CLAIMS

1. A method for compressing nucleic acid sequence data, the method comprising:

obtaining, by one or more computers, nucleic acid sequence data representing:

(i) a read sequence comprising data that corresponds to a plurality of base calls generated by a nucleic acid sequencing device, and

(ii) a plurality of quality scores, wherein each quality score of the plurality of quality scores indicates a likelihood that a particular base call of the read sequence was correctly generated by a nucleic acid sequencing device;

determining, by one or more computers, whether the read sequence includes at least one “N” base;

based on a determination that the read sequence does not include at least one “N” base, generating, by one or more computers, a first encoded data set by using a first encoding process to encode each of the quality scores of the read sequence using a base-(x minus 1) number, where x is an integer representing a number of different quality scores used by the nucleic acid sequencing device; and

using, by one or more computers, a second encoding process to encode the first encoded data set, thereby compressing the data to be compressed.

2. The method of claim 1, wherein x is equal to 3, and/or

wherein the first encoding process comprises encoding, by one or more computers, each set of five quality scores of the plurality of quality scores of the read sequence into a single byte by representing each quality score of the set of five quality scores as a base-3 number.

3. The method of claim 1, further comprising:

based on a determination that the read sequence includes at least one “N” base, generating, by one or more computers, a second encoding data set by using a third encoding process to encode each set of four quality scores of the read sequence into a single byte of memory; and

using, by one or more computers, a fourth encoding process to encode the second encoding data, and/or

wherein the second encoding process and the fourth encoding process are the same.

4. The method of claim 1, wherein the obtained data includes a FASTQ file.

5. The method of claim 1, wherein the first encoded data set is a compressed version of the plurality of quality scores.

6. The method of claim 1, wherein the second encoding process is a compression process; and/or

wherein the compression process comprises a Prediction by Partial Matching (PPMD) implementation of a range encoder, optionally wherein, for a given value of the first encoded data set, the given value is compressed according to a 4-bit context relative to the position of the given value within the first encoded data set.

7. A system for compressing nucleic acid sequence data, the system comprising:

one or more data processing apparatus; and

one or more non-transitory computer-readable storage devices having stored thereon instructions that, when executed by the one or more data processing apparatus, cause the one or more data processing apparatus to perform operations, the operations comprising:

obtaining, by the one or more computers, nucleic acid sequence data representing:

(i) a read sequence comprising data that corresponds to a plurality of base calls generated by a nucleic acid sequencing device, and

(ii) a plurality of quality scores, wherein each quality score of the plurality of quality scores indicates a likelihood that a particular base call of the read sequence was correctly generated by a nucleic acid sequencing device;

determining, by the one or more computers, whether the read sequence includes at least one “N” base;

based on a determination that the read sequence does not include at least one “N” base, generating, by the one or more computers, a first encoded data set by using a first encoding process to encode each of the quality scores of the read sequence using a base-(x minus 1) number, where x is an integer representing a number of different quality scores used by the nucleic acid sequencing device; and

using, by the one or more computers, a second encoding process to encode the first encoded data set, thereby compressing the data to be compressed.

8. The system of claim 7, wherein x is equal to 3, and/or

wherein the first encoding process comprises encoding, by the one or more computers, each set of five quality scores of the plurality of quality scores of the read sequence into a single byte by representing each quality score of the set of five quality scores as a base-3 number.

9. The system of claim 7, the operations further comprising:

based on a determination that the read sequence includes at least one “N” base, generating, by the one or more computers, a second encoding data set by using a third encoding process to encode each set of four quality scores of the read sequence into a single byte of memory; and

using, by the one or more computers, a fourth encoding process to encode the second encoding data, and/or

wherein the second encoding process and the fourth encoding process are the same.

10. The system of claim 7, wherein the obtained data includes a FASTQ file.

11. The system of claim 7, wherein the first encoded data set is a compressed version of the plurality of quality scores.

12. The system of claim 7, wherein the second encoding process is a compression process, and/or

wherein the compression process comprises a Prediction by Partial Matching (PPMD) implementation of a range encoder, optionally wherein, for a given value of the first encoded data set, the given value is compressed according to a 4-bit context relative to the position of the given value within the first encoded data set.

13. A non-transitory computer-readable storage device having stored thereon instructions, which, when executed by a data processing apparatus, cause the data processing apparatus to perform operations, the operations comprising:

obtaining nucleic acid sequence data representing:

(i) a read sequence comprising data that corresponds to a plurality of base calls generated by a nucleic acid sequencing device, and

(ii) a plurality of quality scores, wherein each quality score of the plurality of quality scores indicates a likelihood that a particular base call of the read sequence was correctly generated by a nucleic acid sequencing device;

determining, by one or more computers, whether the read sequence includes at least one “N” base;

based on a determination that the read sequence does not include at least one “N” base, generating a first encoded data set by using a first encoding process to encode each of the quality scores of the read sequence using a base-(x minus 1) number, where x is an integer representing a number of different quality scores used by the nucleic acid sequencing device; and

using a second encoding process to encode the first encoded data set, thereby compressing the data to be compressed.

14. The computer-readable storage device of claim 13, wherein x is equal to 3, and/or

wherein the first encoding process comprises encoding each set of five quality scores of the plurality of quality scores of the read sequence into a single byte by representing each quality score of the set of five quality scores as a base-3 number.

15. The computer-readable storage device of claim 13, the operations further comprising:

based on a determination that the read sequence includes at least one “N” base, generating a second encoding data set by using a third encoding process to encode each set of four quality scores of the read sequence into a single byte of memory; and

using a fourth encoding process to encode the second encoding data, and/or

wherein the second encoding process and the fourth encoding process are the same.

16. The computer-readable storage device of claim 13, wherein the obtained data includes a FASTQ file.

17. The computer-readable storage device of claim 13, wherein the first encoded data set is a compressed version of the plurality of quality scores.

18. The computer-readable storage device of claim 13, wherein the second encoding process is a compression process.

19. The computer-readable storage device of claim 18, wherein the compression process comprises a Prediction by Partial Matching (PPMD) implementation of a range encoder.

20. The computer-readable storage device of claim 19, wherein, for a given value of the first encoded data set, the given value is compressed according to a 4-bit context relative to the position of the given value within the first encoded data set.