Large Memory High Performance Computing Enables Comparison Across Human Gut Microbiome of Patients with Autoimmune Diseases and Healthy Subjects

Sitao Wu[1], Weizhong Li[1], Larry Smarr[2, 3], Karen Nelson[4], Shibu Yooseph[5], Manolito Torralba[4]
University of California San Diego, 9500 Gilman Drive, La Jolla, CA, USA: [1]Center for Research for Biological Systems, [2]California Institute for Telecommunication and Information Technology, [3]Department of Computer Science and Engineering; [4]J. Craig Venter Institute, Rockville, MD, USA; [5]J. Craig Venter Institute, San Diego, CA, USA
(SW) siw006@ucsd.edu, (WL) liwz@sdsc.edu, (LS) lsmarr@ucsd.edu, (KN) Kenelson@jcvi.org, (SY) SYooseph@jcvi.org, (MT) MTorralba@jcvi.org

ABSTRACT

Microbial communities that live on the outside and inside of the human body dramatically influence human health and diseases. In recent years, major progress has been made in understanding the human microbiome communities through projects such as the Human Microbiome Project (http://commonfund.nih.gov/hmp/), using next generation sequencing technologies and metagenomic approaches. In this paper, we describe a comparative computational analysis of 183 human gut microbiome sequence datasets, drawn from healthy individuals as well as those with autoimmune diseases. About 2.4 TB of Illumina deep sequencing metagenomic data were analyzed using computational workflows we developed, which run multiple steps of data- and computing-intensive analyses such as mapping, sequence assembly, gene identification, clustering and functional annotations. The analyses were carried out on the Gordon supercomputer at the San Diego Supercomputer Center (SDSC), using ~180,000 core hours and tens of TB storage space. Our analysis reveals the detailed microbial composition, dynamics, and functional profiles of the samples and provides new insight into how to correlate microbial profiles with human health and disease states.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences – *biology and genetics, health.*

General Terms

Algorithms, Performance, Design

Keywords

Human Microbiome Project, HMP, Next Generation Sequencing, NGS, Metagenomics, Bioinformatics, Computational Biology, Human Gut, XSEDE, High Performance Computing, Parallelization

XSEDE '13, Jul 22-25 2013, San Diego, CA, USA
ACM 978-1-4503-2170-9/13/07.
http://dx.doi.org/10.1145/2484762.2484828

1. INTRODUCTION

The microbes that live in and on the human body outnumber the human cells by 10-fold. The collective human microbial communities, known as the human microbiome, play a profound role in human health and disease. Although a few disease-related microbial species have been extensively studied using culturing techniques, most species of human gut microbes cannot be cultured and have remained unknown. With the orders-of-magnitude reduction in sequencing cost since the Human Genome Program, a revolution in understanding the ecological structure of the microbiome using genomic techniques has become possible through a new approach termed metagenomics [1].

The genomic study of the diverse microbial ecology of the human gut took off with a study [2] using 13,355 prokaryotic ribosomal RNA gene sequences from the human gut. This was followed by the earliest metagenomic study [3] of human gut microbiome, which used traditional Sanger sequencing technologies. This landmark study generated about 80 megabases (MB) of DNA sequence data. In the last few years, the transition to Next Generation Sequencing (NGS) technologies [4] has resulted in dramatic advances, promoting a large wave of new studies in human microbiome [5-8]. Especially, projects such as MetaHIT [6] and Human Microbiome Project (HMP) [7, 8], which utilized the Illumina sequencing platforms, generated many terabases of metagenomic sequences, four orders of magnitude larger than the earliest gut microbiome study [3].

The vast amount of metagenomic data produced by the NGS platforms provide a much deeper, wider and more comprehensive view of the human microbiome. However, this data-intensive approach raises corresponding challenges for researchers in data analysis. In this study, we explore the use of large memory supercomputers to accelerate scientific discoveries from comparative gut microbiome analysis across healthy and diseased human subjects.

