



Qingdao Institute of Bioenergy and Bioprocess Technology
Chinese Academy of Sciences

Parallel-META

A High-Performance Computational Pipeline for Metagenomic Data Analysis

Xiaoquan Su

Qingdao Institute of BioEnergy and Bioprocess Technology

Chinese Academy of Sciences

GTC Asia, Dec 14-15, 2011



I. Introduction to Metagenome and Metagenomic Analysis

II. 16S rRNA Extraction

III. 16S rRNA Mapping

IV. Performance of Parallel-META

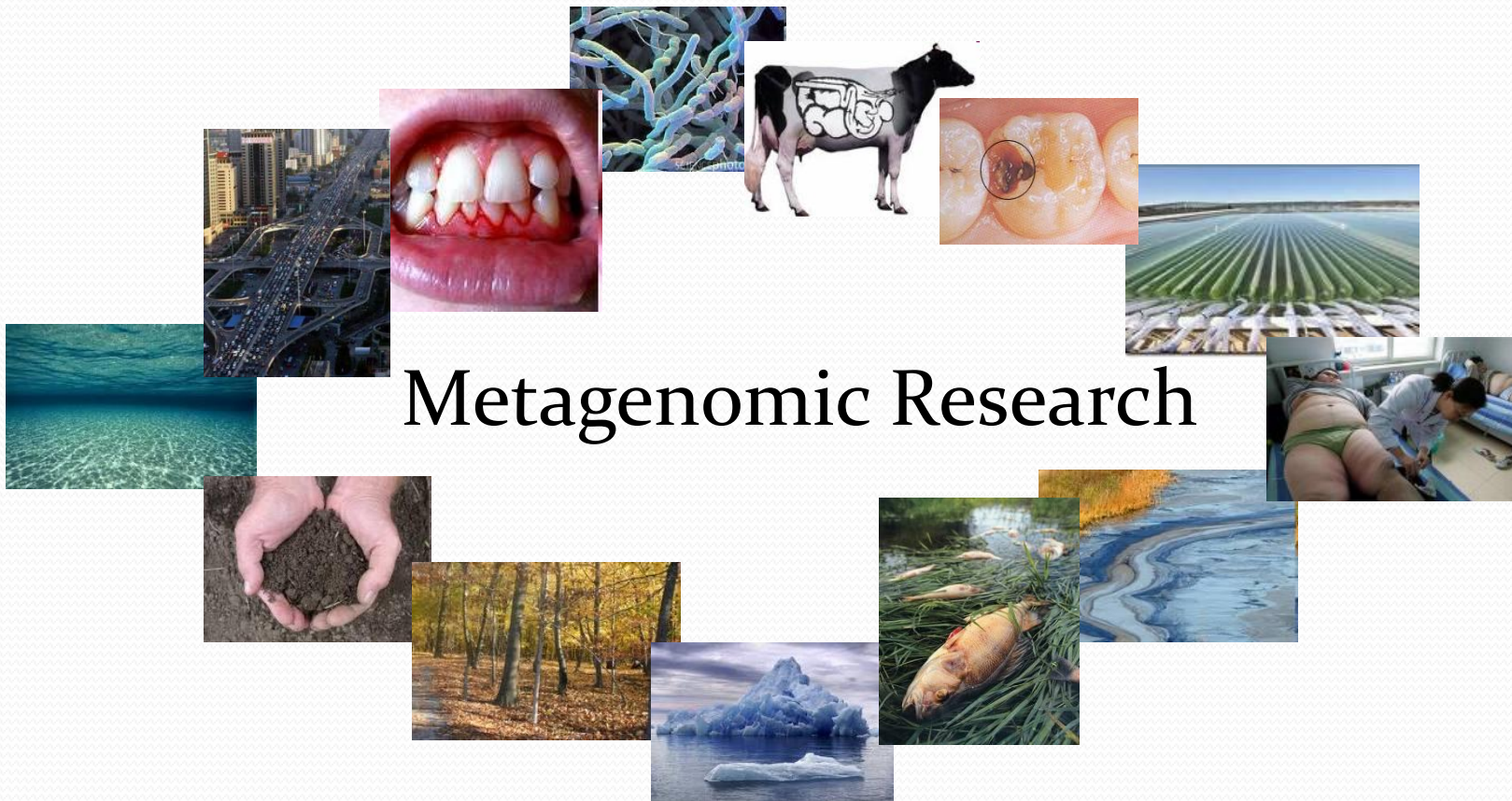
V. Online Service

VI. Further Works



QIBEBT

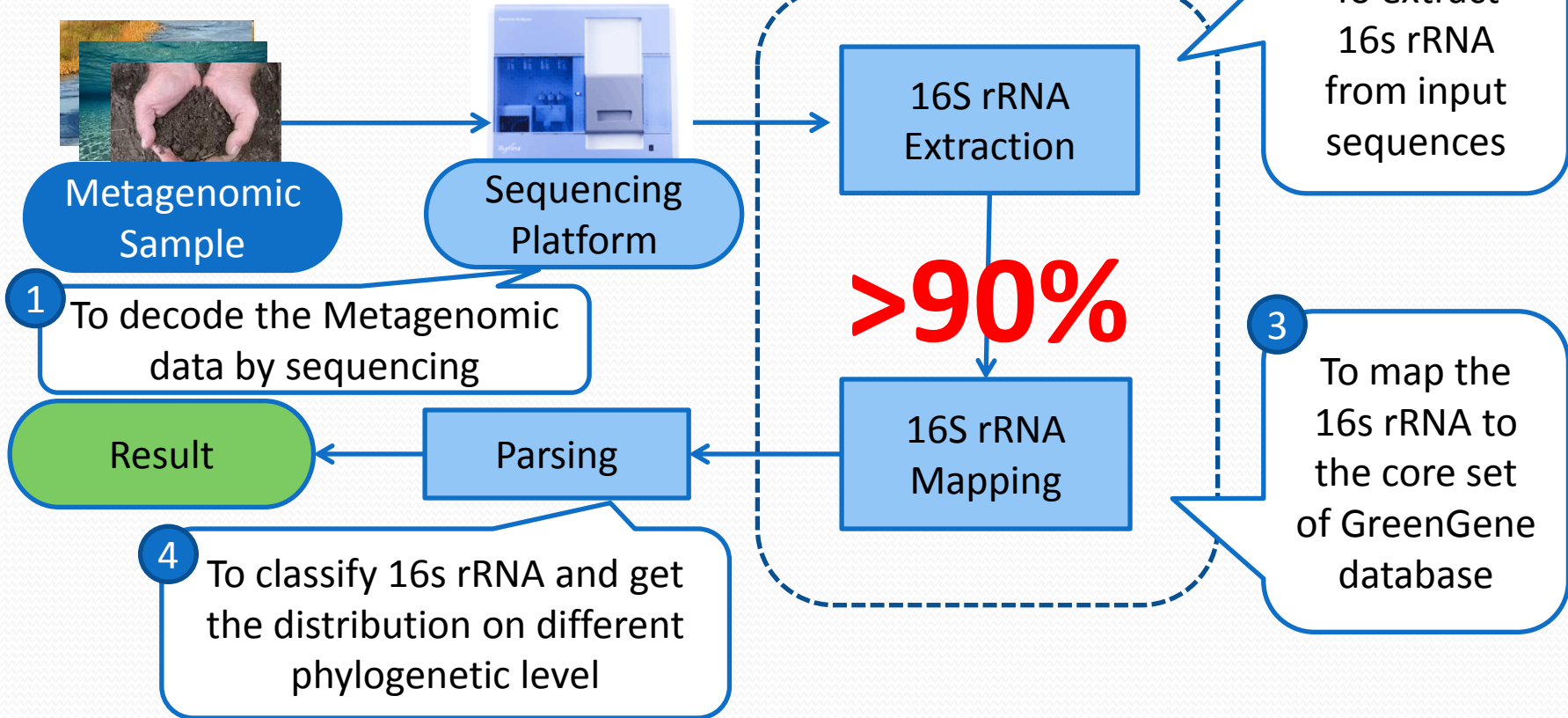
I. Introduction to Metagenome and Metagenomic Analysis



