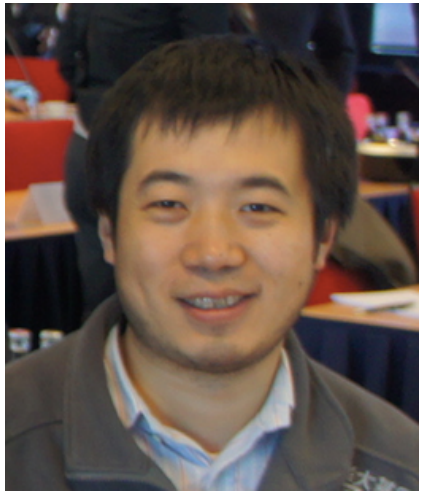
The Time of Peta-byte Is Coming

-Challenges and Opportunities in Big BioData

Yong ZHANG (yongzhangcn@qq.com)
Co-Convenor of TC 276/WG 5 (data processing and integration)
2016.01.23

Yong ZHANG, Ph.D., Prof.



- Bachelor in Software, Peking University, China
- Ph.D. in Bioinformatics, Peking University, China
- Postdoc. in Proteomics, Max-Planck-Institute of Biochemistry, Germany
- 2001-2015, Group Leader, Assistant Director, VP, BGI
- 2011-2014, Head of China National GeneBank (Founder)
- 2013-NOW, Professor, Huazhong Sci & Tech University, China
- 2013-NOW, Adjunct Professor, University of Alberta, Canada
- 2014-NOW, Co-Convenor of WG 2 (biobank and bioresources) and Co-Convenor of WG 5 (data processing and integration) of ISO/TC 276.

15 years in bioinformatics, focus on OMICS.

Achievements

- Papers: 65 SCI papers, on Nature, Science, Cell, etc.
- Books: 6 Books in Biobank, Bioinformatics Database, etc.
- Patents: 23
- Standards: 3 Shenzhen City Standards
- Funding obtained: 61 Million USD (from NDRC, MOST, MOH, etc.)

Invitation for this meeting and WG 5 Paris Meeting

- ISO/TC 276/WG 5 Convenor: Martin Golebiewski; Co-Convenor: Yong ZHANG.
- Thanks a lot to Prof. Dr. –Ing. Joern Osterman, for this online presentation.

- Marco Mattavelli marco.mattavelli@epfl.ch from MPEG consortium attended ISO/TC 276/WG 5 Paris meeting last week.

Outlines

- ISO TC276, WG 5
- China National GeneBank
- Challenges and Opportunities
- Compression Work
- Big Datasets

