Rapid metagenomic workflow using annotated 16S RNA dataset

Naoaya Oec¹ and Hidemasa Bono²

BioHackathon series:
[DBCLS domestic BioHackathon](#)
Aomori, Japan, 2021
Metagenome analysis workflow group

Submitted: 25 Sep 2021

License
Authors retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Published by [BioHackrXiv.org](#)

¹ Dogrun Inc, Shizuoka, Japan ² Graduate School of Integrated Sciences for Life, Hiroshima University, Hiroshima, Japan

Abstract

Thanks to the dramatic progress in DNA sequencing technology, it is now possible to decipher sequences in a mixed state. Therefore, the subsequent data analysis has become important, and the demand for metagenomic analysis is very high. Existing metagenomic data analysis workflows for 16S amplicon sequences have been mainly focused on sequences from short reads sequencers, while researchers cannot apply those workflows for sequences from long read sequencers. A practical metagenome workflow for long read sequencers is therefore really needed. In a domestic version of the BioHackathon called BH21.8 held in Aomori, Japan (23-27 August 2021), we first discussed the reproducible workflow for metagenome analysis. We then designed a rapid metagenomic workflow using annotated 16S RNA dataset (Ref16S) and the practical use case for using the workflow developed. Finally, we discussed how to maintain Ref16S and requested Life Science Database Archive in JST NBDC to archive the dataset. After a stimulus discussion in BH21.8, we could clarify the current issues in the metagenomic data analysis. We also could successfully construct a rapid workflow for those data specially from long reads by using newly constructed Ref16S.

Keywords: metagenome, workflow, long reads

Introduction

In the year 2021, metagenomic analysis has been getting attention. Metagenomic analysis, which involves DNA sequencing of the bacteria that make up the bacterial flora, including previously undetectable bacteria, is being conducted on a variety of samples, including human intestinal bacteria. In the past, microorganisms had to be isolated and cultured, followed by genome sequencing and data analysis, but thanks to the dramatic progress in DNA sequencing technology, it is now possible to sequence microorganisms while they are still mixed together. However, thanks to the dramatic progress in DNA sequencing technology, it is now possible to decipher sequences in a mixed state. Therefore, the subsequent data analysis has become important, and the demand for it is very high.

For metagenomic data analysis, QIIME2 (<https://qiime2.org>), next-generation microbiome bioinformatics platform that is extensible, free, open source and community developed, has been widely used (Bolyen et al., 2019). However, including QIIME2, metagenomic data analysis workflow for 16S amplicon sequences mainly focused on sequences from short reads sequencers such as MiSeq. Researchers therefore cannot apply those workflows for sequences from long read sequencers, and thus a practical metagenome workflow for long read sequencers is really needed.

NanoGalaxy (<https://nanopore.usegalaxy.eu/>) is a webserver to process, analyze and visualize Oxford Nanopore Technologies (ONT) data and similar long-reads technologies (Koning et al., 2020), and also integrates metagenome analysis for long reads by Kraken2 (<https://github.com/DerrickWood/kraken2>), taxonomic classification system using exact k-mer matches to archive

high accuracy and fast classification speeds (Wood, Lu, & Langmead, 2019). Krona is used as a metagenomic visualization in NanoGalaxy, which allows intuitive exploration of relative abundances and confidences within the complex hierarchies of metagenomic classifications (Ondov, Bergman, & Phillippy, 2011).

However, a modification of the reference database is not easy in NanoGalaxy system. In addition, database build for Kraken2 is not successful in many cases because much memory, time and storage were required and database for the reference dataset is too huge to be handled in the Docker version of NanoGalaxy. Moreover, some 16S RNA reference dataset in NanoGalaxy seems to be applicable not only for the analysis of genus level but also for that of species level, but the database is not updated to be used for routine work. Furthermore, the visualization in NanoGalaxy is not suitable for comparisons among multiple samples. This workflow in NanoGalaxy is not suitable for practical use.

Therefore, we developed the rapid metagenomic workflow using an annotated reference 16S RNA sequence set for Kraken2 to allow species-level identification and their comparison.

Results

Due to COVID-19, we were forced not to hold the BioHackathon in 2021 despite the Olympic games was held in Tokyo, Japan. Therefore, a domestic version of the BioHackathon called BH21.8 was hold instead in Aomori, Japan (23-27 August 2021). During BH21.8, we discussed the following points about the workflow of metagenome data.

1. Reproducible workflow for metagenome analysis
2. Practical use case of the workflow
3. Maintaining annotated reference 16S RNA sequence dataset

Reproducible workflow for metagenome analysis.

