CLAIMS

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

(a) applying adapters to DNA fragments in the sample to obtain DNA-adapter products, wherein

each adapter comprises a nonrandom unique molecular index,

nonrandom unique molecular indices of the adapters have at least two different molecular lengths and form a set of variable-length, nonrandom unique molecular indices (vNRUMIs), and

each nonrandom unique molecular index differs from every other nonrandom unique molecular index by at least a defined edit distance;

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

(c) sequencing the plurality of amplified polynucleotides, thereby obtaining a plurality of reads associated with the set of vNRUMIs;

(d) identifying, among the plurality of reads, reads associated with a same variable-length, nonrandom unique molecular index (vNRUMI), comprising:

obtaining, for each read of the plurality of reads, alignment scores with respect to the set of vNRUMIs, each alignment score indicating similarity between a subsequence of a read and a vNRUMI, wherein the subsequence is in a region of the read in which nucleotides derived from the vNRUMI are likely located;

selecting, for each read of the plurality of reads, at least one vNRUMI from the set of vNRUMIs based on the alignment scores; and

associating each read of the plurality of reads with the at least one vNRUMI selected for the read; and

(e) determining a sequence of a DNA fragment in the sample using the reads associated with the same vNRUMI.

2. The method of claim 1, wherein the alignment scores are based on matches of nucleotides and edits of nucleotides between the subsequence of the read and the vNRUMI.

3. The method of claim 2, wherein the edits of nucleotides comprise substitutions, additions, and deletions of nucleotides.

4. The method of claim 2, wherein each alignment score penalizes mismatches at the beginning of a subsequence but does not penalize mismatches at the end of the subsequence.

5. The method of claim 4, wherein obtaining an alignment score between a read and a vNRUMI comprises:

(a) calculating an alignment score between the vNRUMI and each one of all possible prefix sequences of the subsequence of the read;

(b) calculating an alignment score between the subsequence of the read and each one of all possible prefix sequences of the vNRUMI; and

(c) obtaining a largest alignment score among the alignment scores calculated in (a) and (b) as the alignment score between the read and the vNRUMI.

6. The method of claim 1, wherein the subsequence has a length that equals to a length of the longest vNRUMI in the set of vNRUMIs.

7. The method of claim 1, wherein selecting the at least one vNRUMI from the set of vNRUMIs comprises selecting a vNRUMI having a highest alignment score among the set of vNRUMIs.

8. The method of claim 1, wherein the at least one vNRUMI comprises two or more vNRUMIs.

9. The method of claim 8, further comprises selecting one of the two or more vNRUMI as the same vNRUMI of (d) and (e).

10. The method of claim 1, wherein the adapters applied in (a) are obtained by:

(i) providing a set of oligonucleotide sequences having at least two different molecular lengths;

(ii) selecting a subset of oligonucleotide sequences from the set of oligonucleotide sequences, all edit distances between oligonucleotide sequences of the subset of oligonucleotide sequences meeting a threshold value, the subset of oligonucleotide sequences forming the set of vNRUMIs; and

(iii) synthesizing the adapters each comprising a double-stranded hybridized region, a single-stranded 5' arm, a single-stranded 3' arm, and at least one vNRUMI of the set of vNRUMIs.

11. The method of claim 10, wherein the threshold value is 3.

12. The method of claim 1, wherein the set of vNRUMIs comprise vNRUMIs of 6 nucleotides and vNRUMIs of 7 nucleotides.

13. The method of claim 1, wherein (e) comprises collapsing reads associated with the same vNRUMI into a group to obtain a consensus nucleotide sequence for the sequence of the DNA fragment in the sample.

14. The method of claim 13, the consensus nucleotide sequence is obtained based partly on quality scores of the reads.

15. The method of claim 1, wherein (e) comprises:

identifying, among the reads associated with the same vNRUMI, reads having a same read position or similar read positions in a reference sequence, and

determining the sequence of the DNA fragment using reads that (i) are associated with the same vNRUMI and (ii) have the same read position or similar read positions in the reference sequence.

16. The method of claim 1, wherein the set of vNRUMIs includes no more than about 10,000 different vNRUMIs.

17. The method of claim 16, wherein the set of vNRUMIs includes no more than about 1,000 different vNRUMIs.

18. The method of claim 17, wherein the set of vNRUMIs includes no more than about 200 different vNRUMIs.

19. The method of claim 1, applying adapters to the DNA fragments in the sample comprises applying adapters to both ends of the DNA fragments in the sample.

