TSPE16020-007-12

균류 LSU 표준염기서열 DB기반 미생물다양성 및 환경유전체 분석기술 개발

Development of microbial diversity and environmental genome analysis method based on fungal LSU sequence DB

한국생명공학연구원

제 출 문

극지연구소장 귀하

본 보고서를 “기후변화에 의한 킹조지섬 생태계 변화 연구”과제의 위탁연구 “균류 LSU 표준염기서열 DB기반 미생물다양성 및 환경유전체 분석기술 개발에 관한 연구”과제의 최종보고서로 제출합니다.

2017. 1. 23

(본과제) 총괄연구책임자 : 홍 순 규

위탁연구기관명 : 한국생명공학연구원

위탁연구책임자 : 박 두 상

위탁참여연구원 : 최 한 나

“ : 김 옥 희

“ : 전 준 형

“ : 정 선 아

“ : 호 리 한

보고서 초록

<table>
<tr><td>위탁연구과제명</td><td colspan="5">균류 LSU 표준염기서열 DB기반 미생물다양성 및 환경유전체 분석기술 개발</td></tr>
<tr><td>위탁연구책임자</td><td>박두상</td><td>해당단계 참여연구원수</td><td>6명</td><td>해당단계 연구비</td><td>60,000,000원</td></tr>
<tr><td>연구기관명 및 소속부서명</td><td colspan="2">한국생명공학연구원 생물자원센터</td><td>참여기업명</td><td colspan="2">없음</td></tr>
<tr><td>국제공동연구</td><td colspan="5">상대국명 : 없음 상대국연구기관명 : 없음</td></tr>
<tr><td colspan="4">요약</td><td>보고서면수</td><td>43</td></tr>
<tr><td colspan="6">본 위탁과제에서는 본과제 ‘기후변화에 의한 킹조시섬 생태계 변화 연구’의 일환으로, 남극미생물생태연구에 필요한 일련의 생물정보학 프로그램 및 DB 개발에 목표를 둠. 이번 3차년도 연구에서는 극지에서 얻어진 환경시료에 대한 메타유전체 및 메타전사체 염기서열 데이터를 이용하여 metabolic potential을 분석하는 시스템을 구축함

1. 메타지놈 내 특이 유전자 클러스터링 분석 및 빈도 비교 분석 파이프라인 구축
● 유전체상 one-copy 유전자의 relative abundance을 기준으로 메타지놈 유전자 빈도 정규화 기능 구현
● 유전자별 FPKM값 계산 기능 구현
● 자동화된 메타지놈 유전자 클러스터링 분석 및 빈도 비교 분석 파이프라인 구축
● 서열 클러스터링 프로그램 논문 발표 (PLoS ONE 11:e0151064)

2. Metabolic pathway 관련 유전자 빈도 비교 분석 파이프라인 구축
● 주요 대사 경로 별 유전자 빈도수 계산 방법 도입
● KEGG enzyme별 빈도수 계산 방법 도입
● 서로 다른 환경 샘플 별 대사 발현 차이 분석 기능 구현
● 환경 샘플 메타데이터 연동 기능 제공
● 물질대사 관련 유전자 빈도 비교 분석 파이프라인 구축 완료</td></tr>
<tr><td rowspan="2">색 인 어
(각 5개 이상)</td><td>한 글</td><td colspan="4">메타유전체, 물질대사, 생물정보, 유전자빈도, 클러스터링</td></tr>
<tr><td>영 어</td><td colspan="4">metagenome, metabolism, bioinformatics, gene abundance, clustering</td></tr>
</table>

요 약 문

Ⅰ. 제 목

균류 LSU 표준염기서열 DB기반 미생물다양성 및 환경유전체 분석기술 개발

Ⅱ. 연구개발의 목적 및 필요성

- 미생물다양성 및 환경유전체 분석에 대한 생물정보학적 연구는 본과제의 세부목표인 “생태계 융합연구 기반기술 개발”의 성공적 수행을 위한 중요 요소 중 하나임
- 본과제에서는 남극 환경 샘플로부터 “미생물 군집”과 “물질대사 관련 유전자” 분석을 통해 기후변화에 따른 남극생태계 변화를 biogeochemical 수준에서 이해하고자 함. 특히, 남극 육상환경에서 중요하게 작용할 것이라고 예상되는 미생물의 기능 중 탄소순환과 질소순환에 관여하는 유전자 종류 및 발현을 metagenomics 및 metatranscriptomics 기법을 통해 분석할 예정임
- 본 과제에서는 습지, 건조 고지대, 구조토, 담수, 연안 등을 포함한 장기모니터링 20지점을 선정하여 생태계 구성 기본 요소 중 하나인 미생물과 기후, 지형, 지화학 특성 간 상호작용을 이해하고자 함
- 이를 위해 본 과제에서는 환경시료 채취 후 whole DNA를 추출한 뒤, pyrosequencing등을 이용하여 특정 분자 마커에 대한 대용량 미생물 염기서열 데이터를 생산할 예정임. 또한, 환경시료 내 전체 유전자들에 대해 Illumina등을 이용하여 시퀀싱을 수행할 예정임
- 따라서 본과제에서 생산된 next generation sequencing 기반 대용량의 미생물 군집 및 whole genome shotgun 데이터를 처리할 수 있는 생물정보학적 인프라 구축 및 지원이 필수적임
- 이에 본 위탁과제에서는 미생물군집 분석 및 shotgun 메타전사체 염기서열 분석과 관련된 일련의 생물정보학적 알고리즘, 프로그램, 파이프라인을 개발하고자 함

- 주어진 환경 내 존재하는 모든 유전자 서열을 시퀀싱하는 whole genome shotgun 메타전사체 염기서열을 처리할 수 있는 자동화 시스템을 구축할 예정임
- 차세대 염기서열 결정법으로 얻어진 원시데이터의 전처리, assembly, 구조 분석, 기능 분석 등을 수행하는 파이프라인 구축
- 주석처리된 유전자 서열들을 물질대사경로와 비교하여 환경의 metabolic potential을 평가할 수 있는 시스템 구축
- 또한, 균류 표준염기서열 DB를 수정 및 보완함으로서 남극 미생물군집 연구의 "dark matter"라 할 수 있는 토양균류 다양성 및 생태연구에 기여하고자 함

Ⅲ. 연구개발의 내용 및 범위

〈제1주제: 메타지놈 내 특이 유전자 클러스터링 분석 및 빈도 비교 분석 파이프라인 구축〉

- RNA-Seq 데이터에서 사용하는 FPKM (fragment per kilobase of transcript per million fragments)의 개념을 차용하여, *recA*, *gyrB*, *rpoB* 및 EF-Tu 등 유전체 내 1 copy만을 지니는 유전자들과 관심있는 특이 유전자들의 reference 유전자 염기서열의 단위 길이당 match되는 read 수를 파악하여 해당 시료 내 특이 유전자를 지니는 미생물의 빈도수를 계산함
- 임의로 단편화되어 align이 되지 않는 메타지놈 reads를 clustering하기 위

하여 해당되는 각 read의 염기서열에 표준화된 정보값을 부여하고 이를 character value로 인지하여 clustering을 수행할 수 있는 체계를 구현함. 이로부터 시료 내 유전자 그룹별 빈도를 파악할 수 있는 자동화 분석 시스템 개발함

- 메타지놈 서열로부터 유전자의 빈도수 계산을 위한 프로그램을 개발하고, 일련의 과정들을 서로 연동하여 standalone 수준에서 분석 파이프라인을 구축함

〈제2주제: metabolic pathway 관련 유전자 빈도 비교 분석 파이프라인 구축〉

- 확장형 KEGG database의 염기서열과 메타지놈 염기서열 함께 CLUSTOM 알고리즘을 이용하여 clustering한 후, 각 cluster에서 KEGG ID를 추출하고 *rpoB*, *recA* 유전자에 match된 수 및 각 reference 유전자 그룹의 size로 정규화하여 메타지놈 내 각 metabolism에 관여하는 유전자들의 빈도수 계산 방법을 구축함
  - 구축된 방법을 이용하여 KEGG ID에 따른 빈도수를 자동으로 계산하는 프로그램을 구현하고 table 형식으로 도출할 수 있도록 설계함
  - Hypergeometric distribution, t-test등의 통계적 방법을 이용하여 샘플별 enriched된 세부 metabolic pathway 정보 표현
  - KEGG ID 및 EC number를 선택하여 해당하는 metagenome reads를 다운로드할 수 있도록 설계함
  - 분석 시료의 메타데이터(시료 채취 날짜, 온도, 위도 및 경도, biogeochemical data 등)를 연동하여 보여줄 수 있도록 설계함

Ⅳ. 연구개발결과

메타지놈 내 특이 유전자 클러스터링 분석 및 빈도 비교 분석 파이프라인 구축

- 유전체상 one-copy 유전자의 relative abundance을 기준으로 메타지놈 유전자 빈도 정규화 기능 구현
- 유전자별 FPKM값 계산 기능 구현
- 자동화된 메타지놈 유전자 클러스터링 분석 및 빈도 비교 분석 파이프라인 구축
- 서열 클러스터링 프로그램 논문 발표 (PLoS ONE 11:e0151064)

Metabolic pathway 관련 유전자 빈도 비교 분석 파이프라인 구축

- 주요 대사 경로 별 유전자 빈도수 계산 방법 도입
- KEGG enzyme별 빈도수 계산 방법 도입
- 서로 다른 환경 샘플 별 대사 발현 차이 분석 기능 구현
- 환경 샘플 메타데이터 연동 기능 제공
- 물질대사 관련 유전자 빈도 비교 분석 파이프라인 구축 완료

