AARD
Allergy Asthma Respir Dis 6(5):255-262, September 2018 https://doi.org/10.4168/aard.2018.6.5.255
pISSN: 2288-0402
eISSN: 2288-0410
ORIGINAL ARTICLE

분석 방법 변화에 따른 음성 대조군과 호흡기 검체 간 미생물 구성 차이 비교

김효정,[1] 이상표,[2] 강신명,[2] 강성윤,[2] 정성원,[3,4] 이상민[2]

[1]가천대학교 일반대학원 융합의과학과, [2]가천대학교 길병원 호흡기알레르기내과, [3]가천대학교 의예과 유전체의과학전공, [4]가천대학교 길병원 가천유전체의과학연구소

Comparison of differences in microbial compositions between negative controls and subject samples with varying analysis configurations

Hyojung Kim,[1] Sang Pyo Lee,[2] Shin Myung Kang,[2] Sung-Yoon Kang,[2] Sungwon Jung,[3,4] Sang Min Lee[2]

[1]Department of Health Sciences and Technology, GAIHST, Gachon University, Incheon; [2]Division of Pulmonology and Allergy, Department of Internal Medicine, Gachon University Gil Medical Center, Incheon; [3]Department of Genome Medicine and Science, College of Medicine, Gachon University, Incheon; [4]Gachon Institute of Genome Medicine and Science, Gachon University Gil Medical Center, Incheon, Korea

Purpose: Identifying microbial communities with 16S ribosomal RNA (rRNA) gene sequencing is a popular approach in microbiome studies, and various software tools and data resources have been developed for microbial analysis. Our aim in this study is investigating various available software tools and reference sequence databases to compare their performance in differentiating subject samples and negative controls.

Methods: We collected 4 negative control samples using various acquisition protocols, and 2 respiratory samples were acquired from a healthy subject also with different acquisition protocols. Quantitative methods were used to compare the results of taxonomy compositions of these 6 samples by varying the configuration of analysis software tools and reference databases.

Results: The results of taxonomy assignments showed relatively little difference, regardless of pipeline configurations and reference databases. Nevertheless, the effect on the discrepancy was larger using different software configurations than using different reference databases. In recognizing different samples, the 4 negative controls were clearly separable from the 2 subject samples. Additionally, there is a tendency to differentiate samples from different acquisition protocols.

Conclusion: Our results suggest little difference in microbial compositions between different software tools and reference databases, but certain configurations can improve the separability of samples. Changing software tools shows a greater impact on results than changing reference databases; thus, it is necessary to utilize appropriate configurations based on the objectives of studies. (*Allergy Asthma Respir Dis 2018;6:255-262*)

Keywords: Microbiota, Metagenome, Computational biology

서 론

실험실에서 배양할 수 있는 미생물은 전체 미생물의 1% 미만으로, 미생물을 분리하고 배양하는 기존의 미생물 동정방법으로는 확인할 수 있는 미생물의 종류가 매우 제한적인 것으로 알려져 있다.[1] 차세대염기서열분석(next generation sequencing, NGS) 기술의 발달로, 다양한 미생물들의 DNA로부터 각 미생물들을 직접 동정하는 것이 가능하게 되어 이를 이용한 메타지노믹스(metage-

Correspondence to: Sungwon Jung https://orcid.org/0000-0001-6002-554X
Gachon Institute of Genome Medicine and Science, Gachon University Gil Medical Center, 38-13 Dokjeom-ro 3beon-gil, Namdong-gu, Incheon 21565, Korea
Tel: +82-32-458-2740, Fax: +82-32-458-2725, E-mail: sjung@gachon.ac.kr

Co-correspondence to: Sang Min Lee https://orcid.org/0000-0002-9568-2096
Division of Pulmonology and Allergy, Department of Internal Medicine, Gachon University Gil Medical Center, 21 Namdong-daero 774beon-gil, Namdong-gu, Incheon 21565, Korea
Tel: +82-32-458-2713, Fax: +82-32-469-4320, E-mail: sangminlee77@naver.com

• This work was supported by the Gachon University Gil Medical Center (grant number: 2016-18).

Received: April 2, 2018 **Revised:** May 2, 2018 **Accepted:** May 9, 2018

nomics)가 발달하게 되었다.[2] 인체에 존재하는 미생물 역시 상당수가 배양이 불가능한 것으로 추정되고 있으나,[3] NGS 기술의 발달로 다양한 미생물을 확인할 수 있게 되면서 인체의 생명현상과 다양한 미생물 간 연관성에 대한 연구가 활발히 진행되고 있다. 다양한 연구들을 통해 인체는 공존하고 있는 미생물들과 서로 영향을 주고받는 것으로 알려져 있으며,[4] 인체와 공존하며 상호작용하는 미생물군과 해당 미생물들의 유전정보 전체를 일컬어 마이크로바이옴(microbiome)이라고 한다.[5,6] 인체 내 미생물의 구성이 사람마다 다르다는 특징을 가진다는 측면에서 마이크로바이옴을 인체의 유전정보(genome)에 비유하여 '제2의 유전체(second genome)'로 표현하기도 한다.[7]

