

CLAIMS

What is claimed is:

~~1. A method for software-accelerated genomic data read mapping, the method comprising:~~

~~obtaining, by one or more computers, a k-mer seed from a genomic data read;~~

~~generating, by the one or more computers, a genomic signature based on the obtained k-mer seed;~~

~~determining, by one or more computers, a reference sequence location that match at least a portion of the k-mer seed using a hash data structure, wherein the hash data structure comprises N data cells comprising a first portion storing a predetermined genomic signature and a second portion storing a value that corresponds to a first occurrence of a reference sequence location that match at least a portion of the k-mer seed from which the predetermined genomic signature was derived; and~~

~~selecting, by one or more computers, the determined reference sequence location as an actual alignment for the obtained k-mer seed based on one or more alignment scores.~~

~~2. The method of claim 1, wherein the predetermined genomic signature occupies one byte of memory storage.~~

~~3. The method of any preceding claim, wherein the value occupies four bytes of memory storage.~~

~~4. The method of any preceding claim, wherein the hash data structure is a single array of N data cells.~~

~~5. The method of any one of the preceding claims, further comprising:~~

~~filtering, by one or more computers, the genomic data read based on a first set of values corresponding to one or more k-mer seeds of the genomic data read.~~

6.——~~The method of claim 5, wherein the first set of values comprises a result of a predetermined operation applied to the one or more k-mer seeds of the genomic data read, and wherein the first set of values is used to obtain the k-mer seed from the genomic data read.~~

7.——~~The method of claim 6, wherein the predetermined operation comprises generating a hash value based on the one or more k-mer seeds of the genomic data read and a hash function.~~

8.——~~The method of any one of the preceding claims, wherein determining the reference sequence locations comprises:~~

~~computing, by one or more computers, a first position for the k-mer seed of the genomic data read, wherein the first position corresponds to a location of the k-mer seed within the genomic data read; and~~

~~computing, by one or more computers, a second position for the k-mer seed, wherein the second position corresponds to a location of the k-mer seed within the reference genomic data, and wherein the second position is computed based on the hash data structure.~~

9.——~~The method of any one of the preceding claims, further comprising:~~

~~sorting, by one or more computers, the one or more reference sequence locations based on the hash data structure and the genomic data read.~~

10.——~~The method of any one of the preceding claims, further comprising:~~

~~generating, by one or more computers, the one or more alignment scores based on sorting the one or more reference sequence locations.~~

11.——~~The method of any one of the preceding claims, wherein selecting at least one of the determined reference sequence locations as the actual alignment for the obtained k-mer seed comprises comparing the one or more alignment scores to a threshold value.~~

12.——~~The method of any one of the preceding claims, wherein the one or more alignment scores comprises a numerical value representing a number of mismatches~~

~~between the obtained k-mer seed from the genomic data read and the reference sequence location.~~

~~13.—The method of claim of any one of the preceding claims, wherein each subsequent occurrences of a reference sequence location that matches at least a portion of the k-mer seed from which the predetermined genomic signature was derived after first occurrence is discarded.~~

~~14.—A system for software-accelerated genomic data read mapping, the method comprising:~~

~~one or more computers; and~~

~~one or more memory devices storing instructions that, when executed by the one or more computers, cause the one or more computers to perform operations, the operations comprising:~~

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

~~generating, by the one or more computers, a genomic signature based on the obtained k-mer seed;~~

~~determining, by the one or more computers, a reference sequence location that match at least a portion of the k-mer seed using a hash data structure, wherein the hash data structure comprises N data cells comprising a first portion storing a predetermined genomic signature and a second portion storing a value that corresponds to a first occurrence of a reference sequence location that match at least a portion of the k-mer seed from which the predetermined genomic signature was derived; and~~

~~selecting, by the one or more computers, the determined reference sequence location as an actual alignment for the obtained k-mer seed based on one or more alignment scores.~~

~~15.—The system of claim 14, wherein the predetermined genomic signature occupies one byte of memory storage.~~

