Accelerate High Throughput Analysis for Genome Sequencing with GPU

ATIP - A*CRC Workshop on Accelerator Technologies in High Performance Computing
May 7-10, 2012 Singapore

BingQiang WANG, Head of Scalable Computing, BGI
Simon SEE, Chief Scientific Computing Advisor, BGI
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**

- Deletion
- Duplication
Sequencing technology turns the secret codes (A, T, C, G) into digital form.

Thus enabling computational research.
Next Generation Sequencing (NGS)

- Indeed 2\textsuperscript{nd} generation sequencing technology
- Low cost \textit{(several K\$ per human genome)}
- High throughput
- Short reads \textit{(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, dynamics
  – Living cells, living life
Genomes can now be sequenced around 50,000 times faster than in 2000.

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
• Mass spectrometers to capture protein information
  – Complement sequencing
  – Proteome

<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 @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
  – ~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 $o(N)$
  – ...
• Growing computing demand – let us mine for “gold”
  – 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
  – Lots, lots of computing
Solution: Disruptive Computing Technology

Scientific and Clinical Interests

Computing Technology

Sequencing Technology
GPU Accelerated Bioinformatics Research

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

• Tackle challenging scientific questions
  – Gene variation and association
    • First step: High resolution genotyping with GAMA-MPI
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
Reference reads

BWT Algorithm

Alignment

Sampled suffix array

Rotation & sorting

BWT Index

One site by one site --- from right to left

backward search

read
2-BWT: Implemented in SOAP2 and SOAP3

1- Mismatch alignment

**Case A:** The mismatch lies on the first X characters

```
  x
```

**Case B:** The mismatch lies on the last m-X characters

```
  x
```

2- Mismatch alignment

**Case A:** Both mismatches lie on P[1..X+Y]

```
  x  x
```

**Case B:** Both mismatches lie on P[X+Y+1..m]

```
  x  x
```

**Case C:** 1st mismatch lies on P[X+1..X+Y], 2nd mismatch lies on P[X+Y+1..m]

```
  x  x
```

**Case D:** 1st mismatch lies on P[1..X], 2nd mismatch lies on P[X+Y+1..m]

```
  x  x
```
SOAP3 core ideas

Reduce memory access

Engineering the index
- 2-level sampling becomes 1-level sampling
- group data items according to retrieval patterns instead of logical functions
- redundancy

Control branching effect

Round 1
only finish those really simples one

Round 2
group those complicated reads together for another round

Round 3
extremely complicated ones are left to the CPU

a search step takes **four 32-bit & two 256-bit** memory accesses

a search step takes **two 32-bit & two 128-bit** memory accesses

SOAP3 V.S SOAP2

- 2-level sampling becomes 1-level sampling
- group data items according to retrieval patterns instead of logical functions
- redundancy

- extremely complicated ones are left to the CPU
Memory-resident data structures

2BWT  SA

Execution

Process complicated reads & load more reads
Process complicated reads & load more reads
Process complicated reads & load more reads

Memory-resident data structures

2BWT

Execution

Process reads with simple alignment structure
Process reads with simple alignment structure
Process reads with simple alignment structure
Process reads with
<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>SOP2 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>

**Total Time (second)**

- Human: 1893.45
- Zebra fish: 10671.39

**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
Compared with its predecessor SOAP3

- Actual alignment time (excluding index and read loading time)

- SOAP3-dp is about 2.5 times more efficient.
Single-end Alignment Compared with Other Tools

- SOAP3-dp is the fastest and reports the highest percentage of aligned reads.
Paired-end Alignment Compared with Other Tools

- SOAP3-dp also show it’s superior in paired-end alignment
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.
GSNP

Data quality
Experimental errors
Alignment

Differences from reference for each read

Alignment

Prior probability of each genotype
Bayesian theory

Alignment quality score for each base

Calculate likelihood for each genotype
Infer genotype

Quality score

Calculate likelihood

Infer genotype

Bayesian theory

Prior probability

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score

Differences from reference

Alignment

Quality score
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
The sparse representation of aligned bases

constant memory on the GPU

64 possible results

Searching for the result

Solution to the consistency of GPU and CPU results

Computing the result

The measured non-zero%: ~0.1%
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.
GPU Accelerated Bioinformatics Research

• Individual tools for routine analysis
  – SOAP3 / SOAP3-DP aligner
  – SNP calling with GSNP
• Tackle challenging scientific questions
  – Gene variation and association
    • First step: High resolution genotyping with GAMA-MPI
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 \( (o(N^2)) \)
The site has a probability for the occurrence of allele ‘a’ and ‘A’.
Different sites represent different alleles

<table>
<thead>
<tr>
<th>Individual 1</th>
<th>Individual 2</th>
<th>Individual 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>...</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Compute allele frequency likelihood for each site
One Site Computation

A thread handles a site?

SFS construction

Bad solution!

Multiple threads handle a site

Thread 0  Thread 1  Thread 2  Thread 3  Thread 4
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
Dataset: Human genome, 512 individuals (1024 input files)
Running on TH-1A supercomputer, GPU accelerated version

<table>
<thead>
<tr>
<th>Node</th>
<th>Sites</th>
<th>Computing Time</th>
<th>Parse Time</th>
<th>Alltoall Time</th>
<th>Output Time</th>
<th>Total Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>3G</td>
<td>1h55m</td>
<td>20h10m</td>
<td>28m</td>
<td>1h41m</td>
<td>26h30m</td>
</tr>
<tr>
<td>8</td>
<td>10M</td>
<td>23.01s</td>
<td>242s</td>
<td>5.61s</td>
<td>20.27s</td>
<td>318s</td>
</tr>
<tr>
<td>8</td>
<td>1M</td>
<td>2.40s</td>
<td>24.08s</td>
<td>0.88s</td>
<td>1.19s</td>
<td>40s</td>
</tr>
</tbody>
</table>
Estimating MAF in a Population with *Multiple* GPUs

- Joint development with Tianjin Supercomputing Center
  - Based on TH-1A
  - CUDA + MPI
  - Parallel file system (Lustre)
  - AllToAll comm (now)
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
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
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
• National Supercomputing Center at Tianjin
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

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

Solutions

Cloud
High Speed Data Exchange
Workflows
+ Resource Management

Easy Genomics

https://www.easygenomics.com

42
EasyGenomics™ is the bioinformatics platform for research and applications on the cloud.
Thank you

wangbingqiang@genomics.cn
Backup Slides
GAMA-MPI Flow Chart

node 1

1

Node N

main

decompress and parse data

thread

Alltoall data exchange

M

decompress and parse data

wait

calculate and output result

wait

calculate and output result

wait

calculate and output result

finish