# MPEG Seminar on Genome Compression Standardization

**14:00 - 18:00, 23rd February 2016**
**San Diego Marriott La Jolla**

# Seminar Program

- Marco Mattavelli – MPEG AhG Chair
  - **MPEG Experience in Genomic Information Representation**
- Come Raczy – Illumina
  - **Compression**
- Bayo Lau – Bina Technology/Roche Sequencing
  - **Practical Considerations of Genomic Data**
- Cenk Sahinalp - Simon Fraser University Vancouver
  - **High Throughput Sequencing Compression – State of the Art**
- Yong Zhang – Co-convenor of TC 276 “Biotechnology” WG 5 “Data Processing and Integration”
  - **The Time of Peta-byte Is Coming. Challenges and Opportunities in Big BioData.**
- Joern Ostermann – MPEG Requirements Chair
  - **MPEG Workplan for Genome Compression Standardization**

# MPEG Experience in Genomic Information Representation

Marco Mattavelli - EPFL

La Jolla – 2016/02/23

# Moving Picture Experts Group (MPEG)

- Established 28 years ago (January 1988)
- Developed standard for media conversion from analogue to digital (MP3, MPEG-2, MPEG-4, DASH etc.)
- Attendance: ~400 experts from 25 countries and ~250 companies and organisations
- Meetings: 3/4 meetings a year

# MPEG is not “just” about compression

- MPEG standards deal also with such issues as:
  - Random access
  - Scalability
  - Complexity
  - Reconfigurability
  - Transport
  - APIs
  - System architecture
  - …..

# MPEG working method

- Identify **areas** needing standards
- Liaise with affected **communities**
- Develop and publish **Requirements**
- Collect **Test Material**
- Issue **Call for Evidence/Proposals**
- Develop and publish **Test Model** (text and software)
- Optimise Test Model, update **Working Draft**
- Develop formal ISO standards while **continuing optimization**

# Where are we with genome compression?

- **Requirements** document produced
- **Contacts** with relevant bodies established
- **Community** of interested experts created
- **Test data set** defined
- One **workshop** held in Geneva (Oct. 2015)
- **Compression experiments** carried out
- **Call for Evidence** issued Oct. 2015 – results available at this meeting.
- All documents available at: http://mpeg.chiariglione.org/standards/exploration/genome-compression

# Main compression and processing **requirements**

- **Requirements classified into 4 categories**:
  - Compression of raw reads
  - Compression of mapped/aligned reads
  - Transport
  - General digital data management

# Test data set

- A common data set for assessing and comparing compression algorithms. It includes:
  - Organisms:
    - Homo Sapiens (low-medium-high coverage), Bacteria, Plants
  - Experiments:
    - Metagenomic, Cancer cell lines
  - Sequencing technologies include:
    - Illumina HiSeq®, Pacific Biosciences SMRT®, Oxford Nanopore, Ion Semiconductor (Life Technologies)
  - Dataset will be further completed at the end of this San Diego meeting

# Purpose of the Call for Evidence

- To assess whether new technologies can achieve better performance than state of the art tools for raw and aligned data (BAM)
- To understand which additional functionalities (e.g. non sequential access, lossy compression efficiency, etc. ) can be provideded by available technologies

# Answers to the CfE

- Comprehensive review/assessment of state-of-the-art methods on the “MPEG genome dataset”
- Several institutions have contributed:
  - EPFL, MIT, Stanford University, Simon Fraser Univ., Wellcome Trust Sanger Institute, EBI, SIB, Leibniz Univ. Hannover, ….
- Results have been just been collected and a document will be soon available
  - Widest known coordinated effort to assess performance and limits of state-of-the-art compression technology applied to genomic data and metadata

# Evaluation Criteria 1/2

- 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”)

# Evaluation Criteria 2/2

- “Reasonable” computational complexity
  - Encoding/decoding time
  - 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

