application WO 2017/151828. Further PCR amplification with primers which are specific for a sequencing platform, e.g., P5 and P7 primers for Illumina's SBS library (Fig. 15), or P1 and A primers for Ion Torrent's library, may generate sequencing ready libraries for the specific sequencing platform. When a library is being made by releasing the barcode tagged fragments from the solid support, a primer with sample specific index may be used. In some cases, the sequence in the barcode template may be used as sample specific index. The released barcode tagged fragments with sample specific index can mix with tagged fragments from other samples with their own sample specific index together for further downstream workflow in order to increase sample preparation throughput and simplify the process. The constructed libraries can be sequenced to determine sequences of both barcode and nucleic acid fragment, and determine the linkage information of the nucleic acid target based on the barcode sequences when at least two fragments from the same nucleic acid target received identical barcode information. The linkage information can be used for haplotype phasing, structural variation detection, CNV detection, etc. The barcode information can also be used to differentiate the sources of duplicated reads from amplification or from sequencing. In some embodiments, the nucleic acid targets are double stranded DNA. In some embodiments, the nucleic acid targets are DNA and RNA hybrid.

Assemble barcode sequencing reads into long reads

[00086] This invention provides methods and compositions to clonally barcode tag nucleic acid samples in an open bulk reaction without sophisticated compartmentation or partition scheme as other methods. The barcode tagged fragments may be from a whole genome sample, or a portion of a genome, or a targeted region, or metagenomic samples. The sequencing reads generated from these barcode tagged fragments contain the barcode information which can be used to identify the original target of these fragments. These short sequencing reads with the same barcode can be grouped together and cluster along the original nucleic acid targets. Depending on which transposase system is used, among these reads with the same barcode, starting ends of two originally adjacent reads from the same nucleic acid target will share some bases of reverse complimentary sequences (5 bases for MuA transposase system and 9 bases for Tn5transposase system). These overlap sequences can further link the barcode reads together. In principle, it can re-construct the original nucleic acid target completely when all the tagged fragments are captured by barcoded solid support and sequenced. They provide useful long-range linkage information to be used for haplotype phasing. The longer the original nucleic acid targets are, the longer the linkage information will be, the more useful they are for phasing application. An analysis pipeline which can be developed for full genome assembly or structural variation analysis using these barcode

reads for both *de novo* sequencing and resequencing. In one case, all the sequencing reads may be used for standard shotgun assembly analysis to establish many initial contigs first. The barcode information can then be used to phase the initial contigs into much longer contigs. These barcode tagging methods can also be used for phasing the targeted gene, genes, or exome. These barcode tagging methods may also be used as a tool for differentiating the duplicated reads in the targeted sequencing application. This method improves sequencing assay detection limit on heterogeneous samples, e.g., somatic mutation detection in a cancer biopsy sample or circulating tumor cell/DNA.

[00087] Although the invention has been explained with respect to an embodiment, it is to be understood that many other possible modifications and variations can be made without departing from the spirit and scope of the invention as herein described.

[00088] Further, in general with regard to the processes, systems, methods, etc. described herein, it should be understood that, although the steps of such processes, etc. have been described as occurring according to a certain ordered sequence, such processes could be practiced with the described steps performed in an order other than the order described herein. It further should be understood that certain steps could be performed simultaneously, that other steps could be added, or that certain steps described herein could be omitted. In other words, the descriptions of processes herein are provided for the purpose of illustrating certain embodiments and should in no way be construed so as to limit the claimed invention.

[00089] Moreover, it is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments and applications other than the examples provided would be apparent to those of skill in the art upon reading the above description. The scope of the invention should be determined, not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled. It is anticipated and intended that future developments will occur in the arts discussed herein, and that the disclosed systems and methods will be incorporated into such future embodiments. In sum, it should be understood that the invention is capable of modification and variation and is limited only by the following claims.

[00090] Lastly, all defined terms used in the application are intended to be given their broadest reasonable constructions consistent with the definitions provided herein. All undefined terms used in the claims are intended to be given their broadest reasonable constructions consistent with their ordinary meanings as understood by those skilled in the art unless an explicit indication to the contrary is made herein. In particular, use of the singular articles such as "a," "the," "said," etc. should be read to recite one or more of the indicated elements unless a claim recites an explicit limitation to the contrary.

Example 1

[00091] This example describes a method of target specific barcode tagging of genomic DNA with simultaneous strand transfer and ligation onto barcoded beads in an open bulk reaction without partition of genomic DNA (Fig. 4). Clonally barcode beads were prepared as described in patent application WO 2017/151828. Each barcode sequence was 18-base in length. All the beads including those without clonally amplified barcode templates were collected directly after BEAMing reaction (Diehl et al, 2005). Two MuA transpososomes were pre-assembled with two different MuA transposable DNA separately. The MuA transposable DNA in one MuA transpososome had a ligatable 5' end transposon joining strand and could hybridize and/or ligate to the barcode template on the barcoded beads. Double stranded barcode templates on the beads were denatured into single stranded. 20 million denatured beads were incubated with 5 ng genomic DNA extracted from human embryonic kidney cells 293FT, the two preassembled MuA transpososomes and T4 DNA ligase in a reaction buffer which enabled both strand transfer reaction and ligation reaction at substantially the same time at 37°C for 30 minutes. Reaction was terminated with 0.5% SDS solution. Washed beads were treated with exonuclease I to remove single stranded polynucleotides, and then used for 15-cycle PCR amplification to release immobilized barcode tagged DNA fragments. PCR products were purified with 0.8X AMPure XP beads to remove small primer dimers and PCR by-products, and examined using a high sensitivity D5000 screentape on a TapeStation (Fig. 16A). The purified PCR products were sequenced on an Illumina MiniSeq instrument. Reads with the same barcode sequence were sorted for each barcode based on the reference genome alignment location. Read distance to the next alignment was calculated and read count frequency along the read distance was plotted in Fig. 16B. When barcoded reads kept the linkage information of reads from tagged DNA fragments, piling up of proximal reads was expected. The read distance from the original different DNA fragments would also pile up as distal reads with much longer distance. A bi-modal distribution of read count frequency plot would be expected, which was exactly observed in the Fig. 16B. Although the sequencing depth was very limited in a multi-sample MiniSeq run, strong enrichment of shorter distance proximal reads demonstrated successful barcode reads contiguity.

Example 2

[00092] This example describes a method of target specific barcode tagging of genomic DNA with simultaneous strand transfer and ligation onto barcoded beads in an open bulk reaction without partition of genomic DNA (Fig. 7). One MuA transpososome comprised a MuA transposable DNA which had a ligatable 5' end transposon joining strand and

could ligate to the barcode template on the barcoded beads. 20 million barcoded beads were incubated with 5 ng genomic DNA extracted from human embryonic kidney cells 293FT, the ligatable MuA transpososomes and T4 DNA ligase in the reaction buffer at 37°C for 30 minutes. Reaction was terminated with 0.5% SDS solution. Washed beads were then reacted with another MuA transpososome and exonuclease I. The reaction was stopped with 0.5% SDS again. The beads with barcode tagged fragments were used for 15-cycle PCR amplification to release immobilized barcode tagged fragments. This method generated fewer PCR by products than the method in the Example 1. PCR products were purified with 0.8X AMPure XP beads to remove small primer dimers and PCR by-products, and examined using a high sensitivity D5000 screentape on a TapeStation (Fig. 17A). The purified PCR products were sequenced on an Illumina MiniSeq instrument. Reads with the same barcode sequence were sorted for each barcode based on the reference genome alignment location. Read distance to the next alignment was calculated and read count frequency along the read distance was plotted in Fig. 17B. When barcoded reads kept the linkage information of reads from tagged DNA fragments, piling up of proximal reads was expected. The read distance from the original different DNA fragments would also pile up as distal reads with much longer distance. A bi-modal distribution of read count frequency plot would be expected, which was exactly observed in the Fig. 17B. Although the sequencing depth was very limited in a multi-sample MiniSeq run, strong enrichment of shorter distance proximal reads demonstrated successful barcode reads contiguity.

