Examples of low quality Illumina sequencing of prokaryotic genomes

17-Jan-2025 / 17-Feb-2025 (v2) / 05-Mar-2025 (v3)
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)	

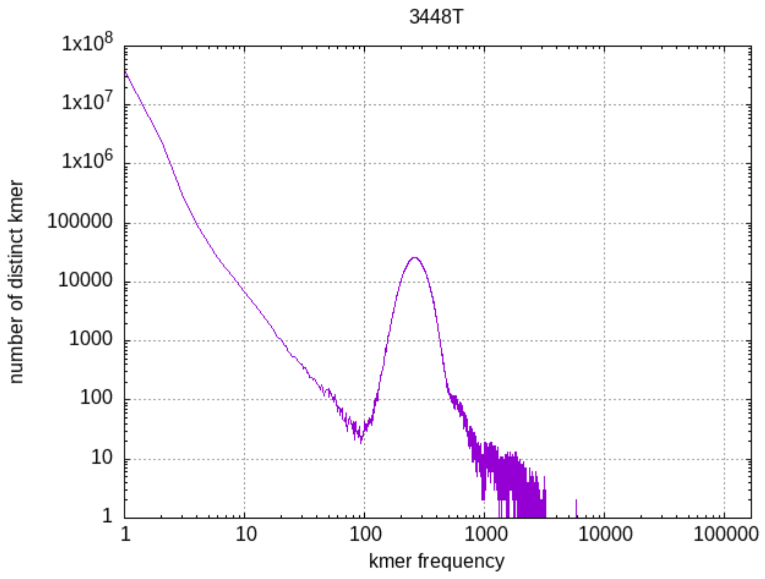
01_reads Paired-end FASTQ files generated using the Illumina HiSeq 2000 platform (2 × 101 nt cycles). For sequencing statistics of each file, including the number of bases, number of sequences, and average sequence lengths, please refer to the `readstats.txt` files.

02_k-mer_analysis Results of 21-mer abundance analysis for the interleaved fastq files (`*.pe.fq`; deleted after the analysis) calculated using Jellyfish v2.3.0

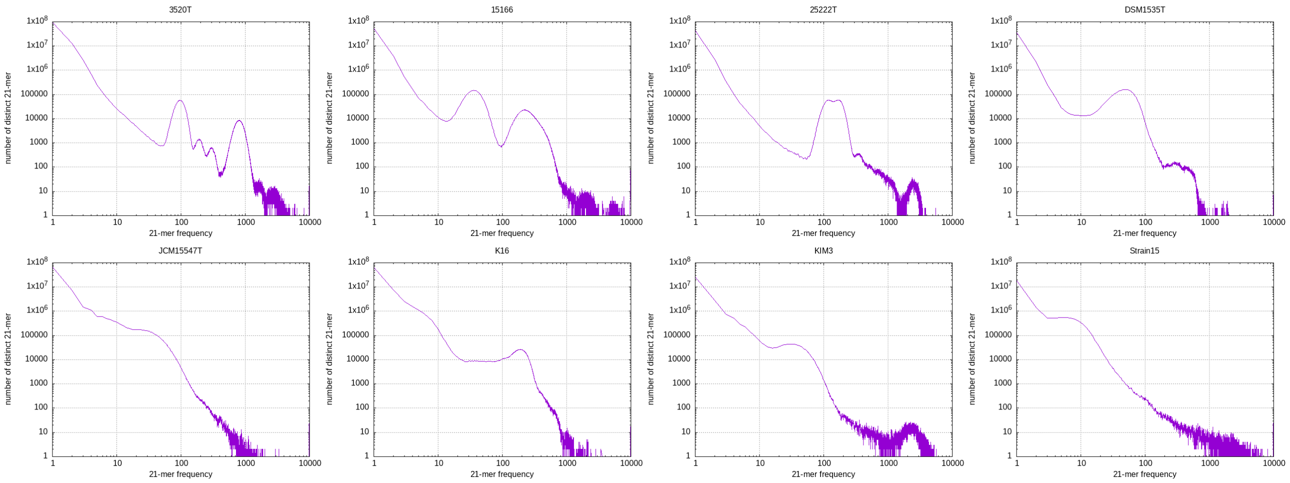
and their plots generated Gnuplot v5.4. For your convenience, use the script shown below.

```
for x in *pe.fq
do
  echo Processing $x...
  x=${x%%.pe.fq}
  jellyfish count -m 21 -s 100M -t 12 -C $x.pe.fq -o $x.counts.jf
  jellyfish histo -o $x.jf.hist $x.counts.jf
  echo Running gnuplot...
  echo set term png > $x.jf.gp
  echo set output \"$x.jf.png\" >> $x.jf.gp
  echo set logscale x >> $x.jf.gp
  echo set logscale y >> $x.jf.gp
  echo set grid >> $x.jf.gp
  echo set xlabel \"21-mer frequency\" >> $x.jf.gp
  echo set ylabel \"number of distinct 21-mer\" >> $x.jf.gp
  echo set key off >> $x.jf.gp
  echo set title \"$x\" >> $x.jf.gp
  echo plot \"$x.jf.hist\" using 1:2 with lines >> $x.jf.gp
  gnuplot $x.jf.gp
done
```

An example of a high-quality k-mer abundance profile from typical prokaryotic haploid genome sequencing (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 (`AllPlots.png`).