V. 연구개발결과의 활용계획

- 세균에 비해 균류의 경우 표준염기서열 DB가 없어 다양성 연구에 어려움이 많음. 따라서 본 과제 수행을 통해 확충 및 업데이트된 LSU 표준염기서열 DB은 균류 다양성 연구를 위한 국제적 표준으로 자리 잡을 것으로 기대함
- 메타지놈 염기서열 분석은 데이터의 방대함으로 인해 고성능의 컴퓨터 자원이 없는 경우 일반 연구자가 접근하기 힘듬. CLOUD 기반의 메타지놈 분석 프로그램 개발은 시스템 메모리 부족을 해결할 수 있는 좋은 학문적 모델이 될 것으로 기대함
- 잘 정립되어 있는 미생물다양성 염기서열 분석 방법에 비해 shotgun 메타지놈 분석은 아직까지 분석 protocol에 대한 consensus가 없는 실정임. 본 과제 수행을 통해 속도 및 정확도를 고려하여 구축될 메타지놈 분석 파이프라인이 일반 생물학자들에게 데이터 분석 표준을 제시할 것으로 기대함
- 물질대사 측면에서 서로다른 환경유전체 샘플을 비교하는 생물정보학적 도구가 부족한게 현실임. 본 과제를 통해 구축될 'web system for evaluating metabolic potential'은 계층구조적 물질대사경로 별 서로 다른 환경 샘플의 metabolic potential를 비교하고, 특정 환경에 specific 또는 enrich된 유전자를 결정해 줄것으로 기대함
- 최근 NGS 기술의 발전으로 유전체학 및 생태학 분야에서 대량의 데이터가 생산되고 있음. 데이터 크기로 인해 단일 연구실에서 분석이 불가능함. 따라서, 본 과제에서 구축 및 개발된 프로그램 및 pipelines을 발전시켜 상용 분석 서비스 제공이 가능함
- 균류 rDNA 레퍼런스 DB가 구축은 정확한 균류 동정 및 환경 내 균류 다양성 정보를 제공하게 되어, 식물병원균류/환경부후균류/생리활성물질생성균류 연구에 기반이 될 것임. 또한, 향후 농업, 환경, 의료, 보건 등의 산업분야에서의 높은 활용도를 가질 것으로 예상됨
- 구축될 메타지놈 분석 프로그램과 물질대사 평가 시스템은 급변하는 전 지구적 기후변화 패턴을 미생물 수준에서 monitoring 하기 위한 기본 도구로 이용될 수 있음

S U M M A R Y

(영 문 요 약 문)

I. Title

Development of microbial diversity and environmental genome analysis method based on fungal LSU sequence DB

II. Purpose and Necessity of R&D

- Bioinformatics research on analysis of microbial diversity and environemntal genomics is one of main goals of the KORPI biological project, focusing on Antarctic regions.

- The main projectin from KORPI is trying to understand ecological change of Antartic ecosystems at the biogeochemical level, by analyzing microbial community and metabolic potential.

- For that purpose, field work including many different natural extreme habitates was conducted and data at the molecular level has been accumulated.

- In this study, we aim to develop a series of bioinformatics tools to facilitate the analysis of those high-throughput DNA(RNA) data for environmental samples.

- Specially, this includes tool development for analysis of microbial community, shotgun metagenome and metatranscriptome.

- The automated pipeline consists of sequence preprocesses, sequence assembly, structural & functional annotations, connect with metabolic pathways, and etc.

- Finally, we expect that this work contributes to more accurately evaluate microbial diversity and their ecological role of the extreme polar regions.

III. Contents and Extent of R&D

<Topic-1: Clustering analysis of specific genes in metagenome and pipeline implementation for comparison of gene relative abundance>

- Calculating relative abundance of individual genes that are calibrated by the abundance of single-copy genes such as *recA*, *rpoB*, etc.
- development of efficient hash table to quickly cluster high-throughput shot reads
- developing programs that calculate relative gene abundance, at a standalone level

<Topic-2: Establishment of Pipeline for Comparative Analysis of Gene Relative abundance Related to Metabolic Pathway>

- After sequence clustering using CLUSTOM followed by building connection with the KEGG database, develop a method that calculates the relative gene abundance that are linked to metabolic pathways.
- Developing a method that tabulates a series of information including gene names, abundance and pathways involved.
- Developing a method that evaluates the statistical significance of gene enrichment
- Establishment of a possible way to integrate sequence information and its metadata

Ⅳ. R&D Results

1. Developing a pipeline for clustering genes from metagenomic sequences and for comparing their gene relative abundance

● Developing a normalization method of gene relative abundance by calibrating their abundance with single-copy genes

● Developing gene abundance in a similar way of FPKM value

● Developing a automated pipeline for clustering high-throughput shotgun metagenome sequences and for calculating gene relative abundance

● Publication of the clustering program (PLoS ONE 11: e0151064)

2. Implementing a pipeline for comparing gene abundance regarding metabolic pathways

● A method development for relative gene abundance from major metabolic pathways

● Calculation of gene abundance per KEGG enzyme

● Implementation of statistically comparing gene expression between different environmental samples

● Integrating other biotic or abiotic metadata with the sequence information

● Completion of the pipeline development for evaluating metabolic potential of given environmental samples

V. Application Plans of R&D Results

- In case of Fungi compared to prokaryotes, there is no standard sequence database for taxonomy. Therefore, it is anticipated that LSU sequence database we here develop will be established as an international standard for studying fungal diversity.
- Analysis of metagenome sequences is difficult for biologists to access if there is no high-performance computer resources due to the vast amount of data. Thus, development of CLOUD-based metagenome analysis program is expected to be a good academic resource to figure out the system memory shortage.
- Methodologies related to shotgun metagenome analyses do not yet consensus because of lack of well-established protocols compared to analysis of microbial diveristy. The metagenome analysis pipeline to be constructed, considering both speed and accuracy, will provide data analysis standard to biologists.
- There is a lack of bioinformatics tools to compare different environmental genomic samples in terms of metabolism. The bioinformatics system for evaluating it to be constructed through this work is expected to compare the metabolic potential of different samples by referring hierarchical metabolic pathways (e.g., KEGG) and to determine specific or enriched genes in a given environment.
- Recent developments in NGS technologies have produced a large amount of sequence data in genomics and ecology. Due to the large size of the data, analysis in a single laboratory is impossible. Therefore, well-designed, user-friendly software should be provided and available.
- Establishment of a fungal rRNA reference DB will provide accurate fungal identification and taxonomy that are useful to evaluate fungal diversity in a given environment. That will be the basis for a number of studies related to Fungi, including biomass-producing, plant pathology, human health, etc.
- Metagenome analysis programs as well as assessment of metabolic potential could be well performed by a series of programs that are developing in this study.

목 차

제 1 장 서론

제 1-1 절: 환경유전체 metabolic potential 분석 시스템 개발 필요성

1. 지난 과제 수행을 통해 Whole metagenome shotgun 염기서열 분석 파이프라인을 구축하였지만 raw data quality control에 이은 read assembly, annotation을 수행하는 기본적인 단계까지만 구축이 되었다.
2. MOTHUR (Schloss et al. 2009), QIIME (Kuczynski et al. 2012) 와 같은 분석 파이프라인으로 많이 수행하는 종 다양성 분석을 whole metagenome shotgun 염기서열 분석에서도 MetaPhlAn2 (Truong et al. 2015) 이라는 분석 툴을 사용하여 가능하다.
3. 단일 유전자 접근법은 주어진 환경이 미생물 다양성 및 군집구조 분석의 수행이 가능하지만, 해당 환경에서 존재하는 미생물이 어떤 역할을 하는지, 그 미생물에 속한 유전자의 환경 내에서의 기능은 무엇인지는 알기 어렵기 때문에 미생물과 해당 환경과의 상호작용을 이해하는데 한계가 있다.
4. Whole metagenome shotgun 염기서열 분석은 기존의 단일 유전자 기반 (예. 16s rRNA) 분석에 비해 얻을 수 있는 유전자정보가 더 많고 이를 통해 유전자 기능 분석을 할 수 있다는 큰 장점이 있지만 기존 파이프라인에는 반영하지 못하였다.
5. 기존 pipeline에는 contig들을 특정 group으로 모으고 taxonomy를 부여하는 binning작업이 빠져있어서 이를 보완해야한다. binning을 함에 있어 기존에 유전체 서열이 밝혀지지 않은 종들을 살펴보기 위해 sequence similarity를 이용하기보다는 coverage와 composition을 이용한 방법을 도입해야할 필요성이 있다.
6. 각 유전자에 존재 유무와 annotation 뿐만 아니라 각 유전자에 mapping되는 read를 통한 유전자의 빈도와 normalization method의 도입이 필요하다.
7. 각 유전자에 해당하는 KEGG와 같은 metabolic pathway 정보를 연결시켜 단순한 개별 유전자의 환경적 차이뿐만 아니라 대사경로의 차이를 살펴볼 필요가 있다.
8. 따라서 로컬 서버 상에서 대용량 whole genome shotgun 메타지놈 데이터를 쉽게 처리 할 수 있을 뿐만 아니라 다양성 분석, 유전자 빈도 분석, metabolic pathway 비교 분석이 가능한 파이프라인이 필요하다.