Example 3

[00093] 10ng genomic DNA from HapMap sample NA12878 was used to generate barcode tagged Illumina sequencing library with the method illustrated in Fig. 4. 2x75bp paired end sequencing was performed on Illumina NextSeq system. Over 600 million paired end reads were pooled for haplotype phasing analysis using a HapCUT2 algorithm (Edge P et al, 2017). There were approximately 22-fold genome coverage depth after removing duplicated reads. The largest phased block size was 9.5Mb and N50 phased block size was 1.7Mb with a switch error rate at 0.14%.

Example 4

[00094] We compared three different transposase-based methods to generate clonal barcode tagged fragments for sequencing library construction (Fig. 19). Method 1 was the simultaneous strand transfer and capture reactions disclosed in this invention. 1ng *E. coli* genomic DNA was mixed with ligatable transpososomes, DNA ligase and 20 million beads among which there were 1 million clonally barcode templated beads in the same reaction buffer to generate clonal barcode tagged fragments and soluble sequencing

library by further PCR amplification. Method 2 and 3 were two methods using separate strand transfer and capture reactions to generate clonal barcode tagged fragments. Method 2 was to generate in-solution STCs first by mixing ligatable transposomes with 10ng *E. coli* genomic DNA; 1/10th of these in-solution STCs which contained 1ng original *E. coli* genomic DNA was then mixed with DNA ligase and 20 million beads among which there were 1 million clonally barcode templated beads in a ligation buffer to capture the in-solution STC onto the beads. Method 3 was to immobilize the ligatable transposomes onto the barcode templated beads first by mixing transposomes and DNA ligase with 20 million beads among which there were 1 million clonally barcode templated beads in a ligation buffer; the reacted beads were washed to remove ligase and then reacted with 1ng *E. coli* genomic DNA in a strand transfer reaction buffer. At last the same number of beads from these three methods were used for PCR amplification to generate soluble sequencing libraries by the same number of PCR cycles. The PCR products were loaded onto a 2% agarose E-gel EX to compare their yield (Fig. 20). Method 1 (Fig. 20, lane M1) produced the most library product as we expected, which demonstrated the superior performance of Method 1 to the other two methods (Fig. 20, lane M2 and lane M3).

Example 5

[00095] 0.5 ng high molecular weight genomic DNA extracted from *E. coli* DH10B cells was used to generate barcode tagged Illumina sequencing library with the method illustrated in Fig. 4 using the TELL-Seq™ WGS Library Prep Kit (Universal Sequencing Technology Corporation, Carlsbad, CA) except replacing the TELL beads in the kit with the clonally barcoded beads described here. Three different preps of clonally barcoded beads were made. The barcode templates on these beads contained three different level (high, medium and low) of common barcode sequence (TAGAGAGGCTCTGGATCG) shared among all these barcoded beads besides the clonal unique barcode sequence each bead has. Based on the sequencing analysis of barcode sequence on these beads, percentage of barcode templates containing the common barcode sequence on each bead is average 14.10%, 8.51% and 0.81% of total barcode template counts on each bead, for high (T519), medium (T522) and low (T524) level, respectively (Table 1).

[00096] Table 1. The level of common barcode sequence on each barcoded bead among three preps of clonally barcoded beads measured based on sequencing.

Sample	Common Barcode Sequence	% of barcode
T519	TAGAGAGGCTCTGGATCG	14.10
T522	TAGAGAGGCTCTGGATCG	8.51
T524	TAGAGAGGCTCTGGATCG	0.81

[00097] The TELL-Seq libraries generated from these beads were sequenced on an Illumina NextSeq system. The general sequencing statistics were summarized in Table 2. The reads associated with the common barcode sequence (TAGAGAGGCTCTGGATCG) were considered as a part of reads with error barcodes, which were filtered out before further downstream analyses.

[00098] Table 2. 2x71 paired end sequencing summary of three TELL-Seq libraries generated from three preps of barcoded beads, T519, T522 and T524

sample	T519	T522	T524
total_reads	7,868,606	7,378,261	7,623,585
%reads_with_error_barcode	39.4%	14.7%	5.0%
final_reads_number	4,779,419	6,306,768	7,263,094
final_correct_barcode_number	512,620	181,469	621,889
barcode_with_multi_reads	223,104	126,359	321,841
read1_reads_mapped_percentage	95.90%	96.24%	98.70%
read2_reads_mapped_percentage	87.65%	90.47%	89.37%

[00099] The barcode read distance plots for these three samples all showed very good bi-modal distribution of proximal linked reads and non-linked distal reads and were very similar as each other (Fig. 21).

[000100] De novo assemblies of these sequencing data were successfully generated using TELL-link software. All three samples showed nearly full length assemble of the *E. coli* DH10B genome with relatively low mismatches and indel errors in their assemblies (Table 3). The data indicated that the level of common barcode sequences among the clonal barcodes beads up to 15% did not have any adverse effect on the linked read quality and do novo assembly results.

[000101] Table 3. QUAST analysis summary of de novo assembly results of sequencing the three TELL-Seq libraries generated from three preps of barcoded beads, T519, T522 and T524

Sample	T519	T522	T524
kmer size	41 / 31	45 / 31	41 / 31
# contigs (>= 0 bp)	83	65	74
# contigs (>= 1000 bp)	66	52	58
# contigs (>= 5000 bp)	2	2	2
Total length (>= 0 bp)	4,897,312	4,885,670	4,863,532
Total length (>= 1000 bp)	4,885,015	4,875,708	4,851,921

Total length (≥ 5000 bp)	4,757,856	4,778,570	4,742,842
Largest contig	4,642,570	4,661,172	4,646,181
Total length	4,897,312	4,885,670	4,863,532
Reference length	4,686,137	4,686,137	4,686,137
N50	4,642,570	4,661,172	4,646,181
Genome fraction (%)	99.40	99.52	99.38
Largest alignment	4,524,892	3,054,128	3,043,049
Total aligned length	4,779,212	4,765,560	4,763,057
NA50	4,524,892	3,054,128	3,043,049
# misassemblies	1	5	4
Duplication ratio	1.027	1.023	1.024
# N's per 100 kbp	33.79	20.88	21.2
# mismatches per 100 kbp	9.9	12.16	12.24
# indels per 100 kbp	0.54	0.86	1.12
GC (%)	50.7	50.7	50.8
Reference GC (%)	50.8	50.8	50.8

All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (e.g., "# contigs (≥ 0 bp)" and "Total length (≥ 0 bp)" include all contigs).

REFERENCES

- [000102] Amini S. et al. 2014. Nature Genetics, 46(12): 1343-1349.
- 5 [000103] Au T et al. 2004. EMBO J., 23: 3408-3420.
- [000104] Bankevich A. et al. 2012. J Comput Biol. 5: 455-77.
- [000105] Boeke J. D. 1989. Transposable elements in *Saccharomyces cerevisiae* in Mobile DNA. pp. 335-374 in Mobile DNA, edited by D. E. BERG and M. M. HOWE.
- [000106] Burton B.M. and Baker T.A. 2003. Chemistry & Biology 10: 463-472.
- 10 [000107] Chen Z. et al. 2017. Foreign Patent Application WO 2017/151828 A1.
- [000108] Craig N. L. 1996. Transposon Tn7. Curr. Top. Microbiol. Immunol. 204: 27-48.
- [000109] Crouse J. and Amorese D. 1987. Focus, 7(4): 1-2.
- [000110] Devine S. E. and Boeke, J. D. 1994. Nucleic Acids Research, 22(18): 3765-3772.
- [000111] Diehl F. et al. 2005. PNAS, 102 (45): 16368-16373.
- 15 [000112] Drmanac R., Peters B. A. and Alexeev A. 2016. United State Patent US9,328,382 B2.
- [000113] Edge P., Bafna V. and Bansal V. 2017. Genome Res., 27(5): 801-812.
- [000114] Haapa S. et al. 1999. Nucleic Acids Research, 27(13): 2777-2784.
- [000115] Ichikawa H. and Ohtsubo E. 1990. J. Biol. Chem., 265(31): 18829-32.
- 20 [000116] Kaufman P. and Rio D. C. 1992. Cell, 69(1): 27-39.
- [000117] Kleckner N. et al. 1996. Curr. Top. Microbiol. Immunol., 204: 49-82.
- [000118] Mizuuchi M., Baker T.A. and Mizuuchi K. 1992. Cell, 70, 303-311.
- [000119] Lampe D. J., Churchill M. E. A. and Robertson H. M. 1996. EMBO J., 15(19): 5470-5479.

