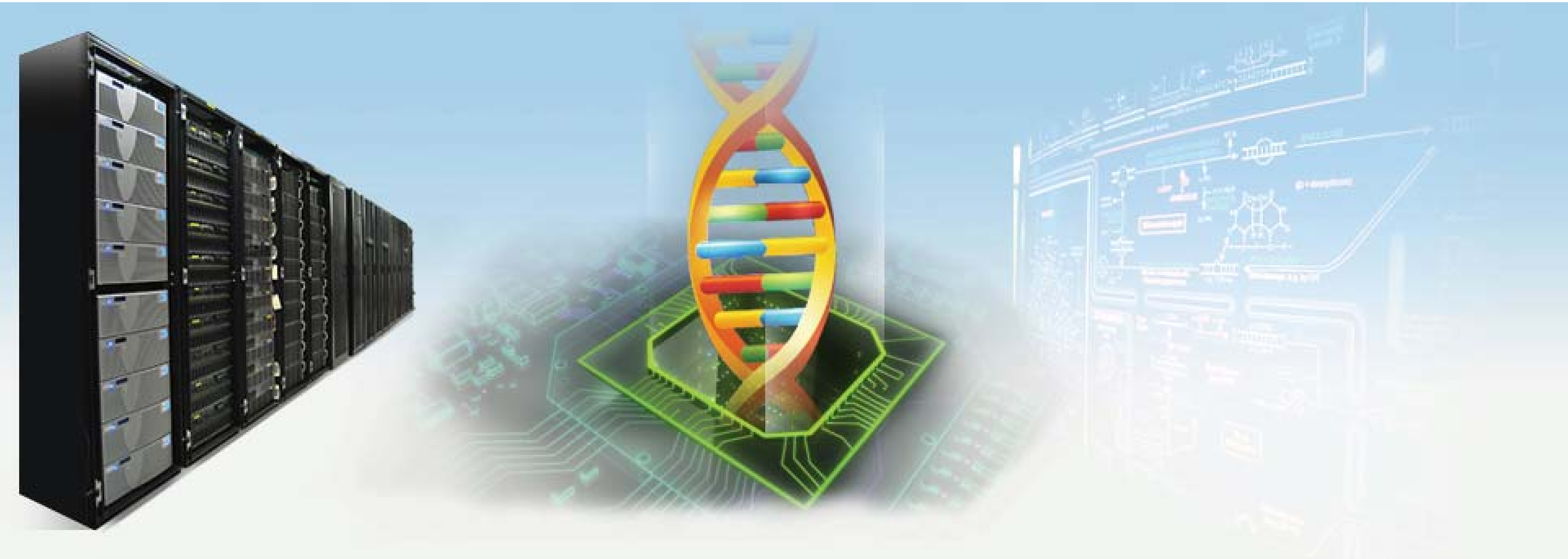
High Performance Computing Enabled Bioinformatics Research at BGI



Workshop on Building Collaborations in Clouds, HPC, and Application Areas
Co-organized by the University of Hong Kong and PRAGMA, 17 July, 2012

华大基因
BGI

BingQiang WANG, Head of HPC, BGI
wangbingqiang@genomics.cn

Agenda

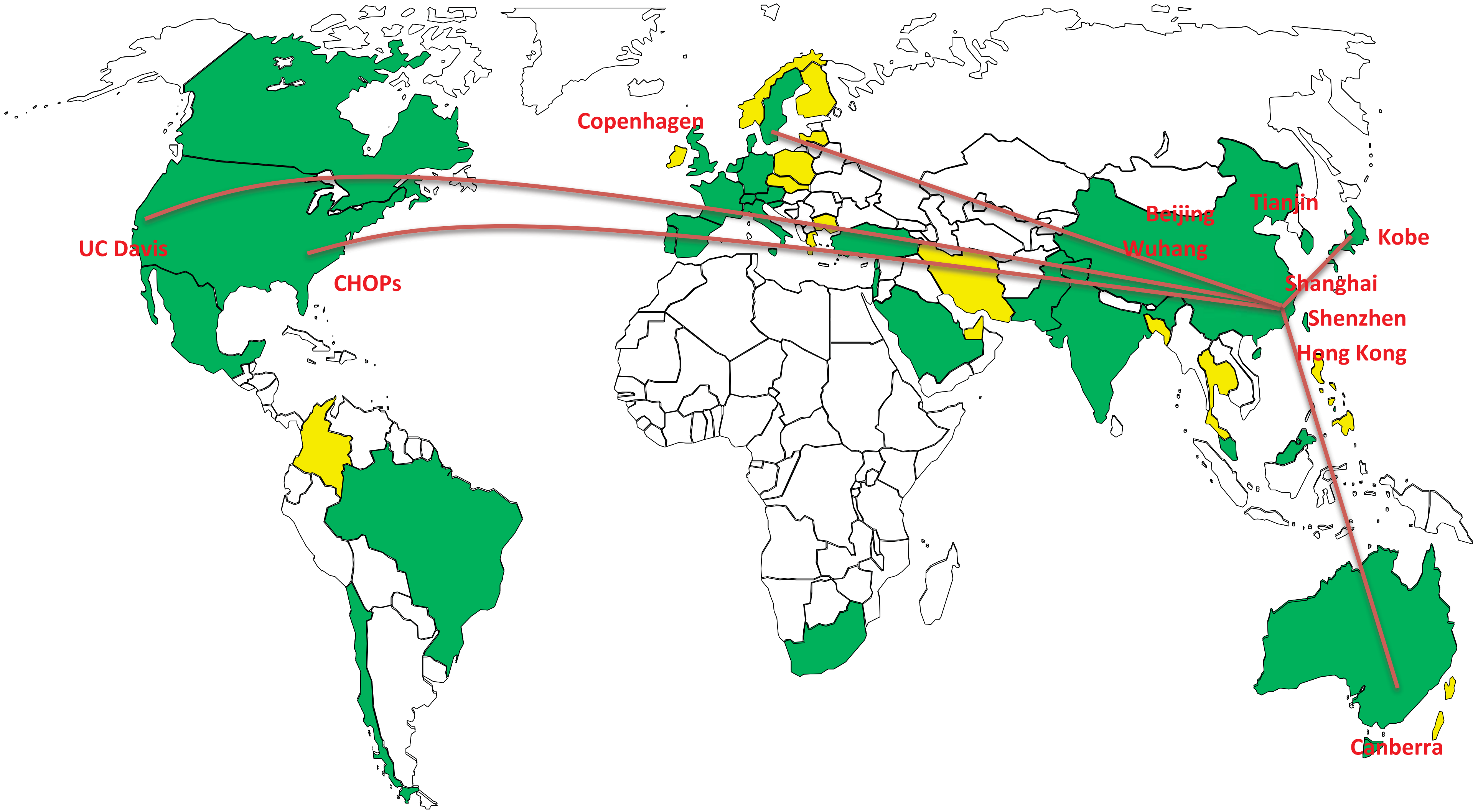
- Short BGI Intro
- Computing @BGI
- Future Genomics - “Big Data”
- Summary

History

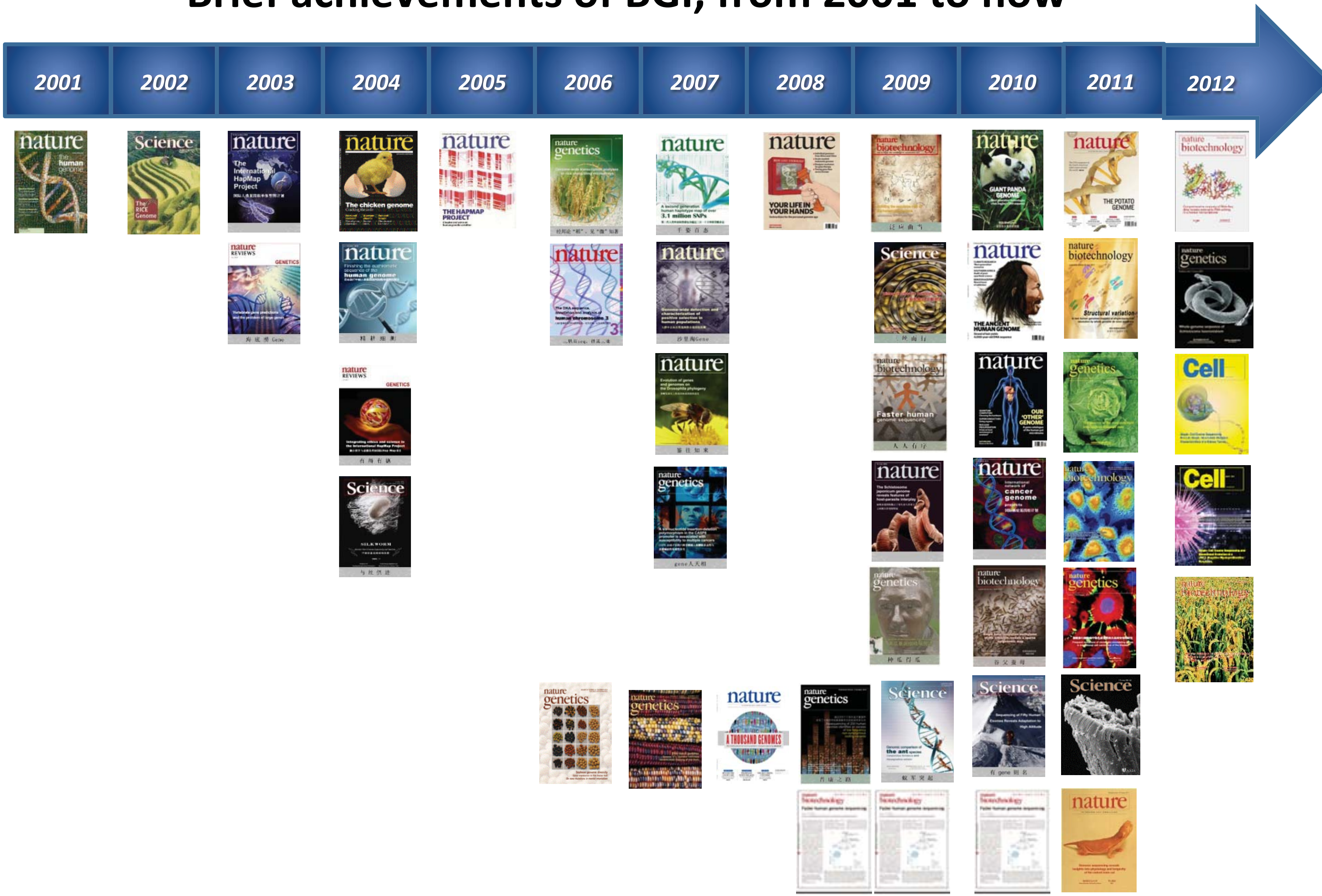
- September 1999 Beijing Genomics Institute, Beijing
- April 2007, Beijing Genomics Institute, Shenzhen
- May 2010, Beijing Genomics Institute, Hong Kong