~~16.—The system of any one of the preceding claims 14 or 15, wherein the value occupies four bytes of memory storage.~~

~~17.—The system of any one of the preceding claims 14 to 16, wherein the hash data structure is a single array of N data cells.~~

~~18.—The system of any one of the preceding claims 14 to 17, the operations further comprising:~~

~~filtering, by the one or more computers, the genomic data read based on a first set of values corresponding to one or more k-mer seeds of the genomic data read.~~

~~19.—The system of claim 18, wherein the first set of values comprises a result of a predetermined operation applied to the one or more k-mer seeds of the genomic data read, and wherein the first set of values is used to obtain the k-mer seed from the genomic data read.~~

~~20.—The system of claim 19, wherein the predetermined operation comprises generating a hash value based on the one or more k-mer seeds of the genomic data read and a hash function.~~

~~21.—The system of any one of the preceding claims 14 to 20, wherein determining the reference sequence locations comprises:~~

~~computing, by the one or more computers, a first position for the k-mer seed of the genomic data read, wherein the first position corresponds to a location of the k-mer seed within the genomic data read; and~~

~~computing, by the one or more computers, a second position for the k-mer seed, wherein the second position corresponds to a location of the k-mer seed within the reference genomic data, and wherein the second position is computed based on the hash data structure.~~

~~22.—The system of any one of the preceding claims 14 to 21, wherein the operations further comprise:~~

~~sorting, by the one or more computers, the one or more reference sequence locations based on the hash data structure and the genomic data read.~~

~~23.—The system of any one of the preceding claims 14 to 22, the operations further comprising:~~

~~generating, by one or more computers, the one or more alignment scores based on sorting the one or more reference sequence locations.~~

~~24.—The system of any one of the preceding claims 14 to 23, wherein selecting at least one of the determined reference sequence locations as the actual alignment for the obtained k-mer seed comprises comparing the one or more alignment scores to a threshold value.~~

~~25.—The system of any one of the preceding claims 14 to 24, wherein the one or more alignment scores comprises a numerical value representing a number of mismatches between the obtained k-mer seed from the genomic data read and the reference sequence location.~~

~~26.—The system of claim one of any of the preceding claims 14 to 25, wherein each subsequent occurrences of a reference sequence location that matches at least a portion of the k-mer seed from which the predetermined genomic signature was derived after first occurrence is discarded.~~

~~27.—A computer-readable medium storing instructions that, when executed by one or computers, cause the one or more computers to perform operations for software-accelerated genomic data read mapping, the operations comprising:~~

~~obtaining a k-mer seed from a genomic data read;~~

~~generating a genomic signature based on the obtained k-mer seed; determining a reference sequence location that match at least a portion of the k-mer seed using a hash data structure, wherein the hash data structure comprises /V data cells comprising a first portion storing a predetermined genomic signature and a second portion storing a value that corresponds to a first occurrence of a reference sequence location that match at least a portion of the k-mer seed from which the predetermined genomic signature was derived; and~~

~~selecting the determined reference sequence location as an actual alignment for the obtained k-mer seed based on one or more alignment scores.~~

~~28.—The computer-readable medium of claim 27, wherein the predetermined genomic signature occupies one byte of memory storage.~~

~~29.—The computer-readable medium of any one of the preceding claims 27 or 28, wherein the value occupies four bytes of memory storage.~~

~~30.—The computer-readable medium of any one of the preceding claims 27 to 29, wherein the hash data structure is a single array of N data cells.~~

~~31.—The computer-readable medium of any one of the preceding claims 27 to 30, the operations further comprising:~~

~~filtering the genomic data read based on a first set of values corresponding to one or more k-mer seeds of the genomic data read.~~

~~32.—The computer-readable medium of claim 31, wherein the first set of values comprises a result of a predetermined operation applied to the one or more k-mer seeds of the genomic data read, and wherein the first set of values is used to obtain the k-mer seed from the genomic data read.~~