16S ribosomal RNA 서열은 계통유전학적으로 보존적인 염기서열을 가지고 있다고 알려져 있으며, 미생물 군집 구성을 확인하기 위한 연구에서 보편적으로 사용되고 있어,[8] 체계화된 분석 방법과[9] 다양한 분석 도구가 존재한다.[10,11] 마이크로바이옴 분석에 사용되는 참조 염기서열 데이터베이스의 경우, 데이터베이스 간 미생물종 정리 체계, 등록된 표준 염기서열 수, 데이터베이스 관리 및 업데이트 시점 등의 차이로 인해 데이터베이스 간 편차가 존재하는 것으로 알려져 있으며,[12] 참조 염기서열 데이터베이스 간 편차는 분석 결과에 영향을 미칠 수 있을 것으로 보인다. 마이크로바이옴 분석의 각 단계에 사용되는 프로그램들의 차이 역시 다음 단계 결과에 영향을 줄 수 있는 가능성이 존재하기 때문에, 서로 다른 프로그램으로 시행된 분석 결과 간 차이가 존재할 가능성이 있다.[13,14]

이에 이 연구에서는 서로 다른 특성을 가진 샘플의 16S rRNA 서열을 여러 조합의 참조 염기서열 데이터베이스들과 분석 도구들을 사용하여 분석하고, 분석 결과 간 차이를 정량적으로 확인하고자 하였다.

Table 1. List of samples and their acquisition protocols

Sample ID	Acquisition protocol
Negative control 1 (NC1)	Normal saline
NC2	Protected brush → Normal saline
NC3	Protected brush → Bronchoscopy → Normal saline
NC4	Bronchoscopy → Normal saline
Subject 1 (S1)	Protected brush → Bronchoscopy → Brushing → Normal saline
S2	Bronchoscopy → Bronchial washing

Normal saline and protected brushes which used for sample acquisition were sterilized. Four negative controls were derived from using multiple sample acquisitions that are normal saline (NC1), immersing protected brush in normal saline (NC2), immersing protected brush which through the bronchoscopy in normal saline (NC3), and washing bronchoscopy channel with normal saline (NC4). Two subject samples were acquired from a healthy subject using protected brush (S1) and bronchial washing (S2).

대상 및 방법

1. 샘플 구성

분석에 사용한 샘플은 음성대조군(negative control, NC) 4개, 호흡기검체(subject sample, S) 2개로, 총 6개로 구성되었다. 호흡기검체는 기관지내시경 채널을 통해 검체 보호용 솔(protected brush)을 통과시켜 정상기관지 점막을 찰과(brushing)하여 얻은 검체와, 동측에 대해 시행한 기관지세척(bronchial washing)으로 획득한 기관지세척액을 사용하여 검체를 획득하였다. 음성대조군 샘플은 각각 멸균생리식염수(NC1), 검체 보호용 솔을 담근 멸균생리식염수(NC2), 기관지내시경 채널을 세척한 멸균생리식염수(NC3), 기관지내시경 채널을 통과시킨 검체 보호용 솔을 담근 멸균생리식염수(NC4)로부터 획득하였다(Table 1). 해당 샘플들을 서로 다른 구성의 파이프라인에 적용하여 분석을 진행하였다.

호흡기검체는 기관지 내 병변을 배제하기 위해 기관지내시경을 시행한 환자로부터 얻어졌으며, 기관지내시경 시행 전 서면으로 연구에 대한 설명이 이루어졌고, 연구에 대한 서면동의서를 획득하였다(IRB number: GAIRB2017-065). 2개의 호흡기검체는 한 명의 subject로부터 채취한 것으로, 병력상 특이사항이 없고 기관지내시경 소견상 병변이 관찰되지 않는 환자를 대상으로 하였다(Table 2). 무작위 배정에 따라 폐좌하엽 전기저부 기관지분절(anterobasal segment of left lower lobar bronchus)에서 정상기관지 점막을 찰과한 후 기관지세척을 시행하여 채취하였다.

2. 16S rRNA gene 시퀀싱

여러 가지 프로토콜을 사용하여 획득한 검체로부터 미생물의 DNA를 추출하기 위해 Mo Bio Laboratories, Inc. (Carlsbad, CA, USA)의 PowerSoil DNA Isolation Kit를 사용하였고, 전체 DNA 추출 과정은 제공되는 제품 프로토콜(Instruction manual version 07272016)에 따라 진행되었다.

16S rRNA 서열은 여러 세대 후에도 염기서열의 변화가 많지 않다는 특징을 가지며, 해당 서열 내에 종 특이적 염기서열을 가지는

Table 2. Clinical characteristics of subject

Variable	Status
Sex	Male
Age (yr)	54
History of the present illness	Negative
Relevant past medical history	Negative
Smoking (pack-years)	40
Chest computed tomography	Stricture of laterobasal segment of right lower lobar bronchus
Bronchoscopy	Benign bronchial stricture without lesions on laterobasal segment of right lower lobar bronchus

초가변영역(hypervariable region)이 존재한다. 16S 영역의 초가변영역은 9개가 존재(V1–V9)하는 것으로 알려져 있으며, 이 연구에서는 V3–V4 영역을 증폭하여 Illumina 사의 MiSeq 시퀀싱 기기를 사용하여 16S rRNA 서열 데이터를 확인하였다.

3. 16S rRNA gene 서열에 기반한 미생물 종 분류 분석

16S rRNA 서열을 이용한 마이크로바이옴 분석은 전처리 단계, 군집 분석 단계, 미생물 종 분류 단계로 수행되며, 필요에 따라 다양성 분석 등을 할 수 있다.