BGI Branches

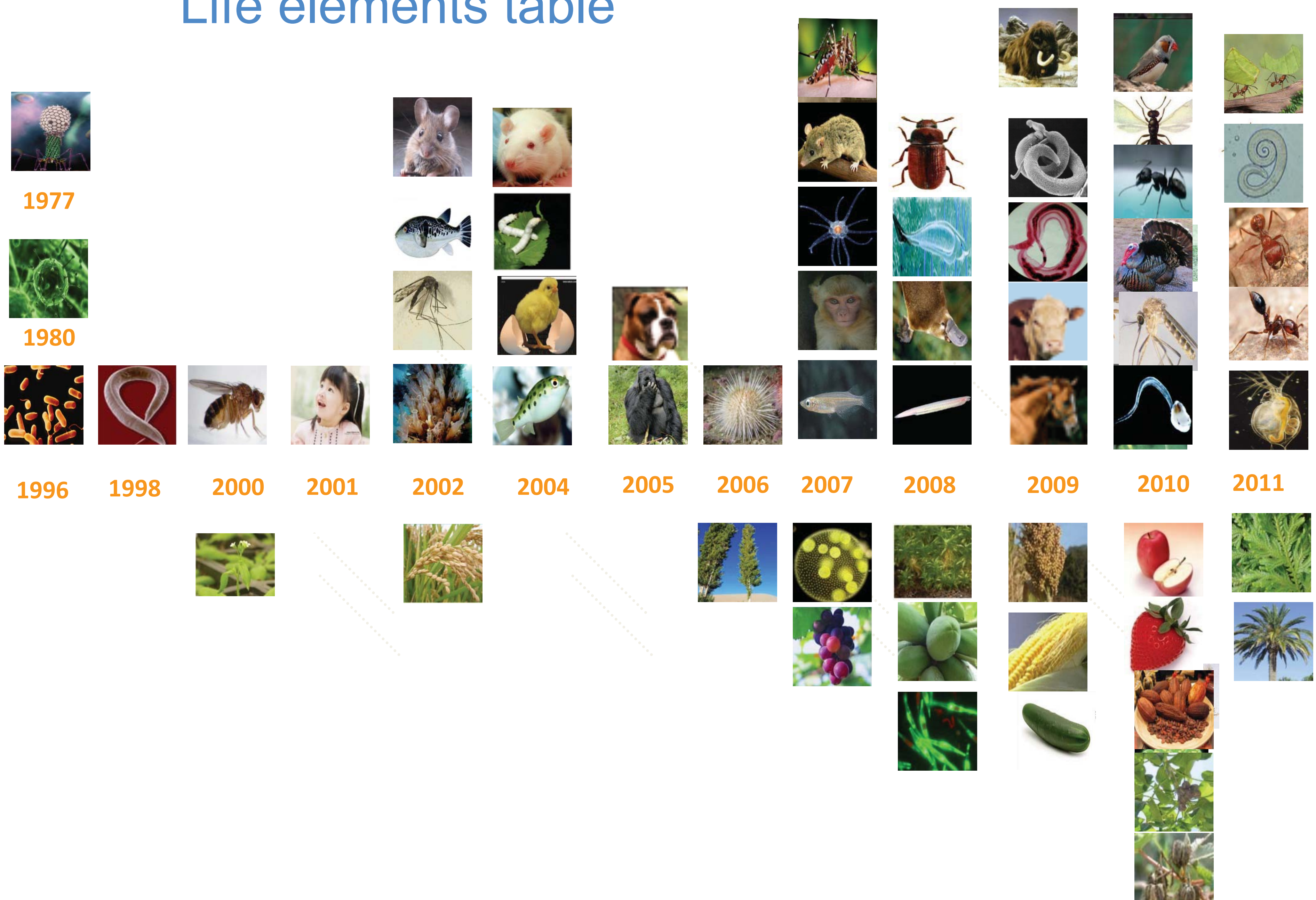


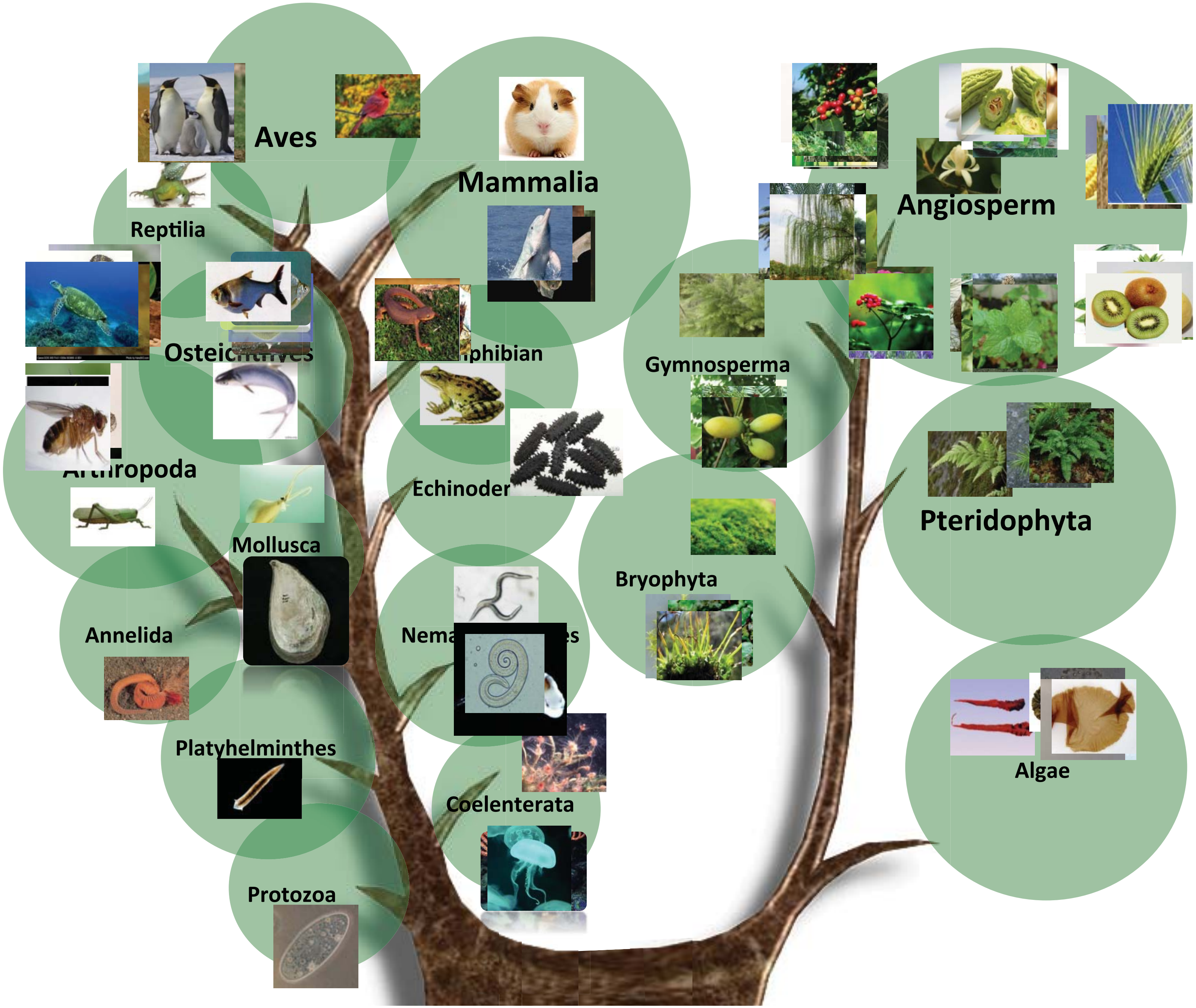
Brief achievements of BGI, from 2001 to now



- 3M Project
  - 1M human genome
  - 1M plants and animals genome
  - 1M microbe ecosystem genome

Life elements table




Aves
Mammalia
Reptilia
Osteic
phibian
ropoda
Echinoder
Mollusca
Annelida
Nema
Platyhelminthes
Coelenterata
Protozoa
Angiosperm
Gymnosperma
Pteridophyta
Bryophyta
Algae

Tree of Life

General Interest

Health care

- Complex disease
  - Metabolic disorder (type 2 disable, obesity)
  - Cancer
  - Neurodegenerative disease
- Personal genome

Animal, plant and microbe genome

- Sequencing new genomes
- Animal & plant
- Bacteria & meta

- Molecular breeding
  - Livestock
  - Crop

Agenda

- Short BGI Intro
- Computing @BGI
- Future Genomics - “Big Data”
- Summary

Award

Best Practices Award for IT Infrastructure

"Flexible green cloud computing framework for De novo assembly and whole genome resequecing" Won "The 10th Bio-IT World Conference & Expo Best Practices Awards for IT Infrastructure"!

IDC Announces New Winners of HPC Innovation Excellence Awards

18 Jun 2012

HAMBURG, Germany, June 18, 2012 -- International Data Corporation (IDC) today announced the third round of recipients of the HPC Innovation Excellence Award at the 2012 International Supercomputing Conference (ISC'12) in Hamburg, Germany. Prior winners were announced at ISC'11 and at the SC'11 supercomputing conference in the U.S.

The HPC Innovation Excellence Award recognizes noteworthy achievements by users of High Performance Computing (HPC) technologies. The program's main goals are to showcase return on investment (ROI) and scientific success stories involving HPC; to help other users better understand the benefits of adopting HPC and justify HPC investments, especially for small and medium-size businesses (SMBs); to demonstrate the value of HPC to funding bodies and politicians; and to expand public support for increased HPC investments.

Also 2011 Winner
Twice in two years

- **BGI Shenzhen (China).** BGI has developed a set of distributed computing appli to process large genome data sets on clusters. By applying advanced software te GlusterFS, and the Platform Symphony MapReduce framework, the institute has s some application workloads, BGI achieved a significant improvement in processin storage, resulting in reduced infrastructure costs while delivering results in less th 2.5 hours. Some of the applications enabled through the MapReduce framework i Human Genome for the International Human Genome Project; contributing 10% t Project; conducting research in combating SARS, and a German variant of the E. the rice genome, the silkworm potato genome, and the human gut metagenome.

