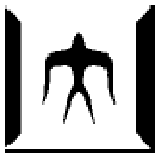
# An ultra-fast computing pipeline for metagenome analysis on TSUBAME 2.0

Yutaka Akiyama



Graduate School of Information Science and Engineering

Tokyo Institute of Technology

2011/12/14 GTC Asia (Beijing)

TOKYO TECH Pursuing Excellence

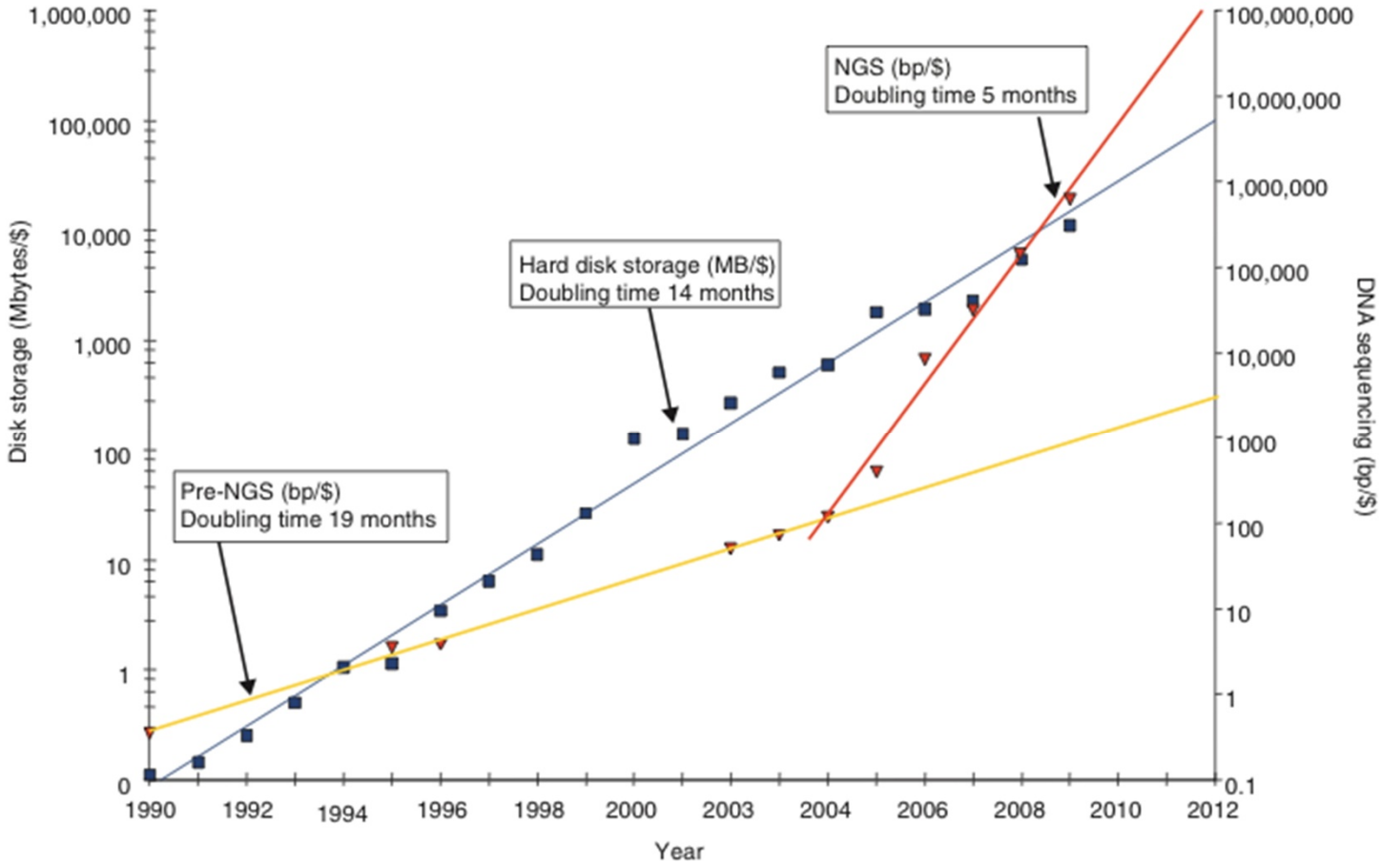
# Agenda

- Background
  - Rapid improvement of DNA sequencing technologies
  - Metagenome analysis

- An automated pipeline system for metagenome analysis on TSUBAME 2.0
- GPU accelerated homology search tool  GHOSTM
- Large-scale performance evaluation on TSUBAME 2.0

TOKYO TECH Pursuing Excellence

# Rapid improvement of DNA sequencing technology




Whole human genome (about 3 Gbp) now can be sequenced by about $1000.

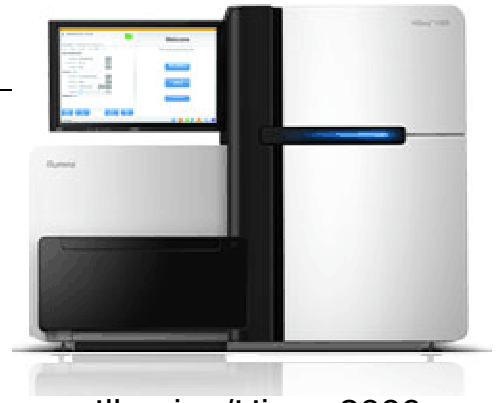
Lincoln Stein, Genome Biology, vol. 11(5), 2010

TOKYO TECH Pursuing Excellence

# Next-generation sequencer

- Next-generation sequencers (NGSs) can produce more than 100Gb genomic data on a single run

| | Pre-NGS (PRISM3730x) | NGS (Hiseq 2000) |
|---|---|---|
| Read length (bp) | 700 | 100 |
| Run time | 1 hour | 11 days |
| Data size (bp) | 67K | 600G |
| Throughput/day (bp) | 1.6M | 55G (55000M) |

