Examples of low quality Illumina sequencing of prokaryotic genomes

17-Jan-2025

Haeyoung Jeong
Korea Bioinformation Center (KOBIC)
Korea Research Institute of Bioscience and Biotechnology (KRIBB)

Unpacking data.tar file will produce five subdirectories: **01_reads, 02_k-mer_analysis, 03_phyloFlash, 04_assembly,** and **05_GTDB-Tk.** This document does not include descriptions for all files. For example, descriptions for self-explanatory files have been omitted. Some subdirectories may have their own documentation files to describe the contents. Please feel free to contact Haeyoung Jeong at hyjeong@kribb.re.kr or jeong0449@gmail.com if you have any questions or comments.

The eight strains used for genome sequencing are listed in the table below. The correct names were verified using the LPSN website (https://lpsn.dsmz.de/) as of January 16, 2025. Sample IDs beginning with numbers indicate KCTC numbers. Please note that some of these strains may not be available from KCTC.

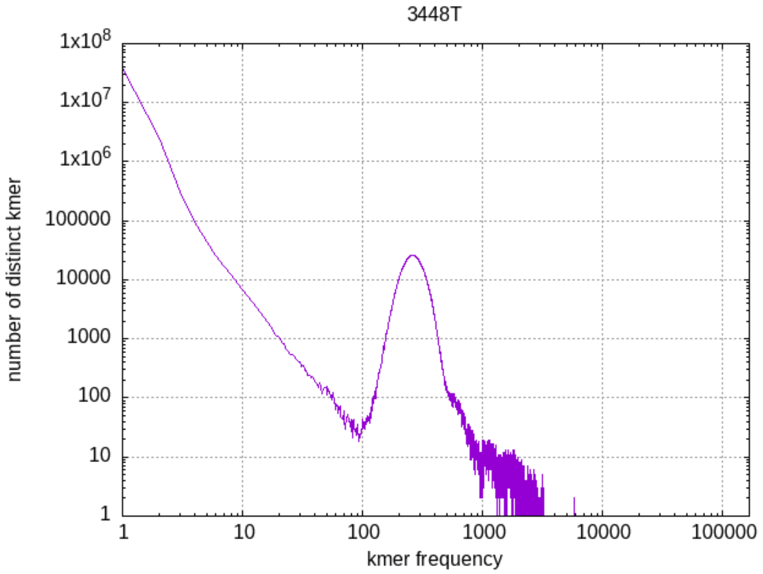
Sample ID	Original label (2013)	Correct name
3520T	*Brochothrix campestris*	
15666	*Bacillus* sp.	
25222T	*Succinivibrio dextrinosolvens*	
DSM1535T	*Methanobacterium formicicum*	
JCM15447T	*Sunxiuqinia faeciviva*	
K16	*Catabacter hongkongensis* (88678)	*Christensenella hongkongensis*
KIM3	*Methanobrevibacter* sp. (Dr. Kim3)	
Strain15	*Geodermatophilus obscurus* (15)	

01_reads - Paired-end FASTQ files generated using the Illumina HiSeq 2000 platform (2 × 101 nt cycles). For sequencing statistics of each file, including the number of bases, number of sequences, and average sequence lengths, please refer to the `readstats.txt` files.

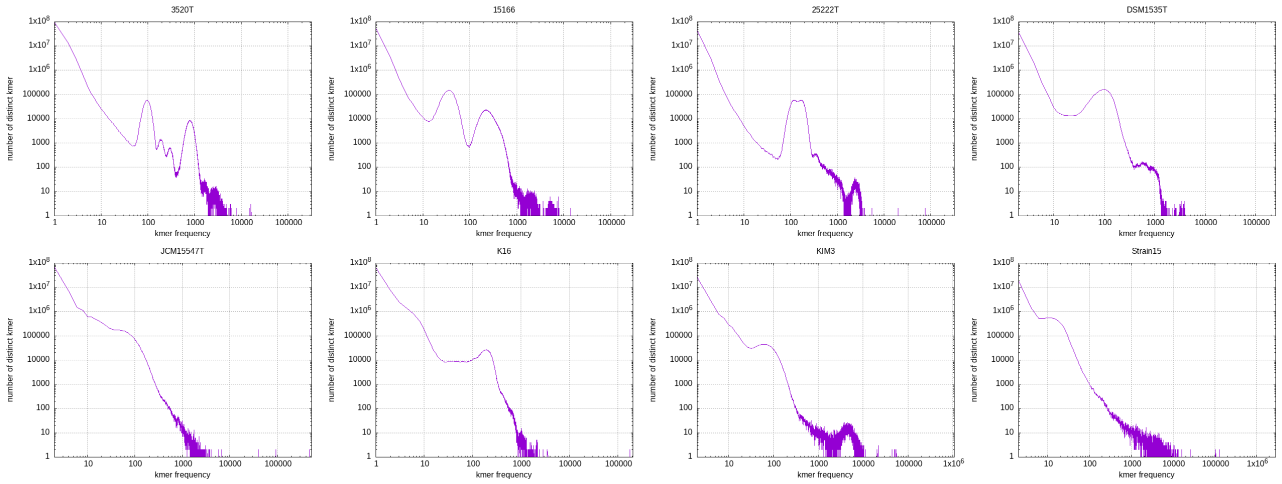
02_k-mer_analysis - Results of 21-mer abundance analysis calculated using Jellyfish v2.3.0 (`*.kmer21.txt.gz` and `*.jf.hist`). Histograms were generated using the commands outlined below.

```
awk '{print $2}' SAMPLE.kmer21.txt | sort -n | uniq -c | awk '{print
$2 "," $1}' > SAMPLE.jf.hist
```

