Overseas Practice on（~~Field Epidemiology~~・Collaborative Research）
report form（For Student）

2016/2/19 (Year/Month/Day)

Name	Jednipit Borthong
Laboratory	Division of Bioinformatics, CZC
Year (Grade)	D3
Place of practice	Center for Genomic Epidemiology, National Food Institute, Technical University of Denmark (DTU), Kongens Lyngby, Denmark
Period of practice	Jan 11th – Feb 4th, 2016
Purpose	Collaborative Research

Summary of activities（about 800 words, provide photos, tables and figures that clearly show the activities during the period）

Location and Purpose for collaborative research

Center for Genomic Epidemiology (CGE), Technical University of Denmark is located in Kongens Lyngby. Main research of this center is focused on public health in relation to human nutrition, food safety, food quality, food technology, environment and health. In addition, the center also conducts basis research in the fields of bioinformatics and system biology, especially metagenomic analysis in water sample. The purpose for my visiting at CGE is to analyze 16S metagenomic sequences and learn how to create a database for identification of unknown sequences. After arriving at CGE, Dr. Rene S. Hendriksen, a supervisor additionally advised me for learning shotgun metagenome and data analysis. I was also provided for attending the progress report of PhD student and postdoc at CGE, science club, and lecture class in every week. Moreover, I was designed to have a meeting once a week for discussion in my experiment. All activities are described in this report and short schedules are shown in Table 1.

Table 1 Schedules of activities at CGE, DTU during Jan 8th to Feb 4th, 2016

Day	Time	Activity
Monday	9:30 – 10:00	Breakfast with CGE members
	10:00 – 11:00	Progress report from PhD student and postdoc
Friday	14:00 – 15:00	Science club
	16:00 – 17:30	Lecture class
As design	30 min (once a week)	Update and discuss for my experiment

Fig. 1 Center of Genomic Epidemiology, National Food Institute, Technical University of Denmark

1. Analysis of 16S rRNA metagenome

Backgrounds

A metagenomic analysis based on 16S rRNA (small subunit of ribosomal RNA) recently becomes genetically important marker for elucidation of bacterial population in environment using next generation sequencing (NGS). Identification of 16S query sequences generated from NGS, those unknown sequences need to compare with reference sequence from database and use some algorithm to identify for bacterial name. At present, several databases are established and widely used for instance Ribosomal Database Project (RDP) and MG-RAST metagenomics analysis server. Nevertheless, those databases remain some limitations. RDP can provide the maximum identification of sequence at genus level, which is now not enough information for public health concern, especially when we want to know pathogenic species. On the other hand, MG-RAST can provide more information of identification at species level. Conversely, some 16S query sequences are often to cross-identify with eukaryotic sequences because MG-RAST pooled reference sequences from all microorganisms: bacteria, archaea, and eukaryote within a MG-RAST database. Therefore, new method for analysis 16S metagenomic is required for identification of bacteria at species level.

1.1 SILVA NGS and QIIME

In order to answer question on above, Silva NGS from SILVA high quality ribosomal RNA databases and Quantitative Insights Into Microbial Ecology (QIIME) were recommended to use. Both are dependently work from each other. SILVA is a web server with high quality data for ribosomal RNA gene and can generated result with good visualization, rarefaction curve or krona. Conversely, SILVA NGS has limited with level of identification at genus level as same as RDP. Furthermore, this web server is also limited with number of sample analysis per month by using credit system. For QIIME, it is specific software for metagenomic analysis but this software is very difficult to install, require installation of several programs for analysis, and contain errors in the analysis. After facing those limitations, I designed to created own database and used BLAST for analyzing 16S metagenomic data.

1.2 Creation of 16S rRNA database

Since the web servers and software cannot be answered to research question aim. Therefore, database creation and BLAST were designed to use for analysis of 16S metagenomic data. The processes of database creation are described here, a fasta file of full-range sequence of small subunit ribosomal RNA (SSU) and Basic Local Alignment Search Tool (BLAST) software version 2.2.29+ were downloaded from Silva database version 123 and NCBI, respectively. Normally, SSU file is contained 16S and 18S sequences from bacteria, archaea, and protozoa, respectively. All sequences from archaea and protozoa were then removed from the experiment and only 16S bacterial sequences were further used for making a database. After BLAST installation into local computer, a command line, namely makeblastdb was used to create database and followed by setting variable environment for database. This database is now specific for 16S analysis. Additionally, an accuracy checking of identification with new database was examined using simulated partial 16S sequence from reference sequences of bacteria. The result of identification demonstrated that all stimulated sequences of partial 16S gene were correctly identified to bacteria species. Therefore, this database and pipeline is ready to use for analysis of 16S metagenomic sequence from NGS. However, Dr. Sünje J. Pamp who is Professor at DTU commented me do not use partial 16S rRNA to identify bacteria in species level, whole gene should be used for this purpose.