The large number of sequence reads (10^8) or sequence database file sizes (100s of GB) in microbiome projects means that computing and storage has become a bottleneck. Beyond the computing bottleneck, the complexity of analysis is another major challenge. NGS microbiome data analysis is a complex process, including many computational procedures (e.g. sequence quality control, filtering, mapping and assembly) to analyze and interpret the data at several different levels such as phylogenetic, gene, function and pathway levels. This requires many different computational tools, each with different computing requirements and running patterns, to be integrated using workflows creating highly flexible and scalable computing platforms.

To overcome these challenges, we developed several robust computational workflows specific to human microbiome metagenomic data deployed on SDSC's large memory Gordon supercomputer. We describe the results of applying our computational workflows to study 183 human gut microbiome sequence datasets from healthy individuals and those with diseases. About 2.4 TB of Illumina deep sequencing metagenomics data were processed.

Our computational analysis has led to a series of discoveries about the microbial composition, dynamics, and functional profiles of the samples and produced new insights into correlating microbial profiles and human diseases. In this paper we will describe the computational analysis and exhibit examples of the microbiome database which results, continuing with biomedical analysis of the large microbial database in further publications.

2. Materials and Data

2.1 Gut Microbiome Metagenomic Data

Our goal was to develop a time series of an individual's (Larry Smarr-"LS") gut microbiome and to compare the dynamics against a broad group of healthy individuals and a selection of patients with inflammatory bowel disease (IBD). IBD is an autoimmune disease, which affects over one million Americans. This study was motivated by LS's detailed biomarker time series study, derived from blood and stool specimens. These biomarkers led to LS being diagnosed with IBD [9]. Here we seek to add the dynamics of LS's gut microbiome. To do this, we analyzed the metagenomics of LS's gut microbiome derived from 3 stool samples (LS001, LS002 and LS003) obtained at three different time points (4 months apart). From the stool samples provided by the subject (LS), the J. Craig Venter Institute extracted and sequenced the microbial DNA using NGS Illumina technology.

Forming a control group were 154 samples from healthy (HE) individuals. We then compared these healthy gut microbial ecologies with those from LS and from patients with IBD, dividing the IBD patients into the two IBD subtypes: Crohn's Disease (CD) and Ulcerative Colitis (UC). We used 15 samples from 5 patients with CD (3 samples per patient) and another 11 samples in the group from patients with UC. All the CD, UC and HE samples had been sequenced using Illumina technology. The HE and IBD raw sequence reads are from the Human Microbiome Project [6] and were downloaded from National Center for Biotechnology Information (NCBI) Sequence Read Achieves (SRA) under BioProject IDs 46321, 46881 and 43021. All the raw reads are pair-ended (PE) and about 100 base pair in length. The total size of these raw reads is about 2.4 TB. The list of samples is summarized in **Table 1**.

For more detailed analyses, we removed 5 samples from the UC group, each of which had less than 2 million reads after filtering out low quality reads. We also down-selected the 154 HE samples to 35 samples that spanned the variation in phyla abundance that we observed in our preliminary analysis of the full samples. We then found that one HE sample has over 95% proteobacteria. This sample may be from an unhealthy subject or it might have been contaminated. We therefore removed it from the calculation. This yielded a high-resolution group of 58 samples (**Table 1**). **Table 1** also lists the number of high-quality reads for these selected samples after filtering out reads with low quality scores, reads from human and the artifically duplicated reads using our analysis workflows (**Figure 1**).

2.2 Reference Genomes

A comprehensive microbial reference genome database is essential for analyzing microbiome data. We compiled a reference database from several resources available as of September 17 2012: NCBI's complete bacteria and archaea genomes (2036 genomes), NCBI's complete virus genomes (1397 genomes), NCBI's complete fungi genomes (39 genomes), NCBI's draft bacteria and archaea genomes (1826 genomes) and HMP eukaryote reference genomes (309 genomes). These 5607 genomes, which have ~15 GB of sequences, are used as reference for aligning the reads and then calculating the taxonomy profiles. The computational details are described in following sections.

