An Open-source Collaboration Environment for Metagenomics Research

Xiaoquan Su[1], Yongzheng Ma[2], Hongwei Yang[2], Xingzhi Chang[1], Kai Nan[2], Jian Xu[1]*, Kang Ning[1]*

[1]*Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, Shandong, China*

[2]*Computer Network Information Center, Chinese Academy of Sciences, Beijing, China*

suxq@qibebt.ac.cn, xujian@qibebt.ac.cn, ningkang@qibebt.ac.cn (*corresponding authors)

Abstract

By analyzing metagenomic data from microbial communities, the taxonomical and functional component of hundreds of previously unknown microbial communities have been elucidated in the past few years. However, metagenomic data analyses are both data- and computation-intensive, which require extensive computational power. Most of the current metagenomic data analysis software were designed to be used on a single PC (Personal Computer), which could not match with the fast increasing number of large metagenomic projects' computational requirements. Therefore, advanced computational environment has to be developed to cope with such needs. In this paper, we proposed an open-source collaboration environment for metagenomic data analysis, which enabled the parallel analysis of multiple metagenomic datasets at the same time. By using this collaboration environment, researchers from different locations could submit their data, collaboratively configure the analysis pipeline, and perform data analysis efficiently. As of now, more than 30 metagenomic data analysis projects have already been conducted based on this environment.

Keywords: metagenomics, collaboration environment, data- and computation-intensive computing

I. INTRODUCTION

The total number of microbial cells on earth is huge: approximate estimation of their number is 10^{30}[1], and the genomes of these vastly unknown microbes might contain a large number of novel genes with very important functions. However, more than 99% of microbe species were unknown and un-culturable [2], making traditional isolation and cultivation process non-applicable. Metagenomics refer to the study of genetic materials recovered directly from environmental samples [3], which has made it possible for better understanding of microbial diversity as well as their functions and interactions. The broad applications of metagenomic research, including environmental sciences, bioenergy research and human health, have made it an increasingly popular research area.

The primary goal of metagenomics is the assessment of taxonomic and functional diversity of microbial communities. Metagenomic research was based on sequencing data from 16S rRNA amplicon, or large-scale shot-gun whole-genome metagenomic sequencing. Early 16S rRNA-based metagenomic survey of microbial communities focused on 16S ribosomal RNA sequences which are relatively short, often conserved within a species, and different between species. These surveys have already produced data for analyses of microbial communities of Sargasso Sea [4], acid mine drainage biofilm [5] and human gut microbiome [6].

Facilitated with Next Generation Sequencing (NGS) techniques [7], current metagenomic projects have been advanced rapidly. NGS techniques could produce millions of reads at very fast speed with relatively low price, thus it enables sequencing at much greater depth. Armed with NGS techniques and high performance computational analysis methods, many large-scale metagenomic research projects have been conducted worldwide [8]. NGS techniques also made the large-scale metagenomic research the mainstream in metagenomic research. In this paper,

we are focusing on shot-gun whole-genome metagenomic sequencing, in which computational methods play very important roles.

Based on NGS techniques, metagenomic data analysis is both data- and computing-intensive. Therefore, high-performance computing environment is need for metagenomic data analysis, in which large metagenomic data analysis projects could be conducted concurrently. Additionally, metagenomic data analysis projects were usually highly-collaborative projects, in which different research groups, sometimes from different continents, have to collaborate to complete the whole analysis process. A well-designed working environment should facilitate their collaborations.

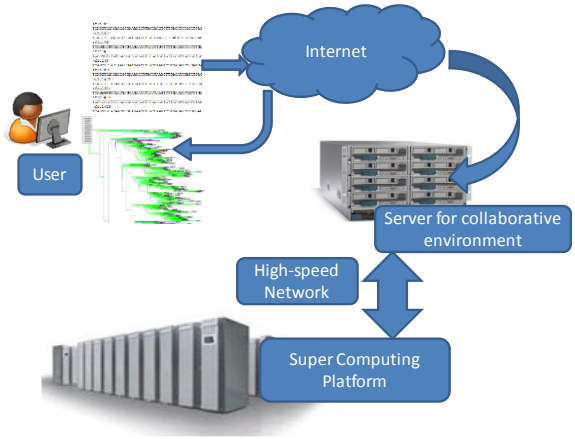


Figure 1. Organization of the whole collaborative system for metagenomic data analysis.