2. Learning and practice for analysis of shotgun metagenome

A shotgun metagenomic analysis is one kind of metagenomic analysis, which can provide more information than 16S metagenomic analysis. This analysis can be given information for microbial population, functional gene, plasmid, and drug resistance gene including the contamination of tissue from plant, vertebrate mammals, and/or invertebrates. Thus, this analysis becoming an important tool for public health research in near future.

Fig. 2 BLAST program execution in terminal

Skill of shotgun metagenomic analysis was provided using their sequence data from water sample, which were collected from toilet in hospital. Pair-end sequences were generated from NGS machine in fastq format. These sequence were subsequently checked quality of sequence using FastQC (Fig. 2). Partial sequence with low quality (score less than 20) were trimmed out from further analysis by cutadapt 1.9.1 before submit to MGmapper 2.1 (Metagenomics mapper) (Fig. 3). In another way, fastq file could be directly uploaded to the web service because cutadapt is a one of optional steps in pipeline for analysis. Database in MGmapper are generally divided on target of microorganisms and genes, which allow users to select some or all of databases for analysis with best mode and full mode. In my case, the bacteria and archeae databases were chosen for best mode and drug resistance gene was selected for full mode. After selection, pair-end sequences were then uploaded to the web server. The result of analysis will be sent to the register email.

The result of analysis for shotgun metagenome by MGmapper was download from the web server in excel file from ling as attach in email. From the result, it was obviously observed that most of sequences in best mode were mapped to bacteria while small amount of sequences were mapped to archaea (Fig. 5A). In best mode, most sequences were identified to *Aeromonas media* followed by *Tolumonas auensis* DSM 9187, *Klebsiella oxytoca* KCTC 1686 (Fig .5B). On the other hand, for archaea database, more than a half of sequences was identified to *Methanobrevibacter smithii* ATCC 35061 (Fig. 5C). For full mode, most sequences were classified

to resistance gene of aminoglycoside drug compound. As result, MGmapper is web server for using to analyze shotgun metagenome, which is easy to use.

