A Review of Compression Technology of Genomics Data

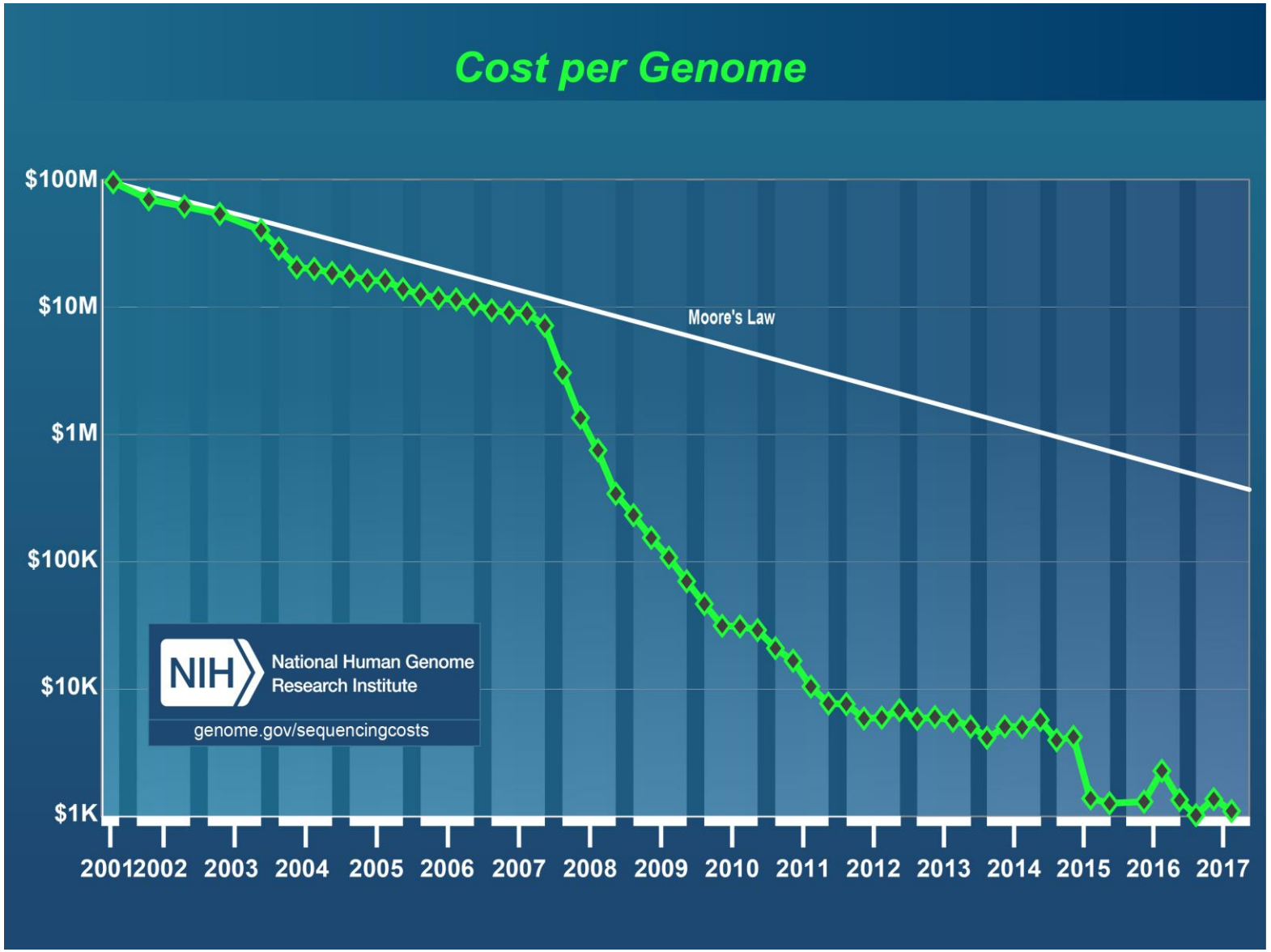
Yong Zhang

Agenda

- **Background: data at BGI & CNGB**
- **Dive into Compression Technology of Genomics Data**
- **Outlook of future**

