Summary Report

The 8th WFCC World Data Centre for Microorganisms (WDCM) Symposium – GCM 2.0 Type Strain Sequencing Project Workshop

Beijing, China

November 21-30, 2018

By

Somsak Likhitrattanapisal, Ph.D.

Thailand Bioresource Research Center (TBRC),

National Center for Genetic Engineering and Biotechnology (BIOTEC),

National Science and Technology Development Agency (NSTDA),

113 Thailand Science Park, Phahonyothin Road, Khlong Nueng,

Khlong Luang, Pathum Thani 12120, Thailand

Personal introduction

Somsak Likhitrattanapisal graduated a doctoral degree (Ph.D.) in biology from Mahidol University, Thailand, in 2016. Since January 2017, he has worked as a researcher at National Center for Genetic Engineering and Biotechnology (BIOTEC) and Thailand Bioresource Research Center (TBRC), Thailand. His areas of research include bioinformatics, biodiversity informatics, and biological database management.

ABSTRACT

On 21st – 30th November 2018, Institute of Microbiology, Chinese Academy of Sciences (IMCAS) hosted The 8th WFCC World Data Centre for Microorganisms Symposium and GCM 2.0 Type Strain Sequencing Project Workshop in Beijing, China. Delegates from culture collections and organizations involved in the GCM 2.0 Project were discussed on several topics. IMCAS and TBRC addressed the technical problems and proposed possible solutions both in the discussion forums and in personal conversation. Moreover, the further collaboration from TBRC was herein discussed.

Key words: ***Microbial taxonomy, genomics, microbiome, TBRC Thailand, GCM 2.0***

Contents

1. Brief introduction of TBRC and BCC Culture Collection

Thailand is located in one of the most biologically diverse hotspots in the world. The biodiversity of microorganisms including bacteria, yeasts, and filamentous fungi, is hence tremendous. BIOTEC conducts research studies on microorganisms in many aspects, varying from basic research (such as biodiversity, ecology, and taxonomy) to application research in several fields, such as medical sciences, agricultural technology, food technology, etc. Therefore, BIOTEC has been collecting a huge variety of microbial strains. Currently, BIOTEC hosts two culture collections namely BIOTEC Culture Collection (BCC) and Thailand Bioresource Research Center (TBRC). BCC and TBRC together hold a large collection of these microorganisms more than 80,000 strains in total (about 7,000 strains of yeasts, 20,000 strains of bacteria, and 50,000 of filamentous fungi). BIOTEC ranks the 6th place in the top-10 list of strain holders worldwide.

BCC established in 1996. The primary objective of BCC is to collect and maintain microorganisms and their relevant data for BIOTEC's in-house research. Approximately 75% of the BCC microorganisms holds are filamentous fungi, which are taxonomically and ecologically diverse. Microorganisms in BCC are routinely tested by BIOTEC researchers to

find valuable products such as secondary metabolites, enzymes, and bioactive short peptides (Figure 3). Almost all strains in the collection are cryopreserved at -80°C as working cultures. Freeze drying, liquid-drying and storage in the vapour phase of nitrogen are used for long-term preservation for strains with special characteristics such as newly described species from Thailand, biologically active compound producers, and safe deposit strains.

Thailand Bioresource Research Center (TBRC) was officially opened in February 2015 as a central hub for bioresource management with the goal to enhance the information availability and accessibility using information technology. Many types of bioresources have been archived at TBRC. Microorganism strains in BCC collection, which show the potentials in industrial applications, are selected and transferred to TBRC collection. Moreover, TBRC has also collected the bioresources from other research units in BIOTEC, including plasmids, monoclonal antibodies, etc.

Furthermore, TBRC serves as a information center of bioresources deposited in other research institutes and organizations both in Thailand and foreign countries. The information from our networks is disseminated to the public via online systems (bioresources in TBRC’s online catalogue

include those in BIOTEC and those in other organization). Customers or clients can directly request bioresources of their interests through TBRC system service. Hence TBRC can be regarded as the first full-service “Bioresource bank” of Thailand.