In this work, an open-source collaboration environment for metagenomics research was proposed, which has two key advantages: (1) parallel metagenomic data analysis for multi-users and (2) collaboration environment that enabled researchers to perform data analysis collaboratively and efficiently. The whole system was illustrated in Figure 1. There were two major components of the system: the collaborative platform based on Duckling server, which enabled collaborative research; and super computing platform, which enabled parallel metagenomic data analysis. This work was a collaborative effort between QIBEBT-CAS (Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences) and CNIC-CAS (Computer Network Information Center, CAS).

II. METAGENOMICS AND COLLABORATIVE RESEARCH

Larger databases of reference sequence, such as Greengenes [9], SILVA [10] and RDP [11] already exist for metagenomic sequence analysis. As most of the microbial communities were unknown and metagenomic analyses on them are still ongoing, these databases are also updating frequently. For computational analysis of metagenomic data, the most important tasks include taxonomical and functional analysis. A crucial step in the taxonomical analysis of large-scale metagenomic data is “binning”, in which the metagenomic sequences were assigned to phylogenetic groups according to their taxonomic origins at different resolutions: from “kindom” to “genus” level. The most frequently used metagenomic data binning methods include MEGAN [12], Sort-ITEM [13], TETRA [14], PhyloPhythia [15], and QIIME [16]. Most of these software could be used on a single PC. The web-based metagenomic annotation platforms, such as MG-RAST [17] and CAMERA [18] were also designed to analyze metagenomic data. However, these pipelines and web-servers did not support collaborative research environment very well: the configurations of the data storage scheme and computational process were quite fixed, which made highly-flexible and highly-efficient collaborative research on metagenomic data analysis difficult.

The Duckling system (http://duckling.sourceforge.net/) was a collaboration environment for the rapid developing academic requirement of the Chinese Academy of Sciences. With the goal of building a virtual work environment to enable the high efficient collaboration of scientists in different locations and academic fields, it was an open-source software system developed by the CERC (Collaboration Environment Research Center) of CNIC. Via the Duckling system, all resources, such as hardware, software, data, information and human, could be organized and integrated to form an efficient and easy-to-use system, supporting and advancing

new research activity mode during the era of information. Constructed by the DAIF (Duckling Application Integration Framework) based on JSP, it included these modules:

- UMT (User Management Tool): A virtual organization oriented user management system for creating, editing and deleting users, groups and roles, providing a unified user management solution for the whole laboratory.
- DCT (Document Collaboration Tool): A collaborated editing, publishing and sharing tool based on virtual organization, which can easily and efficiently manage documents.
- CLB (Collaboration Library): A resource library directed by a search tool, to arrange and share all digital resources and data with high flexibility.
- Extendable Components: A flexible component that support various kinds of scientific resources and software as plug-ins.

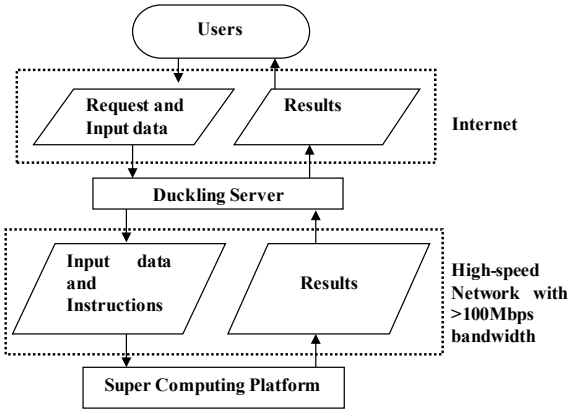


Figure 2. The data flow chart for metagenomic data analysis

The Duckling system has played an important role as an integrated working platform center for users within and outside CAS. In the Duckling virtual working environment, users can not only easily edit, publish and share almost all the digital resources to implement the daily work, but also use various scientific plug-ins to fulfill the requirements in different academic fields.