Sequencing @ 华大基因 BGI

- World’s leading sequencing and genomics research center
- Started with Human Genome Project in 1999
  - Several sequencers at that time
  - Now more than 150 sequencers
  - Consider the trend …
- Mass spectrometers to capture protein information
  - Complement sequencing
  - Proteomics, so on

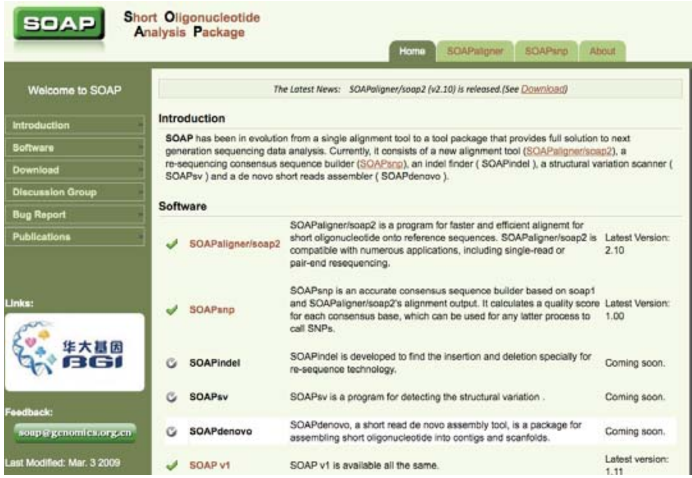
MODEL	ABI 3730XL	Roche 454	ABI SOLiD 4	Solexa GA IIx	Illumina HiSeq 2000
INSTALLATION	16	1	27	6	135

Computing @

- Sequencing throughput
  - 6T base pairs per day (upgraded from 4T)
  - ~20 PB data storage
- Connecting raw data and scientific discovery
  - Analysis tools
  - High performance computing is the key
- Computing horsepower
  - ~20,000 cores
  - ~20 GPUs
  - ~220 Tflops peak performance
- Still increasing ...

SOAP

- Next- generation sequencing data analysis software package



- Website:
  http://soap.genomics.org.cn

- >10,000 users

- SOAP:

SOAP: short oligonucleotide alignment program. **Bioinformatics**. **2008** 24: 713-714

- SOAP2:

SOAP2: an improved ultrafast tool for short read alignment. **Bioinformatics**. **2009**

- SOAPsnp:

SNP detection for massively parallel whole genome resequencing. **Genome Research**. **2009**

- SOAPdenovo:

De novo assembly of the human genomes with massively parallel sequencing. **Genome Research**. **2009**

Novel Approaches

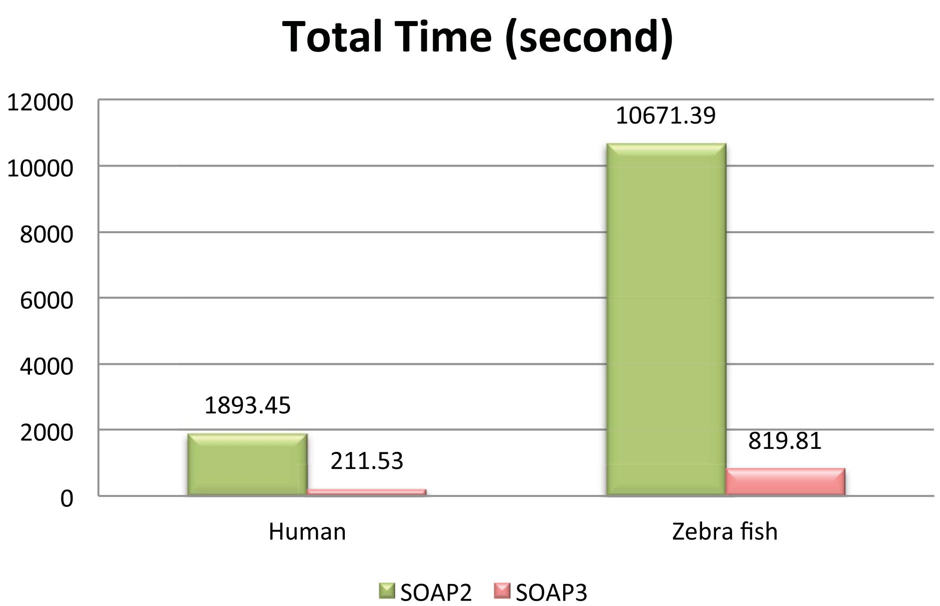
- Using heterogeneous computing to accelerate bioinformatics analysis
  - SOAP3/SOAP3-DP
  - GSNP
  - GAMA
- Scale up with Cloud computing
  - Gaea
  - Hecate

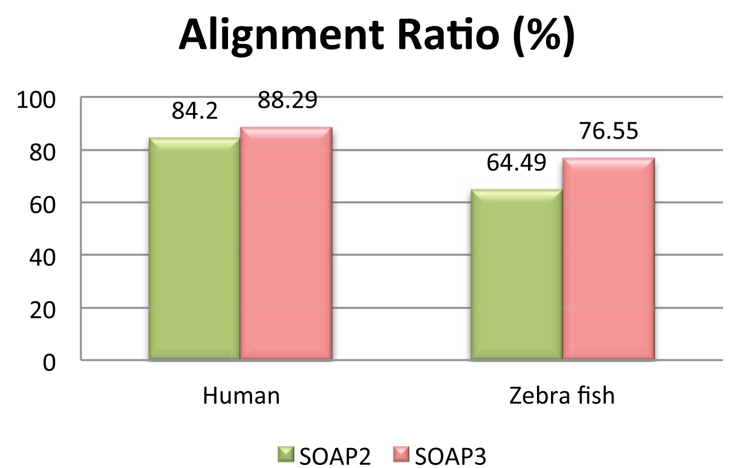
SOAP3 Aligner – History and Intro

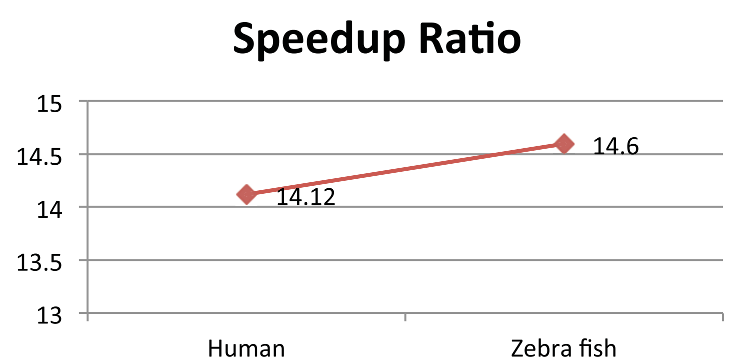
- Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. (from Wikipedia)
- SOAP: first-generation short read alignment tool
- SOAP2 (2008): 20 to 30 times faster than SOAP, less memory
  - Collaboration between BGI & HKU
  - Compressed indexing: bidirectional BWT (2BWT)
- SOAP3 (2011): 10 to 30 times faster than SOAP2
  - Collaboration from HKU
  - GPU's parallel processing power
  - CPU memory: increase from a few to tens GB
  - GPU-based indexing: GPU-2BWT

SOAP3

Data type	Reads length (bp)	Total Number of Reads (million)	Mismatch number	SOAP3 (Total Time: second)			SOP2 Total time (second)	Alignment Speed-up ratio (second)
				Time for reading reads	Time for alignment and output	Total time		
Human	100	16	3	83.30	128.23	211.53	1893.45	14.12
Zebra fish	76	21	3	95.50	724.32	819.81	10671.39	14.6

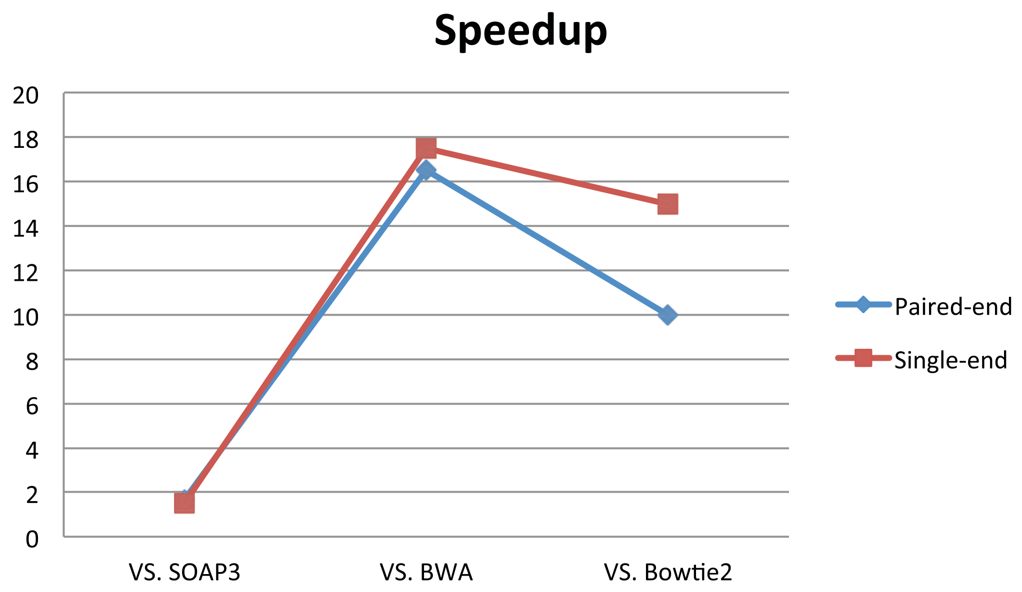






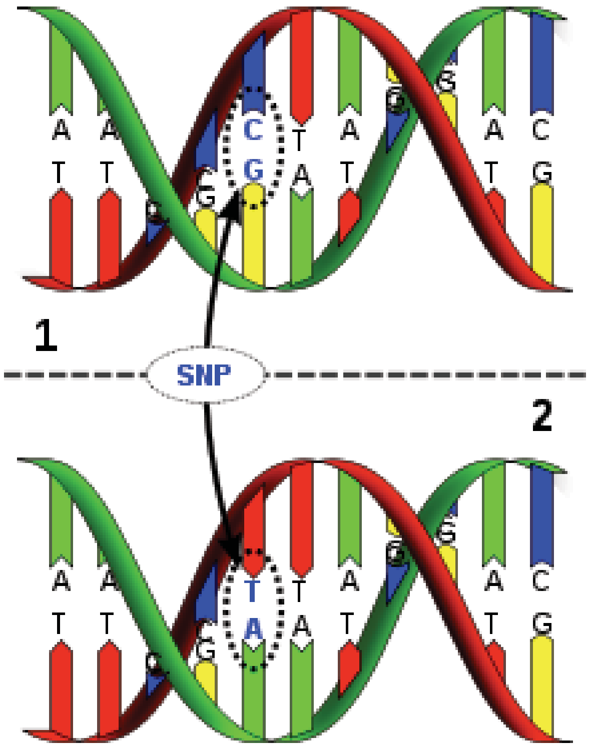
Speedup Compared with Other Tools

- SOAP3-dp is about 2 times faster than SOAP3, while at least 10 times faster when comparing with other tools