3. Implementation

3.1 Computing Resources

We used SDSC's Gordon for all the computational data analysis because it is a dedicated data-intensive supercomputer sponsored by the National Science Foundation's (NSF) XSEDE program, well matched to the large data requirements of all software tools in our workflows. Gordon has 1024 compute nodes and 64 I/O nodes. Each compute node contains two 8-core Intel Sandy Bridge processors and 64 GB of DDR3–1333 RAM. Each I/O node has two 6-core Intel Westmere processors, 48 GB of DDR3–1333 RAM, and sixteen 300 GB solid state drives. Its large memory supernodes have over 2 TB of cache coherent memory. Gordon also features dual rail QDR InfiniBand network and data Oasis high performance parallel file system with over 4 PB capacity and sustained rates of 100 GB/s. The theoretical peak performance of Gordon is 341 TFlop/s. Gordon runs Rocks as the cluster management software, CentOS as the operating system, Catalina and TORQUE as the job scheduling and resource managing systems. Mapping Gordon's capabilities against our detailed data and software requirements will be discussed in section 3.3.

3.2 Workflow Integration

The computational workflows we implemented are illustrated in **Figure 1**.

The first step in our analysis is quality control for raw sequencing reads using our internal QC scripts that remove low quality reads based on quality scores. Human sequences are then removed by comparing to human genome and mRNA sequences with Bowtie [10, 11], BWA [12, 13] or FR-HIT [14]. Artificial duplicates, which are common in NGS raw reads, are also removed using program CD-HIT-DUP [15], a program from the CD-HIT package [16, 17]. The filtered reads are mapped against the curated microbial reference genome sequences described above using FR-HIT [14] and BLAT [18]. Taxonomic profiles are then computed based on the mapping results.

The filtered reads are further denoised and clustered to remove sequence errors and redundancy [15] before sequence assembly using Velvet [19], Soapdenovo [20] or Abyss [21]. Filtered reads are mapped to contigs to calculate the abundance of contigs. ORFs are called from assembled contigs by 6-reading frame translation or by using Metagene [22, 23] or FragGeneScan

[24]. ORFs are annotated through comparison to Pfam, Tigrfam, COG, KOG, GO and KEGG databases using Hmmer3 [25], RPS-BLAST and BLASTP [26].

3.3 Workflow implementation

The major computational procedures in our workflow are data-intensive, since they all deal with GB of input files and generate the same scale of output files for a single microbiome sample. In addition, the software tools for these procedures have different computing requirements on CPU, memory and I/O (**Table 2**).

The unique configuration of Gordon resources greatly facilitates the development of our workflows. Two important efforts in our workflow implementation are to allocate jobs in proper compute nodes and to parallelize computationally intensive tasks, especially for two types of processes.

The first type of process, including filtering human DNA, mapping reads to reference genomes, function and pathway annotation against COG, KOG, Pfam, Tigrfam and KEGG databases, require large reference databases. In our analysis, the top largest reference databases are the microbial genome database (15GB) used for mapping, KEGG protein sequence database (6GB) for pathway annotation, and human genome (3GB) for filtering human DNA. A normal compute node on Gordon has 64 GB RAM. So all reference data or the indexed reference data built by aligning or mapping algorithms (e.g. Bowtie, BWA and BLAST) can be fully loaded into computer memory and be concurrently used by all threads across the compute cores. These tasks are further parallelized by dividing the input data, submitting them to multiple compute nodes, and merging the results (**Table 2)**. In this setting, the huge reference databases, which will be read by all the compute nodes, are placed in Gordon's ultrafast Oasis file system. Due to the high I/O speed of Gordon, it is possible to submit a large number of jobs without causing notable delay.

