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~~.~~

~~3. The method of claim 2~~, 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.

~~4~~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~~.~~

~~5. The method of claim 4~~, and/or wherein the second encoding process and the fourth encoding process are the same.

~~6~~4. The method of claim 1, wherein the obtained data includes a FASTQ file.

~~7~~5. The method of claim 1, wherein the first encoded data set is a compressed version of the plurality of quality scores.

~~8~~6. The method of claim 1, wherein the second encoding process is a compression process~~.~~

~~9. The method of claim 8~~, and/or wherein the compression process comprises a Prediction by Partial Matching (PPMD) implementation of a range encoder~~.~~

~~10. The method of claim 9~~, 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.

~~11~~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.

~~128~~. The system of claim ~~117~~, wherein x is equal to 3~~.~~

~~13. The system of claim 12~~, 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.

~~149~~. The system of claim ~~117~~, 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~~.~~

~~15. The system of claim 14~~, and/or wherein the second encoding process and the fourth encoding process are the same.

~~16~~10. The system of claim ~~11~~7, wherein the obtained data includes a FASTQ file.

~~17~~11. The system of claim ~~11~~7, wherein the first encoded data set is a compressed version of the plurality of quality scores.

~~18~~12. The system of claim ~~11~~7, wherein the second encoding process is a compression process~~.~~

~~19. The system of claim 18~~, and/or wherein the compression process comprises a Prediction by Partial Matching (PPMD) implementation of a range encoder~~.~~

~~20. The system of claim 19~~, 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.

~~21~~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.

~~22~~14. The computer-readable storage device of claim ~~21~~13, wherein x is equal to 3~~.~~

~~23. The computer-readable storage device of claim 22~~, 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.

~~24~~15. The computer-readable storage device of claim ~~21~~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~~.~~

~~25. The computer-readable storage device of claim 24~~, and/or wherein the second encoding process and the fourth encoding process are the same.

~~26~~16. The computer-readable storage device of claim ~~21~~13, wherein the obtained data includes a FASTQ file.

~~27~~17. The computer-readable storage device of claim ~~21~~13, wherein the first encoded data set is a compressed version of the plurality of quality scores.

~~28~~18. The computer-readable storage device of claim ~~21~~13, wherein the second encoding process is a compression process.

~~29~~19. The computer-readable storage device of claim ~~28~~18, wherein the compression process comprises a Prediction by Partial Matching (PPMD) implementation of a range encoder.

~~30~~20. The computer-readable storage device of claim ~~29~~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.

31. 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, a frequency of occurrence for each quality score group in the plurality of quality scores, wherein each quality score group comprises a subset of quality scores of the plurality of quality scores;

for each particular quality score in a first subset of the plurality of quality scores:

determining, by one or more computers, that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold;

based on a determination that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, generating, by one or more computers and using a predetermined group mapping, first data that is to be used as a single entry in a reduced sequence, wherein the first data that is to be used as the single entry in the reduced sequence represents the particular quality score group; and

generating, by one or more computers, the reduced sequence by aggregating the generated first data for each of the particular quality score groups.

32. The method of claim 31, wherein the obtained data includes a FASTQ file.

33. The method of claim 31, wherein each quality score in the plurality of quality scores is data representing an ASCII value for the quality score.

34. The method of claim 31, further comprising:

for each particular quality score in a second subset of the plurality of quality scores:

determining, by one or more computers, that the particular quality score in the second subset of the sequence of the quality scores is not a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and

generating, by one or more computers and using a predetermined single mapping, second data that is to be used as a single entry in a reduced sequence, wherein the second data that is be used as the single entry in the reduced sequence represents the quality score that is not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, wherein the predetermined single mapping defines a one-to-one mapping between each of a plurality of single quality scores and a corresponding single entry.

35. The method of claim 34,

wherein generating, by one or more computers, the reduced sequence comprises:

aggregating, by one or more computers, the generated first data for each of the particular quality score groups; and

aggregating, by one or more computers, the generated second data for each of the quality scores that are not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold.

36. The method of claim 31, further comprising identifying, by one or more computers, a plurality of quality score groups in the plurality of quality scores.

37. The method of claim 31, wherein the predetermined group mapping defines a one-to-one mapping between each of a plurality of different quality score groups and a corresponding single entry.

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

one or more data processing apparatus; and

one or more 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, a frequency of occurrence for each quality score group in the plurality of quality scores, wherein each quality score group comprises a subset of quality scores of the plurality of quality scores;

for each particular quality score in a first subset of the plurality of quality scores:

determining, by the one or more computers, that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold;

based on a determination that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, generating, by the one or more computers and using a predetermined group mapping, first data that is to be used as a single entry in a reduced sequence, wherein the first data that is to be used as the single entry in the reduced sequence represents the particular quality score group; and

generating, by the one or more computers, the reduced sequence by aggregating the generated first data for each of the particular quality score groups.