1 Background: data at Genomics

1 Genomics Data in future



Apr-15	$4,211
Jul-15	$1,363
Oct-15	$1,245
Apr-16	$1,297
Jul-16	$2,257
Oct-16	$1,356
Jan-17	$1,015
Apr-17	$1,323
Jul-17	$1,121
BGI-seq	$600

100GB* 10,000,000 = 1 EB= 1000PB

100GB* 3,500,000,000 = 350EB

Half of the living life…

2 Dive into Compression Technologies of Genomics Data

2 Genomics Data format

•**readID:**
@HWUSI-EAS100R:6:73:941:1973#0/1

•**Sequence:**
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

•**Sign:**
+

•**Base quality:**
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

$$Q_{\text{sanger}} = -10 \log_{10} p$$

HOW TO COMPRESS ?

3 STAGES

2 Stage 1: general-purpose tools

- gzip (LZMA, Huffman)
- bzip2 (BWT, Huffman)
- 7zip (LZMA, Arithmetic)

Compress Ratio: ~ 3.3x

Processing Speed: Dozens to a hundred MB/s

Acceleration method： GPU or FPGA

2 Stage 2: ID & (Seq+Qual)

Similarity of ID

@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=72
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=72
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=72
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=72

- Method 1:
- Format + variation

@SRR001666.[] 071112_SLXA-EAS1_s_7:5:1:[]:[] length=72
---
1 817 345
1 817 345
2 801 338
2 801 338

- Method 2:
- Delta encoding

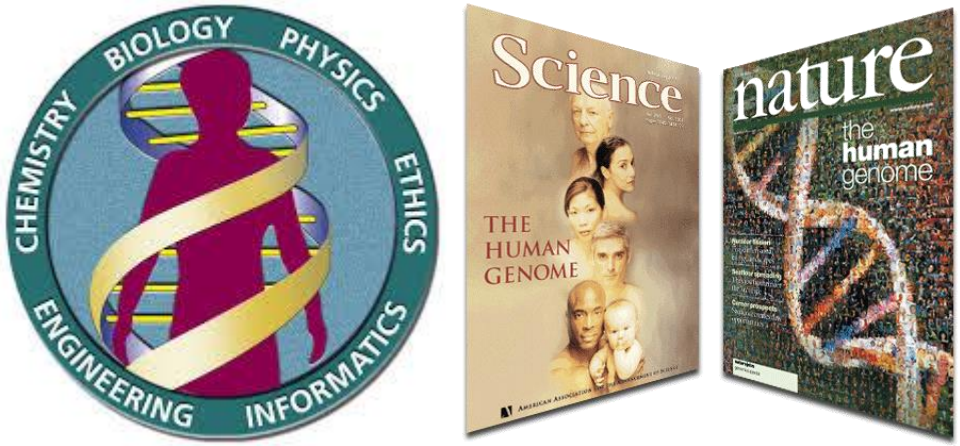
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=72
[55]
[11]”2” [38]”01” [2]”38” [10]
[55]

Compression ratio: ~ 8x

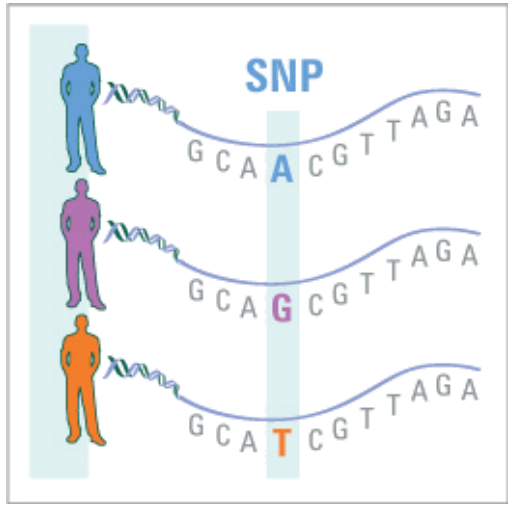
2 Stage 2: ID & (Seq+Qual)

