**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 N16147**

**San Diego, USA – February 2016**

|  |  |
| --- | --- |
| **Source:** | **Requirements** |
| **Status:** | **Draft** |
| **Title:** | **Results of the Evaluation of the CfE on Genomic Information Compression and Storage** |

[1 Introduction 2](#_Toc444306455)

[2 Raw data 2](#_Toc444306456)

[2.1 Low coverage human sample 2](#_Toc444306457)

[2.2 High coverage human sample 4](#_Toc444306458)

[3 Aligned data 5](#_Toc444306459)

[3.1 High coverage human sample 5](#_Toc444306460)

[4 Computational complexity 6](#_Toc444306461)

[5 Conclusions 6](#_Toc444306462)

[6 Acknowledgements 7](#_Toc444306463)

[7 References 7](#_Toc444306464)

# Introduction

The purpose of the issued Call for Evidence (CfE) on genome compression [1] was to assess whether new technologies can achieve better performance in terms of compression efficiency compared with currently used file formats such as gzipped FASTQ (used for raw data) and BAM (used for aligned data). Additional functionalities (e.g. non-sequential access, lossy compression efficiency, etc.) are provided by available technologies.

There were two prerequisites for the evaluation of answers to the CfE: the proposals/tools considered had to be under active development and maintenance, and they had to be reliable.

The CfE defined a number of criteria for the evaluation of proposals:

* Compression factor
* Separate assessment of performance for each class of data in the compressed bitstream:
  + Reads headers/identifiers
  + Sequence reads
  + Quality scores
  + Any other metadata (identified as “auxiliary data”)
* “Reasonable” computational complexity
  + Encoding and decoding time, respectively
  + Peak and average memory usage
* Support of a minimal set of functionalities
  + Non-sequential access
  + More than 5 symbols (A, C, G, T, N) alphabets
  + Encoding of additional metadata (extensibility)
  + Lossy compression of metadata
    - Quality scores
    - Reads identifiers

In the scope of the CfE, 22 tools have been evaluated. Eight tools have been developed by four providers who directly answered the CfE.

# Raw data

## Low coverage human sample

Figure 1 shows the compression results for unaligned data, for the human low coverage (8x) sample ERR174310.

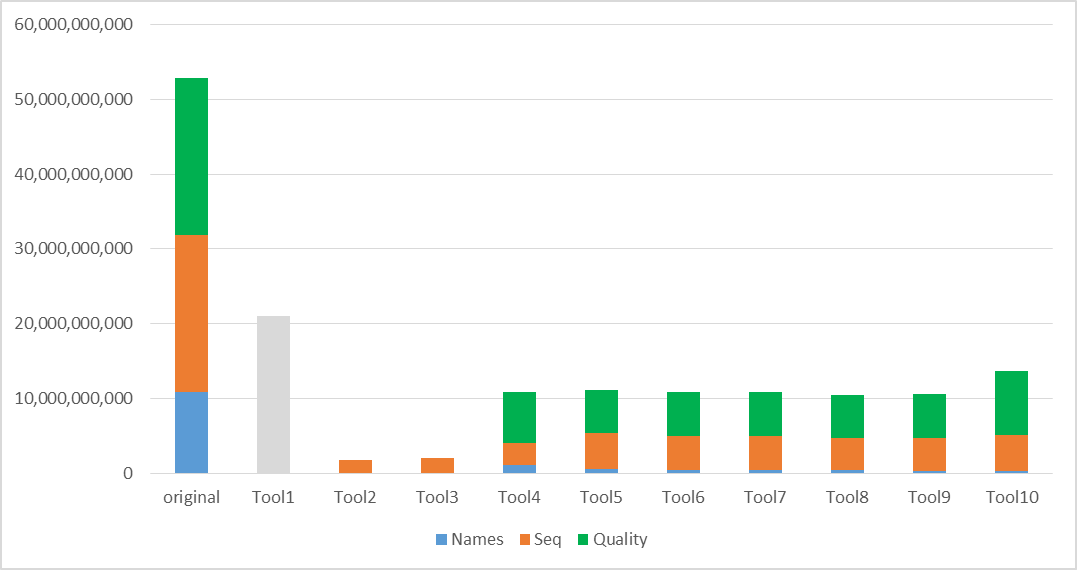


Figure 1 – Compression results for ERR174310

Tool1 in figure 1 represents the currently most used compression solution for raw sequence data and is represented in grey as it does not allow separation of the four classes of data in the encoded file.

Figure 1 shows that some of the tools only support compression of one class of data (sequences), but they have been retained since their performance is the best among all the measured tools.