SNP Calling with GSNP

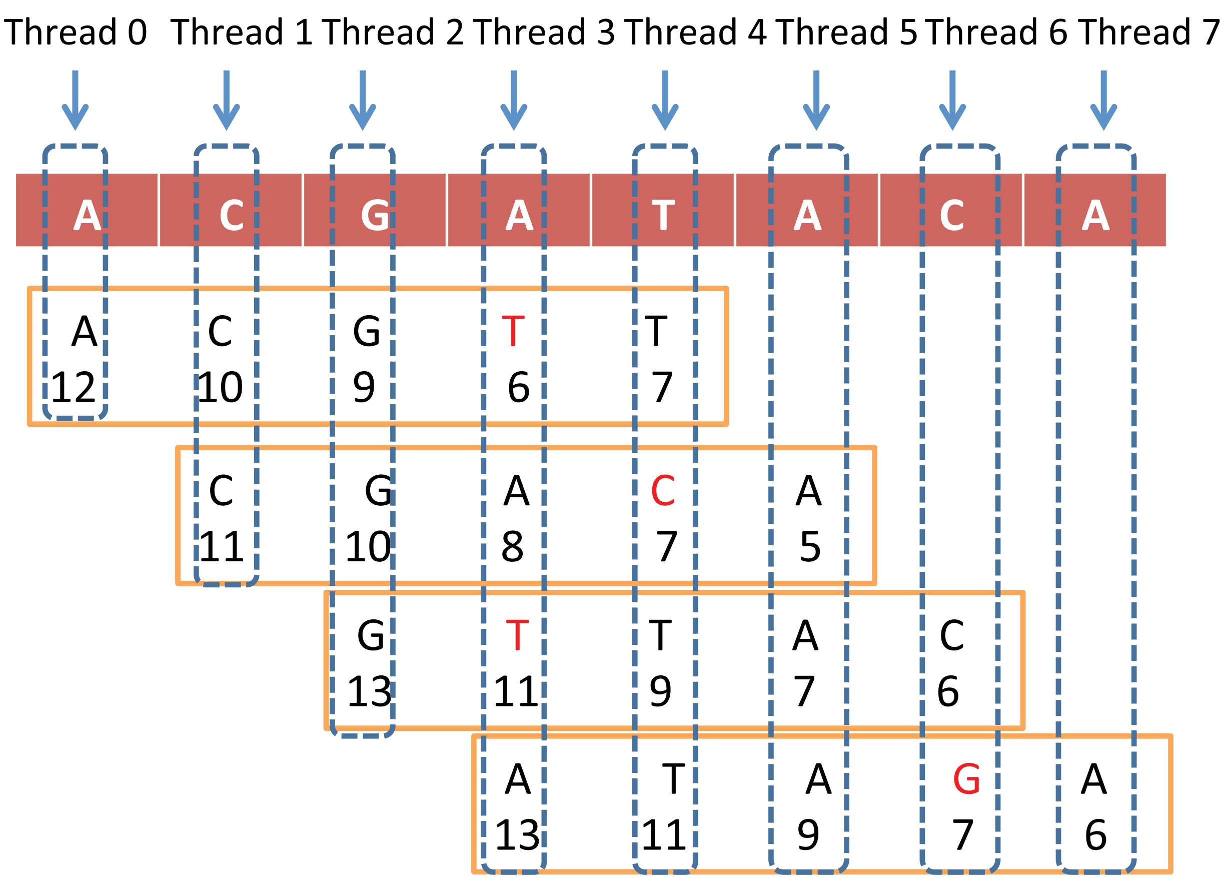
- A single-nucleotide polymorphism (SNP, pronounced snip) is a DNA sequence variation occurring when a single nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in an individual.
- Collaboration with Hong Kong University of Science and Technology (HKUST)
  - Professor Qiong Luo
  - Mian Lu
  - Jiuxin Zhao
- Based on SOAPsnp
  - BGI's home made standard SNP calling tool



GSNP

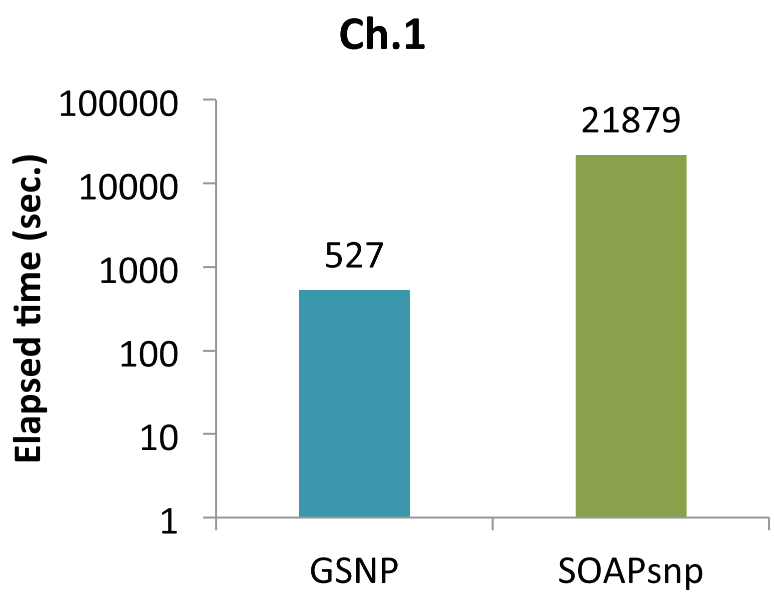
The parallelization strategy on the GPU: one thread handles one site

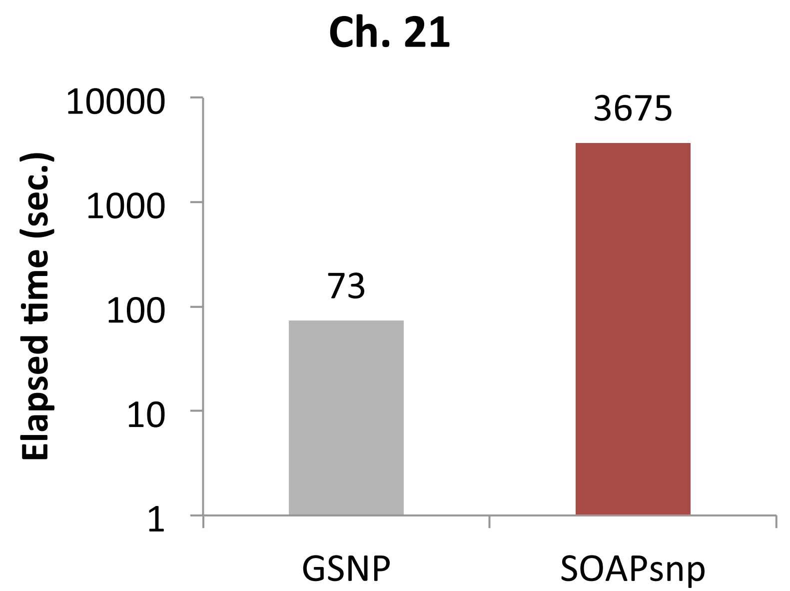
→ Optimization techniques of GSNP



- Reduce memory overhead and branch divergence
- Balance workloads
- The Consistency of GPU and CPU Results
- Reduce I/O cost

End-to-End Performance Comparison of GSNP



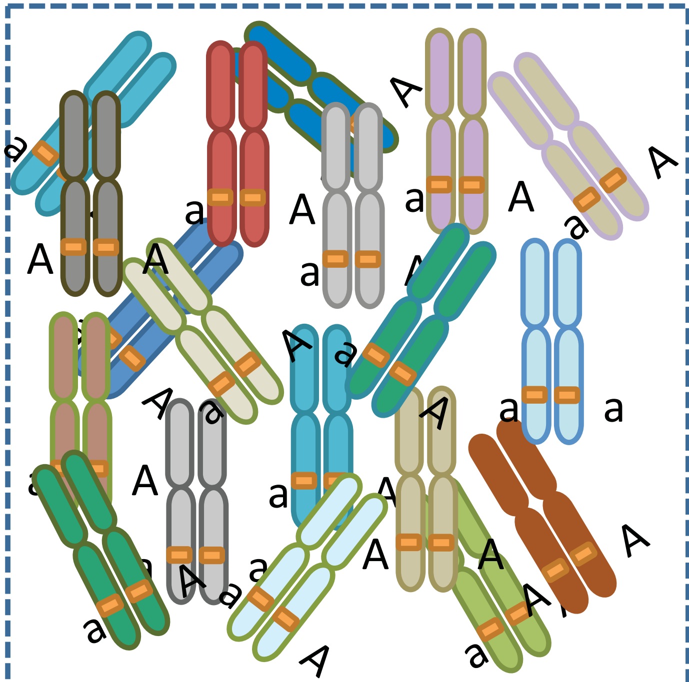


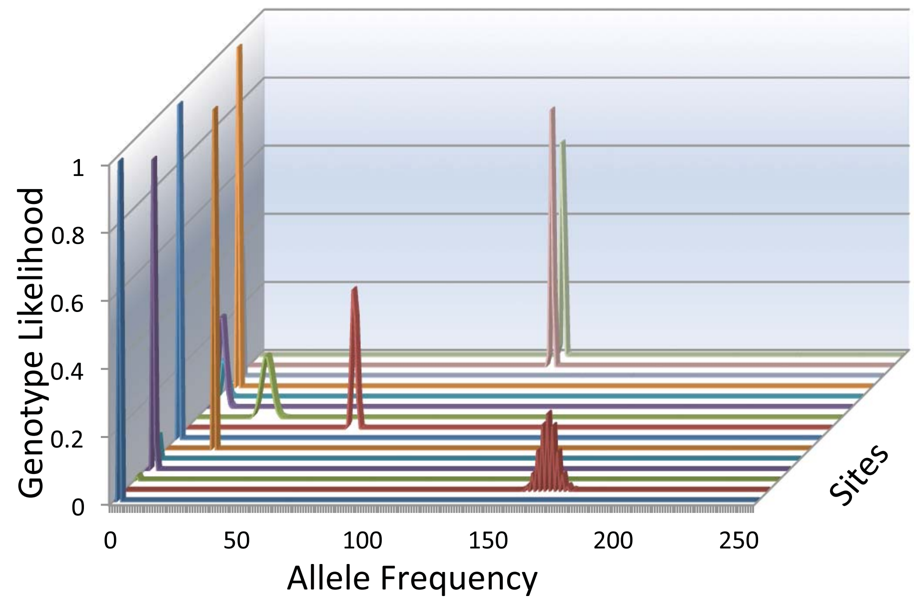
The elapsed time of all components are included.
GSNP is around 50X faster than the single-thread CPU-based SOAPsnp.

