CLAIMS

1. A system for determining a nucleotide sequence from nucleotide sequencing reads, comprising:

a non-transitory memory configured to store executable instructions and a first hash data structure for storing nucleotide sequencing reads in a plurality of bins; and

a hardware processor programmed by the executable instructions to perform:

receiving a plurality of first nucleotide sequencing reads;

for each first nucleotide sequencing read:

generating a plurality of first identifier subsequences from a first identifier sequence of the first nucleotide sequencing read that are consecutive subsequences of the first identifier sequence;

for each first identifier subsequence, determining a plurality of hashes using a plurality of hash functions;

generating a first signature for the first nucleotide sequencing read comprising a plurality of first signature hashes each for a different first position, wherein each first signature hash is the minimum of the hashes determined for the plurality of first identifier subsequences at the first position; and

assigning the first nucleotide sequencing read to at least one first particular bin of the first hash data structure based on the first signature; and

determining a nucleotide sequence for each first particular bin of the first hash data structure with one or more first nucleotide sequencing reads assigned.

2. The system of claim 1, wherein assigning the first nucleotide sequencing read comprises:

determining a plurality of subsequences of the first signature from the first signature of the first nucleotide sequencing read; and

assigning the first nucleotide sequencing read to a first particular bin of each first hash data structure of a plurality of first hash data structures based on a subsequence of the first signature.

3. The system of claim 1, wherein assigning the first nucleotide sequencing read comprises:

determining a plurality of subsequences of the first signature from the first signature of the first nucleotide sequencing read; and

assigning the first nucleotide sequencing read to a plurality of first particular bins of the first hash data structure based on the plurality of subsequences of the first signature.

4. The system of claim 1, wherein the first particular bin is an existing bin of the first hash data structure, and wherein an alignment score of the first nucleotide sequencing read and another first nucleotide sequencing read assigned to the first particular bin of the first hash data structure is above an alignment score threshold.

5. The system of claim 1, wherein the first particular bin is an existing bin of the first hash data structure, and wherein the highest alignment score of the first nucleotide sequencing read and any first nucleotide sequencing read assigned to the first particular bin of the first hash data structure is above an alignment score threshold.

6. The system of claim 1, wherein the first particular bin is a new bin of the first hash data structure, and wherein an alignment score of the first nucleotide sequencing read and any first nucleotide sequencing read assigned to any existing bin of the first hash data structure is below an alignment score threshold.

7. The system of claim 1, wherein the first signature matches a key of the first particular bin of the first hash data structure.

8. The system of claim 1, wherein the first signature and the key of the first particular bin of the first hash data structure are identical.

9. The system of claim 1, wherein each first nucleotide sequencing read is associated with a second nucleotide sequencing read, and wherein the first nucleotide sequencing read and the second nucleotide sequencing read are paired-end nucleotide sequencing reads.

10. The system of claim 1, wherein determining the nucleotide sequence comprises determining a consensus sequence of the one or more first nucleotide sequencing reads assigned to the first particular bin.

11. The system of claim 10, wherein determining the consensus sequence comprises determining a first nucleotide sequencing read with a highest quality score assigned to the first particular bin as the consensus sequence of the first particular bin.

12. The system of claim 1, wherein determining the nucleotide sequence comprises selecting a sequence of the one or more first nucleotide sequencing reads assigned to the first particular bin as a representative sequence of the first particular bin.

13. The system of claim 1, wherein determining the nucleotide sequence comprises determining an alignment score of two of the one or more first nucleotide sequencing reads assigned to the first particular bin is above an alignment score threshold.

14. The system of claim 1, wherein the plurality of nucleotide sequencing reads is associated with an identical physical identifier sequence.

15. The system of claim 1, wherein the plurality of nucleotide sequencing reads is not associated any physical identifier sequence.

16. A computer-implemented method for determining a nucleotide sequence from nucleotide sequencing reads, comprising:

receiving a plurality of first nucleotide sequencing reads;

for each first nucleotide sequencing read:

generating a plurality of first identifier subsequences from a first identifier sequence of the first nucleotide sequencing read;

for each first identifier subsequence, determining a plurality of hashes using a plurality of hash functions;

generating a first signature for the first nucleotide sequencing read comprising a plurality of first signature hashes each for a different first position, wherein each first signature hash is selected from the hashes determined for the plurality of first identifier subsequences at the first position of the first signature hash; and

assigning the first nucleotide sequencing read to a first particular bin of a first data structure based on the first signature; and

determining a nucleotide sequence for each first particular bin of the first data structure with one or more first nucleotide sequencing reads assigned.

17. The method of claim 16, wherein generating the plurality of first identifier subsequences comprises generating a plurality of *k*-mers from the first identifier sequence of the sequencing read.

18. The method of claim 16, wherein the subsequence comprises a nucleotide insertion, a nucleotide deletion, a nucleotide substitution, or a combination thereof.

19. The method of claim 16, wherein two consecutive first identifier subsequences overlap.

20. The method of claim 16, wherein the plurality of first identifier subsequences comprises a plurality of 4-mers, and wherein the first identifier sequence comprises about 25 nucleotides.

21. The method of claim 17, wherein the first identifier sequence is a subsequence of the first nucleotide sequencing read.

