Metagenomics Analysis Report

Sponsor:	Swiss Federal Institute of Technology Lausanne
Project ID:	s1886g03001
Report date:	19.Mar.2022
Contract number:	MBEPFL-20220215
Sample number:	48
Sample type:	Mouse feces
Host species:	Mus musculus
Analysts:	Liangliang Chen
Reviewer:	Jiawei Wang

Sequanta Technologies Co., Ltd.

1. Data analysis methods

1.1. Quality Filtering and Trimming
Low quality bases and adapters were trimmed. Short (length <35 bp) reads and low-quality reads were removed.

1.2. Host Sequences Removal
Host sequences were eliminated by mapping to the host reference genome with bowtie2[1], then mapped host sequences will be removed.

1.3. Taxonomic Classification
The clean data was aligned to in house develop microbial databases, and identified by Kraken2[2], using a memory-intensive algorithm that associates short genomic substrings (k-mers) with the lowest common ancestor (LCA) taxa.

1.4. Abundance reestimation
Producing accurate species- and genus-level abundance estimates by Bracken[3].

1.5. Microbial annotation
Identified microorganism is annotated with scientific name, gram and taxonomy information.

2. Database range

2.1. Bacteria: 9471
2.2. Fungi: 1854
2.3. Virus: 15752
2.4. Parasite: 88
2.5. Total: 27165

3. Data Summary

case	total_reads	total_bases	q30_rate	gc_content	total_clean_reads	total_microbial_reads	microbial_reads_percents
C22001830LD01-1	90817622	13622643300	92.04%	47.20%	82371055	13150028	15.96%
C22001831LD01-2	73664528	11049679200	91.60%	47.34%	68283513	9218596	13.50%
C22001832LD01-3	69320618	10398092700	91.81%	48.28%	65882311	2952457	4.48%
C22001833LD01-4	87074424	13061163600	91.58%	47.91%	81493768	9530338	11.69%
C22001836LD01-7	87970482	13195572300	91.47%	49.36%	85432491	4534232	5.31%
C22001837LD01-8	67017996	10052699400	91.82%	47.94%	65028886	2105433	3.24%
C22001838LD01-9	88136822	13220523300	92.66%	47.88%	79720914	6144721	7.71%
C22001839LD01-10	66956074	10043411100	91.83%	48.70%	65591661	1977901	3.02%
C22001842LD01-13	93956104	14093415600	91.67%	50.15%	91752071	4323712	4.71%
C22001843LD01-14	80965010	12144751500	92.61%	48.11%	78465303	5985436	7.63%
C22001844LD01-15	77776550	11666482500	91.39%	47.85%	72158393	6784981	9.40%
C22001845LD01-16	84531978	12679796700	92.86%	47.76%	76654720	5904282	7.70%
C22001848LD01-19	102438006	15365700900	92.04%	48.41%	97747911	8458350	8.65%
C22001849LD01-20	77432542	11614881300	92.11%	47.79%	76320586	1766405	2.31%
C22001850LD01-21	67931212	10189681800	91.05%	49.46%	64027174	2209141	3.45%
C22001851LD01-22	72391532	10858729800	90.11%	47.78%	70186253	1891099	2.69%
C22001854LD01-25	87752174	13162826100	92.12%	48.24%	84548187	8346701	9.87%
C22001855LD01-26	71478734	10721810100	91.08%	46.05%	50199016	6998331	13.94%
C22001856LD01-27	96239348	14435902200	92.51%	47.51%	91367363	6716595	7.35%
C22001857LD01-28	74433018	11164952700	91.78%	48.99%	72218203	2026996	2.81%
C22001860LD01-31	104831918	15724787700	91.41%	49.06%	99873852	7564047	7.57%
C22001861LD01-32	68986000	10347900000	91.97%	48.33%	67035855	2678378	4.00%
C22001862LD01-33	83506388	12525958200	91.24%	49.07%	78831977	2698516	3.42%
C22001863LD01-34	73674928	11051239200	91.40%	49.11%	70743054	2963846	4.19%
C22001866LD01-37	100392008	15058801200	92.08%	48.50%	95014916	7555980	7.95%
C22001867LD01-38	72214214	10832132100	91.73%	48.99%	69212233	4524058	6.54%
C22001868LD01-39	91681086	13752162900	91.53%	48.24%	87659974	2847806	3.25%
C22001869LD01-40	87808944	13171341600	91.40%	48.23%	82357973	4146753	5.04%
C22001872LD01-43	81266242	12189936300	92.44%	48.28%	76289335	10498231	13.76%
C22001873LD01-44	75389270	11308390500	91.33%	48.73%	70989661	11806353	16.63%
C22001874LD01-45	71476092	10721413800	91.55%	45.85%	67455885	23751485	35.21%
C22001875LD01-46	75071452	11260717800	91.91%	47.89%	68950023	12720239	18.45%
C22001878LD01-49	91980370	13797055500	91.60%	47.97%	87459879	7453241	8.52%
C22001879LD01-50	80650978	12097646700	91.54%	48.35%	74901197	3880462	5.18%
C22001880LD01-51	105802176	15870326400	91.92%	47.76%	94527153	12809524	13.55%
C22001881LD01-52	73841028	11076154200	91.91%	48.08%	69457324	5318677	7.66%
C22001884LD01-55	78563156	11784473400	91.74%	48.34%	66729156	6820851	10.22%