In addition, TBRC also builds the collaboration network with universities and other bioresource centers in both national and international scales. For example, TBRC has established strong network with more than 60 departments from universities and companies in Thailand. In order to accelerate scientific research on sustainable utilization of microorganism, TBRC is one of the founding member of AnMicro network, in which 15 research institutes from 6 countries in ASEAN collaborate financially and technically on capacity building, data sharing, culture collection management, etc.

The operation services of TBRC are divided into 3 workgroups as follows;

1. Bioresource workgroup: provides high-quality bioresource-related services through standard operation and management system, for example, paid deposit service for microbial cultures, technical services for maintenance and identification of microbial strains, training workshops for maintenance and

identification of microbial strains.

2. Bioresouce information workgroup: is a information service center of bioresources. It provides data access to public via information technology and software training to network members and clients.
3. Biotechnology law workgroup: manages legal commissions regarding biodiversity and bioresources, for example, benefit sharing and access as well as provides legal counseling services and training to other organizations.

TBRC offers a set of standard protocols for microbial culture maintenance as follows;

1. Freezing at -80^{o}C. If the temperature inside an incubator rises above the safety threshold, an alarm SMS will be automatically pushed to administrators' devices for immediate actions.
2. Liquid drying. The cultures are maintained in small vacuum glass tubes. They can be preserved at 4^{o}C for 10-20 years. This method is suitable for bacteria, actinomycetes, most yeast strains, and some strains of spore-producing fungi.
3. Freezing in liquid nitrogen. The cultures are maintained at the temperature lower than -150^{o}C. This method is suitable for

virtually all microorganisms, especially non spore-producing fungi.

For safe measures, TBRC has created 2 backup repositories in Bangkok and Nakhon Pathom.

According to aforementioned statements, TBRC’s high-quality bioresource management system assures that your bioresources and cultures are and will always be perfectly kept and maintained, even during emergency situations including power blackout, fire, flooding, etc. With the state-of-the-art technology developed at TBRC, we also offer the high-efficiency information services to research and corporate sectors.

You can get more information about TBRC bioresource collection, technical services, legal counseling services, and training at out website, www.tbrcnetwork.org

2. Benefits from the training courses

The 8th WFCC World Data Centre for Microorganisms (WDCM) Symposium

On 21st - 22nd November 2018, 51 participants from 27 culture collections and research institutes had been participating in the 8th WDCM Symposium at Grand Skylight Catic Hotel Beijing. On 21st November, the keynote presentations addressed a wide range of topics covering large-scale genomics data management, backgrounds on genomic analyses, and microbial taxonomy and systematics. The participants also gave brief talks introducing their culture collections and organizations. On 22nd November, Dr. Linhuan Wu from Institute of Microbiology, Chinese Academy of Sciences (IMCAS) presented the first-year progress report of GCM 2.0 Type Strain Sequencing Project. The project aims to complete the genomic sequencing and analyses of 10,000 microbial type strains in 5 years. The pilot phase of the project, including infrastructure, capacity building, analysis pipeline and database creation, was planned to finish in the first year (2018). Now ~300 culture and DNA samples were submitted to the GCM 2.0 project and sequenced. The forum discussed about the legal issues and technical assessment concerning the genome sequence data both in general and specific to the

GCM 2.0 project. The GCM 2.0 Project has considered to open the project's data and protocols to public databases including NCBI GenBank, EMBL, and DDBJ. The default permission of data usage will be under the free-for-non-commercial-uses condition and compliant with CBD and Nagoya Protocol.