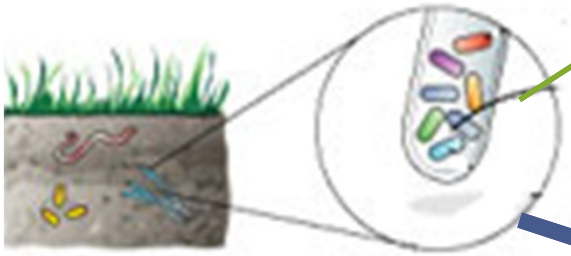


Illumina/Hiseq 2000

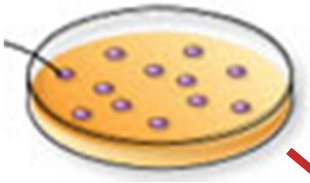
# Metagenome analysis

## General genome analysis

Reveals genomic data of a single organism



Environment
(including various organisms)



Select one organism and culture it

Sequence

Sequence directly

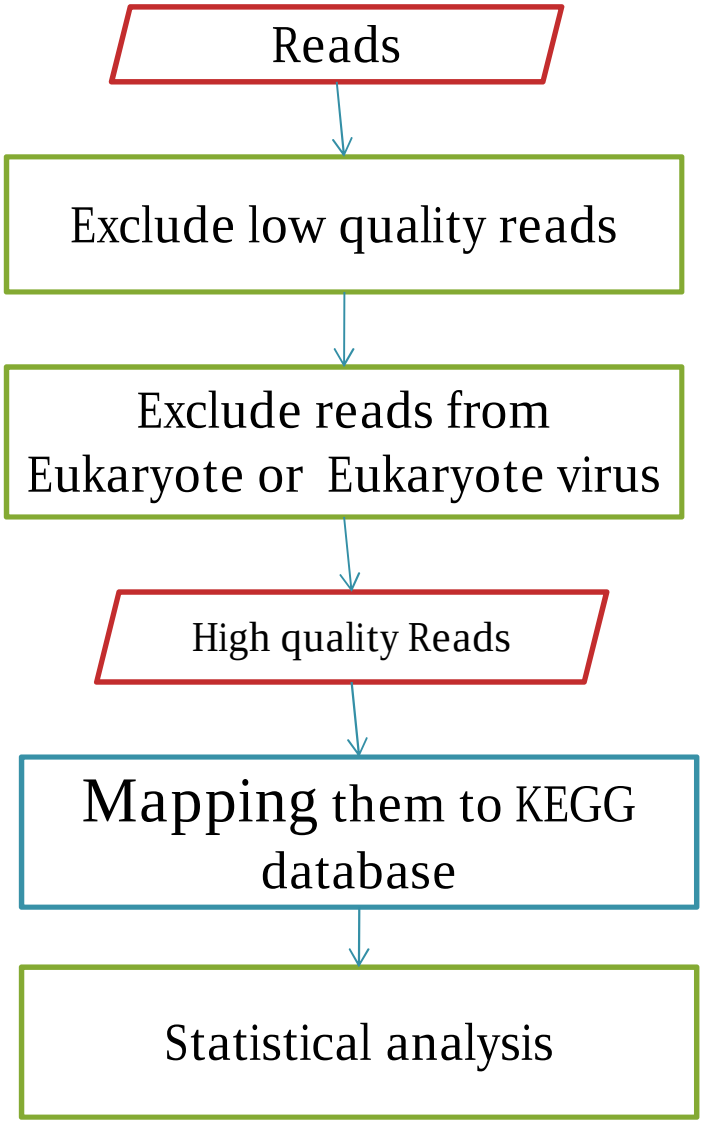
## Metagenome analysis

Directly analyze genetic materials of **all organisms** sampled from environment

- Identify genes and metabolic pathways of the environment
- Compare them to other environment