- [000120] Lis J. T. and Schleif R. 1975. Nucleic Acid Research, 2(3): 383-389.
- [000121] Ohtsubo E. and Sekine Y. 1996. Curr. Top. Microbiol. Immunol., 204:126.
- [000122] Park B. T., Jeong M. H. and Kim B. H. 1992. Taehan Misaengmul Hakhoechi, 27(4): 381-9.
- 5 [000123] Pelta J., Livolant F. and Sikorav J. L. 1996. J. Biological Chemistry, 271: 5656-5662.
- [000124] Savilahti H., Rice P. A., and MiZuuchi K. 1995. EMBO J., 14: 4893-4903.
- [000125] Surette M., Buch S.J. and Chaconas G. 1987. Cell, 70: 303-311.
- [000126] Varmus H. and Brown. P. A. 1989. Retroviruses, in Mobile DNA. Berg D. E. and
10 Howe M. eds. American Society for Microbiology, Washington D. C. pp. 53-108.
- [000127] Vos J. C., Baere I. and Plasterk R. H. A. 1996. Genes Dev., 10(6): 755-61.
- [000128] Zhang F. et al. 2017. Nature Biotechnology, 35(9): 852-857.
- [000129] Zheng G.X. et al. 2016. Nature Biotechnology, 34(3): 303-311.

What is claimed:

1. A method for tracking an origin of a nucleic acid fragment by barcoding comprising:
 - a. providing a reaction mixture comprising a plurality of double stranded nucleic acid fragments and a plurality of beads,
5 wherein each bead comprises at least two different immobilized barcode templates from at least two different populations of barcode templates, wherein each population of barcode template comprises multiple copies of the same barcode template,
wherein each barcode template comprises a barcode sequence,
10 wherein said barcode sequence is configured to be an identifier of the barcode template,
 - b. producing at least two barcode-attached subfragments from said nucleic acid fragment, wherein the at least two barcode-attached subfragments from the same nucleic acid fragment are each attached to the barcode sequence with a
15 same sequence from the same bead; and
 - c. tracking/identifying the origin of said barcode-attached subfragments by their said barcode sequence,
wherein barcode-attached subfragments with the same sequence tracks to the same nucleic acid fragment.
20
2. The method of claim 1, wherein said reaction mixture is not compartmentalized into aliquots or droplets.
3. The method of claim 1, wherein said beads in said reaction mixture comprise at least
25 about 1000 different barcode sequences in total.
4. The method of claim 1, wherein at least one of said barcode template populations on each bead is also present on at least another bead as a common shared barcode template population among said plurality of beads.
30
5. The method of claim 4, wherein the amount of said common shared barcode template is less than about 50% of total barcode template on said bead.
6. The method of claim 4, wherein the amount of said common shared barcode template is
35 less than about 10% of total barcode template on said bead.
7. The method of claim 1, wherein said double stranded nucleic acid comprises a double stranded DNA, or a DNA/RNA hybrid, or a combination thereof.

8. The method of claim 7, wherein said double stranded nucleic acid is greater than about 1000bp.
- 5 9. The methods of claim 1, wherein said double stranded nucleic acid fragment comprises a nucleic acid molecule comprising DNA or RNA in natural, modified, amplified, or other chemically treated forms or a combination thereof.
- 10 10. The method of claim 1, wherein said double stranded nucleic acid fragment is nonspecifically bound to said bead first before any reactions of claim 1.
- 15 11. The method of claim 1, wherein said double stranded nucleic acid fragment is strand transferred with a transpososome and forms a strand transfer complex before interacting with said bead.
- 20 12. The method of claim 1, wherein said producing barcode-attached subfragment comprises steps of ligation, hybridization, strand transfer reaction, tagmentation, amplification, primer extension, or a combination thereof.
- 25 13. The method of claims 11 or 12, wherein said strand transfer reaction or said tagmentation reaction comprises utilizing a transposase, wherein said transposase is selected from a group consisting of Tn, Mu, Ty, and Tc transposases in a wildtype, a mutant or a tagged version thereof, and a combination thereof.
- 30 14. The method of claim 1, wherein said tracking/identifying the origin of said barcode-attached subfragments comprises sequencing to determine haplotype phasing information and/or structural variation of the nucleic acid fragment.
- 35 15. The method of claim 1, wherein said tracking/identifying the origin of said barcode attached subfragments comprises sequencing to determine the identity of duplicated nucleic acid fragments or copy number variation information.
16. A system for tracking an origin of a nucleic acid fragment by barcoding comprising:
a reaction mixture comprising a plurality of double stranded nucleic acid fragments and a plurality of beads,
wherein each bead comprises at least two different immobilized barcode templates from at least two different populations of barcode templates,

wherein each population of barcode template comprises multiple copies of the same barcode template,

wherein each barcode template comprises a barcode sequence,

wherein said barcode sequence is configured to be an identifier of the barcode template,

wherein at least two barcode-attached subfragments are produced from said nucleic acid fragment, wherein the at least two barcode-attached subfragments from the same nucleic acid fragment are each attached to the barcode sequence with a same sequence from the same bead;

wherein the origin of said barcode-attached subfragments are configured to be tracked/identified by their said barcode sequence, and

wherein said barcode-attached subfragments with the same sequence tracks to the same nucleic acid fragment.

17. The system of claim 16, wherein said reaction mixture is not compartmentalized into aliquots or droplets; and
wherein said beads in said reaction mixture comprise at least about 1000 different barcode sequences in total.

17. The system of claim 15, wherein at least one of said barcode template populations on each bead is also present on at least another bead as a common shared barcode template population among said plurality of beads.

18. The system of claim 17, wherein the amount of said common shared barcode template is less than about 50% of total barcode template on said bead.

19. The system of claim 18, wherein the amount of said common shared barcode template is less than about 10% of total barcode template on said bead.

20. The system of claim 15, wherein said double stranded nucleic acid comprises a double stranded DNA, or a DNA/RNA hybrid, or a combination thereof.