## High coverage human sample

In order to evaluate how the different solutions are able to exploit the redundancy of sequence data at very high (> 30x) coverage, compression tests have been performed on a set of FASTQ files of about 1TB (files ERR174324 to ERR174341 in set ERS179577). Results for the overall encoded file and the 4 classes of data are shown in figure 2.

|  |  |
| --- | --- |
| Tool1  Tool2  Tool3  Tool4  Tool5  Tool6  original data |  |

Figure 2 - Compression performance on high coverage human samples

# Aligned data

## High coverage human sample

Figure 3 shows the compression results for aligned data, specifically for the human high coverage (50x) sample NA12878\_S1.

Figure 4 reports the same results, but without quality scores. This shows how relevant is considering solutions supporting lossy compression of metadata.

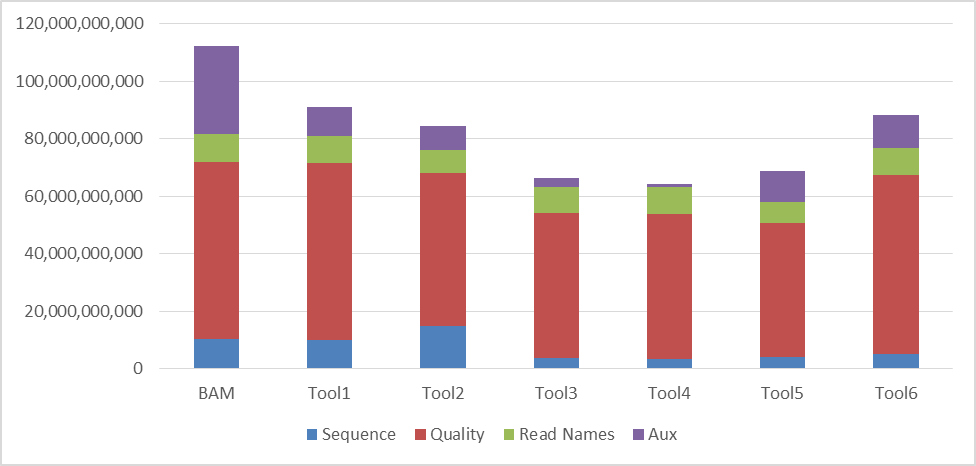


Figure 3 – Compression results for NA12878\_S1

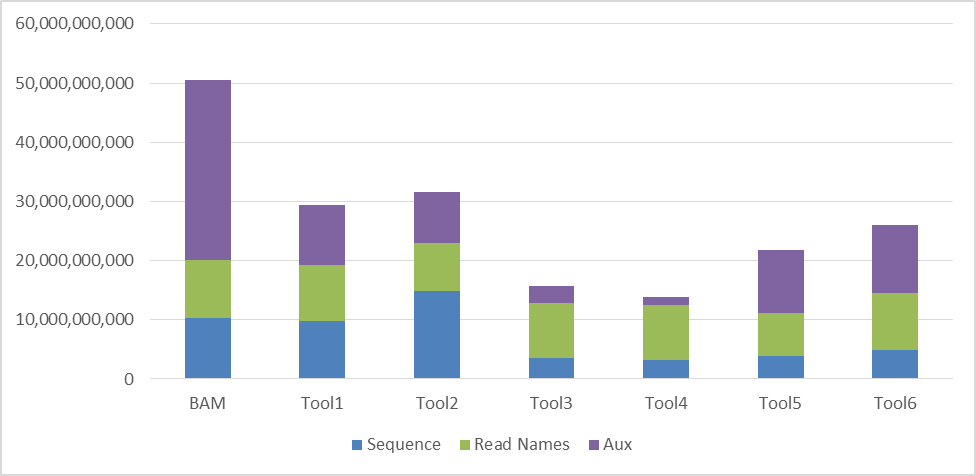


Figure 4 – Compression results (without quality scores) for NA12878\_S1

# Computational complexity

The Call for Evidence respondents provided information to assess the complexity of the proposed solutions. This included encoding and decoding time and average memory footprint on a typical Intel x86 architecture for several available configurations of the tested solutions. Table 1 and Table 2 show the tools performance for the encoding and decoding of a FASTQ file from items 01 and 02 listed in document N16145 [2].

In terms of encoding and decoding time the tools performance were comparable without any tool being outstanding with respect to the others. The fastest encoder was 5 times faster than the slowest while the fastest decoder was 10 time faster than the slower. The latter is due to the fact that some of the tools have already been optimized for parallel processing while others are still under development and optimization.