We made a reproducible workflow for metagenome analysis of 16S amplicon sequences. We first construct 16S RNA reference sequence dataset called Ref16S for the analysis.

Procedure to construct 16S RNA reference dataset

16S RNA sequence data in Ref16S is originally from DNA Data Bank of Japan (DDBJ), a member of International Nucleotide Sequence Database Collaboration (INSDC) (Fukuda, Kodama, Mashima, Fujisawa, & Ogasawara, 2021). DDBJ collects nucleotide sequence data and provides freely available nucleotide sequence data and supercomputer system, to support research activities in life science. Procedure to construct Ref16S is as follows. The BioPython library for python is required to run the scripts below.

1. Compressed version of 16S.fasta downloaded from DDBJ FTP site at `ftp://ftp.ddbj.nig.ac.jp/ddbj_database/16S/16S.fasta.gz`
2. Accession numbers of INSDC (DDBJ/ENA/NCBI) in headers of the FASTA file above collected (`convert_accession2taxonomy.py`)
3. Using Entrez to create 'accession-taxonomy' correspondences
4. `kraken:taxid|(corresponding tax id)`(Example: `>kraken:taxid|3017012|... ..`) added to the head of headers of the FASTA file (`add_kraken_taxonomy.py`)

The latest version of constructed Ref16S can be downloaded from <https://doi.org/10.18908/1sdba.nbd02583-001>. Details of the data repository is described in the chapter 'Maintaining annotated reference 16S RNA sequence dataset'.

Building library

Using Ref16S constructed, we can build library for Kraken2. Procedure to build this library is as follows. While the size of Ref16S is around 200MB, total size of this library for Kraken2 is around 34GB. This is why whole set of libraries is not provided in FTP site.

1. Set working directory

```
% DBNAME="/home/{yourname}/kraken2/"
```

2. Downloading required files (taxonomy and so on: names.dmp and nodes.dmp)

```
% kraken2-build --download-taxonomy --db $DBNAME
```

3. Installing reference library

```
% kraken2-build --add-to-library 16S_kraken2.fasta --threads {n} --db $DBNAME
```

You may see error message for dustmasker. We now ignore this error.

4. Building database. Specify {n} for the number of threads to use.

```
% kraken2-build --build --threads {n} --db $DBNAME
```

5. Classification & output. FASTQ file (sample1.fq.gz) in fq directory. The classification report will be written to report.txt.

```
% kraken2 --db $DBNAME --gzip-compressed fq/sample1.fq.gz --threads {n} --report report.txt --output -
```

6. Visualization. Multiple results from kraken2 above can be included and compared in this command. jinja2 and pandas libraries for python are required to run the script.

```
% python karaken2_report_formatter.py -n sample_name_1 sample_name_2 -f sample_report_1 sample_report_2
```

Because some additional steps are needed to be used in Kraken2, the complete commands were summarized as reproducible workflow. The workflow is integrated into existing data analysis workflow called Systematic Analysis for Quantification of Everything (SAQE; <https://github.com/bonohu/SAQE>) (Bono, 2021).

Practical use case of the workflow

The constructed workflow was compared with existing data analysis pipeline in metagenome analysis.

Data resources used in the comparative analysis

- Ref16S originally DDBJ 16S data from ftp://ftp.ddbj.nig.ac.jp/ddbj_database/16S/
- MiniKraken 2: <https://ccb.jhu.edu/software/kraken2/downloads.shtml>
- RefSeq_bacteria: Kraken 2 Custom Database (RefSeq bacterial genome) <https://github.com/DerrickWood/kraken2/wiki/Manual#custom-databases>
- HumGut: Human gut genome collection <https://github.com/larssnip/HumGut>

Comparison of various databases for metagenome analysis

We used human gut metagenome data from 16S RNA amplicon sequencing by MinION to test our workflow.

Metagenome sequence data from MinION sequencers in Matsuo *et al.* (SRA Run ID: DRR225 048) was used as a test data (Matsuo *et al.*, 2021).

Stacked charts were depicted using our self-made visualization tool for 4 reference data described above (Figure 1)

Ref16S has more detailed taxonomy mapping than other indexes especially in species level (Figure 1D) while not so great differences in Ref16S, MiniKraken2 and refseq_bacteria in order, family and genus levels (Figure 1 A,B,C).

Figure 1: Figure for genus level comparison

Figure 1. Comparison among the output of data analysis pipelines. Stacked charts in the level of (A) Order, (B) Family, (C) Genus and (D) Species.

