GPU Accelerated Life Science Research at BGI

– Reshaping Bioinformatics with GPU Technology

GTC Asia Beijing, December 15, 2011

BingQiang WANG

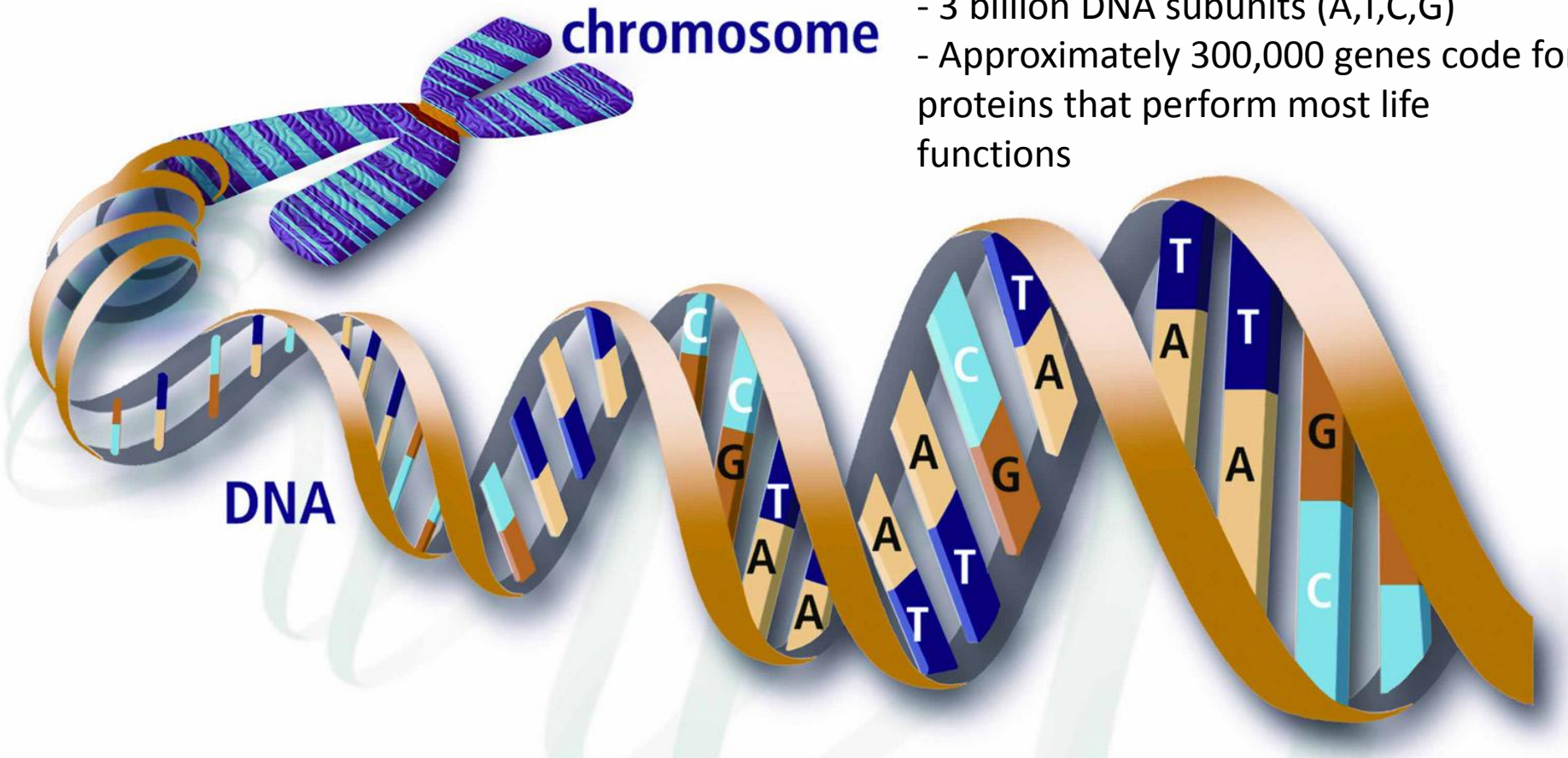
wangbingqiang@genomics.cn

BGI-Shenzhen

DNA double helix encodes secrets of life

Facts: Human Genome

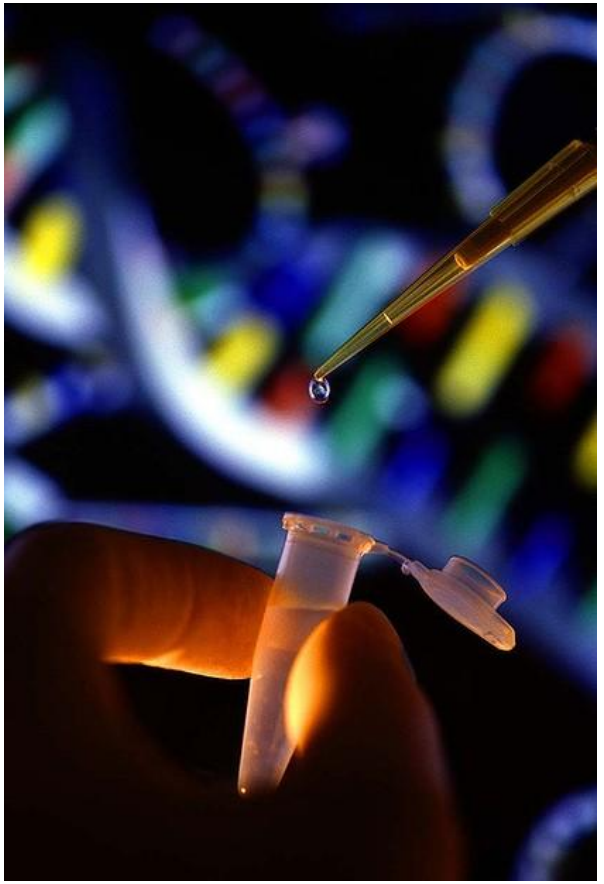
- Trillions of cells
- 23 pairs of chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (A,T,C,G)
- Approximately 300,000 genes code for proteins that perform most life functions