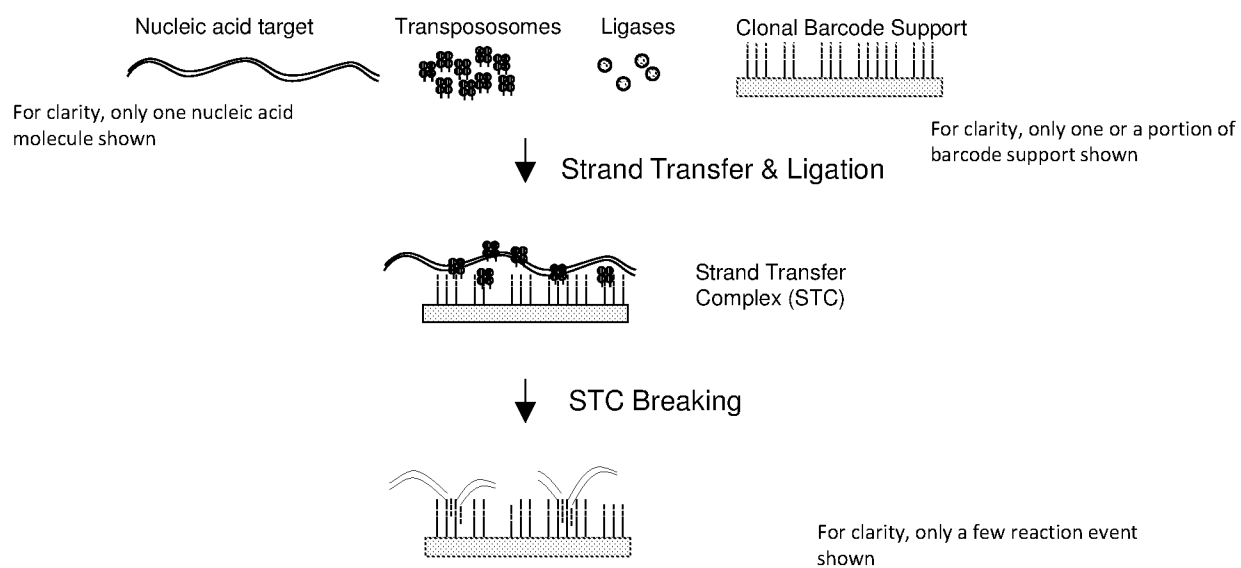


Fig. 1

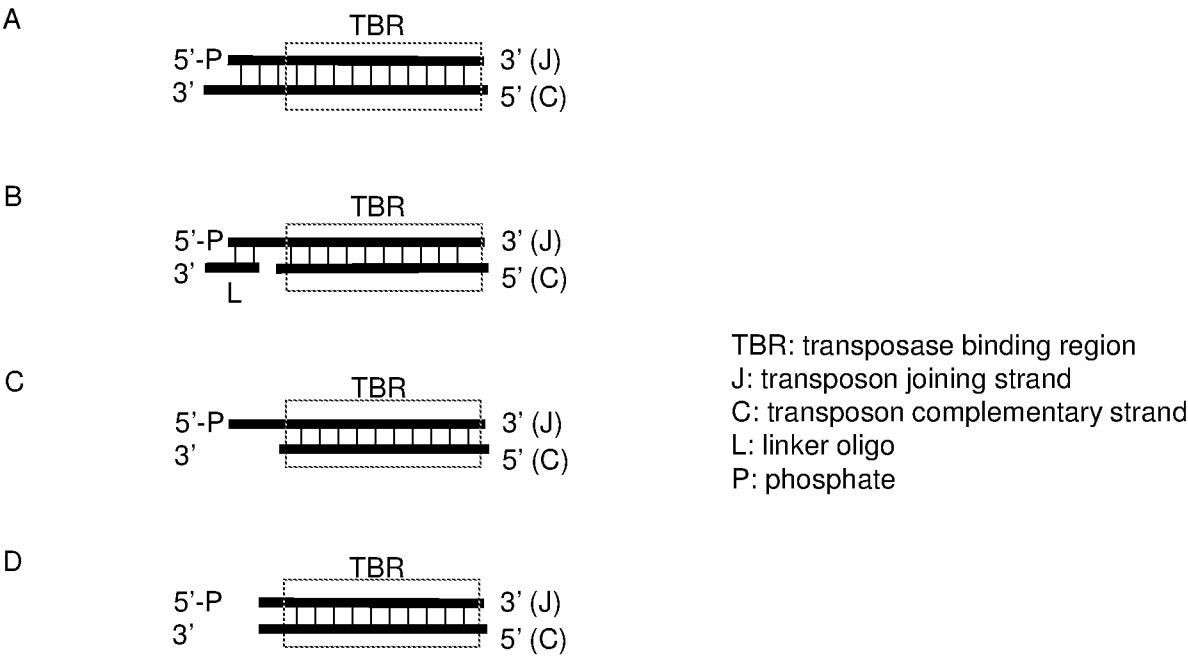


Fig. 2

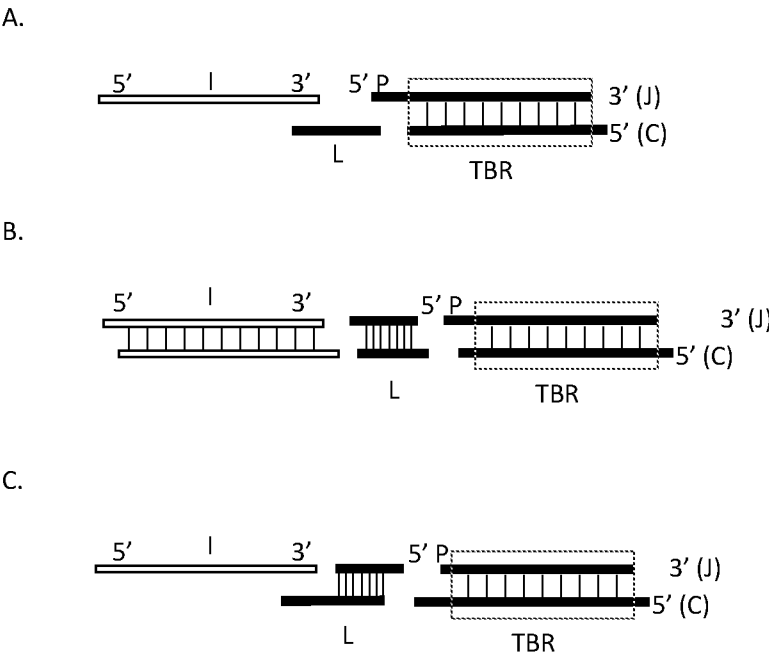


Fig. 3

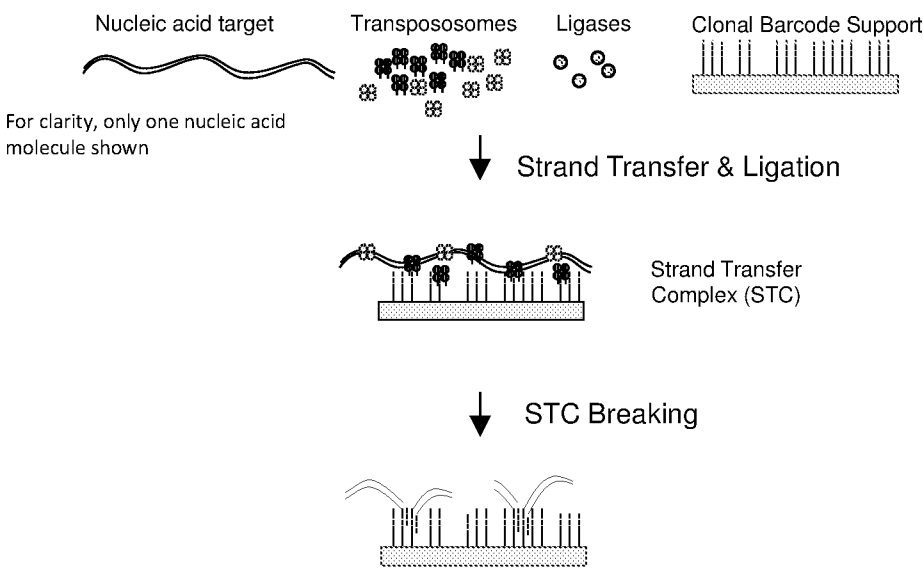


Fig. 4

