Draft Genome Sequence of *Lactobacillus fermentum* 47-7 - A Probiotic Strain Isolated from an Infant

Atthaphon Konyanee* Dr.Viraphong Lulitanond** Dr.Kiatichai Faksri*** Dr.Nipaporn Sankuntaw****

ABSTRACT

The purpose of this study was to determine the draft genome sequence of *Lactobacillus fermentum* 47-7, a probiotic strain isolated from healthy infant in Khon Kaen, Thailand. Next generation whole-genome shotgun sequencing of *L. fermentum* 47-7 was done by using two platforms consists of Illumina Hiseq 2000 and Ion Torrent Personal Genome Machine (PGM). The merged sequence reads of both platforms showed the highest consensus sequences mapped with *L. fermentum* IFO 3956, the selected reference genome. The draft genome was submitted to Whole Genome Shotgun projects (WGS) of National Center for Biotechnology Information (NCBI) database. *L. fermentum* 47-7 genome with G+C content of 52.2% has 1,814 genes consisting of 1,626 coding sequences (CDS), 57 tRNAs, 15 rRNAs, 4 ncRNAs, 112 pseudogenes. This draft genome sequence provides useful data for further exploring various interesting characteristics of this organism.

Keywords: *Lactobacillus*, Draft genome, Probiotics

** Student, Master of Science Program in Medical Microbiology, Department of Microbiology, Faculty of Medicine, Khon Kaen University*

*** Professor, Department of Microbiology, Faculty of Medicine, Khon Kaen University*

**** Assistant professor, Department of Microbiology, Faculty of Medicine, Khon Kaen University*

***** Lecturer, Chulabhorn International College of Medicine, Thammasat University*

Introduction

Probiotics is defined as “live microorganisms which, when administered in adequate amounts, confer a health benefit on the host” (Archer, Halami, 2015). Over the past 10 years the probiotics have becomes a major topic of lactic acid bacteria (LAB) (Bao et al., 2010). Lactobacilli, a member of LAB group, have drawn significant attention because of theirs health-promoting effects such as increasing resistance to *Streptococcus pneumoniae* infection in mice model (Villena et al., 2005), stabilization of the intestinal microflora (Amy E. Foxx-Orenstein , and William D. Chey 2012), enhancing the immunosurveillance to prevent intestinal infections (Tsai et al., 2012), lowering cholesterol (Pereira et al., 2003) and anti-carcinogenic properties (Mital, Garg, 1995). *Lactobacillus* has a long history of safe use in fermented foods and generally regarded as safe (GRAS). With respect to *Lactobacillus fermentum*, it is gram-positive bacteria belonging to the genus *Lactobacillus*, Phylum Firmicutes. *L. fermentum* belong to the group of heterofermentative LAB, which ferment carbohydrates into major end product which is lactic acid and other compounds such as ethanol, acetic, butyrate, etc. It was isolated from various sources including traditionally fermented milk (Bao et al., 2010), fermented foods, plant materials, as well as gastrointestinal tract of humans (Archer, Halami, 2015). We have isolated *L. fermentum*, designated as *L. fermentum* 47-7, from feces of healthy infant in Khon Kaen, Thailand. This isolate showed a good *in vitro* probiotic properties including acid and bile salts tolerance, anti-microbial activity against food borne pathogen *Samonella typhimurium.* In addition, it has been successfully engineered to express green fluorescent protein (GFP) by using *L. casei* replicon-based expression vector (pRCEID-LC13.9) (Yotpanya et al., 2016). To further understanding this isolate, it is worth to genetically analyze and characterize this isolate through sequencing its genome by next generation sequencing (NGS). NGS has been used to explore the gut microbiome and their interaction with the human host, through the analysis of the microbial genome sequences. Several genomes of *Lactobacillus* species have been sequenced by NGS, such as the genome of *L. fermentum* 3872 has been sequenced and revealed genes associated with the mucus-binding protein and fibronectin-binding protein, both of which are important for adhesion to the host cell receptors in gastrointestinal tract (Karlyshev et al., 2013). Currently, a few of the draft genome sequence of *L. fermentum* are available in the databases. Thus, the draft genome sequence of *L. fermentum* 47-7 will be determined by NGS and compared with that of *L. fermentum* IFO 3956.