제 2 장 국내외 기술개발 현황

제 2-1 절: 환경유전체 염기서열 분석 파이프라인 개발 현황

1. 최근 NGS 기술의 발전으로, 다양한 환경 샘플에 대해 싼 가격으로 deep sequencing 이 가능하게 됨에 따라, whole 메타전사체 shotgun 데이터를 분석할 수 있는 여건이 마련되었다. whole 메타전사체 shotgun 데이터는 단일 유전자 기반 data에서 분석하지 못하는 해당 환경에서 존재하는 미생물의 기능과 역할을 파악할 수 있기 때문에 그 필요성이 점점 증대되고 있다.
2. 현재까지 whole genome shotgun 기반의 메타지놈 데이터를 분석하기 위한 분석 파이프라인은 유럽에서 개발된 EBI Metagenomics Portal (Mitchell et al. 2015), 미국 연구그룹에서 개발된 MG-RAST (Wilke et al. 2016), IMG/MER4 (Markowitz et al. 2014) 가 각각 웹상에서 분석파이프라인을 제공하며, MEGAN6 (Huson et al. 2015) 가 GUI를 탑재한 PC 설치 프로그램을 제공하고 있다.
3. EBI Metagenomics Portal의 경우 metagenome과 metatranscriptome data를 분석 가능하며 현재 12,000여개의 metagenome data set을 보유하고 있다. Raw data quality control후에 RNA를 식별하고 작업을 rRNASelector (Lee et al. 2011) 를 사용하여 수행하며 16S rRNA분석과 rRNA가 masking된 read들의 분석을 나눠서 수행한다. 16S rRNA분석은 QIIME과 GreenGenes database (DeSantis et al. 2006) 를 활용하며 기능분석은 InterProScan (Jones et al. 2014) 과 InterPro database (Mitchell et al. 2014) 를 활용하여 수행한다.
4. MG-RAST의 경우 264,000여개의 metagenome dataset을 보유하고 있으며 Fraggenescan으로 feature prediction, Uclust로 protein clustering을 수행한다. Protein clustering후 COGs (Galperin et al. 2014)나 KEGG (Kanehisa et al. 2015), GO (Gene Ontology Consortium. 2015), SEED subsystems (Overbeek et al. 2014) 등 다양한 데이터베이스를 기반으로 annotation 정보를 제공해주나 protein family 분석에 있어 주로 사용되는 pfam (Finn et al. 2016) 이 아닌 FIGfams (Meyer et al. 2009) 을

사용한다.

5. IMG/MER4도 EBI Metagenomics, MG-RAST와 같이 웹 기반의 분석 파이프라인과 데이터 deposit이 가능한 portal이지만 이러한 웹 기반의 파이프라인은 사용자가 데이터를 업로드해야 해야 하기 때문에 대용량 데이터를 업로드하기 어려우며, 사용자가 자신의 모듈이나 외부 리소스를 탑재하지 못하는 단점이 존재한다.
6. MEGAN6의 경우 data deposit의 어려움을 해소하기 위해 기존 BLASTX에 비해 20,000배 정도 속도가 빠른 DIAMOND alignment (Buchfink et al. 2015) 를 사용하여 Raw data를 NR database에 mapping을 시킨 후 daa2rma tool을 사용하여 taxonomic and functional binning을 수행한 후 LCA algorithm으로 NCBI taxonomy를 부여한다. 또한 SEED, COG, KEGG database에 mapping하면 RMA(MEGAN Read Match Archive) format의 binary화한 결과 파일을 MEGAN server에서 사용가능하게 된다. 결과 RMA file을 가지고 기존의 network에 존재하는 다른 RMA file들을 이용하여 비교분석이 가능하다.
7. MEGAN6 의 경우 PC 상에서 설치가 가능하나 컴퓨터 리소스 한계로 인해 대용량 데이터를 다루기 힘들며, 외부 데이터베이스 리소스를 탑재할 수 없는 단점이 존재한다.
8. IMG/MER4, EBI Metagenomics portal, MG-RAST, MEGAN6 모두 metagenome dataset을 수집하고 웹 환경이나 GUI환경에서 분석을 제공하지만 pipeline을 유동적으로 사용할 수 없어 기존의 다양한 방식의 assembly, binning, taxonomic profiling tool들을 연결시킬 수 없다. 또한 각종 database의 분석 (특히 KEGG와 같은 metabolic pathway분석) 에서 보다 정확한 gene abundance normalization 방법이 중요하지만 raw count value를 사용하거나 단순한 방법을 사용하여 부정확한 결과를 보여준다.

제 2-2 절: 환경유전체 대사 포텐샬 분석 도구 개발 현황

1. Assembly

초기 메타지놈 assembly에는 single genome assembly에 사용되었던 tool들을 사용한다. 하지만 contig 길이가 매우 짧게 조립되었으며 다른 종끼리 섞여버리는 경우가 많이 발생하였다. 이를 극복하기 위해 크게 두 개의 방식을 사용하고 있는데 이는 reference-based assembly와 de novo assembly이다. 분석하고자 하

는 dataset에 따라 선택하여야 하지만 reference genome이 존재 하지 않는 많은 종들을 분석하기 위해서는 de novo assembly방식을 써야하지만 이에 따라 computational power가 더 많이 요구된다.

가. Reference-based assembly

reference genome들을 기본 틀로 사용하여 contig를 생성하는 방식이며 특정 genus나 species에 속한 genome들을 target으로 assembly를 수행가능하다. Newbler (Roche), MIRA4 (Chevreux et al. 2007), AMOS (Treangen et al. 2011), MetaAMOS (Treangen et al. 2013) 와 같은 tool들이 사용되고 있다. 이 방식은 계산 자원을 적게 먹고 잘 밝혀지고 조립된 genome에 한해서는 좋은 결과물을 얻을 수 있다. 또한 근연종을 이용할 수 도 있다.

나. De novo assembly

reference-based assembly와 반대로 reference genome 정보가 없이 assembly를 수행하는 방식이기 때문에 사전 정보가 없어도 된다는 장점이 있다. 하지만 계산 자원을 많이 써야하며 de-Brujin graph나 overlap theory같은 graph-based theory algorithm을 사용하게 된다. EULER (Chaisson et al. 2008), Velvet (Zerbiono et al. 2008), SOAP (Liu et al. 2012), Abyss (Simpson et al. 2009) 가 초기 단계에 많이 사용된 tool들이며, 굉장히 큰 memory과 긴 계산시간을 요구한다. 하지만 앞에 언급된 tool들은 single genome assembly를 위한 방식이기 때문에 metagenome assembly에서는 좋은 결과를 내지 못하였다. 특히 유사한 subspecies의 variation과 다른 종간의 genomic sequence의 유사함, sample내의 종들 간에 abundance차이 등으로 인해 graph에 kink나 branch가 발생하게 되고 부정확한 결과를 주었다.

다. 차세대 assembly tool

Meta-IDBA (Peng et al. 2011), MetaVelvet (Namiki et al. 2012), IDBA-UD (Peng et al. 2012), Ray Meta (Boisvert et al. 2012) 와 같은 tool들의 1세대 assembly tool의 문제점을 해결하기 위해 개발되었다. MetaVelvet과 Meta-IDBA는 k-mer frequency를 이용하여 de-Bruijin 의 kink를 수정했으며 k-mer threshold를 이용하여 graph를 subgraph로 나누었다. IDBA-UD 의 경우 불규칙적인 sequencing depth를 더 고려하였으며 multiple depth-relative k-mer threshold를 사용하여 low-depth나 high-depth에서의 잘못된 k-mer를 제거하였다. 마지막으로 Ray Meta의 경우에

도 k-mer와 depth를 이용하지만 de Bruijin subgraph를 수정하지 않고 heuristics-guided graph 탐색으로 assembly를 수행한다.

2. Binning

Binning은 read나 contig들을 각각의 genome으로 추정되는 group으로 모아서 이에 해당하는 taxonomy를 부여하는 것을 말한다. Binning 방법은 다루고자하는 정보에 따라서 크게 두 가지로 나뉘는데 Composition-based binning과 Similarity 또는 Homology-based binning이다.

가. Composition-based binning

각각의 genome들이 고유의 k-mer sequence 분포를 가지고 있다는 점을 이용하여 각각의 genome group으로 sequence들을 묶는 방식이다. TETRA (Teeling et al. 2004), S-GSOM (Chan et al. 2008), Phylopythia (McHardy et al. 2007), ESOM (Deng et al. 2000), ClaMS (Pati et al. 2011), CONCOCT (Alneberg et al. 2013) 와 같은 tool들이 해당 방식을 사용하였다. 이 방식은 짧은 길이의 read나 contig의 경우 정확도가 매우 떨어지는 단점을 가지고 있기 때문에 assemble이 된 contig나 scaffold 중에서 일정길이 이상만 (보통 1kb) 분석에 사용을 하게 된다.

나. Similarity-based binning

BLAST나 profile hidden Markov Model (pHMMs)같은 alignment algorithm을 이용하여 public database의 sequence나 유전자에 mapping하여 유사도를 바탕으로 group을 묶는 방식이다. CARMA (Krause et al. 2008), MetaPhyler (Liu et al. 2010), SOrt-ITEMS (Haque et al. 2009) 와 같은 tool들이 이 방식을 사용하며 IMG/MER 4, MG-RAST, MEGAN pipeline 역시 이 방식을 사용하고 있다. 이 방식 역시 단점을 가지고 있는데 sample내에 근연종이 매우 많을 경우 정확도가 떨어지며 mapping할 database가 없는 알려지지 않은 수많은 종들은 분석에서 제외가 된다.

다. 분석 algorithm에 따른 분류

사용한 algorithm에 따라 다시 두 가지로 나눌 수 있는데 ab initio unsupervised classifier와 supervised/training-based classifier이다. Unsupervised의 경우 사용자의 개입이 없이 algorithm만으로 grouping을 하게 되지만 supervised의 경우 classifier training 단계에서 사용할 data의 특성이

나 범주를 선정하는 등의 사용자의 개입이 필요하다. support vector machine (PhylopythiaS), hidden Markov models (PhymmBL, TETRA), self-organizing maps (ESOM), gaussian mixture models (CONCOCT) 과 같은 algorithm들이 사용된다. Supervised의 경우 높은 신뢰도의 data를 바탕으로 training을 수행하면 일반적으로 unsupervised 방식보다 높은 정확도를 보이지만 신뢰도가 높은 많은 양의 input data를 얻기 힘든 경우에는 unsupervised 방식이 더 적합하다.