Metagenomic Research



Metagenomic Analysis

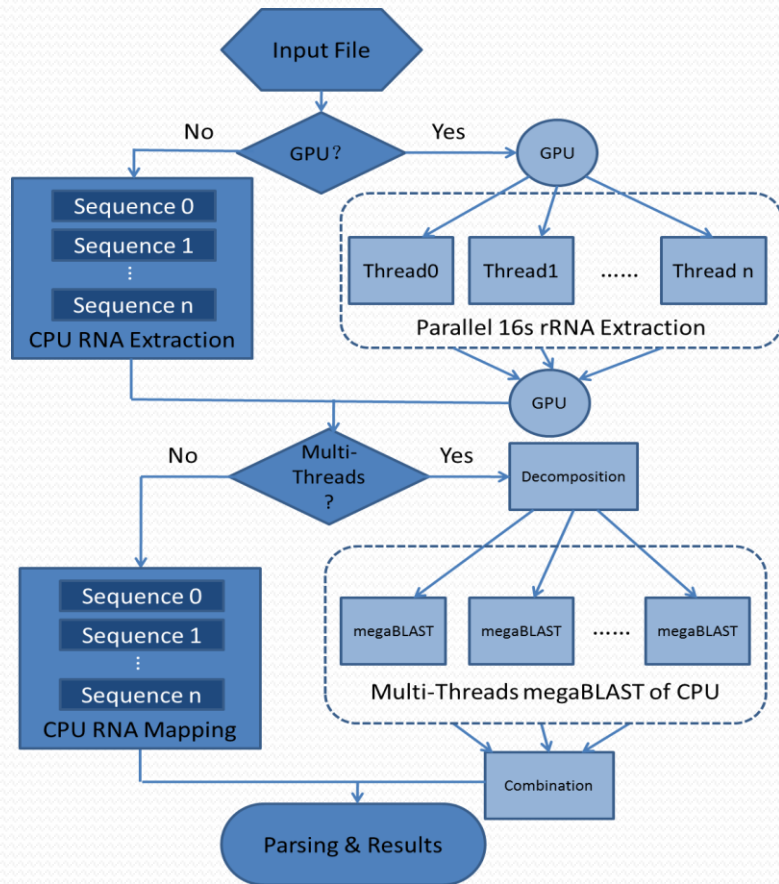




Parallel-META

Using single core of single CPU, scanning each sequence serially

Using single core of single CPU, mapping each sequence serially

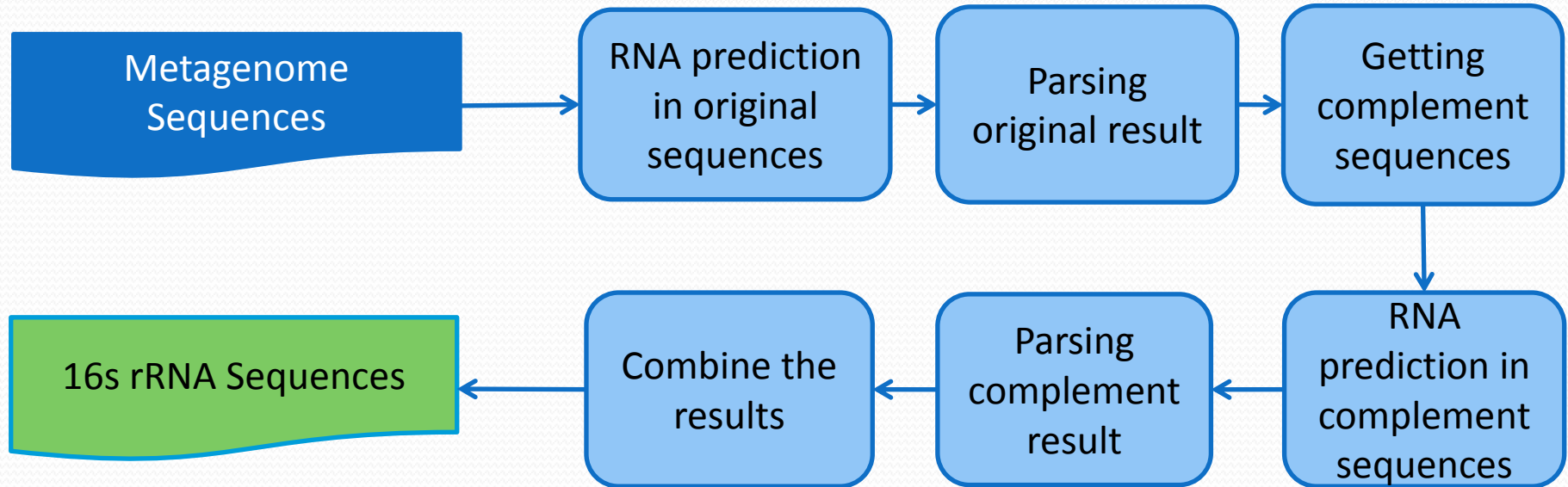


Based on GPGPU and CUDA, distributing the sequences to large number of stream processors of GPU

Based on Multi-core CPU and Multi-thread programming, mapping each sub sequences set by one thread of one core



16S rRNA Extraction





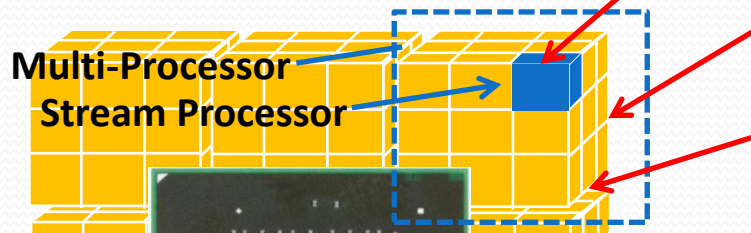
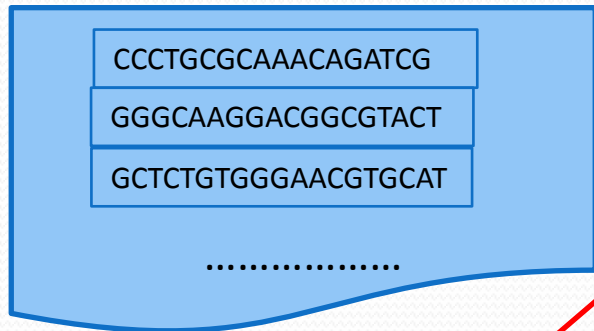
Traditional Method

```
CCCTGCGCAAACAGATCAAAGCACG  
GGGCAAGGACGGCGGTACGGTCAAA  
GCTCTGTGGGCGTTCTGCTCCAAAA  
.....
```