~~33.—The computer-readable medium of any one of the preceding claims 27 to 32, wherein determining the reference sequence locations comprises:~~

~~computing a first position for the k-mer seed of the genomic data read, wherein the first position corresponds to a location of the k-mer seed within the genomic data read; and~~

~~computing a second position for the k-mer seed, wherein the second position corresponds to a location of the k-mer seed within the reference genomic data, and wherein the second position is computed based on the hash data structure.~~

~~34.—The computer-readable medium of any one of the preceding claims 27 to 33, wherein the operations further comprise:~~

~~sorting the one or more reference sequence locations based on the hash data structure and the genomic data read.~~

~~35.—The computer-readable medium of any one of the preceding claims 27 to 34, the operations further comprising:~~

~~generating the one or more alignment scores based on sorting the one or more reference sequence locations.~~

~~36.—The computer-readable medium of any one of the preceding claims 27 to 35, wherein selecting at least one of the determined reference sequence locations as the actual alignment for the obtained k-mer seed comprises comparing the one or more alignment scores to a threshold value.~~

~~37.—The computer-readable medium of any one of the preceding claims 27 to 36, wherein the one or more alignment scores comprises a numerical value representing a number of mismatches between the obtained k-mer seed from the genomic data read and the reference sequence location.~~

~~38.—The computer-readable medium of claim of any one of the preceding claims 27 to 37, wherein each subsequent occurrences of a reference sequence location that matches at least a portion of the k-mer seed from which the predetermined genomic signature was derived after first occurrence is discarded.~~

Claim 139. A method for generating a hash table for software-accelerated genomic data read mapping, the method comprising:

receiving, by one or more computers, genomic data, wherein the genomic data is derived from parent genomic data;

generating, by ~~the~~ one or more computers, a first set of values based on the genomic data;

generating, by one or more computers, a subset of the genomic data based on the first set of values;

computing, by one or more computers, a signature for each k-mer of the subset of the genomic data, wherein the signature is computed based on a first hash function;

computing, by one or more computers, a first attribute for each k-mer of the subset of the genomic data, wherein the first attribute comprises a position of a given k-mer of the genomic data within a sequence of the genomic data;

computing, by one or more computers, an index for each k-mer of the subset of the genomic data, wherein the index is computed based on a second hash function; and

storing, by one or more computers, the signature and the first attribute for each k-mer of the subset of the genomic data within a hash data structure based on the index for each k-mer of the subset of the genomic data.

Claim 240. The method of claim 391, wherein each k-mer of the subset of the genomic data is a k-mer comprising k letters representing a string of one or more nucleotides.

Claim 413. The method of any one of claims 39-1 or 402, wherein the first set of values comprises a representation of a number of times that a given k-mer of the genomic data occurs within the parent genomic data.

Claim 442. The method of any one of the preceding claims 39-1 to 41-3, wherein the first set of values comprises a representation of a hash value computed based on a corresponding k-mer of the genomic data.

Claim 543. The method of any one of the preceding claims 39-1 to 424, wherein a memory allocation size used to store the signature for a given k-mer of the subset is smaller than a memory allocation size used to store the given k-mer.

Claim 644. The method of any one of the preceding claims 39-1 to 434, the method further comprising:

sending, by ~~the~~ one or more computers, data corresponding to the hash data structure as a data package to a first device.

Claim 745. The method of claim 446, wherein the first device is a memory storage device.

Claim 846. The method of claim 44-6 or claim 457, wherein a second device reads the data corresponding to the hash data structure from the first device, and wherein the second device performs a series of operations to generate a second hash data structure based on the data corresponding to the hash data structure.