전처리 단계에서는 양질의 서열을 얻기 위해 염기서열을 양방향으로 읽은 서열정보를 하나의 서열로 만드는 작업과, 키메라를 포함한 분석 결과에 영향을 줄 수 있는 시퀀스상의 오류를 제거하는 작업이 시행된다. 군집 분석은 샘플 내 존재하는 미생물의 종류와 그 비율을 알기 위해 염기서열들을 특정한 기준에 따라 분류하는 과정으로, 접근 방식에 따라 미생물 유전체 표준 염기서열을 사용한 방법(reference-based)과 서열 간 유사도에 근거하여 분류하는 방법(similarity-based)이 존재한다. 미생물 종 분류단계에서는 군집 분석을 통해 분류된 염기서열들을 참조 염기서열 데이터베이스의 표준 염기서열(reference sequence)을 기준으로 어떤 종에 해당되는지 확인하게 된다.

16S rRNA 서열의 종 분류 분석을 위해 소프트웨어 구성을 달리한 두 가지 파이프라인(파이프라인 1, 파이프라인 2)을 준비하였으며, 두 파이프라인은 서로 다른 분석 도구를 사용해서 분석된 결과 간 차이를 확인하기 위해, 전처리 단계와 군집 분석 단계, 그리고 미생물 종 분류 단계에 적용되는 분석 도구를 다르게 하여 구성하였다. 각 파이프라인은 다음과 같이 구성하였다.

1) 파이프라인 1의 구성

파이프라인 1의 전처리 단계는 PANDAseq[15]과 VSEARCH[16]로 구성되었으며, 각각 양방향의 서열을 하나로 취합하는 과정과 미생물 유전체 표준 염기서열 기반의 키메라 서열 확인 및 제거 과정에 사용되었다. 군집 분석 단계는 VSEARCH 알고리즘을 적용하여 분석하고자 하는 서열들을 대상으로 미생물 유전체 표준 염기서열을 기반으로 분류한 후, 분류되지 않은 서열들을 서열 간의 유사도에 근거하여 분류하는 방법으로 구성되었다. 미생물 종 분류는 마이크로바이옴 통합 분석도구인 QIIME[17] (version 1.9.1)을 통해 분석이 진행되었으며, 이 과정에서 UCLUST[18] 알고리즘을 사용하고, 서열 참조 데이터베이스로 Greengenes[19]와 SILVA[20]를 선택적으로 사용하도록 구성되었다.

2) 파이프라인 2의 구성

파이프라인 2는 전처리 단계가 FLASH[21]와 CD-HIT[22]의 CD-HIT-DUP로 구성되었다. 파이프라인 2에서는 키메라 서열 제거에 샘플 내 서열 간 비교를 통한 방법을 적용하였기 때문에, 파이프라인 2에서는 파이프라인 1과 달리 전처리 단계에서 참조 염기서열 데이터베이스를 사용하지 않도록 구성하였다. 군집분석 단계 또한 참조 염기서열 데이터베이스에 의존하지 않고, 샘플 내 존재하는 서열 간의 유사도에 근거하여 분류하는 CD-HIT 알고리즘을 적용하였다. 미생물 종 분류 단계는 파이프라인 1과 동일하게 QIIME을 이용하고, 서열 참조 데이터베이스를 선택적으로 사용하여 진행하도록 구성하였다.

4. 미생물 종 분류 결과의 정량분석

1) 미생물 종 비율에 기반한 Jensen–Shannon divergence

서로 다른 파이프라인과 데이터베이스를 사용하여 분석한 동일 샘플 사이 미생물 종 분류 결과 간 정량적 차이는 두 분석 결과의 미생물 종 비율에 기반한 Jensen-Shannon divergence (JSD) 계산을 통해 평가되었다. JSD는 두 확률 분포의 차이를 계산하는 방법인 Kullback-Leibler divergence의 비대칭성을 해결한 방법으로, 계산한 결과는 항상 양수이며 두 분포가 일치할 때 0의 값을 가진다. 서로 다른 파이프라인과 데이터베이스를 사용하여 분석한 미생물 종 분류 결과 간 차이를 확인하기 위해, 각 구성 사이 동일 샘플 간 미생물 구성비를 비교하여 서로 다른 분석과정이 동일 샘플 결과에 어느 정도의 차이를 유발하는지 평가하였다.

각 미생물 종이 하나의 미생물 종 분류 결과에서 차지하는 비율에 대해 다른 미생물 종 분류 결과에서 차지하는 비율의 차이를 두 미생물 분류 결과에 대해 교차하여 두 미생물 종 분포 간 거리를 대칭적으로 계산하였다.

2) 주성분분석(principle component analysis)

Principle component analysis (PCA)는 전체적인 샘플의 구분 경향을 확인하기 위해 시행하였다. 6개의 샘플을 서로 다른 구성의 파이프라인과 데이터베이스를 사용하여 분석한 결과로부터 얻은 종 구성 비율을 바탕으로, 24개(6개의 샘플, 2개의 파이프라인, 2개의 참조 염기서열 데이터베이스의 조합) 결과들의 분포를 확인하기 위해 PCA를 시행하였다.

3) 미생물의 계통 발생에 기반한 샘플 간 거리 계산에 따른 샘플 분포 분석

미생물의 계통발생학적 특성을 고려한 방법인 weighted-UniFrac[23] distance를 이용하여 계산된 샘플 간 종 비율 유사도는, 분석 대상들 사이의 유사성 또는 비유사성을 측정하여 각 분석 대상들의 거리를 좌표상에 표현하여 공간적으로 시각화하는 분석 방법인 다차원척도법(multidimensional scaling, MDS)을 이용하여 시각화하였다.