Estimating MAF in a Population with GPU

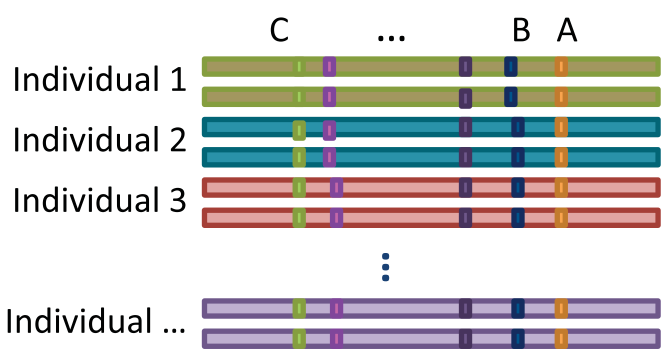
- Within a population, SNPs can be assigned a minor allele frequency — the lowest allele frequency at a locus that is observed in a particular population. There are variations between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in another. (from Wikipedia)
- MAF is the foundation of genome wide association study (GWAS), e.g. HapMap project
- Our approach is a highly accurate yet computationally very expensive one ($O(N^2)$)
- Collaboration with Hong Kong University of Science and Technology (HKUST), as well as National Supercomputing Center at Tianjin (Tianhe-1A)

GAMA





Different sites represent different alleles



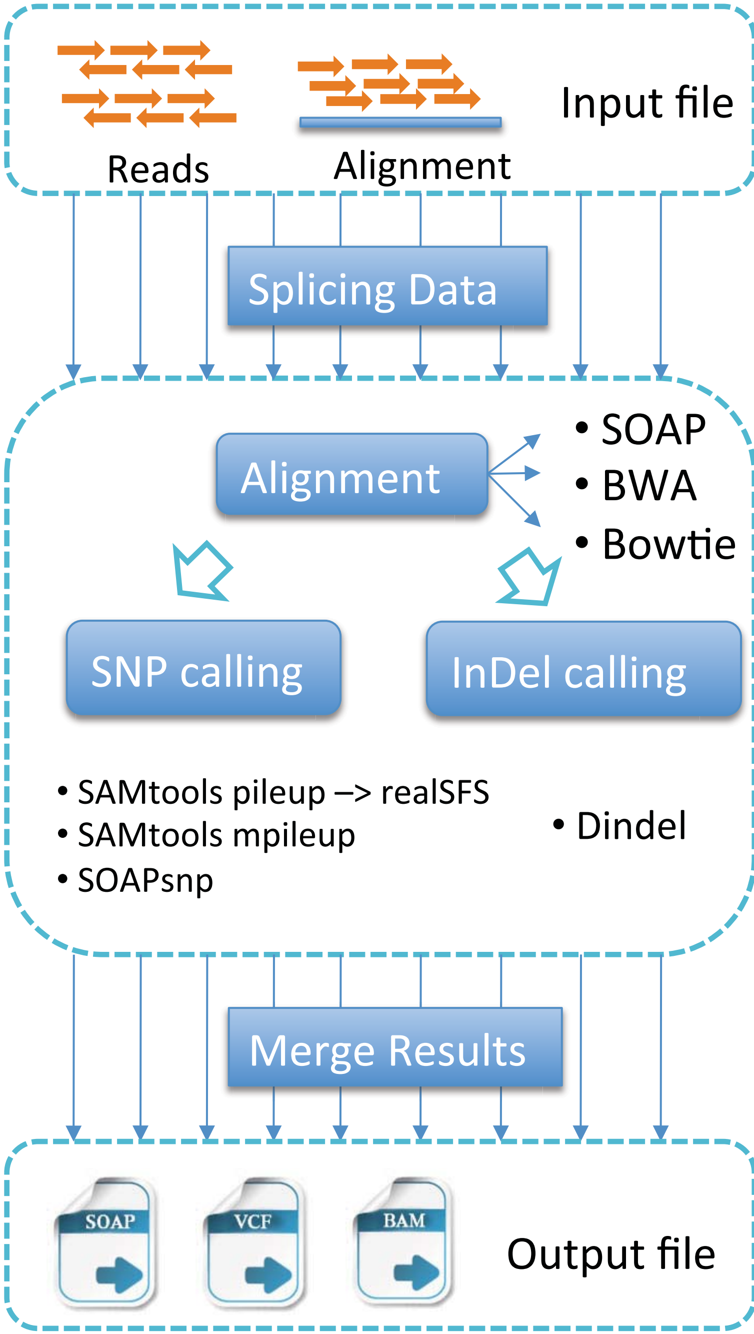
Compute allele frequency likelihood for each site

GAMA

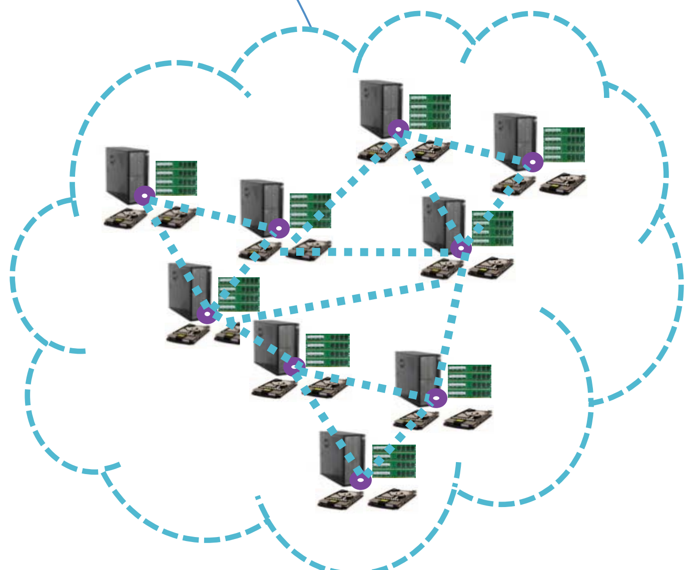
Dataset: Human genome, 512 individuals (1024 input files) , full scan of 3G sites

Version	Computing time	Total Time	Computing Speedup	Total Speedup	Note
CPU	~ 1518 days	~ 1619 days			
GPU (Single)	~ 15.75 hours	~ 101 days	2313	16	against CPU
GPU (86 with MPI *)	~ 717 seconds	~ 5.4 hours	79	449	against single GPU

* 86 nodes x 12 cores per node = 1032 cores , with one core processing one file



Gaea1 is designed to distribute re-sequencing computation to a cluster of nodes based on the Hadoop-Streaming.



It can improves the efficiency of cluster usage by more than 30%.

