Genome annotation pipelines for prokaryotic and eukaryotic microorganisms using *de novo* short read genome assembly

Mr. Songtham Anuntakarun

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Bioinformatics and Computational Biology
Inter-Department of Bioinformatics and Computational Biology
GRADUATE SCHOOL
Chulalongkorn University
Academic Year 2020

Copyright of Chulalongkorn University

ไพป์ไลน์การแอนโนเทตจีโนมสำหรับจุลชีพโปรคาริโอตและยูคาริโอตโดยใช้การแอสแซมเบิลจีโนมแบบ *de novo* จากลำดับนิวคลีโอไทด์ขนาดสั้น

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

นายทรงธรรม อนุนตการุณ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาชีวสารสนเทศศาสตร์และชีววิทยาเชิงคอมพิวเตอร์ สหสาขาวิชาชีวสารสนเทศศาสตร์และชีววิทยาทางคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Thesis Title	Genome annotation pipelines for prokaryotic and eukaryotic microorganisms using *de novo* short read genome assembly
By	Mr. Songtham Anuntakarun
Field of Study	Bioinformatics and Computational Biology
Thesis Advisor	Associate Professor SUNCHAI PAYUNGPORN, Ph.D.
Thesis Co Advisor	CHUREERAT PHOKAEW, Ph.D.

Accepted by the GRADUATE SCHOOL, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

.... Dean of the GRADUATE SCHOOL
(Associate Professor THUMNOON NHUJAK, Ph.D.)

DISSERTATION COMMITTEE

.... Chairman
(Associate Professor TEERAPONG BUABOOCHA, Ph.D.)

.... Thesis Advisor
(Associate Professor SUNCHAI PAYUNGPORN, Ph.D.)

.... Thesis Co-Advisor
(CHUREERAT PHOKAEW, Ph.D.)

.... Examiner
(Assistant Professor MONNAT PONGPANICH, Ph.D.)

.... Examiner
(Associate Professor NARAPORN SOMBOONNA, Ph.D.)

.... External Examiner
(Associate Professor Onrapak Reamtong, Ph.D.)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ทรงธรรม อนุนตการุณ : ไพป์ไลน์การแอนโนเทตจีโนมสำหรับจุลชีพโปรคาริโอตและยูคาริโอตโดยใช้การแอสแซมเบิลจีโนมแบบ *de novo* จากลำดับนิวคลีโอไทด์ขนาดสั้น. (Genome annotation pipelines for prokaryotic and eukaryotic microorganisms using *de novo* short read genome assembly) อ.ที่ปรึกษาหลัก : รศ. ดร.สัญชัย พยุงภร, อ.ที่ปรึกษาร่วม : อ. ดร.จุรีรัตน์ โพธิ์แก้ว

การวิเคราะห์ลำดับเบสด้วยวิธี Next-generation sequencing (NGS) เป็นเทคโนโลยีที่ใช้หาลำดับนิวคลีโอไทด์ได้เป็นปริมาณมากในเวลาเดียวกัน ซึ่งเทคโนโลยีนี้ได้มีบทบาทในการปฏิวัติวิทยาศาสตร์ชีวภาพ ปัจจุบันมีการศึกษาจีโนมจุลชีพอย่างกว้างขวางด้วยเทคโนโลยี NGS อย่างไรก็ตามพบว่าปัญหาทั่วไปที่เกิดขึ้นในการวิเคราะห์ข้อมูลจีโนมคือ การขาดข้อมูลหรือรายละเอียดของยีนในจีโนมอ้างอิง ดังนั้นวัตถุประสงค์ของการศึกษาคือพัฒนาไพป์ไลน์การวิเคราะห์จีโนมยูคาริโอตและโปรคาริโอตโดยการใช้เครื่องมือและฐานข้อมูลทางชีวสารสนเทศที่สามารถดาวน์โหลดได้อย่างอิสระ ซึ่งในการศึกษาครั้งนี้ใช้จีโนมของ *Leishmania matiniquensis* และ *Leptospira interrogans* เป็นโมเดลในการแอสแซมเบิลจีโนมและศึกษาคุณลักษณะของยูคาริโอตและโปรคาริโอตตามลำดับ ไพป์ไลน์ในโครงการนี้เลือกใช้เครื่องมือ SPAdes ในการแอสแซมเบิลจากลำดับนิวคลีโอไทด์ขนาดสั้น สำหรับ AUGUSTUS และ Prokka จะใช้สำหรับทำนายยีนในจีโนมจุลชีพยูคาริโอตและโปรคาริโอตตามลำดับ นอกจากนี้ ยังมีฐานข้อมูลที่หลากหลายและฐานข้อมูลยีนก่อโรคของทั้งจุลชีพยูคาริโอตและโปรคาริโอตได้รวบรวมไว้ในไพป์ไลน์ สุดท้ายนี้พวกเราหวังว่าจะเป็นประโยชน์ต่อนักวิจัยที่ต้องการวิเคราะห์และต้องการข้อมูลของยีนในจุลชีพ

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สาขาวิชา	ชีวสารสนเทศศาสตร์และ ชีววิทยาเชิงคอมพิวเตอร์	ลายมือชื่อนิสิต ....
ปีการศึกษา	2563	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม ....

6087843820 : MAJOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY
KEYWORD: Genome assembly, Genome annotation, Next-generation sequencing, Virulence factor genes

Songtham Anuntakarun : Genome annotation pipelines for prokaryotic and eukaryotic microorganisms using *de novo* short read genome assembly. Advisor: Assoc. Prof. SUNCHAI PAYUNGPORN, Ph.D. Co-advisor: CHUREERAT PHOKAEW, Ph.D.

Next-generation sequencing (NGS) is the massively parallel sequencing technology that has revolutionized biological sciences. Currently, microorganism genomes have been widely studied using NGS. However, the lack of details in the draft or reference genome is a common problem in genome analysis. Therefore, this study aims to develop genome analysis pipelines for eukaryotic and prokaryotic microorganisms using public bioinformatics software and public databases. *Leishmania matiniquensis* and *Leptospira interrogans* were used as models for genome assembly and annotation in eukaryote and prokaryote, respectively. Our pipelines used SPAdes for short read assembled, AUGUSTUS and Prokka for gene prediction in eukaryotic and prokaryotic microorganisms, respectively. The various functional annotation databases and the eukaryotic and prokaryotic virulence factor gene databases were included in our pipelines. Finally, we hope these pipelines can be useful for the researcher who need to analyze and get the insight into gene information in the microorganism.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Field of Study: Bioinformatics and Computational Biology
Academic Year: 2020

Student's Signature ....
Advisor's Signature ....
Co-advisor's Signature ....

ACKNOWLEDGEMENTS

While a completed dissertation just only has the single name of students, but the long process to complete dissertation is always achieved with the dedicated work of so numerous people during this work. So, I would like to thank with my deepest heart to all the people who contributed in some way to complete my success.

I would like to thank my advisor and co-advisor, Associate Professor Dr. Sunchai Payungporn and Dr. Chureerat Phokaew who give such valuable lesson and opportunity to gain experience in the research field of bioinformatics. I would like to thank my thesis committee members, Associate Professor Dr. Teerapong Buaboocha, Associate Professor Dr. Naraporn Somboonna, Assistant Professor Dr. Monnat Pongpanich, and Associate Professor Dr. Onrapak Reamtong for their support, valuable advice and suggestions that have been important for the development of this work. Without their supervision and constant comment, my thesis would not have been completed.

Thanks, are also due to past and present members of Research Unit of Systems Microbiology and Assistant Professor Dr. Rungrat Jitvaropas for supporting, sharing their knowledge, having a good time, and helping each other for problem solving.

I would like to express my sincere gratitude to all lecturers in Bioinformatics and Computational Biology program at Chulalongkorn University for all knowledge. A special debt of gratitude is expressed to the 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship and the 90th Anniversary of Chulalongkorn University Ratchadaphiseksomphot Endowment Fund from Graduate School, Faculty of Science and Faculty of Medicine, Chulalongkorn University for full scholarship.

Lastly, I would like to convey my appreciation to my family for their supports and advice in every step I have taken in my life.

Songtham Anuntakarun

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

TABLE OF CONTENTS

Page

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

LIST OF TABLES

Page

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

LIST OF FIGURES

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Part 1
INTRODUCTION

Next-generation sequencing (NGS) is the massively parallel sequencing technology that has revolutionized biological sciences, especially in genomic research. There are several platforms of NGS technologies, including 454 Life sciences, Illumina, Ion torrent, and BGI sequencing. They used different techniques to produce an enormous number of short reads from DNA samples. NGS can be used to sequence-specific interested areas or whole genomes. Currently, NGS is used in various research fields in biology, including clinical genetics, microbiology, and oncology [1].

The Human Genome Project (HGP) had been started in 1990 and then was declared complete in 2003. This project aims to map the fragment of nucleotide sequences and assemble to complete reference chromosomes in humans. This project can help the researchers to understand the disease, including the study of genome alteration on oncogenes in a different type of cancers [2], the validation of mutation landscape which was applied for cancer precision medicine [3] and others beneficial applications.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

According to the success of the human genome project, many genome annotation projects were launched after. In 2017, the 100K Pathogen Genome Project was established with the internationalization ally cooperation with many countries, namely China, South Korea, and Mexico. This project provides variety of pathogen draft genomes from many areas which include human and animal disease, food, environmental reservoirs of those pathogens and wildlife. There are many species involved in the project such as *Campylobacter*, *Shigella*, *Salmonella*, *Listeria*, *Helicobacter*, and *Vibrio* species, and more are in progress [4].

Due to the Human Genome Project and Pathogen Genome Project, they provide many draft and reference genomes in several eukaryotic and prokaryotic organisms. However, the lack of details such as the function of genes in some genes in the draft or reference genome is a common problem in genome analysis. Normally, almost of draft genomes in the NCBI public database provides common annotation of genes such as rRNA, tRNA, and some common genes that predicted from programs or related to closely species. Functional annotation is an important step to provide much insight knowledge of genes after predicting gene locations from draft genome sequences. Several databases give details of genes such as pathway, gene ontology, virulence proteins, and function. Therefore, the integration of data from many databases is necessary for gene annotation.

In this study, we aim to develop genome analysis pipelines in eukaryotic and prokaryotic organisms using public Bioinformatics software and public databases. *Leishmania* spp. and *Leptospira* spp. will be used as models for genome assembly and studied the characteristics of their genomes. The improvement of the genome analysis pipelines will be useful for obtaining the insight knowledge of genomes about the functional characterization of the genome, using Bioinformatics tools integrating multiple data sources from various databases, to annotate functional genes in the genome.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Research question

- Is it possible to integrate analysis pipelines using various annotation databases for prokaryotic and eukaryotic microorganisms?
- Are there any genes different between mild and severe strain of *Leptospira interrogans*?

Objectives

To develop pipelines for microbial genome annotations

1. To evaluate pipeline for genome annotation in eukaryotic microorganism
2. To evaluate pipeline for genome annotation in prokaryotic microorganism
    2.1 To compare between mild and severe strain of *Leptospira interrogans*

Keywords

Genome assembly, Genome annotation, Virulence factor genes, *Leptospira interrogans*, *Leishmania martiniquensis*

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Conceptual framework

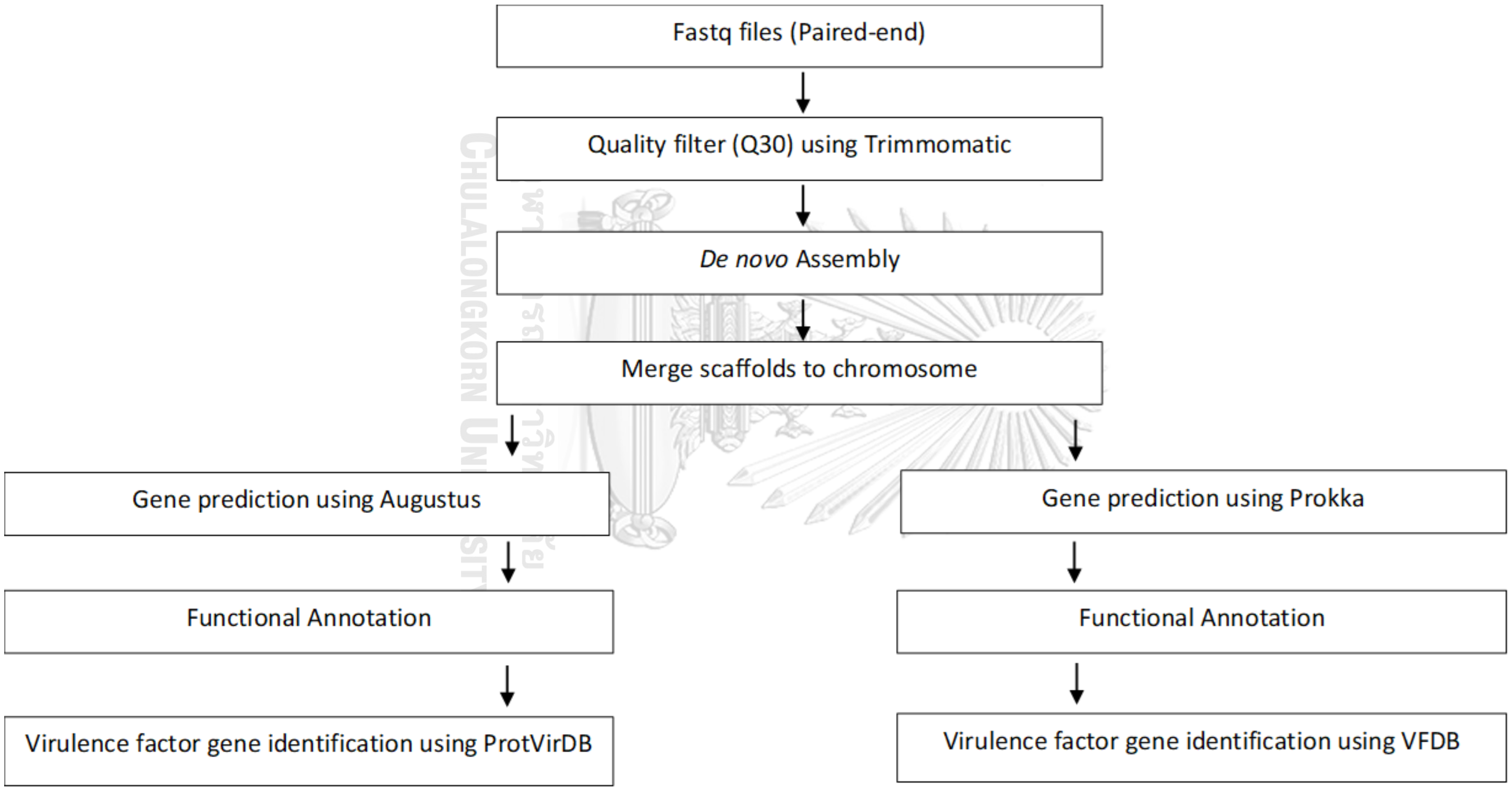

Fastq files (Paired-end)
Quality filter (Q30) using Trimmomatic
De novo Assembly
Merge scaffolds to chromosome
Gene prediction using Augustus
Functional Annotation
Virulence factor gene identification using ProtVirDB
Gene prediction using Prokka
Functional Annotation
Virulence factor gene identification using VFDB