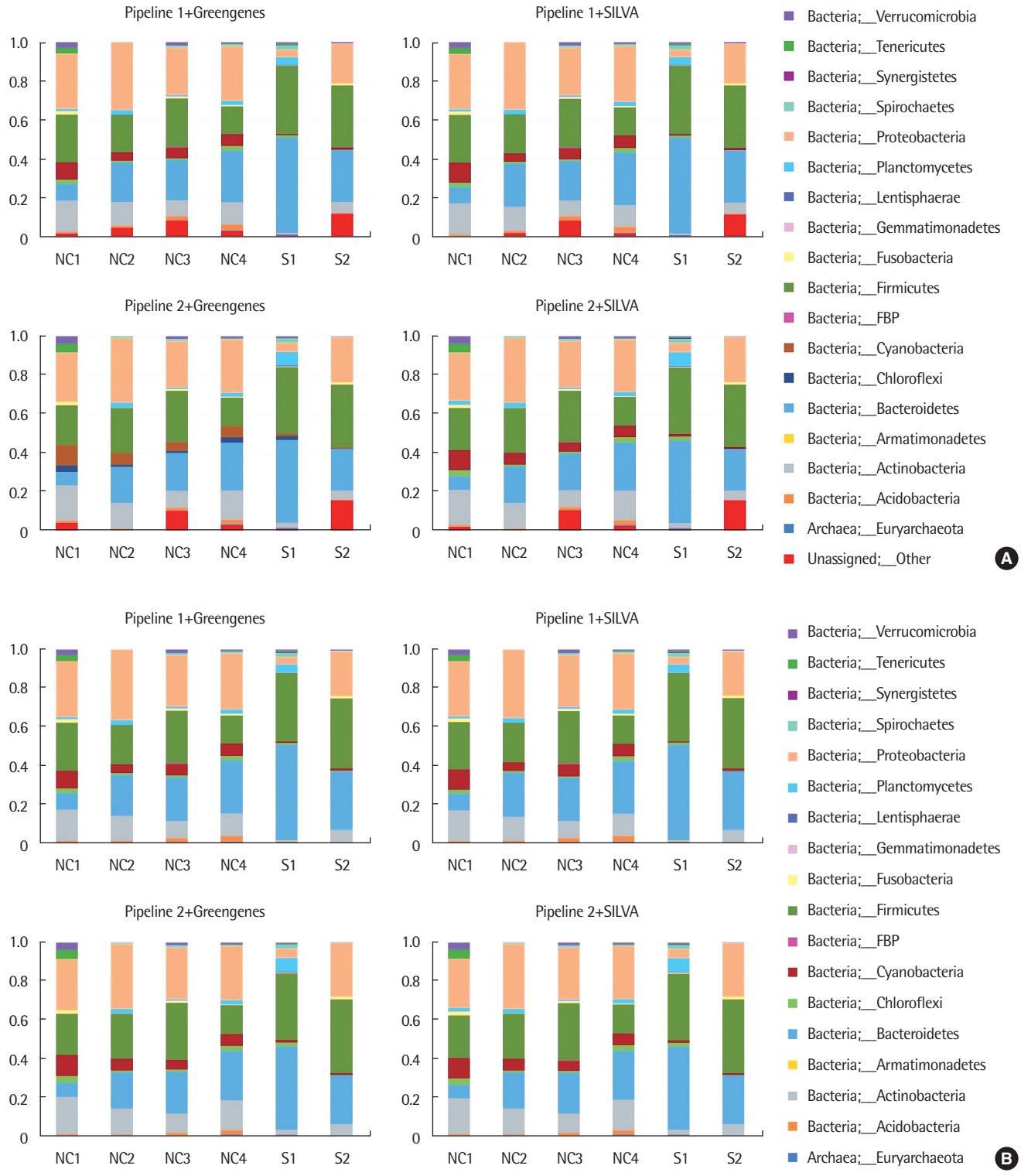

Fig. 1. Normalized compositions of common microbial communities of 6 samples, from each configuration of pipeline and reference sequence database. Compositions of common microbial communities are plotted with unassigned proportions (A), and without unassigned proportions (B). The presented compositions are in phylum level. NC, negative control; S, subject sample.

결 과

1. 전체 샘플의 종 분류 결과

각 샘플에서 얻은 16S rRNA gene 서열 데이터를 서로 다른 파이프라인과 데이터베이스를 사용하여 각 샘플의 미생물 종 분류를 확인하였다. 사용한 파이프라인과 데이터베이스에 따라 미생물 분류 결과에서 발견되는 미생물 종들이 완전히 일치하지는 않으므로, 공통적으로 관찰되는 미생물을 대상으로 각 샘플 내에서의 비율을 정규화(normalization)하여 비교하였고(Fig. 1A), 각 방법의 결과가 종을 판별하지 않는 서열의 양에 의해 편향되는 것을 줄이기 위한 관점에서 어떤 종으로도 분류되지 않은 결과(unassigned)를 제외한 후 정규화한 결과도 함께 확인하였다(Fig. 1B). 미생물 구성비 비교를 위한 공통 미생물 종 선별 과정에서 제외된 미생물의 구성 비율은 파이프라인과 데이터베이스 조합에 따른 네 가지 분석 방법 중 파이프라인 2와 SILVA를 함께 사용한 경우 8%로 가장 높게 나타났고, 각 샘플의 경우에는 최대 13% (NC4), 최소 0% (NC2)로 확인되어, 분석에 제외된 미생물 종들의 경우 전체 미생물 구성에서 전반적으로 낮은 비율을 차지하는 것으로 확인되었다.