GCM 2.0 Type Strain Sequencing Project Workshop

The training course and workshop was held on 23rd – 30th November 2018 at IMCAS. The topics are as follows;

1. "Combination of lab & paper work: culture collection management" by Prof. Phillippe Desmeth – This topic explained the legal frameworks concerning the management of culture collections and data sharing. The implementation of CBD and Nagoya Protocol was discussed using EU regulation as an example. This session also covered the introduction of LIMS (Laboratory Integrated Management System). Most commercial and free LIMS softwares have built-in functions of quality control and customer services which are compliant with standard protocols and regulations such as ISO 20387.
2. "Genome sequencing of data analyses of Ustilaginomycotina"

by Prof. Qinming Wang – The secretome and pathogenicity analysis results of fungal species in Ustilaginomycotina were shown and discussed. The analyses were performed using several bioinformatics softwares and databases such as SignalP, TargetP, TMHMM, Phobius, WolfP, Interproscan, and CAZy. Principal component analyses (PCA) were also conducted.

3. “Genome sequencing and genome analysis report” by Prof. Bangzhuo ‘Ben’ Tong – The genomic reads from BGI were assembled, assessed, and analyzed using the in-house pipeline. FASTQ reads were filtered using SOAPnuke_filter. The filtered reads then were assemble into contigs and scaffolds using SOAPdenovo, SPADES, PE_reads. The k-mer parameters for genome assemblages were assessed in order to obtain the optimum values (usually 15-60 mers). The quality, correctness, and contamination of genomic sequences were assessed using N50, GC-depth plots and k-mer distribution graph. The genomic features (genes, rRNA, tRNA, repeat elements) were predicted and annotated using Glimmer, RNAMMER, Infernal, BLASTN, tRNAscan-SE, TRF. The functions of proteins were annotated against several databases including VFDB, ARDB,

SwissProt/trEMBL, COG, GO, KEFF, T3DB, NCBI's nr.

4. "Introduction of INSDC, data submission, and genome annotation" by Prof. Yasukazu 'Yaz' Nakamura and Dr. Yasuhiro 'Hiro' Tanizawa – The framework and structure of BioProject in SRA database were explained. The DDBJ data submission workflow and tools were demonstrated using web interface and DFAST. DDBJ/DFAST automatically annotated and packaged the submitted genome sequence data into DDBJ format, thus reduce manual operations. DDBJ infrastructure is now based on NIG Supercomputer which has 42.5 PB of storage and more than 300 TFLOPS of processing power.
5. "BIG Data Center" by Prof. Lina Ma – Beijing Institute of Genomics (BIG) began operation in 2003 and officially opened in 2016. BIG has developed many databases such as Genome Sequence Archive (GSA), ScienceWiki, LncRNAwiki, EWA-Atlas, iDOG, NucMap. BIG also collaborates with other organization in various projects related to precision medicine, public health big data, global biodiversity, etc. Now the supercomputer cluster of BIG has 10,000 CPU cores and 150 TFLOPS in total. After presentation and discussion, the

workshop participants were allowed to visit BIG facilities including an exhibition room, BIG supercomputer, and sequencing machines.

6. “Bioinformatics: algorithm in genome analysis” by Dr. Wenyu Shi – Today bioinformatics is one of the top trends in computer sciences. Dr. Shi explained the WDCM genome analysis pipelines which are to be used in the GCM 2.0 Project. The WDCM pipelines have been developed on Linux and Docker platform. The WDCM pipelines were integrated into the modular platform, gcMETA, which can be accessed via web, API and command-line client interfaces. The gcMETA consists standard and common bioinformatics analysis components such as FASTQC, Trimmomatic, KEGG, MetaCyc, Pfam, Rfam, etc.
7. “Microbiome analysis based on 16S rDNA” by Prof. Jun Wang – The taxonomic analysis of microbiome data are basically assessed in 2 levels; alpha diversity (richness of species in a community) and beta diversity (shared and unique taxa between communities). The conventional biodiversity indices such as Shannon-Wienner’s, Mangalef’s, and Simpson’s indices, can be used to describe alpha diversity of microbiome

data. The beta diversity can be represented using Bray-Curtis and Jaccard dissimilarities. In the analyses, the complexity of microbiome can be reduced using PCA and clustering algorithms. There are several means to statistically test the difference between microbial communities such as MANOVA, SIMPER analysis, BEST analysis. The analysis results of microbiome taxonomy are usually formatted in the matrix format which can be visualized using heatmaps or interaction networks. After the presentation, the workshop participants practiced on microbiome analyses using R (package 'vegan') and a sample data (http://enterotype.embl.de/MetaHIT_Sanger Samples.genus.txt).