Figure 1 Conceptual workflow of this study

Workflow of study

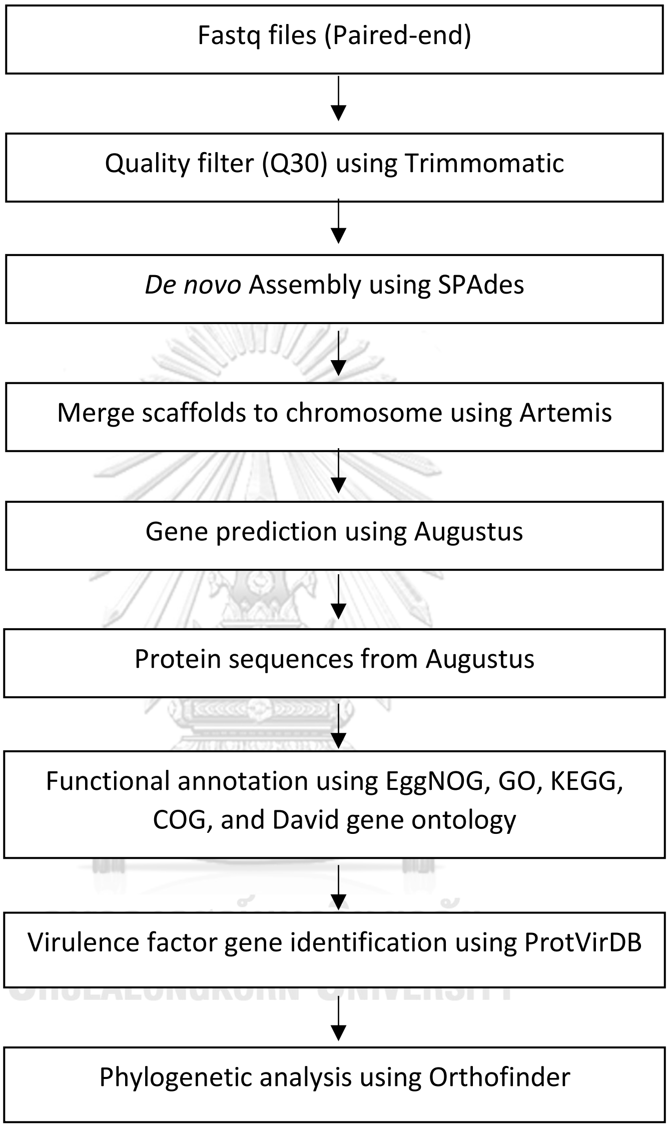
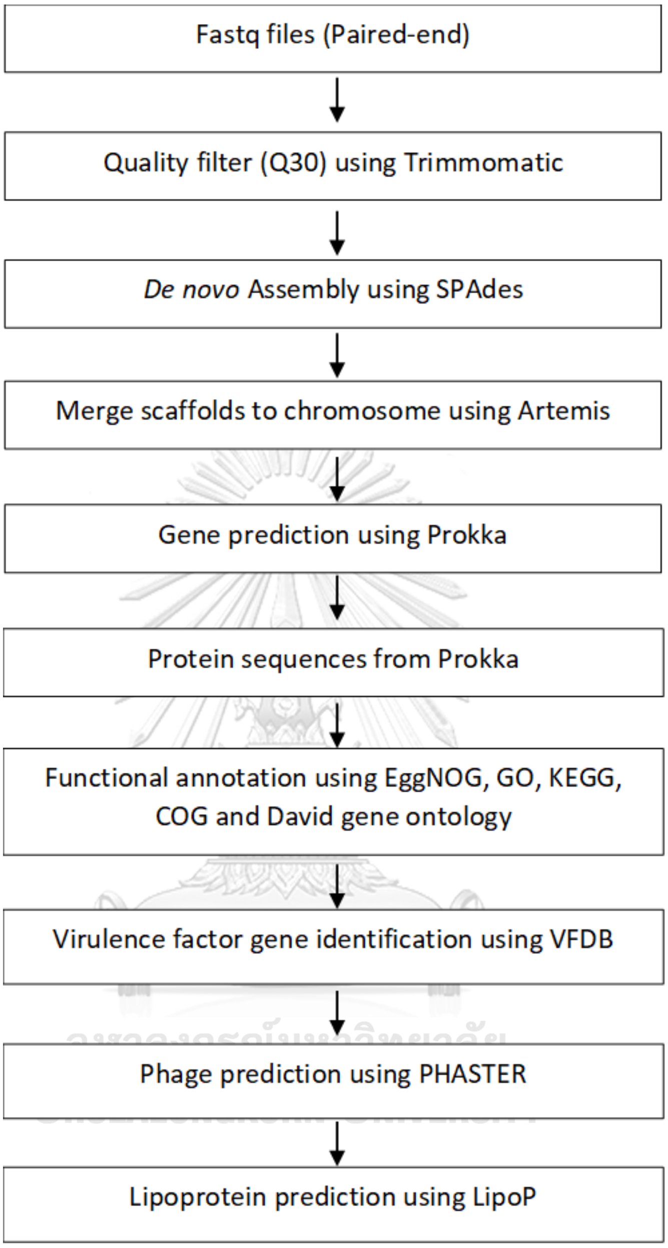

Fastq files (Paired-end)
Quality filter (Q30) using Trimmomatic
De novo Assembly using SPAdes
Merge scaffolds to chromosome using Artemis
Gene prediction using Augustus
Protein sequences from Augustus
Functional annotation using EggNOG, GO, KEGG, COG, and David gene ontology
Virulence factor gene identification using ProtVirDB
Phylogenetic analysis using Orthofinder

Figure 2 Workflow of study *Leishmania martiniquensis* genome


Fastq files (Paired-end)
Quality filter (Q30) using Trimmomatic
De novo Assembly using SPAdes
Merge scaffolds to chromosome using Artemis
Gene prediction using Prokka
Protein sequences from Prokka
Functional annotation using EggNOG, GO, KEGG, COG and David gene ontology
Virulence factor gene identification using VFDB
Phage prediction using PHASTER
Lipoprotein prediction using LipoP

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Figure 3 Workflow of study *Leptospira interrogans* genome

Expected benefits of the study

The benefits of our study to assemble genome from short reads and gain insight of genes from various database. The information of functional annotation and virulence factor gene prediction can guide the researcher focus on interesting genes.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

REVIEW OF RELATED LISTERATURES

Next-generation sequencing

Nowadays, Next-generation sequencing (NGS) is becoming an important role in the study of genomic science. DNA templates in the genome are read randomly from the NGS platform. NGS produces many short reads in range 35-500 bp depend on the platform and experimental design. The Bioinformatics challenging in genomic science interprets short reads and generates the sequencing reads to scaffolds or chromosomes of genomes. There are many bioinformatics approaches to interpret short reads including alignment, assembly, etc. [5].

De novo assembly

De novo assembly is the method for assembling short nucleotide sequences into longer ones without using reference genome. There are three main algorithms used in *De novo* assembly including greedy strategy, the overlap layout consensus, and the de bruijn graph. There is a research [6] suggests that the overlap layout consensus algorithm is more suitable for the low-coverage long reads, on the other hand the de bruijn graph algorithm is more suitable for high-coverage short reads. Building the de bruijn graph starts by collecting all substrings of length k (referred to as k-mers) of all reads; then building a graph with k-mers as nodes and edges connecting two k-mers a and b if the suffix of length (k − 1) of a match the prefix of length k−1 of b and the k+1-mer obtained by overlapping a and b appears in the reads. The de bruijn graph can be built in linear time but storing it requires very large amounts of memory, typically much larger than the string overlap graph. After building the de bruijn graph, each assembler uses several heuristics to simplify graph structures such as cycles and bulges, which mainly induced by repeats in the genome, and bubbles and tips, which mainly induced by sequencing errors and

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

heterozygous sites. Lastly, assemblers select a set of simple paths in the de bruijn graph that would eventually form the contigs.

Gene prediction

After merging contigs or scaffolds to chromosome. Gene prediction and annotation are important steps to identify coding regions and labeling all relevant features on genome sequences [7]. There are several tools and databases used in genome annotation. In this study, Prokka [8] and AUGUSTUS [9] will be used for gene prediction in prokaryote and eukaryote microorganism respectively. Prokka is a command-line software tool to rapidly annotates bacterial, archaeal, and viral genomes and produce standards-compliant output files. Prokka utilizes the external feature prediction tools for identification of coding sequences, rRNA genes, tRNA genes, signal peptide and noncoding RNAs using external software including Prodigal [10], RNAmmer [11], Aragorn [12], SignalP [13] and Infernal [14] respectively. The output files from Prokka represented in Table 1. In the eukaryotic genome, AUGUSTUS is a tool to predict protein-coding genes and their exon-intron structure in genomic sequences using hidden Markov model. Moreover, there are many organisms models for predict gene locations in AUGUSTUS tool. The output files from AUGUSTUS consist of .gff file. There nine columns in gff3 format. The description of each column is shown in Table 2.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 1 Description of the Prokka output file extension

Extension	Description
.gff	This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV.
.gbk	This is a standard Genbank file derived from the master.gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.
.fna	Nucleotide FASTA file of the input contig sequences.
.faa	Protein FASTA file of the translated CDS sequences.
.ffn	Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
.sqn	An ASN1 format "Sequin" file for submission to Genbank. It needs to be edited to set the correct taxonomy, authors, related publication etc.
.fsa	Nucleotide FASTA file of the input contig sequences, used by "tbl2asn" to create the .sqn file. It is mostly the same as the .fna file, but with extra Sequin tags in the sequence description lines.
.tbl	Feature Table file, used by "tbl2asn" to create the .sqn file.
.err	Unacceptable annotations - the NCBI discrepancy report.
.log	Contains all the output that Prokka produced during its run. This is a record of what settings you used, even if the --quiet option was enabled.
.txt	Statistics relating to the annotated features found.
.tsv	Tab-separated file of all features: locus_tag, len_bp, gene, EC_number, COG, product

จุฬาลงกรณ์มหาวิทยาลัย

Table 2 Description of gff3 file format

Column	Header	Description
1	seqid	name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. Important note: the seq ID must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
2	source	name of the program that generated this feature, or the data source (database or project name)
3	type	type of feature. Must be a term or accession from the SOFA sequence ontology
4	start	Start position of the feature, with sequence numbering starting at 1
5	end	End position of the feature, with sequence numbering starting at 1
6	score	A floating point value
7	stand	defined as + (forward) or - (reverse)
8	phase	One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on
9	attributes	A semicolon-separated list of tag-value pairs, providing additional information about each feature

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Functional annotation

Functional Annotation is the technique for describing and collecting the function of genes. The Gene Ontology (GO) [15] is the most comprehensive and extensive functional annotation of gene and protein sequences. There are three terms in the gene ontology including Molecular Function, Cellular Component and Biological process. Molecular Function is the molecular activities of individual gene products. Cellular Component is the parts of a cell or the extracellular environment region, which gene products are active. And Biological process is the process and the pathways in which the activity of gene product is involved. KEGG [16] is a comprehensive resource for understanding high-level functions and utility of biological systems including cells, organisms, and ecosystems from molecular-level data, particularly large-scale molecular datasets produced by genome sequencing and other high-throughput experimental methods. eggNOG [17] is a publicly database that contain various resources including functional annotations, orthology relationship, and history of gene evolutionary.