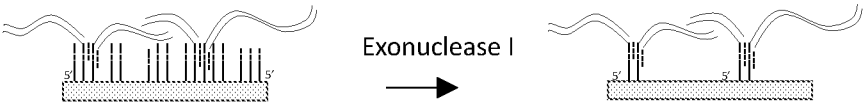


Fig. 5

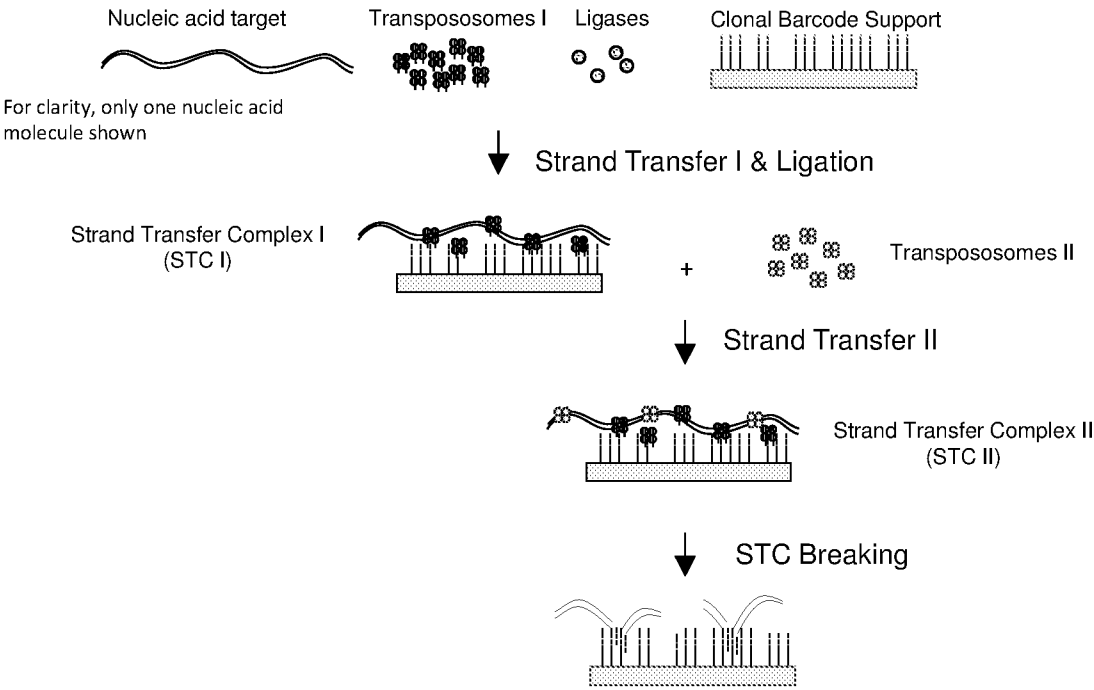


Fig. 6

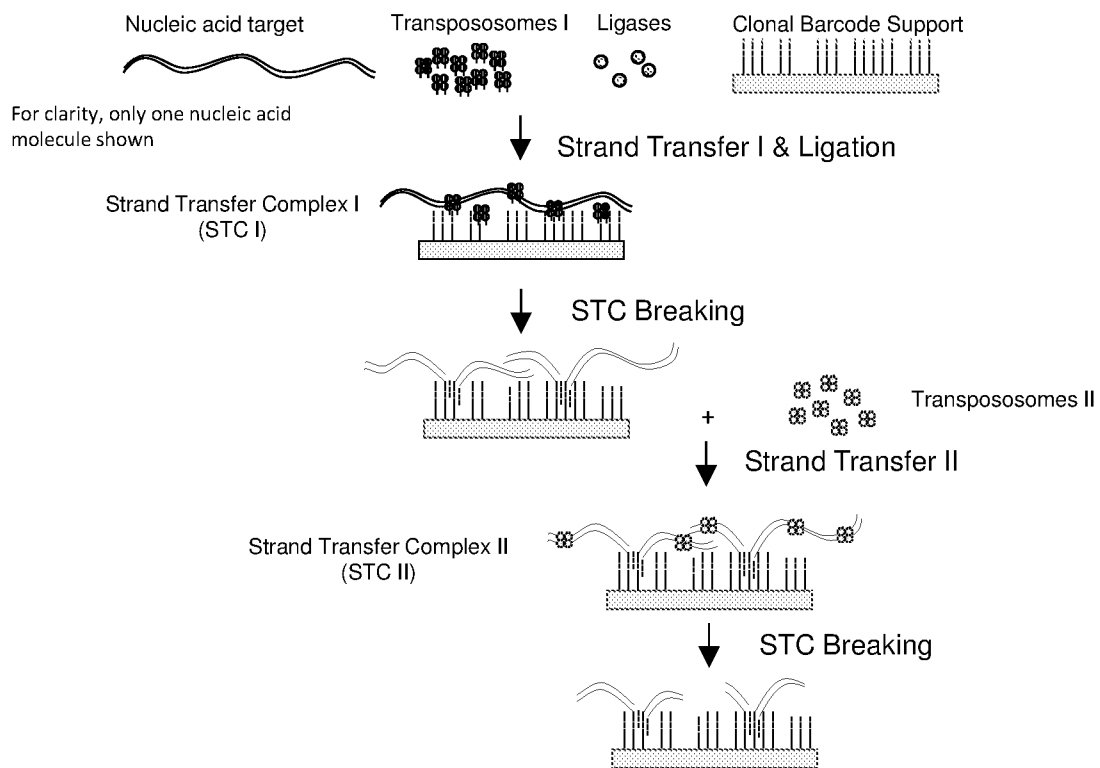


Fig. 7

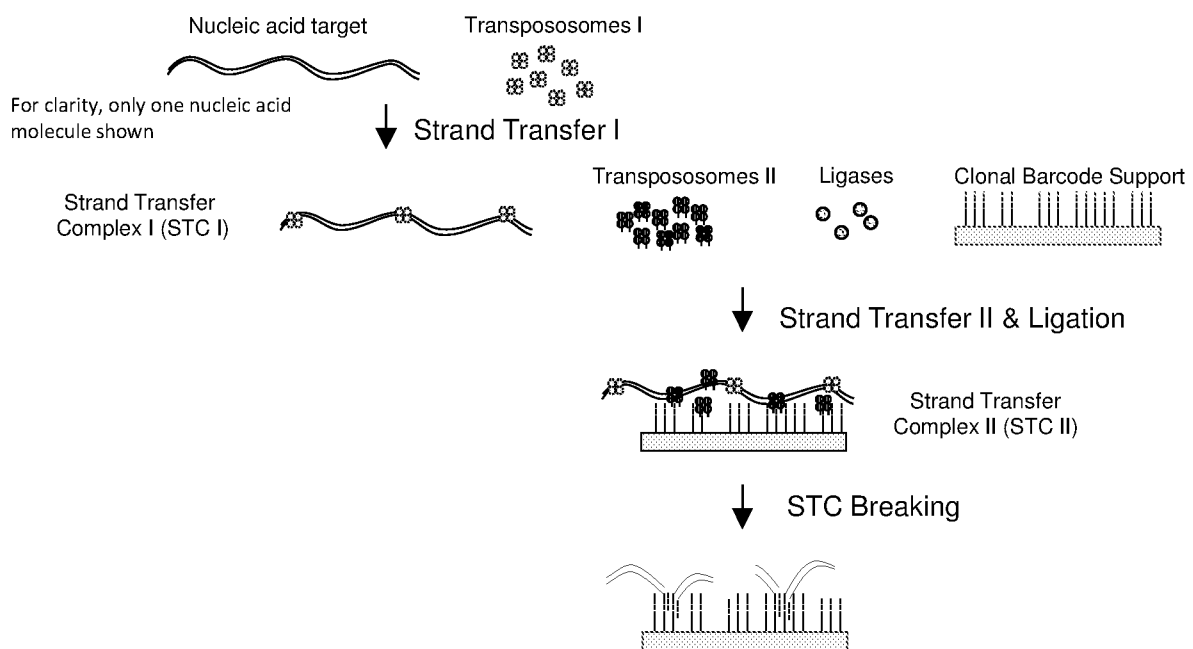


