CLAIMS

1. A computer-implemented method for the storage of a representation of genome sequence data in a genomic file format, said genome sequence data comprising reads of sequences of nucleotides, comprising the steps of:

- aligning said reads to one or more reference sequences thereby creating aligned reads,
- classifying said aligned reads according to whether:
    - class P: said aligned reads match with said one or more reference sequences without any error;
    - class N: said aligned reads match with said one or more reference sequences a number of mismatches constituted by a number of positions in which the sequencing machine was not able to call any base or nucleotide;
    - class M: said aligned reads match with 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 or nucleotide or a different base than the one reported in the reference genome has been called;
    - class I: said aligned reads include the same mismatches of class M plus the presence of insertions or deletions;
    - class U: said aligned reads have no valid mapping on the reference genome according to the specific alignment constraints;

  thereby creating classes of aligned reads,
- encoding said classified aligned reads as layers, each layer comprising data blocks, each data block comprising descriptors as a vector of data elements of the same type, said descriptors comprising for class P a starting position (pos) on the reference genome, a distance to the mate pair in case of paired reads (pair), a flag signaling if the read has to be considered as a reverse complement versus the reference (rcomp), additional flags describing characteristics of the read and the value of the read length (rlen) in case a sequencing technology produces variable length reads, for class N all descriptors of class P and a mismatch position (nmis), for class M all descriptors of class P, a mismatch type (snpt) and a mismatch position (snpp), for class I all descriptors of class P, a mismatch position (indp), a mismatch type (indt) and optionally a soft clipped nucleotides string when present (indc), wherein the encoding is adapted according to the statistical properties of the data carried by the layer;

- structuring said layers of data blocks with header information thereby creating successive access units,
- creating a master index table, containing one section for each class of aligned reads, comprising the mapping positions on said one or more reference sequences of the first read of each access unit of each class of aligned reads; and
- jointly storing said master index table and said access unit data.

2. The method of claim 1, wherein said master index table further comprises a vector of pointers to the physical position of each subsequent access unit in an encoded bitstream.

3. The method of claim 1, wherein said master index table further contains one section for each reference sequence.

4. The method of claim 1, wherein the encoding of said classified aligned reads as layers associates a source model of a descriptor with a specific entropy coder.

5. The method of claim 4, wherein the source model adopted in one Access Units is independent from the source model used by other Access Units for the same data layer.

6. A method to extract reads of sequences of nucleotides stored in a genomic file, wherein said genomic file comprises a master index table and access units data stored according to the method of claim 1, said method comprising the steps of:

- receiving user input identifying the type of reads to be extracted,
- retrieving the master index table from said genomic file,
- retrieving the access units corresponding to said type of reads to be extracted, the retrieving further comprising extracting vectors of data indicating the position on the reference genome of the first read in each access unit and retrieving data indicating a physical position of access units per each class of data, and
- reconstructing said reads of sequences of nucleotides mapping said retrieved access units on one or more reference sequences.

7. The method of claim 6, wherein the genomic file further comprises the one or more reference sequences.

8. The method of claim 6, wherein the one or more reference sequences are provided via an out of band mechanism.

9. A Genome Sequencing Machine comprising:

- a genome sequencing unit (130), configured to output reads of sequences of nucleotides (131) from a biological sample,
- an alignment unit (132), configured to align said reads to one or more reference sequences thereby creating aligned reads (133),
- a classification unit (134), configured to classify said aligned reads according to whether:
    - class P: said aligned reads match with said one or more reference sequences without any error;
    - class N: said aligned reads match with said one or more reference sequences a number of mismatches constituted by a number of positions in which the sequencing machine was not able to call any base or nucleotide;
    - class M: said aligned reads match with 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 or nucleotide or a different base than the one reported in the reference genome has been called;
    - class I: said aligned reads include the same mismatches of class M plus the presence of insertions or deletions;
    - class U: said aligned reads have no valid mapping on the reference genome according to the specific alignment constraints;

  thereby creating classes of aligned reads (135);
- an encoding unit (136), configured to encode said classified aligned reads as layers, each layer comprising data blocks, each data block comprising descriptors as a vector of data elements of the same type, said descriptors comprising for class P a starting position (pos) on the reference genome, a distance to the mate pair in case of paired reads (pair), a flag signaling if the read has to be considered as a reverse complement versus the reference (rcomp), additional flags describing characteristics of the read and the value of the read length (rlen) in case a sequencing technology produces variable length reads, for class N all descriptors of class P and a mismatch position (nmis), for class M all descriptors of class P, a mismatch type (snpt) and a mismatch position (snpp), for class I all descriptors of class P, a mismatch position (indp), a mismatch type (indt) and optionally a soft clipped nucleotides string when present (indc), wherein the encoding of said classified aligned reads as layers is further adapted according to the statistical properties of the data carried by the layer,
- a subdividing unit (138), configured to structure said layers of data blocks with header information thereby creating successive access units (139),

- an index table processing unit (1310), configured to create a master index table, containing one section for each class of aligned reads, comprising the mapping positions on the reference sequence of the first read of each access unit of each class of data;
- a storage unit (1312-1316) configured to jointly store said master index table and said access unit data (1311).

10. The Genome Sequencing Machine of claim 9, wherein the master index table further comprises a vector of pointers to the physical position of each subsequent access unit in an encoded bitstream.

11. The Genome Sequencing Machine of claim 9, wherein the encoding of said classified aligned reads as layers associates a source model of a descriptor with a specific entropy coder.

12. An extractor (140) to extract reads of sequences of nucleotides stored in a genomic file, wherein said genomic file comprises a master index table and access units data stored according to the method of claim 1, said extractor (140) comprising:

- user input means (141) configured to receive input parameters (142) identifying the type of reads to be extracted,
- retrieving means (143) configured to retrieve said master index table (144) from said genomic file,
- retrieving means (145) configured to retrieve the access units (146) corresponding to said type of reads to be extracted,
- reconstructing means (147) configured to reconstruct said reads of sequences of nucleotides (148) mapping said retrieved access units on one or more reference sequences.

13. A machine readable medium comprising a plurality of instructions that in response to being executed on a computing device, causes the computing device to perform the method of claims 1-8.