Virulence factor gene prediction

Virulence factor is a molecule produced by bacteria, virus, fungi, and protozoa used to assist, promote colonization, and bring damage to the host. In prokaryotic, virulence factor database (VFDB) [18] provided up-to-date information of virulence factor genes from various bacterial pathogens. In eukaryotic, protozoan virulent proteins (ProtVirDB) [19] was database provided information of protozoa virulent protein with categories function, based on literature Currently, machine learning techniques were used to apply in various predictions of pathogenic proteins tools [20 - 23]. VirulentPred [21] is a classification tool for predicting virulent protein of bacteria. This tool was built on the Support Vector Machine (SVM) algorithm based on the composition of protein sequence features. This tool was able to achieve a significantly higher accuracy of 81.8%, covering 86% area under curve (AUC) plot. In addition, this tool was used to predict in eukaryotic species. However, the accuracy of prediction in eukaryote is lower than prokaryote. MP3 [24] is a prediction of virulent proteins in both metagenomics and genomics datasets. Support Vector Machine (SVM) and Hidden Markov Model (HMM) approaches were used to develop this tool. This is available as a stand-alone tool and publicly webserver.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Part 2

Genome assembly and genome annotation of *Leishmania martiniquensis* isolated from a leishmaniasis patient in Thailand

(Submitted to Journal of Parasitology Research)

Songtham Anuntakarun[1], Atchara Phumee[2], Vorthon Sawaswong[1], Kesmanee Praianantathavorn[3], Witthaya Poomipak[4], Rungrat Jitvaropas[5], Padet Siriyasatien[6], Sunchai Payungporn[3,7]

[1] Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok, Thailand

[2] Department of Medical Technology, School of Allied Health Sciences, Walailak University, Nakhon Si Thammarat, Thailand

[3] Department of Biochemistry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[4] Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[5] Division of Biochemistry, Department of Preclinical Science, Faculty of Medicine, Thammasat University, Pathum Thani, Thailand

[6] Vector Biology and Vector Borne Disease Research Unit, Department of Parasitology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[7] Research Unit of Systems Microbiology, Chulalongkorn University, Bangkok, Thailand

Correspondence should be addressed to Sunchai Payungporn; sunchai.p@chula.ac.th

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Abstract

Leishmaniasis is a parasitic disease, caused by Leishmania, with worldwide distribution. *Leishmania martiniquensis* is a major cause of autochthonous leishmaniasis in Thailand. For better understanding the genome characteristics of *L. martiniquensis*, high-throughput sequencing was applied for whole-genome sequencing. The FASTQ paired-end reads were trimmed based on Trimmomatic. Pass filtered reads were *de novo* assembled to generate contigs and scaffolds using Spades. Augustus gene prediction tool for eukaryotic annotation was applied for genome annotation of *L. martiniquensis*. Predicted amino acid sequences were searched in EggNOG and David gene ontology databases. In addition, annotated protein sequences that passed the criteria of e-value < $10e^{-5}$ using blastP were searched against the protozoa virulence protein database. From this study, 359 potential virulence factor genes were found in the protozoa virulence protein database. However, these genes should be validated in further study.

Introduction

Leishmania species are members of the Class Kinetoplastea, Order Trypanosomatida. They are intracellular protozoa that are transmitted through vertebrate hosts by infected female phlebotomine sandflies. There are three major clinical presentations of the disease including cutaneous leishmaniasis (CL), mucocutaneous leishmaniasis (MCL), and visceral leishmaniasis (VL). Symptoms of CL occurs on the skin with wet or dry ulcers that are usually painless and localized lesions, while MCL produces sores on mucosal surfaces, especially the nose, mouth, or throat. VL is the most severe form which occurs in internal organs including the spleen, liver, lymph nodes, and bone marrow. The reports of new subgenus Mundinia of Leishmania parasites consist of *L. martiniquensis*, *L. orientalis n. sp.* (previously called *L. siamensis*), *L. enriettii*, and, *L. macropodum* (previously called "*Leishmania* sp. AM-2004") [25-28]. However, only *L. martiniquensis* and *L. orientalis n. sp.* (*L. siamensis*) have been reported to infect humans [25,29-30].

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

In Thailand, autochthonous leishmaniasis was caused by *L. martiniquensis* and *L. orientalis n. sp.* (*L. siamensis*). The *L. martiniquensis* cases in Thailand have dramatically increased in recent years [31-32]. Indigenous leishmaniasis cases in Thailand were diagnosed with CL and VL. Most of the cases were found in immunocompromised patients especially those with AIDS, and these patients also present a poor response to medical treatment. Amphotericin B is the only anti-leishmanial agent available for the treatment of indigenous leishmaniasis in Thailand. Cases of relapsed leishmaniasis caused by *L. martiniquensis* were found after receiving amphotericin B treatment [33]. Therefore, the whole-genome sequencing of *L. martiniquensis* would be useful for the understanding of virulence factor genes and interpretation of clinical severity and manifestations.

There have been many studies of the Leishmania genome in various species based on next-generation sequencing during the past few years [34-35]. Currently, it is known that there are virulence factor genes in protozoans including Leishmania species. These genes are related to parasite survival and infection of the host cell. For example, proteins such as chaperones and endoribonuclease L-PSP can improve the survival rate of the parasite. In addition, some enzymes are related to migrating host cells [36]. Proteinase is also known as a virulence factor in *Leishmania* spp. Proteins and peptides are degraded by protease enzymes that hydrolyze peptide bonds. Moreover, they have a wide range of biological roles, including the mechanism of infection [37].

In this study, the *L. martiniquensis* genome was assembled and explored for a better understanding of its genome characterization. Subsequently, virulence factor genes in this genome were predicted and analyzed. The candidate virulence factor genes will be validated in further studies.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Materials and Methods

Ethics statement

This study was approved by the Institutional Review Board of the Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand (COA No. 768/2012). Patients were not involved in this study.

Promastigote of *L. martiniquensis* culture

The promastigotes of *L. martiniquensis* (CU1 isolated) were isolated from the bone marrow of a leishmaniasis patient in Southern Thailand [38-39]. The promastigotes were cultured in Schneider's Insect Medium (Sigma-Aldrich, Missouri, USA) at a pH of 6.7 supplemented with 10% heat-inactivated fetal bovine serum, 100 U/ml penicillin and 100 μg/ml streptomycin. The promastigotes were incubated at 25±2°C in an incubator and inspected for parasite viability everyday under an inverted microscope (Olympus, Tokyo, Japan).

DNA extraction

The *L. martiniquensis* promastigotes (10^6 parasites/ml) were washed with 1X Phosphate buffer saline (PBS) three times (Sigma-Aldrich, Missouri, USA) and centrifuged at 11,000×g for 10 min. The sample was ground in lysis buffer and used for DNA extraction by using an Invisorb Spin Tissue Mini Kit (STRATEC Molecular, Berlin, Germany), following the manufacturer’s instructions. The DNA concentration and purity were quantified by a Qubit 2.0 Fluorometer (Invitrogen, Massachusetts, USA). The extracted DNA samples were used for sequencing immediately and the rest of the samples were stored at – 80°C.

Library preparation and high-throughput sequencing

DNA (1 μg) was fragmented by using Covaris M220 focused-ultrasonicator (Covaris, Brighton, UK) with 20% duty factor, 50 unit of peak incident power (W) and 200 cycles per burst for 150 seconds. Then the fragmented DNA was used for DNA library preparation based on TruSeq DNA LT Sample Prep Kit (Illumina, California,

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

USA) following the manufacturer’s instructions. The DNA library was cleaned up and the size selected by AMPure XP beads (Beckman Coulter, USA). The concentration of library DNA was measured by using the KAPA Library Quantification Kit (Kapa Biosystems, Massachusetts, USA). The DNA library was diluted to 6 pM and then paired-end sequenced (2x150 bp) based on the MiSeq platform (Illumina, California, USA) by using MiSeq Reagent Kits V2 (300 cycles) according to the standard protocol.

Quality filter and Genome assembly

FASTQ files with 150 bp paired-end reads were checked for the quality of sequences by FastQC [40]. Trimmomatic version 0.39 [41] was used to trim and remove low-quality reads. The processing reads were qualified with high-quality scores (>Q30). De novo assembly was performed using SPAdes version 3.12.0 [42]. The scaffolds sequences from the previous step were used to align with the *Leishmania martiniquensis* genome from the NCBI database (accession number CM030396.1 – CM030431.1 for chromosome 1 – 36) using Artemis comparison tool (ACT) [43].

Gene prediction and functional annotation

AUGUSTUS (Galaxy version 3.3.3) [9] was used to predict genes in the *L. martiniquensis* genome. In this work, the *Leishmania tarantolae* model organism was used in the species parameter for the prediction of gene locations and Protein-coding genes. Putative protein-coding sequences from AUGUSTUS were performed in the functional annotation. The EggNOG-mapper version 2 [44] (default parameters) was used to predict functional annotation against EggNOG 5.0 [17]. This database contains functional information from many sources including a Cluster of orthologous groups of proteins (COGs) [45], KEGG pathway [46], and GO annotation [47].

Prediction of the virulence factor gene

Putative protein-coding sequences were analyzed by blastP with Protozoa virulence protein database (ProtVirDB) [19] and Pathogen host interaction database (PHI-base) [48] for predicting candidate virulence factor proteins and interaction between hosts and pathogens, respectively. In this study, the criteria for the determination of candidate virulence sequences were using criteria e-value of $10e^{-5}$. For proteinase gene analysis, proteinase genes of *L. martiniquensis* were predicted using sequences from the previous report [37] as a reference.

Phylogenetic tree analysis

OrthoFinder version 2.5.2 [49] with default parameter was used for finding single-copy orthologous genes and alignment of single-copy orthologous genes. In this study, the protein sequences dataset from various species including *Trypanosoma brucei* TREU927 (GCF_000002445.2), *Trypanosoma vivax* Y486 (CA_000227375.1), *Trypanosoma grayi* (GCF_000691245.1), *Trypanosoma cruzi* strain CL Brener (GCF_000209065.1), *Trypanosoma rangeli* (GCF_003719475.1), *Phytomonas sp.* isolate EM1 (GCA_000582765.1), *Leptomonas seymouri* (GCA_001299535.1), *Leptomonas pyrrhocoris* (GCF_001293395.1), *Leishmania enriettii* (GCA_017916305.1), *Leishmania martiniquensis* (GCA_017916325.1), *Leishmania tarentolae* (GCA_009731335.1), *Leishmania mexicana* MHOM/GT/2001/U1103 (GCF_000234665.1), *Leishmania major* strain Friedlin (GCF_000002725.1), *Leishmania donovani* (GCF_000227135.1), *Leishmania infantum* JPCM5 (GCF_000002875.1), *Leishmania panamensis* (GCF_000755165.1), *Leishmania braziliensis* MHOM/BR/75/M2904 (GCF_000002845.1) and our *Leishmania martiniquensis* were used as input of OrthoFinder tool. The newick format of phylogenetic tree from OrthoFinder was visualized using Interactive Tree Of Life (iTOL) (https://itol.embl.de/) [50].

Comparison of *L. martiniquensis* genome with other Leishmania species

The analysis percentage's identity of *Leishmania* chromosomes was performed on representative *Leishmania* species including *Leishmania major* strain Friedlin (GCF_000002725.1), *Leishmania infantum* JPCM5 (GCF_000002875.1), *Leishmania donovani* (GCF_000227135.1), *Leishmania mexicana* MHOM/GT/2001/U1103 (GCF_000234665.1), and *Leishmania martiniquensis*

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

LU_Lmar_1.0 (GCA_017916325.1) using Clustal Omega version 1.2.4 with default parameter [51].

Results

Genome characteristics of *L. martiniquensis* genome

Paired-end FASTQ files were used for *de novo* assembly using SPAdes. After assembly, there were 6,939 scaffolds with N50 63,362 bp. The statistics of *L. martiniquensis* data are shown in Table 3. After the gene prediction step, there were 8,209 protein-coding genes in the final assembly of chromosome 1 to chromosome 36. The chromosome size ranges from 0.24-2.8 Mb. The existence of regions in the genome with large variations in the CG content may be caused by over-or under-fragmentation during the library construction. The *L. martiniquensis* genome had an average GC content of 59.77%. The details of our genome were compared with other *Leishmania* species collected from previous research [52], as shown in Table 4.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 3 Statistics of *L. martiniquensis* data and de novo assembly

Genome features of *L. martiniquensis*	
Length (bp)	150
Raw reads	26,205,720
Q30 reads	23,836,943
Number of Scaffolds	6,939
N50 (bp)	63,362
Number of protein coding-genes	8,209

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 4 Comparison genome characteristics of *L. martiniquensis* with other Leishmania species

Feature	*L. major* Friedlin	*L. infantum* JPCM5	*L. mexacana* U1103	*L. brazillensis* M2904	*L. donovani*	*L. martiniquensis*	*L. martiniquensis* LU_Lmar_1.0
Number of chromosomes	36	36	34	36	36	36	36
Chromosome size range (Mb)	0.27-2.68	0.28-2.67	0.27-3.34	0.23-2.6	0.27-2.6	0.24-2.8	0.26-2.6
Total size (Mb)	32.85	31.92	30.94	31.98	32	30.78	32
GC content (%)	59.72	59.58	59.72	57.76	59.50	59.77	59.85
N content (%)	<0.01	0.06	0.11	0.25	3.81	0.025	<0.01
Protein-coding genes	8,400	8,199	8,106	8,160	8,014	8,209	7,993
Tranfer RNA	83	67	83	66	64	80	NA

*NA = data not available

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Comparison of *L. martiniquensis* with other Leishmania species

The genome (36 chromosomes) of *L. martiniquensis* was compared with other *Leishmania* species. The percentages of identity were approximately 17% to 21% compared with *L. infantum*, *L. donovani*, *L. braziliensis*, *L. major* strain Friedlin and *L. mexicana*. In addition, the result of identity percentages compared with *L. martiniquensis* LU_Lmar_1.0 was highly percentages with others (approximately 19% to 57%). The result of identity is shown in Table 5. The COG functional category in *L. martiniquensis* was compared with other *Leishmania* spp. including *L. infantum*, *L. donovani*, *L. braziliensis*, *L. major* and *L. mexicana*. Our result showed that the functional category based on COG of *L. martiniquensis* was similar to other *Leishmania* spp (Table 6). The KEGG pathway analysis and GO annotation are represented in Figure 4. In the KEGG pathway analysis (Figure 4A), the top three pathways include ribosome, metabolic pathways, and RNA polymerase. Functional annotation is the process of collecting information about the function of genes. The Gene Ontology (GO) is the most widespread and extensive functional annotation for gene and proteins sequences. There are three terms in gene ontology. First, the molecular function comprises the molecular activities of individual gene products. Second, the cellular component comprises the region of active gene products. Third, the biological process comprises the process and the pathways in which the activity of gene products is involved. The result of GO analysis in Figure 4B-4D shows that the top three molecular functions were structural constituent of ribosome, poly (A) RNA-binding, and DNA-directed RNA polymerase activity. The top three cellular component functions were cytosolic large ribosomal subunit, motile cilium, and intraciliary transport particle B. The top three biological process functions were translation, rRNA processing, and ribosomal large subunit assembly.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 5 Comparison of percent identity in Leishmania spp. including *L. infantum, L. donovani, L. major* Friedlin, *L. Mexicana, L. braziliensis, L. martiniquensis* and our *L. martiniquensis*