Objectives of the study

The aim of this study is to determine the draft genome sequence of *L. fermentum* 47-7 and to submit it to the Whole Genome Shotgun (WGS) projects.

Materials and Methods

Isolation genomic DNA of *L. fermentum* 47-7

The genomic DNA was extracted with the method described by De et al. (De et al., 2010) with some modifications. In brief, the bacterial cells from the overnight culture of *L. fermentum* 47-7 in MRS (de Man, Rogosa and sharpe) broth were harvested by centrifugation at 14,000 rpm at 4°C, washed three times with STE buffer pH 8.0

(10 mM Tris-HCl, 5 mM EDTA, 0.1% sucrose) and resuspended in 100 μl of freshly prepared lytic solution (10 mg/ml of lysozyme, 100 U/ml of mutanolysin, 100 μg/ml of RNaseA in STE buffer). The mixture was incubated at 37 °C for 3 hrs with intermittent shaking. After incubation, the mixture was added with 500 μl STE buffer, 50 μl of 10% SDS solution and 10 μl of proteinase K solution (20 mg/ml), mixed thoroughly and incubated at 55 °C for 1 hr. After incubation, the cell lysate was extracted once with buffered-saturated phenol (pH 8.0) and repeated with phenol:chloroform:isoamyl alcohol (25:24:1, v/v). The upper aqueous phase was harvested in a sterile microtube. The DNA in the supernatant was precipitated out with 0.8 volumes of isopropanol, washed once with 70% ethanol and air-dried. The final DNA pellet was dissolved in 50 μl Tris-EDTA (10:1, pH 8.0), and stored at 4 °C until use.

Sequencing of *L. fermentum* 47-7 genome by Ion Torrent Personal Genome Machine (PGM) and Illumina Hiseq 2000

The genomic DNA of *L. fermentum* 47-7 was subjected to next generation whole-genome shotgun sequencing using Ion Torrent PGM platform (Thermofisher, USA) with single-end mode with the read length of 200-300bp at Faculty of Medicine, Ramathibodi hospital, Mahidol university, Thailand. The library including fragmentation and purification was performed by using Ion Xpress^{TM} plus Fragment Library Kit and Agencourt$^{®}$ AMPure$^{®}$XP reagent respectively. The Ligation adaptors, nick-repair and purification was performed by using Ion Xpress^{TM} Barcode adaptors kits and Agencourt$^{®}$ AMPure$^{®}$XP reagent, respectively. Size-selected library was done with the E-gel$^{®}$ Sizeselect^{TM} Agarose gel. Purification of the size-selected DNA by using Agencourt$^{®}$ AMPure$^{®}$XP reagent. Amplification and purification of the library by using Ion plus fragment library kit. Bioanalyzer$^{®}$ was used to analyse the quality of genomic DNA fragmented library. Finally, DNA templates for sequencing was performed by using Ion PGM^{TM} OT2 200 kit v2 in order to sequencing in semiconductor Ion 314^{TM} chip by using Ion Torrent PGM^{TM} platform *(*Thermofisher, USA***)***. The Torrent Suite and Ion PGM^{TM} systems software were used to convert raw signal to base calls and generated output data to FASTQ files format for subsequent analysis**.** For sequencing by using Illumina Hiseq 2000 platform (Illumina, USA) with paired-ends mode with the read length of 101bp was done at Macrogen Inc. Company, Seoul, South Korea. The library preparation was performed by using TruSeq$^{®}$ DNA PCR-Free Library preparation in order to sequence refined DNA as short read fragmentation is performed. DNA fragments are repaired to blunt end using end-repair enzyme. “A” base is added to 3’-end (A-tailing). Ligation of “Y”-shaped adapters is performed at both ends. After library is annealed to the flowed cell, clusters are generated by bridge amplification method. In sequencing using SBS (Sequencing By Synthesis), signal intensity of four types fluorescence bases in each cluster of flow cells could be identified by laser and save on images. Base calling process of image files from sequencer is performed and FASTQ raw data files are generated.