Sequencer, Sequencing, Sequence

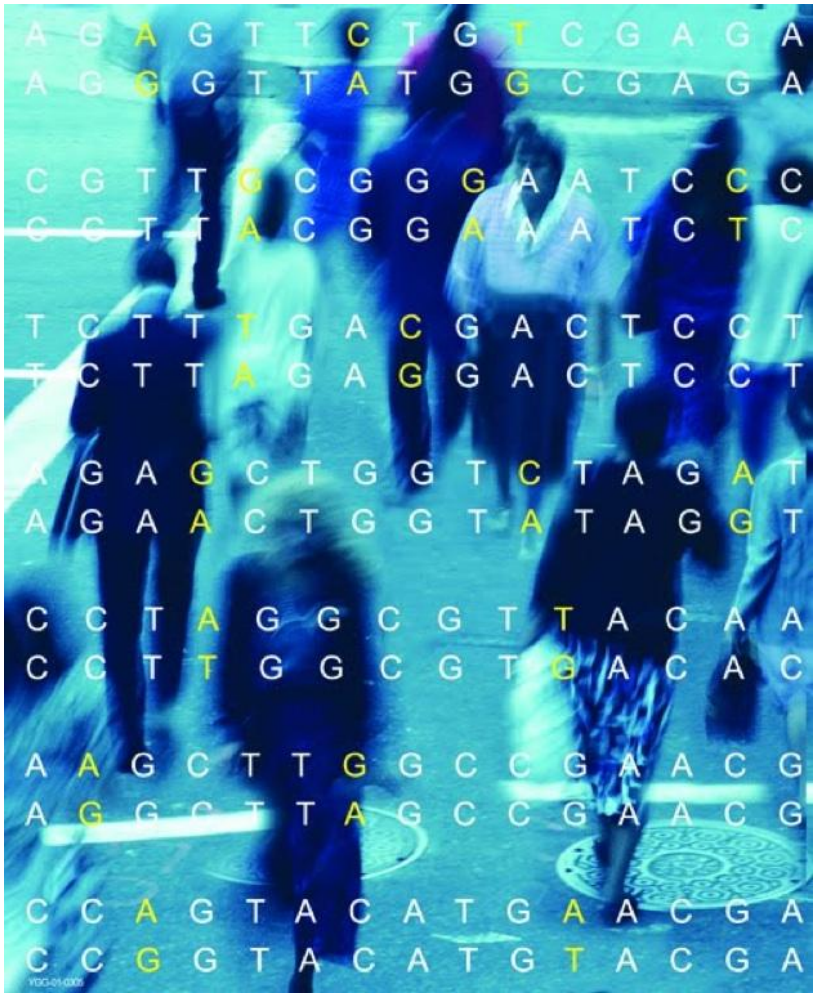
Sample	Preparation	Library	Sequencing	Analysis	Results

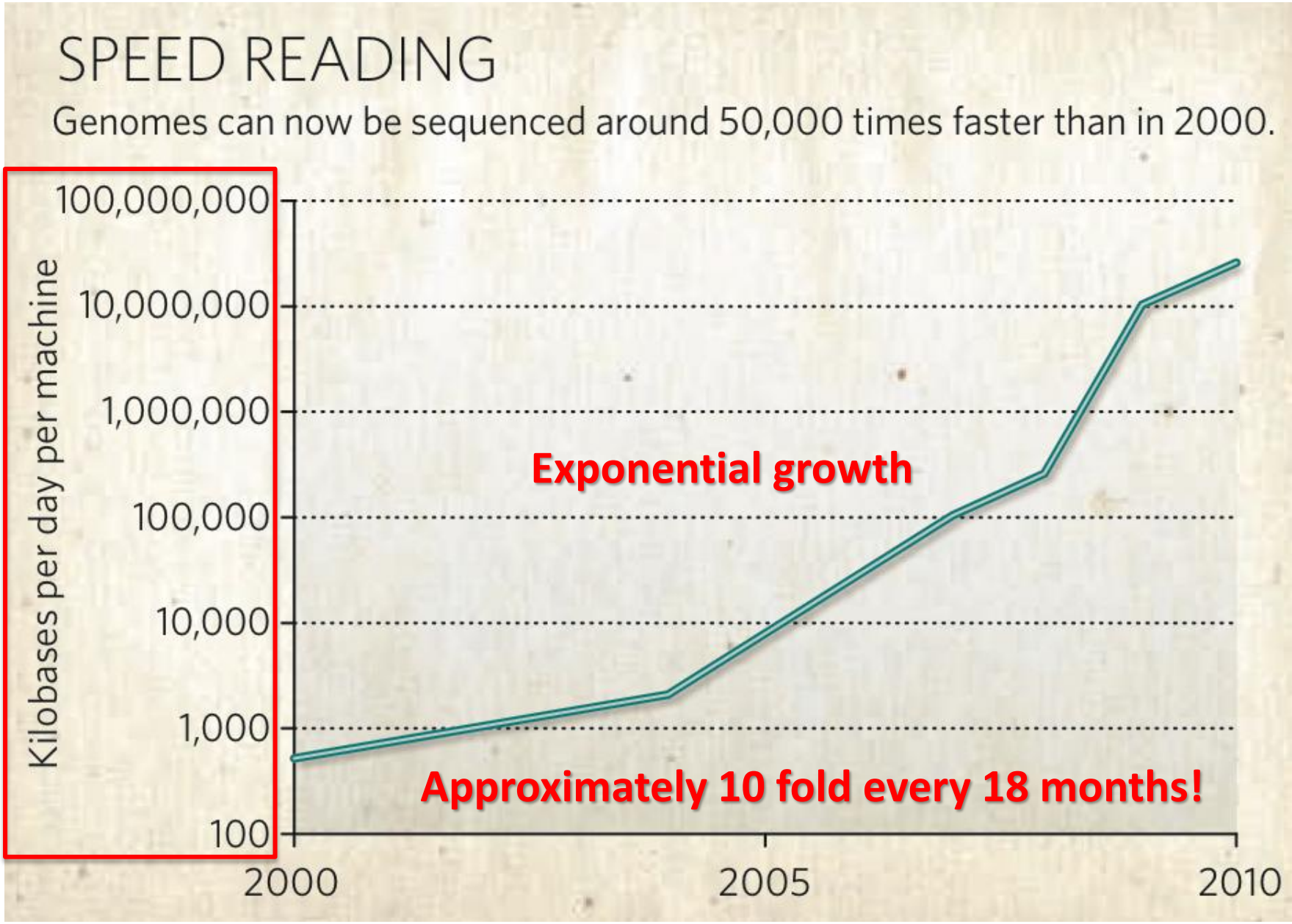
- Sequencing technology turns the secret codes (A, T, C, G) into digits
- Thus enabling computational research