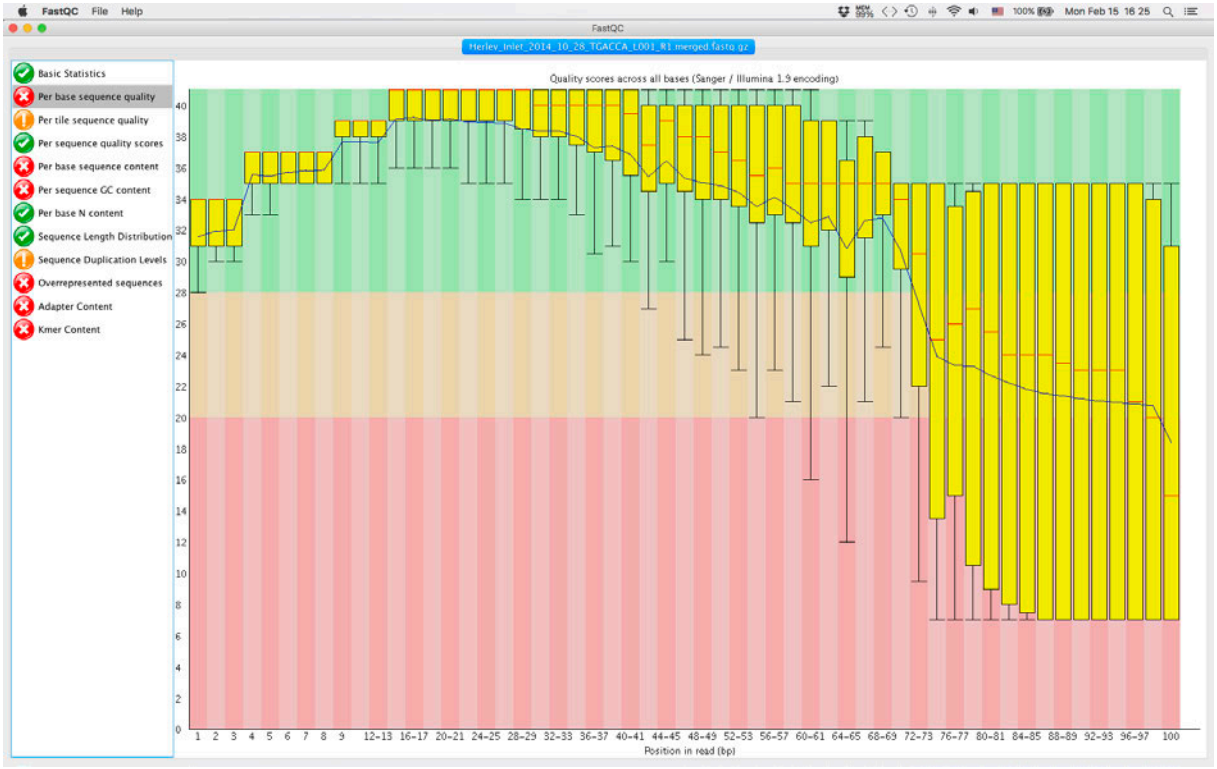
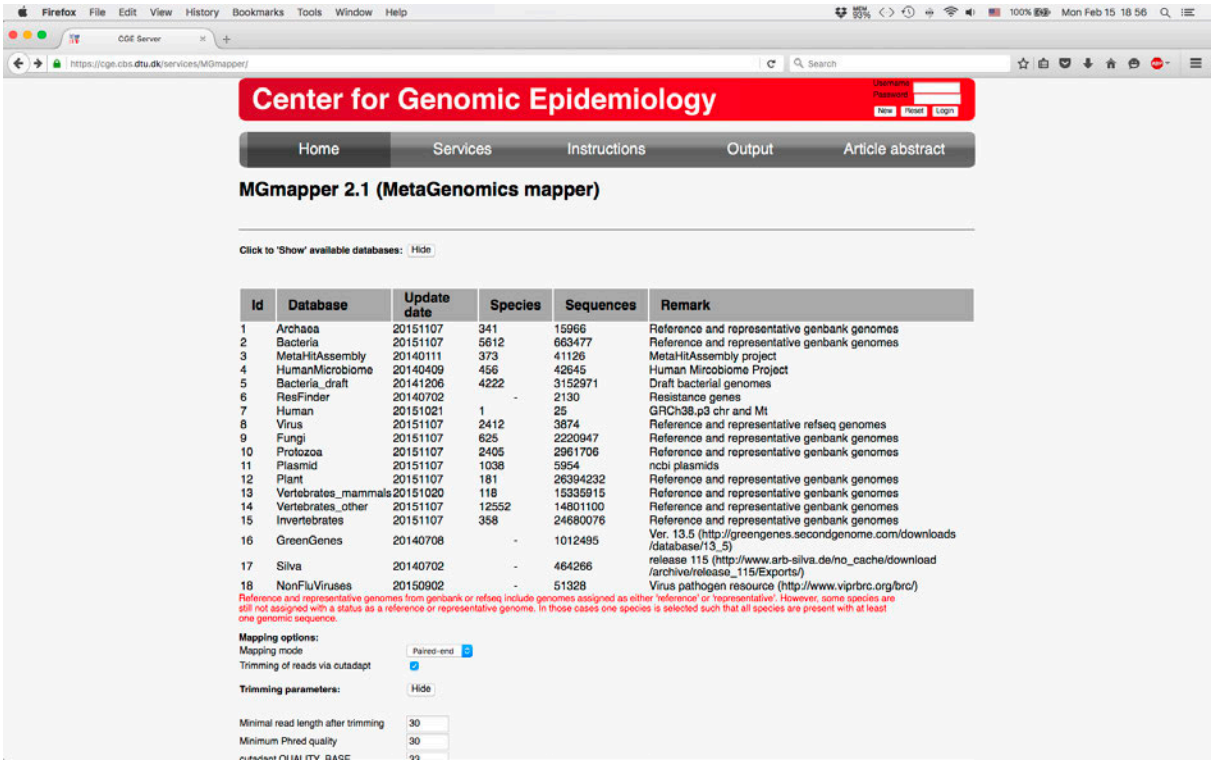