In the metagenomic data analysis collaboration system, the metagenomic computing component was running as a plug-in of Duckling (server located at CNIC). This plug-in communicated with the computational server (located at QIBEBT) by the interface provided by the metagenomic data analysis component via a high-speed network with >100Mbps bandwidth. When users sent requests and input data to the Duckling server on CNIC side, Duckling would initiate the metagenomics plug-in, and transfer the input data via the high-speed network to the server at QIBEBT. Then the corresponding data analysis pipeline would be executed on QIBEBT computing platform. Once the data analysis has been completed, the results would be transferred to the CNIC server and then to the users (Figure 2).

By combining the metagenomic data analysis pipeline on super computing platform and the collaborative research environment Duckling, the high-performance collaborative metagenomic data analysis environment has been established.

III. SYSTEM ARCHITECTURE AND DESIGN

The collaborative metagenomic data analysis environment was built upon an e-Science framework (Duckling), which enabled the submission, database searching and classification of metagenomic data. The advantage of this environment is that it could "intelligently" distribute the computational resources so that the parallel processing of multiple metagenomic data at the same time was feasible. Moreover, the metagenomic data analysis pipeline in this environment was highly-configurable, so that researchers could collaborate together to work on different stages of large metagenomic project. For this system design purpose, the hardware and software were organized as follow:

The super computing platform of QIBEBT was consisted of Dawning TC2600 Blade Servers, A950 r-F Rack Servers, A620r-FX Rack Servers and Lenovo ThinkStation desktop workstations, with 3.7 Tflops total float computing capability (**Figure 3**). Detailed configuration as below:

- 38 TC 2600 Blade Servers for massive computing, on each there were two 2.1GHz AMD Opteron 2352 x86_64 quad core processors, 16GB DDR3 ECC RAM, 2.55 Tflops computing capability in total.

- Two A950 r-F Rack Server for massive computing, on each there were eight 2.2 GHz AMD Opteron 8354 x86_64 quad core processors, 64GB DDR3 ECC RAM, 0.56 Tflops computing capability in total.
- Six A620 r-FX Rack Server for I/O control and normal computing, on each there were two 2.1GHz AMD Opteron 2352 x86_64 quad core processors, 8GB DDR3 ECC RAM, 0.13 Tflops computing capability in total.
- One Lenovo ThinkStation workstation for special metagenomic software with Dual Intel Xeon X5650 processors 12 cores in total, 72GB DDR3 ECC RAM, one nVIDIA Tesla C2070 GPU card, 0.5 Tflops computing capability in total.

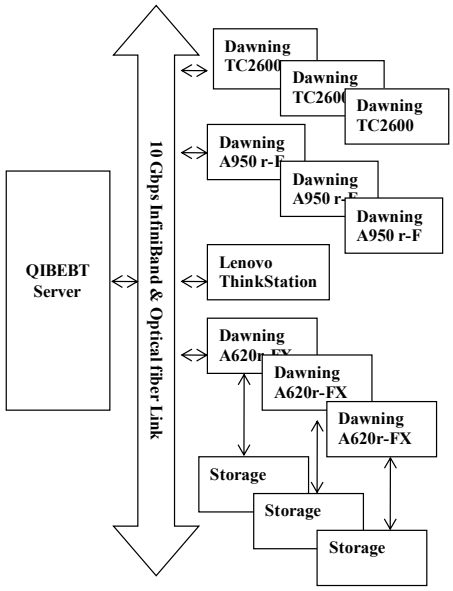


Figure 3. The hardware architecture of the metagenomic data analysis platform

On this super computing platform, we have used a metagenomic data analysis pipeline – Parallel-META [19] as the framework to implement the metagenomic data analysis. Parallel-META was a parallel metagenomic analysis pipeline based on GPGPU and multi-core CPU, developed by QIBEBT. It could extract 16s rRNA gene fragments form metagenomic sequences, report the taxonomic assignment of the identified 16S rRNA fragments, and visualize the taxonomy distribution after alignment. The pipeline included three computation steps:

1) 16S rRNA Extraction based on GPGPU and HMMER [20].