The second type of process, such as removal of duplicated reads by cd-hit-dup and sequence assembly using velvet or SOAPdenovo, do not need a reference database, but the software tools run algorithms on all the input sequences and store the data structure like graph or hash table in computer memory. Gordon's large RAM setting is a good match to this requirement.

Table 1. Gut Microbiome Metagenomic Datasets

Sample Group	Samples	Average PE reads per sample	Total PE reads	Selected samples	Average filtered high quality PE reads per selected sample
LS	3	151,401,139	454,203,418	3	119,694,209
CD	15	91,477,052	1,372,155,785	15	57,624,823
UC	11	14,330,989	157,640,888	6	2,905,681
HE	154	68,609,164	10,565,811,331	34	39,692,865
Total	183	68,578,204	12,549,811,422	58	44,662,870

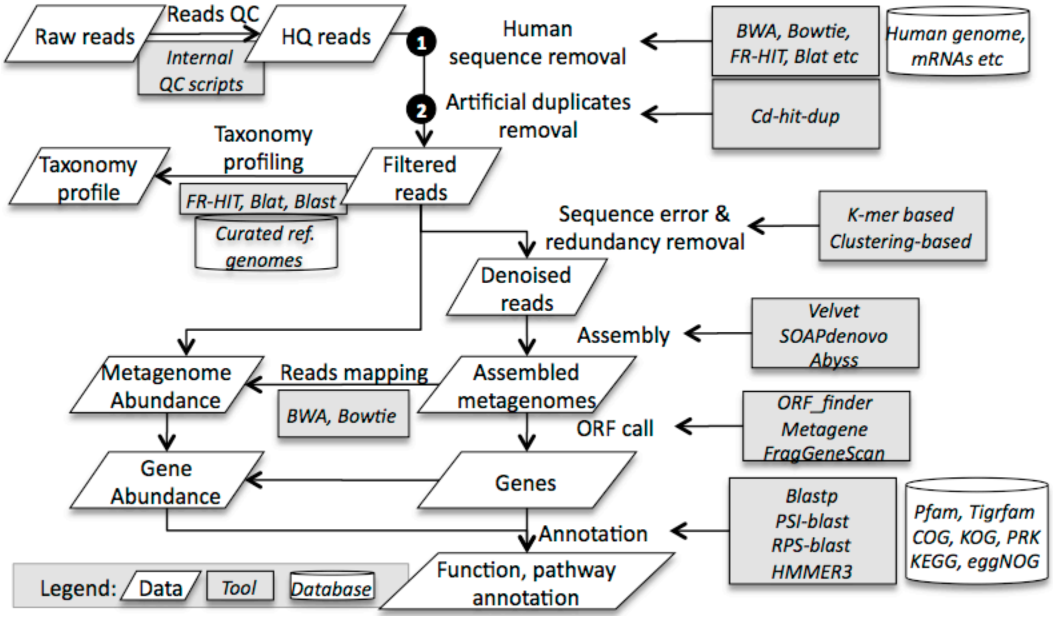

Figure 1. Read-based and assembly-based workflows for Illumina metagenomic data

In addition, most programs in our workflows like cd-hit and velvet have built-in multi-threading feature, so they can take advantage of the multiple cores in Gordon compute nodes. In our analysis, all procedures, except for quality control and ORF call, utilized all 16 cores of Gordon's compute nodes.

4. Results and Discussion

4.1 Computing requirements

The 183 samples were all analyzed using our workflows. The whole analyses took ~180,000 core-hours on Gordon resources and consumed tens of TB of storage.

Among all the analysis steps, pathway annotation against KEGG database is the most time-consuming one, which accounts for about half of the core-hours used. This step was accelerated by aggressive parallelization of over 800 compute cores. The following time-consuming procedures are mapping, duplicates removal, and assembly, which take about 20%, 10% and 10% of the core hours respectively.

Assembly is the most memory-consuming step, which requires up to 256GB memory for some data sets. Duplicates removal and mapping are the next memory-demanding processes, but they all run well within 64 GB memory.