Data processing and genome assembly and quality assessment

Each FASTQ file format was obtained from both sequencing platforms (Illumina Hiseq 2000 and Ion Torrent PGM). Short reads FASTQ preprocessing was performed by using Fastx-toolkit version 0.0.13 (Gordon, Hannon, 2016) before mapping of sequences to the reference genome. All sequence reads of each FASTQ file were filtered based on quality with setting parameters [-t20: nucleotides with lower quality will be trimmed from the end of the sequence, -l100: sequences shorter than this after trimming will be discarded] and filtering sequence reads based on quality with setting parameters [-q20: minimum quality score to keep, -p80: minimum percent of bases that must have -q quality]. The quality control checking of raw sequence data was performed by FastQC version 0.11.4 (Andrews, 2016). The filtered FASTQ files from both platform was mapped with reference genome of *L. fermentum* IFO 3956 genome, accession number NC_010610 by using Burrows-Wheeler aligner (BWA) version 0.7.5 to generate sequence alignment/map format (SAM) file (Li, Durbin, 2010). The SAM files from Illumina Hiseq 2000 and Ion Torrent PGM were converted into binary alignment/map format (BAM) file. The BAM files from Illumina Hiseq 2000 and Ion Torrent PGM were merged by Picard tool version 2.5.0 in order to generate the merged BAM file format (Broadinstitute, 2016). The merged BAM file was performed PCR duplicate marking process by Genome Analysis Toolkit (GATK) version 3.3 to ignore duplicates in subsequent processing (Auwera et al., 2013; DePristo et al., 2011;McKenna et al., 2010). Afterward, the local realignment and base quality score recalibration (BQSR) were performed by using Genome Analysis Toolkit (GATK) (Auwera et al., 2013;DePristo et al., 2011; McKenna et al., 2010). The consensus sequence was generated from the merged BAM file by using Samtools version 0.1.19 with setting parameters [-q30: minimum mapping quality for an alignment to be used, -Q30: minimum base quality for a base to be considered, -C50: coefficient for downgrading mapping quality for reads containing excessive mismatches, the recommended value for BWA is 50], bcftools version 0.1.17 and vcfutils.pl vcf2fq with setting parameters [-d20: minimum read depth] (Li et al., 2009). The genome coverage depth and coverage region was performed by using BAMstats version 1.25 (Nestornotabilis, 2016). The coverage graph was performed by using Bedtools version 2.25.0 (Quinlan, Hall, 2010). The BAM file was visualized by using Integrative Genomics Viewer (IGV) version 2.3.80 (Thorvaldsdottir et al., 2013). The circular genome map was performed by using DNAPlotter version 1.11 (Carver et al., 2009). The nucleotides distribution was performed by using awk command line version 20070501.

Submission of the draft genome of *L. fermentum* 47-7 to Whole-Genome Shotgun (WGS) project of the NCBI database

The draft genome was registered its sources information in the Biosample database (https://www.ncbi.nlm.nih.gov/biosample/) and Bioproject (https://www.ncbi.nlm.nih.gov/bioproject/). The raw reads subsequently were submitted to Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra). The consensus sequence was used as the input file for generating the output file by using Tbl2asn command line version 25.0 (.sqn file format for submission in WGS project). Finally, the output was submitted in to WGS of the NCBI database with the requirement for annotation of the draft genome by using Prokaryotic Genome Annotation Pipeline (PGAP) version 3.3. The submitted data was evaluated by NCBI GenBank Staff before releasing to the public and providing the accession number of draft genome.

Results

Isolation genomic DNA of *L. fermentum* 47-7

The genomic DNA of *L. fermentum* 47-7 was primarily determined for its integrity by 1% agarose gel electrophoresis. From figure 1, its genomic DNA presents as intense band above 20 Kb DNA marker band without any visible shearing smear. In addition, no plasmid was found in this bacteria.