2) Mapping of 16S rRNA to the Greengenes [9] database using Multi-thread megaBLAST [21] based on multi-core CPU.

3) Classification of 16S rRNA fragments based on the results of megaBLAST search.

After these steps, the pipeline reported the classification, length distribution and the summary of the taxonomy assignments of 16S rRNA sequences at different phylogenetic levels. And these results would be transferred to the Duckling server and then to the end-users.

The web-interface for this metagenomic data analysis platform is located at http://159.226.2.248/dct/page/65688.

IV. EXPERIMENTS AND DISCUSSIONS

In this part we tested the performance of collaboration environment for metagenomic data analysis, including data transfer analysis via high-speed network, metagenomic data process analysis on super computing platform and multi-user parallel processing analysis.

A. Experiment data

To analyze data transfer efficiency, computing capability and throughput of the computing platform for metagenomic data analysis, we have selected a series of test data with different sizes and number of users. These datasets included 5 shotgun-sequence datasets and 5 16S rRNA targeted-sequence datasets with increased sequence numbers for the computing capability test (Table 1); 1 shotgun-sequence data and 1 16S rRNA targeted-sequence data to inspect the throughput of parallel processing by different number of users at the same time (data available at ftp://124.16.151.190/). For the shotgun-sequence input, Parallel-META would process the whole pipeline, but for the 16S rRNA targeted-sequence input data, Parallel-META would skip the 16S rRNA sequences, and just process the last 2 steps of the pipeline.

Table 1. Test datasets details, and the average data analysis time for each dataset

Data size (MB)	Sequence Type	Sequence amount	16S rRNA amount
63.491	***Shotgun***	***229875***	***494***
317.455	Shotgun	1194370	2470
634.91	Shotgun	2298750	4940
952.365	Shotgun	3448120	7400
1269.82	Shotgun	4597500	9880
3.474	***Targeted***	***25140***	***25140***
18.735	Targeted	125700	125700
37.472	Targeted	251400	251400
56.207	Targeted	377100	377100
69.481	Targeted	477980	477980

Datasets in bold were used for multi-user test.

B. Data transfer analysis

In order not to conflict with other works, the transfer bandwidth of input data between CNIC server and QIBEBT sever is specially restricted to 20Mb/s. Here we transferred 10 files with different sizes via this high-speed network three times to get the average transfer time and bandwidth. From Figure 4 we could see that as the size of test file increased, the transfer time rose as a liner function of file size.

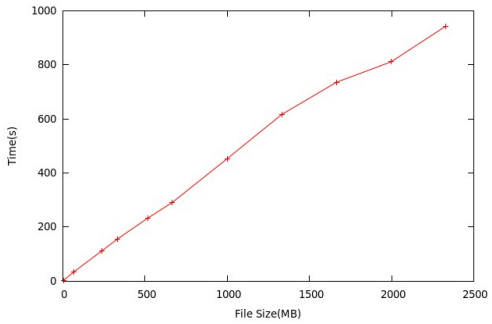


Figure 4. Average transfer time for test files with different size

The bandwidth of the actual data transfer could be represented as the function of file size and transfer time by this formula:

$$\text{Bandwidth} = \text{File Size} / \text{Transfer Time} \quad (1)$$

Figure 5 showed the relation between file size and bandwidth. When the size of the test file was comparative small, the bandwidth was also low due to the high proportion of the network and conveyance latency. As the test file became larger, more and more time was used for data transfer. The proportion of latency decreased, and therefore the utilizing rate was close to the upper-bound of the designed bandwidth of this network. In this test, the peak bandwidth was 19.75 Mbps and the average bandwidth of the network between CNIC and QIBEBT was about 18.2Mbps.