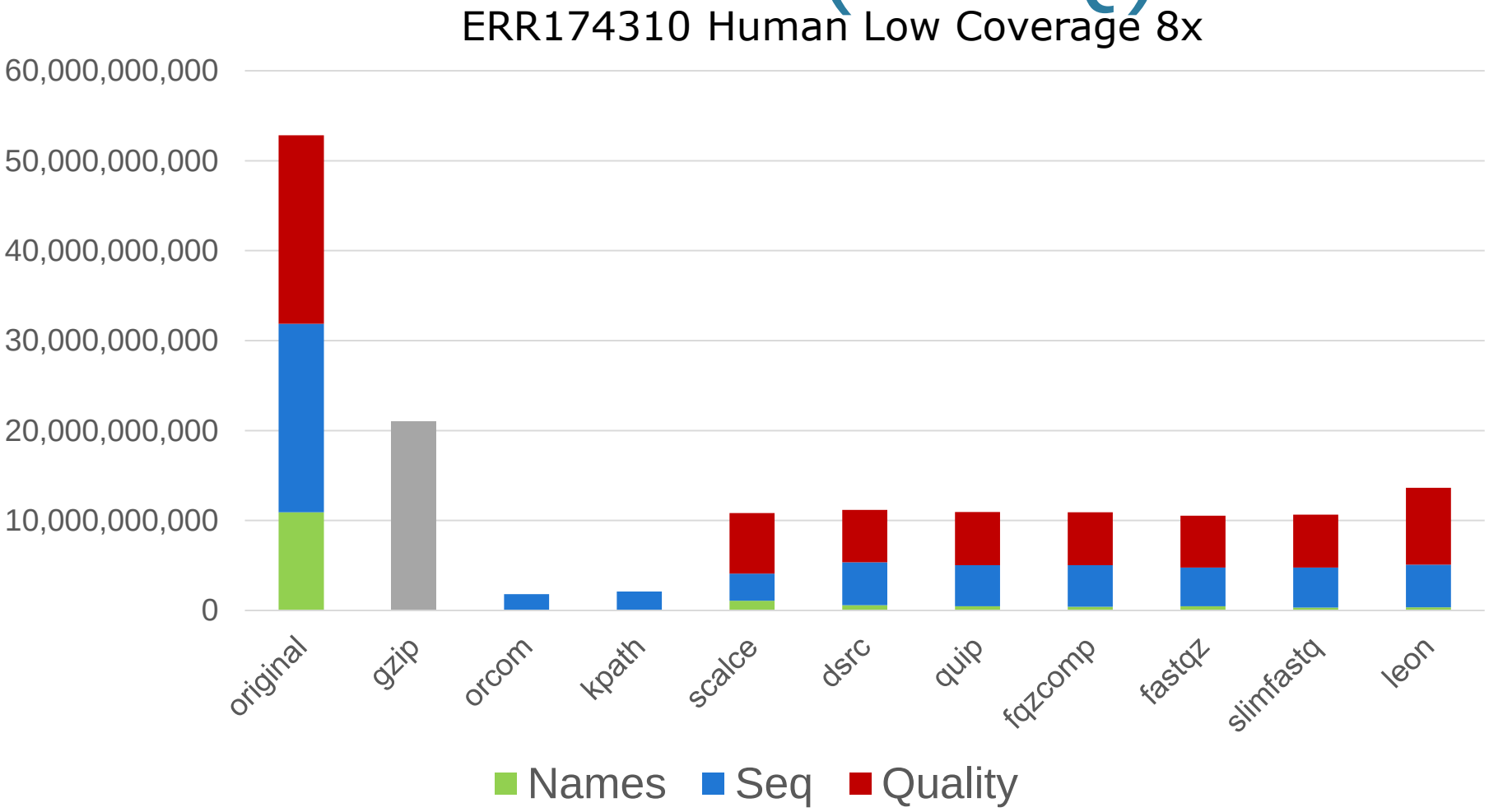
# Results

- 22 tools identified and evaluated

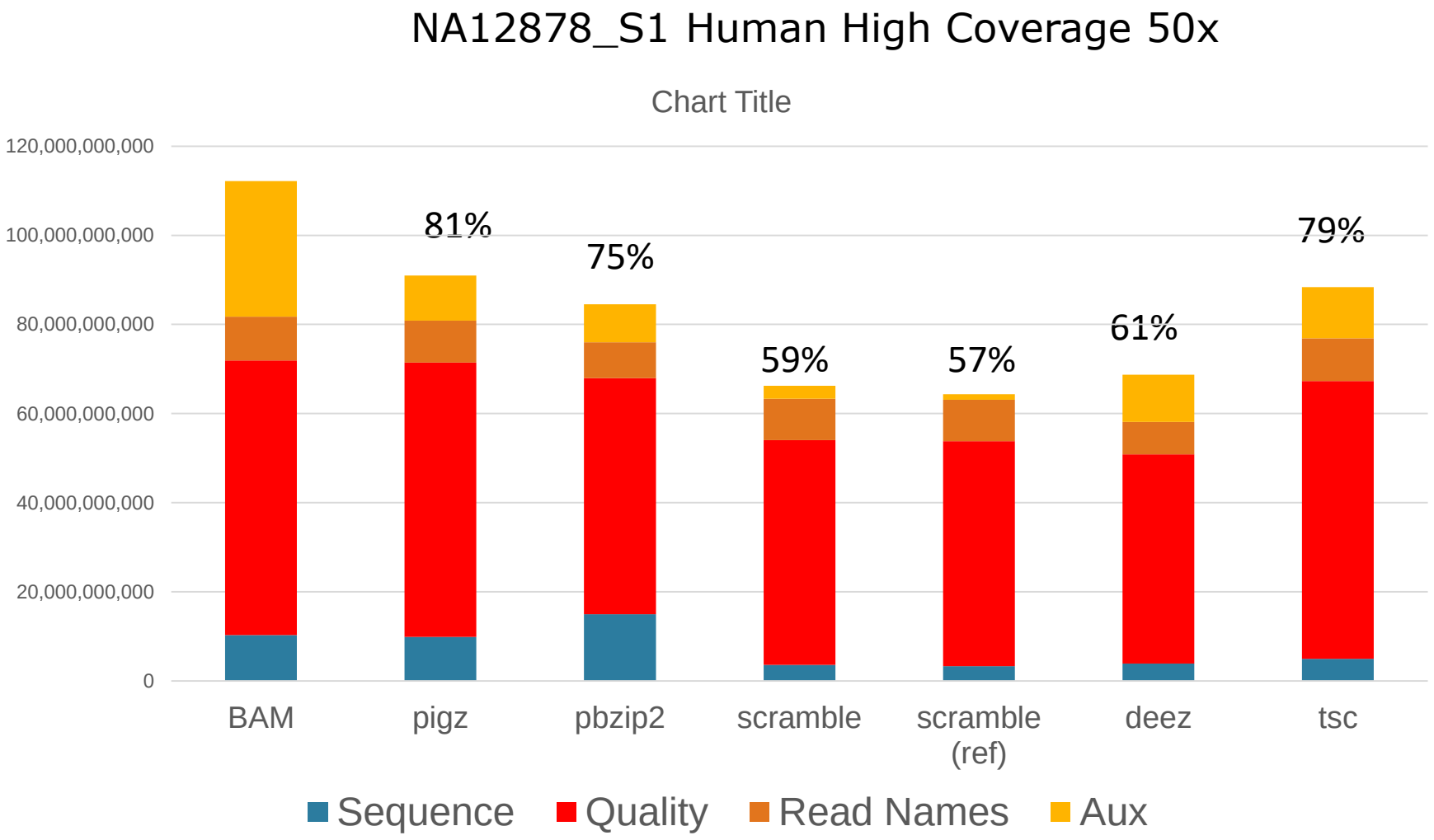
- 8 tools from 4 groups who directly answered the CfE (with improvements!)
  - Wellcome Trust Sanger Institute
  - Simon Fraser University Vancouver
  - Stanford University
  - Leibniz University Hannover

