INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2017/M40271
April 2017, Hobart, AU

Source: **Requirements**
Title: **Use cases for an efficient genomic information representation**
Authors: Nicolas Guex, Christian Iseli, Thierry Schuepbach, Ioannis Xenarios (SIB/Vital-IT), Daniele Renzi, Giorgio Zoia (GenomSys), Claudio Alberti, Marco Mattavelli (EPFL), Jaime Delgado, Silvia Llorente, Daniel Naro (Distributed Multimedia Applications Group – Universitat Politècnica de Catalunya)

Table of Contents

1 Purpose

This document reports a collection of different use cases involving handling of genomic data; the document intends to clearly define modes of usage in terms of functionality of the sub-system and interactions with other (sub-) systems or actors during usage.
These use cases originate from the current procedures and needs of the daily genomic data processing activity that takes place at several bioinformatics data processing centers.

2 Terminology

Term	Definition
Alignment	A sequence read mapped on a reference DNA sequence
BAM	Compressed binary version of SAM
CRAM	GIR that includes SAM + Compression configuration
FASTA	GIR that includes header and sequence reads (nucleotides sequence)
FASTQ	GIR that includes FASTA + Quality Scores
GIR	Genomic Information Representation
Indel	An additional or missing nucleotide in a DNA sequence with respect to a reference DNA sequence.
MAF	Mutation Annotation Format. File format used to mark the genes and other biological features in a DNA sequence
Mate pairs	Two reads from the same (long) DNA strand extracted by sequencing machines. The orientation is the opposite of paired ends.
Paired ends	Couple of reads produced by the same (short) DNA fragment by sequencing both ends. The orientation is the opposite of mate pairs.
Quality score	A quality score is assigned to each nucleotide base call in automated sequencing processes. It expresses the base-call accuracy.
Read header	Each sequence read stored in FASTA and FASTQ format starts with a textual field called “header” containing a sequence identifier and an optional description
SAM	GIR that is human readable and includes FASTQ + Alignment and analysis information
Sequence read	The readout, by a specific technology more or less prone to errors, of a continuous part of a segment of DNA extracted from an organic sample

Some of the concepts described in this document are currently supported by tools used by MPEG members to process large datasets of genomic data on production systems used for scientific research. Other concepts have not been implemented yet but are required in order to enable usage scenarios currently not supported by existing formats such as SAM and CRAM.
The current Use Case were identified and described by:

- the Swiss Institute of Bioinformatics (SIB) and Vital-IT
- EPFL
- GenomSys
- DMAG-UPC

3 Use cases

3.1 Selective access to aligned data

The types of selective access to aligned data described in this Use Case are important in genomic analysis. The new genomic information representation structure should support partitioning and efficient querying of data according to the criteria listed in the actions below.

UCId	GI-UC01
Use Case Name	Selective Access to Aligned Data
Description	Genomic Sequencing Information that has been aligned is accessed in many different ways during a typical analysis session.
Actors	Analyst
Assumptions	The sequencing data to be analysed have been aligned prior to the current session and stored in a suitable format
Actions	1. The analyst opens the file 2. The analyst filters unmapped reads, cuts them from the current file and saves them in a new file. 3. The analyst filters duplicate reads and removes them from the file. 4. The analyst filters reads with no mismatches and remove them from the current session view 5. The analyst filters reads with mismatches only (no indels) and decodes reads with a number of mismatches of 3 or less 6. The analyst filters reads with mismatches, indels and soft clips and decode reads with a distance of X or less 7. The analyst browses the selected read subsets and selects regions of interest in the genome being analysed 8. The analyst labels the selected sub-regions and saves the results in a new file.
Inclusion or Extensions	The analyst reloads the previously saved file and can easily decode and access only the sub-regions selected in the previous session.
Issues	

The typical actions of an analysis session described above have precise rationales for the bioinformaticians. They are made more explicit in the following table:

	Description	Rationale
0	Efficiently retrieve both reads to a pair and reconstruct a FASTQ file equivalent to the original input	Useful to remap with a different tool, or visualize aligned pairs on a genomic browser
1	Access reads with no mismatches	Useful to determine the coverage of any particular nucleotide or genomic region (CNV, RNAseq) but not as a source for the discovery of new variants
2	Access reads with mismatches only	They contain the most useful information for SNP calling or to identify premature stop codons
2.1	Decode only reads with a number of mismatches below/above a given threshold	Used in variant calling to filter reads with too many mismatches which may indicate poor sequencing quality and / or incorrect mapping
3	Access reads with indels	They contain the most useful information to detect frameshifts, or to identify chromosome rearrangements
4	Access unmapped reads only	Useful to detect potential contamination or infections
5	Group genomic sub-regions uniquely identified with user-defined criteria	
5.1	Count the reads (or pairs in the case of paired-end experiments) that fall within each defined region	Used in gene expression analysis
5.2	Extract all the reads with mismatches or indels that fall within a specific region	Used to perform a local re-alignment or assembly of the reads to improve SNP calling or indel detection
6	Remove/filter duplicates	Crucial to prevent over-counting PCR effects during expression analyses
7	Perform the query on several genomic datasets (i.e. the results of a sequencing run)	Used in population genetics, for example to count the penetrance of a given trait / SNP
8	Remove or properly weight reads that map equally well at more than one location	Used to normalize gene expression analysis
9	Access pair-end reads whose ends map on distinct chromosomes	Used to detect cancer chromosome rearrangements and also used in HiC experiments

3.2 Files concatenation

It should be possible to concatenate compressed files without the need to decompress and recompress them. This is particularly useful to transmit, merge and aggregate small subparts of a sequencing experiment, such as gene panels or regions that cover a few SNPs of interest. It would also be very useful in the common case where a library is re-sequenced in order to get additional sequencing depth. It however causes potential issues with duplicate reads (see issues)

UCId	GI-UC02
Use Case Name	Files concatenation
Description	Concatenate compressed files without the need to decode and re-encode them
Actors	Analyst
Assumptions	
Actions	1. The analyst obtains a genomic sample of raw reads from a human individual 2. The analyst performs alignment and compression against reference sequences representing the 24 human chromosomes (22 + X and Y) 3. The alignment is performed in parallel per each chromosome and the output is a compilation of 24 compressed files. The analyst opens the 24 files and concatenates them into a single compressed bitstream very quickly with no need of decompression and re-compression of each individual file.
Inclusion or Extensions	Other examples where files concatenation is useful is when different "experiences" are performed on the same genomic sequence data and the results shall be concatenated. For example alignment of the same raw data using different configurations of the same tool or different alignment approaches. Another application is the common case where a library is re-sequenced in order to get additional sequencing depth.
Issues	

3.3 Genomic studies aggregation

It should be possible to encapsulate in the same compressed file several related studies that can be separately accessible. Additionally transversal queries on all the compressed files should be possible (e.g. “select chr1 of all compressed samples”). This is particularly useful when a study is performed on large populations of individuals of the same species or when the same individual is sequenced/analyzed several times during his life time.

For example when aggregating several datasets from the same individual/species, the available reference sequences (e.g. human genome) should be stored only once with a significant gain in storage efficiency in case of studies on large populations (up to several thousand individuals) or when sequencing several times a single individual.

UCId	GI-UC03
Use Case Name	Genomic Studies Aggregation
Description	Encapsulate in the a compressed file several related studies that are maintained separately accessible
Actors	Analyst 1, Analyst 2
Assumptions	The analyst has performed various types of analysis (such as alignment, variant calling, gene expression analysis) on different datasets of genomic sequence data. The results are in the form of compressed genomic bitstreams containing sequence data and related annotations/metadata.
Actions	1. Analyst 1 opens the several files he has available from previous analysis sessions. a. The reference sequences used for alignment should be stored only a minimum number of times (one in case of single species). 2. Analyst 1 selects the previous studies that he/she wants to aggregate in a single file 3. Analyst 1 saves the aggregation and transmits the file to Analyst 2 4. Analyst 2 opens the file he received from Analyst 1 5. Analyst 2 browse one or more specific dataset of the aggregation and all the related results 6. Analyst 2 identifies an interesting result and browse all the datasets of the aggregation for same or similar types of results. 7. Analyst 2 specifies a matching function on one region of the available

	datasets for results such as: i. variant calling on chr2 of all dataset ii. retrieve all dataset where a specific mutation is present above a given threshold iii. count the number of genes in all datasets and retrieve only those above a given threshold
Inclusion or Extensions	Each aggregated dataset should be looked as one dimension of a “hypercube” with the possibility to randomize access according to any (set of) dimensions. This can be seen as well in terms of "experiences" (i.e. with several concatenated files) where the "hypercube" can be browsed according to dimensions such as genomic region, experience, individual, etc.
Issues	