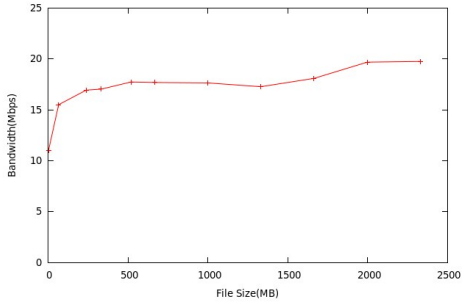
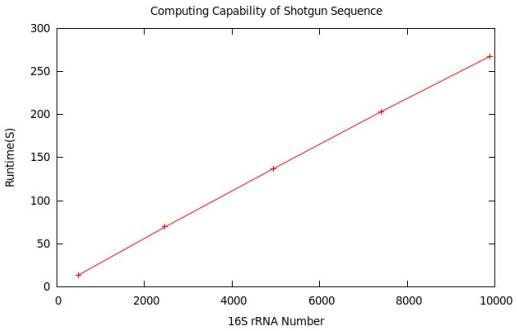


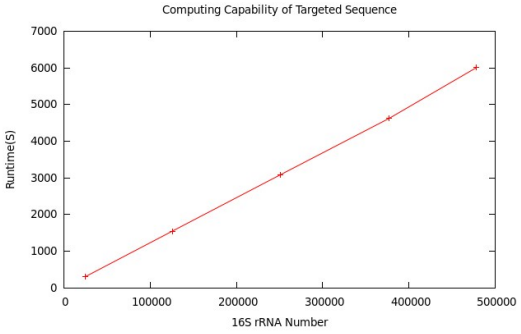
Figure 5. Average bandwidth of the high-speed network between CNIC and QIBEBT

C. Metagenomics data analysis and multi-user parallel processing analysis

In this section we presented the performance of the collaboration environment for metagenomic data analysis from two aspects: the speed of computing for single user, and the throughput of parallel computing for multi-users.



(a)



(b)

Figure 6. Computing capability on (a) shotgun sequence and (b) targeted sequence

1) Computing Capability Test

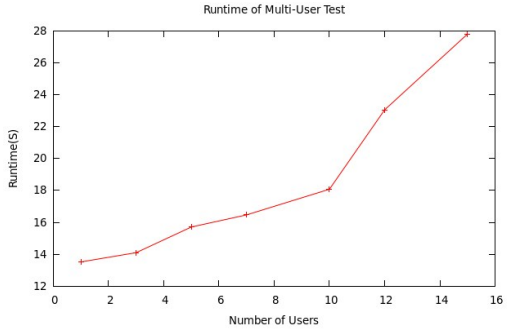
Again, in this section, we computed each test dataset for 3 times and then presented the average results.

From the running time on shotgun and targeted sequences (Figure 6), we could observe that there was almost a liner correlation between the sequence number and the runtime. For shotgun sequences of metagenomic data, the running time was consisted by the runtime of 16S rRNA extraction, 16S rRNA mapping and classification, and for 16S rRNA targeted sequences, the runtime was only the mapping and classification time.

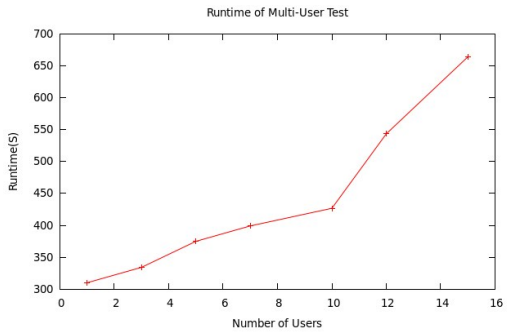
2) Multi-user Throughput Test

In this part, we executed the program with each test datasets for multi-user test (Table 1) by 1, 3, 5, 7, 10, 12 and 15 users at the same time to get the difference between their average runtime.

From Figure 7 (a) we could observe that on shotgun sequences, when the number of users was less than 10, the runtime of the pipeline increased slowly (less than 1.4 times up). However, if more than 12 users executed the program at the same time, the efficiency would be reduced very fast.



