

CLAIMS

1. A method for sequencing nucleic acid molecules from a sample using unique molecular indices (UMIs), wherein each unique molecular index (UMI) is an oligonucleotide sequence that can be used to identify an individual molecule of a double-stranded DNA fragment in the sample, comprising

(a) applying adapters to both ends of double-stranded DNA fragments in the sample to obtain DNA-adapter products, wherein each adapter comprises a double-stranded hybridized region, a single-stranded 5' arm, a single-stranded 3' arm, and a physical UMI on one strand or each strand of the adapter, thereby obtaining DNA-adapter products, wherein the physical UMI is selected from a plurality of physical UMIs, wherein the plurality of physical UMIs comprises nonrandom UMIs, wherein every nonrandom UMI differs from every other nonrandom UMI of the adapters by at least two nucleotides at corresponding sequence positions of the nonrandom UMIs, wherein each double-stranded DNA fragment in the sample comprises a virtual UMI on one strand or each strand of the double-stranded DNA fragment, and wherein the virtual UMI is a unique sub-sequence in a DNA fragment in the sample;

(b) amplifying both strands of the DNA-adapter products to obtain a plurality of amplified polynucleotides;

(c) sequencing the plurality of amplified polynucleotides, thereby obtaining a plurality of reads each UMI comprising a physical UMI sequence corresponding to a physical UMI on an adapter and a virtual UMI sequence corresponding to a virtual UMI on a double-stranded DNA fragment in the sample;

(d) identifying a plurality of physical UMI sequences for the plurality of reads;

(e) identifying a plurality of virtual UMI sequences for the plurality of reads; and

(f) determining sequences of the double-stranded DNA fragments in the sample using the plurality of reads obtained in (c), the plurality of physical UMI sequences identified in (d), and the plurality of virtual UMI sequences identified in (e) wherein (f) comprises:

5 (i) combining, for each double-stranded DNA fragment, a first plurality of reads each comprising a first physical UMI sequence of the plurality of physical UMI sequences and a first virtual UMI sequence of the plurality of UMI sequences to determine a consensus nucleotide sequence; and

10 (ii) determining a sequence of the double-stranded DNA fragment using the consensus nucleotide sequence.

2. The method of claim 1, wherein (f)(i) comprises:

combining the first plurality of reads and a second plurality of reads to determine the consensus nucleotide sequence, wherein the second plurality of reads each comprising a second physical UMI sequence and the first virtual UMI sequence.

15 3. The method of claim 1, wherein the plurality of physical UMIs includes no more than about 10,000 unique nonrandom UMIs; preferably no more than about 1,000 unique nonrandom UMIs; more preferably no more than about 500 unique nonrandom UMIs; more preferably no more than about 100 unique nonrandom UMIs; and most preferably about 96 unique nonrandom UMIs.

20 4. The method of claim 1, wherein the plurality of physical UMIs includes fewer than 12 nucleotides; preferably no more than 6 nucleotides; and more preferably no more than 4 nucleotides.

25 5. The method of claim 1, wherein the adapters each comprise a physical UMI on each strand of the adapters in the double-stranded hybridized region; optionally wherein the physical UMI is at or near an end of the double-stranded hybridized region, said end of the double-stranded hybridized region being opposite from the 3' arm or the 5' arm; optionally wherein the physical UMI is at said end of the double-stranded hybridized region, or is one nucleotide away from said end of the double-

stranded hybridized region; optionally wherein the adapters each comprise a 5'-TGG-3' trinucleotide or a 3'-ACC-5' trinucleotide on the double-stranded hybridized region adjacent to a physical UMI; optionally wherein the adapters each comprise a read primer sequence on each strand of the double-stranded hybridized region.

5 6. The method of claim 1, wherein the adapters each comprise a physical UMI on only one strand of the adapters on the single-stranded 5' arm or the single-stranded 3' arm.

7. The method of claim 6, wherein (f) comprises:

10 (i) collapsing reads having a same first physical UMI sequence into a first group to obtain a first consensus nucleotide sequence;

 (ii) collapsing reads having a same second physical UMI sequence into a second group to obtain a second consensus nucleotide sequence; and

15 (iii) determining, using the first and second consensus nucleotide sequences, a sequence of one of the double-stranded DNA fragments in the sample; optionally wherein (iii) comprises: (1) obtaining, using localization information and sequence information of the first and second consensus nucleotide sequences, a third consensus nucleotide sequence, and (2) determining, using the third consensus nucleotide sequence, the sequence of one of the double-stranded DNA fragments.

20 8. The method of claim 7, wherein (e) comprises identifying the plurality of virtual UMI sequence, while the adapters each comprise the physical UMI on only the single-stranded 5' arm or the single-stranded 3' arm; optionally wherein (f) comprises:

25 (i) combining reads having a first physical UMI sequence and at least one virtual UMI sequence in a read direction and reads having a second physical UMI sequence and the at least one virtual UMI sequence in the read direction to determine a consensus nucleotide sequence; and

(ii) determining a sequence of one of the double-stranded DNA fragments in the sample using the consensus nucleotide sequence.