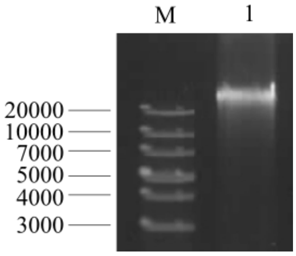

Figure 1 SYBR gold staining gel of the genomic DNA of *L. fermentum* 47-7. Lane M = GeneRuler^{TM} 1 kb Plus DNA ladder with the size in bp. Lane 1 = genomic DNA isolated from *L. fermentum* 47-7

Data processing and genome assembly and quality assessment

Whole genome sequencing of *L. fermentum* 47-7 was performed by two platforms. i) Illumina Hiseq 2000 platform (Thermofisher, USA), this system generated a total reads of 10,293,934 with the average length of 100 to 101bp. ii) Ion torrent PGM platform (Thermofisher, USA) generated a total reads of 583,615 with the average length of 100 to 369bp. The quality improvement by Fastx-toolkit version 0.0.13 and FastQC tool version 0.11.4 showed that the total sequence reads from Illumina Hiseq 2000 and Ion Torrent PGM was 9,340,868 and 488,674 reads after filtering based on quality sequence reads as showed in Table 1. The sequencing results of both platforms were merged by using Picard tool version 2.5.0 and obtained a total reads of 9,829,542. The merged BAM file which aligned with reference genome of *L. fermentum* IFO 3956 was 8,191,533 reads (83.33% from the total reads) and number of PF reads (passing illumina's filter) that were aligned to the reference sequence with a mapping of Q20 or higher was 7,140,640 reads. The consensus sequence was generated by Samtools version 0.1.19 and Bcftools version 0.1.17 with the length of 1,830,836 bp. The consensus sequence has the coverage region about 87.23% of the reference genome and the coverage depth of 411.1X which was performed by using BAMstats version 1.25 and awk command line version 20070501. The histogram of the coverage region and coverage depth of *L. fermentum* 47-7 genome was showed in Figure 2. Finally, the validated merging sequence from both platforms was used for submission into WGS portal. Table 2 shows the characteristics and resources of the draft *L. fermentum* 47-7 and accession numbers of each submission type that are required before submission of the draft genome to WGS. The chromosome features of *L. fermentum* 47-7 was showed in Figure 3.

Table 1 General information of data processing, genome assembly and quality assessment

Platforms/results	**Illumina Hiseq2000**	**Ion Torrent PGM**	**Illumina Hiseq 2000 + Ion Torrent PGM**
Total Reads Before Filtering	10,293,934	583,615	-
Total Reads After Filtering	9,340,868	488,674	-
PF_Reads*	9,340,868	488,674	9,829,542
PF_Reads_aligned*	7,779,337	412,196	8,191,533
PF_HQ_Aligned_reads*	6,813,025	327,615	7,140,640
Mean_Read_Length	100.944822	206.533437	153.7391295
Minimum_Read Length	100	100	100
Maximum_Read_Length	101	369	369
Mean_Coverage_Depth	371.2	39.9	411.1
Minimum coverage depth	0	0	0
Maximum coverage depth	7,646	1,274	8,911
Consensus sequence	2,098,685	2,098,685	2,098,685
Consensus sequence splitting Ns (bp)	1,830,669	1,776,801	1,830,836
Ns (bp)	268,016 (12.8%)	329,398 (15.7%)	267,849 (12.8%)
Ambiguity bases	433	45	430

*__PF_Reads__: The number of PF reads where PF is defined as passing Illumina's filter.

*__PF_Reads_Aligned__: The number of PF reads that were aligned to the reference sequence. This includes reads that aligned with low quality (i.e. their alignments are ambiguous).

*__PF_HQ_Aligned_Reads__: The number of PF reads that were aligned to the reference sequence with a mapping of Q20 or higher signifying that the aligner estimates a 1/100 (or smaller) chance that the alignment is wrong.

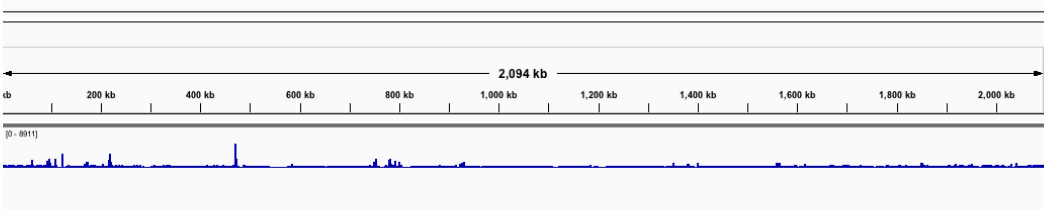

Figure 2 Histogram of the coverage region and coverage depth for the *L. fermentum* 47-7 genome with reference genome *L. fermentum* IFO 3956 (NC_010610.1:1-2,098,685bp) performed by using Bedtools version 2.25.0 and visualization with Integrative Genomics Viewer (IGV) version 2.3.80