(a)



(b)

Figure 7. Run time for multi-users of (a) shotgun sequences and (b) targeted sequences

The rate for increased time used (compared to the single-user case) of targeted sequences (Figure 7 (b)) was quite similar to the result on shotgun sequences for the multi-user test (Figure 7 (a)). For the node of Lenovo Workstation, when the number of users was below or equal to 10, the extra cost time was comparatively low (less than 1.4). However, if more than 12 users ran the software at the same time, for each user the running time would be at least twice as much as single user. That might be due to the fact that the this computing node has 12 cores in total, therefore 12 processes could be activated in the same period, but if there were more than 12 processes to be processed, CPU would use the transition algorithm to manage the processes and activate them serially for the limitation of CPU cores.

D. Theoretical analysis of metagenomic data analysis pipeline's speed

In this collaboration system, for an input data D with N sequences and M MB size, we could predict

the total time cost by such formula:

$$T = T_{transfer} + T_{analysis} * \text{User Rate} \quad (2)$$

In this formula, $T_{transfer} = M * 8 / BW$, here BW was the average bandwidth of the high-speed network, that was 18.2 Mbps. $T_{analysis} = T_{extraction} + T_{mapping} + T_{classification}$. Here $T_{classificaton}$ was very low(less than 1/100 of total analysis time), therefore it could be considered negligible. If the input data was 16S rRNA targeted sequences, then the $T_{extraction} = 0$, else the $T_{extraction} = 3.15E\text{-}05 * N$. After the extraction, suppose the count of 16S rRNA sequences was E(N), the mapping time $T_{mapping} = 0.0125 * E(N)$. The "User rate" referred to the number of users, meaning that the more users doing the computation on the same node at the same time, the higher the "User rate" would be. Therefore, the total time for metagenomics data analysis pipeline could be formulated as:

$$T=M*8/BW + (3.15E\text{-}05 *N + 0.0125 * E(N))*\text{User Rate} \quad (3)$$

V. CONCLUSION AND FUTURE WORKS

Traditional metagenomic data analyses were based on single PC, which also need time-consuming communications between collaborators. In this work, we have proposed a novel e-Science environment (based on Duckling system) that could not only facilitate collaborations, but also enable the parallel processing of the metagenomic data.

The novel collaborative environment proposed in this work targeted the data- and computation-intensive problem of metagenomic data analysis. There are some web-based metagenomic annotation platforms, such as MG-RAST [17] and CAMERA [18] designed to analyze metagenomic data for multi-users too. However, these pipelines and web-servers did not support collaborative research environment very well, as previously stated. This work represented a new e-Science oriented approach, which provided very extendable collaborations environment that could not only serve for current metagenomic research needs, but also future needs with massive data and more users.

Currently, more than 30 metagenomic data analysis projects have been conducted based on this collaboration environment. These projects include soil microbial community analysis [22], oral disease-causing microbial community analysis [23] and rumen lignocellulose degradation microbial community analysis [24].

Current collaboration environment still needs improvements. One of the limitations is the centralized computer server, based on which only up to 10 metagenomic data analysis projects would be conducted efficiently. When the concurrent number of metagenomic data analysis request reaches to several tens or up to hundreds, a significant amount of delay in time would be incurred. For such a lot of concurrent requests, scalable computational resources, such as the de-centralized computer clusters, are needed.

GPU computing for bioinformatics data analysis is becoming more and more popular due to its significant speed boost compared to current methods. Due to the scalable hardware architecture of our metagenomic data analysis environment, GPU computing hardware could be easily installed onto current system. Additionally, current metagenomic data analysis pipeline could also be parallelized by GPU CUDA programming. These GPU hardware and software could be implemented into the next-generation collaboration environment for metagenomic data analysis, so that much more metagenomic data analysis projects could be conducted in parallel [19].

Compliment to the collaborative data analysis environment is the collaborative database management system. As one of the most important roles of collaboration environment is the sharing of software as well as data, it would be beneficial to also have a database system for metagenomics scientists to share the original data as well as the results. Such collaborative database management system would also be incorporated into the next-generation collaboration environment for metagenomic data analysis.