C22001885LD01-56	74738858	11210828700	91.61%	47.72%	71519168	4733935	6.62%
C22001886LD01-57	85700252	12855037800	91.46%	48.11%	80855895	6193297	7.66%
C22001887LD01-58	75721992	11358298800	92.04%	48.65%	72573026	9163772	12.63%
C22001890LD01-61	89405550	13410832500	92.25%	47.78%	76477862	9823739	12.85%
C22001891LD01-62	85563772	12834565800	90.90%	48.38%	66086610	3770538	5.71%
C22001892LD01-63	86082680	12912402000	92.03%	47.45%	77547301	8798250	11.35%
C22001893LD01-64	82310520	12346578000	91.38%	47.48%	63131202	4829361	7.65%
C22001896LD01-67	106073364	15911004600	92.43%	47.70%	103418443	8411936	8.13%
C22001897LD01-68	73384866	11007729900	91.12%	49.04%	65885608	7101451	10.78%
C22001898LD01-69	87930454	13189568100	91.68%	43.50%	46971779	15212295	32.39%
C22001899LD01-70	85597318	12839597700	92.55%	48.55%	80367452	11003857	13.69%

Table Notes:
case: Sample ID
total_reads: total sequenced read counts.
total_bases: total sequenced bases.
q30_rate: Ratio of bases over Q30.
gc_content: Percentage of GC bases in total bases.
total_clean_reads: total read counts after quality filtering, trimming and host sequence removal.
total_microbial_reads: Sum of all microbial species read counts.
microbial_reads_percents: Percentage of microbial reads in total clean reads.

4. Result files

4.1. Raw fastq
Raw sample reads.
Data directory: /s1886g03001/raw_fastq/

4.2. Clean fastq
Sample reads after quality filtering, trimming and host sequence removal.
Data directory: /s1886g03001/clean_fastq/

4.3. Quality control report
A Microsoft Excel table presenting quality control information, such as Q30, microbial reads percents, total reads, gc content and so on.
Data directory: /s1886g03001/qc_report/

4.4. Kraken2 output file
Each sequence classified by Kraken results in a single line of output. Output lines contain five tab-delimited fields; from left to right, they are:
"C"/"U": one letter code indicating that the sequence was either classified or unclassified.
The sequence ID, obtained from the FASTA/FASTQ header.
The taxonomy ID Kraken used to label the sequence; this is 0 if the sequence is unclassified.

The length of the sequence in bp.

A space-delimited list indicating the LCA mapping of each k-mer in the sequence.

Data directory: /s1886g03001/kraken2_output/

4.5. Kraken2 taxonomic report

A plain text table, with one line per taxon. The fields of the output, from left-to-right, are as follows:

Percentage of reads covered by the clade rooted at this taxon

Number of reads covered by the clade rooted at this taxon

Number of reads assigned directly to this taxon

A rank code, indicating (U)nclassified, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies. All other ranks are symbolized as '-'.

NCBI taxonomy ID

indented scientific name

The scientific names are indented using spaces, according to the tree structure specified by the taxonomy.

Data directory: /s1886g03001/kraken2_report/

4.6. Bracken output file

Kraken2 taxonomic report after abundance reestimation.

Data directory: /s1886g03001/bracken_output/

4.7. Report xlsx file

Summary file of metagenomics analysis.

A Microsoft Excel table, with one line per species. The fields of the output, from left-to-right, are as follows:

case:	Sample ID
taxonomy_id:	NCBI taxonomy ID
name:	NCBI taxonomy scientific name of species
genus_id:	NCBI taxonomy ID of genus
name_g:	NCBI taxonomy scientific name of genus
domain:	NCBI taxonomy scientific name of domain
new_est_reads:	Species read counts estimates by Bracken
microbial_reads:	Sum of all species read counts
rpm:	Species read counts / Sum of all species read counts (Millions)
new_est_reads_g:	Genus read counts estimates by Bracken
gram:	Gram positive or gram negative, only for bacteria
virus_strand_type:	Virus strand type, only for virus

Data directory: /s1886g03001/report_table/

4.8. Reference pdf files

Data directory: /s1886g03001/reference/

5. Reference

[1] Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

[2] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biology. 2019, 76230.
[3] Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science. 2017, 3:e104.