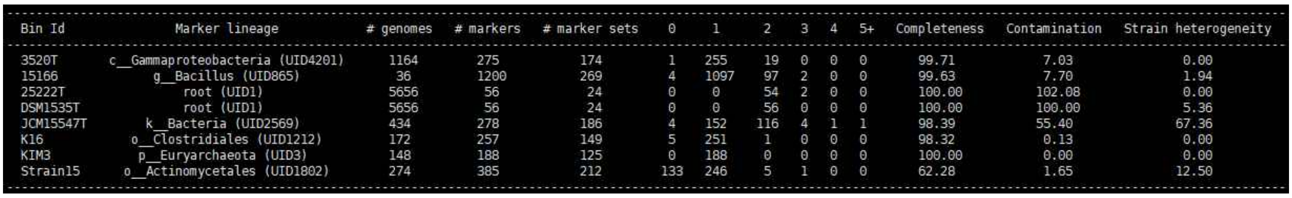
03_phyloFlash Reports (`*.html`) and output files (`*.tar.gz`) generated using phyloFlash v3.4.1. The SILVA 138.1 database was used as the reference. The identification of multiple taxa from a single sample may be a sign of contamination (3520T, 15166, 25222T, DSM1535T, and K16).

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)	None
Strain15	*Geodermatophilus obscurus* (15)	None

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 (‘`zga -1 SAMPLE_1.fastq.gz -2 SAMPLE_2.fastq.gz --calculate-genome-size --threads 16 -minimum-contig-length 200 -o zga_SAMPLE`’). For KIM3, ZGA pipeline was executed with the ‘`--domain archaea`’ option. High-quality genomes are typically defined as having greater than 90% completeness and less than 5% contamination. Six low-quality genomes, as identified by CheckM in the ZGA pipeline (WARNING message in the `zga.log` file), are marked with an asterisk (*).

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*	167 7180529 1201701 245079 42997	100.0	98.28	2.91
JCM15447T*	1101 8022938 155331 15967 7286	96.55	37.73	85.71
K16*	237 4013022 391154 84475 16932	96.55	0.0	0.0
KIM3	41 1786384 314723 102328 43570	100.0	0.0	0.0
Strain15*	1347 3403316 18977 2937 2526	67.49	0.0	0.0

The quality of the assembled genomes was reassessed using CheckM v1.2.0, the same version incorporated in the ZGA pipeline, with the command 'checkm lineage_wf -t 16 -x fasta INPUT_DIR OUTPUT_DIR'. The results files are located in the 04_assembly/rerun subdirectory.



```
------------------------------------------------------------------------------------------------------------------------------------------------------------------
  Bin Id                 Marker lineage            # genomes   # markers   # marker sets    0     1      2    3    4    5+   Completeness   Contamination   Strain heterogeneity
------------------------------------------------------------------------------------------------------------------------------------------------------------------
  3520T       c__Gammaproteobacteria (UID4201)       1164         275           174          1    255     19    0    0    0       99.71           7.03               0.00
  15166             g__Bacillus (UID865)               36         1200          269          4    1097    97    2    0    0       99.63           7.70               1.94
  25222T                root (UID1)                  5656          56            24          0     0      54    2    0    0       100.00         102.08              0.00
  DSM1535T              root (UID1)                  5656          56            24          0     0      56    0    0    0       100.00         100.00              5.36
  JCM15547T         k__Bacteria (UID2569)             434         278           186          4    152    116    4    1    1       98.39          55.40              67.36
  K16             o__Clostridiales (UID1212)          172         257           149          5    251     1     0    0    0       98.32           0.13               0.00
  KIM3             p__Euryarchaeota (UID3)            148         188           125          0    188     0     0    0    0       100.00          0.00               0.00
  Strain15       o__Actinomycetales (UID1802)         274         385           212         133   246     5     1    0    0       62.28           1.65              12.50
------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

The 21-mer abundance spectrum of the KIM3 sample appears somewhat unusual, but since the assembly statistics are excellent and the CheckM assessment results are also satisfactory, we consider it not to be of low quality.

05_GTDB-Tk Taxonomic assignments were performed using GTDB-Tk v2.4.0. 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. Warning messages generated by GTDB-Tk were displayed in the footnotes.

Sample ID	GTDB-Tk classification
3520T	d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales; f__Moraxellaceae;g__Moraxella_A;s__Moraxella_A cinereus
15166[1)]	d__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;g__Bacillus; s__Bacillus velezensis

25222T[2]	Unclassified Bacteria
DSM1535T[3]	Unclassified Bacteria
JCM15447T[4]	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__Klenkia taihuensi

Change Log

v3 (bug fixes) - Some FASTQ files were found to contain identical read IDs and sequences (DSM1535T, JCM15547T, KIM3, and Strain15). Specifically, entire reads appeared repeatedly from the middle of the FASTQ file onward. This issue was likely caused by an error during the manual merging of split FASTQ files generated by the HiSeq 2000 (‘`cat SAMPLE_R1_001.fastq SAMPLE_R1_002.fastq > SAMPLE_1.fastq`’). In version 3 (v3), these FASTQ files were corrected, and all analyses were re-conducted using the updated files. Additionally, GTDB-Tk was upgraded from version 2.1.1 to 2.4.0. We sincerely appreciate the KOBIC quality control team for identifying this issue in the submitted FASTQ files.

v2 (bug fixes) - Table in the **04_assembly** section was corrected.

v1 (first release)

[1] Genome has more than 12.5% of markers with multiple hits

[2] Insufficient number of amino acids in MSA (0.8%)

[3] Insufficient number of amino acids in MSA (1.7%)

[4] Genome has more than 43.3% of markers with multiple hits;Genome not assigned to closest species as it falls outside its pre-defined ANI radius