

Claims

What is claimed is:

1. A computer-implemented method for the compression of genome sequence data produced
5 by a sequencing machine, said genome sequence data comprising reads of sequences of
nucleotides or bases that have been aligned to a reference sequence, thereby creating aligned
reads, said aligned reads being stored as a list of reads in an initial file, said method comprising:
for each aligned read, determining whether said read is perfectly or imperfectly mapped
with said reference sequence or whether said read is unmapped with said reference sequence,
10 encoding the reads according to said determination, wherein the reads that are determined
to be perfectly mapped are encoded according to a first encoding process and the reads that are
determined to be unmapped are encoded according to a second encoding process,
wherein the determining step comprises comparing, for each imperfectly mapped read, the
number of mismatches between said read and said reference sequence with a threshold value,
15 wherein, in the encoding step, the reads that are determined to be imperfectly mapped are
encoded according to the second encoding process or to a third encoding process, the imperfectly
mapped reads being encoded according to the second encoding process when said number of
mismatches is greater than the threshold value, and the imperfectly mapped reads being encoded
according to the third encoding process when said number of mismatches is lower than the
20 threshold value,
wherein, in said second encoding process, each nucleotide or base of the read is
individually encoded,
wherein said first and third encoding processes comprise distinct sets of descriptors, each
set of descriptors univocally representing the reads associated to a corresponding encoding
25 process, each of said first and third encoding processes being a reduced information source
entropy encoding process.
2. The method of claim 1, wherein the determining step comprises, when a read is
determined to be imperfectly mapped with the reference sequence and has a number of
30 mismatches lower than the threshold value, a further determination as to whether the read is

globally or locally mapped with said reference sequence, and wherein the third encoding process comprises a first encoding subprocess and a second encoding subprocess, the reads that are determined to be globally mapped being encoded according to the first encoding subprocess, the reads that are determined to be locally mapped being encoded according to the second encoding
5 subprocess, said first and second encoding subprocesses comprising distinct sets of descriptors, each set of descriptors univocally representing the reads associated to a corresponding encoding subprocess.

3. The method of claim 2, wherein said descriptors of said first encoding subprocess
10 comprise an alignment start position in the reference sequence, a read length and a list of mismatches by substitutions of symbols, and wherein said descriptors of said second encoding subprocess comprise a local alignment start position in the reference sequence, a read length, a list of mismatches by substitutions of symbols, and a length of the clipped portions of the read that are not part of the alignment.

15

4. The method of claim 3, wherein, in the encoding step, the clipped portions of a read that is to be encoded according to the second encoding subprocess are concatenated, each nucleotide or base of said clipped portions being individually encoded.

20 5. The method of any of the preceding claims, wherein, in the encoding step, each mismatch of an imperfectly mapped read is encoded on 1 byte.

6. The method of claim 5, wherein, in the encoding step, each mismatch of an imperfectly mapped read is encoded as follows:

- 25
- two first bits of the byte are used to encode an alternate nucleotide or base present in the read instead of a corresponding reference nucleotide or base in the reference sequence; and
 - six last bits of the byte are used to encode a position of the mismatch in the reference sequence, said position being computed as an offset from a previous
30 mismatch of the read.

7. The method of claim 6, wherein, in the encoding step, if the offset computed between a given mismatch and the previous mismatch is greater than a maximum encodable value, then at least one fake mismatch is inserted between said two mismatches until every offset between each of said mismatches and said at least one fake mismatch is lower than said maximum encodable value, a fake mismatch being defined as a mismatch for which bits of the byte used to encode the mismatch or to encode a nucleotide or base that is equal to the corresponding reference nucleotide or base in the reference sequence.
8. The method of any of the preceding claims, further comprising an initial step of dividing the list of reads into blocks of reads, with each block beginning with a header containing information needed to decode the block, wherein said compression method is performed block by block.
9. The method of claim 8, wherein the blocks of reads have the same block size.
10. The method of any of the preceding claims, further comprising a final step of providing a compressed file comprising a list of encoded reads, said encoded reads being stored in the compressed file in the same order as that of the reads stored in the initial file.
11. The method of any of the preceding claims, wherein said threshold value is equal to 31.
12. The method of any of the preceding claims, further comprising, for each aligned read, a step of determining whether said read comprises at least one mismatch corresponding to a case in which the sequencing machine was unable to call any base or nucleotide.
13. The method of claim 12, further comprising, for each read comprising at least one mismatch corresponding to a case in which the sequencing machine was unable to call any base or nucleotide, a step of determining the number of such mismatches and a step of comparing said number with a reference threshold value.