전체 샘플의 공통된 미생물에 대한 구성비는 파이프라인과 데이터베이스의 차이가 있음에도 unassigned 비율의 포함 여부와 관계없이 유사한 비율을 보이는 경향이 관찰되었다.

2. 동일 샘플 간 구성비의 정량적 비교

각 파이프라인과 참조 염기서열 데이터베이스의 조합에 따른 결과 차이를 보다 정량적으로 비교하기 위하여, 서로 다른 분석 방법을 사용한 동일 샘플의 미생물 구성비 간 차이를 계산하였다. 정량적 비교 결과, 서로 다른 파이프라인과 데이터베이스 구성에 의한 결과들 사이의 차이는 일관되게 낮은 것이 확인되었다(Fig. 2A). 또한 모든 동일 샘플의 미생물 구성비에서 unassigned 비율을 제외하였을 때 더 낮은 차이를 보였다. 파이프라인의 차이와 unassigned 비율 포함 여부에 관계없이, 동일한 파이프라인에 상이한 데이터베이스를 적용한 결과에 비해 상이한 파이프라인에 동일한 데이터베이스를 적용한 결과가 상대적으로 큰 차이를 보이는 것을 확인하였다(Fig. 2B).

3. 서로 다른 검체 특성을 가진 샘플 구분 경향 분석

6개의 샘플을 두 가지 파이프라인(파이프라인 1, 파이프라인 2)에 두 가지 참조 서열 데이터베이스(Greengenes, SILVA)를 적용하여 분석한 총 24개의 샘플 종 분류 결과들의 주성분분석(PCA)을 시행하여 파이프라인 차이, 참조 서열 데이터베이스 차이, 검체 종류에 따른 전체적인 샘플의 구분 경향을 확인하고자 하였다.

두 가지 파이프라인에 두 가지 미생물 데이터베이스를 조합하는 네 가지 분석 방법은 공통적으로 unassigned 비율 포함 여부와 관계없이 음성대조군(NC)과 호흡기 검체(S)의 구분이 가능하였다(Fig. 3).

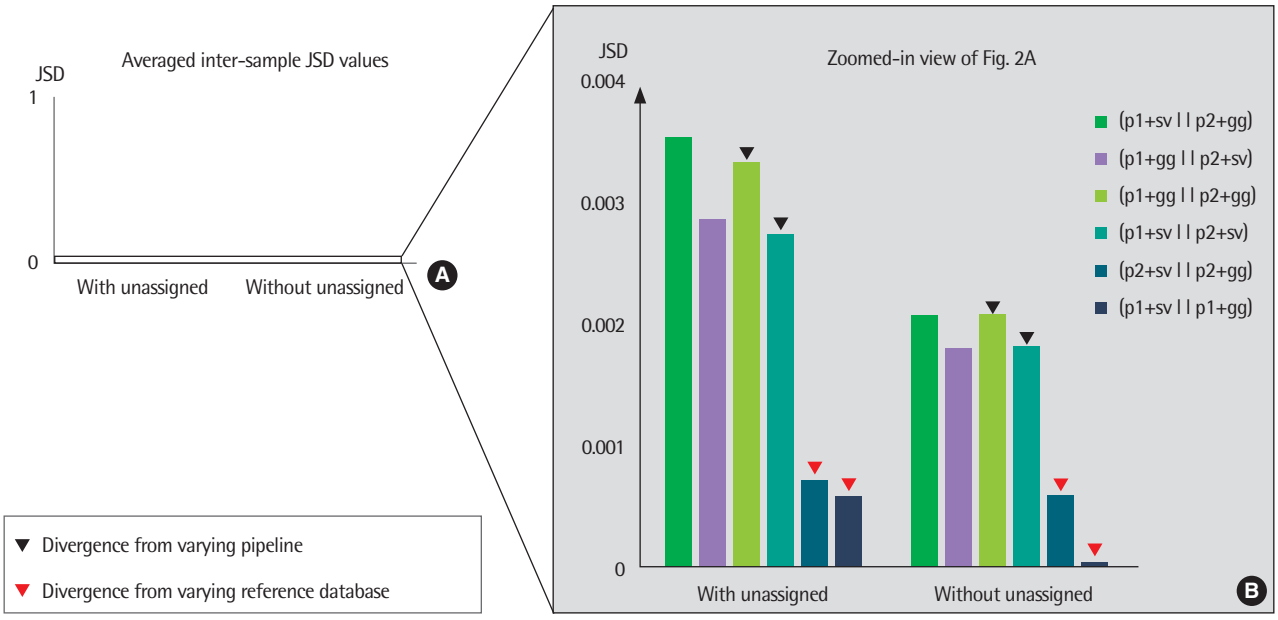

Fig. 2. A bar plot of averaged Jensen-Shannon divergence (JSD) values between the results of 2 different configurations of pipelines and reference sequence databases. The averaged JSD values were computed with and without unassigned portions. The averaged JSD values between 0 to 1 were plotted (A) and zoomed to observe the details (B). Markers indicate the averaged JSD from varying pipeline (black), and from varying database (red). p1, Pipeline1; p2, Pipeline2; gg, Greengenes; sv, SILVA.