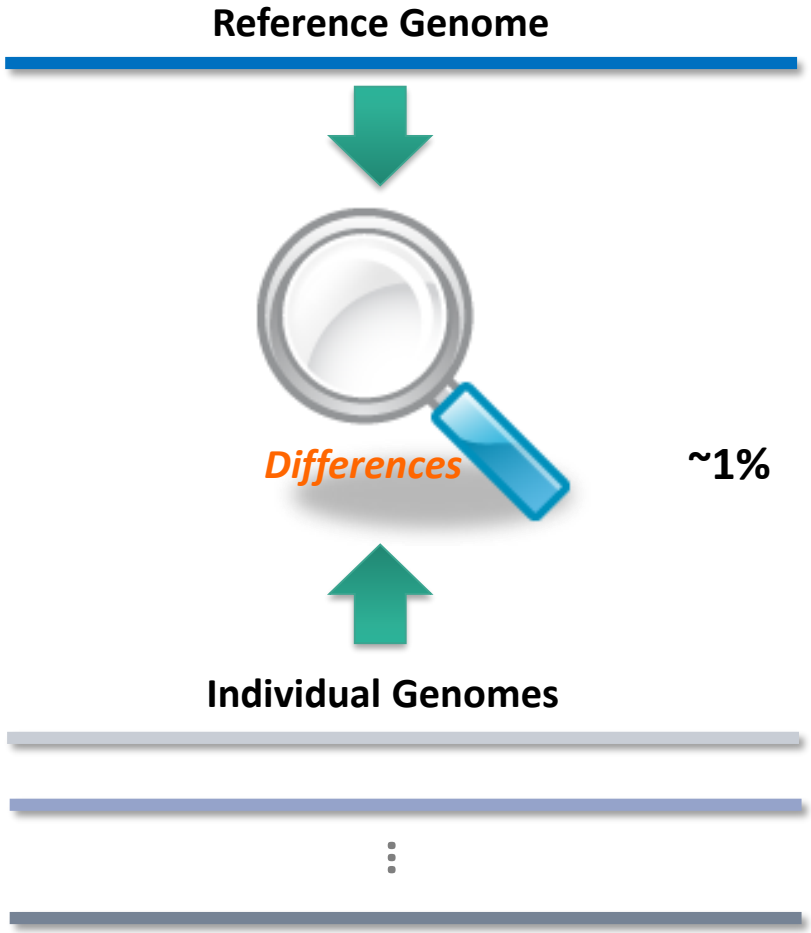
- Seq+Qual compression
  - G-SQZ
    - <base, quality> as unit & Huffman coding
  - SeqDB
    - 1 byte (256) = ACGTN (5) * Q (51)

2 Feature of sequence



Human Genome Project (HGP) 1990–2003





2 Stage3 : ID & Seq & Qual

- Seq (ATCGN)
  - Method 1: Ref-based (alignment & index)
    - Seq --> Ref_Pos + Strand + Cigar
    - e.g. ref: GTAGTATCGACC seq: ATCGAT -> 6, positive, 5T
    - Alignment
    - Encode Ref_Pos , Strand, Cigar separately
      - GenCompress
        - Pos/RelPos
        - Distance & nucleotide substitution
      - LWFQZip
        - Pos, flag, Mlength, Mtype, MisValues

Compression ratio: ~12.5x from fqzcom

2 Stage3 : ID & Seq & Qual

- Seq (ATCGN)
  - Method 2: De novo (redundancy removal)
    - 1. entropy encoding
      - KungFQ: 1 byte (7 for 3mer/RLC, 1 for diff) + LZMA
      - Fqzcomp: k-order model for prediction
    - 2. build ref and index
      - Vishal's work*: dynamic Dict following FIFO
      - ReCoil: similar graph based on all reads
    - 3. cluster and encoding
      - SCALCE: 4-mers obeying LCP for clustering
      - ORCOM: minimizer

Compression ratio: 5.5x from fqzcom

* Bhola, Vishal, et al. "No-reference compression of genomic data stored in fastq format." *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*. IEEE, 2011.