ACKNOWLWDGMENTS

This research is supported in part by Ministry of Science and Technology's high-tech (863) grant 2009AA02Z310 and Chinese Academy of Sciences' e-Science grant INFO-115-D01-Z006.

REFERENCES

[1] G. N. Proctor, "Mathematics of microbial plasmid instability and subsequent differential growth of plasmid-free and plasmid-containing cells, relevant to the analysis of experimental colony number data," *Plasmid,* vol. 32, pp. 101-30, Sep 1994.

[2] A. Jurkowski, *et al.*, "Metagenomics: a call for bringing a new science into the classroom (while it's still new)," *CBE Life Sci Educ,* vol. 6, pp. 260-5, Winter 2007.

[3] J. A. Eisen, "Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes," *PLoS Biol,* vol. 5, p. e82, Mar 2007.

[4] J. C. Venter, *et al.*, "Environmental genome shotgun sequencing of the Sargasso Sea," *Science,* vol. 304, pp. 66-74, Apr 2 2004.

[5] G. W. Tyson, *et al.*, "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature,* vol. 428, pp. 37-43, Mar 4 2004.

[6] M. Arumugam, *et al.*, "Enterotypes of the human gut microbiome," *Nature,* vol. 473, pp. 174-80, May 12 2011.

[7] E. R. Mardis, "Anticipating the 1,000 dollar genome," *Genome Biol,* vol. 7, p. 112, 2006.

[8] J. Xu, "Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances," *Mol Ecol,* vol. 15, pp. 1713-31, Jun 2006.

[9] T. Z. DeSantis, *et al.*, "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB," *Appl Environ Microbiol,* vol. 72, pp. 5069-72, Jul 2006.

[10] E. Pruesse, *et al.*, "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," *Nucleic Acids Res,* vol. 35, pp. 7188-96, 2007.

[11] J. R. Cole, *et al.*, "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis," *Nucleic Acids Res,* vol. 37, pp. D141-5, Jan 2009.

[12] D. H. Huson, *et al.*, "MEGAN analysis of metagenomic data," *Genome Res,* vol. 17, pp. 377-86, Mar 2007.

[13] M. Monzoorul Haque, *et al.*, "SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences," *Bioinformatics,* vol. 25, pp. 1722-30, Jul 15 2009.

[14] H. Teeling, *et al.*, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics,* vol. 5, p. 163, Oct 26 2004.

[15] A. C. McHardy, *et al.*, "Accurate phylogenetic classification of variable-length DNA fragments," *Nat Methods,* vol. 4, pp. 63-72, Jan 2007.

[16] J. G. Caporaso, *et al.*, "QIIME allows analysis of high-throughput community sequencing data," *Nat Methods,* vol. 7, pp. 335-6, May 2010.

[17] E. M. Glass, *et al.*, "Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes," *Cold Spring Harb Protoc,* vol. 2010, p. pdb prot5368, Jan 2010.

[18] R. Seshadri, *et al.*, "CAMERA: a community resource for metagenomics," *PLoS Biol,* vol. 5, p. e75, Mar 2007.

[19] X. Su, *et al.*, "Parallel-META: A High-Performance Computational Pipeline for Metagenomic Data Analysis," *ISB2011,*2011.

[20] Y. Huang, *et al.*, "Identification of ribosomal RNA genes in metagenomic fragments," *Bioinformatics,* vol. 25, pp. 1338-40, May 15 2009.

[21] A. Morgulis, *et al.*, "Database indexing for production MegaBLAST searches," *Bioinformatics,* vol. 24, pp. 1757-64, Aug 15 2008.

[22] W. Xia, *et al.*, "Autotrophic growth of nitrifying community in an agricultural soil," *ISME J,* Feb 17 2011.

[23] F. Yang, *et al.*, "Saliva microbiomes distinguish caries-active from healthy human-populations," *ISME Journal,* vol. Accepted, 2011.

[24] X. Dong, *Personal communications,* 2011.