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 30,000 genes code for proteins that perform most life functions
Different Types of Gene Variation

- Single nucleotide polymorphism (SNP)
- Insertion and deletion polymorphism (indel)
- Nucleotide repeat polymorphism

Copy number variation

- Gene
- Gene

Deletion

Duplication
Sequencer, Sequencing, Sequence

- Sequencing technology turns the secret codes (A, T, C, G) into digital form
- Thus enabling computational research
Sequencing long DNA

Multiple copies of the same sequence

Randomly fragment the copies
Next Generation Sequencing (NGS)

- Indeed 2\textsuperscript{nd} generation sequencing technology
- Low cost (\textit{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
SPEED READING
Genomes can now be sequenced around 50,000 times faster than in 2000.

Exponential growth

Approximately 10 fold every 18 months!

Collecting and integrating large-scale, diverse types of data

a Many different types of data can be systematically scored
- Different gene isoforms
- Histone modification
- DNA methylation
- Protein phosphorylation
- Gene expression and non-coding RNA
- Metabolites
- Protein expression

b These data can be integrated to build predictive models
- DNA variation
- Protein–protein interactions and protein complexes
- Gene expression
- DNA–protein binding

c Networks over multiple tissues can be combined to model the system
... 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 @ 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

<table>
<thead>
<tr>
<th>MODEL</th>
<th>ABI 3730XL</th>
<th>Roche 454</th>
<th>ABI SOLiD 4</th>
<th>Solexa GA IIx</th>
<th>Illumina HiSeq 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSTALLATION</td>
<td>16</td>
<td>1</td>
<td>27</td>
<td>6</td>
<td>135</td>
</tr>
</tbody>
</table>
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 ...
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 \textbf{NP hard}
  – Gene association study \textit{high dimensional}
  – Systems biology with various levels of data \textbf{NP hard}
  – ...

• Sequencing cost down leads to more and more high dimensional analysis
  – Lots, lots of computing
Solution: Disruptive Computing Technology

Computing Technology

Sequencing Technology

Scientific and Clinical Interests
GPU Accelerated Bioinformatics Research

• Individual tools for routine analysis
  – SOAP3 / SOAP3-DP aligner
  – SNP calling with GSNP

• Tackle challenging scientific questions
  – Gene association study
    • First step: High resolution genotyping with GAMA-MPI
  – More incoming
SOAP3 / SOAP3-DP Aligner

• 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.
• Collaboration between University of Hong Kong (HKU) and BGI
  – Professor T.W. Lam
• SOAP3 and SOAP3-DP
  – Designed for GPU, great performance
  – SOAP3: GPU accelerated version of SOAP2
    • Compressed indexing: bidirectional Burrows-Wheeler transform
  – SOAP3-DP: index-assisted dynamic programming (semi-global alignment)
    • Alignment with INDELs (insert/delete) & mismatches
• **SOAP3: GPU-based Compressed Indexing and Ultra-fast Parallel Alignment for Short Reads**
  – Wednesday (Tomorrow) 16:00-16:30, Hall B
  – On behalf of Professor T.W. Lam and team members
<table>
<thead>
<tr>
<th>Data type</th>
<th>Reads length (bp)</th>
<th>Total Number of Reads (million)</th>
<th>Mismatch number</th>
<th>SOAP3 (Total Time: second)</th>
<th>SOAP2 Total time (second)</th>
<th>Alignment Speed-up ratio (second)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Time for reading reads</td>
<td>Time for alignment and output</td>
<td>Total time</td>
</tr>
<tr>
<td>Human</td>
<td>100</td>
<td>16</td>
<td>3</td>
<td>83.30</td>
<td>128.23</td>
<td>211.53</td>
</tr>
<tr>
<td>Zebra fish</td>
<td>76</td>
<td>21</td>
<td>3</td>
<td>95.50</td>
<td>724.32</td>
<td>819.81</td>
</tr>
</tbody>
</table>