  - Compression Comparison Table.xlsx

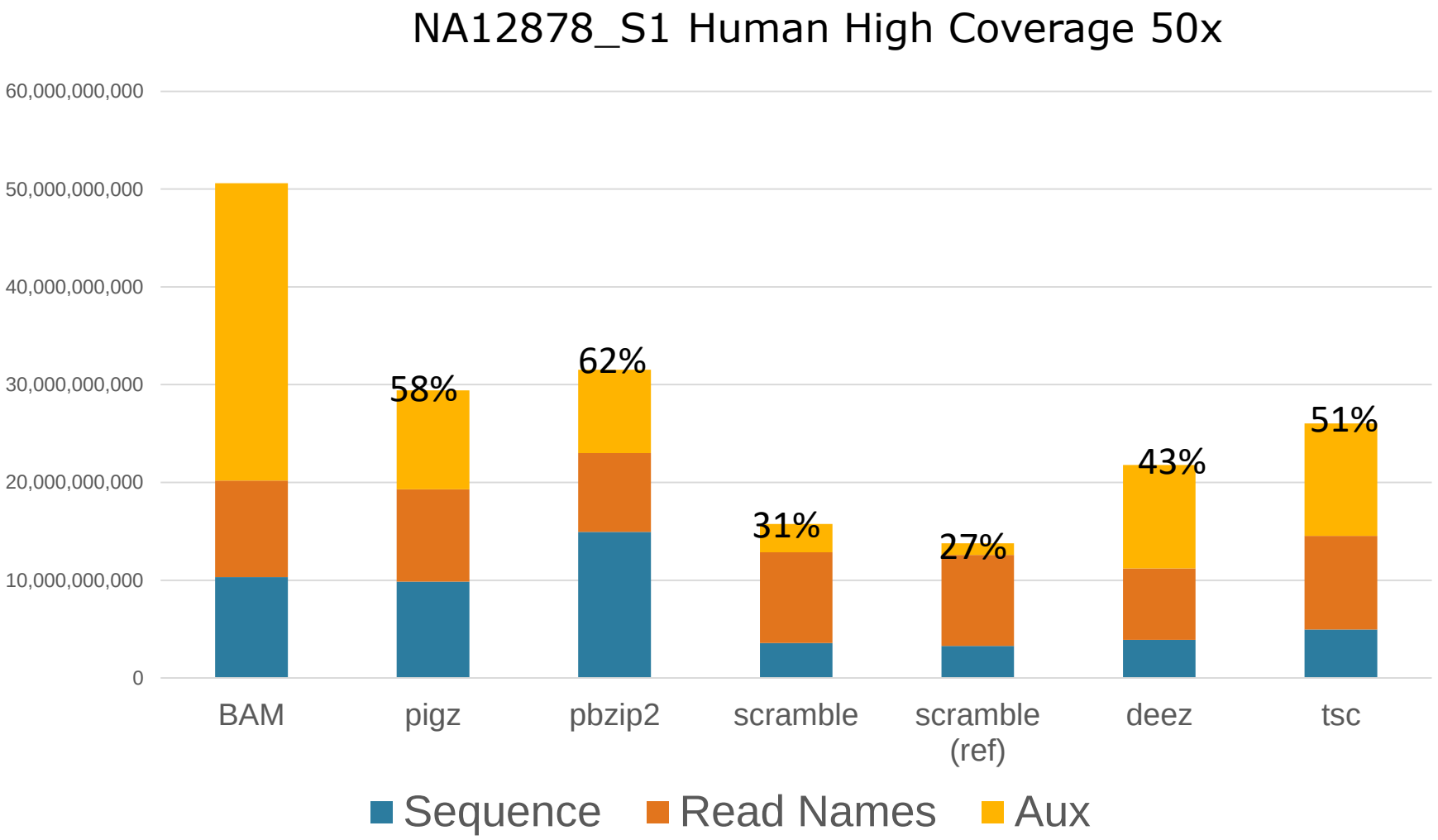
# Raw data (FastQ)



# Aligned data (BAM)



# Aligned data (BAM) without QS



# Features for aligned data

| | Random Access | More than 5 symb | Add. Meta-data | Lossy Comp | Req. 1 | Req. 2 | Req. …. | Req. n |
|---|---|---|---|---|---|---|---|---|
| pigz | N | Y | Y | N | N | ? | ? | ? |
| pbiz2 | N | Y | Y | N | N | ? | ? | ? |
| Scramble | Y | Y | Y | Y | N | ? | N | N |
| Scramble (ref) | Y | Y | Y | Y | N | N | ? | ? |
| deez | Y | N | Y | Y | N | N | N | ? |
| tsc | Y | Y | Y | N | N | N | N | N |

| | Random Access | More than 5 symb | Add. Meta-data | Lossy Comp | Req. 1 | Req. 2 | Req. …. | Req. n |
|---|---|---|---|---|---|---|---|---|
| **MPEG-G** | **Y** | **Y** | **Y** | **Y** | **Y** | **Y** | **Y** | **Y** |

# Lossy Compression

- Lossy is applied to metadata only!!! Not to nucleotides!!
- The main idea is to compress metadata so that results of downstream, analysis is not affected:
  - Variant Calling
  - Alignment
- An evaluation framework has been defined and the goal is the “intrinsic” definition of a “rate distortion” function for Quality Values
- Comparison of lossy compressors will be based on such implicit rate distortion function