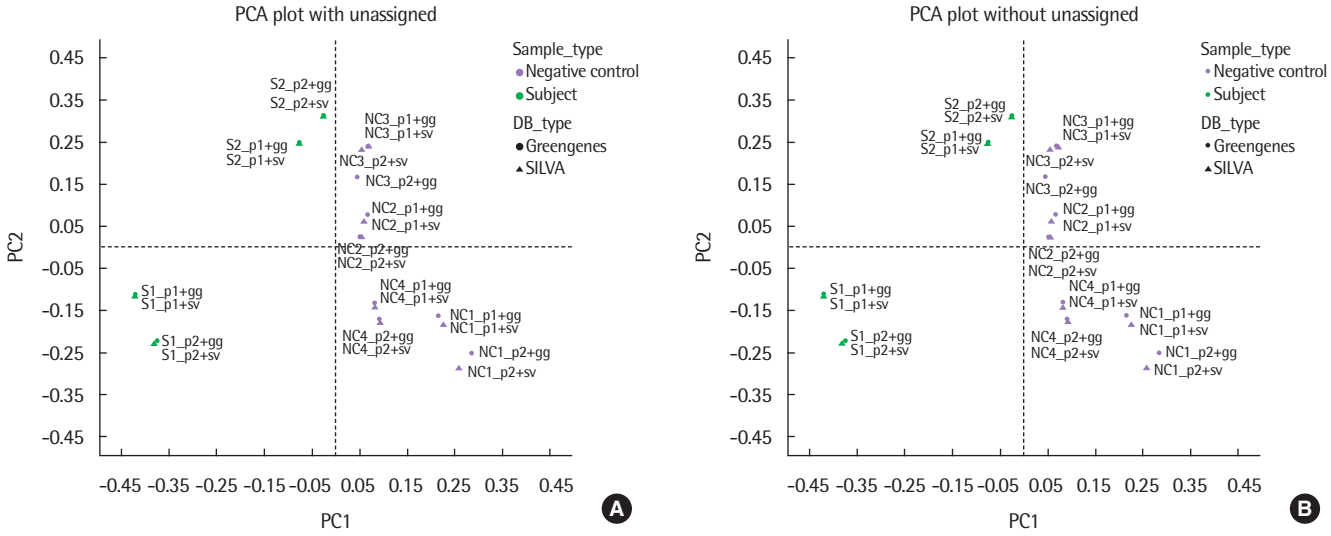

Fig. 3. Principle component analysis (PCA) plots of all samples using different pipelines and databases (with and without unassigned). NC and S are separable either with (A) or without unassigned portion (B). NC, negative control; S, subject sample. p1, Pipeline1; p2, Pipeline2; gg, Greengenes; sv, SILVA.

종 비율만이 아닌, 계통발생학적 연관성을 고려한 미생물 종 비율을 비교하기 위하여, 미생물의 계통발생학적 특성을 고려한 방법인 weighted-UniFrac distance를 이용하여 각 파이프라인과 데이터베이스의 조합으로 구성된 분석 방법을 적용한 결과별 샘플 간 종 비율의 유사도를 계산하였다. 구성이 다른 분석 방법을 사용한 결과별로 얻은 샘플 간 weighted-UniFrac distance에 의한 샘플 구분의 시각화를 위하여, 데이터 간 거리 matrix가 있는 경우 시각화가 가능한 MDS를 이용하였다. 그 결과, 단순 종 비율에 기반한 PCA를 한 경우와 유사하게 음성대조군(NC)과 호흡기검체(S)가 구분되는 것을 확인하였다(Fig. 4).

NC와 S 사이 weighted-UniFrac distance의 평균을 통해 파이프라인의 차이에 따른 NC와 S 샘플 구분 정도를 확인하였고(Fig. 5), 파이프라인 1을 사용한 NC와 S의 거리가 더 큰 차이를 보였으며 이를 통해 특정 샘플군 구분에 보다 유리한 파이프라인 구성의 발견이 가능하였다.

고 찰

서로 다른 참조 염기서열 데이터베이스와 파이프라인 차이에 따른 결과 사이에 발생하는 차이의 정량적 계산을 통해, 분석 방법의 차이에 관계없이 낮은 차이를 보이는 것을 알 수 있었다. JSD의 평균을 지표로 하여 산출된 동일 샘플 간 구성비의 정량적 비교 결과는 최대 0.0035의 차이를 보여, 동일 샘플 간 미생물 구성비는 파이프라인과 데이터베이스 변화에도 절대적인 구성 비율의 차이가 크지 않았다. 그러나, 동일 샘플에 대해 서로 다른 파이프라인과 데이터베이스를 사용한 결과 간 정량적 비교에서, 데이터베이스의 변화에 따른 차이보다는 파이프라인의 변화에 따른 차이가 큰 것을 확인하였다. 16S rRNA 서열을 사용한 마이크로바이옴 분석에서, 참조 염기서열 데이터베이스보다 파이프라인 구성 중 데이터 전처리와 군집분석 단계에 사용된 소프트웨어의 차이가 분석 결과에 미치는 영향이 더 큰 것으로 보인다.

전체 샘플의 공통 미생물 종 구성비에 대하여 시행한 주성분 분석의 결과를 통해, 분석에 사용한 파이프라인, 데이터베이스와 관계없이 모든 경우에서 unassigned 비율의 포함 여부와 무관하게, 음성대조군과 호흡기 검체 간 구분 경향이 관찰되었다. 이는 사용한 파이프라인과 데이터베이스의 변화에 관계없이, 검체 특성에 따른 샘플 구분이 가능한 것으로 해석할 수 있다.

