CLAIMS

1. A computer-implemented method for the compression of genome sequence data produced by a sequencing machine, said genome sequence data comprising reads of sequences of nucleotides,
said method comprising the steps of:

- aligning said reads to one or more reference sequences thereby creating aligned reads,
- classifying said aligned reads into classes according to whether
    - class P: said aligned reads match said one or more reference sequence without any error,
    - class N: said aligned reads match a region in said one or more reference sequences with a number of mismatches constituted by a number of positions in which the sequencing machine was not able to call any base,
    - class M: said aligned reads match a region in said one or more reference sequences with a number of mismatches constituted by a number of positions in which the sequencing matching was not able to call any base and the presence of insertions or deletions or clipped nucleotides, or
    - class I: said aligned reads do not find any valid mapping on one or more reference sequences according to specified alignment constraints,

  thereby creating classes of aligned reads, and
- encoding said classified and aligned reads as a multiplicity of layers of syntax elements comprising descriptors, said descriptors comprising for class P at least the start position on the reference sequence, a flag signaling if the read has to be considered as a reverse complement versus the reference, a distance to the mate pair in case of paired reads, the value of the length in case of the sequencing technology produces variable length reads, said descriptors comprising for class N at least the descriptors of class P and a mismatch position for each mismatch, said descriptors comprising for class M the descriptors of class N and a mismatch position and a mismatch type for each mismatch, said descriptors comprising for class I the descriptors of class P and a mismatch type for each mismatch,

wherein encoding said classified aligned reads as a multiplicity of layers of syntax elements comprises selecting said syntax elements comprising said descriptors according to said classes of aligned reads,
wherein the encoding of said classified aligned reads as a multiplicity of layers of syntax elements is adapted according to the statistical properties of the data carried by the layer,
wherein the encoding of said classified aligned reads as a multiplicity of layers of syntax elements comprising said descriptors associates a specific source model and a specific entropy coder to each layer, the source model

being characterized by the definition of the syntax elements emitted by each source, the definition of an associated probability model and the definition of the associated entropy coder,
wherein there is decomposition of the sequence read data and metadata into homogeneous layers of syntax elements in order to obtain distinct information sources with reduced information entropy.

2. The method of claim 1, wherein said layers of syntax elements comprise a position along the reference sequence, a distance between two positions on the reference sequence of two reads belonging to a read pair and information indicating if sequence reads are reverse complemented.

3. The method of claim 2, wherein said layers of syntax elements further comprise the position of a variant with respect to the reference sequence, the type of variant, the position of a deletion with respect to the reference sequence, the position of one or more symbols not present in the reference sequence, but present in the aligned reads, the type of insertion at a given position.

4. The method of claim 1, wherein said entropy coder is a context adaptive arithmetic coder.

5. A computer-implemented method for the decompression of a genomic stream compressed according to the method of claim 1, said method comprising the steps of:

- parsing and decoding (212-214) said compressed genomic stream into genomic layers of syntax elements (215),
- decoding said genomic layers into classes of data (216-217),
- expanding said genomic layers into classified reads of sequences of nucleotides, and
- selectively decoding, using class decoders (219), said classified reads of sequences of nucleotides and merging the result on one or more reference sequences so as to produce uncompressed reads of sequences of nucleotides.

6. A genomic encoder (2010) for the compression of genome sequence data (209), said genome sequence data (209) comprising reads of sequences of nucleotides, said genomic encoder (2010) comprising:

- an aligner unit (201), configured to align said reads to one or more reference sequences thereby creating aligned reads,
- a data classification unit (204), configured to classify said aligned reads according to whether
    - class P: said aligned reads match said one or more reference sequence without any error;

- class N: said aligned reads match a region in said one or more reference sequences with a number of mismatches constituted by a number of positions in which the sequencing machine was not able to call any base;
- class M: said aligned reads match said a region in said one or more reference sequences with a number of mismatches constituted by a number of positions in which the sequencing machine was not able to call any base and the presence of insertions or deletions or clipped nucleotides; or
- class I: said aligned reads do not find any valid mapping on one or more reference sequences according to specified alignment constraints; thereby creating classes of aligned reads; and

- one or more layers encoding units (205-207), configured to encode said classified aligned reads as layers of syntax elements comprising descriptors which univocally represent said classified and aligned reads, said descriptors comprising at least data related to the mapping position of the reads with respect to the reference sequence by selecting said syntax elements according to said classes of aligned reads, said descriptors comprising for class P at least the start position on the reference sequence, a flag signaling if the read has to be considered as a reverse complement versus the reference, a distance to the mate pair in case of paired reads, the value of the length in case of the sequencing technology produces variable length reads, said descriptors comprising for class N at least the descriptors of class P and a mismatch position for each mismatch, said descriptors comprising for class M the descriptors of class N and a mismatch position and a mismatch type for each mismatch, said descriptors comprising for class I the descriptors of class P and a mismatch type for each mismatch,

wherein the encoding of said classified aligned reads as a multiplicity of layers of syntax elements is adapted according to the statistical properties of the data carried by the layer,

wherein the encoding of said classified aligned reads as a multiplicity of layers of syntax elements associates a specific source model and a specific entropy coder to each layer, the source model being characterized by the definition of the syntax elements emitted by each source, the definition of an associated probability model and the definition of the associated entropy coder,

wherein there is decomposition of the sequence read data and metadata into homogeneous layers of syntax elements in order to obtain distinct information sources with reduced information entropy.

7. A genomic decoder (218) for the decompression of a genomic stream (211) compressed by the genomic encoder of claim 6, said genomic decoder (218) comprising:

- parsing and decoding means (210, 212-214) configured to parse said compressed genomic stream into genomic layers of syntax elements (215),
- one or more layer decoders (216-217), configured to decode the genomic layers into classes of data, and further configured to process said genomic layers into classified reads of sequences of nucleotides (2111), and
- genomic data classes decoders (213) configured to selectively decode said classified reads of sequences of nucleotides and merge the results on one or more reference sequences so as to produce uncompressed reads of sequences of nucleotides.

8. The genomic decoder of claim 7, wherein the one or more reference sequences are stored in the compressed genome stream (211).

9. The genomic decoder of claim 7, wherein the one or more reference sequences are provided to the decoder via an out of band mechanism.

10. The genomic decoder of claim 7, wherein the one or more reference sequences are built at the decoder.

11. A computer-readable medium comprising instructions that when executed cause at least one processor to perform the method of any one of claims 1 to 5.