CONTACT NAME
Shuji Suzuki: suzuki@bi.cs.titech.ac.jp



# An ultra-fast computing pipeline for metagenome analysis with GPUs

Shuji Suzuki[1], Takashi Ishida[1] and Yutaka Akiyama[1]
[1] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Japan

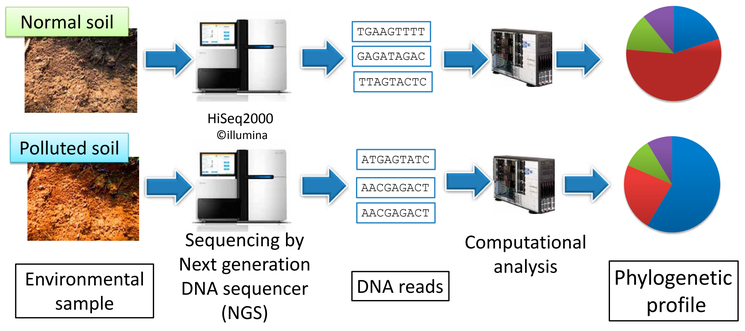
## Abstract

Metagenome analysis is useful for not only understanding symbiotic systems but also watching environment pollutions. However, metagenome analysis requires sensitive sequence homology searches which require large computation time and it is thus a bottleneck in current metagenome analysis based on the data from the latest DNA sequencers generally called a next-generation sequencer. To solve the problem, we developed a large-scale computing pipeline for metagenome analysis on TSUBAME2.

## Introduction

**Metagenome analysis** is the study of the genomes of uncultured microbes obtained directly from microbial communities in their natural habitats such as soils, seas, and human bodies.

For example:
The phylogenetic profile of microbes in a polluted soil is different from that in normal soils. Thus, we can monitor whether the soil is normal or polluted by analyzing the phylogenetic profile of microbes




## Problems :

Metagenome analysis is useful but it requires a huge amount of computational resources

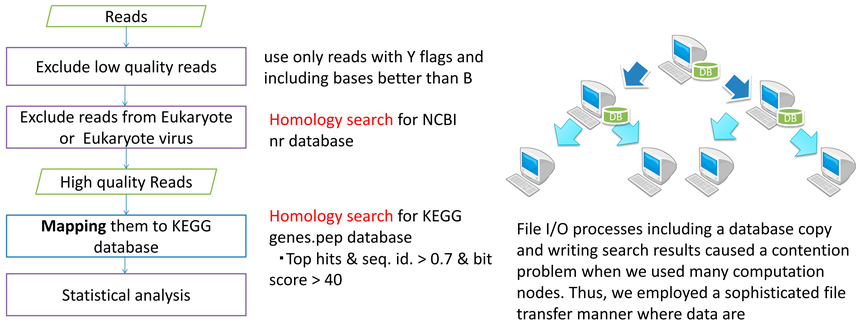
1. Large computation of sensitive homology searches
2. A huge amount of genomic data
   The size of NGS (Hiseq2000) result is about 600 GB / run and about 10,000 time larger than that of the previous generation sequencers.

   → Approximately **25,000 CPU days** are needed for NGS (Hiseq2000) result with NCBI BLASTX[1] program

In this study, we developed a large-scale automated computing pipeline with GPUs for analyzing a huge amount of metagenomic data obtained from a next-generation sequencer. The pipeline enables us to analyze metagenomic data from a next generation sequencer in real time by utilizing huge computational resources on TSUBAME2 which was ranked 5th in Top500 on November, 2011 .

## Automated computing pipeline on TSUBAME2

The pipeline includes twice homology search processes and they are the bottleneck of the pipeline because sensitive homology searches used in metagenome analysis require high computing cost. To solve the problem, the pipeline performs the homology search on many computation nodes.
For homology search, the pipeline can utilize GHOSTM[2], which use GPUs.