Next Generation Sequencing (NGS)

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





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

Sequencing @BGI

- World's leading sequencing and genomics research center

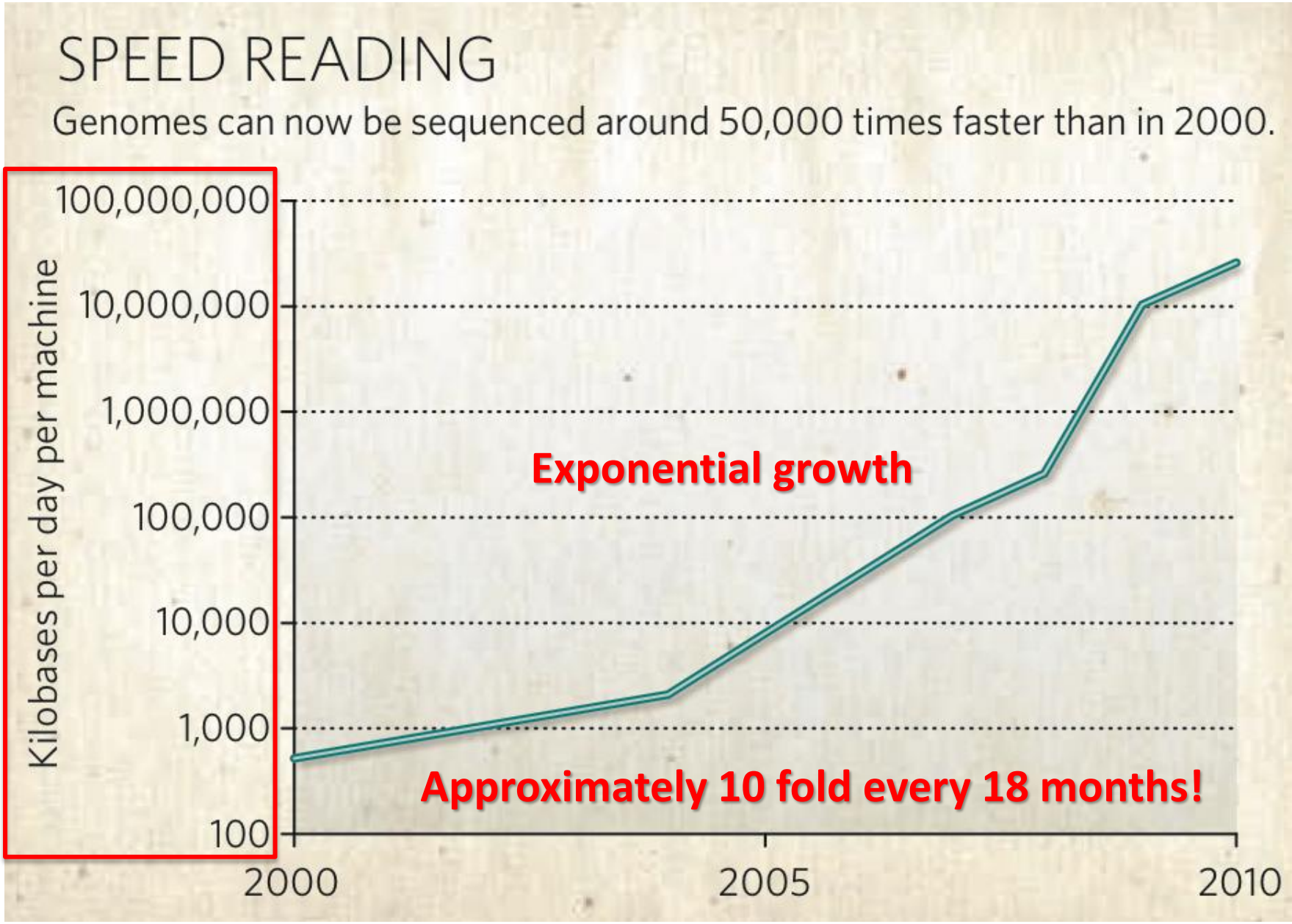




MODEL	ABI 3730XL	Roche 454	ABI SOLiD 4	Solexa GA IIx	Illumina HiSeq 2000
INSTALLATION	16	1	27	6	135