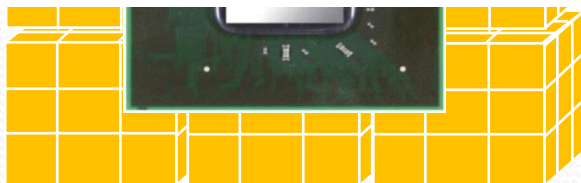
Using **One** Core of CPU
Transform **One** sequence one time



GPU Method



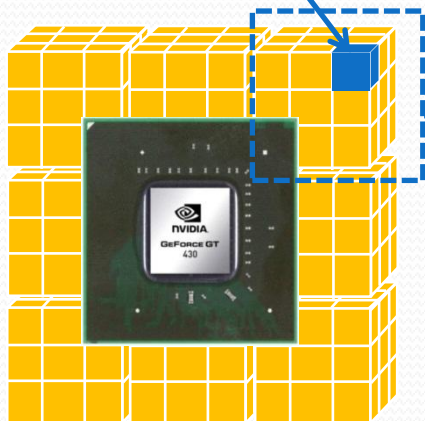
Using **Hundreds** of Processors of GPU





GPU Method

CCCTGCGCAAACAGATCG



```
__shared__ char complement_table[26];  
__shared__ char seq_A[BLOCK_DIM];  
__shared__ char seq_B[BLOCK_DIM];  
seq_A[tidx] = seqs[ offset ]  
__syncthreads();  
seq_B[length - tidx-1]  
=complement_table[seq_A[tidx] ];  
__syncthreads();  
seqs[offset= temp_seq_B[tidx];
```

Getting Complement Sequences

```
char original_seq[];  
char complement_seq[];  
char original_rna[]  
char complement_rna[];  
__device__ substr(original_seq,  
original_rna );  
__device__ substr(complement_seq,  
complement_rna);
```

16S rRNA Extraction

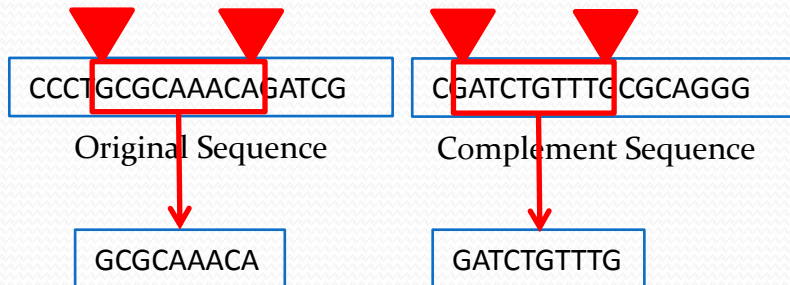
CCCTGCGCAAACAGATCG

Original Sequence



CGATCTGTTTGC GCAGGG

Complement Sequence





- Shared Memory

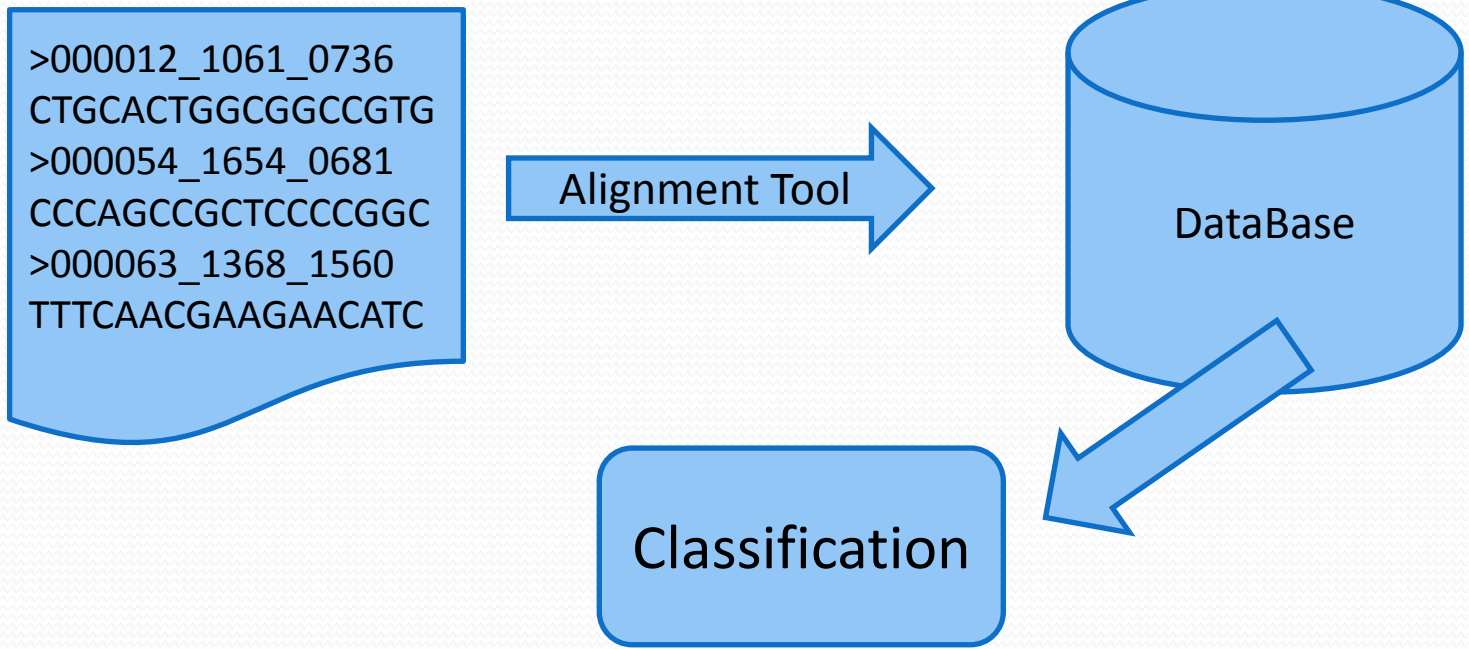
Using shared memory to store the common complement table for each thread to do the transformation