Memory footprint presented a higher variance (up to a factor of 1,000), but the measured memory requirements are largely satisfied by average computer architectures (6 GB max required).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Encoding** | | **Decoding** | |
|  | | **Time** | **Memory (MB)** | **Time** | **Memory (MB)** |
| **Tool1** | | 2:08:45 | 5.70 | 0:15:48 | 5.70 |
| **Tool2** | | 1:44:03 | 12.00 | 0:44:44 | 5.70 |
| **Tool3** | | 1:15:46 | 5,322.10 | 1:12:19 | 1,006.90 |
| **Tool4** | **V1** | 0:25:43 | 20.00 | 0:19:13 | 29.40 |
| **V2** | 1:26:21 | 485.60 | 1:31:24 | 392.50 |
| **Tool5** | **V1** | 0:59:01 | 399.00 | 1:22:28 | 389.60 |
| **V2** | 1:08:08 | 775.00 | 1:24:02 | 772.10 |
| **Tool6** | **V1** | 0:40:41 | 77.50 | 0:52:06 | 65.40 |
| **V2** | 0:44:05 | 330.50 | 0:54:53 | 318.40 |
| **Tool7** | | 1:00:02 | 79.00 | 1:15:15 | 78.90 |
| **Tool8** | | 2:37:02 | 682.70 | 1:30:42 | 163.00 |
| **Tool9** | | 1:25:27 | 527.20 | 0:13:07 | 547.00 |

Table 1 - Tools performance for the compression of a FASTQ file of 53 GB (ERR174310\_1.fastq)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Encoding** | | **Decoding** | |
|  | | **Time** | **Memory (MB)** | **Time** | **Memory (MB)** |
| **Tool1** | | 2:13:38 | 5.90 | 2:22:07 | 5.90 |
| **Tool2** | | 7:50:19 | 45.80 | 2:19:10 | 65.50 |
| **Tool3** | **V1** | 1:44:21 | 1,143.90 | 1:58:34 | 595.00 |
| **V2** | 1:42:08 | 384.70 | 1:51:16 | 275.80 |
| **Tool4** |  | 7:32:05 | 465.40 | 8:37:55 | 459.20 |
| **Tool5** | **V1** | 4:20:09 | 2,269.20 | 3:58:47 | 5,642.90 |
| **V2** | 6:46:54 | 2,713.30 | 7:45:14 | 5,906.80 |

Table 2 - Tools performance for the compression of a SAM file of 589 GB (SAM file decoded from NA12878\_S1.bam)

# Conclusions

One conclusion of this activity is that all tools have different performances for the 4 classes of data. By integrating multiple tools it is possible to improve the compression of up to 27% with respect to the best state-of-the-art tool. During the time frame of the CfE, existing tools as well as new tools improved the obtainable compression factors by about 4%.

Results demonstrate the possibility to achieve compression rates above the state of the art through the combination of new and improved algorithms.

Technology is therefore mature enough to move forward and issue a Call for Proposals at the next MPEG meeting in Geneva (30th May – 3rd June 2016).

# Acknowledgements

MPEG would like to thank Simon Fraser University, the Wellcome Sanger Trust Institute, the M.I.T., Stanford University, Leibniz Universität Hannover and EPFL for submitting technology and contributing to the evaluation.

# References

|  |  |
| --- | --- |
| [1] | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2015/N15740 - Call for Evidence on Genome Compression and Storage,” Geneva, 2015. |
| [2] | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2016/N16145 - Database for Evaluation of Genome Compression and Storage,” San Diego, 2016. |
| [3] | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2014/N14553 - Issues in genome compression and storage,” Sapporo, 2014. |
| [4] | R. Canovas, A. Moffat and A. Turpin, “Lossy compression of quality scores in genomic data,” *Bioinformatics,* vol. 30, no. 15, pp. 2130-2136, 2014. |
| [5] | M. Hernaez, I. Ochoa, R. Goldfeder, T. Weissman and E. Ashley, “Effect of lossy compression of quality scores on SNP,” *Biorxiv,* 2015. |
| [6] | S. D. Kahn, “On the Future of Genomic Data,” *Science,* vol. 331, pp. 728-729, 2011. |
| [7] | Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron and G. E. Robinson, “Big Data: Astronomical or Genomical?,” *PLOS Biology,* 2015. |
| [8] | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2016/N15XXX - Results of the Call for Evidence on Genome Compression and Storage,” San Diego, 2016. |
| [9] | MPEG Requirements, “ISO/IEC JTC 1/SC 29/WG 11 MPEG2016/N15XXX - Evauation Criteria of the Call for Proposals on Genome Compression and Storage,” San Diego, 2016. |