39. The system of claim 38, wherein the obtained data includes a FASTQ file.

40. The system of claim 38, wherein each quality score in the plurality of quality scores is data representing an ASCII value for the quality score.

41. The system of claim 38, the operations further comprising:

for each particular quality score in a second subset of the plurality of quality scores:

determining, by the one or more computers, that the particular quality score in the second subset of the sequence of the quality scores is not a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and

generating, by the one or more computers and using a predetermined single mapping, second data that is to be used as a single entry in a reduced sequence, wherein the second data that is be used as the single entry in the reduced sequence represents the quality score that is not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, wherein the predetermined single mapping defines a one-to-one mapping between each of a plurality of single quality scores and a corresponding single entry.

42. The system of claim 41,

wherein generating, by the one or more computers, the reduced sequence comprises:

aggregating, by the one or more computers, the generated first data for each of the particular quality score groups; and

aggregating, by the one or more computers, the generated second data for each of the quality scores that are not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold.

43. The system of claim 38, further comprising identifying, by the one or more computers, a plurality of quality score groups in the plurality of quality scores.

44. The system of claim 38, wherein the predetermined group mapping defines a one-to-one mapping between each of a plurality of different quality score groups and a corresponding single entry.

45. 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, 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 a frequency of occurrence for each quality score group in the plurality of quality scores, wherein each quality score group comprises a subset of quality scores of the plurality of quality scores;

for each particular quality score in a first subset of the plurality of quality scores:

determining that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold;

based on a determination that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, generating, using a predetermined group mapping, first data that is to be used as a single entry in a reduced sequence, wherein the first data that is to be used as the single entry in the reduced sequence represents the particular quality score group; and

generating the reduced sequence by aggregating the generated first data for each of the particular quality score groups.

46. The computer-readable storage device of claim 45, wherein the obtained data includes a FASTQ file.

47. The computer-readable storage device of claim 45, wherein each quality score in the plurality of quality scores is data representing an ASCII value for the quality score.

48. The computer-readable storage device of claim 45, the operations further comprising:

for each particular quality score in a second subset of the plurality of quality scores:

determining that the particular quality score in the second subset of the sequence of the quality scores is not a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and

generating, using a predetermined single mapping, second data that is to be used as a single entry in a reduced sequence, wherein the second data that is be used as the single entry in the reduced sequence represents the quality score that is not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, wherein the predetermined single mapping defines a one-to-one mapping between each of a plurality of single quality scores and a corresponding single entry.

49. The computer-readable storage device of claim 48,

wherein generating the reduced sequence comprises:

aggregating the generated first data for each of the particular quality score groups; and

aggregating the generated second data for each of the quality scores that are not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold.

50. The computer-readable storage device of claim 45, the operations further comprising identifying a plurality of quality score groups in the plurality of quality scores.

51. The computer-readable storage device of claim 45, wherein the predetermined group mapping defines a one-to-one mapping between each of a plurality of different quality score groups and a corresponding single entry.

52. 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 identified by a nucleic acid sequencing device;

determining, by one or more computers, a frequency of occurrence for each quality score group in the plurality of quality scores, wherein each quality score group comprises a subset of quality scores of the plurality of quality scores;

for each particular quality score in a first subset of the plurality of quality scores: determining, by one or more computers, that the particular quality score in the first subset of the plurality of the quality scores is not a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and

generating, by one or more computers and using a predetermined single mapping, first data that is to be used as a single entry in a reduced sequence, wherein the first data that is be used as the single entry in the reduced sequence represents the quality score that is not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold; and

generating, by one or more computers, the reduced sequence by aggregating the generated first data for each of the quality scores that are not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold.

53. The method of claim 52, wherein the obtained data includes a FASTQ file.

~~54. The method of claim 52, wherein each quality score in the plurality of quality scores is data representing an ASCII value for the quality score.~~

~~55 The method of claim 52, the method further comprising:~~

~~for each particular quality score in a second subset of the plurality of quality scores:~~

~~determining, by one or more computers, that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and~~

~~based on a determination that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, generating, by one or more computers and using a predetermined group mapping, second data that is to be used as a single entry in a reduced sequence, wherein the second data that is to be used as the single entry in the reduced sequence represents the particular quality score group, wherein the predetermined group mapping defines a one-to-one mapping between each of a plurality of different quality score groups and a corresponding single entry.~~

~~56. The method of claim 55,~~

~~wherein generating, by one or more computers, the reduced sequence comprises:~~

~~aggregating, by one or more computers, the generated first data for each of the quality scores that are not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold; and~~

~~aggregating, by one or more computers, the generated second data for each of the particular quality score groups.~~

~~57. The method of claim 52, the method further comprising:~~

~~identifying, by one or more computers, a plurality of quality score groups in the plurality of quality scores.~~

58. The method of claim 52,

wherein the predetermined single mapping defines a one-to-one mapping between each of a plurality of single quality scores and a corresponding single entry.

