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

ISO/IEC JTC 1/SC 29/WG 11
MPEG2017/m39945
January 2017, Geneva, Switzerland

Source: DMAG-UPC, GenomSys, EPFL
Status: Proposal
Title: Towards a WD for a format on genomic information compression and storage

Authors: Jaime Delgado, Daniel Naro, Silvia Llorente (Distributed Multimedia Applications Group – Universitat Politècnica de Catalunya),
Łukasz Roguski (Centro Nacional de Análisis Genómico – Centre de Regulació Genòmica (CNAG – CRG)),
Giorgio Zoia, Daniele Renzi (GenomSys SA),
Claudio Alberti, Marco Mattavelli (EPFL)

Contents

1 Introduction

This document arises after the development of Core Experiment CE4 and the cross-checking between two of the proposals. It identifies common features from the HEGIF [1] and GENIFF [2] format proposals that could be incorporated in a first Working Draft for a standard format in the context of Genomic information compression and storage.

2 Proposal for a WD based on HEGIF and GENIFF after CE4 and cross-checking

The main contribution of this document is a detailed comparison of some technical details of both formats, mainly focusing in the header elements. The identification of the differences paves the way for a common proposal.

2.1 Format syntax

The following table compares the encapsulation levels and headers mapping.

Hierarchy	Type	HEGIF	GENIFF	Map	Comments
0	Container	Genomic File (GF)	FILE	OK	Equivalent.
0	Header	Genomic File Header (gfh)	filetype	OK	Ok. Equivalent. In GENIFF, number of compatible brands is inferred from box size
0	Container	Genomic Multiplex (GM)	genstudy	OK	A way to aggregate.
0	Header	Genomic Multiplex Header (gmh)	studyhea	OK	Length, Id, Version Nr., Flag (info. usable or not), Metadata Metadata, Access rules, Encryption attributes Compatible.
0	Header child	Genomic Dataset List (GDL)	-	-	GENIFF: Computed on the fly Decide best option.
1	Container	Genomic Dataset (GD)	gendatas	OK	Aggregate. Equivalent.
1	Header	Genomic Dataset Header (gdh)	datahead	OK	- Unique Id, Versions, Size, Reads length, reference count, number of blocks per reference, reference Ids, MIT, parameters set, Metadata

					- Id, label (id string), Metadata info, type of content (aligned or not), compression info (algorithm used), privacy rules, encryption info. Compatible.
2	Header child	Master Index Table (MIT)	dataindx	?	MIT: Embedded in gdh (child of the header of the dataset) Child of dataset container (one level above). Decide where to put it.
2	Container	Genomic Layer (GL)	gendecod	OK	Aggregators.
2	Header	Genomic Layer Header (glh)	recohead	OK	- Id, header size, payload size, number of blocks, LIT, metadata - Id, label, metadata, type, compression, security info Compatible. Same comments as previous header. Decide about possible duplications between headers.
3	Header child	Local Index Table (LIT)	recoindx	?	Similar position issue as with MIT.
3	Container	Genomic Block	-	-	GENIFF: External vs. Internal file.

2.2 Syntax mapping

The four following subsections compare HEGIF and GENIFF headers.

2.2.1 HEGIF Genomic File Header (gfh) vs GENIFF filetype

HEGIF	**GENIFF**
genomic_file_header() {	*filetype {*
Major Brand	Major brand
Minor Version	Minor version
number of_compatible_brands	*Inferred from box size*
for (i=0;i<number of_compatible_brands;i++) {	for (i=0;i<number of_compatible_brands;i++) {

compatible_brand	Compatible brand
}	}
}	}

2.2.2 HEGIF Genomic Multiplex Header (gmh) vs GENIFF studyhea

HEGIF	**GENIFF**
genomic_multiplex_header() {	*studyhea*
gmh_length	
multiplex_id	
version_number	
applicable_gdl_flag	
list_number	
gd_number	
for (i=0; i<gd_number;i++) {	
genomic_dataset_ID	
}	
gm_metadata	Metadata box (XML with information from the SAM header, or information about Medical Center, Sample)
	Encryption type
	Privacy Rules
}	

2.2.3 HEGIF Genomic Dataset Header (gdh) vs GENIFF datahead

HEGIF	GENIFF
genomic_dataset_header() {	*Datahead {*
unique_ID	ID
genomic_dataset_ID	
maj_version	
min_version	
gdh_length	//indicated before the beginning of the datahead box
reads_length	
ref_count	
for (i=0; i<ref_count;i++) {	
blocks_counter	
total_blocks += blocks_counter * ref_count	
}	
for (i=0; i<ref_count;i++) {	
ref_id	
}	
Master_Index_Table	//as separated box
parameters_sets (PS)	
gd_metadata	Metadata box (XML with information from the SAM header, or information about Medical Center, Sample)
	Label size

	Label
	Dataset type (aligned or unaligned)
	Extra data length
	Extra data
	Compression mechanism for unaligned data
	Compression mechanism for aligned data
	Encryption type
	Privacy rules box
}	}

}	

2.2.4 HEGIF Genomic Layer Header (glh) vs GENIFF recohead

HEGIF	**GENIFF**
genomic_layer_header() {	*Recohead{*
	Flags
genomic_layer_ID	ID
glh_length	//indicated before the beginning of the recohead box
layer_length	
number_of_blocks	
Local_Index_Table	//as separated box
gl_metadata	Metadata box (XML with information from the SAM header, or information about Medical Center, Sample)

	Label size
	Label
	Name size
	name
	Dataset type
	Extra data length
	Extra data
	Compression method for unaligned data
	Compression method for aligned data
	Encryption type
	Salt length
	Salt
	Privacy rules
}	

3 Streaming

Fine as long as the results of HEGIT format, described in [3] and demonstrated in [4], are compatible to the file format. To discuss once a first version of format agreed.

4 Acknowledgements

The work done in this proposal has been partially supported by the Spanish Government under the project: Secure Genomic Information Compression (GenCom, TEC2015-67774-C2-1-R and TEC2015-67774-C2-2-R).

5 References

[1] Giorgio Zoia, Daniele Renzi, ISO/IEC JTC 1/SC 29/WG 11 M39149, Core Technology Proposal for Genomic Information Coding, October 2016.

[2] Jaime Delgado, Silvia Llorente, Daniel Naro et al., ISO/IEC JTC 1/SC 29/WG 11 M39940, GENIFF v2, December 2016.

[3] Giorgio Zoia, Daniele Renzi, ISO/IEC JTC1/SC29/WG11 MPEG2016 M38961, Coding and Transport Framework for Genomic Information, October 2016

[4] Giorgio Zoia, Daniele Renzi, ISO/IEC JTC 1/SC 29/WG 11 M39871, Tools, Technology and Results for CE4 on Genomic Access Abstract Layer, January 2017.