International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

General Analysis Process of Bacterial Communities

Hyun Seung Kong

Undergraduate Student, Department of Systems Biology, College of Life Science and Biotechnology, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, South Korea

Abstract: ***This paper generalizes the process of analysis bacterial communities from tissues and natural environment for people who are not familiar with the analysis using the computer. This paper is a short article consisting of introduction and procedure. Described in the Introduction what is bacterial community analysis, procedure describes the techniques used in each process: DNA extraction, Sequencing, Data trimming, Taxonomic assignment, diversity indices and Taxonomic analysis.***

Keywords: systems biology, sequencing, trimming, taxonomic analysis, microbial community

1. Introduction

In many areas of life sciences, analyses of bacterial communities play an important role. Comparative analysis of the microbial community and microbial community composition may be an important clue in explaining the effects caused by any of the microorganisms. The high demand of low cost sequencing promoted the development of the next generation sequencing (or high throughput sequencing) that is capable of producing millions of sequences at one time, increasing the producibility of sequencing process by thousands times.

In particular, these microbes cause disease in humans and they cause changes in the ecosystem through interaction with the environment. Since such bacterial communities analysis require large amount of data processing, systems biological method is used.

This paper was written for those who are not familiar with the analysis method using a computer and generalize the method to find out procedures.

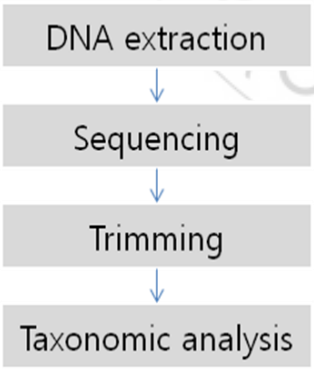

Figure 1: Flowchart of procedure

Following this general procedure, you can easily analyze microbial communities.

2. Procedure

2.1 DNA extraction

Extraction/isolation methods would much different depending on samples. For the sequencing, however, DNA must be quantificated. The amount of DNA extract is not enough: you cannot correctly represent the real environment. So it is necessary a large amount of the DNA as much as possible at this stage.

2.2 Sequencing

We used the next g eneration sequencing (NGS). Sequences produced by the next-generation sequencing are shorter in length than the sequences which are produced by the sanger sequencing. But, the amount of sequences many more. There are a number of NGS sequencing such as Illumine-Solexa sequencing and Roche/454 pyrosequencing. After sequencing is finished, the sequencing files will have been created. The file extension is probably the NGS data format: fasta, fastq, fq, csfast, SFF.

2.3 Data trimming

Trimming is the removal process of potential vector contamination and/or poor quality parts of reads. There are several methods for trimming: quality trimming, ambiguity trimming, adapter sequence trimming, base trimming and length trimming.

In quality trimming, if the sequence files contain quality scores from modified-Mott trimming algorithm, this information can be used for trimming sequence ends.

Ambiguity trimming is removal of stretches of Ns meanwhile, adapter sequence trimming removes sequence adapters. adapter sequence is a kind of barcode sequences that are attached on the sequencing process. So the adapter sequence is different depending on the type of sequencing.

Other methods are base trim (remove a specified number of bases at either 3' or 5' end of the reads) and length trimming (remove reads shorter or longer than a specified threshold). Generally, the program which called CLC workbench is used to trim sequence

2.4 Taxonomic analysis

Taxonomic analysis is the process of analyzing the distribution of the community by using the analysis sequence. The bacterial community is analyzed using the conserved sequence

of the 16s rRNA. Algorithm of Bayesian method is used for analysis. Generally, the program which called RDP classifier is used for this step. Using this data, the distribution of microorganisms can graph and display as a table. However, since there cannot be trusted if the reliability of the data is low, the reliability verification process is required.

2.5 Taxonomic assignment & Diversity indices

This process is reliability verification process. After Trimming and Taxonomic analysis, undergo Taxonomic assignment. The Taxonomic assignment is to calculate the richness and evenness, coverage and rarefaction curve. The meanings are as follows:

Richness: Refers to the richness of species. The number of species in the microbial community
- ACE
- Chao

Evenness: richness of species refers to the extent consistent with the environment. Mathematically, there are various types of calculation.
- Jackknife
- Shannon
- Simpson

Rarefaction curve: Rarefaction curve is drawn repeatedly to the average species count out a random sampling from 1 to N.

* Appendix: Formulas of diversity indices [ACE]

$$S_{ace} = S_{abund} + \frac{S_{rare}}{C_{ace}} + \frac{F_1}{C_{ace}}\gamma^2_{ace}$$

[Chao]

$$S_{chao1} = S_{obs} + \frac{n_1(n_1-1)}{2(n_2+1)}$$

[Jackknife]

$$S_{jack2} = S_{obs} + \left[\frac{Q_1(2m-3)}{m} - \frac{Q_2(m-2)^2}{m(m-1)}\right]$$

[Shannon]

$$H_{shannon} = -\sum_{i=1}^{S_{obs}} \frac{n_i}{N} ln \frac{n_i}{N}$$

[Simpson]

$$D_{simpson} = \frac{\sum_{i=1}^{S_{obs}} n_i(n_i-1)}{N(N-1)}$$

[Rarefaction curve]

$$f_n = E[X_n] = K - \binom{N}{n}^{-1} \sum_{i=1}^{K} \binom{N-N_i}{n}$$

References

[1] T.A.Brown, Gene cloning & DNA analysis: an introduction, 6th edition, Wiley-Blackwell publication, 2010, chapter 3 p.26-27
[2] Greg Gibson & Spencer W.Muse, A primer of Genome Science, 3rd edition, Sinuar Associates, Inc. Publisher, 2008, Chapter 2 p.65-83
[3] "Taxonomic analysis" Available: https://rdp.cme.msu.edu [Accessed: Feb. 10, 2015].
[4] Ik-Young Choi, "The principle and application of NGS to genome biology"