9. The method of claim 1, wherein the adapters each comprise a physical UMI on each strand of the adapters in a double-stranded region of the adapters, wherein the physical UMI on one strand is complementary to the physical UMI on the other strand; optionally wherein (f) comprises:

(i) combining reads having a first physical UMI sequence, at least one virtual UMI sequence, and a second physical UMI sequence in the 5' to 3' direction and reads having the second physical UMI sequence, the at least one virtual UMI sequence, and the first physical UMI in the 5' to 3' direction to determine a consensus nucleotide sequence; and

(ii) determining a sequence of one of the double-stranded DNA fragments in the sample using the consensus nucleotide sequence.

10. The method of claim 1, wherein the adapters each comprise a first physical UMI on a 3' arm of the adapter and a second physical UMI on a 5' arm of the adapter, wherein the first physical UMI and the second physical UMI are not complementary to each other; optionally wherein (f) comprises:

(i) combining reads having a first physical UMI sequence, at least one virtual UMI sequence, and a second physical UMI sequence in the 5' to 3' direction and reads having a third physical UMI sequence, the at least one virtual UMI sequence, and a fourth physical UMI sequence in the 5' to 3' direction to determine a consensus nucleotide sequence; and

(ii) determining a sequence of one of the double-stranded DNA fragments in the sample using the consensus nucleotide sequence.

11. The method of claim 1, wherein at least some of the virtual UMIs derive from subsequences at or near the ends of the double-stranded DNA fragments in the sample; optionally wherein one or more physical UMIs and/or one or more virtual

UMIs are uniquely associated with a double-stranded DNA fragment in the sample; optionally wherein the double-stranded DNA fragments in the sample comprise more than about 1,000 DNA fragments; optionally wherein the plurality of virtual UMIs comprise UMIs of about 6 bp to about 24 bp, preferably of about 6 bp to about 10 bp.

- 5 12. The method of claim 1, wherein obtaining the plurality of reads in operation (c) comprises: obtaining two pair-end reads from each of the amplified polynucleotides, wherein the two pair-end reads comprise a long read and a short read, the long read being longer than the short read; optionally wherein (f) comprises:

10 combining read pairs comprising a first physical UMI sequence into a first group and combining read pairs comprising a second physical UMI sequence into a second group, wherein the first and the second physical UMI sequences are uniquely associated with a double-stranded fragment in the sample; and

15 determining the sequence of the double-stranded fragment in the sample using sequence information of long reads in the first group and sequence information of long reads in the second group; optionally wherein the long read has a read length of about 500 bp or more; optionally wherein the short read has a read length of about 50 bp or less.

- 20 13. The method of claim 1, wherein the amplified polynucleotides include an allele having an allele frequency lower than about 1%; optionally wherein the amplified polynucleotides include a cell free DNA molecule originating from a tumor, and the allele is indicative of the tumor.

14. A method for sequencing nucleic acid molecules from a sample, comprising

(a) attaching fork-shaped adapters to both ends of double-stranded DNA fragments in the sample,

- 25 wherein the fork-shaped adapters each comprise a double-stranded hybridized region, a single-stranded 5' arm, a single-stranded 3' arm, and a physical unique molecular index (UMI) on one strand or each strand of the adapter, the physical UMI is selected

from a plurality of physical UMIs, and wherein the plurality of physical UMIs comprises nonrandom UMIs, wherein every nonrandom UMI differs from every other nonrandom UMI of the adapters by at least two nucleotides at corresponding sequence positions of the nonrandom UMIs, wherein each double-stranded DNA fragment in the sample comprises a virtual UMI on one strand or each strand of the double-stranded DNA fragment, wherein the virtual UMI is a unique sub-sequence in a DNA fragment in the sample, and wherein a UMI is an oligonucleotide sequence that can be used to identify an individual molecule of a double-stranded DNA fragment in the sample;

10 (b) amplifying both strands of ligation products from (a), thereby obtaining a plurality of single-stranded, amplified polynucleotides;

(c) sequencing the plurality of amplified polynucleotides, thereby obtaining a plurality of reads each comprising a physical UMI sequence corresponding to a physical UMI on an adapter and a virtual UMI sequence corresponding to a virtual UMI on a double-stranded DNA fragment in the sample;

(d) identifying a plurality of physical UMI sequences for the plurality of reads;

(e) identifying a plurality of virtual UMI sequences for the plurality of reads; and;

20 (f) determining sequences of the double-stranded DNA fragments in the sample using the plurality of sequences obtained in (c) and the plurality of physical UMIs identified in (d), wherein (f) comprises:

(i) combining, for each double-stranded DNA fragment, a first plurality of reads each comprising a first physical UMI sequence of the plurality of physical UMI sequences to determine a consensus nucleotide sequence; and

(ii) determining a sequence of the double-stranded DNA fragment using the consensus nucleotide sequence.