3. Annotation

가. Gene calling

Annotation 단계에서의 주된 목적은 read나 assemble된 contig에서 gene을 찾아내는 것이며 이를 gene calling이라고 한다. Gene들은 coding DNA sequence (CDS), noncoding RNA gene 로 크게 구분되며 clustered regularly interspaced short palindromic repeats (CRISPRs)와 같은 regulatory element를 예측할 수도 있다. MetaGeneMark (Zhu et al. 2010), Metagene (Noguchi et al. 2006), Prodigal (Hyatt et al. 2010), Orphelia (Hoff et al. 2009), FragGeneScan (Rho et al. 2010) 과 같은 tool들이 ab initio gene prediction algorithm을 이용하여 CDS를 예측하고 있다. CRISPR element의 경우 CRT (Bland et al. 2007) 나 PILER-CR (Edgar 2007) 와 같은 tool로 살펴볼 수 있다. tRNA와 같은 noncoding RNA는 tRNAscan (Lowe and Eddy 1997) 을 사용한다. rRNA를 위해서는 IMG/MER의 경우 내부 rRNA model을 사용하고 MG-RAST는 SILVA (Quast et al. 2013), Greengenes, RDP (Cole et al. 2009) 와 같은 database를 이용한다.

나. Functional assignment

Gene calling으로 찾아낸 CDS의 function을 알아내기 위해 public database에서 homology기반으로 찾게 된다. 일반적으로 metagenome raw data가 매우 크기 때문에 이 단계에서도 계산비용이 매우 많이 들어간다. BLAST와 같은 sequence-similarity-based algorithm이 많이 쓰이며 multi-threading이나 parallel programming 방식으로 속도를 높일 수 있다. 많이 사용되는 public database로는 KEGG, SEED, eggNOG (Huerta-Cepas et al. 2015), COG/KOG 와 같은 functional annotation database가 있고 PFAM, TIGRFAM (Haft et al. 2013) 과 같은 protein

domain database가 있다. Interpro의 경우 여러 가지 database를 종합적으로 포함하고 있다.

4. Taxonomic profiling

초기에는 shotgun metagenome data에서도 amplicon sequencing에서와 같이 단일 유전자 기반(예, 16s rRNA)의 샘플에 대한 미생물 다양성 분석은 annotation을 수행하여 rRNA를 찾거나 rRNA만 선택적으로 선별하여 MOTHUR (Schloss et al. 2009), QIIME (Kuczynski et al. 2012) 와 같은 분석 파이프라인을 통해 수행되었다. 하지만 이 후 metagenome data에서 taxonomic profiling을 위한 tool이 많이 개발되었고 크게 두 개의 그룹으로 나눌 수 있다. dataset의 모든 sequence 정보를 사용하는 방식 (예, CLARK (Ounit et al. 2015), Genometa (Davenport et al. 2012), GOTTCHA (Freitas et al. 2015), Kraken (Wood and Salzberg 2014), LMAT (Ames et al. 2013), MEGAN, MG-RAST, the One Codex webserver, taxator-tk (Droge et al. 2014)) 과 marker gene set에 집중하는 방식 (예, MetaPhlAn, MetaPhyler, mOTU, QIIME, MOTHUR) 이다. 이 중에서도 CLARK와 Kraken은 taxonomy profile의 정확성이나 계산속도 면에서 좋은 성능을 보인다. 하지만 MetaPhlAn2는 종 수준에서 나아가서 strain 수준으로 profiling을 할 수가 있게 되었고 prokaryote뿐만 아니라 eukaryote까지 포함하는 marker sequence database를 구축했다. 자세히 보자면 Archaea 46,649개, Bacteria 767,167개, Virus 38,809개, Eukaryote 22,371개의 marker sequence를 구축하였다. 정확도는 Kraken을 앞섰고 계산속도 면에서도 유사한 수준으로 향상되었다.

5. Abundance normalization

shotgun metagenome 분석을 수행함에 있어 community의 functional capacity를 알아내고 비교하는 것은 매우 중요한 부분이다. Sequencing된 read들을 gene이나 functional marker database에 mapping하여 대략의 abundance를 얻게 된다. 하지만 raw read count는 각 sample의 sequencing depth에 따라 다르고 sample들 간의 비교를 위해서는 normalization 방법이 필요하다. 단순한 compositional normalization 방법은 sample내의 abundance의 총합으로 각 gene의 abundance를 나눈다. 하지만 compositional 방법은 하나의 abundance값이 다른 모든 gene의 abundance의 영향을 받게 되고 각 gene의 특성이나 차이를 반영하지 못한다는 단점이 있다. 이를 해결하기 위해

GAAS (Angly et al. 2009), Raes et al. (Raes et al. 2007), Average genome size (Beszteri et al. 2010), MUSiCC (Manor and Borenstein 2015) 과 같은 method들이 개발되었다. 그중에서 MUSiCC은 76개의 universal single-copy gene정보와 machine learning method를 사용하여 intra-sample과 inter-sample 간의 bias를 줄였다. Universal single-copy gene들의 KEGG, GC contetnt, Consevation, Annotation 특성들을 이용하여 elastic-net regularization을 통해 linear model을 learning하여 sample내의 gene들의 abundance값을 normalization한다. 또한 sample간의 비교를 위해 모든 abundance의 총합으로 나누는 방법이 아니라 Universal single-copy gene의 median abundance를 사용하여 나눠서 normalization을 수행한다. 다른 방식들에 비해 정확도가 더 높고 Coefficient of variation 수치도 낮아서 본 파이프라인 고도화에 사용을 하게 되었다.

제 3 장 연구개발수행 내용 및 결과

제 3-1절: 환경유전체 metabolic potential 분석 시스템 개발

1. 자동화된 메타지놈 염기서열 어셈블리(assembly) 파이프라인 구축

가. 일반적으로 일루미나 기반 시퀀서는 적은 비용으로 많은 데이터의 가능하기 때문에 높은 coverage가 담보되야하는 whole genome shotgun sequencing을 통한 metagenome 연구에 많이 사용된다. 그러나 일루미나 기반 데이터는 전체 유전자 영역을 커버하지 못하는 short read 이기 때문에 정확하게 데이터를 분석하기 위해서는 어셈블리(assembly)과정이 필수적으로 동반되어야 한다.

나. 현재 short read 기반의 메타지놈 서열 어셈블리를 위해 Meta-IDBA (Peng et al. 2011), MetaVelvet (Namiki et al. 2012), Meta-velvetSL (Sato et al. 2015), IDBA-UD (Peng et al. 2012), Ray Meta (Boisvert et al. 2012) 등의 여러 가지 도구들이 개발되었다.

다. 지난 2015년 과제에서는 Meta-velvetSL을 사용하여 chimeric contig들을 줄여서 정확도를 올리려는 방식을 채용했었다. 하지만 Meta-velvetSL은 기존 reference genome sequence를 curation 및 database화하여 support vector machine의 training과정에 사용을 해야 하는데 human microbiome project와 같이 특정 environment에서 심도있게 연구된 분야에는 적합하지만 초기단계의 분야에서는 신뢰도 있는 database가 없거나 적은 경우에는 training 후의 model의 신뢰도가 떨어진다. 또한 알려진 strain의 genome정보를 최대한 사용을 한다고 해도 unknown strain이 압도적으로 많다고 여겨지기 때문에 여전히 bias가 있다고 볼 수 있다.

라. 또한 Meta-velvetSL은 multithread 나 parallel computing을 지원하지 않고 memory를 사용하는 양이 굉장히 크기 때문에 계산속도나 필요한 memory를 고려했을 때 통상적으로 생산되는 Illumina 기반의 input data를 분석이 가능한지에 대한 의문이 들었다.

마. 따라서 본 연구팀은 이제 까지 소개된 모든 어셈블리 도구들의 장단점을 분석하여, genome database와 같은 사전정보가 없어도 되며 multithread 나 parallel computing을 지원하고 다른 tool에 비해 assembly결과에서 16s rRNA를 잘 조립했던 Ray Meta를 기반으로 하는 어셈블리 파이프라

인을 구축하였다.

바. Meta-velvetSL은 여러 외부 도구가 필요하고 여러 가지 단계를 거쳐야하는 복잡한 구동 과정으로 인해 사용이 매우 어려운 단점이 있었으나 본 파이프라인은 Ray Meta를 기반으로 하였기 때문에 fastq input만 입력하면 되는 간단한 구동방식과 parallel computing을 지원한다.

사. 그리고 지난 과제에는 없었지만 보완한 부분은 binning을 위해 CONCOCT를 도입한 것과 taxonomy profiling을 위해 MetaPhlan2를 도입, Metabolic pathway 비교 분석을 위해 HUMAnN2와 MUSiCC을 도입한 것이다.

아. 많은 분석 단계들과 복잡한 input, output구조 때문에 본 연구팀은 자동화된 어셈블리 파이프라인을 구축하여 사용자가 편리하게 구동이 가능하도록 하였다. (그림 1).