22. The method of claim 16, wherein the first data structure comprises a hash table.

23. A system for identifying similar nucleotide sequencing reads, comprising:

non-transitory memory configured to store: executable instructions, a first hash data structure and a second hash data structure for storing a plurality of pairs of sequencing reads; and

a hardware processor programmed by the executable instructions to perform:

receiving a pair of a first query nucleotide sequencing read and a second query nucleotide sequencing read, wherein the first query nucleotide sequencing read and the second query nucleotide sequencing read are paired-end nucleotide sequencing reads;

generating (i) a plurality of first query identifier subsequences from the first query nucleotide sequencing read, and (ii) a plurality of second query identifier subsequences from the second query nucleotide sequencing read;

determining (i) for each first query identifier subsequence, a plurality of first query hashes, and (ii) for each second query identifier subsequence, a plurality of second query hashes, using a plurality of hash functions;

generating a first query signature for the first query nucleotide sequencing read comprising a plurality of first query signature hashes each for a different first position, and (ii) a second query signature for the second query nucleotide sequencing read comprising a plurality of second query signature hashes each for a different second position, wherein each first query signature hash is selected from the first hashes determined for the plurality of first query identifier subsequences at the first position of the first query signature hash, and wherein each second query signature hash is selected from the second hashes determined for the plurality of second query identifier subsequences at the second position of the second query signature hash;

retrieving (i) one or more first stored pairs from the first hash data structure using the first query signature, and (ii) one or more second stored pairs from the second hash data structure using the second query signature, wherein each first stored pair comprises a first stored nucleotide sequencing read and a second stored nucleotide sequencing read that are paired-end nucleotide sequencing reads, and wherein each second stored pair comprises a first stored nucleotide sequencing read and a second stored nucleotide sequencing read that are paired-end nucleotide sequencing reads; and

determining, for each pair of a first stored nucleotide sequencing read and a second stored nucleotide sequencing read present in both the first stored pairs and second stored pairs, the first

stored nucleotide sequencing read and the second stored nucleotide sequencing read of the pair as being similar to the first query sequencing read and the second query sequencing read, respectively.

24. The system of claim 23, wherein each pair of stored sequencing reads comprises a first stored nucleotide sequencing read and a second stored nucleotide sequencing read, wherein each pair of stored sequencing reads is assigned to one of a plurality of first bins of the first hash data structure based on a first signature of a first stored nucleotide sequencing read of the pair generated by hashing first identifier subsequences of a first identifier sequence of the first stored nucleotide sequencing read, and wherein each pair of stored sequencing reads is assigned to one of a plurality of second bins of the second hash data structure based on a second signature of a second stored nucleotide sequencing read of the pair generated by hashing second identifier subsequences of a second identifier sequence of the second nucleotide sequencing read.

25. The system of claim 23, wherein the hardware processor is programmed by the executable instructions to perform:

for each pair of sequencing reads:

generating a plurality of first identifier subsequences from a first identifier sequence of the first nucleotide sequencing read of the pair of sequencing reads;

generating a first signature for the first nucleotide sequencing read by applying the plurality of hashing functions to the plurality of first identifier subsequences; and

assigning the pair of sequencing reads to at least one first particular bin of the first hash data structure based on the first signature; and

determining a first nucleotide sequence and a second nucleotide sequence for each first particular bin of the first hash data structure with one of more pairs assigned to the particular bin of the first hash data structure.

26. The system of claim 23, wherein each pair of sequencing reads is associated with a first identifier sequence and a second identifier sequence, and wherein the hardware processor is programmed by the executable instructions to perform:

determining the first identifier sequence and the second identifier sequence of a first pair of sequencing reads and the second identifier sequence and the first identifier sequence of a second pair of sequencing reads are identical, respectively; and

determining a first nucleotide sequence and a second nucleotide sequence for the first pair of sequencing reads and the second pair of sequencing reads.

27. A method for identifying similar nucleotide sequencing reads, comprising:

receiving a first query nucleotide sequencing read;

generating a plurality of first query identifier subsequences from the first query nucleotide sequencing read;

for each first query identifier subsequence, determining a plurality of first query hashes using a plurality of hash functions;

generating a first query signature for the first nucleotide sequencing read comprising a plurality of first query signature hashes each for a different first position, wherein each first query signature hash is selected from the first hashes determined for the plurality of first query identifier subsequences at the first position of the first query signature hash; and

retrieving one or more first stored nucleotide sequencing reads from a first hash data structure using the first query signature, wherein each of the first stored nucleotide sequencing reads is similar to the first query nucleotide sequencing read.

28. The method of claim 27, wherein receiving the first query nucleotide sequencing read comprises receiving a pair of the first query nucleotide sequencing read and a second query nucleotide sequencing read, wherein generating the plurality of first query identifier subsequences comprises generating a plurality of second query identifier subsequences from the second nucleotide sequencing read, wherein for each first query identifier subsequence, determining a plurality of first query hashes using a plurality of hash functions comprises for each second query identifier subsequence, determining a plurality of second query hashes using the plurality of hash functions, wherein generating the first query signature comprises generating a second query signature for the second nucleotide sequencing read comprising a plurality of second query signature hashes each for a different second position, wherein each second query signature hash is selected from the second hashes determined for the plurality of second query identifier subsequences at the second position of the second query signature hash, and wherein retrieving one or more first stored nucleotide sequencing reads comprises retrieving one or more first stored pairs from the first hash data structure, wherein each of the first stored pairs comprises a first stored nucleotide sequencing read and a second stored nucleotide sequencing

read similar to the first query nucleotide sequencing read and the second query nucleotide sequencing read, respectively.

29. The system of claim 9, wherein assigning the first nucleotide sequencing read comprises assigning a pair of sequencing reads comprising the first nucleotide sequencing read and the second nucleotide sequencing read to the first particular bin of the first hash data structure based on the first signature.