- Global Memory Coalition

Making the start address of each sequence in the Global memory to be a multiple of 256

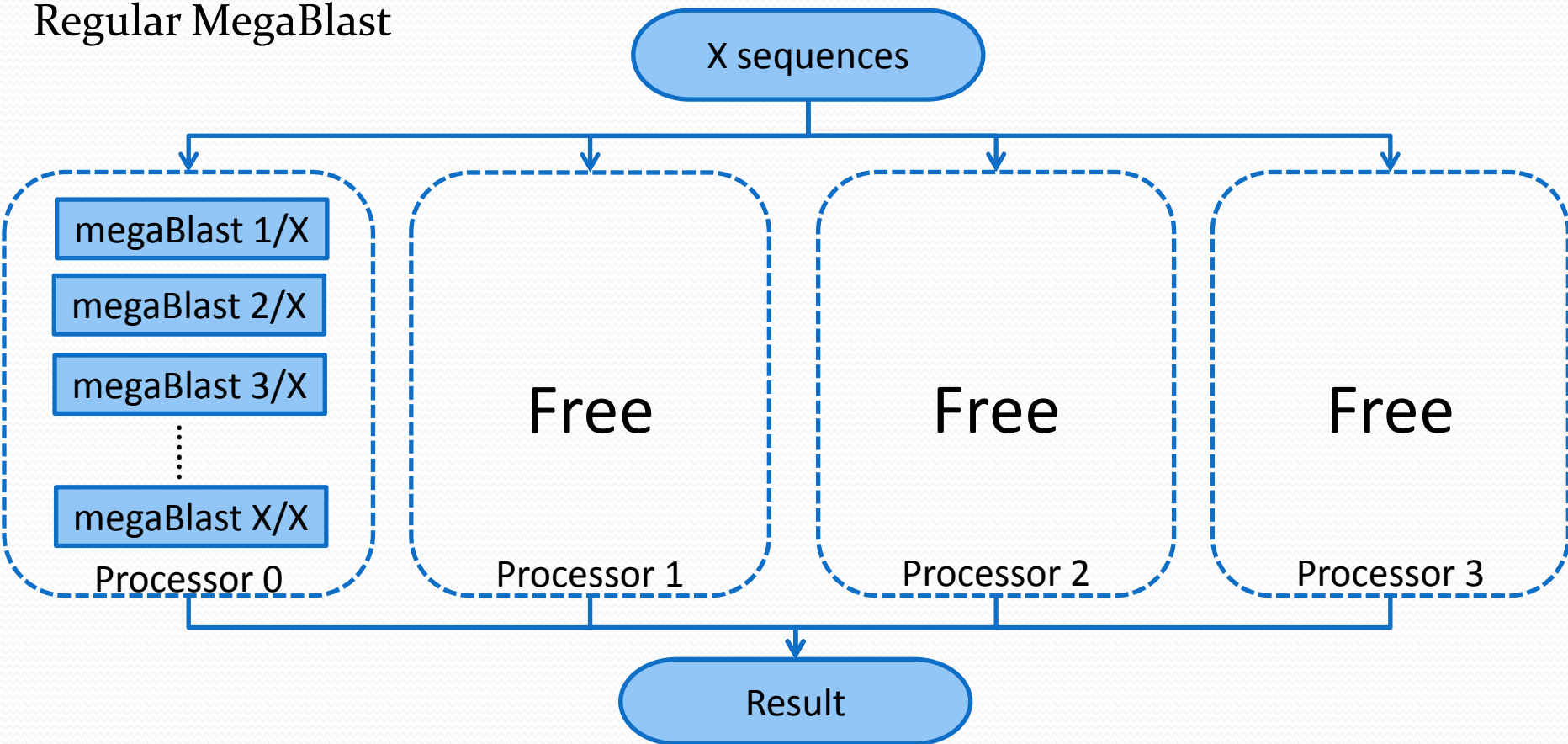


16S rRNA Mapping





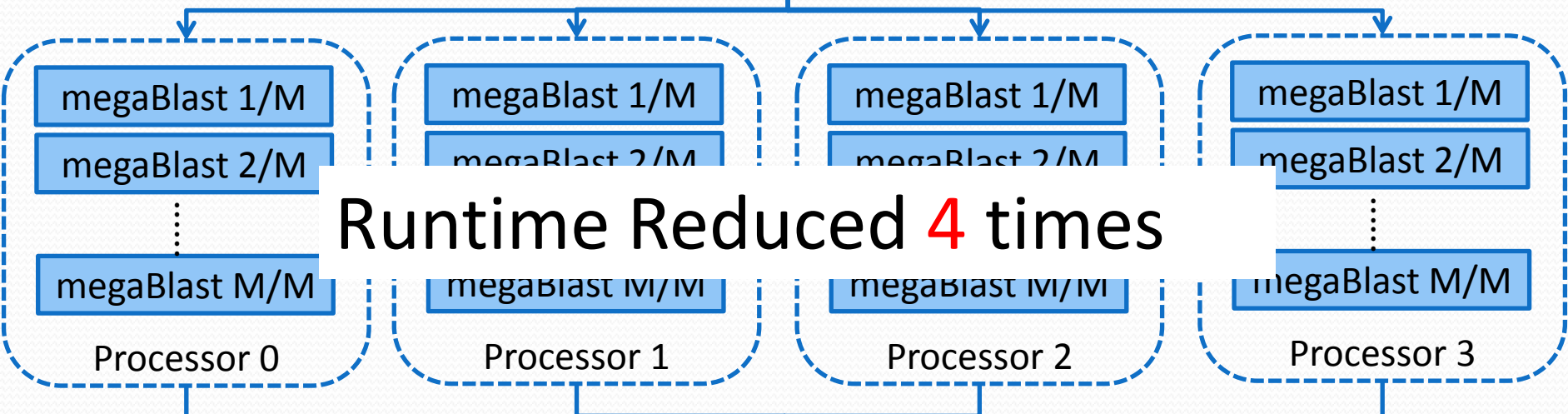
Regular MegaBlast





Parallel-megaBlast

X sequences
 $M = X / 4$



Runtime Reduced 4 times

Synchronization
Result

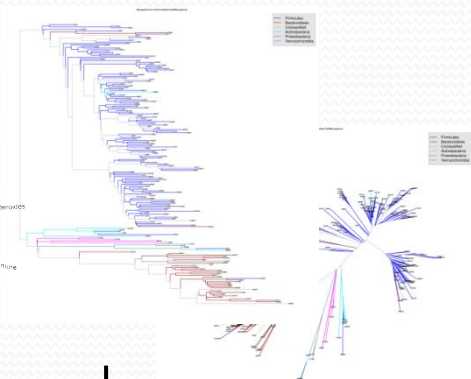
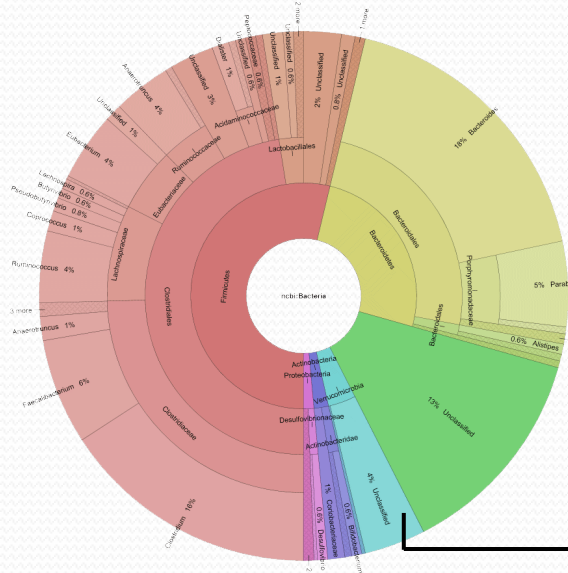


QIBEBT

IV. Performance of Parallel-META



```
>000012_1061_0736
CTGCACTGGCGGCCGTG>
000054_1654_0681
CCCAGCCGCTCCCCGGC>
000063_1368_1560
TTTCAACGAAGAACATC
```



Phylogenetic Tree
Taxonomy Distribution



Parallel-META





- Test Data: 2 health and 2 decayed teeth

	Size(MB)	Sequences	16S rRNA
Input 1	531.86	7,544,950	2406
Input 2	1576.96	17,591,235	2118
Input 3	2775.04	34,405,667	6468
Input 4	2928.64	28,854,628	17119

- Workstation Hardware

CPU	Dual Intel Xeon X5650, 12 core supporting 24 threads
RAM	72 GB DDR3 RDIMM
GPU	nVIDIA Tesla C2070, 448 cores, 6GB DDR5 ECC
DISK	1 TB * 2 RAID 0