8. "Applying advanced bioinformatic tools and machine learning to big data problems in microbiome research" by Prof. Zhenjiang 'Zech' Xu – The next-generation metagenomic sequencing reads from microbiome samples can be analyzed using several software suites, such as QIIME, SciKit-Bio, Oecophylla, micronota, CALOUR, etc. The microbiome studies can be applied in microbial forensics, medical prognoses, etc.
9. "Characterization of prokaryote strains for taxonomic

purposes” by Prof. Man Cai – The prokaryote type strains are important in classification and new species discovery. The classification of prokaryote strains should rely on genetic-based and phenotypic-based characteristics. After the presentation, the participants had a lab-visit tour at China General Microbiological Collection Center (CGMCC).

10. “Diversity of yeast community isolated from crater lakes, plant leaves and soil, and proposal of novel species and genera” by Prof. Aihua Li – The yeast communities were analyzed using multi-gene analyses. The isolation of new taxon candidates were assessed based on physiological and morphological properties.
11. “Bioinformatics: Genome comparative analysis” by Dr. Wenyu Shi – The NGS reads from the GCM 2.0 Project were assembled into contigs and scaffolds using several programs. The assemblies were compared using MUMMER alignment and assessed the consensus.

3. Suggestion on WDCM work

TBRC had sent DNA samples of 15 type strains of bacteria and actinomycetes to the GMC 2.0 Project. The genome assemblages of the majority of these strains are not satisfactory. The possible causes of the problem might be grouped into 2 reasons;

1. Unusual GC-content – The BGI and IMCAS genome analysis pipelines, as well as parameter settings, were designed for and tested with bacterial strains with usual GC-content (between 30% and 70%). The unusually high GC-content of some bacteria and actinomycete strains (>75%) might result in poor assembly results (number of contigs > 200). Therefore, these strains will be re-sequenced and the reads will be assembled using the PCR-free library-based pipeline. It is strongly advised that, if available, TBRC submit the known GC-content values and/or verified 16S rDNA sequences of the microbial strains which were and will be sent to the GCM 2.0 Project. However, IMCAS will derived GC-content from closely-related strains if the GC-content of the strain is not known.
2. Contamination – Some DNA samples showed evident contamination. The contamination were assessed visually with k-

mer frequency distribution histograms and GC-depth plots (noted 2 peaks and hotspots in the example Figure 1. and Figure 2, respectively). The perfectly satisfactory results would produce normal distribution graphs and GC-depth plots. TBRC and BGI are to be responsible to check possible causes of contamination in their sample-handling processes. The new DNA samples or culture samples may be re-sent to the GCM 2.0 Project after the cause(s) of contamination are addressed and solved.