Fig. 3 Results of quality checking of shotgun metagenomic sequence by FastQC



Id	Database	Update date	Species	Sequences	Remark
1	Archaea	20151107	341	15966	Reference and representative genbank genomes
2	Bacteria	20151107	5612	663477	Reference and representative genbank genomes
3	MetaHitAssembly	20140111	373	41126	MetaHitAssembly project
4	HumanMicrobiome	20140409	456	42645	Human Mircobiome Project
5	Bacteria_draft	20141206	4222	3152971	Draft bacterial genomes
6	ResFinder	20140702	-	2130	Resistance genes
7	Human	20151021	1	25	GRCh38.p3 chr and Mt
8	Virus	20151107	2412	3874	Reference and representative refseq genomes
9	Fungi	20151107	625	2220947	Reference and representative genbank genomes
10	Protozoa	20151107	2405	2961706	Reference and representative genbank genomes
11	Plasmid	20151107	1038	5954	ncbi plasmids
12	Plant	20151107	181	26394232	Reference and representative genbank genomes
13	Vertebrates_mammals	20151020	118	15335915	Reference and representative genbank genomes
14	Vertebrates_other	20151107	12552	14801100	Reference and representative genbank genomes
15	Invertebrates	20151107	358	24680076	Reference and representative genbank genomes
16	GreenGenes	20140708	-	1012495	Ver. 13.5 (http://greengenes.secondgenome.com/downloads/database/13_5)
17	Silva	20140702	-	464266	release 115 (http://www.arb-silva.de/no_cache/download/archive/release_115/Exports/)
18	NonFluViruses	20150902	-	51328	Virus pathogen resource (http://www.viprbrc.org/brc/)

Reference and representative genomes from genbank or refseq include genomes assigned as either 'reference' or 'representative'. However, some species are still not assigned with a status as a reference or representative genome. In those cases one species is selected such that all species are present with at least one genomic sequence.