TOKYO TECH Pursuing Excellence

# A metagenome analysis pipeline for NGS reads




use only reads with Y flags and including bases better than B

Homology search for NCBI **nr** database

Homology search for **KEGG** genes.pep database

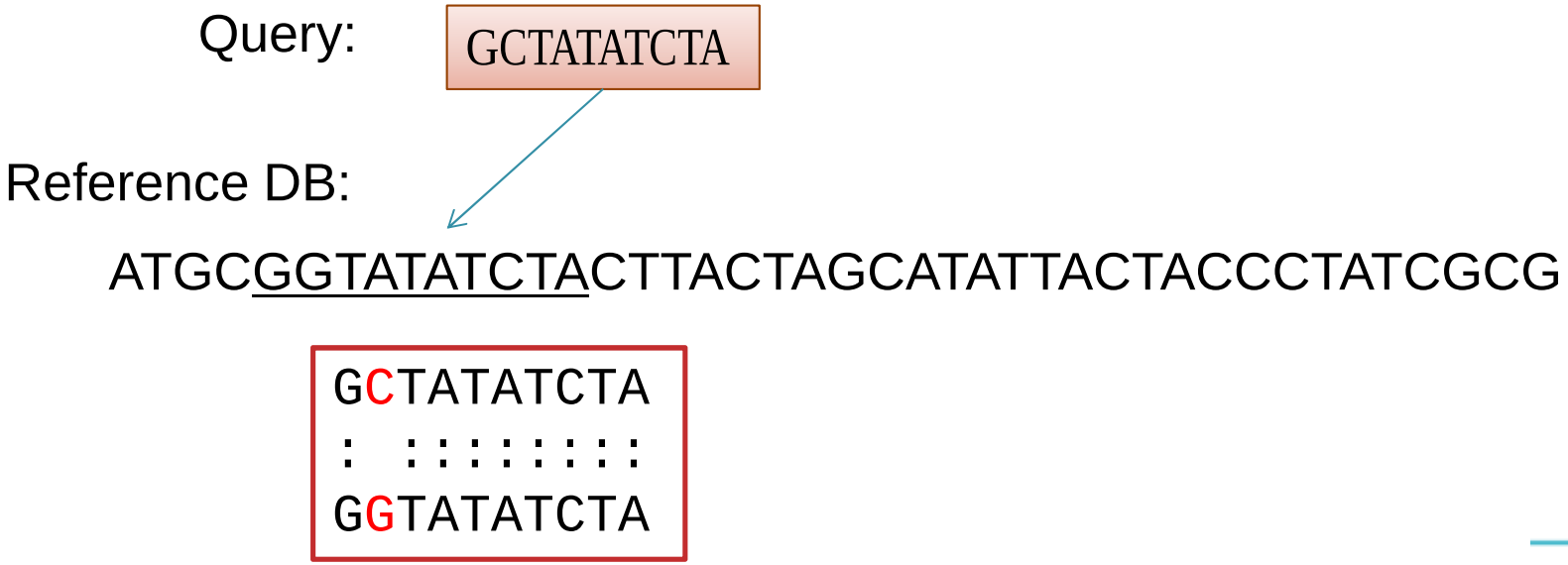
- Top hits & seq. id. > 0.7 & bit score > 40

(designed by Prof. Ken Kurokawa, Tokyo Tech.)

TOKYO TECH Pursuing Excellence

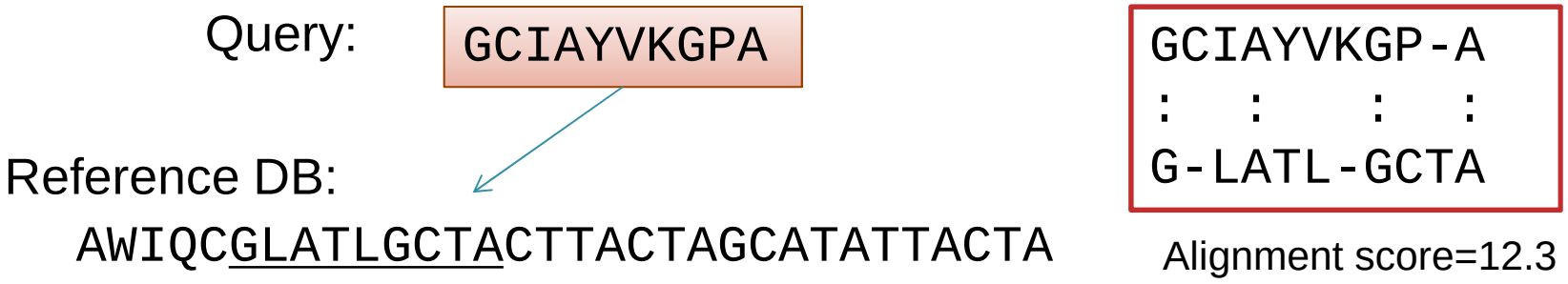
# Genome mapping

- Search same sequence for databases
  - Whole genome of the organism is already known
  - Search DNA sequence database (4 types, ATGC)
  - Only accept small (1, 2 or 3) errors
    - Sequencing errors
    - SNPs



# Metagenome mapping

- Search similar sequence for databases
  - Genomic data of same organism is unavailable
  - Sensitive homology searches are required
    - Have to search similar sequence for the sequences of homologues (similar species)
    - Search for amino acid sequence database (20 types)
    - Accept many mutations, insertion and deletion
    - Evaluate each match by using a score matrix



# Metagenomic analysis requires vast amount of computation

- BLASTX program (Altschul *et al.*, *JMB*, 1990) is generally used for metagenome mapping
  - Standard efficient sequence homology search software developed and maintained by NCBI
  - Requires large computational power

About 400 hours are required for an output of a single run of a NGS (20,000,000 reads) by using a small PC cluster (144 cores) (Prof. Ken Kurokawa)

Development of an automated metagenomic pipeline system on TSUBAME 2.0

# An automated pipeline system on TSUBAME 2.0

- Can utilize hundreds of computation nodes (thousands of CPU cores)
- Efficient DB copy
  - DB data are simultaneously copied from local disk (SSD) of a node to another in a binary-tree manner
- Web-interface (under development)

- Homology search tools;
  - BLASTX (Altschul *et al.*, *JMB*, 1990)
  - GHOSTM (Suzuki *et al.*, submitted.)