Chr	*L. braziliensis*	*L. donovani*	*L. infantum*	*L. major*	*L. mexicana*	*L. martiniquensis*
Chr1	18.82	19.01	19.18	19.55	19.20	19.13
Chr2	19.09	18.73	21.53	21.71	18.87	38.23
Chr3	18.63	18.91	19.13	19.38	19.20	57.50
Chr4	18.16	19.33	18.41	18.47	18.57	32.05
Chr5	18.14	18.48	18.71	18.56	18.62	19.46
Chr6	18.34	18.68	18.78	18.71	18.94	45.44
Chr7	18.32	18.40	19.06	18.62	18.66	26.46
Chr8	18.51	17.87	18.69	18.75	18.70	25.95
Chr9	19.89	18.17	18.99	19.77	19.55	28.77
Chr10	19.35	20.06	20.71	20.03	21.45	26.58
Chr11	18.50	17.26	18.13	18.05	18.13	19.88
Chr12	18.12	16.82	18.32	18.58	18.63	29.20
Chr13	18.27	17.85	18.29	20.11	20.24	25.00
Chr14	18.88	19.38	19.92	19.13	19.23	59.70
Chr15	18.30	18.41	18.29	18.44	18.29	20.46
Chr16	18.26	17.88	18.98	18.44	18.54	26.26
Chr17	18.17	18.14	18.46	18.40	18.41	40.39
Chr18	18.08	17.86	18.03	17.95	17.89	23.20
Chr19	17.89	17.41	18.96	19.17	18.01	22.55
Chr20	18.13	20.59	19.79	20.22	19.51	57.58
Chr21	18.64	18.47	18.58	19.79	19.84	42.76
Chr22	17.92	17.89	18.17	18.40	18.17	20.74
Chr23	17.83	18.16	18.43	18.49	18.66	24.29
Chr24	18.18	17.98	18.25	18.21	18.19	24.57
Chr25	18.14	19.36	19.51	19.63	19.43	33.01
Chr26	17.98	18.42	18.69	18.80	18.44	23.42
Chr27	17.84	16.95	18.19	18.06	18.20	19.23
Chr28	17.89	17.84	18.18	18.11	18.20	25.06
Chr29	17.81	17.29	17.83	17.82	17.80	24.91
Chr30	17.76	17.52	17.88	17.88	17.86	29.76
Chr31	17.78	18.30	18.08	18.42	17.88	23.19
Chr32	17.63	17.76	18.01	17.94	17.80	22.23
Chr33	17.67	17.52	18.21	18.07	18.04	23.79
Chr34	18.17	17.85	17.38	17.86	18.72	29.40
Chr35	17.54	17.04	17.81	17.86	NA	20.76
Chr36	17.53	17.46	18.60	17.77	NA	19.22

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 6 Comparison of functional category of putative protein-coding genes in *Leishmania spp.* genome. The alphabet A-G represent name of Leishmania species including *L. martiniquensis* (A), *L. infantum* (B), *L. donovani* (C), *L. braziliensis* (D), *L. major* strain Friedlin (E), *L. mexicana* (F) and *L. martiniquensis* LU_Lmar_1.0 (G).

Functional category based on COG		Number of genes in *Leishmania spp.*						
		A	B	C	D	E	F	G
Information storage and processing								
J	Translation, ribosomal structure, and biogenesis	374	391	347	383	392	392	376
A	RNA processing and modification	227	231	227	224	229	230	224
K	Transcription	83	82	81	81	85	82	78
L	Replication, recombination, and repair	156	153	152	150	153	150	146
B	Chromatin structure and dynamics	45	50	45	59	55	54	49
Cellular processes and signaling								
D	Cell cycle control, cell division, chromosome partitioning	59	59	59	60	60	61	59
Y	Nuclear structure	3	3	3	3	3	3	3
V	Defense mechanisms	33	35	34	29	32	33	32
T	Signal transduction mechanisms	317	313	307	306	309	309	290
M	Cell wall/membrane/envelope biogenesis	15	14	14	16	15	15	13
N	Cell motility	13	14	11	13	14	14	13
Z	Cytoskeleton	148	143	139	149	172	151	146
W	Extracellular structures	2	2	2	2	2	2	2
U	Intracellular trafficking, secretion, and vesicular transport	218	227	223	212	219	221	208
O	Posttranslational modification, protein turnover, chaperones	404	427	412	436	452	429	425
Metabolism								
C	Energy production and conversion	150	155	152	155	159	155	132
G	Carbohydrate transport and metabolism	219	200	186	224	212	198	190
E	Amino acid transport and metabolism	168	166	159	163	164	160	144
F	Nucleotide transport and metabolism	74	72	71	68	68	69	73
H	Coenzyme transport and metabolism	134	133	132	126	129	131	134
I	Lipid transport and metabolism	203	201	198	198	204	200	176
P	Inorganic ion transport and metabolism	87	94	87	87	87	88	84
Q	Secondary metabolites biosynthesis, transport, and catabolism	100	87	87	89	95	90	89
Poorly characterized								
S	Function unknown	1515	1530	1480	1521	1636	1559	1429

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

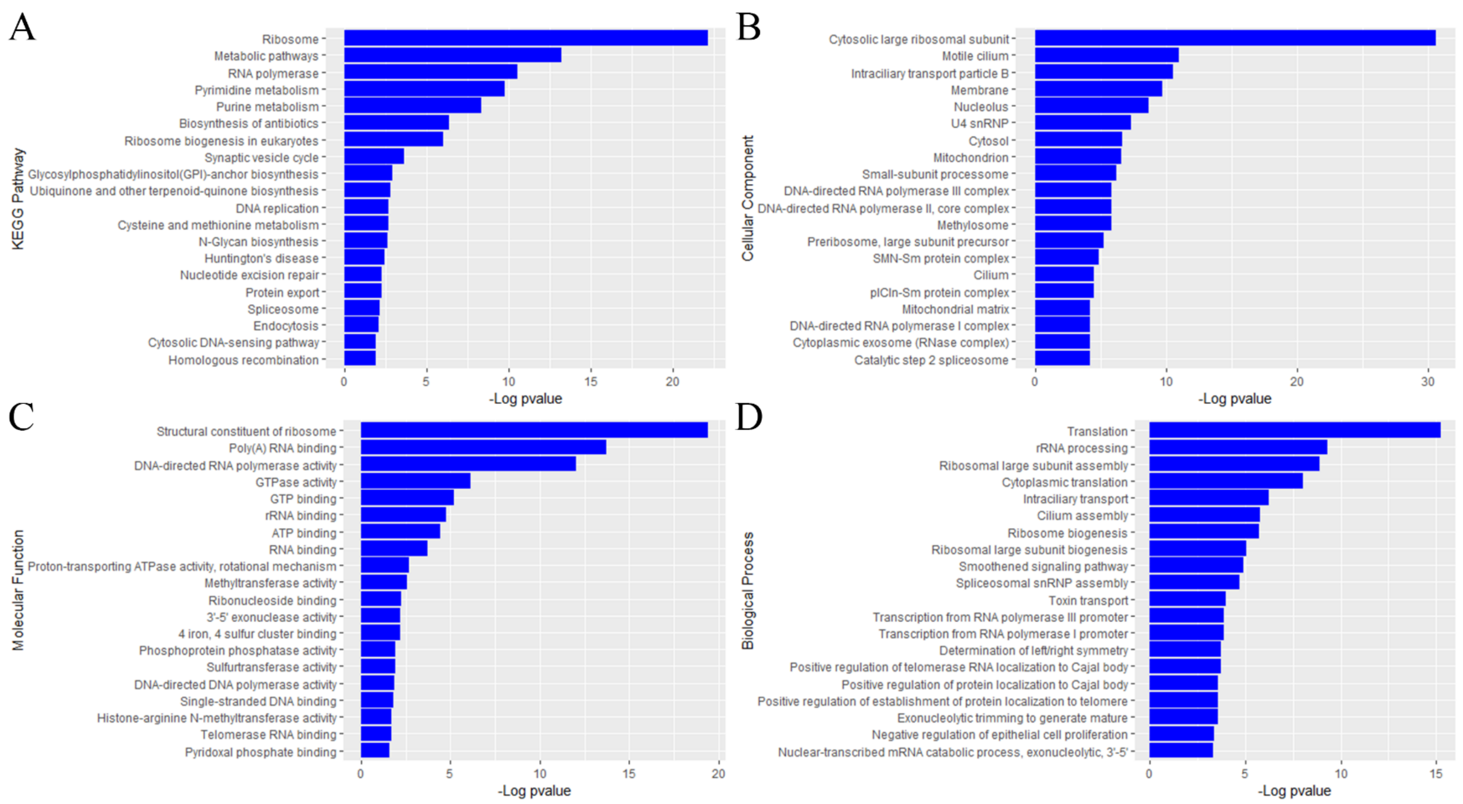


Figure 4 The functional annotation of *L. martiniquensis*. (A) The KEGG pathway annotation, (B) The cellular component annotation, (C) The molecular function annotation, and (D) The biological process annotation

Virulence factor gene analysis

Predicted genes in protein sequences from AUGUSTUS tool were blastP with protozoa virulence protein database (ProtVirDB) using the criteria of e-value < $10e^{-5}$. A total of 359 genes were found as candidate virulence factor genes. These genes were then analyzed for COG functional annotation. The top three COG functions were signal transduction mechanism, carbohydrate transport and metabolism, and intracellular trafficking, secretion, and vesicular transport, while the remaining COG functions are shown in Figure 5. The annotation lists of the 359 genes from ProtVirDb were shown in supplementary material1. Moreover, forty-three predicted protein sequences that passed the criteria from blastP with PHI-base were related with Homo sapiens organisms. The annotation lists of the 43 genes from PHI base are shown in supplementary material2. However, the predicted virulence factor gene should be validated in further study.

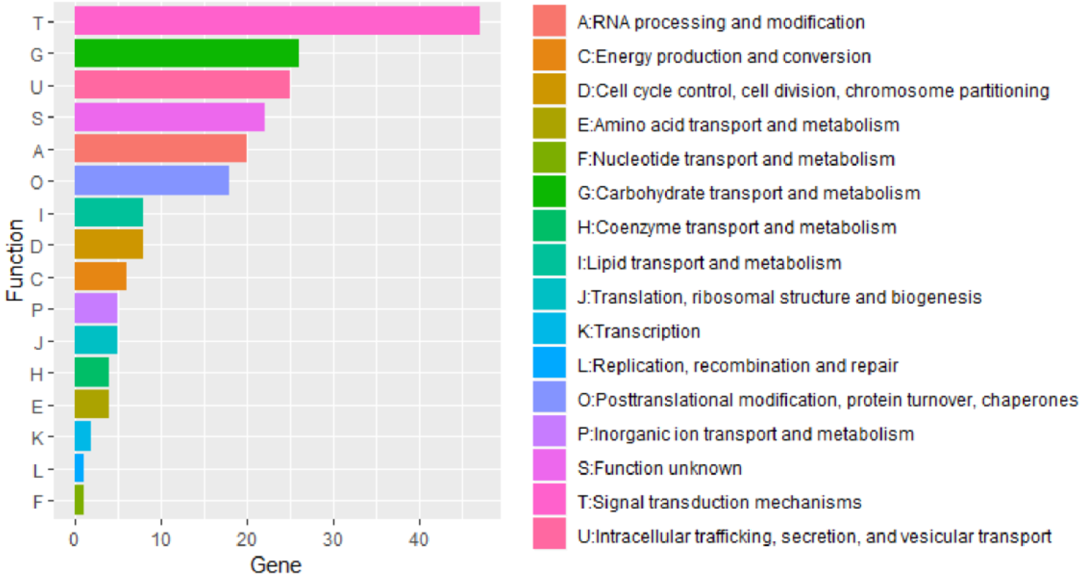


Figure 5 The COG functional analysis of candidate virulence factor protein-coding genes

Phylogenetic tree analysis

The concatenated protein sequences of up to 17 single-copy orthologous genes were used to create a phylogenetic tree. In Figure 6, the phylogenetic tree indicates that *L. martiniquensis* is related to *Leishmania* spp. Moreover, the outgroup including *Trypanosoma brucei* TREU927, *Trypanosoma vivax* Y486, *Trypanosoma grayi*, *Trypanosoma cruzi* strain CL Brener, and *Trypanosoma rangeli* is a more distinctly related group of the *Leishmania* species. This result suggests that *L. martiniquensis* is closely related with *L. martiniquensis* LU_Lmar_1.0 that published in April 2021 on NCBI website.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

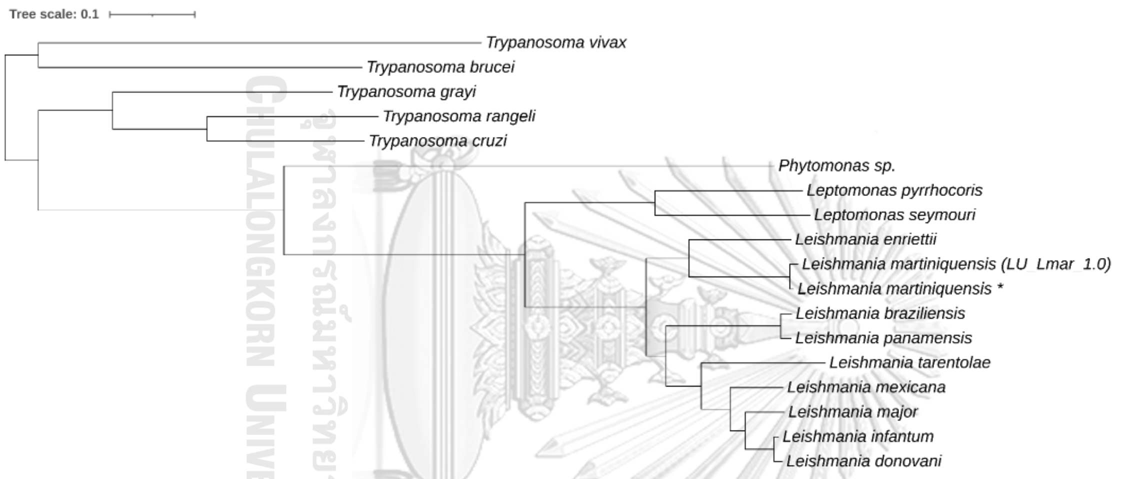


Figure 6 Phylogenetic tree of 17 single copy orthologous genes. The star symbol represents our Leishmania genome.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Discussion

In this study, genome assembly and gene prediction of *Leishmania martiniquensis* were performed. The results showed that the COG functional category of *L. martiniquensis* was similar to other *Leishmania* species. However, there was a slight difference in the number of genes in each functional group. The importance of parasite virulence factors has become apparent in recent years [53]. The variability of virulence factor genes within the *Leishmania* species is largely unknown. In our virulence factor gene prediction of *L. martiniquensis*, the result showed that 359 candidate virulence factor genes were found in *L. martiniquensis*. Some of these genes are discussed below.