2 Stage3 : ID & Seq & Qual

- Seq (ATCGN)
  - Method 1: Ref-based (alignment & index)
  - Method 2: De novo (redundancy removal)
  - Pros and Cons

  - Combination of Method 1&2
    - Fritz's work*
      - Aligned: index
      - Not aligned: build a new ref and index

*Fritz, Markus Hsi-Yang, et al. "Efficient storage of high throughput DNA sequencing data using reference-based compression." *Genome research* (2011).

2 Stage3 : ID & Seq & Qual

- Qual (0~40)
  - Feature: nearly no pattern
  - Feature: decrease with read position; similar with nearby qualities

  - Method: entropy encoding
    - Huffman (LFQC) or RunLength (KungFQ)
    - Quip: 3-order Markov model
    - Fqzcomp: Q_{i-1}

      $\max(Q_{i-2}, Q_{i-3})$

      $[Q_{i-2} = Q_{i-3}]$

      $\min(7, [\frac{1}{8}\sum_{j-2}^{i} \max(0, Q_{j-2} - Q_{j-1})])$

      $\min(7, [i/8])$

2 Stage3 : ID & Seq & Qual

- Qual (0~40)
  - Lossy compression
    - Method1: merge
      - equally, e.g. [01234] -> 2
      - unequally
        - lower quality, more bins
        - lower quality, less bins
    - Method2: replace with similar value within a block
      - 1. PBlock: max-min < threshold
      - 2. RBock: max/min < threshold
    - Other Methods

2 BAM & CRAM formats

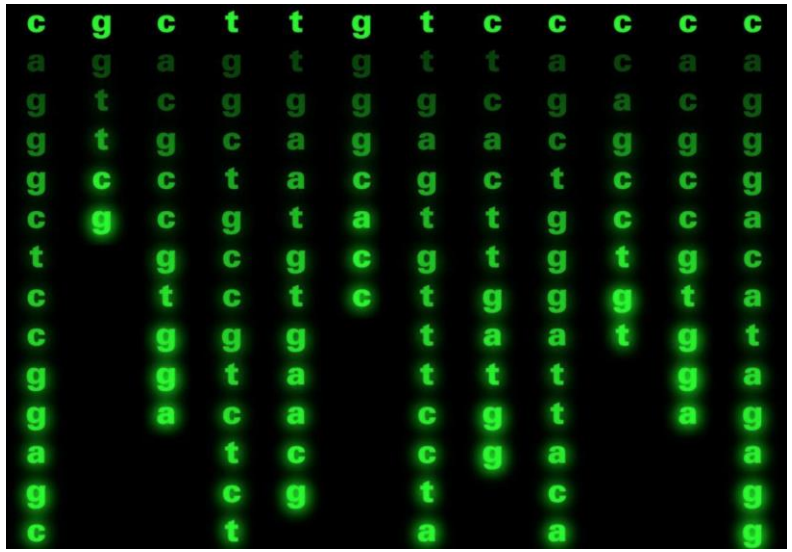
- BAM
  - Combine alignment info with fastq, compression ratio: ~3x
  - Easy to index and query by alignment info

- CRAM
  - Optimized for BAM: reference based sequence compression
  - Compression ratio: ~5.5x