TOKYO TECH Pursuing Excellence

# GHOSTM

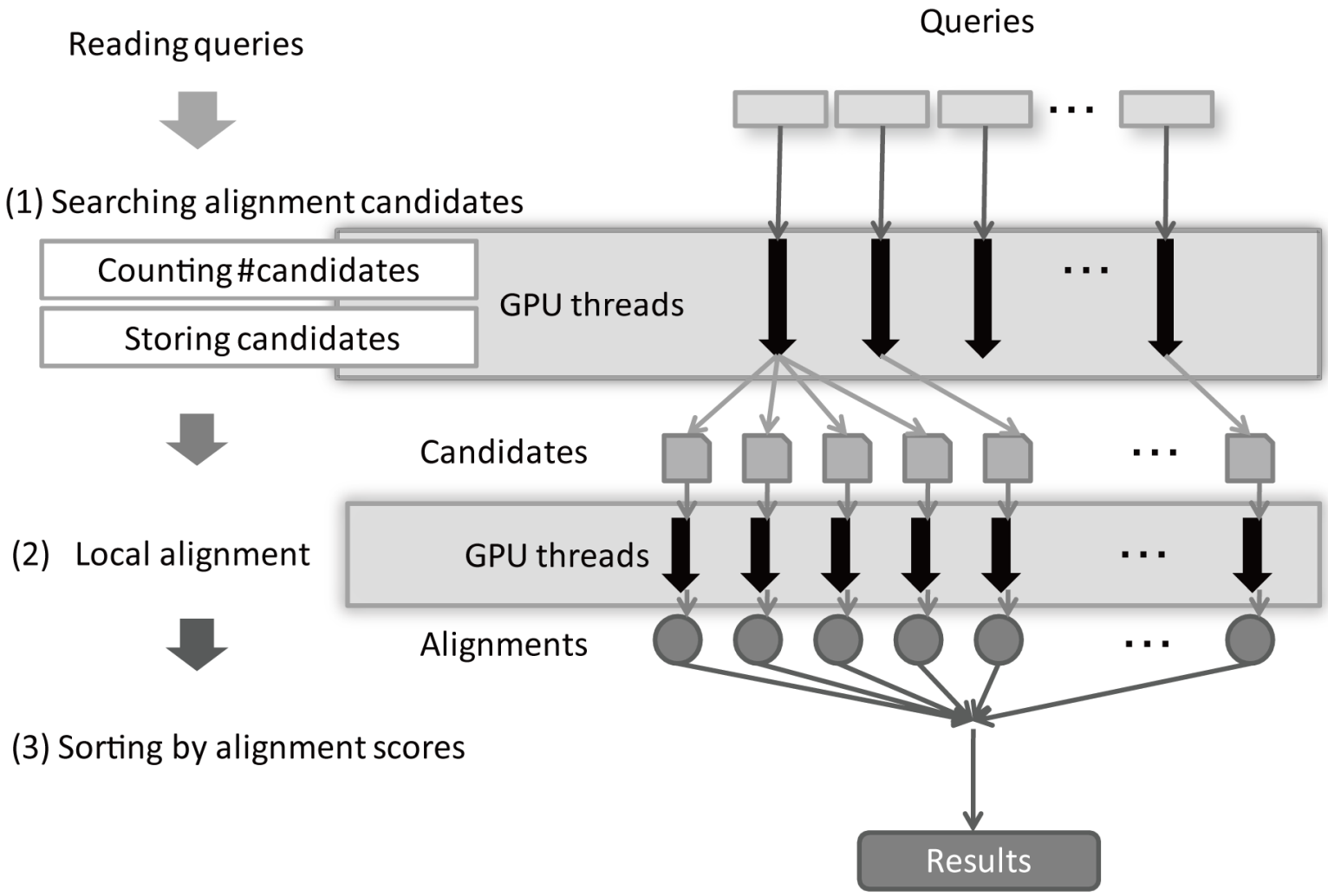
- GPU-based HOmology Search Tool for Metagenomics

- Perform fast and sensitive homology search by using GPU computing technique
  - Implemented by NVIDIA's CUDA (required ver. 2.2 or higher)