4.2 Taxonomy profiles

In this paper, we will give an overview of the results from our taxonomy analysis. The taxonomy profiles are calculated based on the mapping results between the reads and the reference genomes using the following procedures. A mapped read is assigned to the top matched genome. When a read is aligned to multiple genomes (Number =N) with identical score, each reference receives a coverage of 1/N * read length. This alignment process results in the abundance values for all major taxonomy ranks: domain, phylum, order, class, family, genus, species and strains.

Table 2. Features of the Computational Tools in Our Workflow

Analysis	Tool	Computing requirement			Parallelization	
		Data	CPU	RAM	Multi-threading	Map Reduce
Quality control	QC script	√				
Human DNA removal	Bowtie	√			√	
Duplicates removal	CD-HIT-DUP	√		√	√	
Mapping	FR-HIT	√	√		√	
Read clustering	CD-HIT-EST	√		√	√	
Assembly	Velvet, SOAPdenovo	√		√	√	
ORF call	Metagene, Fraggenescan	√				
ORF clustering	CD-HIT	√		√	√	
COG, KOG annotation	RPS-BLAST	√	√		√	√
KEGG annotation	BLAST	√	√		√	√
Pfam, Tigrfam annotation	HMMER3	√	√		√	√

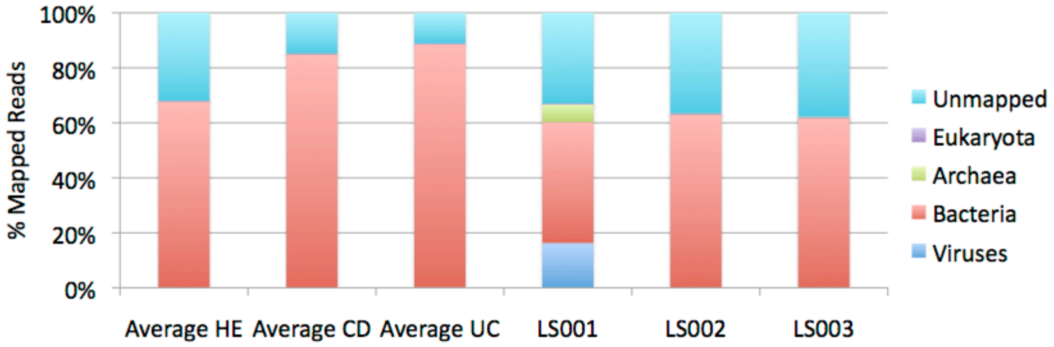

Figure 2. Percentage of reads mapped to domains

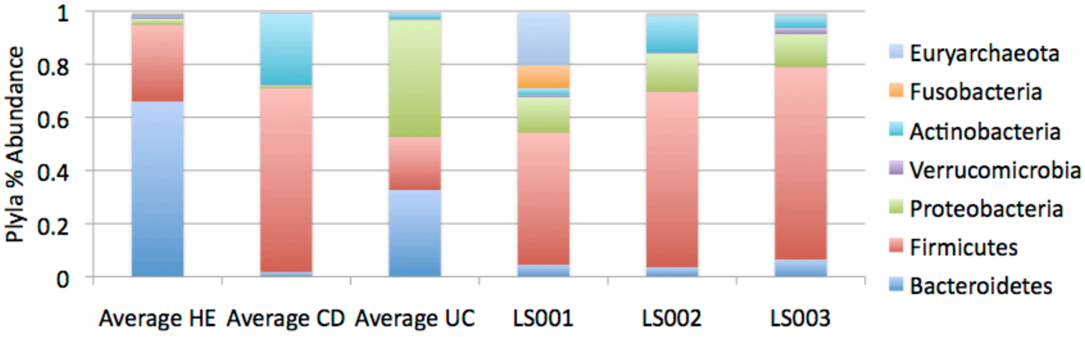

Figure 3. Relative species abundance for the samples at phylum level

We calculated the read coverage for each genome, which is the total coverage divided by genome length. Plasmid sequences were excluded from coverage calculation due to variable copy number. The genome coverage values were normalized against all reference genomes to calculate the relative abundance for each reference genome. For the most abundant microbial species, we found a genome coverage as high as 391.