Fig. 8

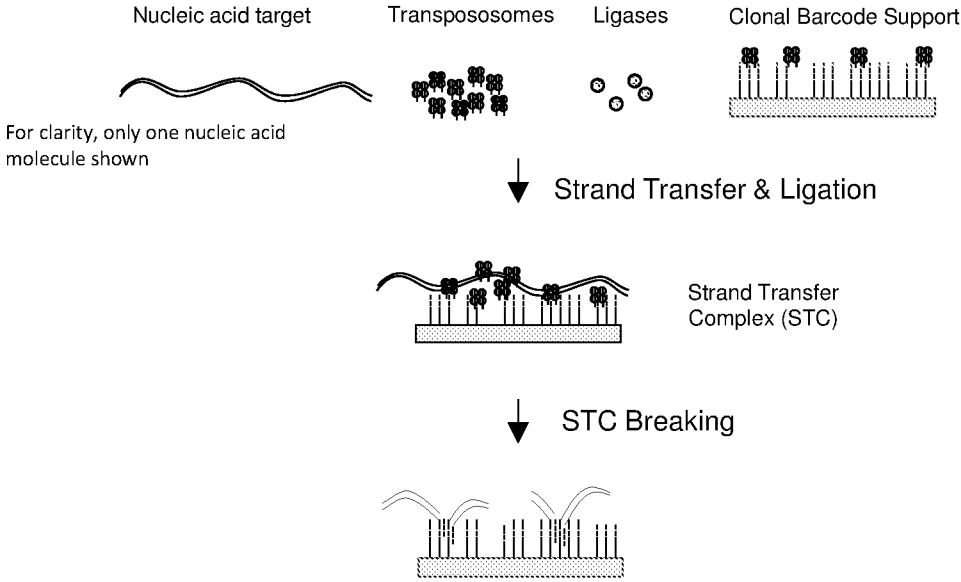


Fig. 9

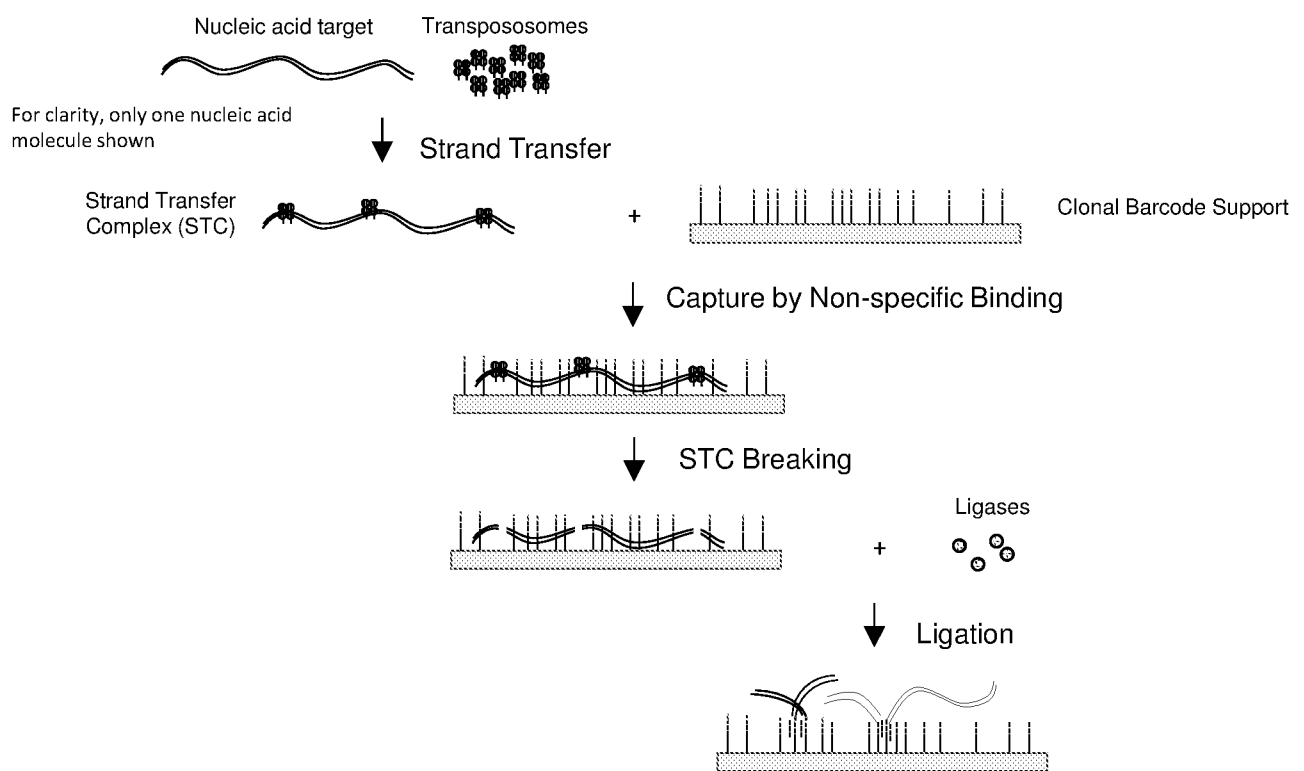


Fig. 10

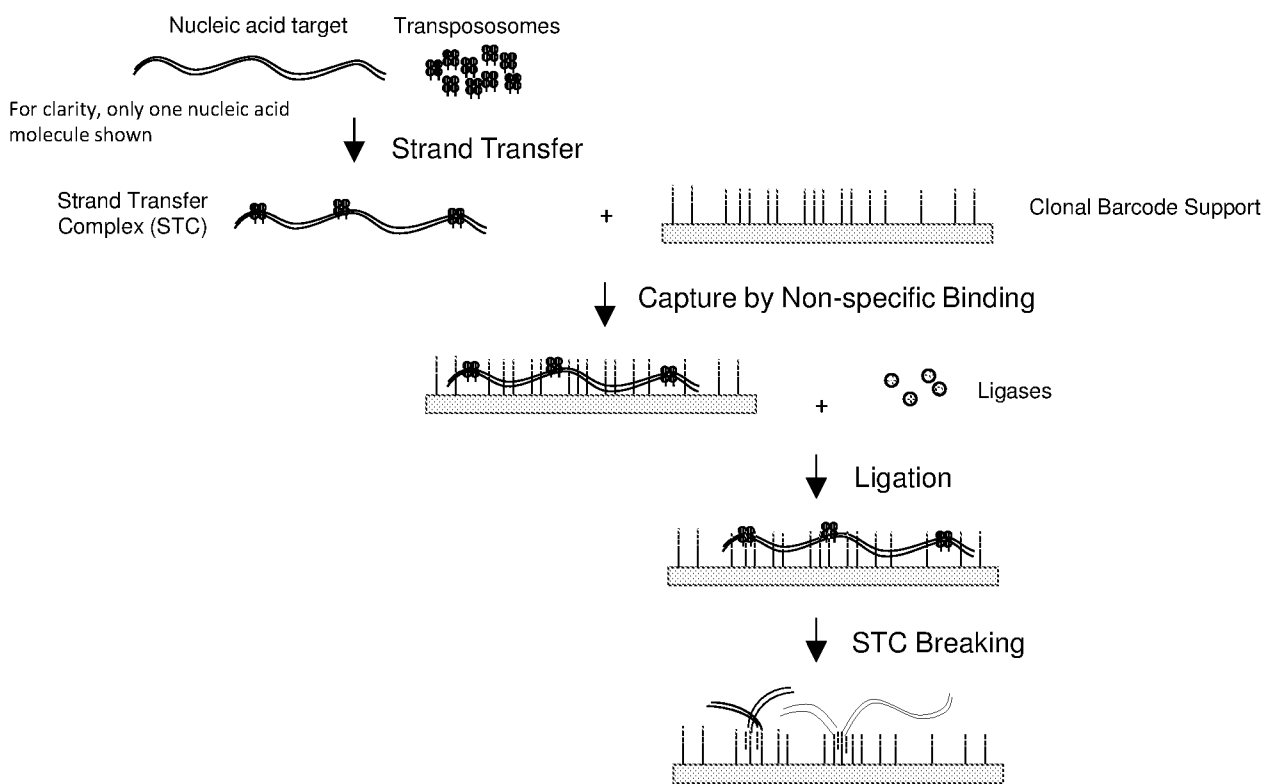


Fig. 11

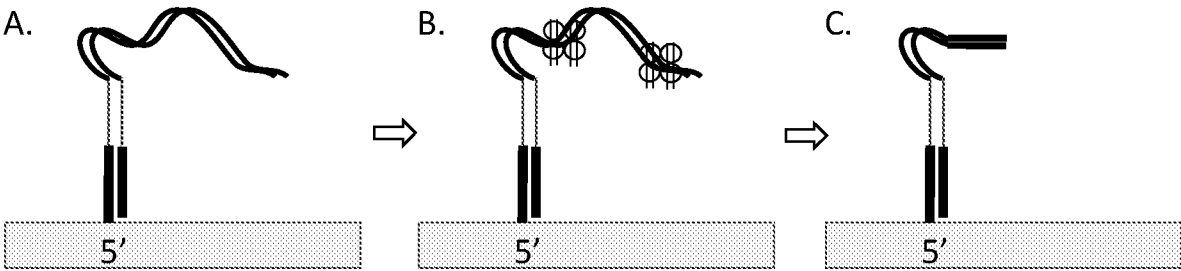