자. 본 파이프라인은 쉘 스크립트 및 리눅스에서 사용하는 명령어, python script, R script를 조합하여 구현하였기 때문에 유지보수가 쉽고 즉각적인 추가기능의 탑재가 가능하다. 각종 설정 값들은 실행 script를 수정하여 설정 가능하다. (그림 2)

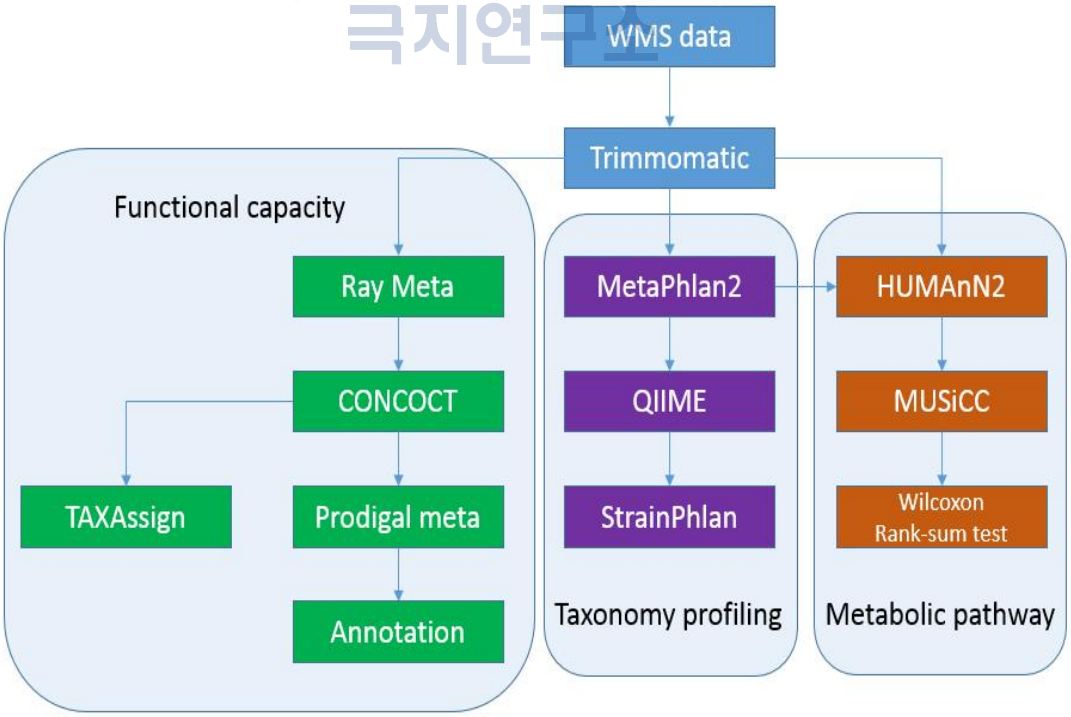

그림 1. 메타지놈 염기서열 어셈블리 파이프라인 워크플로우

```
[kdlsh@localhost ~]$ bash meta_pipe.sh

#!/bin/bash

dataPath=/home/kdlsh/Database
minlength=70
contig_min_len=500

#input1=/home/kdlsh/project/metagenome/test_data/SRR2198976_1.fastq.gz
input1=/home/kdlsh/project/metagenome/test_data/SRX343841_1.fastq.gz
input2=/home/kdlsh/project/metagenome/test_data/SRX343841_2.fastq.gz
```

그림 2. 염기서열 어셈블리 파이프라인 쉘 스크립트

2. Test data 및 계산시간

가. 실제 metagenome shotgun 데이터를 가지고 본 파이프라인의 구동 테스트를 수행하였다. dataset은 NCBI SRA로부터 다운 받은 Illumina HiSeq 2000 기반 paired-end data이며 두 sample간의 비교를 위해서 해양 sample과 토양 sample을 선정하였다. 토양은 SRX343841이며 10 Gb 용량의 압축된 fastq format이고, 해양은 ERR599025로 8Gb 용량의 압축된 fastq format 이다.

다. 실험 환경은 AMD Opteron(TM) Processor 6274 2.2 MHz 64 core CPU와 512GB 메모리를 탑재한 리눅스 서버에서 수행하였으며, 40 core를 사용하여 5 시간 만에 성공적으로 파이프라인 계산이 가능함을 보였다.

다. 같은 환경에서 40 Gb의 ERR1051325를 파이프라인 계산을 수행하였을 때, 40 core를 사용하여 56 시간이 걸렸다.

3. Input data 및 Quality control

가. 본 파이프라인은 gzip으로 압축된 Paired-end fastq file을 입력 값으로 받아들이며 추가적인 fasta 파일 형태로의 변환과정 없이 바로 실행이 가능하도록 구현하였다.

나. 대부분의 shotgun metagenome library는 throughput을 고려하여 Illumina 계열의 sequencer를 사용한다. Illumina 기반의 short read 서열 데이터에는 여러 가지 오류가 있거나 quality가 낮고, 길이가 짧은 서열이 섞여 있을 수 있다. 이러한 데이터는 추후 assembly 과정에서 잘못된 결과를 도출할 수 있기 때문에 제거되어야 한다. 본 파이프라인에서는 요즘 가장 널리 사용되며, 정확하다고 알려진 Trimmomatic (Bolger et al.

2014) 을 탑재하여 서열 전처리 과정을 통해 정확한 assembly가 가능하도록 하였다.

다. 실제 ERR599025에 대해 Trimmomatic 처리 전과 후를 fastqc tool을 이용해서 비교해 보았다. (그림 3)

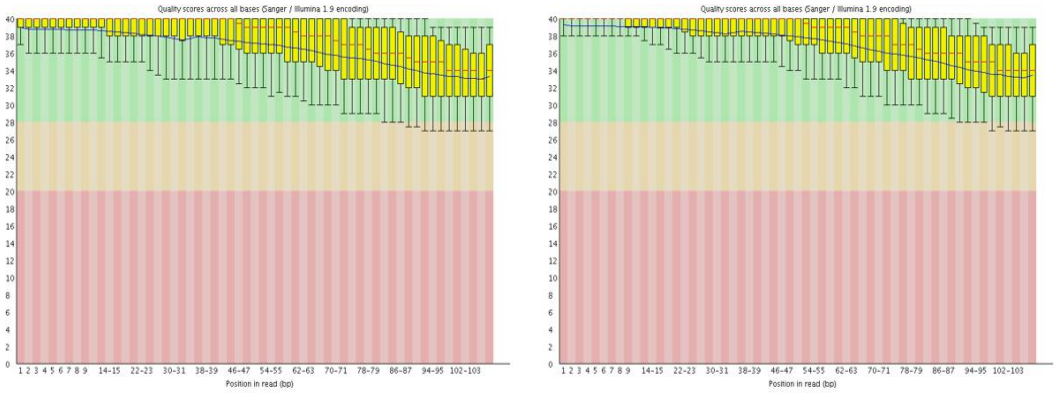

그림 3. Trimmomatic 계산 전후의 base quality.

4. Functional capacity part

가. 이 부분은 서열 어셈블리 과정을 통해 생성된 contig 서열을 가지고 유전자 예측 및 annotation을 위한 분석 파이프라인을 구축하였다.

나. Ray Meta로 assembly후에 CONCOCT로 binning, TAXAssign으로 operational taxonomic unit (OTU) taxonomy부여, Prodigal로 structural annotation, COG와 같은 functional annotation 으로 이루어져 있다.

다. Ray Meta는 de bruijn graph를 사용하여 k-mer정보를 바탕으로 assembly를 수행한다. Metavelvet-SL와 같이 training을 위한 사전 genome database가 필요 없으며 parallel computing을 지원하고 memory 사용량이 더 적다. 또한 16S rRNA를 더 정확하게 assembly하는 것으로 확인되어 사용하게 되었으며 input으로 gzip으로 압축된 fastq file을 입력하고 parallel computing을 위한 option, k-mer option만 주어주면 계산을 수행하게 된다.

라. Ray Meta로 계산을 수행하는 과정에서 최적의 k-mer option을 찾는 것이 중요한데 이는 k-mer option (-k) 에 21bp부터 41bp까지 2bp단위로 계산을 해서 N50 값과 assembly된 contig total size등의 수치를 보고

상황에 맞는 방식을 선택할 수 있다. 아래의 표는 해양 sample인 ERR599025를 Ray Meta로 계산한 결과이며 default 값으로 주어지는 31에서 N50이 184로 가장 최고치에 도달한 것을 알 수 있다. 실제 input size가 매우 클 경우에는 default값을 사용하거나 위아래로 5가지 정도 option으로 test해보면 좋을 것이다. (표 1)

k-mer	count	size	GC	min	max	mean	median	N50
23	1555107	245888691	45.0	100	23108	158	116	149
25	1752424	270151693	44.2	100	24529	154	113	144
27	1140348	204682095	44.0	107	37628	179	140	172
29	859603	170978045	44.1	115	33572	198	158	181
31	744798	154914706	43.9	123	37629	207	166	184
33	744764	154908355	43.9	123	37629	207	166	184
35	744803	154928957	43.9	123	37629	208	166	184
37	744820	154916282	43.9	123	37629	207	166	184
39	744816	154922833	43.9	123	37629	208	166	184

표 1. k-mer 수치에 따른 Ray Meta로 계산된 contig의 각종 통계 값.

마. CONCOCT는 unsupervised binning tool로써 nucleotide composition, kmer frequency, coverage data를 사용한다. CONCOCT species level 까지 binning을 하며 binning을 수행하기 전에 여러 가지 준비 사항들이 있다. 우선 길이가 너무 긴 contig들은 하나의 bin으로 지정되어 정보손실이 되는 것을 줄이기 위해 10 - 20 kb의 크기로 잘라주게 되며 그 뒤에 input read를 bowtie2를 이용하여 mapping하고 PICARD의 MarkDuplicates를 사용해 duplicated read를 제거하여 coverage data를 생성한다. 생성된 bam file과 contig를 이용해 input table을 만들고 concoct를 실행하여 binning을 수행한다. (그림 4) 이 부분에서 중요한 option은 contig minimum length인데 binning을 수행할 최소 길이가 default로 1kb로 지정되어 있다. 하지만 composition 기반의 binning 방법은 contig의 길이가 길수록 더 정확한 결과를 보여주기 때문에 사용자의 목적에 따라 길이를 더 높여주는 것도 false positive를 줄일

수 있는 방법이다. binning을 수행 한 후에 TAXAssign이라는 외부 프로그램을 이용해 taxonomy를 부여 할 수 있으며 single copy gene 정보를 이용해 binning이 잘 되었는지 평가를 해볼 수도 있다.