We created an output spreadsheet of microbial taxonomic rank against all high resolution 58 samples, which leads to very large output spreadsheets. For instance, at the strain taxonomic level, we have ~4000 rows of NCBI taxid identified strains vs. 58 samples, producing ~230,000 filled spreadsheet cells. We then look for patterns within these spreadsheets.

Figure 2 shows the percentage of reads mapped to different domains: viruses, bacteria, archaea and eukaryota. About 12-38% of reads couldn't be mapped to any currently existing reference species, indicating these reads may be from novel species. With the fast growth of the available reference genomes in public databases, many of these unmapped reads will be assignable to known species in the near future, and thereby will allow refinement of our preliminary results shown here. An unusual result is that at the first time sample for LS (LS001), the gut microbiome had a large viral load and a significant portion of archaea, in addition to the bacteria seen in all the other samples.

A high level summary of our results is shown in **Figure 3,** which shows the taxonomy profile at phylum level for LS and averages of healthy and IBD samples. For this comparison all viral reads have been removed, so we are only comparing abundance within the bacteria, archaea and eukaryota domains. Even at this averaged phyla level, a number of results can be seen.

First, in our data, there is a microbial ecology signature differentiating the two types of IBD from each other and from healthy individuals. In the average of healthy subjects, 95% of the microbes are in two dominant phyla - the Bacteroidetes and Firmicutes, with the former in greater abundance than the latter. In the Crohn's Disease subjects, the Bacteroidetes are reduced by over 95% compared to the average of healthy subjects, while the Firmicutes are doubled in their percentage and the Actinobacteria increased by over 30-fold. In the UC subjects the Proteobacteria (predominately E. coli) is increased over 20-fold compared to the average of the healthy subjects, while the fraction of both the Bacteroidetes and the Firmicutes is decreased. A bigger sample size will be needed to determine if this microbial ecology discriminator between the two forms of IBD holds up.

Second, there are also differences in the time series samples from individual LS. Sample LS001 (December 28, 2011) was taken at his largest value of overall inflammation (as measured in the blood by Complex Reactive Protein). The second sample (LS002) was taken (April 3, 2012) after LS had antibiotics for one month and a corticosteroid for two months. The third sample was obtained four months later (August 7, 2012) with no additional pharmaceutical intervention.

The overall pattern of the three LS microbial abundance (Figure 3) is similar to CD in that the Bacteroidetes phylum is greatly reduced and the Actinobacteria phylum is increased. However, the Proteobacteria fraction (~10% in all three samples) is more similar to UC. We might expect this since the CD samples in our study are all from patients[1] with ileal CD, where the primary inflammation is located in the end of the small intestine. In contrast, LS has colonic CD, with inflammation confined to 16 cm. of the sigmoid colon of the large intestine[2]. In UC inflammation is restricted to the large intestine, so we might expect the LS samples to be intermediate between our UC and CD subjects.

Interestingly, the first sample LS001 is significantly different from all the other subjects in that a) archaea methanogens compose 20% of the total archaea/bacterial abundance and b) Fusobacteria are 8% of the total. Significantly after the combined antibiotic/corticosteroid therapy, the Fusobacteria were reduced 90-fold and the archaea were reduced 50-fold. These results are intriguing and needs to be studied further with a wider range of IBD patients receiving therapy.

More detailed biomedical results will be fully discussed in a future publication, including a description of the microbial ecology for each sample and an analysis down to the species and strain level.

4.3 Discussion

To address the great computational challenges in analyzing next generation microbiome sequence data, we developed effective bioinformatics workflows using SDSC's Gordon data-intensive supercomputer, an NSF XSEDE resource.