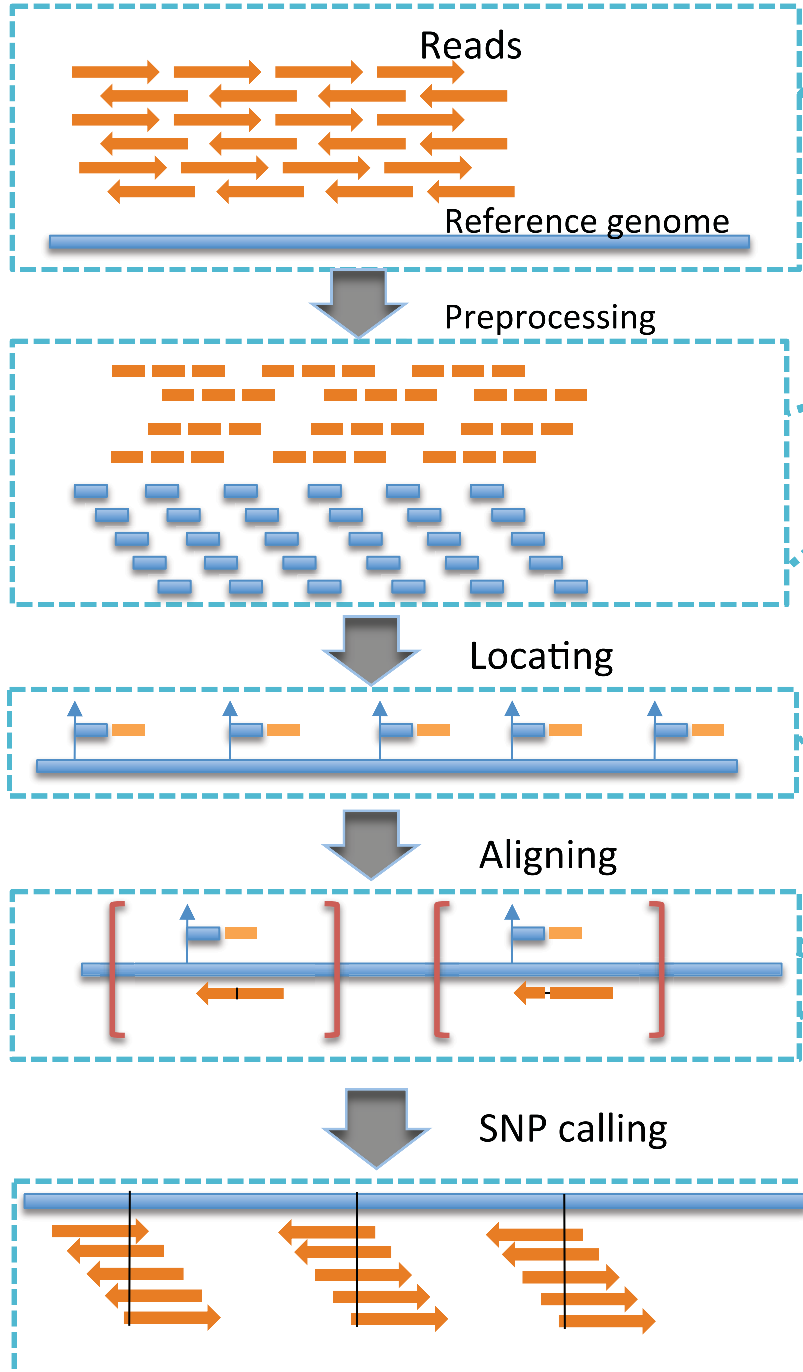
Gaea 2.1

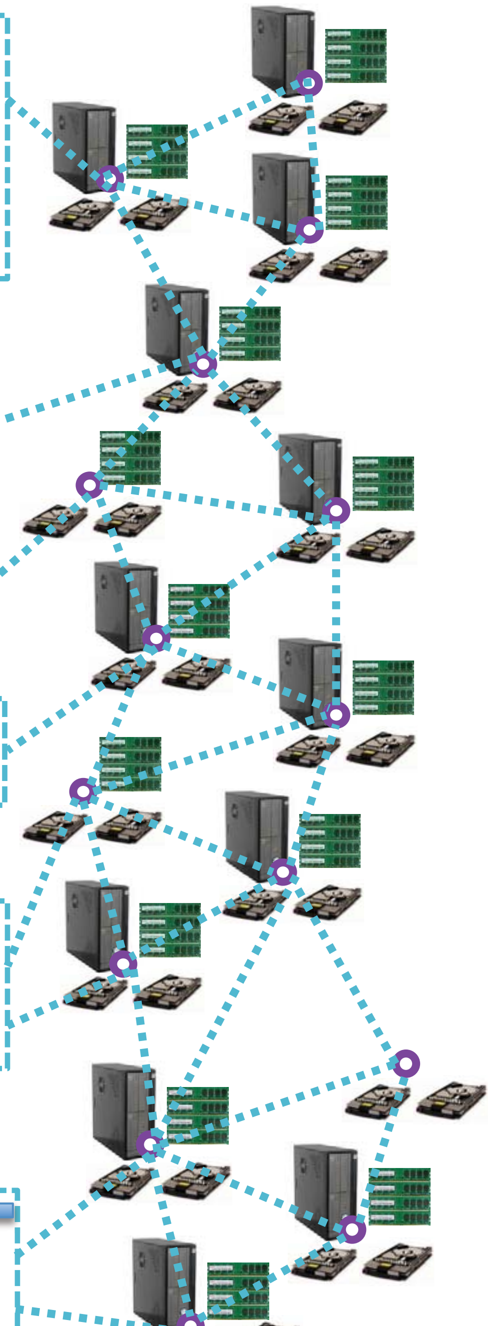
Distributed Indexing for load balancing

Flexible splitting tolerates more mistmatches

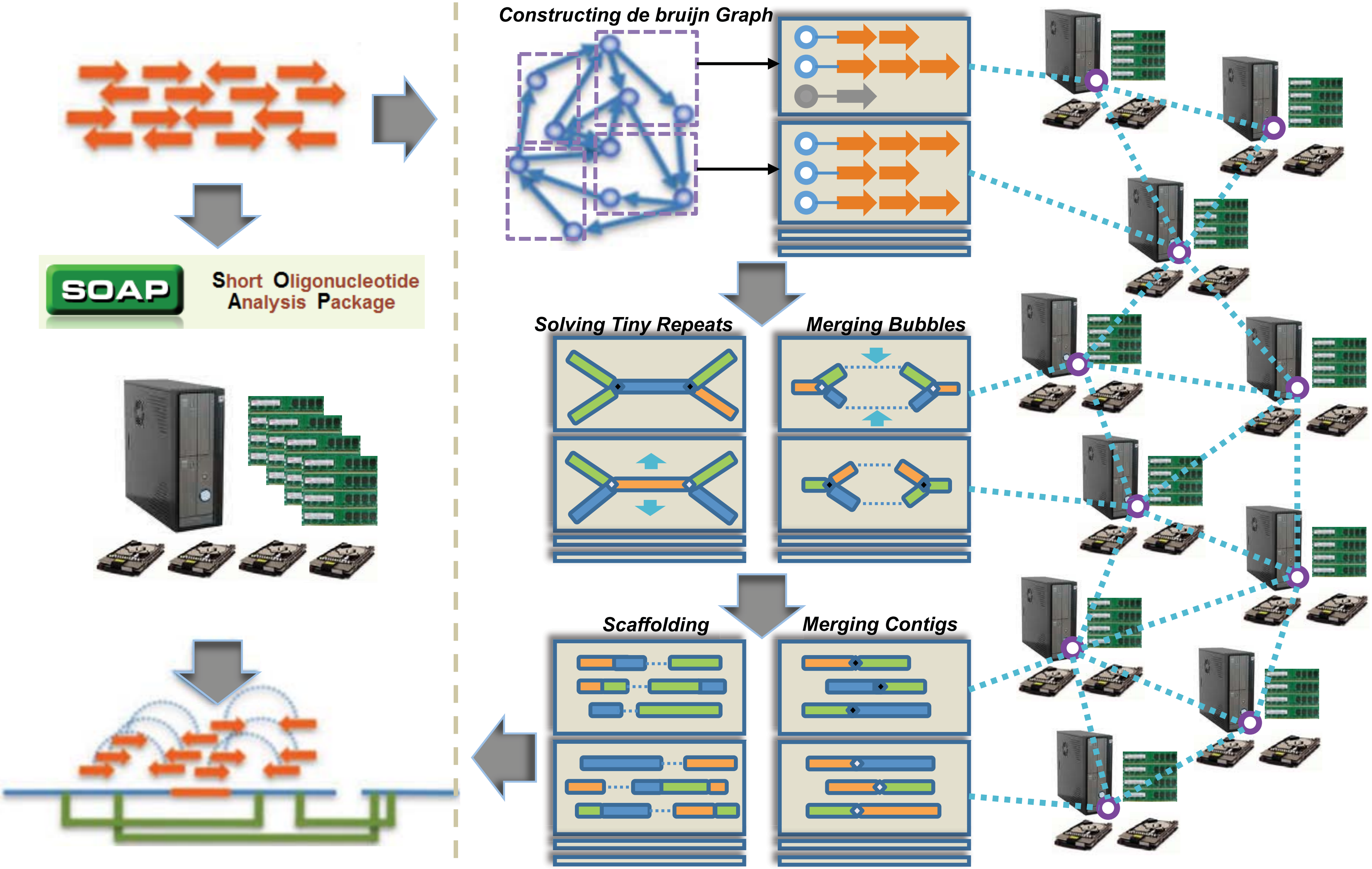
Dynamic Programming for robust gap alignment

Standard mapping quality for SNP calling





Hecate



Constructing de bruijn Graph
SOAP
Short Oligonucleotide Analysis Package
Solving Tiny Repeats
Merging Bubbles
Scaffolding
Merging Contigs

Agenda

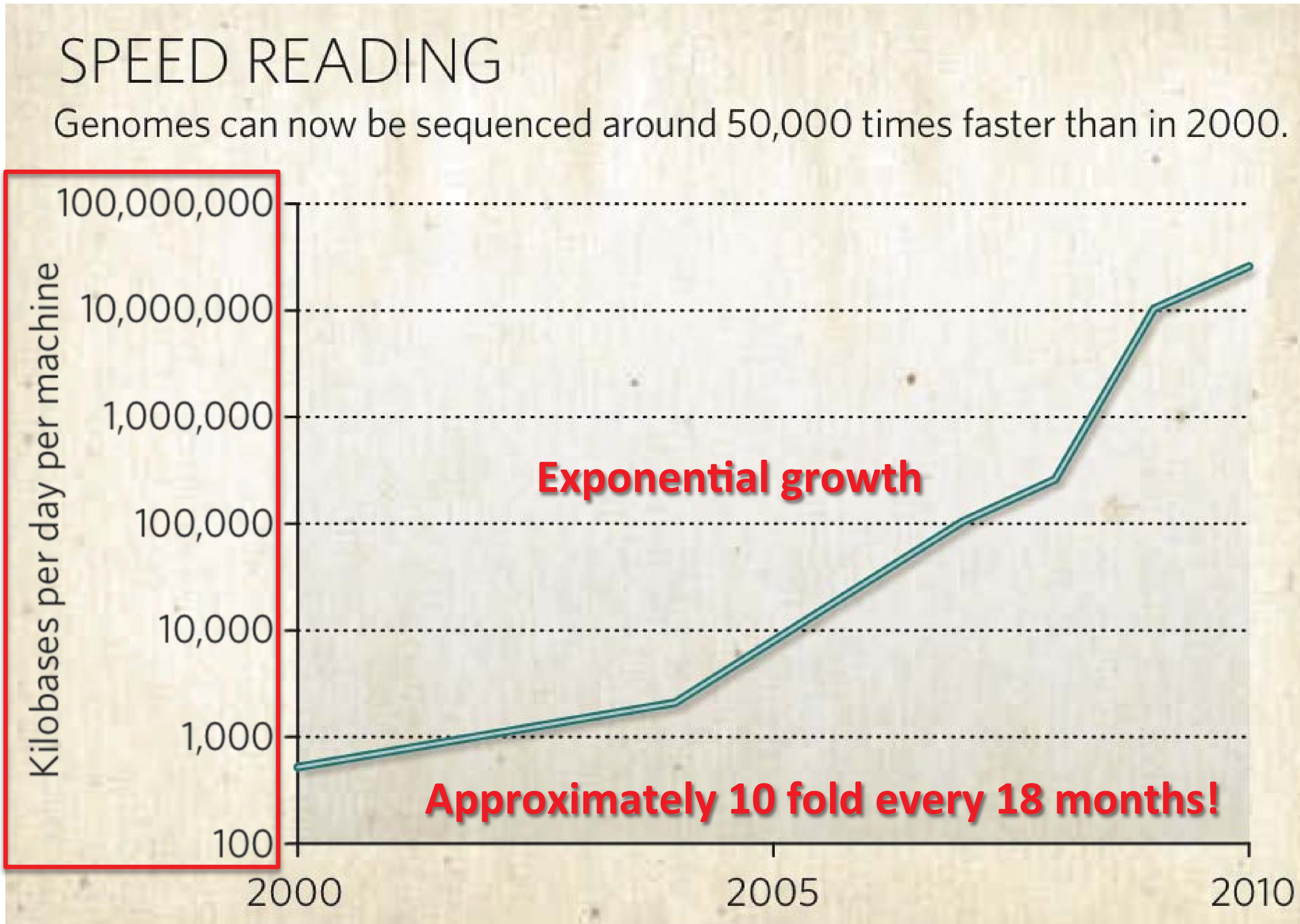
- Short BGI Intro
- Computing @BGI
- Future Genomics - “Big Data”
- Summary