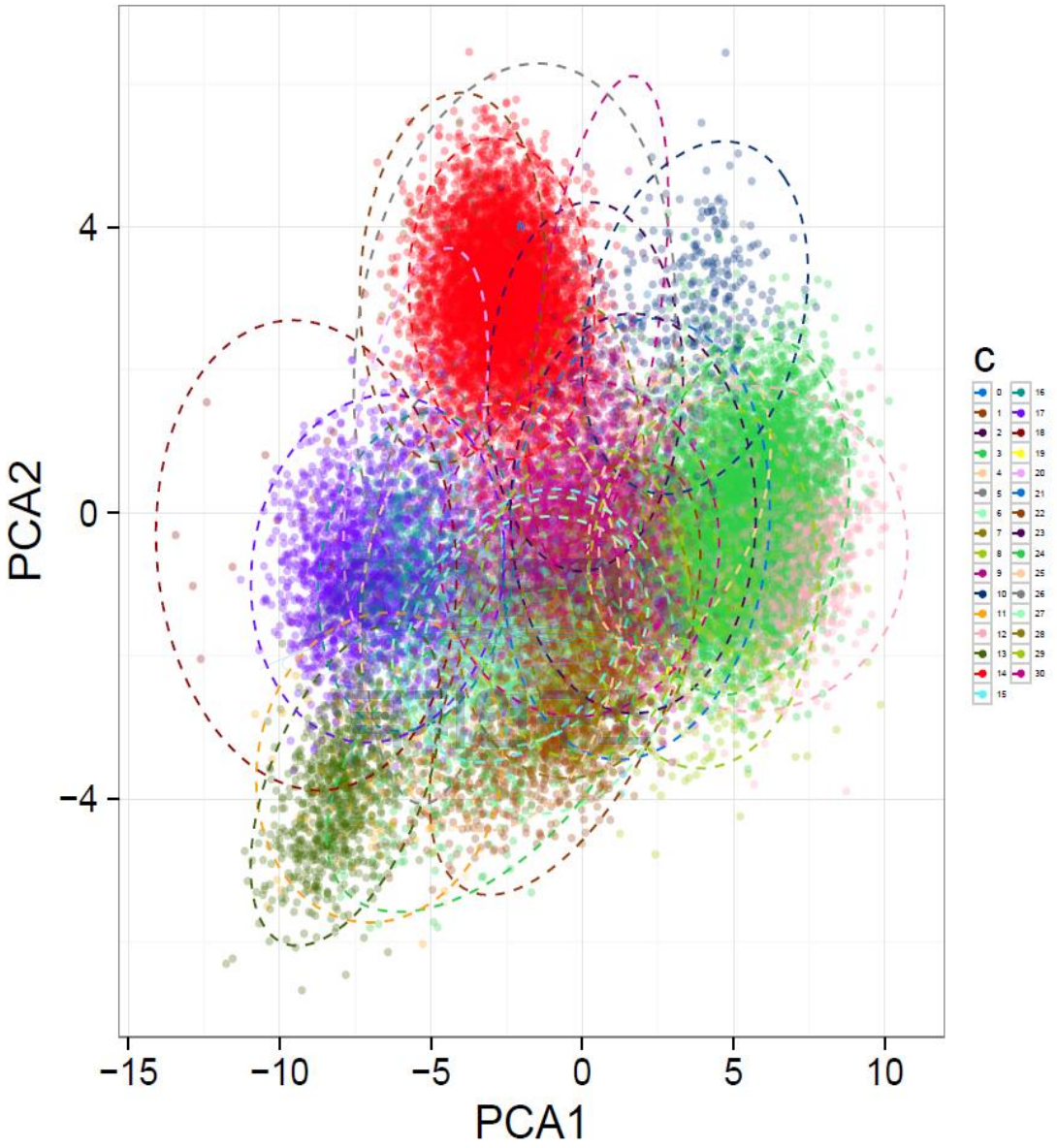

그림 4. CONCOCT의 binning 결과 PCA plot (ERR599025).

바. concoct로 binning을 수행 한 후에 prodigal meta로 structural annotation을 수행하게 되며 COG database를 이용해 RPSBLAST로 amino acid의 function을 mapping 한다. (그림 5)

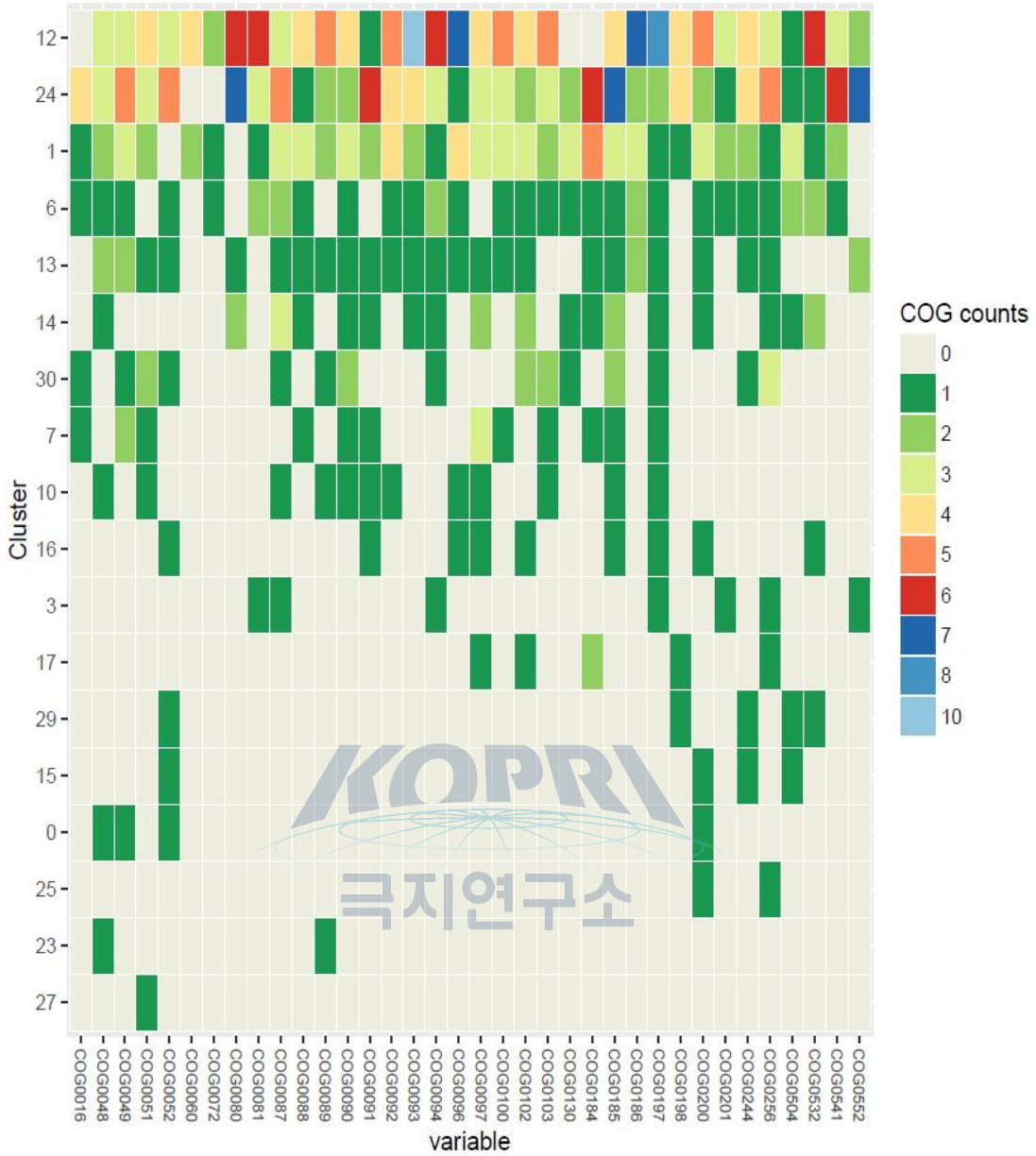

그림 5. CONCOCT에 의해 계산된 cluster(bin) 의 COG intensity plot (ERR599025).

5. 메타지놈 내 특이 유전자 클러스터링 분석 및 빈도 비교 분석 파이프라인 구축

가. 이 부분은 전처리 과정을 거친 fastq data를 바로 marker sequence database에 mapping하여 taxonomy profile을 수행하며 그 이후에 QIIME의 function script들을 이용해 결과 그림이나 alpha diversity와 같은 계산이 가능하다. 또한 MetaPhlan2 내부 tool인 strainphlan을 이용하여 strain단계의 profiling을 할 수 있다.

나. MetaPhlan2는 metagenome shotgun sequencing data에서 Bacteria, Archaea, Eukaryotes, Viruses 와 같은 미생물 community의 composition profiling을 위한 tool이다. 2.0 version부터 특정 species을 선정하여 strain level까지 비교할 수 있게 되었고 약 17,000여개의 reference genome (~13,500 bacterial과 archaeal, ~3,500 viral, ~110 eukaryotic) 에서부터 백만 개에 달하는 clade-specific marker gene을 구축하였다. 이는 불명확했던 taxonomy 선정에 도움을 주고 relative abundance의 정확도를 올리며 bacteria를 넘어서 archaea, eukaryotes, viruses에서도 species-level의 taxonomy를 알 수가 있게 해주었다. 또한 strain을 variation을 통해 구분하고 1버전에 비해 계산 속도가 상승하였으며 strain-level population genomics가 가능하게 되었다.