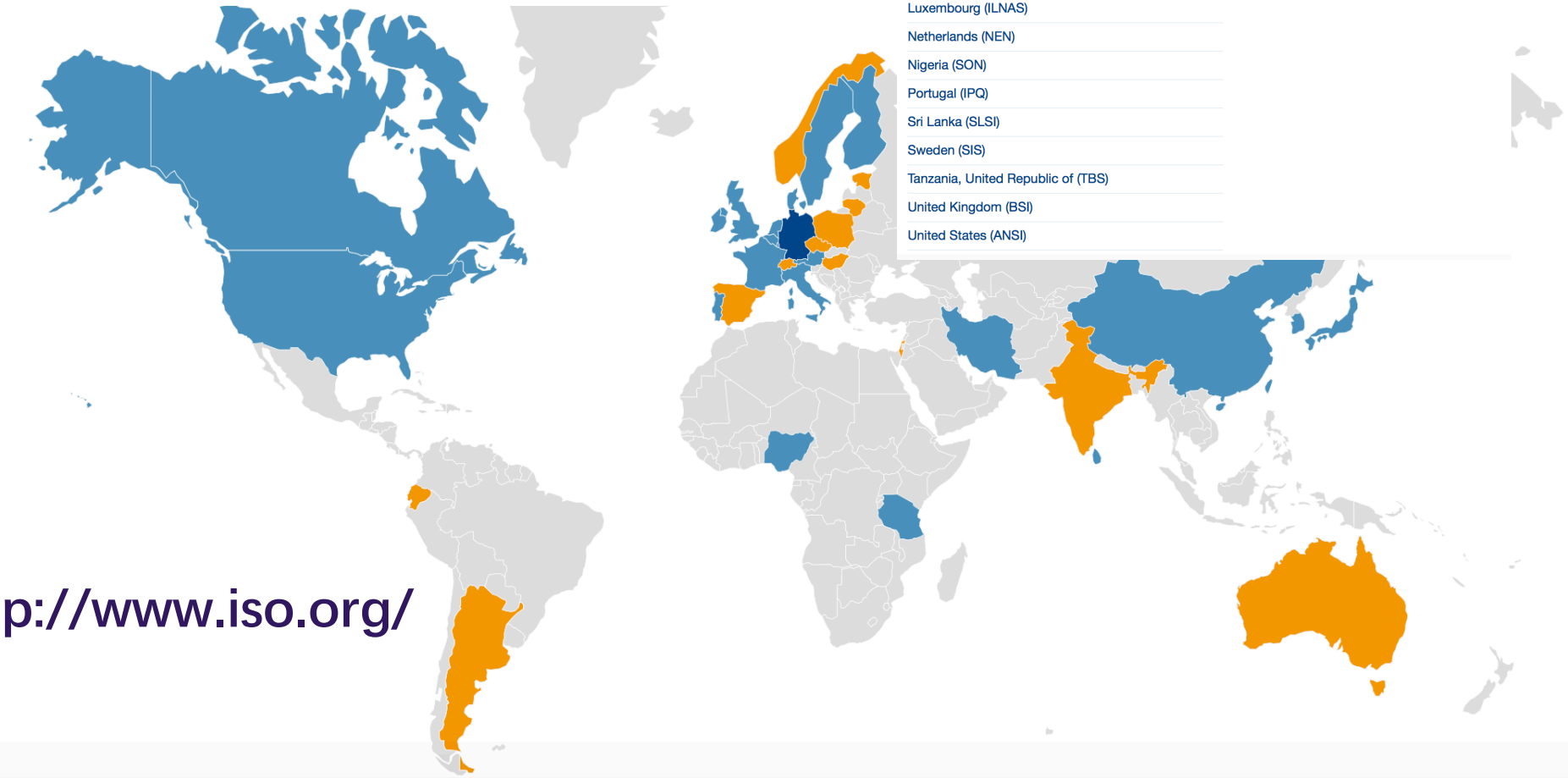
ISO/TC 276 - Biotechnology

Participating Countries (22)

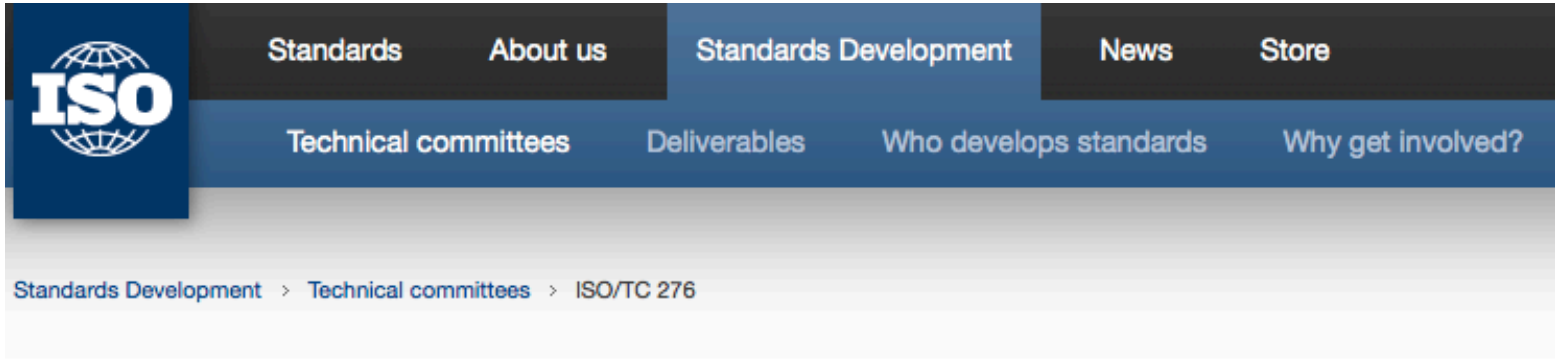
- Austria (ASI)
- Belgium (NBN)
- Canada (SCC)
- China (SAC)
- Denmark (DS)
- Finland (SFS)
- France (AFNOR)
- Germany (DIN)
- Iran, Islamic Republic of (ISIRI)
- Ireland (NSAI)
- Italy (UNI)
- Japan (JISC)
- Korea, Republic of (KATS)
- Luxembourg (ILNAS)
- Netherlands (NEN)
- Nigeria (SON)
- Portugal (IPQ)
- Sri Lanka (SLSI)
- Sweden (SIS)
- Tanzania, United Republic of (TBS)
- United Kingdom (BSI)
- United States (ANSI)