공통 미생물의 구성비에 대한 분석에서 제외한 미생물의 구성비를 포함한, 각 분석 방법별 결과의 전체 미생물 구성비에 대한 검체 구분 정도를 확인하고자, 미생물의 계통발생학적 특성을 고려한 방법으로 각 샘플의 미생물 구성비 차이를 계산하고(weighted-UniFrac distance), 다차원척도법(MDS)을 사용해 각 경우별 샘플 구분 경향을 분석하였다. 각 경우별 샘플 구분 경향 분석을 통해, 이 연구에서 사용한 모든 데이터베이스와 파이프라인의 조합에 대해 검체 특성에 따른 샘플의 구분이 가능함을 확인할 수 있었다.

각 분석방법별 샘플 구분 경향 분석에서 검체 획득 방식에 따른 구분 경향의 차이를 확인할 수 있었다. 모든 분석방법 공통적으로 검체 보호용 솔(protected brush)을 사용한 S1 샘플이 음성대조군(NC)과의 구분이 보다 명확한 경향을 보이는 것이 확인되었고, 검체 보호용 솔을 사용하지 않은 S2의 경우에는 음성대조군과의 구분 정도가 비교적 덜한 경향을 보이는 것이 관찰되었다(Fig. 4). 이는 동일 subject의 동측 정상기관지 점막으로부터 얻어진 두 호흡기

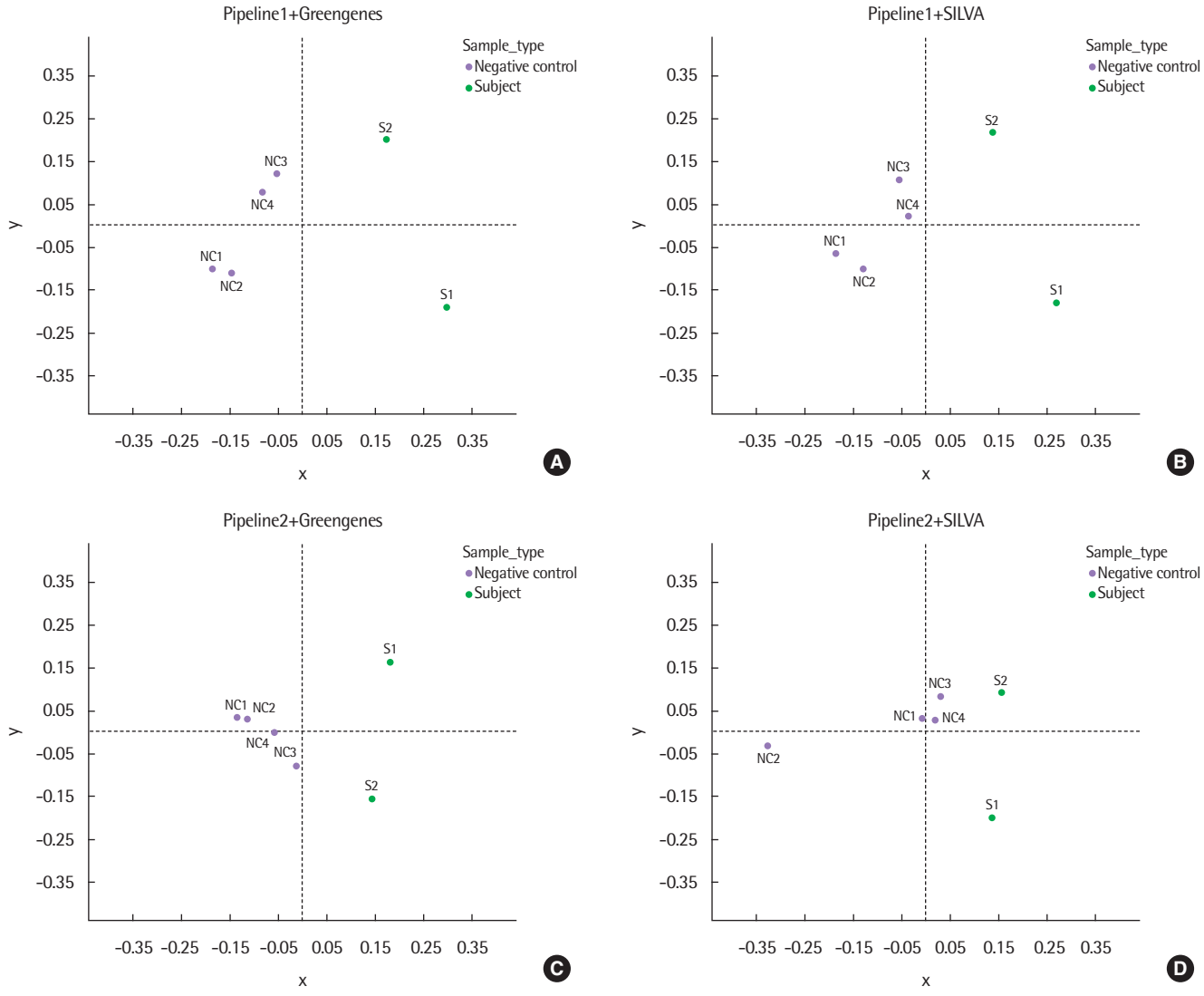

Fig. 4. Multidimensional scaling (MDS) plots of 6 samples from each configuration of pipelines and databases based on the weighted-UniFrac distance. Pipeline1 with Greengenes (A), Pipeline1 with SILVA (B), Pipeline2 with Greengenes (C), and Pipeline2 with SILVA (D). NC, negative control; S, subject sample.