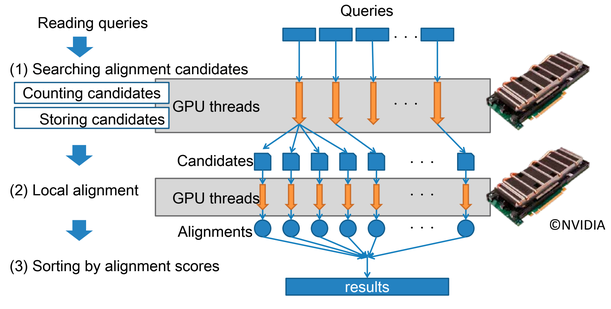

**Overview of our metagenomic analysis**

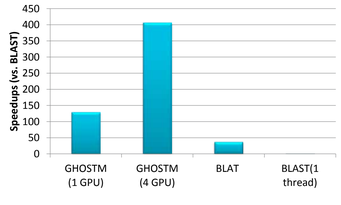



File I/O processes including a database copy and writing search results caused a contention problem when we used many computation nodes. Thus, we employed a sophisticated file transfer manner where data are simultaneously copied from a local disk of a node to another in a binary-tree manner.

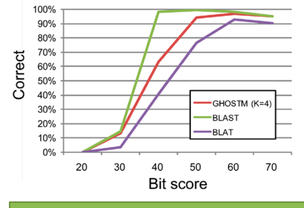
## GHOSTM: A GPU-Accelerated Homology Search Tool for Metagenomics

GHOSTM [2] is a homology search software developed by our group and accelerated by using GPU-computing technique. The algorithm is similar to BLAST but optimized for GPU-calculation. We used NVIDIA CUDA to implement the GPU-calculation.




130 times faster than BLASTX




Enough sensitivity for metagenomic analysis

## Results

We performed a large-scale metagenome analysis by using our pipeline on whole TSUBAME2. We used metagenomic data sampled from polluted soils and obtained by using a next-generation sequencer.

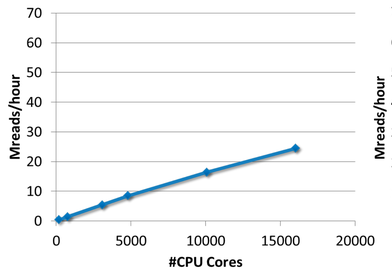
**TSUBAME2**

**#computing nodes:**
1432 nodes

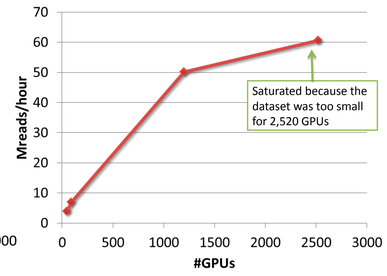
**1 computing node:**
CPU : Intel Xeon 2.93 GHz (6 cores) x 2
MEM: >54GB
GPU : NVIDIA Tesla M2050 x 3

**Data set:**
Queries :
Metagenomic data DNA reads (75 bp, 71 million, excluded low-quality data)

DB :
NCBI nr amino-acid sequence database (4.2GB)




**Speedup of the BLASTX-based system for the number of CPU cores**




**Speedup of the GHOSTM-based system for the number of GPUs**

the pipeline achieved to process

- about 24 million reads per an hour with 16,008 CPU cores (the BLASTX-based system , 1,334 computing nodes)
- about 60 million reads per an hour with 2,520 GPUs (the GHOSTM-based system, 840 computing nodes)

→ The calculation speed of the GHOSTM-based system was about 2.5 times faster than the BLASTX-based system.

## References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *Journal of molecular biology*. 215, 3 (Oct. 1990), 403–10.
[2] Suzuki, S., Ishida, T., Kurokawa, K. and Akiyama, Y. 2012. GHOSTM: A GPU-Accelerated Homology Search Tool for Metagenomics. *PLoS ONE*. 7, 5 (May. 2012), e36060.

## Acknowledgments

This research was supported in part by HPCI Strategic Program Computational Life Science and Application in Drug Discovery and Medical Development by MEXT of Japan, Cancer Research Development funding by National Cancer Center, Japan, and the CUDA COE Program by NVIDIA.