Observing Countries (13)

- Argentina (IRAM)
- Australia (SA)
- Czech Republic (UNMZ)
- Ecuador (INEN)
- Estonia (EVS)
- Hungary (MSZT)
- India (BIS)
- Israel (SII)
- Lithuania (LST)
- Norway (SN)
- Poland (PKN)
- Spain (AENOR)
- Switzerland (SNV)



http://www.iso.org/



ISO/TC 276 Biotechnology

About | Contact details | Structure | Liaisons | Meetings | Tools

Chair: **Ricardo Gent** , Deutsche Industrievereinigung Biotechnologie im VCI e.V. (Germany)

Subcommittees/Working Groups:

Subcommittee/Working Group	Title
ISO/TC 276/WG 1	Terminology Convenor: **Pablo Serrano**, BPI e.V. (Germany)
ISO/TC 276/WG 2	Biobanks and bioresources Convenor: **Georges Dagher**, INSERM (France)
ISO/TC 276/WG 3	Analytical methods Convenor: **Sheng Lin-Gibson**, NIST (USA)
ISO/TC 276/WG 4	Bioprocessing Convenor: **Tatsuo Heki**, FUJIFILM (Japan)
ISO/TC 276/WG 5	Data processing and integration Convenor: **Martin Golebiewski**, HITS (Germany)

Developing an ISO standard for applying and connecting community modelling standards

ISO/TC 276 Biotechnology WG 5 (Data Processing and Integration) has started to work on a draft for a new ISO standard in the life sciences:

„Minimal requirements for downstream data processing and integration workflows for interfacing and linking heterogeneous data, models and corresponding metadata "

Framework („hub ") standard making references to existing community standards:

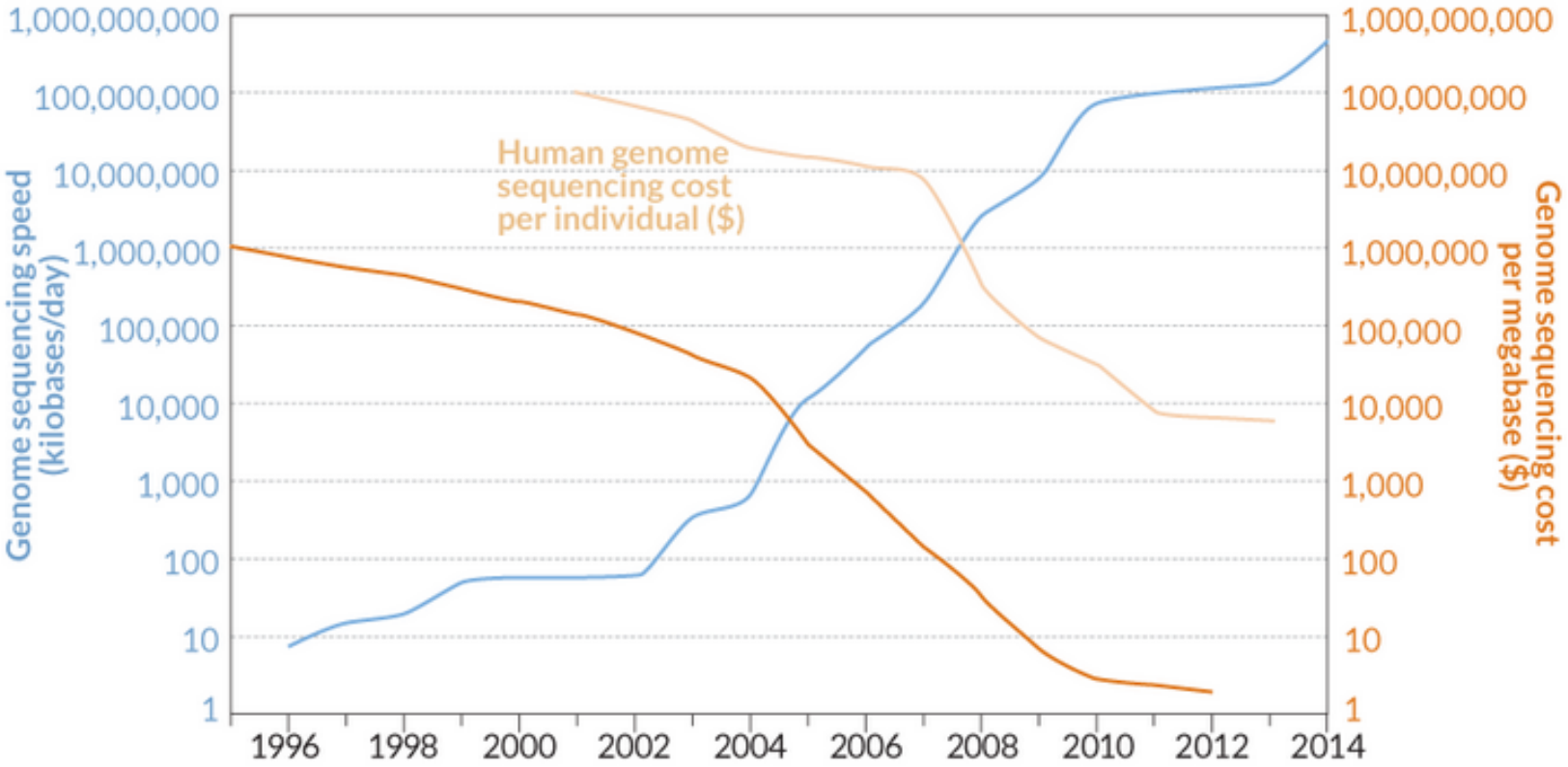
- References to community formats for life science data and computer models
- References to minimal reporting guidelines for data, models and metadata
- Definition of interfacing between data sets, models and metadata
- Standards for describing workflow elements and their interplay in modelling biological systems