- http://www.bi.cs.titech.ac.jp/ghostm/

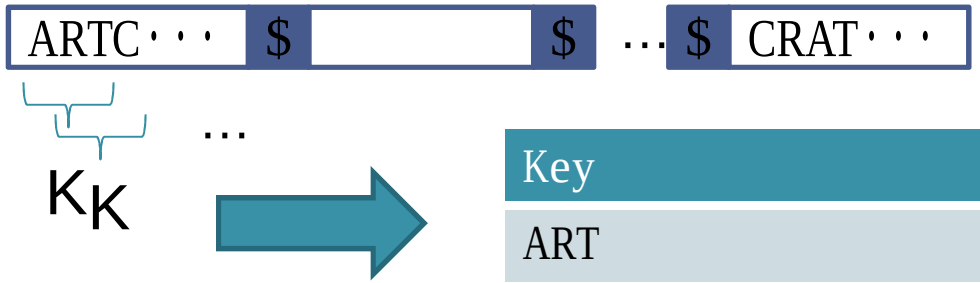

TOKYO TECH Pursuing Excellence

# GHOSTM: flowchart

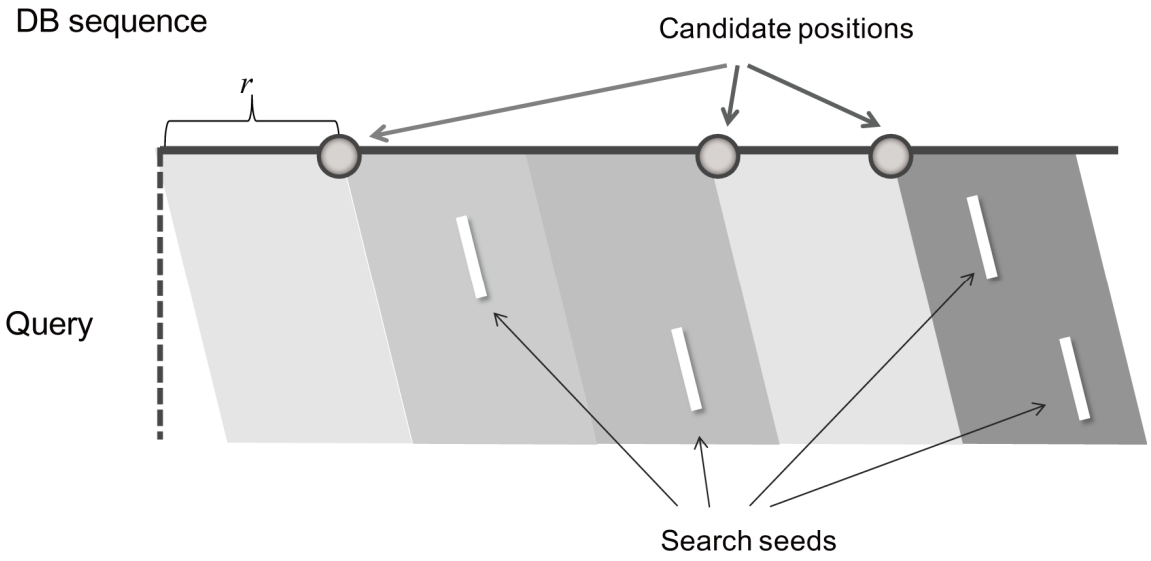


# GHOSTM: searching alignment candidates

- Search *k*-mer matches (seeds) between query and DB sequence

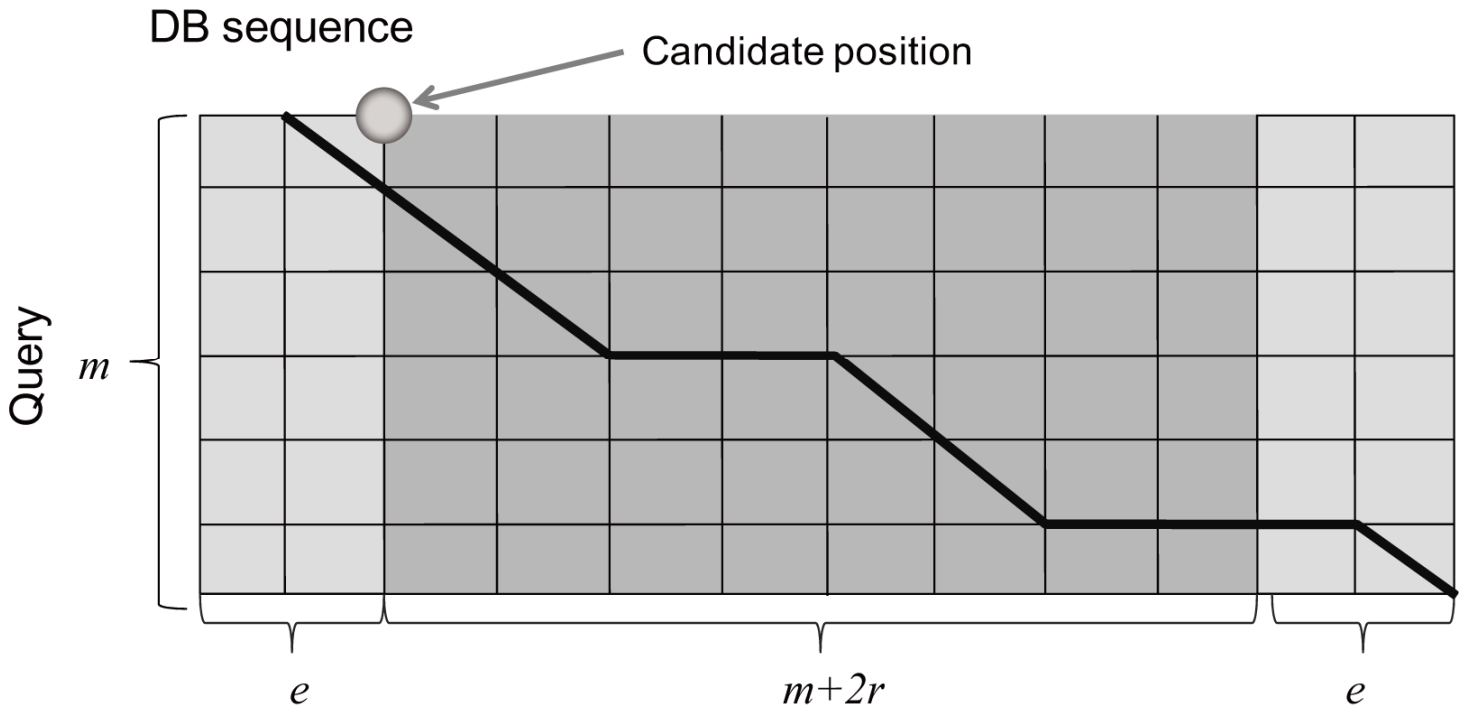



| Key | Positions |
|---|---|
| ART | 0, ・・・・・ |
| RTC | 1, ・・・・・ |

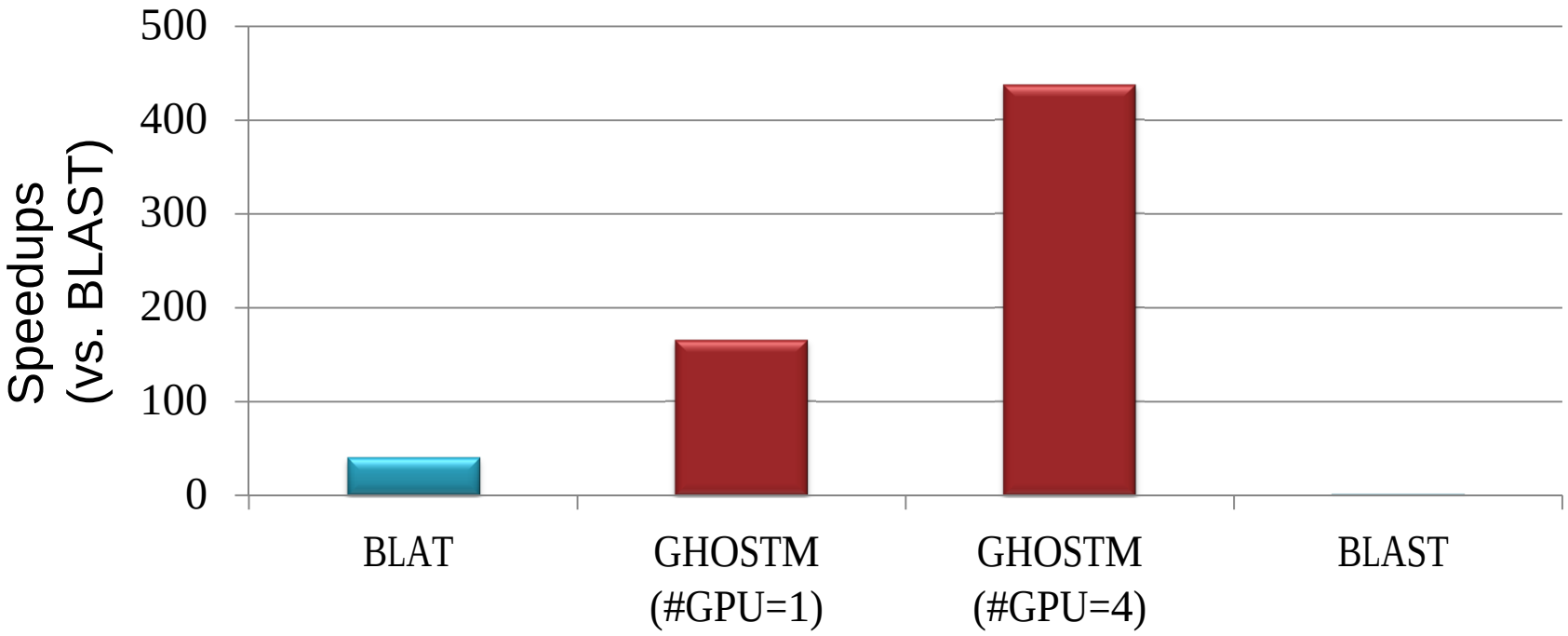


# GHOSTM: local alignment

- Calculate alignment score by dynamic programming (Smith-Waterman algorithm) for each alignment candidates