Heat shock protein (HSP) comprises intracellular molecules of varying molecular weights. They are a large family of molecular chaperones. The role of this protein is maturation, degradation and refolding [54]. They also play an important role in immune biological functions, especially in hsp70. There was a report which showed that hsp70 induces dendritic cells to generate pro-inflammatory cytokines [55] and is related to the enhancement of adaptive immunity [56]. In our results, the hsp70 protein-coding gene in 359 candidate virulence factor genes was found. This gene might be related to the infection of host cells.

Proteinase is an enzyme that hydrolyzes peptide bonds in proteins, and participates in a wide range of biological functions, including the process of infection [37]. There are many classes of proteinase based on catalytic domains [57]. There are only 3 classes, including aspartyl-, metallo- and cysteine-proteinase, which have been extensively studied in Leishmania organisms [58-59]. In a previous review, cysteine proteases were considered to play a crucial role in the pathogenesis of other parasitic protozoan infections [60]. CPA, CPB and CPC genes in a group of cysteine proteases have been widely studied in Leishmania species. In our analysis result, CPC gene in *L. martiniquensis* was found. CPC played a relevant role in the defending mechanism, by resisting killing by macrophages, as described in a previous report [61].

The phylogenetic analysis showed that *L. martiniquensis* is closely related with the latest *L. martiniquensis* (LU_Lmar_1.0) reference genome in the NCBI

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

database. However, there is a previous report about comparative genomics of *L. mundinia* (*L. martiniquensis*) in 2019 [62]. This research reported genomes of *L. mundinia*. Unfortunately, the protein sequences of predicted genes are not available for download. For this reason, the phylogenetic result was not including protein dataset from the *L. mundinia* genome in 2019.

Conclusions

In this study, *L. martiniquensis* genomic DNA was successfully sequenced and assembled to chromosomes. A total of 30,784,469 bases in 36 chromosomes of the *L. martiniquensis* genome were analyzed. The analysis results showed that the general features of *L. martiniquensis* were similar to other *Leishmania* species, including chromosome sizes, the number of protein-coding genes and the GC contents. In addition, the results of COG functional annotation were shown to be similar to other *Leishmania* species. In the virulence factor gene prediction result, 359 potential candidate virulence factor genes were found in this study. Most predicted virulence factor genes were related to RNA processing and modification function. However, candidate potential virulence factor genes should be validated in a further study using experimental study.

Data Availability

All data analysed during this study are included within this article. In addition, The DNA sequences were deposited in Sequence Read Archive (SRA) data of NCBI server (BioProject ID PRJNA674467).

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding Statement

This work was supported by Thailand Research Fund (TRF) [grant number RSA6180035]; National Science and Technology Development Agency (NSTDA) [grant number P-17-51377] and Program Management Unit for Human Resources &

Institutional Development, Research and Innovation [grant number B16F630071] and Thailand Science Research Innovation (TSRI) [grant number FRB640001].

Acknowledgments

The authors would like to acknowledge the supports from Graduate School, Faculty of Science and Faculty of Medicine, Chulalongkorn University (the 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship; the 90th Anniversary of Chulalongkorn University Ratchadaphiseksomphot Endowment Fund.

Supplementary Materials

Supplementary material 1: The annotation lists of the candidate predicted virulence factor genes from ProtVirDb.

Supplementary material 2: The annotation lists of the candidate predicted virulence factor genes from PHI base.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Comparative genome characterization of *Leptospira interrogans* from mild and severe leptospirosis patients

(Submitted to Genomics & Informatics)

Songtham Anuntakarun[1], Vorthon Sawaswong[1], Rungrat Jitvaropas[2], Kesmanee Praianantathavorn[3], Witthaya Poomipak[4], Yupin Suputtamongkol[5], Chintana Chirathaworn[6], Sunchai Payungporn[2,7,*]

[1] Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok, Thailand

[2] Division of Biochemistry, Department of Preclinical Science, Faculty of Medicine, Thammasat University, Pathum Thani, Thailand

[3] Department of Biochemistry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[4] Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

[5] Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand.

[6] Department of Microbiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.

[7] Research Unit of Systems Microbiology, Chulalongkorn University, Bangkok, Thailand

* Corresponding Author:

Associate Professor Sunchai Payungporn, PhD.

Research Unit of Systems Microbiology, Department of Biochemistry,

Faculty of Medicine, Chulalongkorn University

Email address: sp.medbiochemcu@gmail.com

Tel: +66 89 108 3179

CHULALONGKORN UNIVERSITY

Abstract

Leptospirosis is a zoonotic disease caused by spirochetes from the genus *Leptospira*. In Thailand, *Leptospira interrogans* is a major cause of leptospirosis. Leptospirosis patients present with a wide range of clinical manifestations from asymptomatic, mild infections to severe illness involving organ failure. For better understanding the difference between *Leptospira* isolates causing mild and severe leptospirosis, illumina sequencing was used to sequence genomic DNA in both serotypes. DNA of *Leptospira* isolated from 2 patients, one with mild and another with severe symptoms, were included in this study. The paired-end reads were removed adapters and trimmed with Q30 score using Trimmomatic. Trimmed reads were constructed to contigs and scaffolds using SPAdes. Cross-contamination of scaffolds was evaluated by ContEst16s. Prokka tool for bacterial annotation was used to annotate sequences from both *Leptospira* isolates. Predicted amino acid sequences from Prokka were searched in EggNOG and David gene ontology database to characterize gene ontology. In addition, *Leptospira* from mild and severe patients, that passed the criteria e-value < 10e^{-5} from blastP against virulence factor database, were used to analyze with Venn diagram. From this study, we found 13 and 12 genes that were unique in the isolates from mild and severe patients, respectively. The 12 genes in the severe isolate might be virulence factor genes that affect disease severity. However, these genes should be validated in further study.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Keywords: Genome annotation, Leptospirosis, *Leptospira interrogans*, virulence factor genes

Introduction

Leptospirosis is a worldwide zoonotic disease that influences humans and animals worldwide [63]. It is a zoonosis caused by bacteria in the genus *Leptospira. Leptospira* can be clustered in three groups including pathogenic, intermediate pathogenic and saprophytic groups. The various clinical manifestations are caused by the pathogenic and intermediate groups, while the saprophytic group does not cause the disease in humans or animals [64]. Human leptospirosis can be acquired by contact with the urine of infected animals or soil and water contaminated with *Leptospira* [63]. There are two chromosomes in the *Leptospira* species with a cumulative length ranging from 3.9 to 4.6 Mb. This variability in the genome length confers the bacteria with an ability to live within diverse environments and adapt to a wide range of hosts [65]. Approximately 60% of the functional genes that affect the unique pathogenic mechanisms caused by *Leptospira* are unknown [66].

In 2017, the 100K Pathogen Genome Project was established with internationalization coprojects by many countries, including China, South Korea, and Mexico. This project provides various pathogen draft genomes from many areas, and which include human and animal diseases, food, environmental reservoirs of those pathogens and wildlife. Several species such as *Campylobacter*, *Shigella*, *Salmonella*, *Listeria*, *Helicobacter*, and *Vibrio* are currently involved in the project [4]. Virulence genes code for virulence factors that are essential for successful infection and pathogenesis, such as invasion, colonization, adaptation in host environments, immune evasion and tissue damage. Comparison of genomes from microorganisms causing the variety of symptoms provides insight into the mechanisms of microbial infection and pathogenesis. The virulence factor database (VFDB) [18] provides up-to-date information of virulence factor genes from various bacterial pathogens.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

In this study, we compared the genomes of *Leptospira* isolated in Thailand from both mild and severe leptospirosis patients. The data provide insight into the genomic characteristics of *Leptospira interrogans*. In addition, virulence factor genes were analyzed using bioinformatics approaches. This research provides information for therapeutic and vaccine development for leptospirosis.

Methods

Isolation of *Leptospira*

Leptospira isolated from human patients in this study were obtained from the Department of Medicine, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand. The protocol was approved by the Ethical Committee of the Ministry of Public Health, Royal Government of Thailand. One isolate was from a mild leptospirosis patient, while the other was from a patient presenting with a severe clinical manifestation. Leptospirosis was laboratory confirmed by detecting IgM antibody to *Leptopsira* by indirect immunofluorescent assay (IFA) and PCR for *lipL32* gene detection. Briefly, the mild case was a 25-year-old male, admitted to Loei Hospital on 21 August 2001. He presented with three days of fever, headache and myalgia. *Leptospira* detected from his blood culture was identified as Serogroup Pyrogenase. The severe case was a 59-year-old male admitted to Nakhon Ratchasima Hospital on 2 July 2012. He presented with septic shock and died within 48 hours of admission. He had a history of 3 days of fever and developed hypotension, jaundice, acute renal failure and upper GI hemorrhage. He had no hemoptysis or acute respiratory distress syndrome (ARDS).

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Library preparation

DNA was extracted from the leptospires grown in EMJH medium using QIAamp DNA mini kit (QIAGEN, USA) according to the manufacturer's instructions. In the fragmentation step, a Covaris M220 focused-ultrasonicator (Covaris, Brighton, UK), with 20% duty factor, 50 unit of peak incident power (W), and 200 cycles per burst for 150 seconds, was used to fragment 1 µg of DNA. In the DNA library preparation, the fragmented DNA was prepared based on the TruSeq DNA LT Sample Prep Kit (Illumina, California, USA) following the manufacturer's instructions. Then, AMPure XP beads (Beckman Coulter, USA) was used to perform clean up and size selection of the DNA library. The concentration of the DNA library was measured using the KAPA Library Quantification Kit (Kapa Biosystems, Massachusetts, USA). The DNA library was diluted to 6 pM. Finally, the diluted DNA library was paired-end sequenced (2x150 bp) with the MiSeq platform (Illumina, California, USA), using MiSeq Reagent Kits V2 (300 cycles) according to the standard protocol.

Quality filter and Genome assembly

MIseq was used to sequence the mild and severe strains of *Leptospira* isolated from the Thai patients. Trimmomatic-0.38 [41] was used to trim and remove low quality reads using default parameter. *De novo* assembly was performed in both strains using SPAdes-3.13.0 [42]. All scaffolds were checked for contamination of 16S rRNA using the ContEST16s database [67]. The Artermis comparison tool (ACT) [43] was used to perform alignment of assembled sequences to a reference genome using *L. interrogans* serovar Lai 56601 as a reference. The DNA sequences were deposited in the Sequence Read Archive (SRA) data of NCBI server (BioProject ID PRJNA716760).

Gene prediction and functional annotation

In the gene prediction step, Prokka 1.13.3 [8] was used to predict genes in the mild and severe *Leptospira* genome. Putative protein coding sequences from Prokka were performed in the functional annotation. The integration of annotation data from the EggNOG database version 1.0.3 [17] and the David gene ontology database [68] represent the function of predicted genes including the cluster of orthologous groups of proteins (COGs), KEGG pathway [46], and GO annotation.

Prediction of virulence factor gene

The putative protein coding sequences were searched using blastP with the virulence factor database (VFDB). The criteria for the determination of candidate virulence sequences was based on an e-value of 10e^{-5}. Venn diagram analysis was used to find unique candidate virulence sequences in a specific strain. Lipoprotein prediction in gram-negative bacteria was performed using LipoP 1.0 [69].

Identification of phages in mild and severe *Leptospira* genomes

PHASTER (PHAge Search Tool Enhanced Release) [70] was performed to identify phages in both the mild and severe genomes.

Results

Genome characteristics of mild and severe strain

There was a total of 5,439,790 and 2,162,355 reads with 150 bp paired-end library using mean Phred score (Q) > 30 in mild and severe strain, respectively. The number of scaffolds more than 500 bp are 165 in the mild strain and 309 in the severe strain. The overview of fastq and de novo data assembly of mild and severe strains is shown in Table 7. After merging and ordering scaffolds with ACT, there are 3,947 and 297 predicted genes in the final assembly of chromosome 1 (4.70 Mb) and chromosome 2 (0.36 Mb), respectively. In the severe strain, there are 4,373 and 236 predicted genes in the final assembly of chromosome 1 (5.14 Mb) and chromosome 2 (0.37 Mb), respectively. The large variations of the CG content regions in the genome may be caused by being over- or under-fragmented during the library construction. The percentage of GC content in *Leptospira interogans* ranges from 35-41% [71]. The mild genome had an average GC content of 35%, and the severe genome had an average GC content of 37%.