ISO/TC 276/WG 5 (data processing and integration)

- ISO/TC276 Biotechnology
  - WG 5: data processing and integration
- ISO/TC 276/WG 5 Paris Meeting (2016-02-19)

- **RECOMMENDATION 1/2016/01 taken by ISO/TC 276/WG 5 on 2016-02-19**
- ISO/TC 276/WG 5 recommends developing a joint project in the field of "genome compression standardization" together with ISO/IEC JTC 1/SC 29/WG 11 (MPEG consortium). ISO/TC 276/WG 5 recommends establishing a joint ad-hoc group to prepare an NWIP for this project.
- **RECOMMENDATION 2/2016/01 taken by ISO/TC 276/WG 5 on 2016-02-19**
- ISO/TC 276/WG 5 recommends having a joint session with WG 3 at the upcoming combined meetings of ISO/TC 276/WGs on Tuesday, May 10th 2016 (3 to 5pm) in Washington DC, USA, to discuss a joint project with the MPEG consortium on "genome compression standardization." Marco Mattavelli (ISO/IEC JTC 1/SC 29/WG 11) will be invited to this meeting.

Data explosion, and Cost drop down



NOW: 1,500 human genome per day.
$1,000 per human genome (30x)
1PB = 1,000 TB = 1,000,000 GB (1 million GB)

Biology Data (Information)

- The international human genome: 10 years, $ 3 billion
- Now, $ 10,000 and 1 day
- Next 2-3 years, $1,000 and 1 hour

- In 2010, BGI generated 500 TB sequence data, equal with 10 times of NCBI total data till 2009
- In 2011, BGI generated 3 PB

* 16 Feb, 2011, NCBI announced that they won’ t accept NGS sequence data any more due to budget limitation

My Slide in 2011

China National GeneBank (CNGB, www.cngb.org), founded in 2011, was approved by the Central Government of China, and operated by BGI-Shenzhen, which combines biological bank, informatics bank and national-wide consortium/alliance.

Phase I (2011-2016) investment: 124 million US dollar.

My Slide in 2011

国家基因库
CHINA NATIONAL GENEBANK

China National GeneBank (Goals)

- Resource DB (BioBank)
  - 1 million samples (up to 30 million)
  - Blood to cell, seed to tissue, of human, non-human
  - Main DB, separated local DB, and sister-BioBank
- Information DB
  - 500 PB (up to EB) data volume (~1 million human genomes)
  - DNA, RNA, Protein, Metabolism, and clinical information and phenotype of species.
  - Main DB, BackUP DB, International DB

My Slide in 2011

TWO KEYs

- Sequencing ➔ Data / Biological Information
- Bioinformatics ➔ Data Mining/ IT issues

- TB -> PB
- Challenges for storage, computing, mining, data sharing, data transfer
- Meaningful discovery (Data mining)

My Slide in 2011

Challenges (1)

- Huge amount of data (PB, EB data volume)
- New computer architecture, big memory usage, high-speed I/O, efficient storage system
- Ultra-speed network transfer
- Low-energy storage system