# GHOSTM: search speed



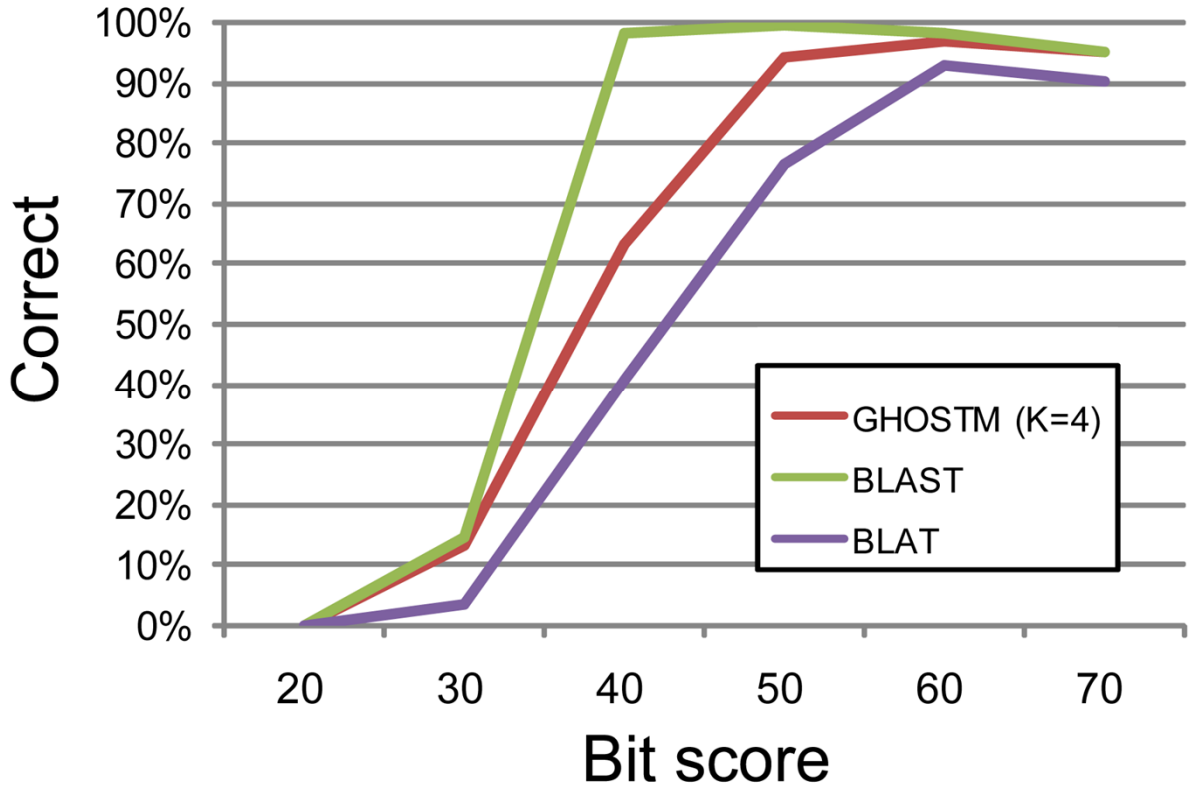

| Program | #GPUs | Time (s) | Speed up |
|---|---|---|---|
| GHOSTM (K = 4) | 1 | 2409 | 165.1 |
| GHOSTM (K = 4) | 4 | 909 | 437.6 |
| BLAT | | 9898 | 40.2 |
| BLAST | | 397798 | 1 |

* GPU: NVIDIA Tesla S1070 (on TSUBAME 1.2)

TOKYO TECH Pursuing Excellence

# GHOSTM: search accuracy

- Correct answer: Smith-waterman local alignment algorithm by using SSEARCH program