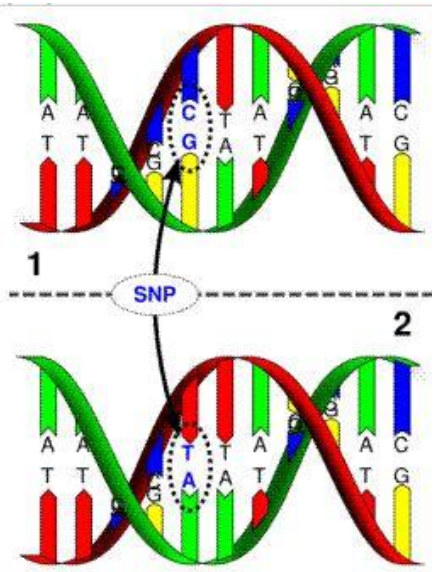
3 Outlook of future

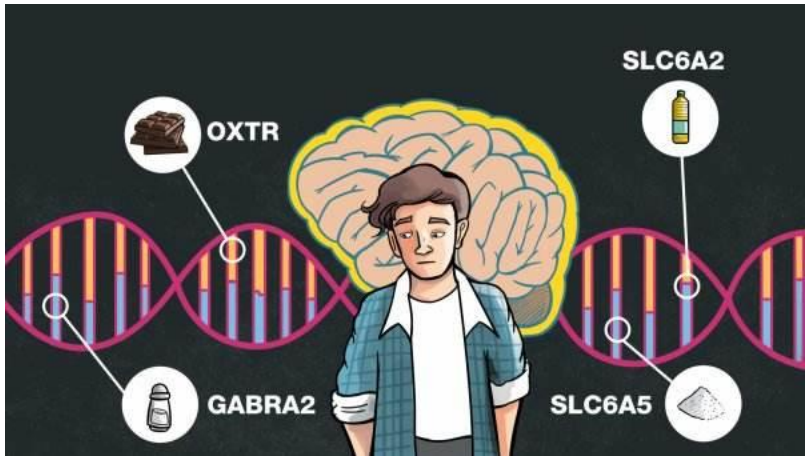
Genomics Data, Information and knowledge

- Nobody understand this…
- Compute one time and become code data for most of data.
- Can not be deleted or lossy compressed for future research purpose.
- **Low cost file format for archiving**



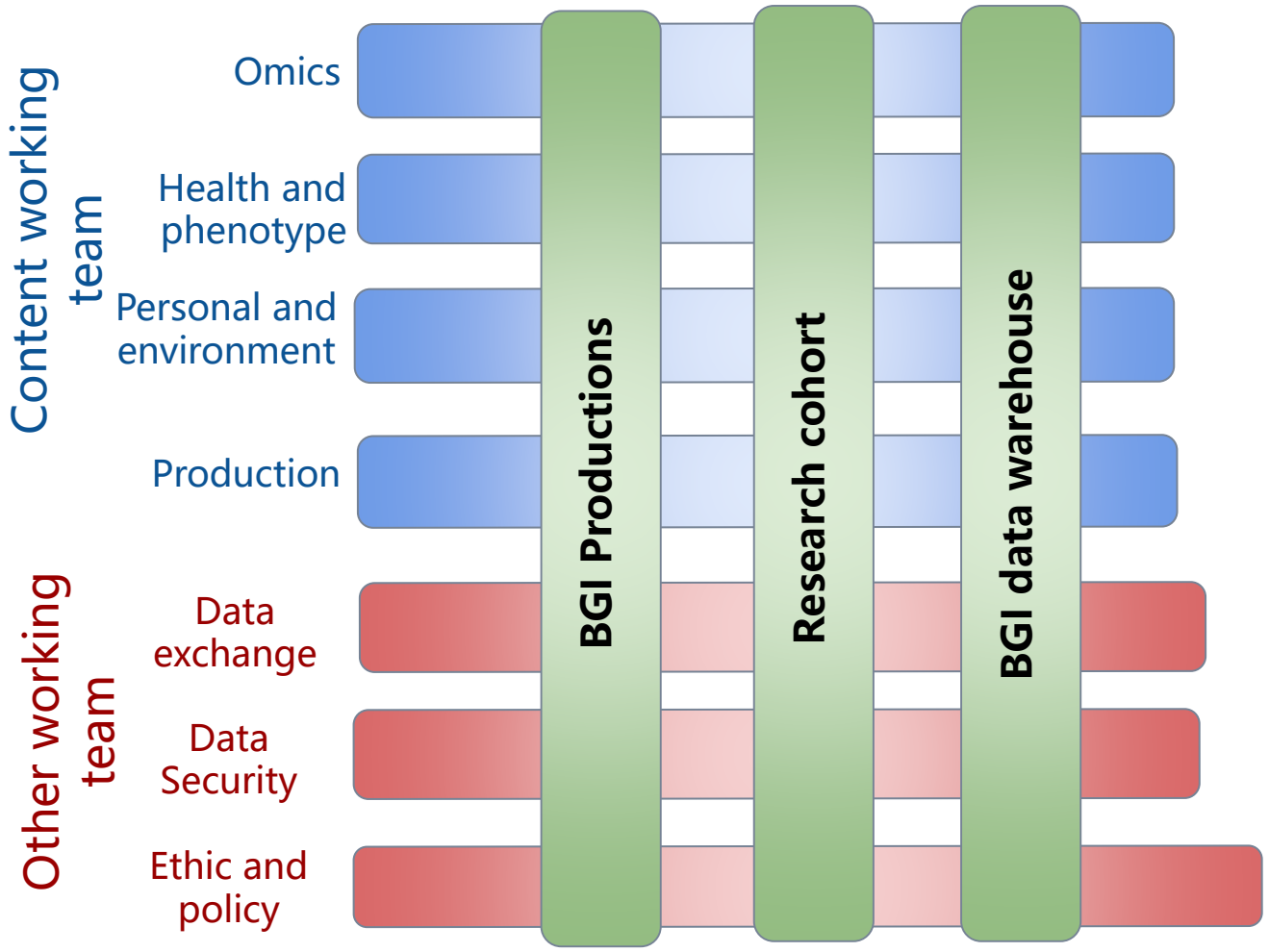
- Professional understand this…
- Cloud be active data for a long time for different research and health production.
- Better keep the Info than delete it and compute again.
- **Easy accessible file format for querying and exchanging**