My Slide in 2011

Challenges (2)

Software & Hardware Issues:

- New database scheme design
- Information sharing protocol
- A series cloud: storage, computing, platform
- New mining tools, and how can easily help biologist working on the data

My Slide in 2011

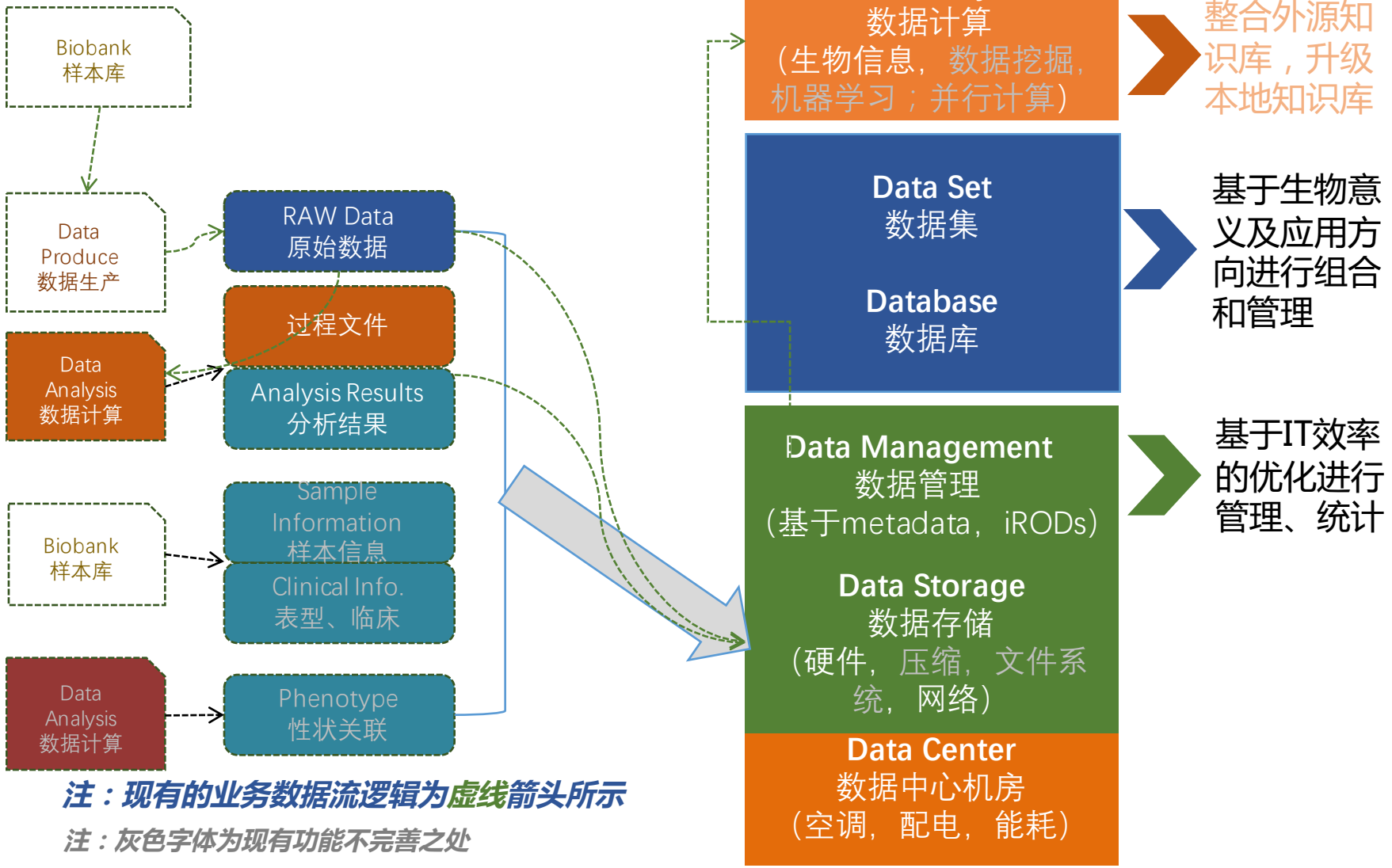
Compression (1)

- Experts in Compression for Data/File. Such as MPEG consortium, best people in compression world.

- My experience is most on bioinformatics analysis. So I will think more about how to help for end-application and analysis pipeline.