Chulalongkorn University

Table 7 Characteristics of mild and severe data and *de novo* assembly

	Mild	Severe
Length	150bp	150bp
Raw reads	5,989,479	2,590,133
Q30 reads	5,439,790	2,162,355
Number of scaffolds	619	1,210
Number of scaffolds (>500bp)	165	309
N50	97,013	185,969

From Clusters of Orthologous Groups of proteins (COGs) analysis of mild and severe strains, the top three categories included function unknown, membrane/envelope biogenesis and signal transduction mechanisms, as indicated in Figure 7. For the KEGG pathway analysis, the top three pathways included metabolic pathways, biosynthesis of amino acids, and 2 -oxocarboxylic metabolism acid, as shown in Figure 8. Functional annotation is the process of collecting information about the function of genes. The Gene Ontology (GO) system [47] was used in this study. There are three distinct categories in gene ontology, namely molecular function, cellular component and biological process. The results of GO analysis given in Figure 9 - 11 show that the top three molecular functions are sigma factor activity, magnesium ion binding, and structural constituent of ribosome. The top three cellular components are cytoplasm, ribosome, and large ribosomal subunit. The top three biological processes are DNA-templated transcription/initiation, translation, and peptidoglycan biosynthetic process. There is no significant difference between mild and severe strains from COGs, KEGG pathway and GO analysis.

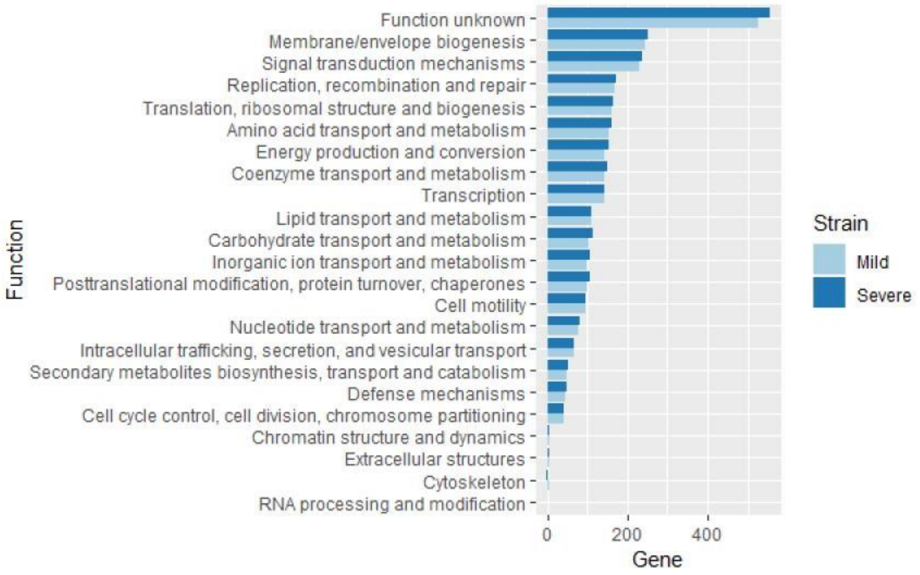


Figure 7 Comparison of Clusters of Orthologous Groups of proteins (COGs) between mild and severe strains

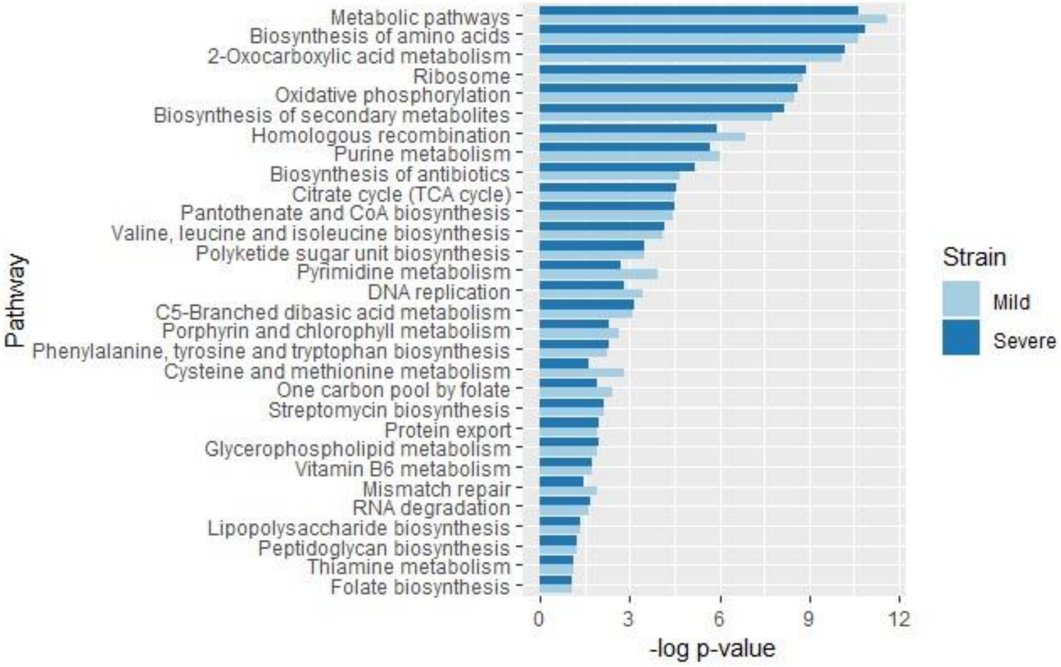


Figure 8 Comparison of KEGG pathway between mild and severe strains

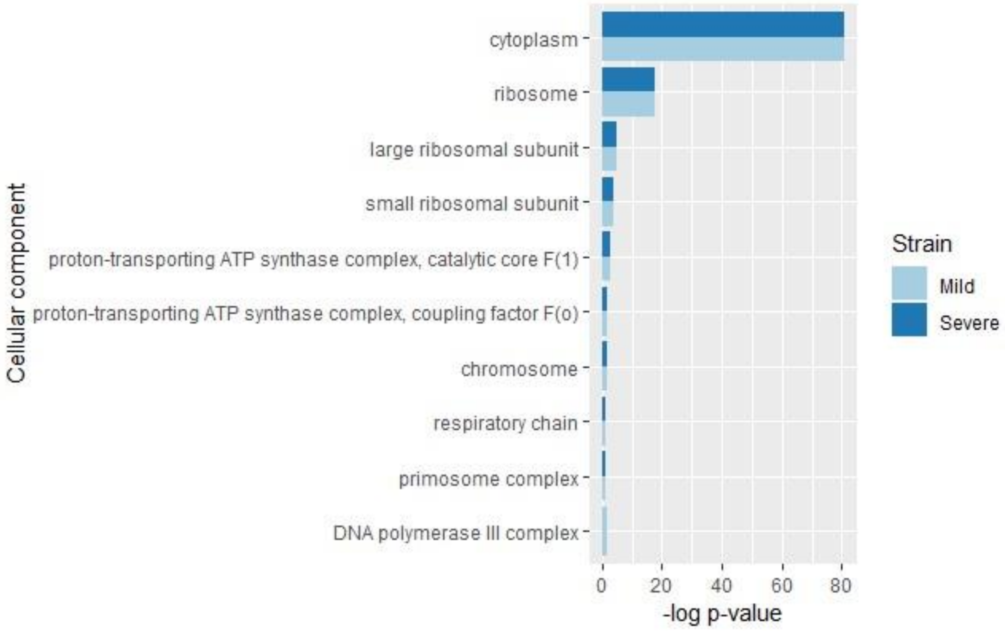


Figure 9 Comparison of cellular component between mild and severe strains

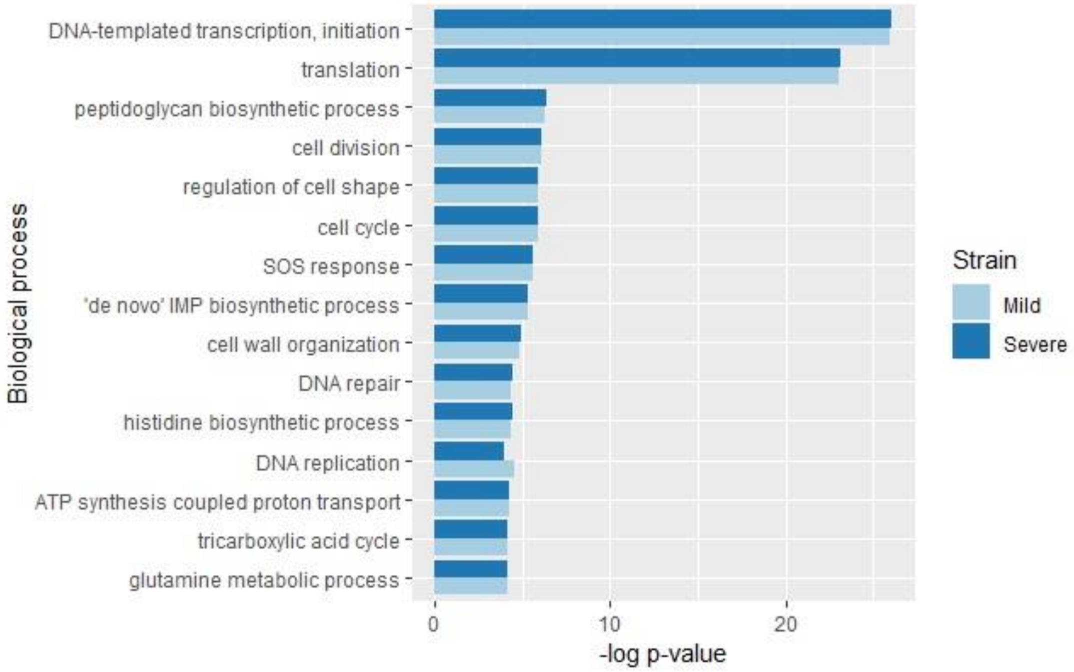


Figure 10 Comparison of biological process between mild and severe strains

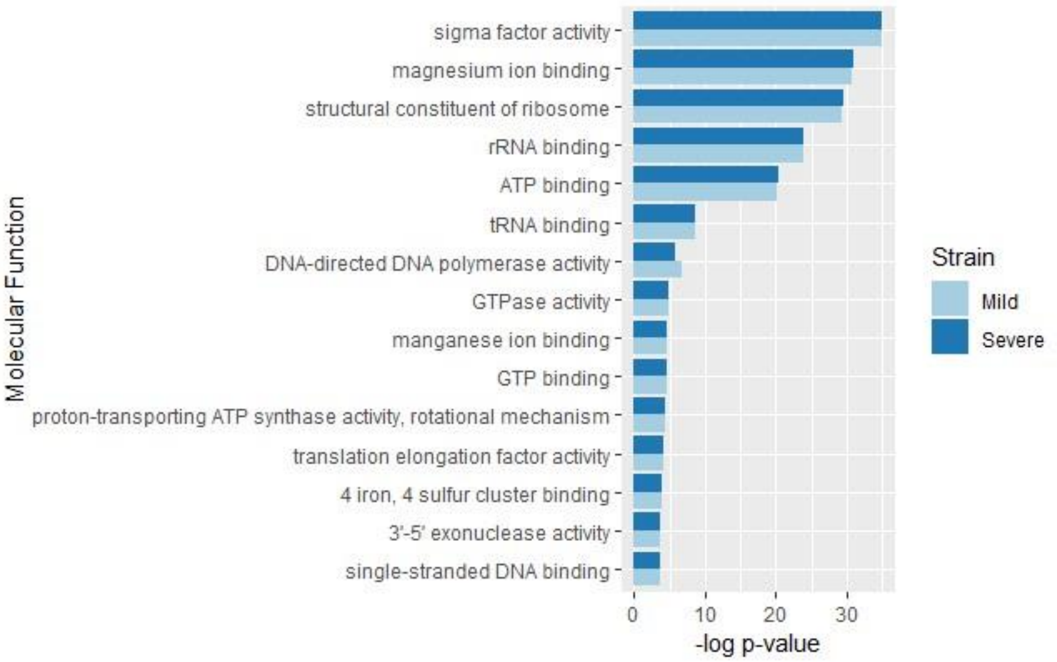


Figure 11 Comparison of molecular function between mild and severe strains

Putative virulence factor analysis

A total of 4,244 and 4,699 predicted genes in mild and severe strains, respectively from Pokka were used to identify virulence factor gene with virulence factor database (VFDB). The 162 and 161 virulence factor genes were found in mild and severe stains, respectively using blastP with an e-value < $10e^{-5}$. Venn diagram analysis was used to compare virulence factor genes between mild and severe strains. Figure 12A shows that 12 genes and 10 genes, respectively, of chromosome 1 were found in only the mild strain and only the severe strain. In chromosome 2, one gene was found in the mild strain only and two genes were found in the severe strain only (Figure 12B). The gene lists that were discovered in only the mild strain included AfaG-VII, neuA/flmD, rhmA, dapH, yhbX, murB, ahpC, flhB, LA_3103, nuc, PS_PT04340, ipaH2.5 and rfaK. Meanwhile, the gene lists found in only the severe strain consist of mntB, iga, flgG, proC, kdnB, neuA_1, neuA_2, pyrB, C8J_1334, rfbB, gtf1 and hemB. The description of virulence factor genes is shown in Table 8 and 9. In Figure 12C, the regions of virulence factor genes were mapped into chromosomes of mild and severe strains. There are many different regions of virulence factor genes found in mild and severe strains, especially in chromosome 1. In chromosome 2 of the severe strain, the group of virulence factor genes were located in the range of 4.8 -5.2 Mb. In addition, nearby virulence factor genes might exhibit co-expression or regulation. However, nearby virulence factor genes will be studied further.