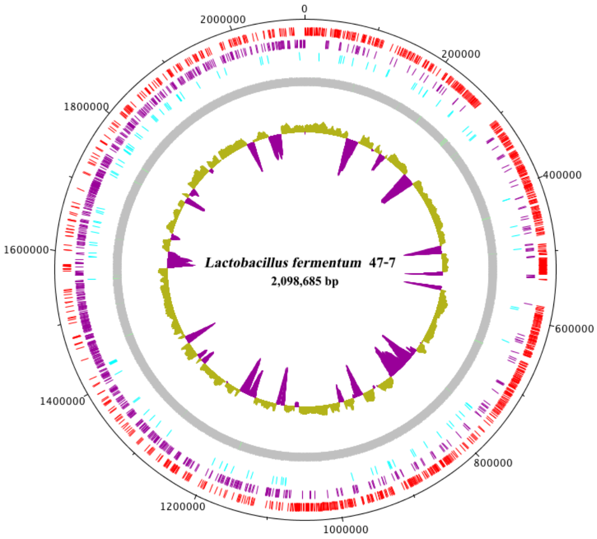

Figure 3 The chromosome features of *L. fermentum* strain 47-7. Track are numbered from outside to inside. Track 1 (black), size. Track 2 (red) and 3 (indigo), coding sequences (CDS) forward and reverse strand. Track 4 (cyan), pseudogenes. Track 5 (green), tRNA. Track 6, GC content. The circular map was drawn using DNAPlotter version 1.11

Table 2 The characteristics and resources of the draft genome of *Lactobacillus fermentum* 47-7

Name	Genome characteristics and resources
NCBI bioproject ID	PRJNA347617
NCBI biosample ID	SAMN05893390
NCBI Sequence Read Archive (SRA) ID	SRX2239653 (Illumina Hiseq 2000); SRX2239654 (Ion Torrent PGM)
Assembly ID	GCA_001854105.1
NCBI genome accession number	CP017712
Sequencer	Illumina Hiseq 2000 and Ion Torrent Personal Genome Machine (PGM)
Total number of reads	9,829,542
Read length (average)	153
Genome coverage depth	411.1
Genome coverage region	87.23%
Mapped reads	8,191,533 or 83.33 %
Total length (bp)	2,098,685
Ungapped genome size (bp)	1,830,930
Scaffolds	1
Contigs	180
N50	37,781
L50	14
G+C content	52.2%
Genes	1,814
Coding sequence (cds) or protein	1,626
tRNA coding genes	57
rRNA coding genes	15
ncRNA coding genes	4
Pseudogenes	112
Closest species reference genome	*Lactobacillus fermentum* IFO 3956 symmetrical identity: 92.6205%
Closest genome	*Lactobacillus fermentum* Lf1 symmetrical identity: 95.7865%