20. A method for preparing sequencing adapters, comprising:

(a) providing a set of oligonucleotide sequences having at least two different molecular lengths;

(b) selecting a subset of oligonucleotide sequences from the set of oligonucleotide sequences, all edit distances between oligonucleotide sequences of the subset of oligonucleotide sequences meeting a threshold value, the subset of oligonucleotide sequences forming a set of variable-length, nonrandom unique molecular indexes (vNRUMIs), comprising:

(i) selecting an oligonucleotide sequence from the set of oligonucleotide sequences;

(ii) adding the selected oligonucleotide to an expanding set of oligonucleotide sequences and removing the selected oligonucleotide from the set of oligonucleotide sequences to obtain a reduced set of oligonucleotide sequences;

(iii) selecting an instant oligonucleotide sequence from the reduced set that maximizes a distance function, wherein the distance function is a minimal edit distance between the instant oligonucleotide sequence and any oligonucleotide sequences in the expanding set, and wherein the distance function meeting the threshold value;

(iv) adding the instant oligonucleotide to the expanding set and removing the instant oligonucleotide from the reduced set;

(v) repeating (iii) and (iv) one or more times; and

(vi) providing the expanding set as the subset of oligonucleotide sequences forming the set of vNRUMIs; and

(c) synthesizing a plurality of sequencing adapters, wherein each sequencing adapter comprises a double-stranded hybridized region, a single-stranded 5' arm, a single-stranded 3' arm, and at least one vNRUMI of the set of vNRUMIs.

21. The method of claim 20, wherein (v) comprises repeating (iii) and (iv) until the distance function no longer meets the threshold value.

22. The method of claim 20, wherein (v) comprises repeating (iii) and (iv) until the expanding set reaches a defined size.

23. The method of claim 20, wherein the instant oligonucleotide sequence or an oligonucleotide sequence in the expanding set is shorter than a longest oligonucleotide sequence in the set of oligonucleotide sequences, the method further comprising, before (iii), (1) appending a thymine base or a thymine base plus any of four bases to the instant oligonucleotide sequence or the oligonucleotide sequence in the expanding set, thereby generating a padded sequence having the same length as the longest oligonucleotide sequence in the set of oligonucleotide sequences, and (2) using the padded sequence to calculate the minimal edit distance.

24. The method of claim 20, wherein the edit distances are Levenshtein distances.

25. The method of claim 20, wherein the threshold value is three.

26. The method of claim 20, further comprising, before (b), removing certain oligonucleotide sequences from the set of oligonucleotide sequences to obtain a filtered set of oligonucleotide sequences; and providing the filtered set of oligonucleotide sequences as the set of oligonucleotide sequences from which the subset is selected.

27. The method of claim 26, wherein the certain oligonucleotide sequences comprise oligonucleotide sequences having three or more consecutive identical bases.

28. The method of claim 26, wherein the certain oligonucleotide sequences comprise oligonucleotide sequences having a combined number of guanine and cytosine bases smaller than 2 and oligonucleotide sequences having a combined number of guanine and cytosine bases larger than 4.

29. The method of claim 26, wherein the certain oligonucleotide sequences comprise oligonucleotide sequences having a same base at the last two positions.

30. The method of claim 26, wherein the certain oligonucleotide sequences comprise oligonucleotide sequences having a subsequence matching the 3' end of one or more sequencing primers.

31. The method of claim 26, wherein the certain oligonucleotide sequences comprise oligonucleotide sequences having a thymine base at the last position of the oligonucleotide sequences.

32. The method of claim 20, wherein the set of vNRUMIs comprises vNRUMIs of 6 nucleotides and vNRUMIs of 7 nucleotides.

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

(a) applying adapters to DNA fragments in the sample to obtain DNA-adapter products,

wherein each adapter comprises a nonrandom unique molecular index, and

wherein nonrandom unique molecular indices of the adapters have at least two different molecular lengths and form a set of variable-length, nonrandom unique molecular indices (vNRUMIs);

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

(c) sequencing the plurality of amplified polynucleotides, thereby obtaining a plurality of reads associated with the set of vNRUMIs;

(d) identifying, among the plurality of reads, reads associated with a same variable-length, nonrandom unique molecular index (vNRUMI), by obtaining, for each read of the plurality of reads, alignment scores with respect to the set of vNRUMIs, each alignment score indicating similarity between a subsequence of a read and a vNRUMI,

wherein the subsequence is in a region of the read in which nucleotides derived from the vNRUMI are likely located, and

wherein the subsequence has a length that equals to a length of the longest vNRUMI in the set of vNRUMIs; and

(e) determining a sequence of a DNA fragment in the sample using the reads associated with the same vNRUMI.

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

(a) applying adapters to DNA fragments in the sample to obtain DNA-adapter products,

wherein each adapter comprises a nonrandom unique molecular index, and

wherein nonrandom unique molecular indices of the adapters have at least two different molecular lengths and form a set of variable-length, nonrandom unique molecular indices (vNRUMIs);

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

(c) sequencing the plurality of amplified polynucleotides, thereby obtaining a plurality of reads associated with the set of vNRUMIs;

(d) identifying, among the plurality of reads, reads associated with a same variable-length, nonrandom unique molecular index (vNRUMI), by obtaining, for each read of the plurality of reads, alignment scores with respect to the set of vNRUMIs, each alignment score indicating similarity between a subsequence of a read and a vNRUMI, wherein the subsequence is in a region of the read in which nucleotides derived from the vNRUMI are likely located, comprising:

(i) calculating an alignment score between the vNRUMI and each one of all possible prefix sequences of the subsequence of the read,

(ii) calculating an alignment score between the subsequence of the read and each one of all possible prefix sequences of the vNRUMI, and

(iii) obtaining a largest alignment score among the alignment scores calculated in (i) and (ii) as the alignment score between the read and the vNRUMI;

(e) determining a sequence of a DNA fragment in the sample using the reads associated with the same vNRUMI.