3.4 Data streaming

UCId	GI-UC04
Use Case Name	Data Streaming
Description	Transmission of compressed data from a datacenter for analysis with real-time access
Actors	Sequencing operator, Analyst
Assumptions	
Actions	1. At the sequencing facility the operator launches a sequencing process 2. Raw sequence data are produced and stored on a local storage device 3. Raw sequence data are compressed and the operator starts a streaming session to an analysis center that subscribed to receive the data, as soon as data are available on the local storage device 4. The analyst at the analysis center starts receiving data from the open link with the sequencing facility 5. The analyst starts analyzing the received data as soon as they are received, even before sequencing and streaming are completed. For

	example: a. alignment and variant calling against an available reference b. if a mutation is detected above a given threshold an alarm is triggered (this could provide the required result before the sequencing process is completed) 6. The analyst forward some portions of the data being received to a subcontractor for further analysis and/or specific refinements a. These data may be selected based on e.g. some specific classification such as non-mapping data or some specific given threshold as above, etc. 7. The analyst terminates his/her analysis after all data have been received from the sequencing facility 8. The analyst waits for further results from the subcontractor and aggregates them with his/her results into a single file.
Inclusion or Extensions	
Issues	

3.5 Checking action conformity with privacy rules

UCId	GI-UC05
Use Case Name	Checking action conformity with privacy rules
Description	The system checks if the action the user intends to perform is compliant with the rules expressed by the owner of the data.
Actors	Data consumer (analyst, practitioner, service provider, forensic scientist).
Assumptions	The sequenced data are linked to privacy rules, defining which conditions the owner imposes on its access and use. The system might operate locally at the user's premises or remotely (for example on the data custodian's repository)
Actions	1. The data consumer submits the relevant information about the action s/he wants to be performed:

	- the intended use (not for further use, commercial use, relatives search, forensics, scientific purposes) - her/his identity as contact information and role (analyst, researcher, healthcare, forensic, not specified) - the action to perform (genetic analysis for diagnostic, analysis for a research project, view information, paternity test, forensic use) - to whom the result will be forwarded (contact information and role) - the set of genomic regions where the action will be performed 2. The system checks if the provided user information can be verified 3. The system checks if the provided privacy rules can be verified 4. The system checks if the provided information meet the privacy rules, including the different uses regulated (not for further use, not for commercial use, not for relatives search use, not for forensics use, forensics use, open access, only for scientific purposes, …), and performs necessary actions, such as informing the owner or requesting a further confirmation from the owner 5. Depending on the result of privacy rules evaluation, the system performs the requested action 6. The system notifies the owner about the result, if requested.
Inclusion or Extensions	In order to enforce the application of the privacy rules, data have to be encrypted and the process has to request the necessary information for decryption.
Issues	

3.6 Multiple alignments

It is necessary to support and properly weight reads that map equally well at more than one location. This is used e.g. to normalize gene expression analysis.
The current existing format supporting Multiple Alignments is the Multiple Alignment Format (MAF, https://cgwb.nci.nih.gov/FAQ/FAQformat.html).

UCId	GI-UC06
Use Case Name	Multiple alignments
Description	Many genes are duplicated in the genome, and there are many gene families where differences between genes from the same family at the DNA level cannot be determined from short reads, if at all. In such cases, it is common practice to spread the weight of reads where there are multiple mapping positions among all possible positions/genes. In some cases, those reads can also be excluded from the downstream analysis
Actors	Analyst, alignment tool
Assumptions	A scoring scheme is adopted to measure the *quality* of each alignment. Existing scoring schemes use: • points to each aligned base • probability (normalized to 1 among all alignments)
Actions	1. The analyst opens files containing non-aligned sequencing data 2. The analyst selects the scoring scheme in case of multiple alignments 3. The analyst runs the first alignment pass 4. When multiple mapping positions are found, each alignment is assigned a score according to the selected scheme 5. The analyst saves the alignment results in a suitable format allowing the storage of sensible information for all multiple alignment (incl. MAF)
Inclusion or Extensions	In case of multiple pass alignments, an additional integer number is associated to each alignment: usually pass 1 indicates the "strongest" alignment and additional passes indicate the "weaker" ones
Issues	It would be important to store the configuration of the tool or toolset used to generate the multiple alignments. This should probably be encoded as metadata associated with the alignment information.

4 Conclusions

Use Cases presented in this document should help the specification of a new file format and genomic data representation and partitioning presented in this document. They have been proposed validated by organizations working on genomic data such as the Swiss Institute of Bioinformatics.
Currently the used compressors are general purpose entropy coders, but the development of compressors exploiting the nature of genomic information would certainly provide better performance.