Computing @BGI

- 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
  - 17,000+ cores
  - ~160 Tflops peak performance
  - Still increasing …



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

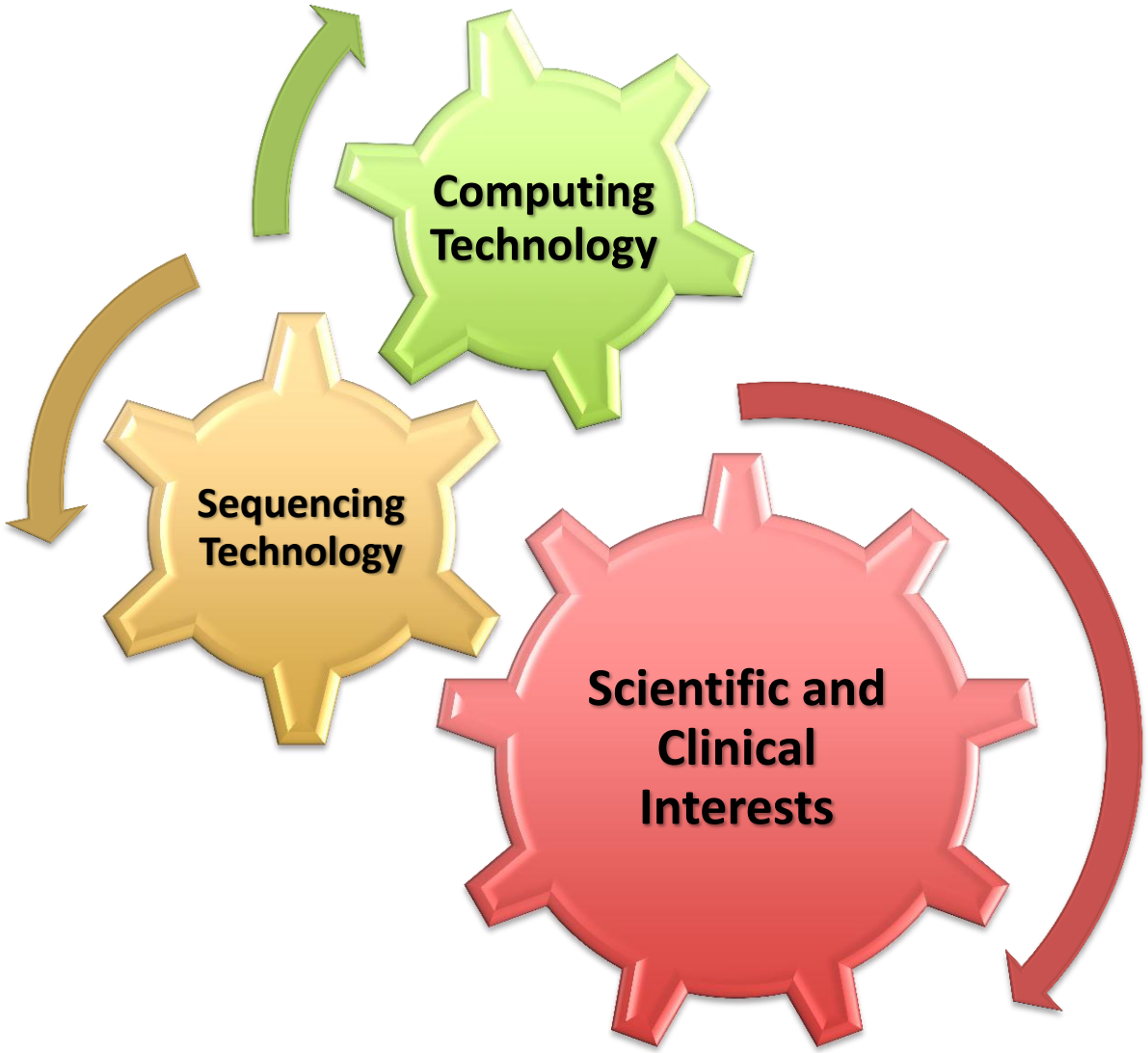
Sequencing vs Computing

- 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)***
  - …
- Growing computing demand
  - Population genomics **as *o(N^2)***
  - Gene association study **as beyond *o(N^2)***
  - Systems biology with various levels of data **as beyond *o(N^2)***
- Sequencing cost down leads to more and more high dimensional analysis
  - Computing intensive

Solution: Disruptive Computing Technology



GPU Accelerated Bioinformatics Research

- Individual tools for routine analysis
- Put them together to speed up workflows
- Leverage data access with compression
- Tackle challenging scientific questions

GPU Accelerated Bioinformatics Research

- **Individual tools for routine analysis**
- Put them together to speed up workflows
- Leverage data access with compression
- Tackle challenging scientific questions

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 Aligner - Design

- Optimized for GPU architecture
  - Reduce memory access
    - 1-level sampling instead of 2
    - Group data items according to retrieval patterns
    - Redundancy
    - A single search step takes two 32-bit & two 128-bit memory accesses instead of four 32-bit & two 256-bit in SOAP2
  - Reduce branching
    - Work with CPU in a multiple pass manner
    - CPU handles rest part of complicated reads leaved by GPU