Discussion

In this study we generated the draft genome of *L. fermentum* 47-7. The FASTQ files from Illumina Hiseq 2000 and Ion Torrent PGM was filtered with minimum quality scores to keep at 20 and above was performed by using Fastx-toolkit version 0.0.13 (Gordon, Hannon, 2016) and FastQC version 0.11.4 (Andrews, 2016). The filtered sequence reads with high quality scores > 20 and above from Illumina Hiseq 2000 and Ion Torrent PGM was used to mapping with reference genome of *L. fermentum* IFO 3956 by using BWA mem algorithm with number of threads (-t 8) with reference genome and in the previous report used parameters the number of threads (-t 10) in human genome (Cornish, Guda, 2015). BWA mem algorithm is robust to sequencing errors and can be applied to a wide range of sequence lengths from 70bp to a few megabases and also exhibits a good compromise between computing speed and sensitivity expressed as percentage of aligned reads (Mielczarek, Szyda, 2016). Therefore, we selected BWA mem for mapping draft genome in this study. The output file for mapping is two SAM files from both Illumina Hiseq 2000 and Ion Torrent PGM and SAM files was converted to BAM file, which reduces the file size and improves computing efficiency by using Picard tool version 2.5.0 (Broadinstitute, 2016). The BAM files were subsequently sorted in the order of chromosomes and indexed by using Picard tool version 2.5.0 (Broadinstitute, 2016). The read group information of the sequencing machine was added to the head of each BAM files by using Picard too version 2.5.0 (Broadinstitute, 2016). Both BAM files were merged together by using Picard tool version 2.5.0 (Broadinstitute, 2016) in order to increase the number of sequence reads, coverage depth and coverage regions resulting in a more accurate and reliable result. During the sequencing process, the same DNA molecules can be sequenced for several times. The results of duplicate reads are not informative and should not be counted as additional evidence for or against a putative variant (Auwera et al., 2013). Therefore, the merged BAM file was performed a PCR duplicate marking process, which enables the GATK version 3.3 (Auwera et al., 2013;DePristo et al., 2011;McKenna et al., 2010) to ignore duplicates in subsequent processing. The realignment process identifies the most consistent placement of the reads with respect to the indels in order to clean up these artifacts (Auwera et al., 2013). Thus in this study, local realignment in merged BAM file prior to recalibration was performed. The per-base estimate of error known as the base quality score is the foundation upon which all statistically calling algorithms are based. The estimates provided by sequencing machines are often inaccurate and/or biased. The recalibration process applies an empirically accurate error model to the bases, producing a BAM file that is suitable for analysis (Auwera et al., 2013). Therefore, base quality score recalibration (BQSR) step was performed in merged BAM file by using GATK version 3.3 (Auwera et al., 2013;DePristo et al., 2011;McKenna et al., 2010) with known sites SNPs and INDELs. After performed recalibration process in merged BAM file, it was showed a significant improvement in the accuracy of quality scores. The parameters used in markduplicates, local realignment and BQSR step in merged BAM file were derived from those previous reported by GATK (Auwera et al., 2013). The merged BAM file was used to generated consensus sequence by using Samtools version 0.1.19 and Bcftools version 0.1.17 (Li et al., 2009). The parameters for generating the consensus sequence of the draft genome of *L. fermentum* 47-7 were derived from those used in case of the draft genome of *Bacillus pumilus* ku-bf1 (Balsingh et al., 2016). The consensus sequence was finally submitted to WGS of NCBI database and request for annotation draft genome by using PGAP version 3.3 and showed the length is 1,830,930 bp with coverage region is 87.23% of reference genome and G+C content of 52.2% has 1,814 genes consisting of 1,626 coding sequences (CDS), 57 tRNAs, 15 rRNAs, 4 ncRNAs, 112 pseudogenes. The draft genome of *L. fermentum* 47-7 showed high coverage depth of 411.1X in merged BAM file generated by BAMstats version 1.25 (Nestornotabilis, 2016) when comparing with draft genome of *L. fermentum* Lf1 mapping with BWA algorithm (Grover et al., 2013) showed the coverage depth of 86.9X. Previous reported suggested that the high depth WGS is the “gold standard” for DNA resequencing because it can interrogate all variant types including SNPs, INDELs, structural variants and copy number variants (CNVs) (Sims et al., 2014). Therefore, the high coverage depth in merged BAM file leads to a more accurate and reliable result for the draft genome of *L. fermentum* 47-7. Submission of this draft genome sequence to the public database provides a useful information for those who are interested to explore various interesting characteristics of *L. fermentum* strains and/or the interaction with the human host.

Conclusion

The draft genome sequence of *L. fermentum* 47-7 has been determined using two different sequencing platforms i.e. Illumina Hiseq 2000 and Ion Torrent PGM, and shows high accurate and reliable result with the genome coverage of 411.1X. This genome sequence has been deposited in NCBI database with the accession number CP017712. The genome with G+C content of 52.2% contains 1,814 genes with 1,626 coding sequences, 57 tRNAs, 15 rRNAs, 4 ncRNAs, 112 pseudogenes. This draft genome sequence provides a useful database of *L. fermentum* for further explore various interesting characteristics e.g. genes involving in probiotic properties.

Acknowledgements

This study was supported by Research and Diagnostic Center for Emerging Infectious Diseases (RCEID), Department of Microbiology, Faculty of Medicine, Khon Kaen University.