Fig. 12

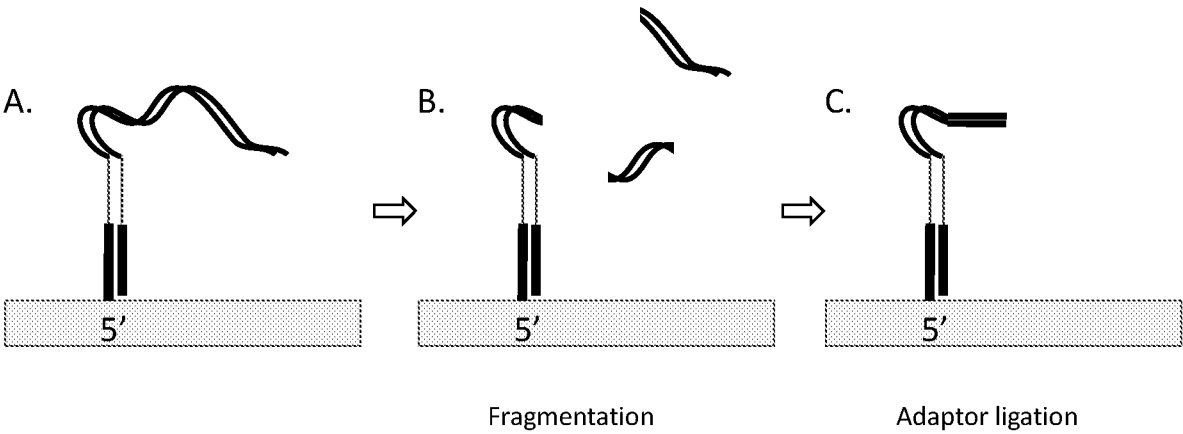


Fig. 13

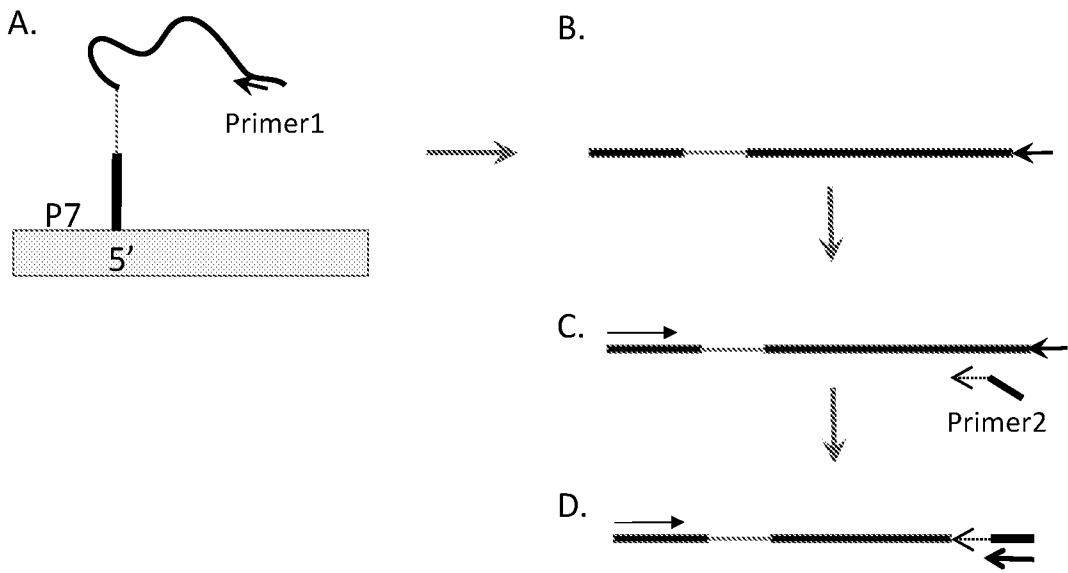


Fig. 14

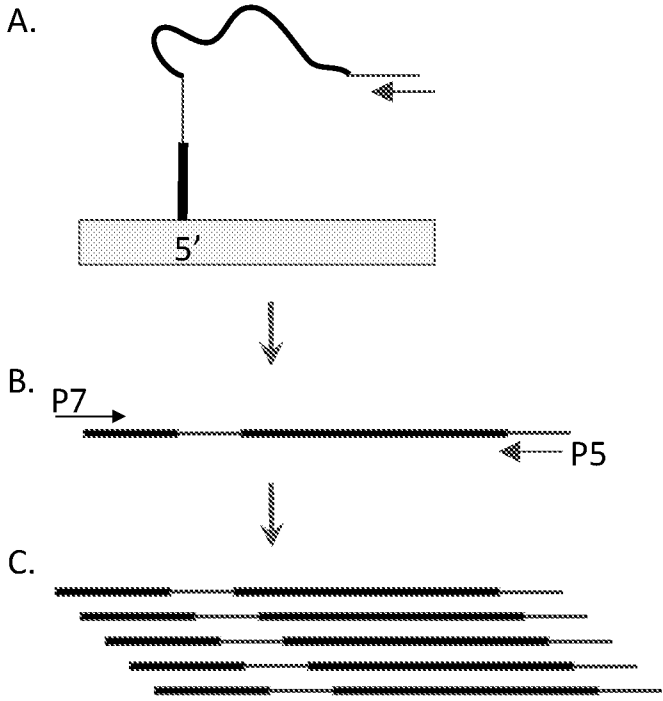
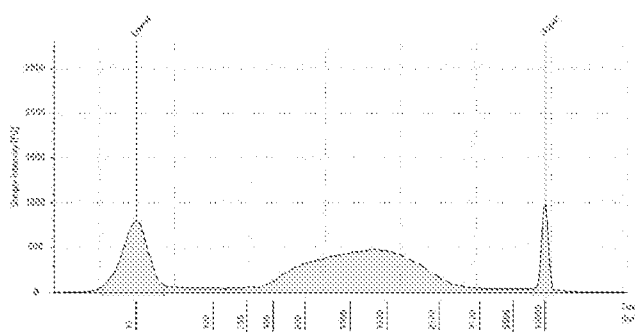