# Large-scale metagenome analysis on TSUBAME 2.0

- Metagenome analysis for organisms in soils
  - 2 conditions (polluted soil / control soil)
  - 7 time series data (0, 1, 2, 3, 6, 12, 20 weeks)
  - Sequenced by Illumina/Solexa

Each dataset contains;

Original metagenomic data:

about 20 million DNA reads (75 bp)

Size after excluding low-quality data:

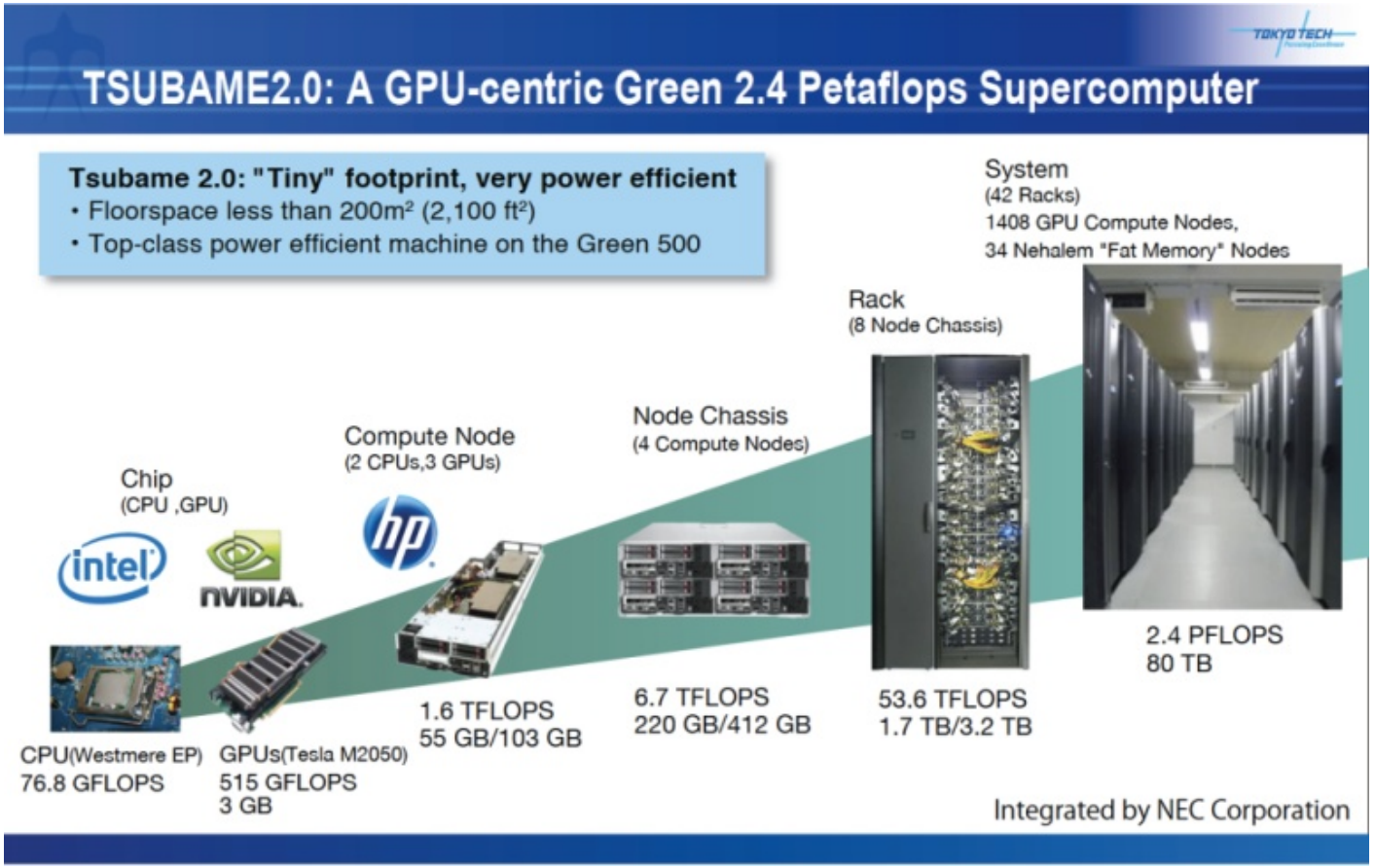
about 7 million DNA reads

Homology search target DB:

NCBI nr amino-acid sequence DB (4.2GB)

TOKYO TECH Pursuing Excellence

# Grand challenge on TSUBAME 2.0




CPU: >17,000 cores (12 cores x 1432 nodes)