The advanced configuration, software and hardware frameworks of Gordon enables fast and reliable execution of our workflows on terabytes of sequence data. Extremely time-consuming jobs can be easily performed through efficient parallelization across hundreds of compute cores. With accelerated utilization of NGS technologies, the ever larger-scale of microbiome data or other types of genomic data in the future will routinely require robust bioinformatics workflows and advanced computing infrastructures like Gordon.

Our project has led to the discovery of novel microbial ecological signatures, potentially creating new medical diagnostics for separating healthy from disease states and differentiating between subtypes of autoimmune diseases such as IBD.

5. ACKNOWLEDGMENTS

Software tool development by SW and WL were supported by Award R01HG005978 from the National Human Genome Research Institute (NHGRI) and the microbial metagenomics CAMERA project funded by Gordon and Betty Moore Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NHGRI or the National Institutes of Health.

We are grateful for the CPU hours on Gordon provided by a grant from Mike Norman, Director of the San Diego Supercomputer Center. Funds for sequencing LS001, LS002 and LS003 at JCVI were provided by the UC San Diego Health Sciences.

LS acknowledges the support of the California Institute for Telecommunications and Information Technology and the Harry

[1] Claire Fraser, private communication.

[2] Dr. William J. Sandborn and Dr. Cynthia Santillan, UCSD

E. Gruber Chair, as well as many useful discussions with colleagues.

6. REFERENCES

1. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms, Microbiol Mol Biol Rev 2004;68:669-685.
2. Eckburg PB, Bik EM, Bernstein CN et al. Diversity of the human intestinal microbial flora, Science 2005;308:1635-1638.
3. Gill SR, Pop M, Deboy RT et al. Metagenomic analysis of the human distal gut microbiome, Science 2006;312:1355-1359.
4. Mardis ER. A decade's perspective on DNA sequencing technology, Nature 2011;470:198-203.
5. Turnbaugh PJ, Hamady M, Yatsunenko T et al. A core gut microbiome in obese and lean twins, Nature 2009;457:480-U487.
6. Qin J, Li R, Raes J et al. A human gut microbial gene catalogue established by metagenomic sequencing, Nature 2010;464:59-65.
7. Peterson J, Garges S, Giovanni M et al. The NIH Human Microbiome Project, Genome Res 2009;19:2317-2323.
8. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome, Nature 2012;486:207-214.
9. Smarr L. Quantifying your body: A how to guide from a systems biology perspective, Biotechnology Journal 2012;7:980-991.
10. Langmead B, Trapnell C, Pop M et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol 2009;10:R25.
11. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2, Nat Methods 2012;9:357-359.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 2009;25:1754-1760.
13. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 2010;26:589-595.
14. Niu B, Zhu Z, Fu L et al. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes, Bioinformatics 2011;27:1704-1705.
15. Li W, Fu L, Niu B et al. Ultrafast clustering algorithms for metagenomic sequence analysis, Brief Bioinform 2012;13:656-668.
16. Fu L, Niu B, Zhu Z et al. CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 2012;28:3150-3152.
17. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 2006;22:1658-1659.
18. Kent WJ. BLAT--the BLAST-like alignment tool, Genome Res 2002;12:656-664.
19. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res 2008;18:821-829.
20. Li R, Zhu H, Ruan J et al. De novo assembly of human genomes with massively parallel short read sequencing, Genome Res 2010;20:265-272.
21. Simpson JT, Wong K, Jackman SD et al. ABySS: a parallel assembler for short read sequence data, Genome Res 2009;19:1117-1123.
22. Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes, DNA Res 2008;15:387-396.
23. Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences, Nucleic Acids Res 2006;34:5623-5630.
24. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads, Nucleic Acids Res 2010.
25. Eddy SR. A new generation of homology search tools based on probabilistic inference, Genome Inform 2009;23:205-211.
26. Altschul SF, Madden TL, Schaffer AA et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res 1997;25:3389-3402.