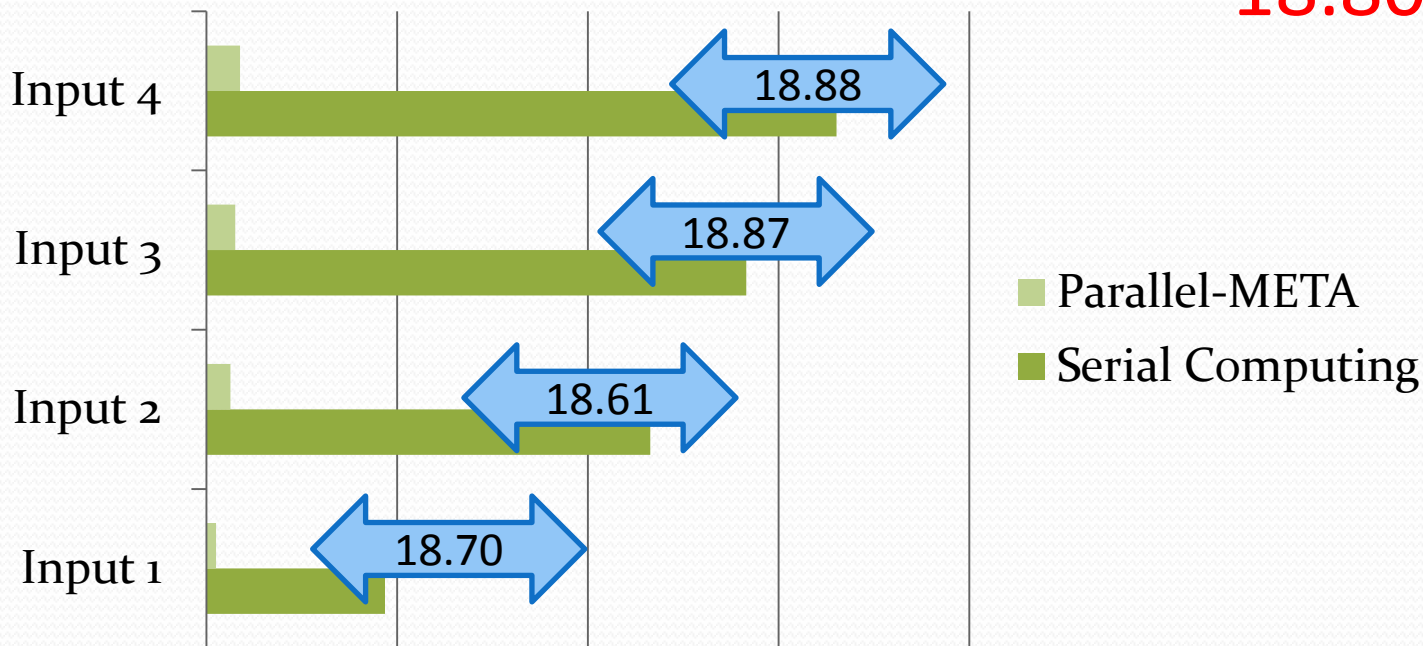
QIBEBT

IV. Performance of Parallel-META

Speed Up

Average Speed Up for RNA Extraction:

18.80





Result



All happy families resemble each other, while unhappy ones each have their own problems.

——Leo Tolstoy

幸福的家庭都是相似的，不幸福的家庭各有各的不幸

——列夫 托尔斯泰

Yang F., et al., Saliva microbiomes distinguish caries-active from healthy human-populations. *ISME J* 2011 (<http://www.nature.com/ismej/index.html>)





QIBEBT

V. Parallel-META online

<http://www.computationalbioenergy.org/parallel-meta.html>



Parallel-META Online 1.0 Beta

Submission A Parallel-META JOB

Parallel-Meta is a GPU/GPU and Multi-Core CPU based software which can parallelly analyse massive metagenomic data. It accepts both metagenomic shotgun data and MS-RNA targeted data, and reports the classification, construction and distribution on different phylogenetic and taxonomy level. The current core version is Parallel-META 1.2-beta.

If you are new to Parallel-META online you might want to start by reading through the [help](#).

Notice:

1. The input file must be in fasta format.
2. The upload file of metagenomic shotgun file is up to 500MB and 16S rRNA targeted file is up to 5MB. Please select the correct type.
3. Please use [help](#) to view the result.

TYPE:

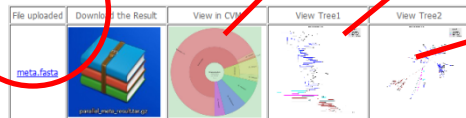
EVALUE:

LENGTH:

1

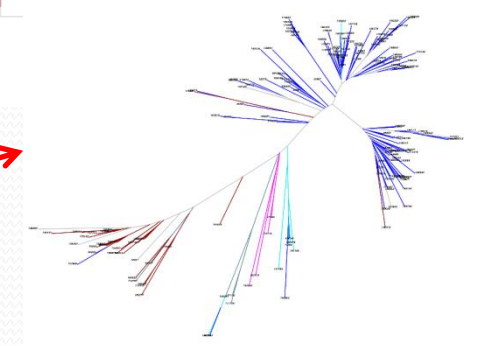
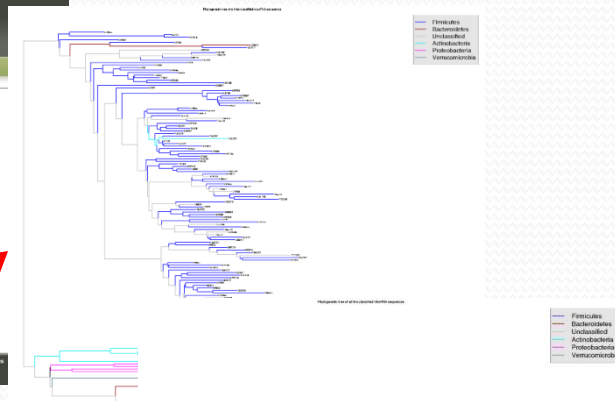
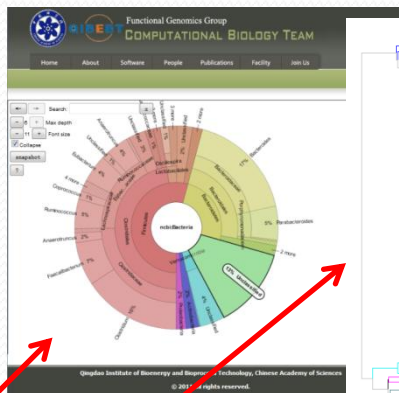
Upload the input data

result



2

Wait a short while for the result



3

View or download the result in any level

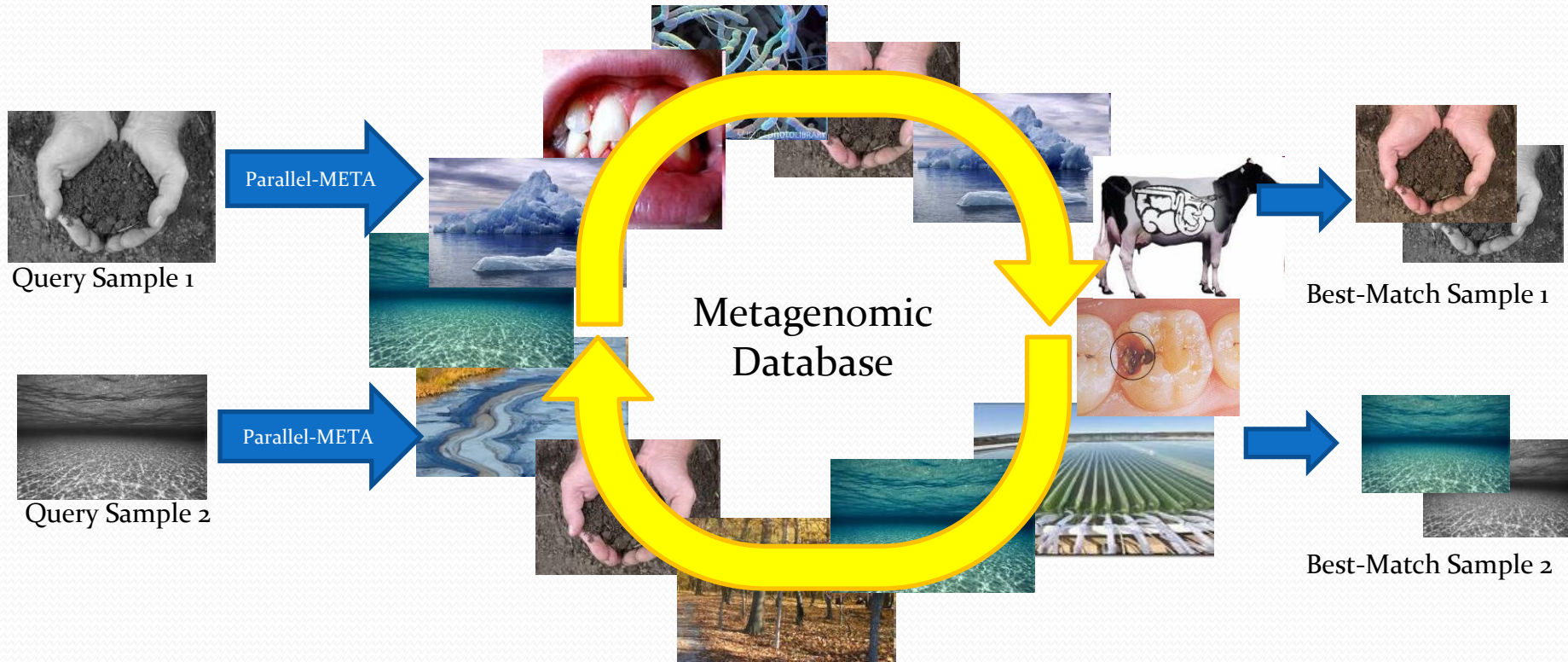


VI. Further Works

- GPU Sequence Alignment Tools
- Metagenomic Database Search Energy based on GPU



Metagenomic Database Search Energy based on GPU





Qingdao Institute of Bioenergy and Bioprocess Technology
Chinese Academy of Sciences

Website:

<http://www.computationalbioenergy.org/>

<http://www.bioenergychina.org/>

<http://www.bioenergychina.org/fg/>

Thanks to nVIDIA



nVIDIA.