数据压缩作用于数据流的每个节点
Data compression can be applied in every node/step

not only for sequence but also for metadata



注：现有的业务数据流逻辑为虚线箭头所示

注：灰色字体为现有功能不完善之处

Compression (2)

- DNA Sequence is most biggest dataset ever in Biology.
- Two compression targets:
  - Sequence (file) compression
    - With reference (for resequencing data) or without reference (de novo, etc.)
  - Quality (file) compression

- Old fashion time
  - gzip

Performance in compression process

Leaderboard

Times are in seconds, memory is average total use in Kbytes. Measured using the time command with %e and %K options. ALL valid entries shown (valid = ran successfully and produced at least one correct fully matching result).

Name, Institute & entry ID (link to source)	Compress. ratio	Compress. time	Compress. memory	Decompress. time	Decompress. memory	Header mismatches	Sequence mismatches	Quality mismatches
Matt Mahoney, Dell Inc.(96)	0.0287	905.79	5293280	394.63	5294704	0	0	24033620
Armando J. Pinho, IEETA / Universidade de Aveiro(78)	0.0536	3035.17	5200	3010.02	5200	16629547	16629548	16629548
Armando J. Pinho, IEETA / Universidade de Aveiro(79)	0.0536	3018.11	5200	3026.36	5200	16628724	16628724	16628724
Armando J. Pinho, IEETA / Universidade de Aveiro(69)	0.0546	3404.17	68986880	3389.73	68984480	16512460	16512461	16512461
James Bonfield, Sanger Institute(101)	0.1141	3280.24	13235808	325.25	4381552	0	0	0
James Bonfield, Sanger Institute(104)	0.1142	10299.99	13272736	342.79	4381552	0	0	0
James Bonfield, Wellcome Trust Sanger Institute(17)	0.1154	208.42	932976	293.05	932880	0	0	24030818

The winner is James Bonfield (Sanger). The Ratio is 1:9.

2012

Reference Based Data

- Genome alignment analysis, etc.

Reference sequence will be used as the baseline,
and all sequence data will be aligned with the reference sequence.

A highly compressed data set will be used for further analysis.
The compression ratio has reached to 1:50 in my group.

A compressed database has been designed for further analysis.

Quality FILE

- FASTA: Quality by “number for each base pair” . (~100)
- FASTAQ: Quality by “ASCII, from [!]→[~]” . (~100)
- Illumina 1.0 Quality: ASCII 59 to 126; score from -5 to 62
- Illumina 1.3~1.8: ASCII 64 to 126; Phred score from 0 to 62.
- Illumina 1.5~1.8: slightly changed*
- Illumina 1.8: Phred + 33.
- Complete Genomics Sequence Quality:

- The score system has been changed for many times.

The compression ratio is significantly increased 10 times.

Big Datasets

- Human Genomes/Sequences: (> 500 TB)
  - 1,000 Human Genome Project
  - 100 Chinese Human Genome Project
  - ICGC (International Cancer Genome Consortium)
- Non-Human Genomes/Sequences: (> 20 TB)
  - Plant
  - Animal (Birds, Fishes, Vertebrates, etc.)

Thanks

- Thanks to Claudio Alberti for technical support.

- Also I would like to invite all experts for any topics in NGS data or relevant fields.
- As Co-convenor of WG 5 (MSG from Convenor Martin, too), we are appreciate to collaborate with MPEG and the other experts for standardization projects.

Yong ZHANG (yongzhangcn@qq.com)

ISO/TC 276/WG 5 (data processing and integration)

- ISO/TC276 Biotechnology
  - WG 5: data processing and integration
- ISO/TC 276/WG 5 Paris Meeting (2016-02-19)

- **RECOMMENDATION 1/2016/01 taken by ISO/TC 276/WG 5 on 2016-02-19**
- ISO/TC 276/WG 5 recommends developing a joint project in the field of "genome compression standardization" together with ISO/IEC JTC 1/SC 29/WG 11 (MPEG consortium). ISO/TC 276/WG 5 recommends establishing a joint ad-hoc group to prepare an NWIP for this project.
- **RECOMMENDATION 2/2016/01 taken by ISO/TC 276/WG 5 on 2016-02-19**
- ISO/TC 276/WG 5 recommends having a joint session with WG 3 at the upcoming combined meetings of ISO/TC 276/WGs on Tuesday, May 10th 2016 (3 to 5pm) in Washington DC, USA, to discuss a joint project with the MPEG consortium on "genome compression standardization." Marco Mattavelli (ISO/IEC JTC 1/SC 29/WG 11) will be invited to this meeting.

Yong ZHANG (yongzhangcn@qq.com, zhangy@genome1k.com)