Gnuplot v5.4 was used to generate PNG files from the *.gp scripts. An example of a high-quality k-mer abundance profile from typical prokaryotic haploid genome sequencing (a single strain, KCTC 3448^{T}) is shown below. The major peak should be located at the sequencing coverage value on the X-axis.



Shown below are the 21-mer abundance profiles from the genome sequencing results of the eight samples analyzed in this study.



03_phyloFlash – Reports (`*.html`) and output files (`*.tar.gz`) generated using phyloFlash v3.4.1. The SILVA 138.1 database was used as the reference.

Sample ID	Original label (2013)	16S rRNA assembly-based taxa
3520T	*Brochothrix campestris*	*Moraxella osloensis* *Brochothrix thermosphacta*
15166	*Bacillus* sp.	*Bacillus velezensis* *Bacillus gottheilii*
25222T	*Succinivibrio dextrinosolvens*	*Clostridium subterminale* *Succinivibrio dextrinosolvens*
DSM1535T	*Methanobacterium formicicum*	*Paraclostridium benzoelyticum* *Clostridium botulinum*
JCM15447T	*Sunxiuqinia faeciviva*	*Sunxiuqinia faeciviva*
K16	*Catabacter hongkongensis* (88678)	*Flavonifractor plautii* *Paraclostridium benzoelyticum* uncultured *Anaerotruncus* sp.
KIM3	*Methanobrevibacter* sp. (Dr. Kim3)	*Methanobrevibacter smithii*
Strain15	*Geodermatophilus obscurus* (15)	*Klenkia brasiliensis*

04_assembly – *De novo* genome assembly results were generated using the ZGA pipeline v0.0.9, with Unicycler v0.5.0 employed for the assembly process. Quality assessment of the assembled genomes was conducted using CheckM v1.2.0 within the ZGA pipeline. High-quality genomes are typically defined as having greater than 90% completeness and less than 5% contamination. In other words, a well-assembled genome should exceed 90% completeness while maintaining contamination levels below 5%. Six low-quality genomes, as identified by CheckM, are marked with an asterisk (*). The other two either exhibit high strain heterogeneity (JCM15447T) or low completeness (Strain 15).

Sample ID	Assembly metrics (count, total, max, n50, average)	Completeness	Contamination	Strain heterogeneity
3520T*	87 2833025 452206 183682 32563	100.0	13.79	0.0
15166*	86 4559427 591661 244180 53016	98.28	8.62	20.0
25222T*	497 6470842 410753 94698 13019	100.0	100.69	0.0
DSM1535T*	227 6984074 526278 117567 30766	100.0	96.55	2.94
JCM15447T	2327 6278725 24715 3442 2698	84.58	20.25	80.95
K16*	237 4013022 391154 84475 16932	96.55	0.0	0.0
KIM3	569 1507056 13090 3440 2648	26.06	2.04	0.0
Strain15*	774 1174504 8889 1471 1517	20.97	0.0	0.0

05_GTDB-Tk – Taxonomic assignments were performed using GTDB-Tk v2.1.1. The `gtdbtk.bac120.summary.tsv` and `gtdbtk.ar53.summary.tsv` files, located in the `05_GTDB-Tk/classify` subdirectory, contain results for bacterial and archaeal genomes, respectively.

Sample ID	GTDB-Tk classification
3520T	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Moraxella_A;s__Moraxella_A cinereus
15166	d__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;g__Bacillus;s__Bacillus velezensis
25222T	Unclassified Bacteria
DSM1535T	Unclassified Bacteria
JCM15447T	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Prolixibacteraceae;g__Sunxiuqinia;s__
K16	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Flavonifractor;s__Flavonifractor plautii
KIM3	d__Archaea;p__Methanobacteriota;c__Methanobacteria;o__Methanobacteriales;f__Methanobacteriaceae;g__Methanobrevibacter_A;s__Methanobrevibacter_A smithii
Strain15	d__Bacteria;p__Actinobacteriota;c__Actinomycetia;o__Mycobacteriales;f__Geodermatophilaceae;g__Klenkia;s__