다. MetaPhlan2는 전처리된 fastq file을 input으로 받는데 forward와 reverse 구분 없이 모든 read를 병합해서 사용한다. Multithread option을 지원하며 bowtie2를 내부에서 사용해 clade-specific marker gene database에 mapping 한 후 bam file과 결과 profile text file을 output으로 얻게 된다. (그림 6) hierarchical 방식으로 결과를 얻게 되며 kingdom level에서부터 species level까지 순차적으로 나열되고 marker gene에 mapping된 read를 normalization한 abundance값을 얻을 수가 있다. 밝혀진 taxonomy의 종류와 각 항목에 해당하는 abundance 결과를 사용해 자체 script로 형식변환 및 hclust heatmap plot을 얻을 수 있다. (그림 7) 그림 7을 보면 해양과 토양에서 얻은 metagenome sequencing data에서 명확하게 다른 종류의 species들이 나타나는 것을 확인 할 수 있었다. 사용한 data가 10 Gb 가 안 되는 적은 양이여서 두 가지 환경에서 동시에 나타나는 species는 볼 수 없었다.

```
ID  profiled_metagenome profiled_metagenome
#SampleID   Metaphlan2_Analysis Metaphlan2_Analysis
k__Archaea  0.07655 0.0
k__Archaea|p__Thaumarchaeota    0.07655 0.0
k__Archaea|p__Thaumarchaeota|c__Thaumarchaeota_noname   0.07655 0.0
k__Archaea|p__Thaumarchaeota|c__Thaumarchaeota_noname|o__Nitrosopumilales   0.06568 0.0
k__Archaea|p__Thaumarchaeota|c__Thaumarchaeota_noname|o__Nitrosopumilales|f__Nitrosopumilaceae  0.06568 0.0
k__Archaea|p__Thaumarchaeota|c__Thaumarchaeota_noname|o__Nitrosopumilales|f__Nitrosopumilaceae|g__Nitrosopumilaceae_unclassified    0.06568 0.0
k__Archaea|p__Thaumarchaeota|c__Thaumarchaeota_noname|o__Thaumarchaeota_noname  0.01087 0.0
k__Archaea|p__Thaumarchaeota|c__Thaumarchaeota_noname|o__Thaumarchaeota_noname|f__Thaumarchaeota_noname 0.01087 0.0
k__Archaea|p__Thaumarchaeota|c__Thaumarchaeota_noname|o__Thaumarchaeota_noname|f__Thaumarchaeota_noname|g__Thaumarchaeota_noname    0.01087 0.0
k__Archaea|p__Thaumarchaeota|c__Thaumarchaeota_noname|o__Thaumarchaeota_noname|f__Thaumarchaeota_noname|g__Thaumarchaeota_noname|s__Thaumarchaeota_noname_unclassified  0.01087 0.0
k__Bacteria 71.16075    100.0
k__Bacteria|p__Acidobacteria    0.0 47.89094
k__Bacteria|p__Acidobacteria|c__Acidobacteriia  0.0 47.89094
k__Bacteria|p__Acidobacteria|c__Acidobacteriia|o__Acidobacteriales  0.0 47.89094
k__Bacteria|p__Acidobacteria|c__Acidobacteriia|o__Acidobacteriales|f__Acidobacteriaceae 0.0 47.89094
k__Bacteria|p__Acidobacteria|c__Acidobacteriia|o__Acidobacteriales|f__Acidobacteriaceae|g__Acidobacteriaceae_unclassified   0.0 43.11224
k__Bacteria|p__Acidobacteria|c__Acidobacteriia|o__Acidobacteriales|f__Acidobacteriaceae|g__Granulicella 0.0 4.7787
k__Bacteria|p__Acidobacteria|c__Acidobacteriia|o__Acidobacteriales|f__Acidobacteriaceae|g__Granulicella|s__Granulicella_unclassified    0.0 4.7787
k__Bacteria|p__Actinobacteria   0.0 10.57841
k__Bacteria|p__Actinobacteria|c__Actinobacteria 0.0 10.57841
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales  0.0 8.32917
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Brevibacteriaceae 0.0 2.04408
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Brevibacteriaceae|g__Brevibacterium   0.0 2.04408
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Brevibacteriaceae|g__Brevibacterium|s__Brevibacterium_unclassified    0.0 2.04408
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Dermatophilaceae  0.0 1.87391
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Dermatophilaceae|g__Dermatophilaceae_unclassified 0.0 1.87391
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Microbacteriaceae 0.0 0.86218
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Microbacteriaceae|g__Leifsonia    0.0 0.86218
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Microbacteriaceae|g__Leifsonia|s__Leifsonia_unclassified  0.0 0.86218
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Micromonosporaceae    0.0 0.75999
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Micromonosporaceae|g__Salinispora 0.0 0.75999
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Micromonosporaceae|g__Salinispora|s__Salinispora_unclassified 0.0 0.75999
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Propionibacteriaceae  0.0 1.54475
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Propionibacteriaceae|g__Propionibacteriaceae_unclassified 0.0 1.54475
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Pseudonocardiaceae    0.0 0.65972
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Pseudonocardiaceae|g__Pseudonocardia  0.0 0.65972
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Pseudonocardiaceae|g__Pseudonocardia|s__Pseudonocardia_unclassified   0.0 0.65972
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Streptomycetaceae 0.0 0.30709
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Streptomycetaceae|g__Streptomyces 0.0 0.30709
```

그림 6. ERR599025, SRX343841 2 sample 의 taxonomic profiling 결과.

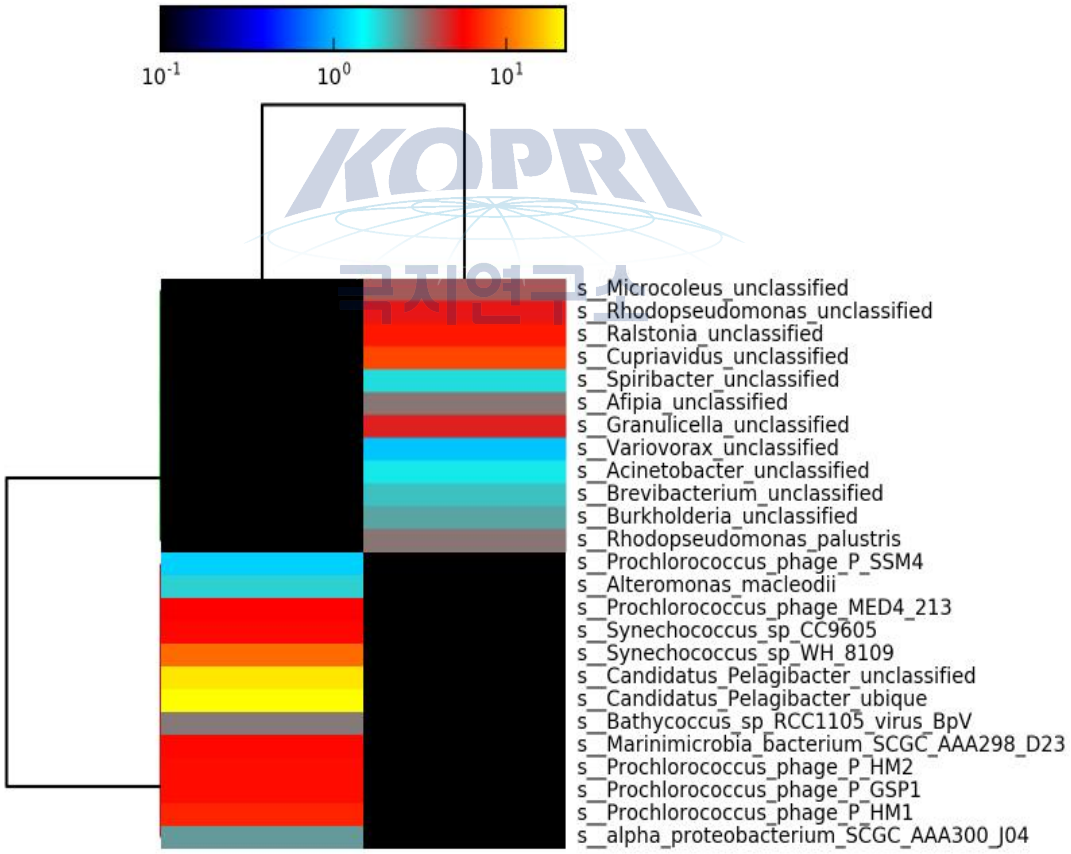

그림 7. ERR599025, SRX343841 2 sample 의 MetaPhlAn2 결과의 hclust heatmap plot.

라. MetaPhlAn2 의 taxonomic profiling결과를 biom format으로 얻을 수

도 있는데 이를 이용해서 qiime의 script의 계산이 가능하다. 예를 들어 plot_taxa_summary.py와 같은 script나 alpha_diversity.py와 같은 function으로 다양성 분석에 사용되는 많은 계산을 수행 가능하다.

마. StrainPhlAn은 여러 개의 sample에서 각각의 strain을 추적하고 분석하기 위한 MetaPhlan2의 내부 tool이다. StrainPhlAn의 input은 metanogme sample들이며 각 species에 대해 MUSCLE 과 같은 tool로 multiple sequence alignment (MSA)을 수행한다. 이 MSA으로부터 RAxML을 사용해 strain evolution을 나타내는 phylogenetic tree를 구성하게 된다.