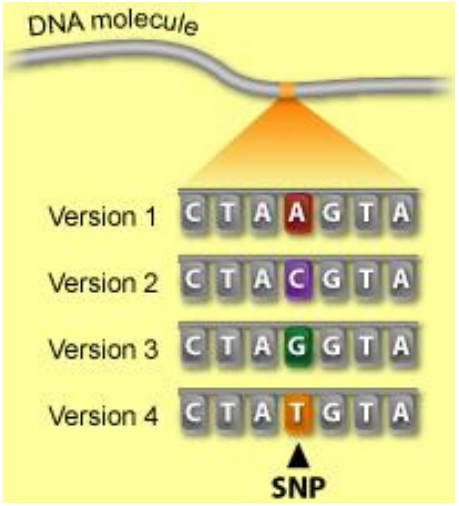
SOAP3 Aligner - Performance

- Performance
  - NVIDIA Tesla C2070 with 6GBytes memory
  - Intel Xeon E5630@2.53GHz, 2-way quad-core

Dataset	Reads Length	Mismatch	Load Timing (s)	Alignment and Output Timing (s)	Total Timing (s)	SOAP2 Timing (s)	Speedup
Human	100	3	83.30	128.23	211.53	2156.99	**16.17**
Drosophila	76	3	95.50	724.32	819.81	8936.27	**12.21**
Zabrafish	45	2	41.20	48.79	89.99	2792.97	**56.40**

SNP Detection Tool GSNP

- Single Nucleotide Polymorphism (SNP) detection for an individual
- Will be introduced by Mian Lu from HKUST