Silva and RDP which are preinstalled in NanoGalaxy, can be analyzed only in genus level. On the other hand, Greengenes also preinstalled in NanoGalaxy can be analyzed in species level, but it uses outdated classification terms (<https://greengenes.secondgenome.com/>). Standard and MiniKraken, which are also preinstalled in NanoGalaxy, can be analyzed in species level. But those databases may not be updated and the latest version of those. We can check those status from links in the Kraken2 website (<https://ccb.jhu.edu/software/kraken2/index.shtml?t=downloads>). According to the website, the latest version of MiniKraken is 3/27/2020, while that in NanoGalaxy is 7/13/2019.

In conclusion, all databases are needed to be install by users to analyze metagenomic data locally. To analyze locally, it takes so much time and is computationally expensive to build databases for Standard and MiniKraken, which consist of genomic sequences. Ref16S only contains 16S RNA sequences from various bacteria and thus it does not take so much time compared with whole genome sequences. It therefore would be relatively easy to update on a regular basis.

Maintaining annotated reference 16S RNA sequence dataset

Large data is very hard to maintain, and users often should request authors for download. It is often obstacle to carry out the research.

Because the file size of Ref16S is around 200MB and not small enough to be handled in GitHub archive, we put that in the Life Science Database Archive (<https://dbarchive.biosciencedbc.jp/index-e.html>) in National Bioscience Database Center (NBDC) in Japan Science and Technology Agency (JST). This enables researchers to replicate the workflow very rapidly. The data file is archived with Digital Object Identifier (DOI) and accessible from URL below.

- 16S RNA sequence set for metagenomic analysis (Database Description) <https://doi.org/10.18908/lbdba.nbd02583-000>
- 16S RNA sequence set for metagenomic analysis (Data) <https://doi.org/10.18908/lbdba.nbd02583-001>

Currently, the version of the data is V001, but we plan to update those because the reference 16S RNA sequence set is updating.

Conclusion

After a stimulus discussion in BH21.8, we could clarify the current issues in the metagenome data analysis. We also could successfully construct a rapid workflow for those data specially from long reads by using newly constructed Ref16S. Ref16S will make it possible to measure composition of bacteria on the fly.

Future work

We will continue current effort to update Ref16S to analyze increasing metagenome sequences. We plan to make fungus version of annotated reference 18S RNA sequence dataset in the near future for metagenome analysis of fungi.

Workflow created

The workflow is integrated into existing data analysis workflow called Systematic Analysis for Quantification of Everything (SAQE; <https://github.com/bonohu/SAQE>).

Acknowledgement

We thank Prof. Kiichi Hirota in Kansai Medical University for stimulus discussion in the initial stage of this study. We deeply thank organizers of BH21.8 (domestic biohackathon in Japan) at Aomori (23-27 August 2021) for giving us a chance to discuss Ref16S dataset. We also thank Shigeru Yatsuzuka in JST NBDC for prompt response to host our dataset in the Life Science Database Archive in JST NBDC.

References

- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8), 852–857. doi:[10.1038/s41587-019-0209-9](https://doi.org/10.1038/s41587-019-0209-9)
- Bono, H. (2021). Meta-Analysis of Oxidative Transcriptomes in Insects. *Antioxidants (Basel, Switzerland)*, 10(3), 345. doi:[10.3390/antiox10030345](https://doi.org/10.3390/antiox10030345)

Fukuda, A., Kodama, Y., Mashima, J., Fujisawa, T., & Ogasawara, O. (2021). DDBJ update: Streamlining submission and access of human data. *Nucleic Acids Research*, 49(D1), D71–D75. doi:[10.1093/nar/gkaa982](https://doi.org/10.1093/nar/gkaa982)

Koning, W. de, Miladi, M., Hiltemann, S., Heikema, A., Hays, J. P., Flemming, S., Beek, M. van den, et al. (2020). NanoGalaxy: Nanopore long-read sequencing data analysis in Galaxy. *GigaScience*, 9(10). doi:[10.1093/gigascience/giaa105](https://doi.org/10.1093/gigascience/giaa105)

Matsuo, Y., Komiya, S., Yasumizu, Y., Yasuoka, Y., Mizushima, K., Takagi, T., Kryukov, K., et al. (2021). Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. *BMC microbiology*, 21(1), 35. doi:[10.1186/s12866-021-02094-5](https://doi.org/10.1186/s12866-021-02094-5)

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1), 385. doi:[10.1186/1471-2105-12-385](https://doi.org/10.1186/1471-2105-12-385)

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1), 257. doi:[10.1186/s13059-019-1891-0](https://doi.org/10.1186/s13059-019-1891-0)