**Alignment Ratio (%)**

- **Human**
  - SOAP2: 84.2%
  - SOAP3: 88.29%

- **Zebra fish**
  - SOAP2: 64.49%
  - SOAP3: 76.55%

**Speedup Ratio**

- **Human**: 14.12
- **Zebra fish**: 14.6

**Total Time (second)**

- **Human**: 1893.45 seconds
- **Zebra fish**: 10671.39 seconds
Speedup Compared with Other Tools

- SOAP3-dp is at least 10 times faster when comparing with other tools (BWA, Bowtie2)
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
### Differences from reference for each read

<table>
<thead>
<tr>
<th>Site</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>6</td>
<td>T</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Alignment

<table>
<thead>
<tr>
<th>Read quality score</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The parallelization strategy on the GPU: one thread handles one site

The consistency of GPU and CPU results

- Reduce memory overhead and branch divergence
- Balance workloads
- The consistency of GPU and CPU results
- Reduce I/O cost
The sparse representation of aligned bases

The measured non-zero%: ~0.1%

GSNP

Site 1

SOAPsnp

Site 2

GSNP

Solution to the consistency of GPU and CPU results

Searching for the result

64 possible results

Constant memory on the GPU

Computing the result
The elapsed time of all components are included. GSNP is around 50X faster than the single-thread CPU-based SOAPsnp.
GPU Accelerated Bioinformatics Research

• Individual tools for routine analysis
  – SOAP3 / SOAP3-DP aligner
  – SNP calling with GSNP

• Tackle challenging scientific questions
  – Gene association study
    • First step: High resolution genotyping with GAMA
  – More incoming
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)
The site has a probability for the occurrence of allele ‘a’ and ‘A’.

One individual

Probability of AA Aa aa

Compute minor allele frequency

Add another individual

The site has a probability for the occurrence of allele ‘a’ and ‘A’.
Different sites represent different alleles.

Compute allele frequency likelihood for each site.
One Site Computation

A thread handles a site?

Multiple threads handle a site

Bad solution!

SFS construction

Thread 0
Thread 1
Thread 2
Thread 3
Thread 4
Estimating MAF in a Population with *Multiple* GPUs

- Based on TH-1A
- CUDA + MPI
- Parallel file system (Lustre)
- AllToAll comm (now)
GAMA-MPI Flow Chart

node 1

- start
- wait
- calculate and output result
- wait
- calculate and output result
- wait
- calculate and output result

Node N

- decompress and parse data
- Alltoall data exchange
- decompress and parse data
- Alltoall data exchange
- decompress and parse data
- Alltoall data exchange
- wait
- finish
Dataset: Human genome, 512 individuals (1024 input files), full scan of 3G sites

<table>
<thead>
<tr>
<th>Version</th>
<th>Computing time</th>
<th>Total Time</th>
<th>Computing Speedup</th>
<th>Total Speedup</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>~ 1518 days</td>
<td>~ 1619 days</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPU (Single)</td>
<td>~ 15.75 hours</td>
<td>~ 101 days</td>
<td>2313</td>
<td>16</td>
<td>against CPU</td>
</tr>
<tr>
<td>GPU (86 with MPI*)</td>
<td>~ 717 seconds</td>
<td>~ 5.4 hours</td>
<td>79</td>
<td>449</td>
<td>against single GPU</td>
</tr>
</tbody>
</table>

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

• GPU is very promising to accelerate bioinformatics analysis and life science research
• Efforts need to be made to leverage data access and computation
• For large-scale data analysis, especially GWAS type analysis, more study needed
Acknowledgement

• Team members and Xing Xu, Lin Fang
• Our research 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 from HKBU
    • Kaiyong Zhao
• National Supercomputing Center at Tianjin
Acknowledgement (cont’d)

• NVIDIA
  – Dr Simon See
  – Allan Hou
  – China DevTech Team
    • Dr Peng Wang
    • Dr Bin Zhou
    • Agatha Hu
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

Problems:

• Big genomic data
• Geological distribution
• Algorithm integration
• Computational demand

EasyGenomics

https://www.easygenomics.com
EasyGenomics™ is the bioinformatics platform for research and applications on the cloud.
Thank you

wangbingqiang@genomics.cn