Fig. 4 Interface of MGmapper 2.1, a web service for shotgun metagenomic analysis

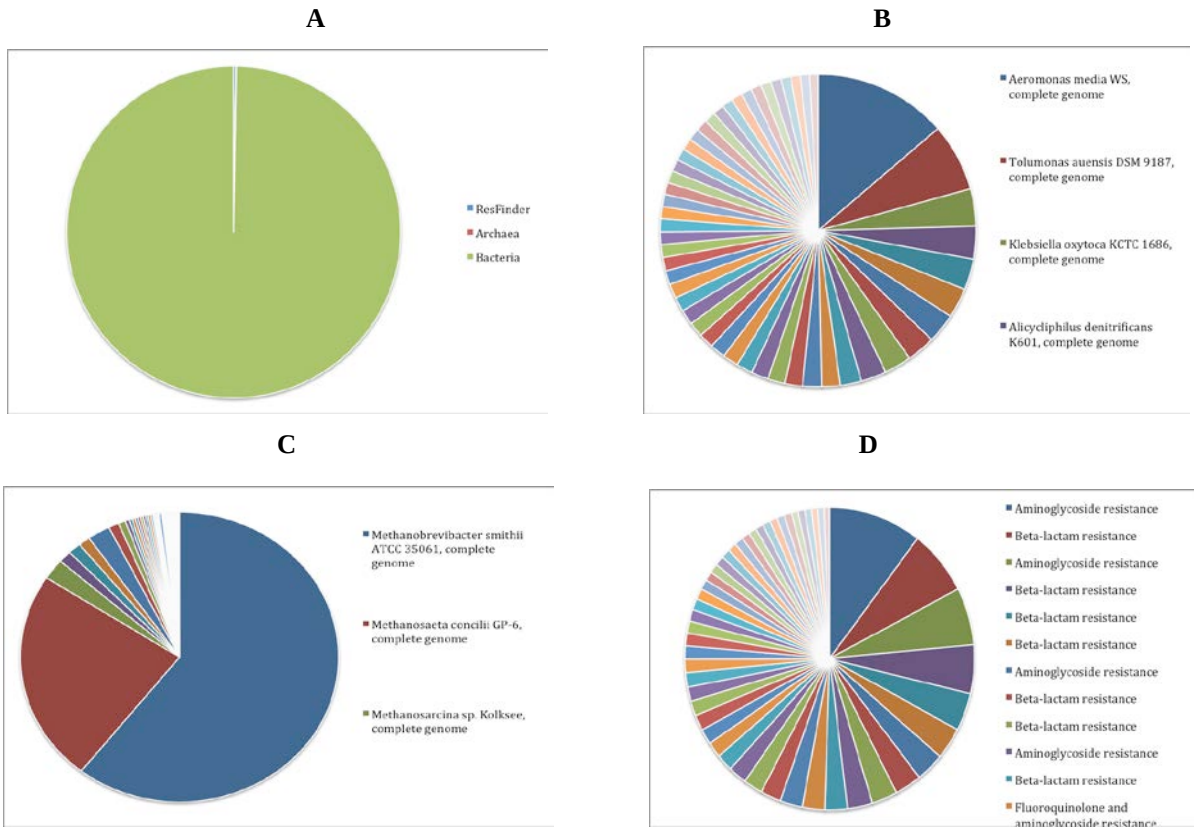


Fig. 5 Result of identification for shotgun metagenomic sequence from MGmapper 2.1, which demonstrate proportion of sequence (A), bacterial (B), archeae (C), and drug resistance gene (D)

3. Participate for discussion section, science club, and lecture class

Three activities were requried to participate during collabolative reaserch at CGE: discussion section, science club, and lecture class.

Progress report by PhD student and Postdoc and breakfast

The section always set up and comprize with two activities on Monday morning in every week. The breakfast will be first started around 9:30 – 10:00, and progress report from PhD student and Postdoc to head of research group, Prof. Frank M. Aarestrup will then started for 30 min – 45 min. In this section, PhD Student and Postdoc will report the activity in previous week, problem in their work, also they will get suggestion from professor or senoir person. Most research in the center can be divide into three group: [1] shotgun metagenomic analysis in water sample, [2] whole genome sequence analysis for etiological agent of foodborne disease, and [3] animal expertment, some PhD student focus on the movement of swine herd for infectious disease, behavivor, feeding, and drug usage from local fram in Denmark.

Science club

Science club will be arraged every Friday in afternoon, in this section will be set up by PhD student and Postdoc for sharing information of their experiment, idea research, or technique for research. In this part, I learnt

following these topics: [1] association between commercial DNA extraction kit and bacteria population. [2] how long we can keep sample at different temperature, and [3] quality of 16S rRNA sequence in Silva and Greengene database. These information are very useful and can be applied for my study in future.

Lecture class

The lecture class is be held on Friday in evening. The speaker will be invited from each reserch group of DTU. I have a chance to participate in the Lecture in titile of Gut feelings – The microbial world with us by Prof Tine R. Licht who is head of resrach group on Gut Microbiology and Immunology. In this lecture, she demonstrated the bacteria population in human intestine and dog intetine including funtional gene. In addition, she also shown the effect of food nutrition to bacteria population in animal.

Fig. 6 Lecture class in topic Gut feeling-The microbial world within us, lectured by Professor Tine R. Licht from Research Group of Gut Microbiology and Immunology from DTU

Summary

The activities at Center of Genomic Epidemiology, DTU improve my knowledge for research study. It allows me to see the different ideas of metagenomic analysis including the process of concentrate microorganisms from water samples. It also shows me to thing about the tiny things that may affect to the study such the quality of DNA sample preparation from commercial kit, sample manipulation, and sample transportation. In addition, this corroborative research automatically teaches me how to due with people from another group and how to live and manage my life in different place.

（~~Field Epidemiology~~・Collaborative Research）Evaluation by supervisor

Institution・Official title・Name	Division of Bioinformatics, CZC 印 Professor Kimihito Ito
Describe overall evaluation on the applicant's activity in overseas practice. Jednipit Borthong has completed his mission successfully. Through his trip, he has learned many things from Dr. Rene S. Hendriksen's laboratory. Especially, analysis using shotgun metagenome would help a lot on his PhD study. Adding to this, he must made many friends there, and the developed network will help his future carrier plan.	

※1 Send the electronic file to the Leading School section, International Affairs Office, also submit the original print out with seal of supervisor to the Leading School section, International Affairs Office.
※2 The Steering Committee of the Leading Program will first confirm the content of this report and the report will be forwarded to the Educational Affairs Committee for credits evaluation.

Submit to：Leading School section, International Affairs Office
Ext: 9545　e-mail: leading@vetmed.hokudai.ac.jp