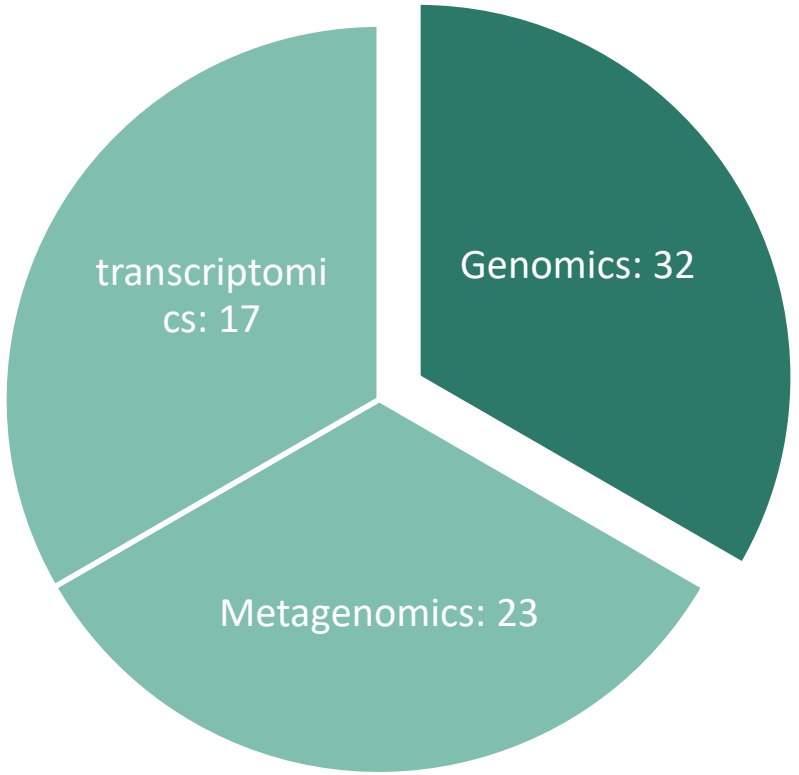


Data → Info → Knowledge

3 BGI-GOLD (Genomics Of Life Data) Standard: from Genomics to Omics integration



GOLD working group



Total: 107 kinds of file formats
Total: 6538 data elements

Thank You

Q&A