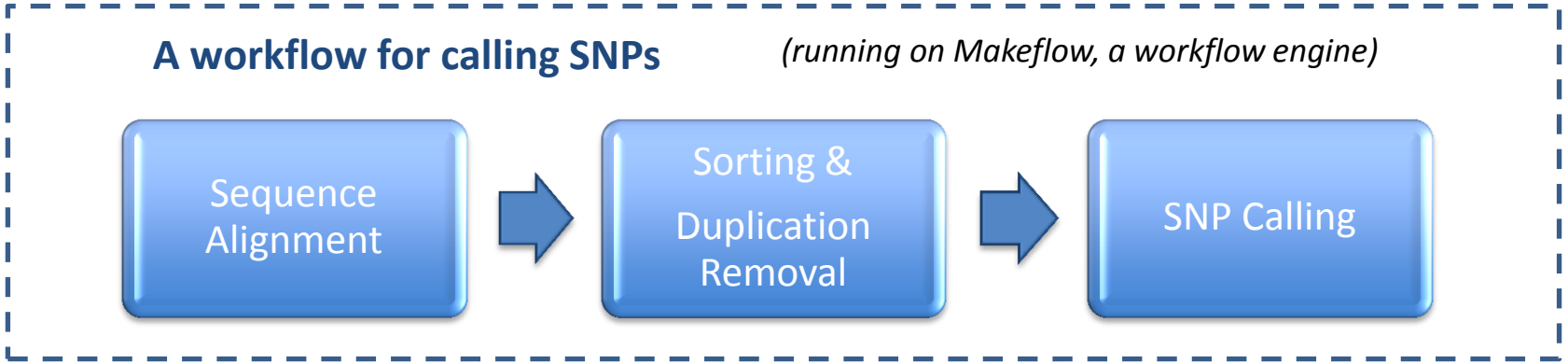


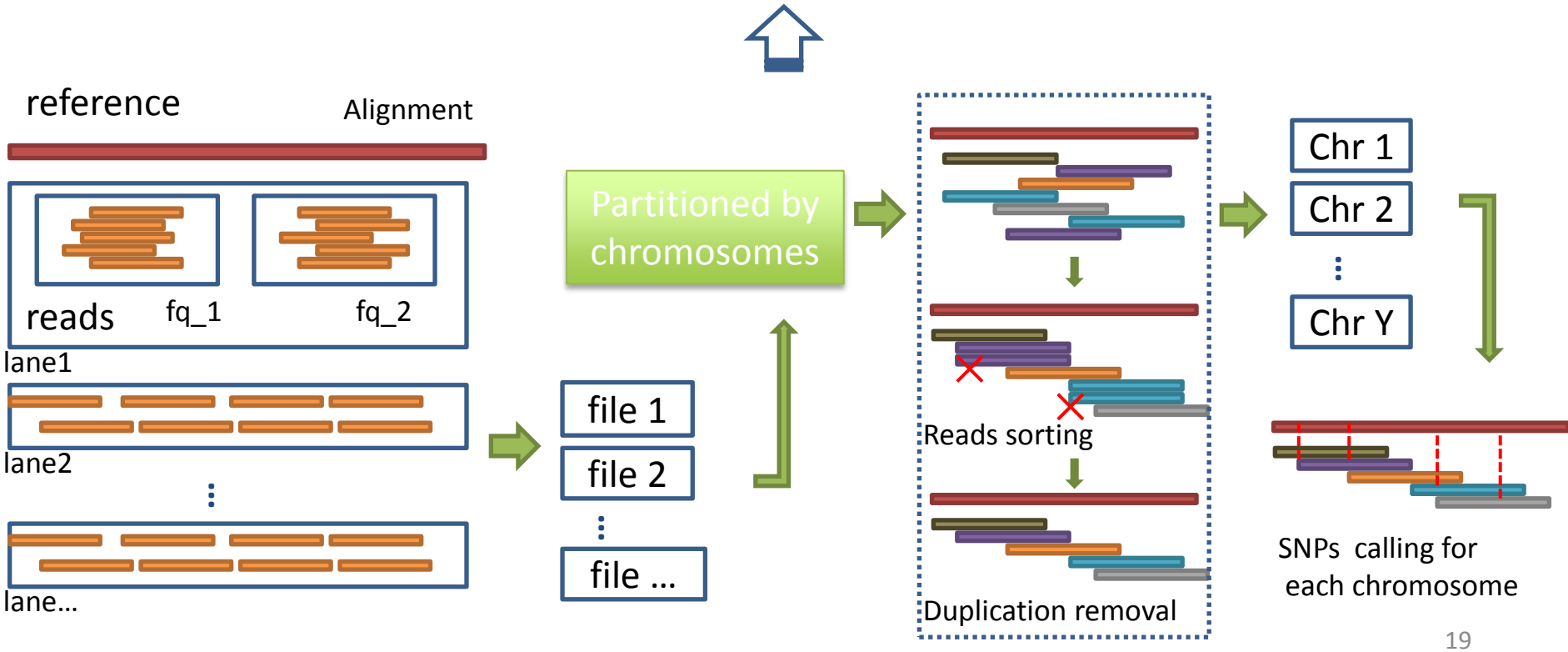
GPU Accelerated Bioinformatics Research

- Individual tools for routine analysis
- **Put them together to speed up workflows**
- Leverage data access with compression
- Tackle challenging scientific questions

A workflow for calling SNPs

(running on Makeflow, a workflow engine)





Timing break down of SNP calling workflow
Human genome, 30x

	Sequence Alignment		Sorting & Duplication Removal		SNP Calling	
CPU	BWA	~7 days	SAMtools	~19 hours	SOAPsnp	~4 days
GPU	SOAP3	9 hours	Adam*	7.5 hours	GSNP	14 hours**

Total timing for SNP calling workflow

CPU	**~12 days**
GPU	**~31 hours**

* Adam currently runs on CPU.

** Timing varies from previous report due to different input / output format and conversion.

Workflow Built with GPU Accelerated Tools

- Observation
  - Computation time drops a lot (11 days to 23 hours)
  - Data manipulation time decrease a little bit (~50%)
  - Now data manipulation step occupies ~25% of total time instead of 8%
  - The (computation time) / (data volume) ratio is low comparing with other application fields
- Solution
  - Using GPU to leverage data access with compression
  - Redundant computation to save data volume as well as loading time

GPU Accelerated Bioinformatics Research

- Individual tools for routine analysis
- Put them together to speed up workflows
- **Leverage data access with compression**
- Tackle challenging scientific questions

Leverage Data Access with Compression

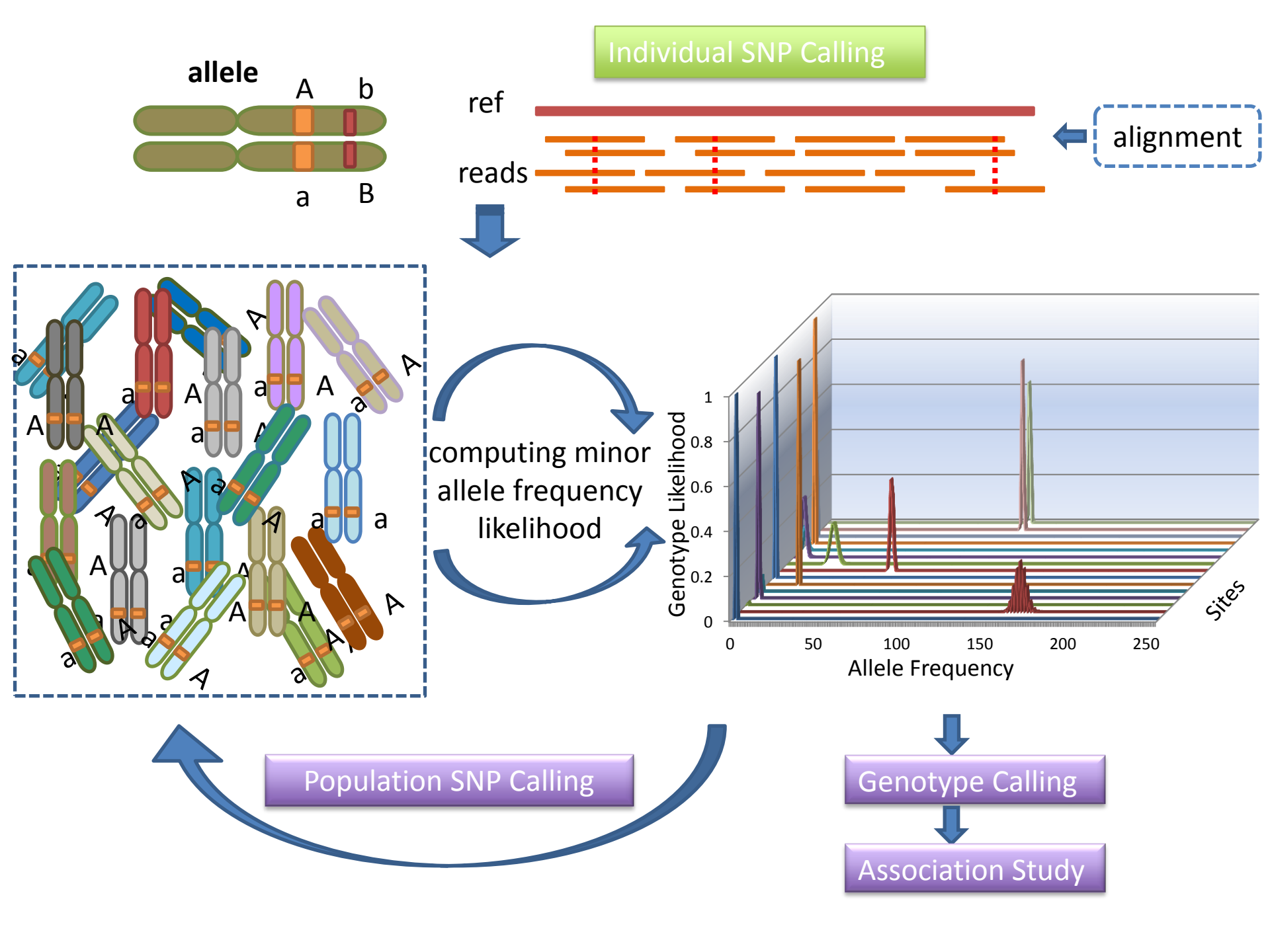
- Why compression?
  - More efficient utilization of storage space
  - Implicitly improve data access bandwidth
    - Shorten workflow execution time
    - Less computing hardware waste by reducing idle time for data to arrive
  - Use customized highly efficient algorithm for typical kinds of files