14. The method of claim 13, wherein, in the encoding step, if the number of such mismatches is greater than the reference threshold value, each nucleotide or base of a read that is to be encoded according to the second encoding process is individually encoded on 4 bits, and, if the
5 number of such mismatches is lower than the reference threshold value, each nucleotide or base of a read that is to be encoded according to the second encoding process is individually encoded on 2 bits and the encoding step further comprises encoding a list of positions along the reference sequence, said positions corresponding to the positions of such mismatches in the reference sequence.

10

15. A computer program product embodied on a computer readable storage medium, said computer program product comprising computer executable instructions which, when executed by a processor, cause the processor to perform operations comprising the steps of the method of any of the preceding claims.

15

16. A computer readable storage medium having computer executable instructions which, when executed by a processor, cause the processor to perform operations comprising the steps of the method of any of claims 1-14.

20 17. An apparatus, comprising:

a processor; and

a memory operatively coupled to the processor to form a computing device, the memory storing processor-executable instructions which, based at least on being executed on the processor, cause the processor to perform operations comprising the steps of the method of claim

25 1.

18. A method for compressing genomic sequence data, the method comprising:

obtaining, by one or more computers, a read record;

determining, by the one or more computers, whether the read record corresponds to a read

30 that is perfectly mapped to a reference sequence or imperfectly mapped to the reference sequence;

based on determining, by the one or more computers, that the read record corresponds to a read that is imperfectly mapped to the reference sequence, determining, by the one or more computers, whether a number of mismatches of the imperfectly mapped read satisfies a predetermined threshold number of mismatches; and

- 5 based on determining that the number of mismatches satisfies the predetermined threshold number of mismatches, encoding, by the one or more computers, each mismatch of the imperfectly mapped read into a record having a size of 1 byte.

19. The method of claim 18, wherein determining, by the one or more computers, whether a
10 number of mismatches of the imperfectly mapped read satisfies a predetermined threshold number of mismatches comprises:

 determining, by the one or more computers, whether the number of mismatches of the imperfectly mapped read is greater than the predetermined threshold number of mismatches.

- 15 20. The method of claim 18, wherein each read record comprises:
 data indicating an absolute starting position of an aligned read with respect to the reference sequence;
 data indicating a length of the read;
 data indicating whether the read is perfectly mapped or imperfectly mapped;
20 data indicating a number of mismatches identified in the read; and
 data indicating a relative position of each of said possible mismatches in the read.

21. The method of claim 18, wherein encoding each mismatch of the imperfectly mapped read into a record having a size of 1 byte comprises for each particular mismatch:

- 25 encoding, by the one or more computers, a first two bits of the byte to include data representing an alternate nucleotide or base present in the read instead of a corresponding reference nucleotide or base in the reference sequence; and
 encoding, by one or more computers, six remaining bits of the byte to include data representing a position of the mismatch in the reference sequence, said position being
30 computed as an offset from a previous mismatch of the read.

22. The method of claim 21, the method further comprising:
determining, by one or more computers, whether the offset is greater than a maximum encodable value; and
- 5 based on determining that the offset is greater than the maximum encoded value, inserting, by one or more computers, at least one fake mismatch between the particular mismatch and the previous mismatch.
23. The method of claim 18, the method further comprising:
10 based on determining that the number of mismatches does not satisfy the predetermined threshold number of mismatches, encoding, by one or more computers, a list of positions of the reference sequence corresponding to a position of each of the mismatches to the reference sequence using a reduced information entropy encoding process.
- 15 24. The method of claim 18, the method further comprising:
based on determining that the read record corresponds to a read that is perfectly mapped to the reference sequence, encoding, by one or more computers, at least a portion of the read record using reduced information entropy encoding.
- 20 25. The method of claim 18, wherein the one or more computers comprises one or more hardware processors.
26. The method of claim 25 wherein the one or more hardware processors comprises one or more field programmable gate arrays (FPGAs).
- 25 27. A hardware processor that includes hardware processing circuitry that is configured to perform one or more operations, the one or more operations comprising:
obtaining, by the hardware processing circuitry, a read record;

determining, by the hardware processing circuitry, whether the read record corresponds to a read that is perfectly mapped to a reference sequence or imperfectly mapped to the reference sequence;