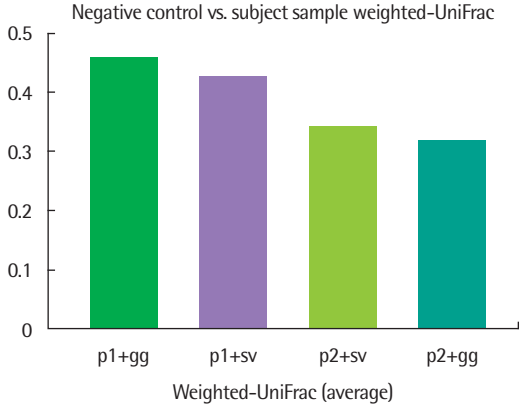

Fig. 5. Average weighted-UniFrac distance. The average of weighted-UniFrac distance between negative control and subject sample were sorted in descending order. p1, Pipeline1; p2, Pipeline2; gg, Greengenes; sv, SILVA.

검체 샘플이 음성대조군에 대해 보이는 샘플 구분 경향 정도의 차이는 검체 획득 방식의 차이에 의한 것으로 보인다. 또한, 음성대조군 샘플 중 기관지내시경을 사용한 NC3과 NC4가, 기관지내시경을 거치지 않은 NC1과 NC2에 비해 S2와 가까운 경향을 보이는 것이 공통적으로 관찰되었다. 음성대조군과 호흡기검체가 기관지내시경 사용에 의해 가까워지는 것을 확인할 수 있었으며, 기관지내시경을 사용한 호흡기 마이크로바이옴 검체 획득 과정에서 기관지내시경 채널 환경의 가능성을 배제할 수 없을 것으로 보인다.

서로 다른 검체 특성을 가진 샘플 간 weighted-UniFrac으로 계산된 거리의 평균을 확인한 결과, 파이프라인 1이 샘플 특성에 따른 구분 정도가 더 명확한 것으로 나타났다. 파이프라인 간 검체 특성 구분 능력의 차이는, 전처리 단계에 사용된 분석 소프트웨어의 구성 차이에 의해, 각 파이프라인의 군집분석 단계에 사용된 염기

서열의 수에 차이가 존재하기 때문인 것으로 보인다.

이 연구에 사용된 전체 샘플의 수가 적기 때문에, 검체 획득 방법에 따른 결과 해석의 일반화에는 한계가 있다. 향후, 검체를 추가적으로 획득하여 분석할 예정이며, 이 연구에서 생물 분류 단위를 문(phylum)을 기준으로 분석을 수행한 것 외에, 더 세분화된 생물 분류 단위에 대한 추가 분석으로 보다 일반화된 결론을 내릴 계획이다.

이 연구에서는 파이프라인과 데이터베이스 변화에 따른 미생물 구성비율의 차이와 검체 특성 구분 능력의 차이를 정량적 방법으로 확인하였고, 이를 통해 분석 방법에 관계없이 기관지내시경을 이용한 검체 획득 프로토콜 차이에 따라 두 호흡기검체가 음성대조군과 구분되는 정도에 차이를 보이는 것을 확인하였다. 기관지내시경만 사용한 검체에 비해, 검체 보호용 솔(protected brush)을 이용하여 얻어진 검체가 음성대조군과 구분이 잘 되는 것을 확인할 수 있었다. 16S rRNA 서열을 이용한 호흡기 마이크로바이옴 연구에서 기관지내시경 이용한 검체 획득 시, 검체 보호용 솔(protected brush)을 사용한 프로토콜의 사용이 검체 획득 과정에 발생할 수 있는 주변 환경의 영향을 줄일 수 있으며, 검체 획득 시 주변 환경의 영향을 배제할 수 없는 경우에는 이 연구에서 시사하는 바와 같이, 분석하고자 하는 데이터의 특성과 분석 목표를 고려한 마이크로바이옴 분석 파이프라인 구축 및 사용이 동반되어야 할 것으로 보인다.

REFERENCES

1. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 1985;39:321-46.
2. Zoetendal EG, Collier CT, Koike S, Mackie RI, Gaskins HR. Molecular ecological analysis of the gastrointestinal microbiota: a review. J Nutr 2004;134:465-72.
3. NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. The NIH Human Microbiome Project. Genome Res 2009;19:2317-23.
4. Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 2006;124:837-48.
5. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell 2012;148:1258-70.
6. Lederberg J, McCray AT. 'Ome sweet' omics: a genealogical treasury of words. Scientist 2001;15:8-10
7. Grice EA, Segre JA. The human microbiome: our second genome. Annu Rev Genomics Hum Genet 2012;13:151-70.
8. Weinstock GM. Genomic approaches to studying the human microbiota. Nature 2012;489:250-6.
9. Morgan XC, Huttenhower C. Chapter 12: human microbiome analysis. PLoS Comput Biol 2012;8:e1002808.
10. Armougom F, Raoult D. Exploring microbial diversity using 16S rRNA high-throughput methods. J Comput Sci Syst Biol 2009;2:74-92.
11. Clarridge JE 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 2004;17:840-62.
12. Balvociute M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? BMC Genomics 2017;18(Suppl 2):114.
13. Schloss PD. Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. mSystems 2016;1(2). pii: e00027-16.
14. Plummer E, Twin J, Bulach DM, Garland SM, Tabrizi SN. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. J Proteomics Bioinform 2015;8:283-91.
15. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics 2012;13:31.
16. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016;4:e2584.
17. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7:335-6.
18. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26:2460-1.
19. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 2006;72:5069-72.
20. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013;41(Database issue):D590-6.
21. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011;27:2957-63.
22. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658-9.
23. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005;71:8228-35.