Leverage Data Access with Compression

- Case study
  - .fasta file with sequences of bases
  - Under development
    - Using align-to-reference compression
    - Up to 0.32 bits per base, 84% space saving
  - Completed
    - Hoffman based compression on GPU
    - Compression ratio **~24%**
    - Compression rate at **1GB/s**, decompression **1.5GB/s**
    - Comparing with gzip
      - ~20MB/s compression, ~60MB/s decompression

GPU Accelerated Bioinformatics Research

- Individual tools for routine analysis
- Put them together to speed up workflows
- Leverage data access with compression
- **Tackle challenging scientific questions**


allele
A
b
a
B
Individual SNP Calling
ref
reads
alignment
computing minor allele frequency likelihood
Genotype Likelihood
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200
250
Allele Frequency
Sites
Population SNP Calling
Genotype Calling
Association Study

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 ($o(N^2)$)
- Details will be presented by Mian Lu from HKUST

Estimating MAF in a Population with *Multiple* GPUs

- Joint development with Tianjin Supercomputing Center
  - Based on TH-1A
  - CUDA + MPI
  - Parallel data loading
  - AllToAll comm

Estimating MAF in a Population with *Multiple* GPU

Dataset: Human genome, 1024 individuals, full scan of 3G sites
Note: All data below are given according to partial scale test timing.
The CUDA + MPI version performance is estimated with model.

Version	Timing	Speedup	Note
CPU	~ 10 years		
GPU (Single)	~ 0.5 years	20	against CPU
GPU (64 with MPI)	~ 51 hours	84	against single GPU
GPU (256 with MPI)	~ 13 hours	329	against single GPU

Conclusion from Preliminary Results

1. Numerical consistency is maintained between different versions.
2. Data access consumes lots of time, and loading data in a parallel manner helps a lot.
3. Computation part speedup is actually much higher.

Summary

- GPU is very promising to accelerate bioinformatics analysis and life science research
- Efforts need to be made to leverage data access and computation
- Workflow engine needs significant improvement to support GPU computing better

Acknowledgement

- Our collaborators
  - Prof T.K. Lam, Dr S.M. Yiu from HKU
  - Prof Qiong Luo and group member from HKUST
    - Mian Lu
    - Jiuxin Zhao
    - Yuwei Tan
  - Prof Xiaowen Chu and Kaiyong Zhao from HKBU
- NVIDIA China DevTech Team
  - Bin Zhou, Peng Wang and Agatha Hu