5 based on determining, by the hardware processing circuitry, that the read record corresponds to a read that is imperfectly mapped to the reference sequence, determining, by the one or more computers, whether a number of mismatches of the imperfectly mapped read satisfies a predetermined threshold number of mismatches; and

10 based on determining that the number of mismatches satisfies the predetermined threshold number of mismatches, encoding, by the hardware processing circuitry, each mismatch of the imperfectly mapped read into a record having a size of 1 byte.

28. The hardware processor of claim 27, wherein each read record comprises:

data indicating an absolute starting position of the aligned read with respect to the reference sequence;

15 data indicating a length of the read;

data indicating whether the read is perfectly mapped or imperfectly mapped;

data indicating a number of mismatches identified in the read; and

data indicating a relative position of said possible mismatches in the read.

20 29. The hardware processor of claim 27, wherein encoding each mismatch of the imperfectly mapped read into a record having a size of 1 byte comprises for each particular mismatch:

encoding, by the hardware processing circuitry, a first two bits of the byte to include data representing an alternate nucleotide or base present in the read instead of a corresponding reference nucleotide or base in the reference sequence; and

25 encoding, by the hardware processing circuitry, a six remaining bits of the byte to include data representing a position of the mismatch in the reference sequence, said position being computed as an offset from a previous mismatch of the read.

30 30. The hardware processor of claim 29, wherein the hardware processor circuitry is further configured to perform operations comprising:

determining, by the hardware processing circuitry, whether the offset is greater than a maximum encodable value;

based on determining that the offset is greater than the maximum encoded value, inserting, by the hardware processing circuitry, at least one fake mismatch between the particular mismatch
5 and the previous mismatch.

31. The hardware processor of claim 27, wherein the hardware processor circuitry is further configured to perform operations comprising:

based on determining that the number of mismatches does not satisfy the predetermined
10 threshold number of mismatches, encoding, by the hardware processing circuitry, a list of positions of the reference sequence corresponding to a position of each of the mismatches to the reference sequence using a reduced information entropy encoding process.

32. The hardware processor of claim 27, wherein the hardware processor circuitry is further
15 configured to perform operations comprising:

based on determining that the read record corresponds to a read that is perfectly mapped to the reference sequence, encoding, by the hardware processing circuitry, at least a portion of the read record using reduced information entropy encoding.

20 33. The hardware processor of claim 24 wherein the hardware processing circuitry comprises one or more field programmable gate arrays (FPGAs).

34. The hardware processor of claim 18, wherein determining, by the hardware processing circuitry, whether a number of mismatches of the imperfectly mapped read satisfies a
25 predetermined threshold number of mismatches comprises:

determining, by the hardware processing circuitry, whether the number of mismatches of the imperfectly mapped read is greater than the predetermined threshold number of mismatches.

35. A system for compressing genomic sequence data, the system comprising:

one or more computers and one or more storage devices storing instructions that are operable, when executed by one or more computers, to cause the one or more computers to perform the operations comprising:

obtaining, by the one or more computers, a read record;

5 determining, by the one or more computers, whether the read record corresponds to a read that is perfectly mapped to a reference sequence or imperfectly mapped to the reference sequence;

based on determining, by the one or more computers, that the read record corresponds to a read that is imperfectly mapped to the reference sequence, determining,
10 by the one or more computers, whether a number of mismatches of the imperfectly mapped read satisfies a predetermined threshold number of mismatches; and

based on determining that the number of mismatches satisfies the predetermined threshold number of mismatches, encoding, by the one or more computers, each mismatch of the imperfectly mapped read into a record having a size of 1 byte.

