MPEG-G: The Emerging Standard for Genomic Data

Jan Voges and Jörn Ostermann
Institut für Informationsverarbeitung, Leibniz Universität Hannover
{voges,office}@tnt.uni-hannover.de

The development of next-generation sequencing (NGS) technologies enables the usage of genomic information as everyday practice in several fields. However, the growing volume of data generated becomes a serious obstacle for the advancement of related applications. To comprehend the volume of data that needs to be represented, stored and transmitted, current sequencing machines are capable of delivering over 18,000 whole human genomes a year, which accounts for almost 5 PB of data per year and system. Therefore, efficient storage and transmission of genomic data is becoming of uttermost importance. Besides compression, which is at the base of any efficient processing of genomic information, there are several other requirements that the current data formats do not fulfill [Joi16].

ISO/IEC JTC 1/SC 29/WG 11 — also known as Moving Picture Experts Group (MPEG) — has the mission to develop standards for coded representation and compression of digital audio, video and related data. In its 29 years of activity MPEG has developed many generations of audio and video compression standards such as MP3 and AVC (also sometimes referred to as H.264). ISO/TC 276 works on standardization in the field of biotechnology processes that include analytical methods (Working Group 3) and data processing and integration (Working Group 5). MPEG and ISO/TC 276/WG 5 (Data Processing and Integration) have combined their respective expertise and missions and are jointly working to develop a new and open standard for genomic information representation, called MPEG-G.

Data type	Compression factor (approx.)
Read identifiers	10
Quality values (lossless)	3.7
Quality values (quasi-lossless)	12.5 (with less than 3% F-score degradation)
Unaligned reads (constant & variable lengths)	25–58 (low to high coverage samples)
Aligned reads (constant length)	12
Aligned reads (variable lengths)	8

This standard will be offering higher levels of compression for all relevant data classes such as reads, quality scores/values, read identifiers, and alignment information [N+16]. The table shows the results of the initial technology performance assessment. The MPEG-G standard will furthermore provide new functionalities such as support for selective access, data protection mechanisms, conversion from/to the SAM/BAM file formats for backwards compatibility, and streaming of genomic data, enabling for example live streaming of data from a sequencing machine to a remote analysis center during sequencing.

The framework for the development of the open source standard MPEG-G is provided by ISO/IEC. Following the identification of requirements and the evaluation of technologies, the standardization process involves the selection and integration of the best performing technologies into a platform, called "General Model", for the evaluation and verification of performance and the validation of requirements fulfillment. This work started in January 2017 and is currently ongoing. The current Working Draft of the standard (ISO/IEC NP 23092) will evolve into a Committee Draft in October 2017 and a Final Draft International Standard in January 2019. Finally, a normative and informative specification (International Standard) in the form of text and reference software will be published. This specification will provide the foundation for interoperable genomic information processing applications enabling the use of genomic data on a large scale in fields such as personalized medicine, where the DNA of the patient will be sequenced and analyzed as part of a standard procedure.

[Joi16] Joint AhG on Genomic Information Compression And Storage. Requirements on Genomic Information Compression and Storage. Technical report, ISO/IEC JTC 1/SC 29/WG 11 (MPEG) and ISO/TC 276/WG 5, Document Number N16323/N97, Geneva (CH), 2016.

[N+16] Ibrahim Numanagić et al. Comparison of high-throughput sequencing data compression tools. *Nature Methods*, 13(12):1005–1008, 2016.