

CLAIMS

1. A 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:
 - 5 aligning said reads to one or more reference sequences thereby creating aligned reads,
 - classifying said aligned reads according to different degrees of matching accuracy with said one or more reference sequences thereby creating classes of aligned reads;
 - encoding said classified aligned reads as layers of syntax elements,
 - structuring said layers of syntax elements with header information thereby creating successive
 - 10 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 units of each class of data;
 - jointly storing said master index table and said access unit data.
- 15 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.
3. The method of claim 1, wherein said master index table further contains one section for each
- 20 reference sequence.
4. The method of claim 1, wherein the encoding of said classified aligned reads as layers of syntax elements is adapted according to the specific features of the data or metadata carried by the layer.
- 25 5. The method of claim 4, wherein the encoding of said classified aligned reads as layers of syntax elements is further adapted according to the statistical properties of the data or metadata carried by the layer.
6. The method of claim 5, wherein the encoding of said classified aligned reads as layers of syntax
- 30 elements associates a source model of a descriptor with a specific entropy coder.
7. The method of claim 6, wherein the the source model adopted in one Access Units is independent from the source model used by other Access Units for the same data layer.

8. 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:

- 5 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,
reconstructing said reads of sequences of nucleotides mapping said retrieved access units on one or more reference sequences.

10

9. The method of claim 8, wherein the genomic file further comprises the one or more reference sequences.

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

11. A Genome Sequencing Machine comprising:

a genome sequencing unit 130, configured to output reads of sequences of nucleotides 131 from a biological sample,

- 20 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 matching accuracy degrees with said one or more reference sequences thereby creating classes of aligned reads 135;

- an encoding unit 136, configured to encode said classified aligned reads as layers of syntax elements
25 137,

a subdividing unit 138, configured to structure said layers of syntax elements 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
30 of the first read of each access units of each class of data;

a storage unit 1312-1316 configured to jointly storing said master index table and said access unit data 1311.

12. The Genome Sequencing Machine of claim 8, wherein the master index table further comprises a vector of pointers to the physical position of each subsequent access unit.

13. The Genome Sequencing Machine of claim 8, wherein the encoding of said classified aligned
5 reads as layers of syntax elements is adapted according to the specific features of the data or metadata carried by the layer.

14. 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,

10 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

15 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.

15. A machine readable medium comprising a plurality of instructions that in response to being
20 executed on a computing device, causes the computing device to perform the method of claims 1-10.