15

36. The system of claim 35, wherein each read record comprises:

data indicating an absolute starting position of an aligned read with respect to the reference sequence;

data indicating a length of the read;

20 data indicating whether the read is perfectly mapped or imperfectly mapped;

data indicating a number of mismatches identified in the read; and

data indicating a relative position of each of said possible mismatches in the read.

37. The system of claim 35, wherein encoding each mismatch of the imperfectly mapped read

25 into a record having a size of 1 byte comprises for each particular mismatch:

encoding, by one or more computers, a first two bits of the byte to include data representing an alternate nucleotide or base present in the read instead of a corresponding reference nucleotide or base in the reference sequence; and

encoding, by one or more computers, six remaining bits of the byte to include data representing a position of the mismatch in the reference sequence, said position being computed as an offset from a previous mismatch of the read.

5 38. The system of claim 37, the operations further comprising:
determining, by the one or more computers, whether the offset is greater than a maximum encodable value; and
based on determining that the offset is greater than the maximum encoded value, inserting,
by one or more computers, at least one fake mismatch between the particular mismatch and the
10 previous mismatch.

39. The system of claim 35, the operations further comprising:
based on determining that the number of mismatches does not satisfy the predetermined
threshold number of mismatches, encoding, by one or more computers, a list of positions of the
15 reference sequence corresponding to a position of each of the mismatches to the reference
sequence using a reduced information entropy encoding process.

40. The system of claim 35, the operations further comprising:
based on determining that the read record corresponds to a read that is perfectly mapped
20 to the reference sequence, encoding, by one or more computers, at least a portion of the read
record using reduced information entropy encoding.

41. The system of claim 35, wherein the one or more computers comprises one or more
hardware processors.

25 42. The system of claim 41 wherein the one or more hardware processors comprises one or
more field programmable gate arrays (FPGAs).

43. A computer-readable storage device having stored thereon instructions, which, when executed by a data processing apparatus, cause the data processing apparatus to perform operations for compressing genomic sequence data, the operations comprising:

obtaining a read record;

5 determining whether the read record corresponds to a read that is perfectly mapped to a reference sequence or imperfectly mapped to the reference sequence;

based on determining that the read record corresponds to a read that is imperfectly mapped to the reference sequence, determining whether a number of mismatches of the imperfectly mapped read satisfies a predetermined threshold number of mismatches; and

10 based on determining that the number of mismatches satisfies the predetermined threshold number of mismatches, encoding each mismatch of the imperfectly mapped read into a record having a size of 1 byte.

44. The computer-readable storage device of claim 43, wherein each read record comprises:

15 data indicating an absolute starting position of an aligned read with respect to the reference sequence;

data indicating a length of the read;

data indicating whether the read is perfectly mapped or imperfectly mapped;

data indicating a number of mismatches identified in the read; and

20 data indicating a relative position of each of said possible mismatches in the read.

45. The computer-readable storage device of claim 43, wherein encoding each mismatch of the imperfectly mapped read into a record having a size of 1 byte comprises for each particular mismatch:

25 encoding, by one or more computers, a first two bits of the byte to include data representing an alternate nucleotide or base present in the read instead of a corresponding reference nucleotide or base in the reference sequence; and

encoding, by one or more computers, six remaining bits of the byte to include data representing a position of the mismatch in the reference sequence, said position being
30 computed as an offset from a previous mismatch of the read.

46. The computer-readable storage device of claim 45, the operations further comprising:
determining, by the one or more computers, whether the offset is greater than a maximum
encodable value; and
- 5 based on determining that the offset is greater than the maximum encoded value, inserting,
by one or more computers, at least one fake mismatch between the particular mismatch and the
previous mismatch.
47. The computer-readable storage device of claim 43, the operations further comprising:
- 10 based on determining that the number of mismatches does not satisfy the predetermined
threshold number of mismatches, encoding a list of positions of the reference sequence
corresponding to a position of each of the mismatches to the reference sequence using a reduced
information entropy encoding process.
- 15 48. The computer-readable storage device of claim 43, the operations further comprising:
based on determining that the read record corresponds to a read that is perfectly mapped
to the reference sequence, encoding at least a portion of the read record using reduced
information entropy encoding.

20