References

Amy E. Foxx-Orenstein D, and William D. Chey M. Manipulation of the Gut Microbiota as a Novel Treatment Strategy for Gastrointestinal Disorders. The American Journal of Gastroenterology supplements 2012: 41-6.

Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. [2016 Oct 22]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2016.

Archer AC, Halami PM. Probiotic attributes of *Lactobacillus fermentum* isolated from human feces and dairy products. Appl Microbiol Biotechnol 2015 Oct; 99(19): 8113-23.

Auwera GAVd, Carneiro MO, Hartl C, Poplin R, Angel Gd, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics 2013; 43: 11.0.1–.0.33.

Balsingh J, Radhakrishna S, Ulaganathan K. Draft Genome Sequence of *Bacillus pumilus* ku-bf1 Isolated from the Gut Contents of Wood Boring Mesomorphus sp. Front Microbiol 2016; 7: 1037.

Bao Y, Zhang YC, Zhang Y, Liu Y, Wang SQ, Dong XM, et al. Screening of potential probiotic properties of *Lactobacillus fermentum* isolated from traditional dairy products. Food Control 2010 May; 21(5): 695-701.

Broadinstitute. Picard tool: A set of command line tools (in java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. [2016 Oct 22]. Available from: https://broadinstitute.github.io/picard/. 2016.

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 2009 Jan 1; 25(1): 119-20.

Cornish A, Guda C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Research International 2015.

De S, Kaur G, Roy A, Dogra G, Kaushik R, Yadav P, et al. A Simple Method for the Efficient Isolation of Genomic DNA from Lactobacilli Isolated from Traditional Indian Fermented Milk (dahi). Indian J Microbiol 2010 Oct; 50(4): 412-8.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 2011 May; 43(5): 491-8.

Gordon A, Hannon G. Fastx-Toolkit. FASTQ/A Short-Reads Pre-Processing Tools. [2016 oct 22]. Available from: http://www.hannonlab.cshl.edu/fastx_toolkit. 2016.

Grover S, Sharma VK, Mallapa RH, Batish VK. Draft Genome Sequence of *Lactobacillus fermentum* Lf1, an Indian Isolate of Human Gut Origin. Genome Announc 2013; 1(6).

Karlyshev AV, Raju K, Abramov VM. Draft Genome Sequence of *Lactobacillus fermentum* Strain 3872. Genome Announc 2013; 1(6).

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010 Mar 1; 26(5): 589-95.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009 Aug 15; 25(16): 2078-9.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 2010 Sep; 20(9): 1297-303.

Mielczarek M, Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet 2016 Feb; 57(1): 71-9.

Mital BK, Garg SK. Anticarcinogenic, hypocholesterolemic, and antagonistic activities of *Lactobacillus acidophilus*. Crit Rev Microbiol 1995; 21(3): 175-214.

Nestornotabilis. BAMStats: an interactive desktop GUI tool for summarising Next Generation Sequencing alignments. [2016 Oct 22]. Available from: http://bamstats.sourceforge.net/. 2016.

Pereira DI, McCartney AL, Gibson GR. An in vitro study of the probiotic potential of a bile-salt-hydrolyzing *Lactobacillus fermentum* strain, and determination of its cholesterol-lowering properties. Appl Environ Microbiol 2003 Aug; 69(8): 4743-52.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010 Mar 15; 26(6): 841-2.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 2014 Feb; 15(2): 121-32.

Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 2013 Mar; 14(2): 178-92.

Tsai YT, Cheng PC, Pan TM. The immunomodulatory effects of lactic acid bacteria for improving immune functions and benefits. Applied Microbiology and Biotechnology 2012 Nov; 96(4): 853-62.

Villena J, Racedo S, Aguero G, Bru E, Medina M, Alvarez S. *Lactobacillus casei* improves resistance to pneumococcal respiratory infection in malnourished mice. Journal of Nutrition 2005 Jun; 135(6): 1462-9.

Yotpanya P, Panya M, Engchanil C, Suebwongsa N, Namwat W, Thaw H, et al. Probiotic characterization of lactic acid bacteria isolated from infants feces and its application for the expression of green fluorescent protein. Malaysian Journal of Microbiology 2016 Mar; 12(1): 76-84.