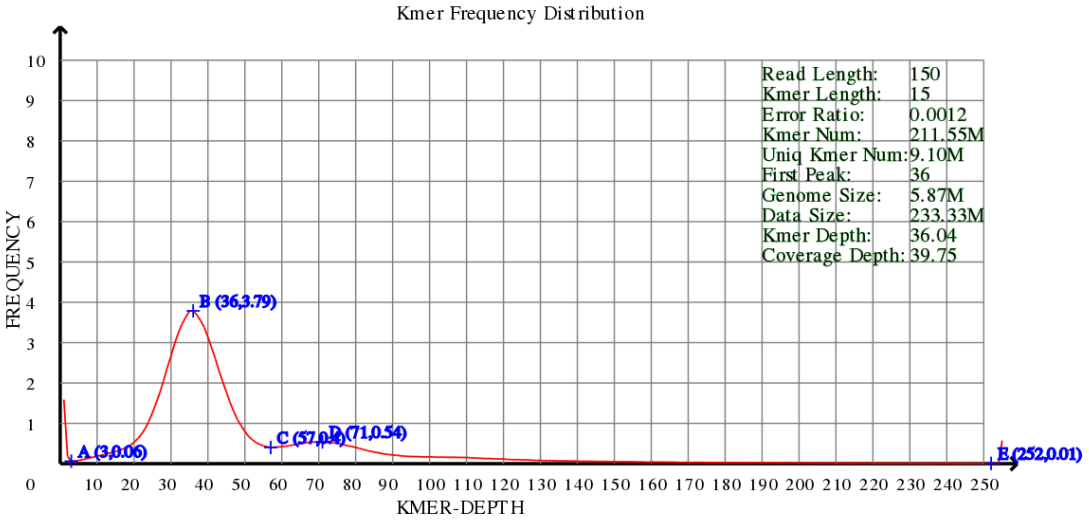

Figure 1. an example k-mer frequency distribution histogram of the contaminated sample TBRC-7906.

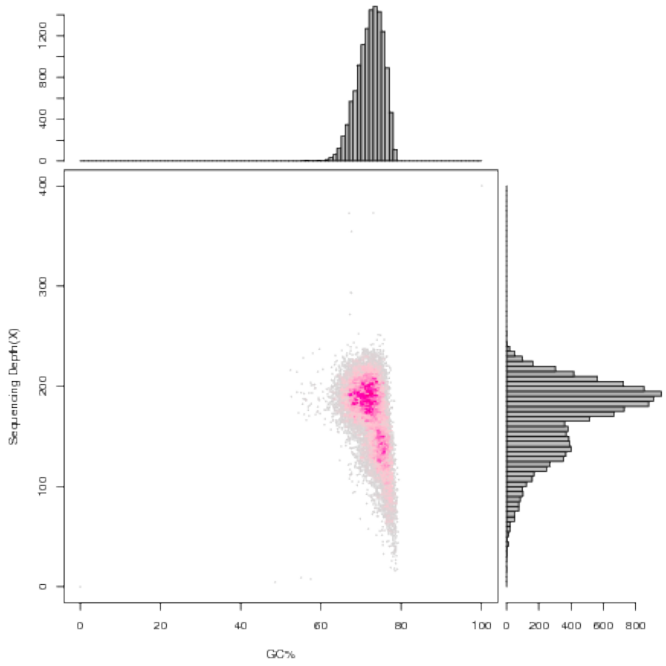

Figure 2. an example GC-depth plot of the contaminated sample TBRC-7906.

In cases of sending new samples for re-sequencing, TBRC, IMCAS, and the related parties may need to consulted the MOU and discuss legal issues later.

4. Comments or suggestion on the training courses

The GCM 2.0 Workshop 2018 presented opportunities for all parties in the GCM 2.0 Type Strain Genome Project to discuss and exchange ideas on a comprehensive range of topics, including legal concerns, technical problems, data sharing, and bioinformatic analyses. The facilities and analysis tools or pipelines developed at IMCAS and BGI showed convinced potentials.

The scale of the GCM 2.0, however, are very ambitious (10,000 genomes in 5-years' time). The success of the project requires good collaboration and strict commitments from all parties involved.

5. Suggestion on further cooperation between WDCM and TBRC

GCM 2.0 Type Strain Sequencing Project is one of the TBRC-WDCM joint collaboration project. In this project TBRC commits to extract and send DNA samples of 56 type strains of Thailand native bacterial species. The genomic sequence data and analytics of TBRC type strains will be shared between Thai and Chinese researchers, according to the mutually-signed agreement.

GCM and TBRC may collaborate further on data sharing. The data in GCM 2.0 Project can be connected via API to AmiBase (ASEAN Microbial Database) which TBRC is leading the development. This connection will add an extra dimension to the genome data of GCM 2.0 Project.