CLAIMS

1. A method for compressing a genomic data read based on a mapping of the genomic data read to a reference genomic sequence, the method comprising:

(a) obtaining, by one or more computers, a first k-mer seed from the genomic data read;

(b) generating, by the one or more computers, a hash value representing a genomic signature by applying a hash function to the first k-mer seed;

(c) determining, by the one or more computers, a reference sequence location that matches at least a portion of the first k-mer seed using a hash data structure, wherein the hash data structure comprises N data cells and wherein a first data cell of the N data cells includes (i) a first portion storing a predetermined genomic signature derived from the portion of the first k-mer seed and (ii) a second portion storing a value that corresponds to a location within the reference genomic sequence that matches at least the portion of the first k-mer seed;

(d) determining, by the one or more computers, a number of mismatches based on comparing genomic data of the genomic data read to genomic data of the reference genomic sequence;

(e) determining, by the one or more computers, whether the number of mismatches satisfies a threshold number of mismatches;

continuing to perform operations corresponding to (a)-(e) for an additional k-mer seed at each iteration, wherein the additional k-mer seed corresponds to a portion of the genomic data read different than the first k-mer seed, until a determination is made that the number of mismatches for the additional k-mer seed satisfies a second mismatch threshold;

determining that the number of mismatches for the additional k-mer seed satisfies the second mismatch threshold;

based on a determination that the number of mismatches satisfies the second mismatch threshold, terminating the continued performance of the operations corresponding to (a)-(e) prior to a subsequent pass;

selecting an actual alignment for the genomic data read based on the additional k-mer seed;

using the selected actual alignment to compress the genomic data read; and

storing the compressed genomic data read in computer memory.

2. The method of claim 1, wherein obtaining the first k-mer seed from the genomic data read comprises:

obtaining a first k bases of the genomic data read as the first k-mer seed.

3. The method of claim 1, comprising obtaining a set of k-mer seeds, that include the additional k-mer seed, from the genomic data read, wherein selecting the actual alignment for the genomic data read comprises:

generating genomic signatures for the set of k-mer seeds; and

selecting a subset of the set of k-mer seeds based on the second genomic signatures.

4. The method of claim 3, wherein selecting the subset of the set of k-mer seeds based on the second genomic signatures comprises:

performing one or more modulo operations including a modulo operation on each genomic signature of the genomic signatures; and

selecting the subset of the set of k-mer seeds based on results of the one or more modulo operations and a predetermined criterion.

5. The method of claim 1, wherein continuing to perform the operations corresponding to (a)-(e) comprises:

determining, by the one or more computers, a reference sequence location for each k-mer seed of the set of k-mer seeds, that include the additional k-mer seed, that matches at least a portion of a given k-mer seed using the hash data structure; and

generating a candidate location list including a reference sequence location for each k-mer seed of the set of k-mer seeds.

6. The method of claim 5, comprising:

sorting the reference sequence locations of the candidate location list according to a number of k-mer seeds paired with the given reference sequence location in the hash data structure; and

determining a number of mismatches for each of the reference sequence locations of the sorted candidate location list compared to the reference genomic sequence in an order of the sorted candidate location list.

7. The method of claim 6, comprising:

determining a number of mismatches for a first candidate location of the reference sequence locations of the sorted candidate location list satisfies the second mismatch threshold; and

selecting the first candidate location as the actual alignment.

8. The method of claim 7, wherein the second mismatch threshold is different than the threshold number of mismatches.

9. The method of claim 1, comprising:

generating second genomic signatures for each k-mer seed of an obtained set of k-mer seeds including the additional k-mer seed;

performing second one or more modulo operations including a modulo operation on each genomic signature of the second genomic signatures; and

selecting a subset of the set of k-mer seeds based on results of the second one or more modulo operations and a predetermined criterion.

10. The method of claim 9, comprising:

determining a reference sequence location for each k-mer seed of the subset that matches at least a portion of a given k-mer seed using the hash data structure; and

generating a second candidate location list including a reference sequence location for each k-mer seed of the subset.

11. The method of claim 10, comprising:

sorting the reference sequence locations of the second candidate location list according to a number of k-mer seeds paired with the given reference sequence location in the hash data structure; and

determining a number of mismatches for each of the reference sequence locations of the sorted second candidate location list compared to the reference genomic sequence in an order of the sorted second candidate location list.

12. The method of claim 11, comprising:

determining a number of mismatches for a second candidate location of the reference sequence locations of the sorted second candidate location list satisfies the second mismatch threshold; and

selecting the second candidate location as the actual alignment.

13. A non-transitory computer-readable medium storing one or more instructions executable by a computer system to perform operations of one or more of claims 1-12.

14. A system, comprising:

one or more processors; and

machine-readable media interoperably coupled with the one or more processors and storing one or more instructions that, when executed by the one or more processors, perform operations of one or more of claims 1-12.