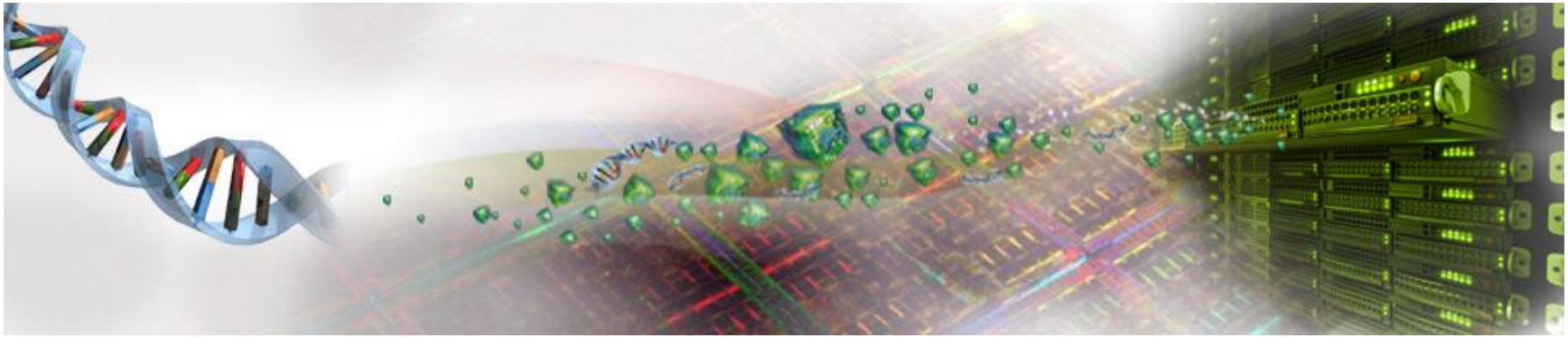
We are looking …

… for guys like you …

… to join us

contact

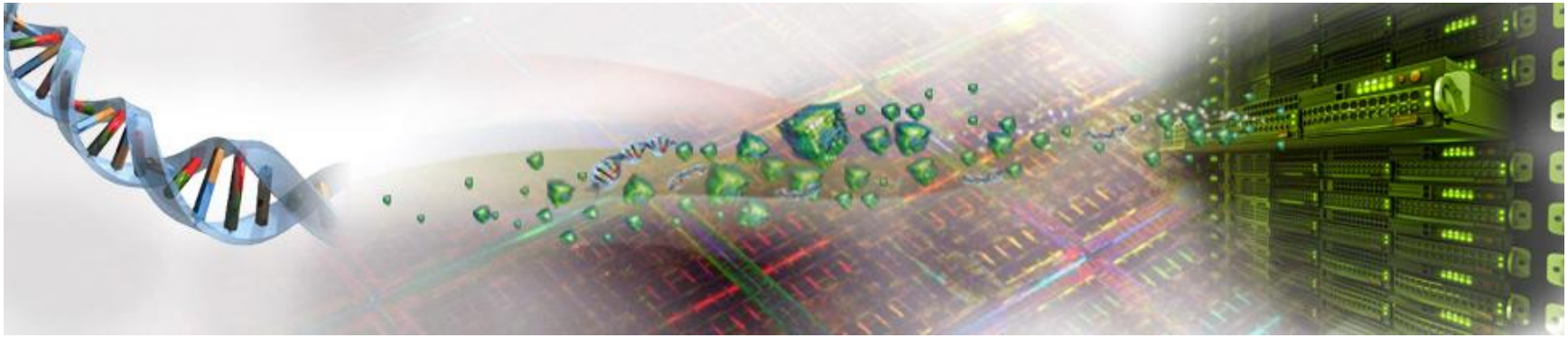
wangbingqiang@genomics.cn



Stay tuned with us

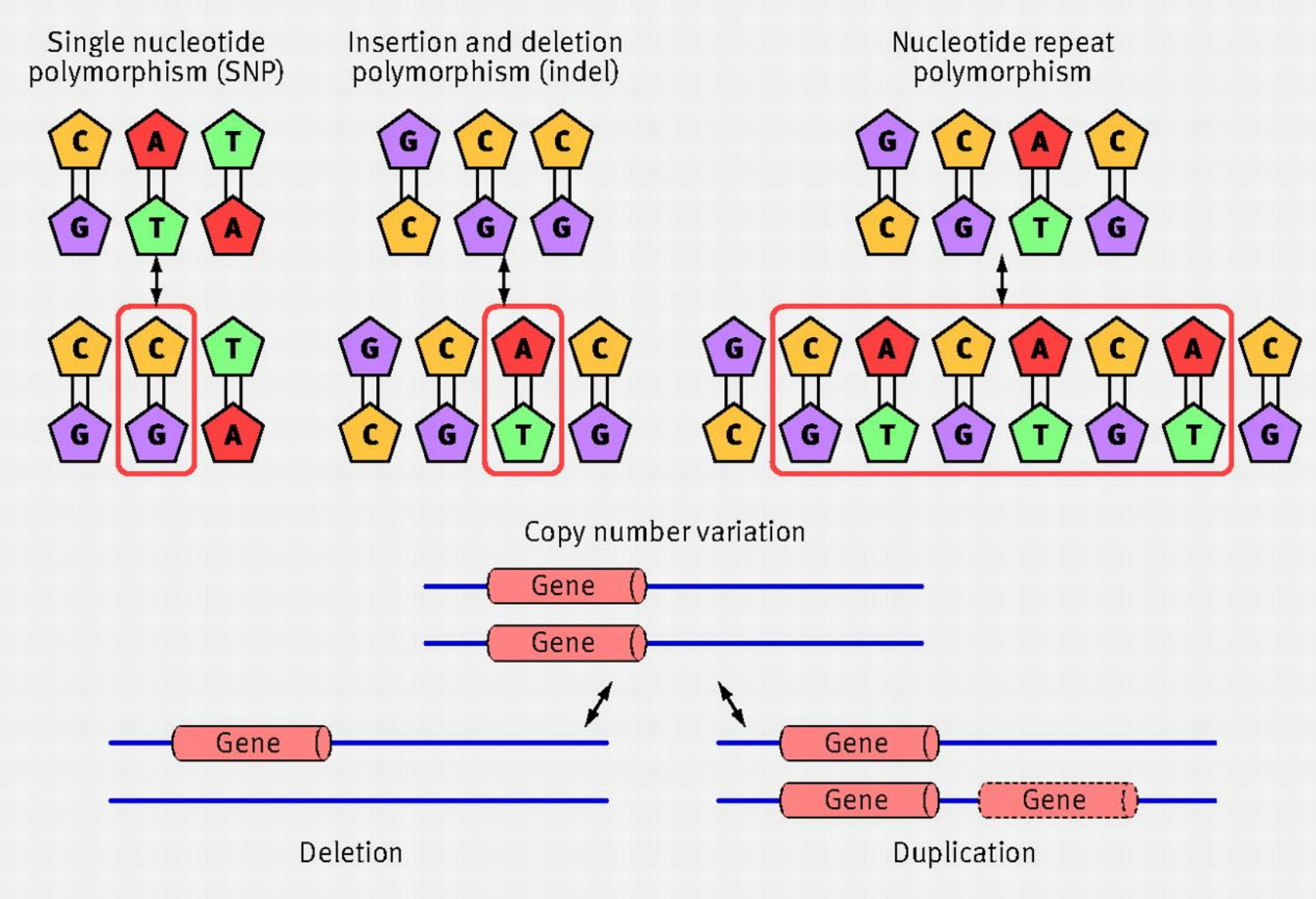
http://jil.genomics.cn/

BGI-NVIDIA Joint Innovation Lab



Thank you

Different Types of Gene Variation



SOAP2 Performance

Software	Reads aligned	Time (paired, s)	Time (single, s)	Memory Usage (GB)
SOAP2	93.6%	828	478	5.4
SOAP	93.8%	19,234	14,328	14.7
MAQ	93.2%	22,506	19,847	1.2
Bowtie	91.7%	-	405	2.3

Query dataset of one million read pairs generated by the Illumina Genome Analyzer on a human DNA sample, read length 44 bp. The paired-end insert size was about 200 bp. The human reference genome was NCBI build 36.1.