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

one or more data processing apparatus; and

one or more 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 identified by a nucleic acid sequencing device;

determining, by the one or more computers, a frequency of occurrence for each quality score group in the plurality of quality scores, wherein each quality score group comprises a subset of quality scores of the plurality of quality scores;

for each particular quality score in a first subset of the plurality of quality scores:

determining, by the one or more computers, that the particular quality score in the first subset of the plurality of the quality scores is not a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and

generating, by the one or more computers and using a predetermined single mapping, first data that is to be used as a single entry in a reduced sequence, wherein the first data that is be used as the single entry in the reduced sequence represents the quality score

that is not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold; and

generating, by the one or more computers, the reduced sequence by aggregating the generated first data for each of the quality scores that are not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold.

60. The system of claim 59, wherein the obtained data includes a FASTQ file.

61. The system of claim 59, wherein each quality score in the plurality of quality scores is data representing an ASCII value for the quality score.

62. The system of claim 59, the operations further comprising:

for each particular quality score in a second subset of the plurality of quality scores:

determining, by the one or more computers, that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and

based on a determination that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, generating, by the one or more computers and using a predetermined group mapping, second data that is to be used as a single entry in a reduced sequence, wherein the second data that is to be used as the single entry in the reduced sequence represents the particular quality score group, wherein the predetermined group mapping defines a one-to-one mapping between each of a plurality of different quality score groups and a corresponding single entry.

63. The system of claim 62,

wherein generating, by the one or more computers, the reduced sequence comprises:

aggregating, by the one or more computers, the generated first data for each of the quality scores that are not a member of a particular quality score

group having a frequency of occurrence that satisfies the predetermined threshold; and

aggregating, by the one or more computers, the generated second data for each of the particular quality score groups.

64. The system of claim 59, the system further comprising:

identifying, by the one or more computers, a plurality of quality score groups in the plurality of quality scores.

65. The system of claim 59,

wherein the predetermined single mapping defines a one-to-one mapping between each of a plurality of single quality scores and a corresponding single entry.

66. 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, 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 identified by a nucleic acid sequencing device;

determining a frequency of occurrence for each quality score group in the plurality of quality scores, wherein each quality score group comprises a subset of quality scores of the plurality of quality scores;

for each particular quality score in a first subset of the plurality of quality scores:

determining that the particular quality score in the first subset of the plurality of the quality scores is not a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and

generating, by one or more computers and using a predetermined single mapping, first data that is to be used as a single entry in a reduced

~~sequence, wherein the first data that is be used as the single entry in the reduced sequence represents the quality score that is not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold; and~~

~~generating, by one or more computers, the reduced sequence by aggregating the generated first data for each of the quality scores that are not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold.~~

~~67. The computer-readable storage device of claim 66, wherein the obtained data includes a FASTQ file.~~

~~68. The computer-readable storage device of claim 66, wherein each quality score in the plurality of quality scores is data representing an ASCII value for the quality score.~~

~~69. The computer-readable storage device of claim 66, the operations further comprising:~~

~~for each particular quality score in a second subset of the plurality of quality scores:~~

~~determining that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies a predetermined threshold; and~~

~~based on a determination that the quality score is a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold, generating, using a predetermined group mapping, second data that is to be used as a single entry in a reduced sequence, wherein the second data that is to be used as the single entry in the reduced sequence represents the particular quality score group, wherein the predetermined group mapping defines a one-to-one mapping between each of a plurality of different quality score groups and a corresponding single entry.~~

~~70. The computer-readable storage device of claim 69,~~

~~wherein generating the reduced sequence comprises:~~

~~aggregating the generated first data for each of the quality scores that are not a member of a particular quality score group having a frequency of occurrence that satisfies the predetermined threshold; and~~

~~aggregating the generated second data for each of the particular quality score groups.~~

~~71. The computer-readable storage device of claim 66, the operations further comprising:~~

~~identifying a plurality of quality score groups in the plurality of quality scores.~~

~~72. The computer-readable storage device of claim 66,~~

~~wherein the predetermined single mapping defines a one-to-one mapping between each of a plurality of single quality scores and a corresponding single entry.~~