Next Generation Sequencing (NGS)

- Indeed 2nd generation sequencing technology
- Low cost (*several K$ per human genome*)
- High throughput
- Short reads (small pieces of DNA strand)
- Lots, lots of data

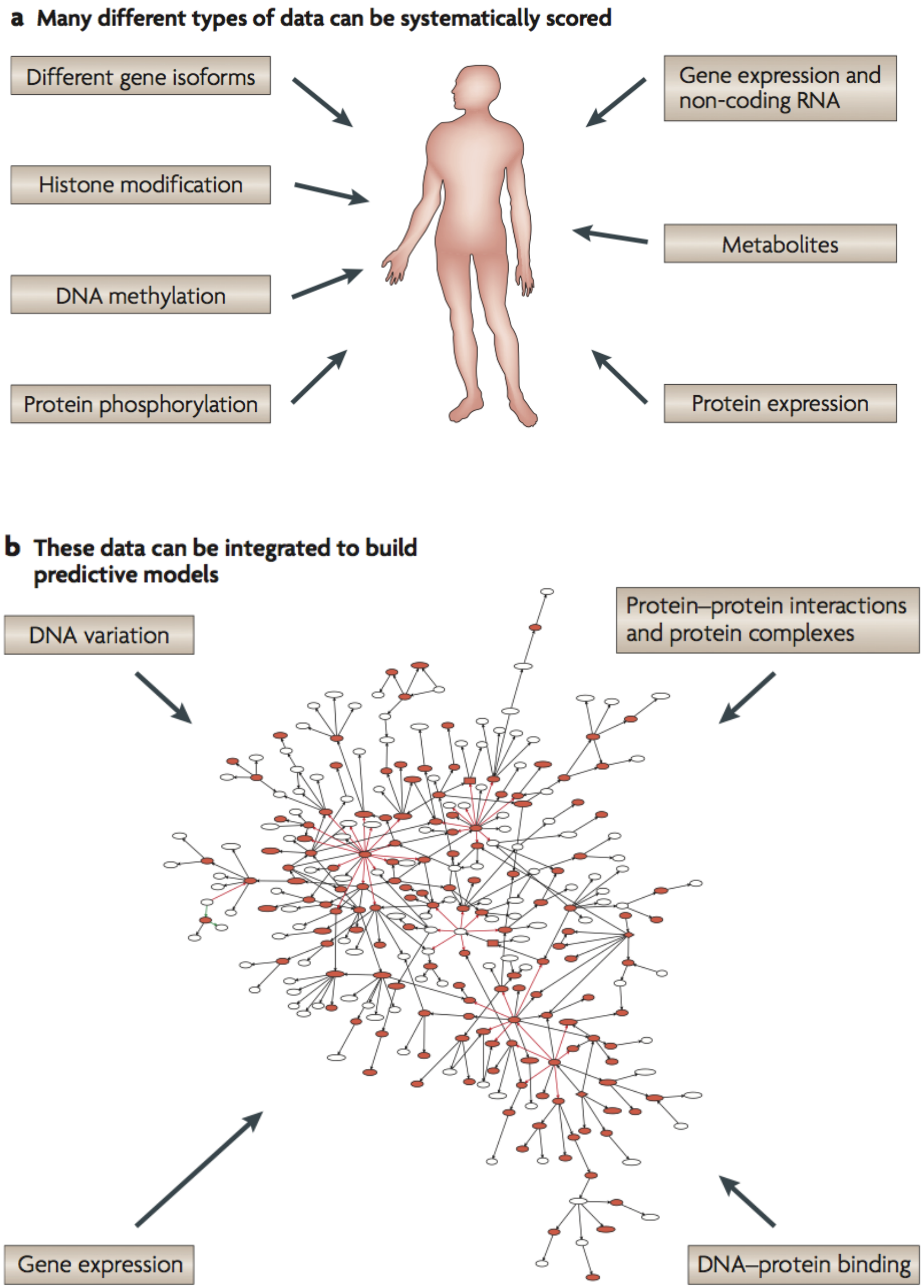
Big Data Incoming

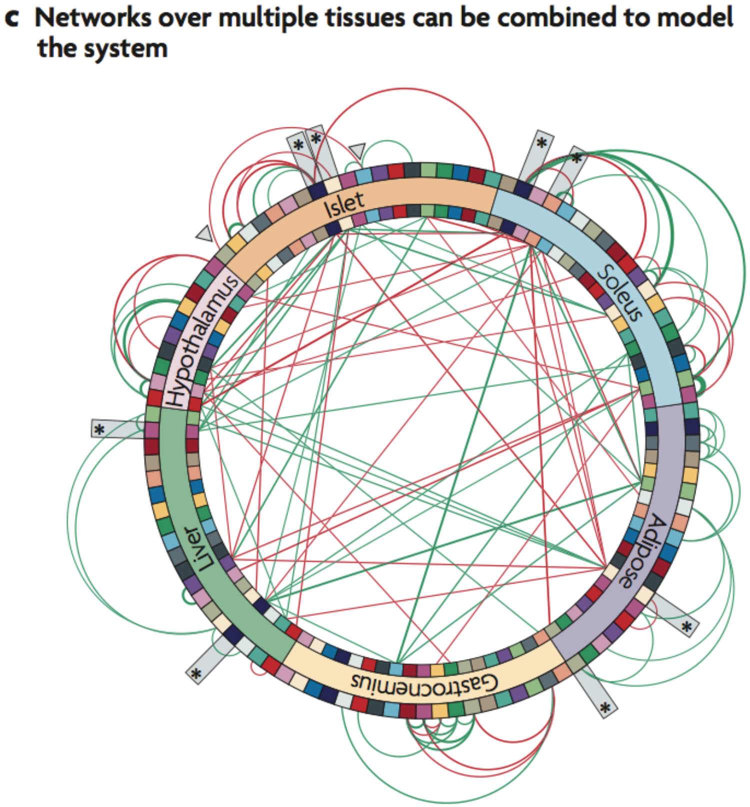
- Breadth
  - As sequencing cost falling falling down
  - More individuals are being sequenced
    - Thousands of human individuals: diagnostics and treatment of diseases
    - Tens of thousands of rice individuals: molecular breeding, more food
- Depth
  - Combining data from other sources / levels
  - DNA, RNA, protein…
- And, dynamically
  - The dimension other than breadth and depth - time
  - Living cells, living life: stem to multiple tissues and organs



Craig Venter, **Multiple personal genomes await**, *Nature*, Vol 464, April 2010

Collecting and integrating large-scale, diverse types of data





… we are able to isolate and sequence individual cells, monitor the dynamics of single molecules in real time and lower the cost of the technologies that generate all of these data, such that hundreds of millions of individuals can be profiled. Sequencing DNA, RNA, the epigenome, the metabolome and the proteome from numerous cells in millions of individuals, and sequencing environmentally collected samples routinely from thousands of locations a day …

Eric E. Schadt *et al*, Computational Solutions to Large-scale Data Management and Analysis, Nature Reviews | Genetics, Vol 11, September 2010

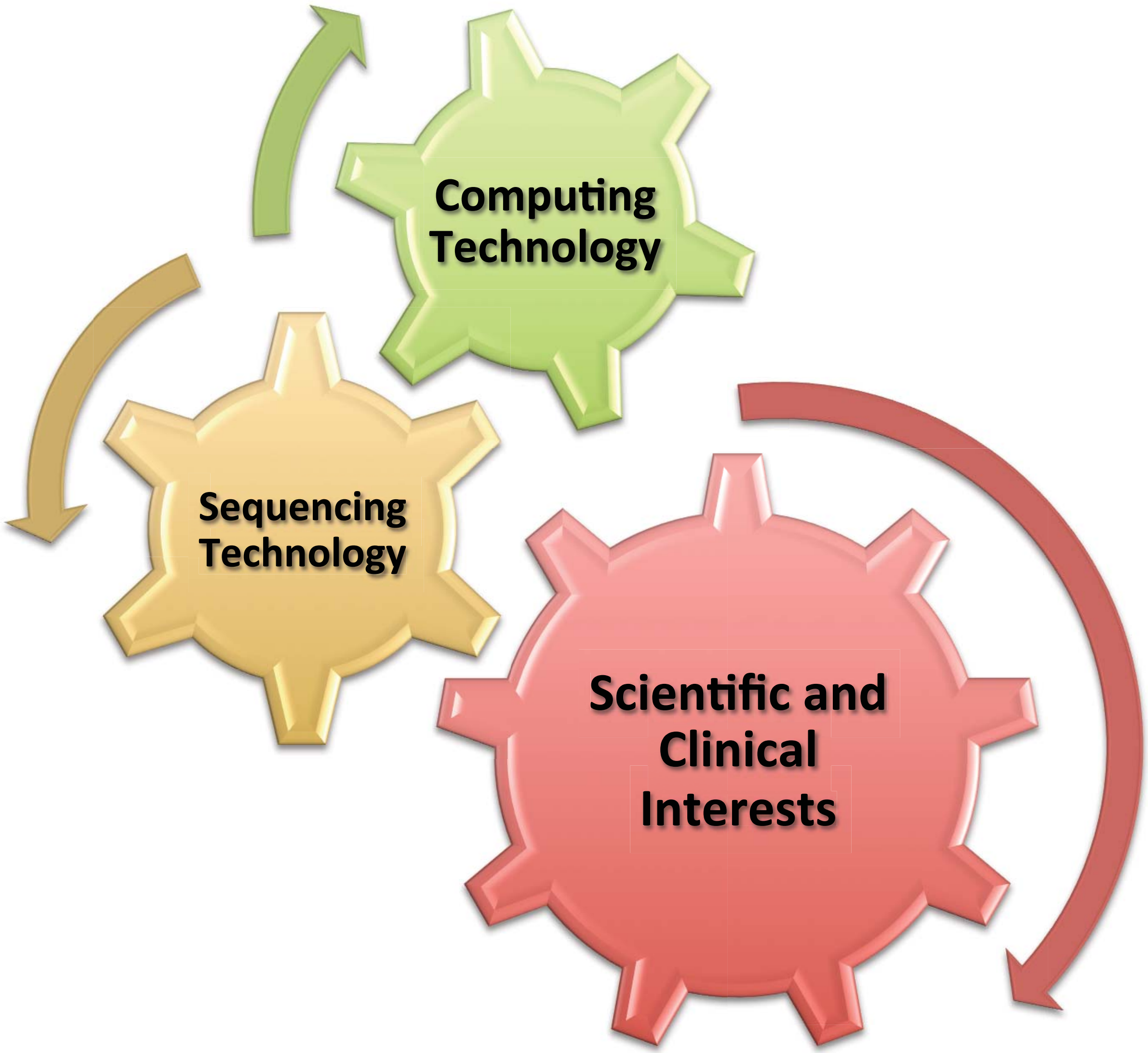
Sequencing vs Computing

- Observation
  - Exponential growth of sequence data output
- What will happen if, demand for computation grows with amount of data, as
  - *O(N)*
  - $O(N^2)$
  - beyond $O(N^2)$?