Fig. 15

A



B

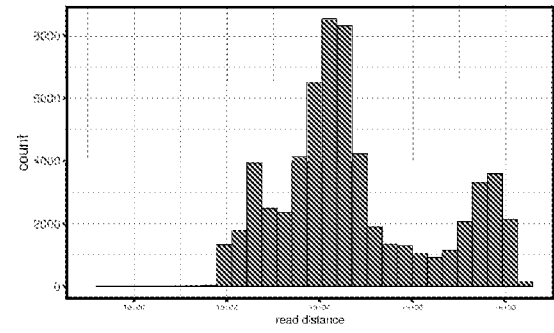


Fig. 16

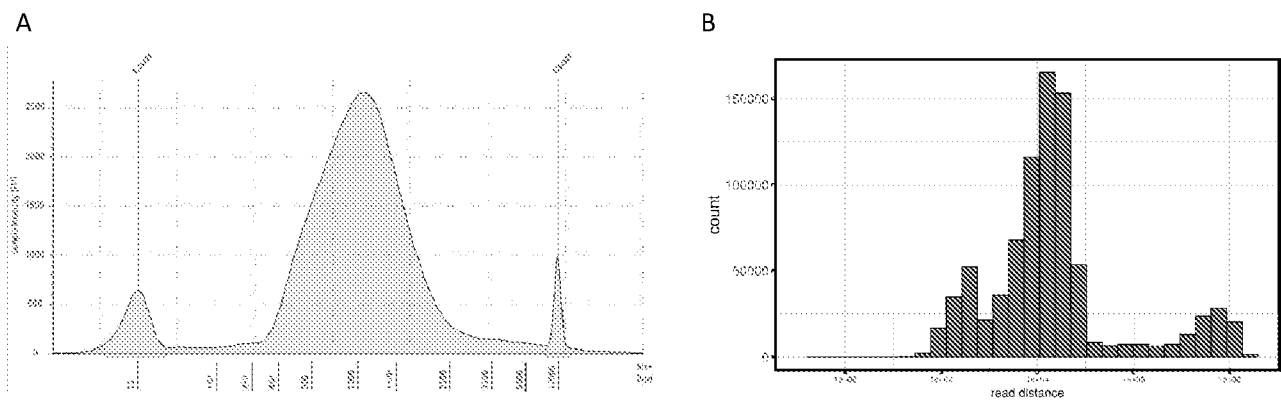


Fig. 17

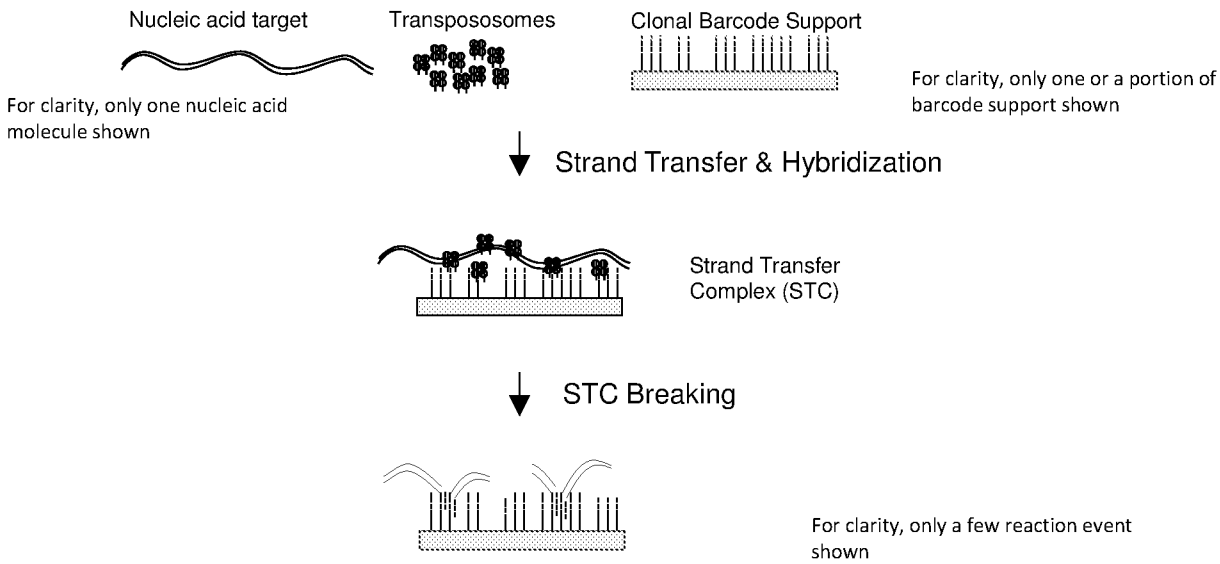


Fig. 18

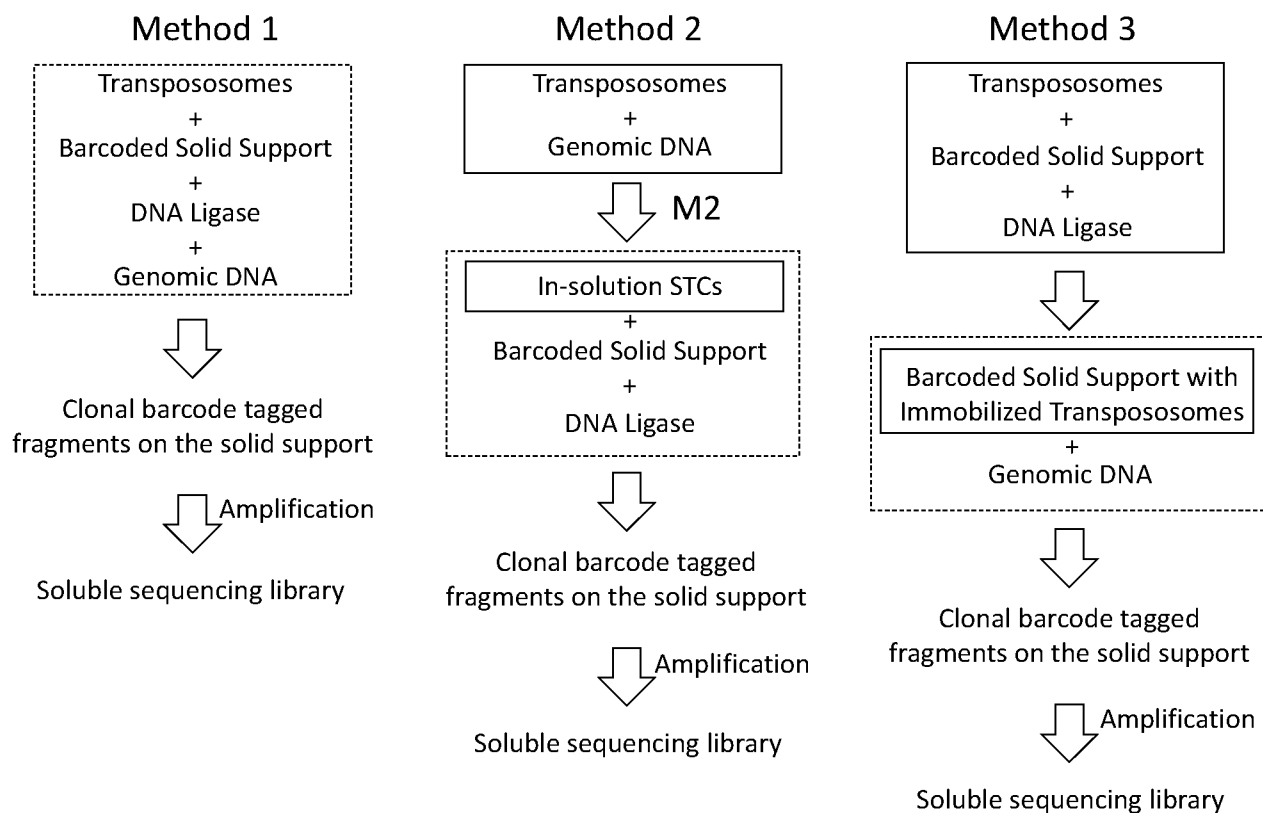


Fig. 19

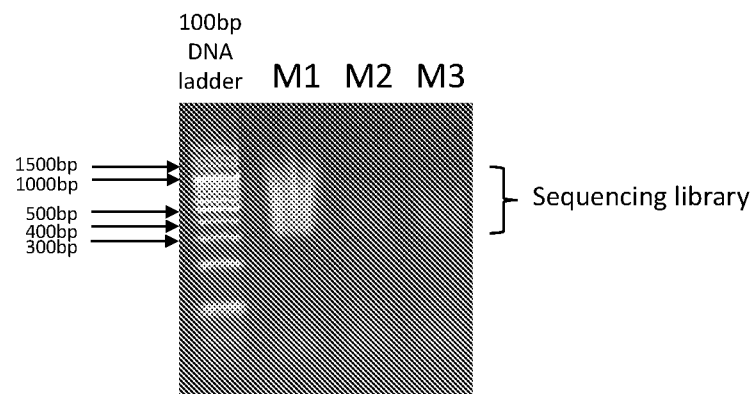
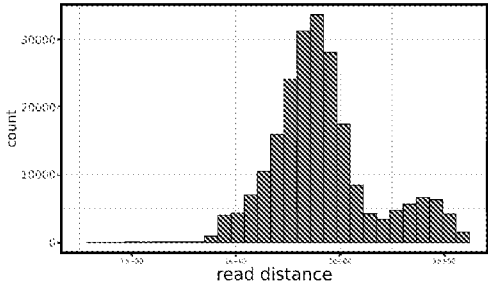
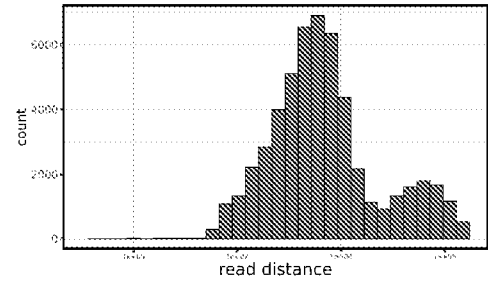


Fig. 20

T519



T522



T524

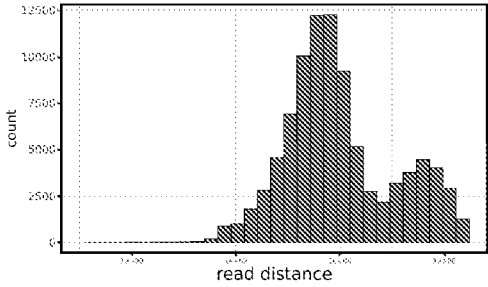


Fig. 21

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/47067

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C12Q 1/6869, C12Q 1/6874, C12Q 1/6876 (2020.01)

CPC - C12Q 1/6883, C12Q 2521/501, G16C 20/60, C12Q 2565/501

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

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

See Search History document

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

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2019/157318 A1 (UNIVERSAL SEQUENCING TECHNOLOGY CORPORATION) 15 August 2019 (15.08.2019); para [0006]-[0007], [0009], [00016], [00021], [00025], [00049], [00067]-[00069], [00071]-[00072], [00083], [00087]-[00088], [00098]-[00099], [000105]; claim 41	1-4, 7-16, 17A, 17B, 20
Y		5-6, 18-19
Y	US 2019/0002970 A1 (COMPLETE GENOMICS, INC.) 03 January 2019 (03.01.2019); para [0092]-[0093]	5-6, 18-19

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

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

12 November 2020

Date of mailing of the international search report

02 DEC 2020

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Lee Young

Telephone No. PCT Helpdesk: 571-272-4300