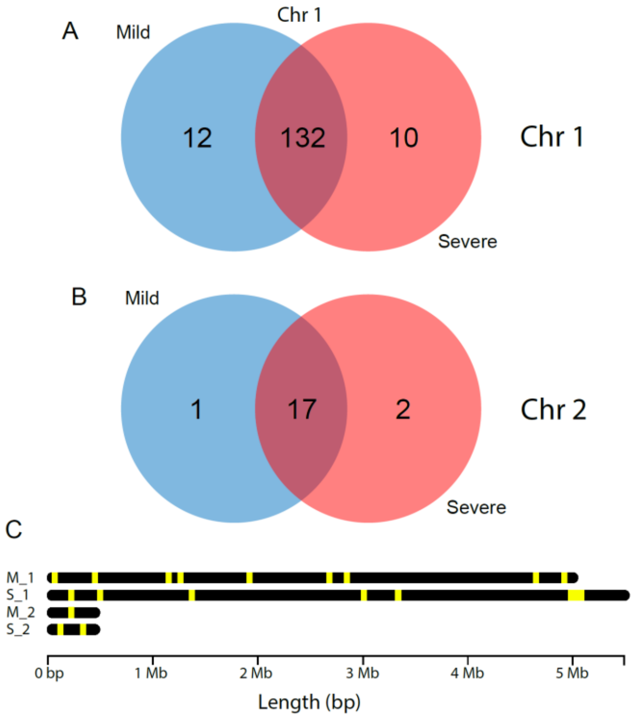


Figure 12 Comparison of virulence factor genes between mild and severe strains. (A) Venn diagram analysis between mild and severe strains in chromosome 1. (B) Venn diagram analysis between mild and severe strains in chromosome 2. (C) Comparison region of predicted virulence factor genes in each chromosome of both mild and severe strains (M_1: chromosome 1 in mild strain, M_2: chromosome 2 in mild strain, S_1: chromosome 1 in severe strain and S_2 chromosome 2 in severe strain. Yellow stripe in the black bar: region of virulence factor genes).

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Table 8 Description of predicted virulence factor genes in mild strain

Gene	Description
AfaG-VII	Afimbrial adhesin
neuA/flmD	CMP-N-acetylneuraminic acid synthetase
rhmA	2-keto-3-deoxy-L-rhamnonate aldolase
dapH	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase
yhbX	outer membrane protein YhbX
murB	UDP-N-acetylenolpyruvoylglucosamine reductase
ahpC	Alkyl hydroperoxide reductase C
flhB	Flagellar biosynthetic protein FlhB
LA_3103	Fibronectin-binding protein
nuc	Thermonuclease
PS_PTO4340	insecticidal toxin protein, putative
ipaH2.5	invasion plasmid antigen
rfaK	alpha 1,2 N-acetylglucosamine transferase

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 9 Description of predicted virulence factor genes in severe strain

Gene	Description
mntB	Manganese transport system membrane protein MntB
iga	IgA-specific serine endopeptidase
flgG	flagellar basal-body rod protein FlgG
proC	Pyrroline-5-carboxylate reductase
kdnB	3-deoxy-alpha-D-manno-octulosonate 8-oxidase
neuA_1	N-acylneuraminate cytidylyltransferase
neuA_2	CMP-N,N'-diacetyllegionaminic acid synthase
pyrB	Aspartate carbamoyltransferase catalytic subunit
C8J_1334	hypothetical protein
rfbB	dTDP-glucose 4,6-dehydratase
gtf1	Glycosyltransferase Gtf1
hemB	Delta-aminolevulinic acid dehydratase

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Phage analysis

For phage investigation, prophage sequences in mild and severe strain genomes were identified and annotated using PHASTER. Prophages play an important role in the evolution of the bacterial host and are commonly found in the bacterial genome [72]. In our results, there is no phage in either mild and severe genomes. However, the size ranges of incomplete phages from 6.9 - 11.3 Kbp were detected in both strains. PHAGE_Synech_S_CAM7_NC_031927, PHAGE_Sphing_PAU_NC_019521, PHAGE_Synech_ACG_2014b_NC_027130, PHAGE_Bacill_Finn_NC_020480, PHAGE_Psychr_pOW20_A_NC_020841 and PHAGE_Shigel_Sf6_NC_005344 were found in the mild genome. Moreover, PHAGE_Acinet_Acj9_NC_014663, PHAGE_Bacill_SP_15_NC_031245, PHAGE_Synech_S_CAM7_NC_031927, PHAGE_Sphing_PAU_NC_019521 ,and PHAGE_Synech_ACG_2014f_NC_026927 were found in the severe genome. Almost all of the incomplete prophages were similar to other *leptospira* species that contained incomplete phages with sizes ranging from 4.1 to 13.8 Kbp [73]. However, PHAGE_Acinet_Acj9_NC_014663 which was found in the severe strain, is the one multiple-drug resistant species [74].

Lipoprotein analysis

Lipoproteins of bacteria are a set of membrane proteins. There are many functions in the role of pathogenesis and host-pathogen interaction, especially the functions of surface adhesion and initiation of inflammatory processes through translocation of virulence factors in the host cytoplasm [75]. In our study, we used 32 and 67 unique genes in mild and severe strains, respectively, from eggNOG annotation to predict lipoprotein signal peptide using LipoP 1.0. This software can discriminate between lipoprotein and other signal peptides. The prediction was separated into 4 groups, including cytoplasmic, signal peptide, N-terminal transmembrane helix and lipoprotein signal peptide. In addition, this result in Figure 13 showed that a protein sequence was assigned to a lipoprotein signal peptide found in the severe strain only.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

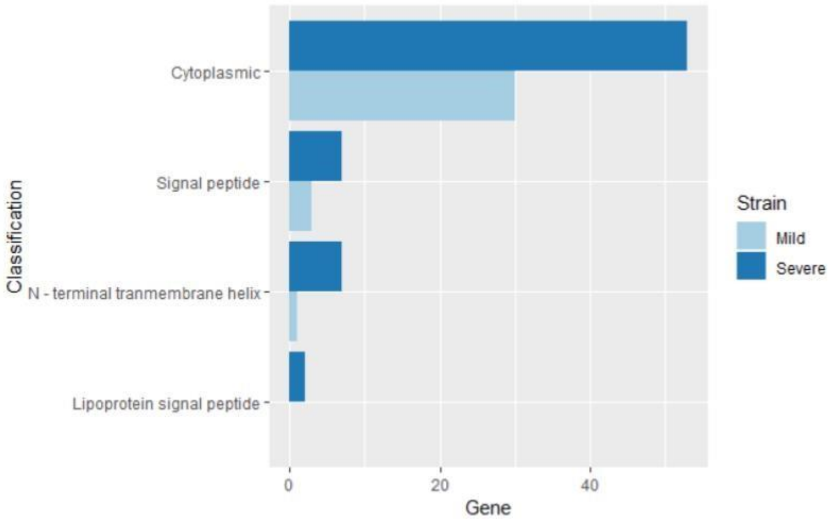


Figure 13 Comparison of lipoprotein predicted gene between mild and severe strains. The class of prediction from LipoP 1.0 was separated into 4 groups including Cytoplasmic, Signal peptide, N- terminal transmembrane helix and Lipoprotein signal peptide.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Discussion

LipoP1.0 predicts lipoproteins and discriminates between lipoprotein signal peptides and other signal peptides in Gram-negative bacteria using a Hidden Markov model (HMM). They report that the accuracy performance of prediction in gram-negative bacteria is 96.8%. Another lipoprotein prediction is called LIPOPREDICT which predicts signal peptides using a support vector machine [76]. The accuracy of this tool is 97%. Support vector machine has a similar performance to HMM. We would like to use LIPOPREDICT to predict lipoproteins in our genomes. Unfortunately, LIPOPREDICT is not available so far.

IgA-specific serine endopeptidase or IgA protease is secreted by gram-negative bacteria. This enzyme plays an important role in human antibodies. They can specifically cleave IgA, which provides an antibody for defending the mucosal surface [77]. The inactivation of IgA protease might have the potential to reduce bacterial colonization on mucosal surfaces [78].

Aminoglycosides are broad-spectrum antibiotics that are used in Gram-negative and Gram-positive organisms [79]. Many reports showed that *leptospira* are sensitive to aminoglycosides [80-81]. dTDP-glucose-4,6-dehydratase genes were related in a gene cluster in an aminoglycoside antibiotics producer [82].

In bacteria, metal ions play an important role in survival in their host environment. Bacteria which cannot maintain proper homeostasis of metals are less virulent [83]. In many biological processes metal ions are needed as metalloprotein materials, which function as enzyme cofactors or structural elements. Manganese (Mn) is one important example. Many bacteria require manganese with eukaryotic host cells to form pathogenic or symbiotic interactions [84]. Currently, there is evidence that the invading microbe uses Mn as the main micronutrient to avoid the effects of host-mediated oxidative stress and thus plays a significant role in the human host's tolerance to pathogenic bacteria [85]. In our study, we found manganese transport system membrane protein MntB (mntB) in the severe *leptospira* strain. This gene encodes transmembrane protein. The mntB gene is part of the ABC transporter system for manganese that mediates the movement of various substrates from microbes to humans across different biological membranes [86]. The lack of the mntB gene might affect the homeostasis of metal in bacteria that are less virulent.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

The flagellum consists of three main sections, including a flagellar filament, a hook complex, and a basal body in both gram-negative and gram-positive bacteria. There are many genes related to flagellar biosynthetic protein such as flhA flhB [87-88]. The results showed that flhB was found in the mild strain. This result came from blastP with a virulence factor database. However, flhB was also found in the severe strain from prokka annotation. In this case, some genes in the mild strain are similar to the flhB gene in other species of bacteria in the virulence factor database.

Conclusion

In this study, two strains of *Leptospira* spp. isolated from mild and severe Thai patients were compared. Our analysis showed 3,947 and 297 predicted genes in the final assembly of chromosome 1 (4.70 Mb) and chromosome 2 (0.36 Mb), respectively, in the mild strain. In addition, there are 4,373 and 236 predicted genes in the final assembly of chromosome 1 (5.14 Mb) and chromosome 2 (0.37 Mb), respectively, in the severe strain. The difference of virulence factor genes was found in both strains. Our results focus on predicting virulence factor genes in the severe strain that is not found in the mild strain. The virulence factor genes in the severe strain are only related to host immune response, and survival in the host environment might be the vital virulence factor genes. However, these genes should be validated in further study.

ORCHID

Songtham Anuntakarun: https://orcid.org/0000-0002-6849-0523

Vorthon Sawaswong: https://orcid.org/0000-0003-2805-6690

Rungrat Jitvaropas: https://orcid.org/0000-0001-7555-0048

Kesmanee Praianantathavorn: https://orcid.org/0000-0002-5368-3015

Witthaya poomipak: https://orcid.org/0000-0002-3282-7219

Yupin Suputtamongkol: https://orcid.org/0000-0001-7324-1698

Chintana Chirathaworn: https://orcid.org/0000-0002-2131-1815

Sunchai Payungporn: https://orcid.org/0000-0003-2668-110X

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Authors' Contribution

Conceptualization: SP, SA. Data curation: CC, YS. Formal analysis: SA, VS. Funding acquisition: SP. Methodology: SA, KP, WP. Writing - original draft: SA. Writing - review & editing: SP, JK, CC, RJ.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgements

The authors would like to acknowledge the supports from Graduate School, Faculty of Science and Faculty of Medicine, Chulalongkorn University (the 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship; the 90th Anniversary of Chulalongkorn University Ratchadaphiseksomphot Endowment Fund).

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Part 3

CONCLUSION LIMITATION AND SUGGESTION

CONCLUSION

In this study, our pipelines can perform analysis both of eukaryotic and prokaryotic microorganisms. In assembly step, our pipelines used SPAdes tool for assembling short reads from *Leishmania martiniquensis* and *Leptospira interrogans*. In the gene prediction step, AUGUSTUS and PROKKA were performed in eukaryotic and prokaryotic microorganisms, respectively. The protein sequences from AUGUSTUS and PROKKA tools were used to discover insight information of sequences using BlastP and eggNOG-mapper with many public biological databases. Moreover, the eukaryotic and prokaryotic virulence factor gene databases (ProtVirDB and VFDB) were including in our pipelines. Finally, we hope these pipelines can be useful for researchers who need to analyze and get the insight into gene information in the microorganism.

LIMITATION

Our pipelines can perform analysis in prokaryotic and eukaryotic microorganisms using only illumina short reads. Some tools in our pipelines are not supported long- read sequencing, such as PacBio and Nanopore platform. In addition, our pipelines are not supported the Windows operating system.

SUGGESTION

In our pipelines, we suggest at least ~200GB space for install various databases and ~64GB of ram for using SPAdes assembly. The requirement of ram depends on the size of fastq files. If the size is less than 1GB, ~32GB of ram works properly. In addition, we suggest at least 16 CPU cores for SPAdes assembly, BlastP and eggNOG-mapper tools.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

REFERENCES

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. 2013;98(6):236–8.

2. Wheeler DA, Wang L. From human genome to cancer genome: The first decade. Genome Res. 2013;23(7):1054–62.

3. Friedman AA, Letai A, Fisher DE, Flaherty KT. Precision medicine for cancer with next-generation functional diagnostics. Nat Rev Cancer. 2015;15(12):747–56.

4. Weimer BC. 100K pathogen genome project. Genome Announc. 2017;5(28).

5. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008;24(3):142–9.

6. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, et al. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics. 2012;11(1):25–37.

7. Richardson EJ, Watson M. The automatic annotation of bacterial genomes. Brief Bioinform. 2013;14(1):1–12.

8. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.

9. Stanke M, Morgenstern B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33(SUPPL. 2):465–7.

10. Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11.

11. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100–8.

12. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11–6.

13. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–6.

14. Kolbe DL, Eddy SR. Fast filtering for RNA homology search. Bioinformatics. 2011;27(22):3102–9.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

15. Carbon S, Douglass E, Dunn N, Good B, Harris NL, Lewis SE, et al. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(1):330–8.

16. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(1):353–61.

17. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(1):309–14.

18. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: A reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33(DATABASE ISS.):325–8.