Claim 947. A system for generating a hash table for software-accelerated genomic data read mapping, the system comprising:

one or more computers; and

one or more memories storing instructions that, when executed by the one or more computers, cause the one or more computers to perform operations, the operations comprising:

receiving, by ~~the~~ one or more computers, genomic data, wherein the genomic data is derived from parent genomic data;

generating, by the one or more computers, a first set of values based on the genomic data;

generating, by the one or more computers, a subset of the genomic data based on the first set of values;

computing, by the one or more computers, a signature for each k-mer of the subset of the genomic data, wherein the signature is computed based on a first hash function;

computing, by the one or more computers, a first attribute for each k-mer of the subset of the genomic data, wherein the first attribute comprises a position of a given k-mer of the genomic data within a sequence of the genomic data;

computing, by the one or more computers, an index for each k-mer of the subset of the genomic data, wherein the index is computed based on a second hash function; and

storing, by the one or more computers, the signature and the first attribute for each k-mer of the subset of the genomic data within a hash data structure based on the index for each k-mer of the subset of the genomic data.

Claim 1048. The system of claim 479, wherein each k-mer of the subset of the genomic data is a k-mer comprising k letters representing a string of one or more nucleotides.

Claim 1149. The system of any one of the preceding claims 47-9 or 4810, wherein the first set of values comprises a representation of a number of times that a given k-mer of the genomic data occurs within the parent genomic data.

Claim 1250. The system of any one of the preceding claims 47-9 to 4911, wherein the first set of values comprises a representation of a hash value computed based on a corresponding k-mer of the genomic data.

Claim 1354. The system of any one of the preceding claims 47-9 to 5012, wherein a memory allocation size used to store the signature for a given k-mer of the subset is smaller than a memory allocation size used to store the given k-mer.

Claim 1452. The system of any one of the preceding claims 47-9 to 51-13, wherein the operations further comprise:

sending, by ~~the~~ one or more computers, data corresponding to the hash data structure as a data package to a first device.

Claim 1553. The system of claim 5214, wherein the first device is a memory storage device.

Claim 1654. The system of claim 52-14 or claim 5315, wherein a second device reads the data corresponding to the hash data structure from the first device, and wherein the second device performs a series of operations to generate a second hash data structure based on the data corresponding to the hash data structure.

Claim 1755. A computer-readable medium storing instructions that, when executed by one or more computers, cause the one or more computers to perform operations for generating a hash table for software-accelerated genomic data read mapping, the operations comprising:

receiving genomic data, wherein the genomic data is derived from parent genomic data;

generating a first set of values based on the genomic data;

generating a subset of the genomic data based on the first set of values; computing a signature for each k-mer of the subset of the genomic data, wherein the signature is computed based on a first hash function;

computing a first attribute for each k-mer of the subset of the genomic data, wherein the first attribute comprises a position of a given k-mer of the genomic data within a sequence of the genomic data;

computing an index for each k-mer of the subset of the genomic data, wherein the index is computed based on a second hash function; and

storing the signature and the first attribute for each k-mer of the subset of the genomic data within a hash data structure based on the index for each k-mer of the subset of the genomic data.

Claim 1856. The computer-readable medium of claim 5517, wherein each k-mer of the subset of the genomic data is a k-mer comprising k letters representing a string of one or more nucleotides.

Claim 1957. The computer-readable medium of claim 55-17 or claim 5618, wherein the first set of values comprises a representation of a number of times that a given k-mer of the genomic data occurs within the parent genomic data.

Claim 2058. The computer-readable medium of any one of the preceding claims 55-17 to 5719, wherein the first set of values comprises a representation of a hash value computed based on a corresponding k-mer of the genomic data.

Claim 2159. The computer-readable medium of any one of the preceding claims 55-17 to 5820, wherein a memory allocation size used to store the signature for a given k-mer of the subset is smaller than a memory allocation size used to store the given k-mer.

Claim 2260. The computer-readable medium of any one of the preceding claims 55-17 to 5920, the operations further comprising:

sending data corresponding to the hash data structure as a data package to a first device.

Claim 2364. The computer-readable medium of claim 6022, wherein the first device is a memory storage device.

Claim 2462. The computer-readable medium of claim 60-22 or claim 6423, wherein a second device reads the data corresponding to the hash data structure from the first device, and wherein the second device performs a series of operations to generate a second hash data structure based on the data corresponding to the hash data structure.