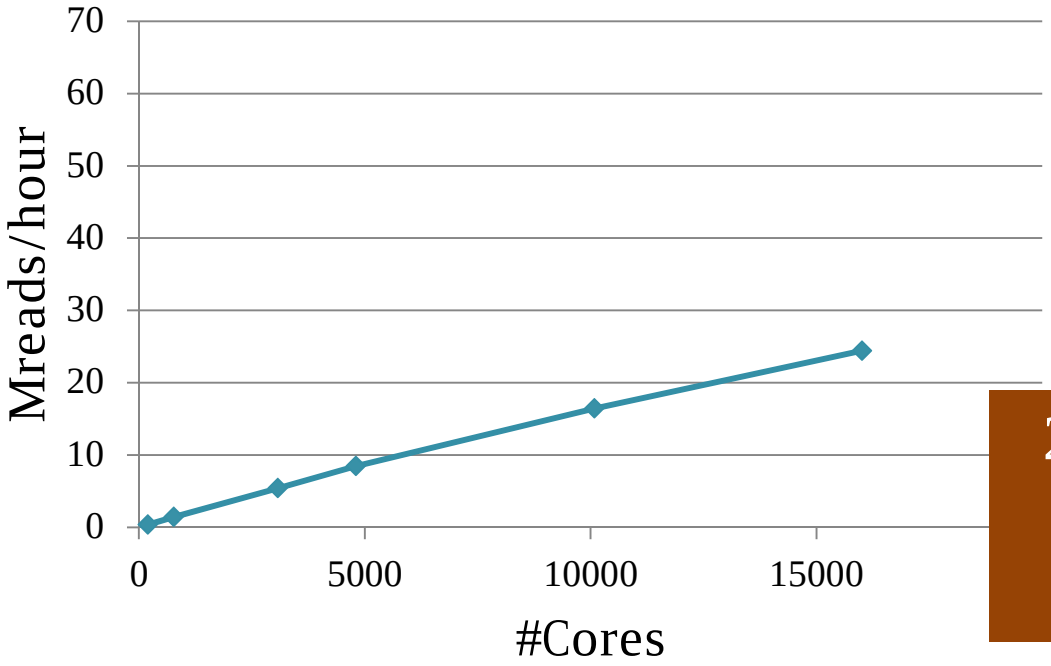
GPU: > 4,000 GPUs (3 GPUs x 1432 nodes)

NVIDIA Tesla M2050

TOKYO TECH Pursuing Excellence

# Performance of the pipeline on TSUBAME 2.0

- BLASTX-based system

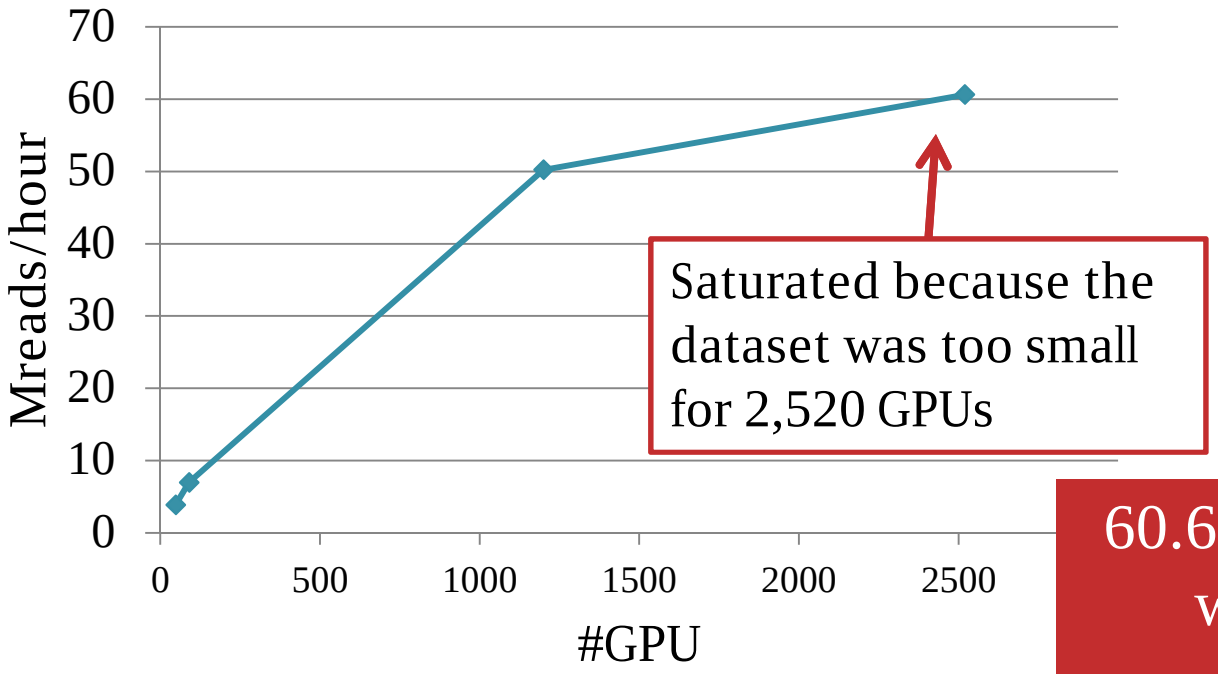



24.4 Million reads/hour
with 16,008 CPU cores
(1,334 nodes)

Achieve to analyze the output of a single run of a next-generation sequencer within 20 minutes.

# Performance of the pipeline on TSUBAME 2.0

- GHOSTM-based system





60.6 million reads/hour with 2,520 GPUs (840 nodes)

Achieve to analyze the output of a single run of a next-generation sequencer within 10 minutes.

# Conclusion

- We developed an automated pipeline system for metagenome analysis on TSUBAME 2.0
  - The system process about 24 million reads per an hour with 16,008 CPU cores and about 60 million reads per an hour with 2,520 GPUs
  - The system can process metagenome data obtained from a single run of a next generation sequencer within a hours
- We developed GPU-based fast homology search tool GHOSTM
  - About 165-times faster than BLAST (1GPU vs. 1CPU)
  - Has enough search sensitivity for metagenome analysis

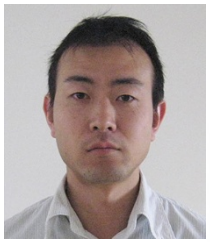
TOKYO TECH Pursuing Excellence

# Acknowledgement


- Mr. Shuji Suzuki
- Dr. Takashi Ishida
- Prof. Fumikazu Konishi
- Prof. Ken Kurokawa

Tokyo Institute of Technology






CUDA COE Program
by NVIDIA

NVIDIA

The Global Scientific Information and Computing Center (GSIC), Tokyo Institute of Technology

GSIC

Tokyo Institute of Technology
Global Scientific Information and Computing Center

TOKYO TECH
Pursuing Excellence