19. Ramana J, Gupta D. ProtVirDB: A database of protozoan virulent proteins. Bioinformatics. 2009;25(12):1568–9.

20. Saha S, Raghava GPS. VICMpred: An SVM-based method for the prediction of functional proteins of gram-negative bacteria using amino acid patterns and composition. Genomics, Proteomics Bioinforma. 2006;4(1):42–7.

21. Garg A, Gupta D. VirulentPred: A SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics. 2008;9:1–12.

22. Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, et al. Sequence-Based Prediction of Type III Secreted Proteins. PLoS Pathog. 2009;5(4).

23. Zou, Lingyun, Chonghan Nan and FH. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. Bioinformatics. 2013;29(24):3135–42.

24. Gupta A, Kapil R, Dhakan DB, Sharma VK. MP3: A software tool for the prediction of pathogenic proteins in genomic and metagenomic data. PLoS One. 2014;9(4).

25. Jariyapan N, Daroontum T, Jaiwong K, Chanmol W, Intakhan N, Sor-Suwan S, et al. Leishmania (Mundinia) orientalis n. sp. (Trypanosomatidae), a parasite from Thailand responsible for localised cutaneous leishmaniasis. Parasites and Vectors. 2018;11(1):3–11.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

26. Espinosa OA, Serrano MG, Camargo EP, Teixeira MMG, Shaw JJ. An appraisal of the taxonomy and nomenclature of trypanosomatids presently classified as Leishmania and Endotrypanum. Parasitology. 2018;145(4):430–42.

27. Barratt J, Kaufer A, Peters B, Craig D, Lawrence A, Roberts T, et al. Isolation of Novel Trypanosomatid, Zelonia australiensis sp. nov. (Kinetoplastida: Trypanosomatidae) Provides Support for a Gondwanan Origin of Dixenous Parasitism in the Leishmaniinae. PLoS Negl Trop Dis. 2017;11(1):1–26

28. Sukmee T, Siripattanapipong S, Mungthin M, Worapong J, Rangsin R, Samung Y, et al. A suspected new species of Leishmania, the causative agent of visceral leishmaniasis in a Thai patient. Int J Parasitol. 2008;38(6):617–22.

29. Chiewchanvit S, Tovanabutra N, Jariyapan N, Bates MD, Mahanupab P, Chuamanochan M, et al. Chronic generalized fibrotic skin lesions from disseminated leishmaniasis caused by Leishmania martiniquensis in two patients from northern Thailand infected with HIV. Br J Dermatol. 2015;173(3):663–70.

30. Pothirat T, Tantiworawit A, Chaiwarith R, Jariyapan N, Wannasan A, Siriyasatien P, et al. First Isolation of Leishmania from Northern Thailand: Case Report, Identification as Leishmania martiniquensis and Phylogenetic Position within the Leishmania enriettii Complex. PLoS Negl Trop Dis. 2014;8(12).

31. Siriyasatien P, Chusri S, Kraivichian K, Jariyapan N, Hortiwakul T, Silpapojakul K, et al. Early detection of novel Leishmania species DNA in the saliva of two HIV-infected patients. BMC Infect Dis. 2016;16(1):1–7.

32. Phumee A, Chusri S, Kraivichian K, Wititsuwannakul J, Hortiwakul T, Thavara U, et al. Multiple Cutaneous Nodules in an HIV-Infected Patient. PLoS Negl Trop Dis. 2014;8(12):6–8.

33. Phumee A, Jariyapan N, Chusri S, Hortiwakul T, Mouri O, Gay F, et al. Determination of anti-leishmanial drugs efficacy against Leishmania martiniquensis using a colorimetric assay. Parasite Epidemiol Control. 2020;9:e00143.

34. Lypaczewski P, Hoshizaki J, Zhang WW, McCall LI, Torcivia-Rodriguez J, Simonyan V, et al. A complete Leishmania donovani reference genome identifies novel genetic variations associated with virulence. Sci Rep. 2018;8(1):1–14.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

35. Coughlan S, Taylor AS, Feane E, Sanders M, Schonian G, Cotton JA, et al. Leishmania naiffi and Leishmania guyanensis reference genomes highlight genome structure and gene evolution in the Viannia subgenus. bioRxiv. 2017;

36. Da Fonseca Pires S, Fialho LC, Silva SO, Melo MN, De Souza CC, Tafuri WL, et al. Identification of virulence factors in leishmania infantum strains by a proteomic approach. J Proteome Res. 2014;13(4):1860–72.

37. Silva-Almeida M, Pereira BAS, Ribeiro-Guimarães ML, Alves CR. Proteinases as virulence factors in Leishmania spp. infection in mammals. Parasites and Vectors. 2012;5(1):1–10.

38. Chusri S, Hortiwakul T, Silpapojakul K, Siriyasatien P. Case report: Consecutive cutaneous and visceral leishmaniasis manifestations involving a novel Leishmania species in two HIV patients in Thailand. Am J Trop Med Hyg. 2012;87(1):76–80.

39. Phumee A, Kraivichian K, Chusri S, Noppakun N, Vibhagool A, Sanprasert V, et al. Short report: Detection of Leishmania siamensis DNA in saliva by polymerase chain reaction. Am J Trop Med Hyg. 2013;89(5):899–905.

40. Simon A. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010.

41. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.

43. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: The Artemis comparison tool. Bioinformatics. 2005;21(16):3422–3.

44. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. bioRxiv 2021;5:2021.06.03.446934.

45. Galperin MY, Makarova KS, Wolf YI, Koonin E V. Expanded Microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43(1):261–9.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

46. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(1):353–61.

47. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Oncology (GO) database and informatics resource. Nucleic Acids Res. 2004;32(DATABASE ISS.):258–61.

48. Urban M, Cuzick A, Rutherford K, Irvine A, Pedro H, Pant R, et al. PHI-base: A new interface and further additions for the multi-species pathogen-host interactions database. Nucleic Acids Res. 2017;45(D1):D604–10.

49. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):1–14.

50. Letunic I, Bork P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49(1):293–6.

51. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7(539).

52. Llanes A, Restrepo CM, Vecchio G Del, Anguizola FJ, Lleonart R. The genome of Leishmania panamensis: Insights into genomics of the L. (Viannia) subgenus. Sci Rep. 2015;5:1–10.

53. Urrea DA, Duitama J, Imamura H, Álzate JF, Gil J, Muñoz N, et al. Genomic Analysis of Colombian Leishmania panamensis strains with different level of virulence. Sci Rep. 2018;8(1):1–16.

54. Miller DJ, Fort PE. Heat shock proteins regulatory role in neurodevelopment. Front Neurosci. 2018;12(11):1–15.

55. Kuppner MC, Gastpar R, Gelwer S, Nössner E, Ochmann O, Scharner A, et al. The role of heat shock protein (hsp70) in dendritic cell maturation: Hsp70 induces the maturation of immature dendritic cells but reduces DC differentiation from monocyte precursors. Eur J Immunol. 2001;31(5):1602–9.

56. MacAry PA, Javid B, Floto RA, Smith KGC, Oehlmann W, Singh M, et al. HSP70 Peptide Binding Mutants Separate Antigen Delivery from Dendritic Cell Stimulation. Immunity. 2004;20(1):95–106.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

57. Rawlings ND, Barrett AJ, Bateman A. MEROPS: The peptidase database. Nucleic Acids Res. 2009;38(SUPPL.1):227–33.

58. Valdivieso E, Dagger F, Rascón A. Leishmania mexicana: Identification and characterization of an aspartyl proteinase activity. Exp Parasitol. 2007;116(1):77–82.

59. Sajid M, McKerrow JH. Cysteine proteases of parasitic organisms. Mol Biochem Parasitol. 2002;120(1):1–21.

60. Vermelho AB, Branquinha MH, D'Ávila-Levy CM, Souza dos Santos AL, de Souza Dias EP, Nogueira de Melo AC. Biological roles of peptidases in trypanosomatids. Open Parasitol J. 2010;4(1):5–23.

61. Frame MJ, Mottram JC, Coombs GH. Analysis of the roles of cysteine proteinases of Leishmania mexicana in the host - Parasite interaction. Parasitology. 2000;121(4):367–77.

62. Butenko A, Kostygov AY, Sádlová J, Kleschenko Y, Bečvá T, Podešvová L, et al. Comparative genomics of Leishmania (Mundinia). BMC Genomics. 2019;20(1):1–12.

63. Bharti AR, Nally JE, Ricaldi JN, Matthias MA, Diaz MM, Lovett MA, et al. Leptospirosis: A zoonotic disease of global importance. Lancet Infect Dis. 2003;3(12):757–71.

64. Adler B, de la Peña Moctezuma A. Leptospira and leptospirosis. Vet Microbiol. 2010;140(3–4):287–96.

65. Picardeau M, Bulach DM, Bouchier C, Zuerner RL, Zidane N, Wilson PJ, et al. Genome sequence of the saprophyte Leptospira biflexa provides insights into the evolution of Leptospira and the pathogenesis of leptospirosis. PLoS One. 2008;3(2):1–9.

66. Xu Y, Zhu Y, Wang Y, Chang YF, Zhang Y, Jiang X, et al. Whole genome sequencing revealed host adaptation-focused genomic plasticity of pathogenic Leptospira. Sci Rep. 2016;6(2):1–11.

67. Lee I, Chalita M, Ha SM, Na SI, Yoon SH, Chun J. ContEst16S: An algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. Int J Syst Evol Microbiol. 2017;67(6):2053–7.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

68. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9).

69. Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. 2003;12(8):1652–62.

70. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016;44(W1):W16–21.

71. Jeremy Farrar, Peter J. Hotez, Thomas Junghanss, Gagandeep Kang, David Lalloo NJW. Manson's Tropical Diseases. Saunders; 2013.

72. Fortier LC, Sekulovic O. Importance of prophages to evolution and virulence of bacterial pathogens. Virulence. 2013;4(5):354–65.

73. Kurilung A, Keeratipusana C, Suriyaphol P, Hampson DJ, Prapasarakul N. Genomic analysis of Leptospira interrogans serovar Paidjan and Dadas isolates from carrier dogs and comparative genomic analysis to detect genes under positive selection. BMC Genomics. 2019;20(1):1–19.

74. Turner D, Ackermann HW, Kropinski AM, Lavigne R, Sutton JM, Reynolds DM. Comparative analysis of 37 Acinetobacter bacteriophages. Viruses. 2018;10(1):1–25.

75. Kovacs-Simon A, Titball RW, Michell SL. Lipoproteins of bacterial pathogens. Infect Immun. 2011;79(2):548–61.

76. Kumari SR, Kadam K, Badwaik R, Jayaraman VK. LIPOPREDICT: Bacterial lipoprotein prediction server. Bioinformation. 2012;8(8):394–8.

77. Plaut AG. The IgA1 proteases of pathogenic bacteria. Annu Rev Microbiol. 1983;37(1):603–22.

78. Mistry D, Stockley RA. IgA1 protease. Int J Biochem Cell Biol. 2006;38(8):1244–8.

79. Krause KM, Serio AW, Kane TR, Connolly LE. Aminoglycosides : An Overview. 2016;

80. Solly Faine, Ben Adler, Carole Bolin PP. Leptospira and leptospirosis (2nd eds). Melbourne, Victoria, Australia.; 1999.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

81. Kobayashi Y. Clinical observation and treatment of leptospirosis. J Infect Chemother. 2001;7(2):59–68.

82. Du Y, Li T, Wang YG, Xia H. Identification and functional analysis of dTDP-glucose-4,6-dehydratase gene and its linked gene cluster in an aminoglycoside antibiotics producer of Streptomyces tenebrarius H6. Curr Microbiol. 2004;49(2):99–107.

83. Hood MI, Skaar EP. Nutritional immunity: Transition metals at the pathogen-host interface. Nat Rev Microbiol. 2012;10(8):525–37.

84. Zeinert R, Martinez E, Schmitz J, Senn K, Usman B, Anantharaman V, et al. Structure–function analysis of manganese exporter proteins across bacteria. J Biol Chem. 2018;293(15):5715–30.

85. Lisher JP, Giedroc DP. Manganese acquisition and homeostasis at the host-pathogen interface. Front Cell Infect Microbiol. 2013;3(12):1–15.

86. Saier MH. Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya. Vol. 40, Advances in Microbial Physiology. 1998. 81–136 p.

87. Lambert A, Picardeau M, Haake DA, Sermswan RW, Srikram A, Adler B, et al. Flaa proteins in Leptospira interrogans are essential for motility and virulence but are not required for formation of the flagellum sheath. Infect Immun. 2012;80(6):2019–25.

88. Cheng C, Wang H, Ma T, Han X, Yang Y, et al. Flagellar basal body structural proteins FlhB, FliM, and FliY are required for flagellar-associated protein expression in Listeria monocytogenes. Front Microbiol 2018; 9:1–11.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

VITA

NAME	Songtham Anuntakarun
DATE OF BIRTH	1 November 1988
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	Bioinformatics and Computational Biology Program, Graduate school, Chulalongkorn University
HOME ADDRESS	3/93, Vipawadeerungsit 16 Rd, Bangkok, 10900
PUBLICATION	1. Anuntakarun S, Phumee A, Sawaswong V, Praianantathavorn K, Poomipak W, Jitvaropas R, Siriyasatien P, Payungporn S. Genome assembly and genome annotation of Leishmania martiniquensis isolated from a leishmaniasis patient in Thailand. (submitted) 2. Anuntakarun S, Sawaswong V, Jitvaropas R, Praianantathavorn K, Poomipak W, Jitvaropas R, Suputtamongkol Y, Chirathaworn C, Payungporn S. Comparative genome characterization of Leptospira interrogans from mild and severe leptospirosis patients. (submitted) 3. Anuntakarun S, Larbcharoensub N, Payungporn S, Reamtong O. Identification of genes associated with Kikuchi-Fujimoto disease using RNA and exome sequencing. Mol Cell Probes [Internet]. 2021;57(December 2020):101728.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University