Computational Challenges

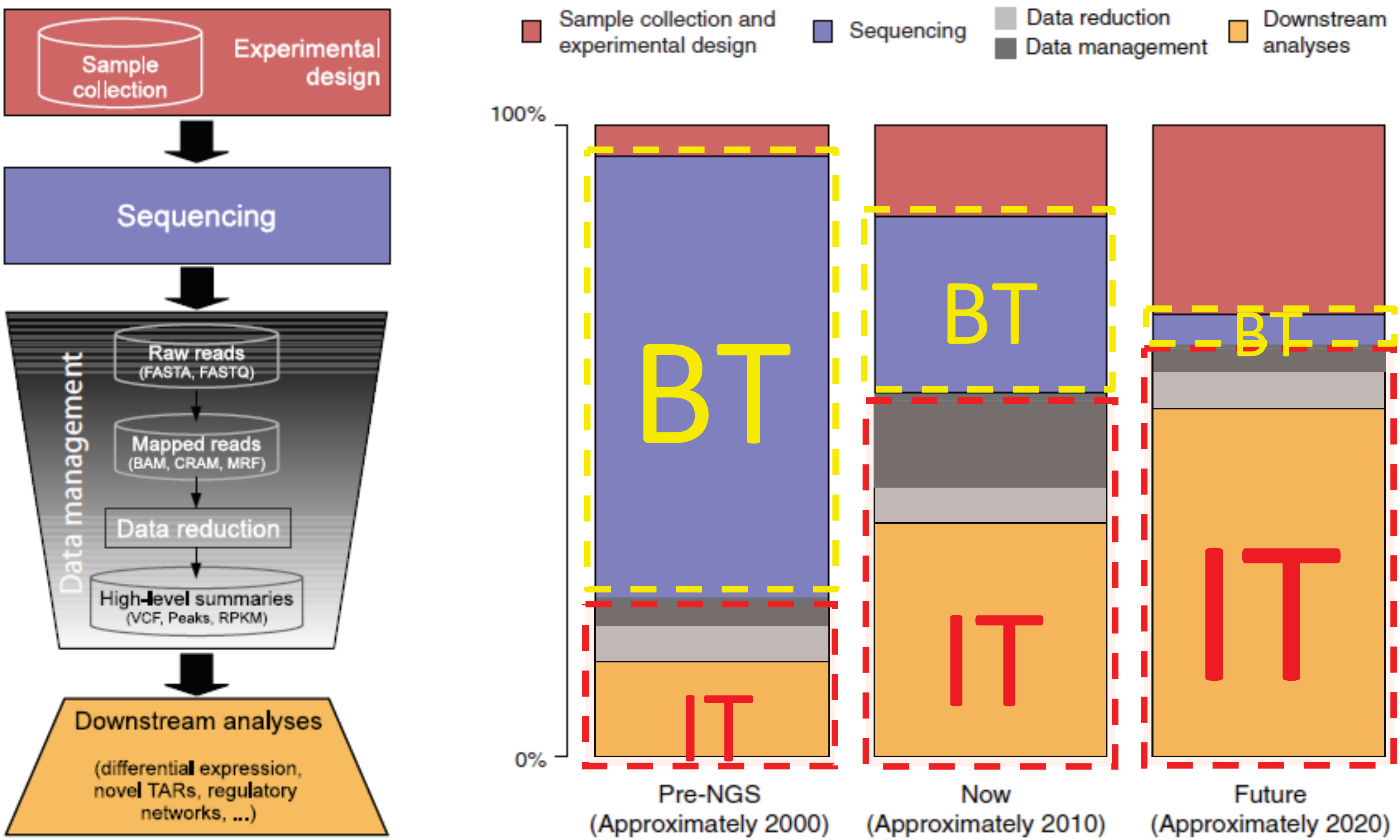
- "Classical" sequence data analysis
  - Alignment **as *O(N)***
  - Variant calling **as *O(N)***
  - Linear as data increasing
- Growing computing demand – let us mine for "sth"
  - Population genomics **as *$O(N^2)$***
  - Phylogenetic study ***NP hard***
  - Gene association study ***high dimensional***
  - Systems biology with various levels of data ***NP hard***
  - …
- Sequencing cost down leads to more and more high dimensional analysis
  - Lots, lots of computing

Solution: Disruptive Computing Technology



Shift of Sequencing Service Profit

The real cost of sequencing: Higher than you think
Sboner et al. *Genome Biology* 2011, 12:125




USA
Bio Tech
Info Tech

Agenda

- Short BGI Intro
- Computing @BGI
- Future Genomics - “Big Data”
- Summary

Our Observation

- Bioinformatics is turning from high throughput computing to data intensive computing ***RIGHT NOW***
- Tools and systems need to be developed
  - See GAMA-MPI example
    - Tens of minutes computation
    - Several hours for data loading / decompression / parsing / filtering
  - Data intensive architecture
  - Data compression technology
  - Data awareness scheduling
  - Manage and mine big data in an efficient manner

EasyGenomics

华大基因 BGI

EasyGenomics

Next Generation Bioinformatics
on the Cloud

http://www.easygenomics.com

Sifei He
Director of BGI Cloud
hesifei@genomics.cn

Xing Xu, Ph.D
Senior Product Manager
EasyGenomics | BGI
xuxing@genomics.cn

Contact Us
info@easygenomics.com

Problems and Solutions

Solutions

Cloud

High Speed Data Exchange

Workflows

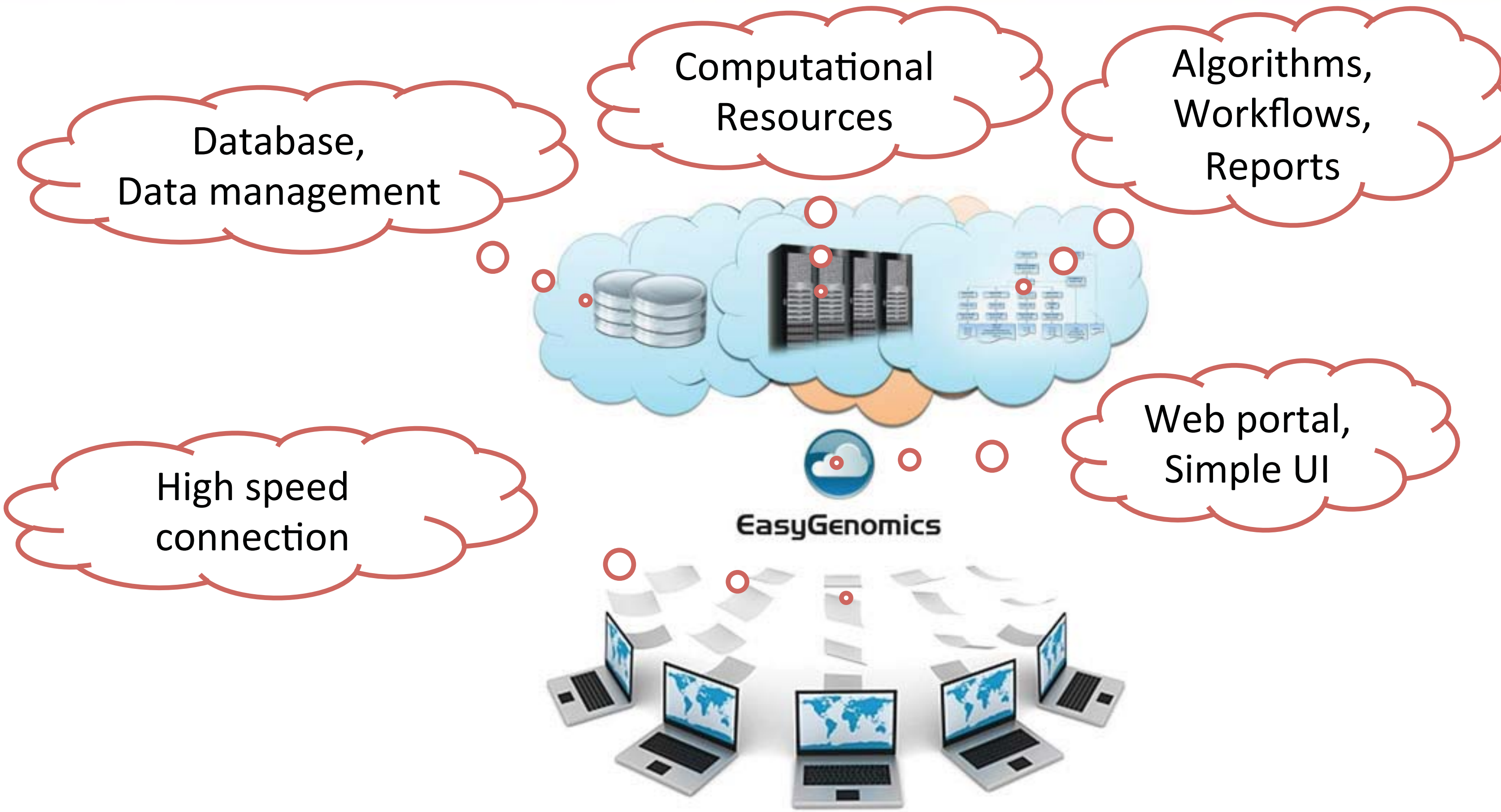
+) Resource Management

EasyGenomics

Problems:

- ~~Big genomic data~~
- ~~Geological distribution~~
- ~~Algorithm integration~~
- ~~Computational demand~~

EasyGenomics™



EasyGenomics is the bioinformatics platform for research and applications on the cloud

Thank you

wangbingqiang@genomics.cn