6. Metabolic pathway 관련 유전자 빈도 비교 분석 파이프라인 구축

가. 이 부분은 KEGG와 같은 metabolic pathway에 연결이 가능한 database를 이용해 해당 sample에서 어떤 항목이 얼마나 있는지 알아낸 후에 정확한 비교를 위한 normalization과 통계 test를 통한 유의성 비교를 포함하고 있다.

나. HUMAnN2로 microbial pathway의 종류와 relative abundance를 계산한 후에 MUSiCC 으로 normalization을 하고 두 그룹의 sample을 Wilcoxon Rank-sum test를 이용해 pathway의 abundance 차이가 유의한지 검사한다.

다. HUMAnN2는 metagenome 이나 metatranscriptome sequencing data에서 sample community내에 microbial pathway의 존재유무와 abundance를 알아내기 위한 tool이다. Functional profiling이라 할 수 있는 이 작업은 sample을 채취한 환경에서의 microbial community의 metabolic potential을 알아내는 것이다. 한두 개의 유전자로 metabolite가 생산될지 안 될지의 문제나 환경의 특성을 확신할 수 없기 때문에 pathway단위의 비교는 보다 의미가 있다.

라. HUMAnN2의 특징은 known과 unclassified 모두에서 community functional profile을 얻을 수 있고 MetaPhlAn2와 ChocoPhlAn pangenome database를 사용해 빠르고 정확하게 organims-specific한 functional profile을 할 수 있다. 또한 여러 가지 database를 활용하는데 UniRef database는 gene family, MetaCyc는 pathway, MinPath는 minimum pathway set 의 정보를 제공한다. Input 으로는 전처리가 되고 interleaved형태의 압축되지 않은 fastq file을 입력하면 되며 nucleotide-level의 탐색에서는 Bowtie2, translated 탐색에서는

Diamond를 사용해 mapping 시간을 단축시켰다. 하지만 KEGG database를 사용하기 위해서는 HUMAnN 버전 1.0 의 database를 형식 변환 및 database 재구축 단계를 거쳐야 한다.

마. HUMAnN2에 interleaved fastq input과 KEGG database, pathway database, protein database를 입력하면 계산을 수행한다. (표 2) 계산 결과는 표2 와 같으며 Gene family에 해당하는 KEGG 번호의 RPK(read per kilobase)값과 relative abundace값, kegg pathway에 해당하는 abundance와 coverage값을 얻을 수 있다.

# Pathway	Abundance	Coverage	# Gene Family	RPKs	Relative abundance
ko00010	450.606	0.225	K00001	11.622	1.69E-05
ko00020	413.225	0.208	K00003	389.111	0.000567069
ko00030	599.231	0.273	K00005	300.880	0.000438485
ko00051	250.851	0.106	K00012	327.026	0.000476589
ko00053	54.498	0.029	K00013	388.036	0.000565501
ko00061	633.377	0.269	K00014	197.226	0.000287426
ko00071	76.849	0.043	K00020	449.013	0.000654367
ko00100	1.826	0.000	K00024	45.543	6.64E-05
ko00130	191.340	0.077	K00029	87.521	0.000127547
ko00190	352.232	0.141	K00030	251.722	0.000366846
ko00195	1567.050	0.508	K00031	880.913	0.00128379
ko00196	479.643	0.341	K00033	1049.123	0.00152893
ko00230	312.628	0.135	K00034	18.764	2.73E-05
ko00240	430.131	0.140	K00036	896.184	0.00130605
ko00260	257.815	0.209	K00052	490.081	0.000714215
ko00270	435.117	0.197	K00053	1593.030	0.00232159
ko00280	140.686	0.070	K00057	14.212	2.07E-05
ko00281	56.680	0.000	K00058	623.789	0.000909074
ko00290	1085.198	0.682	K00059	364.841	0.000531698
ko00300	249.720	0.125	K00067	259.458	0.000378119
ko00330	215.166	0.104	K00075	432.719	0.000630619
ko00340	281.534	0.158	K00082	239.481	0.000349005
ko00362	14.841	0.000	K00088	524.205	0.00076394

ko00364	11.505	0.000	K00097	333.680	0.000486287
ko00400	256.174	0.145	K00099	611.120	0.000890611
ko00430	73.642	0.000	K00100	307.695	0.000448418
ko00450	678.349	0.345	K00101	9.628	1.40E-05

표 2. HUMAnN2 metabolic pathway abundance 계산 결과의 일부 (ERR599025).

바. HUMAnN2를 사용해 function의 relative abundance값을 얻을 수 있지만 이런 방식의 compositional normalization은 sample 내의 gene들의 copy수가 모두 다르고 gene마다 특성이 다르다는 점에서 매우 부정확하다. 때문에 MUSiCC을 사용하였는데 이 tool은 univeral single copy gene을 이용한 intra-sample gene normalization과 gene-specific model을 이용한 inter-sample gene correction을 수행한다.

사. MUSiCC은 input으로 abundance count file을 받으며 generic model을 그냥 사용하거나 (-c use_generic) data로부터 linear model을 learning시킬 수도 있다 (-c learn_model). (표 3)

KO	relative-abun	non-model	generic	learn-model
K00116	0.00057	0.49771	0.51319	0.51434
K00121	0.00026	0.22462	0.22168	0.16232
K00128	0.00093	0.81284	0.84916	0.73444
K00133	0.00066	0.57779	0.57614	0.58337
K00134	0.00075	0.65773	0.63019	0.52894
K00145	0.00059	0.52066	0.54316	0.61027
K00147	0.00072	0.63220	0.64816	0.60448
K00161	0.00234	2.05874	2.15073	2.26542
K00162	0.00222	1.95026	2.02896	2.04143
K00218	0.00141	1.23669	1.23127	0.99436
K00226	0.00055	0.47934	0.48714	0.52627
K00228	0.00127	1.11221	1.12698	0.95429
K00239	0.00108	0.95298	0.96128	0.76990
K00241	0.00016	0.14416	0.14341	0.13013
K00259	0.00057	0.49871	0.51570	0.51440
K00274	0.00021	0.18059	0.18866	0.21469

표 3. MUSiCC normalization 계산 결과의 일부 (ERR599025).

사. MUSiCC을 통해 normalization과 correction이 이루어진 abundance값을 얻게 되면 이제 어떤 pathway가 유의미하게 차이가 있는지를 비교해 봐야 하는데 wilcoxon rank-sum test를 이용해서 두 pathway의 abundance값 distribution을 비교한다.

아. 실제 data를 가지고 비교를 위해서 ERR599025, SRX343841 두 sample에서 Valine, leucine and isoleucine degradation pathway인 ko00280를 선정해서 분석을 하였다. 각각의 sample에서 나타나는 gene과 abundance 차이를 보였다. (그림 8)

자. ko00280 pathway 상의 gene 들의 MUSiCC abundance 값을 input으로 하여 R의 wilcox.test function을 사용하여 wilcoxon rank sum test를 수행하였다. (그림 9) wilcoxon rank sum test는 정규분포를 띄지 않는 nonparametric data에서 독립적인 두 sample을 비교하는 test로 두 모집단의 중심차이가 있는지에 대한 검사이다. Test결과 p-value 가 0.0001761로 나와 두 sample간의 ko00280 pathway의 abundance가 유의하게 차이가 있다는 것으로 결론 낼 수 있다.

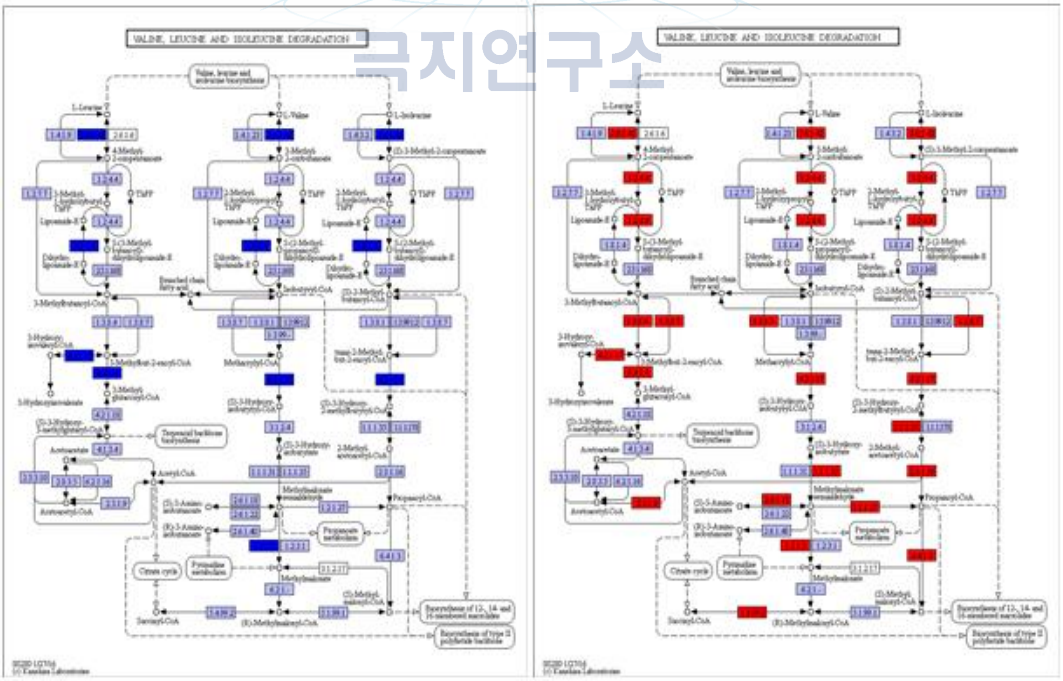

그림 8. ERR599025, SRX343841 2 sample 의 ko00280 pathway 비교.

```
> wilcox.test(x,y)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 30, p-value = 0.0001761
alternative hypothesis: true location shift is not equal to 0
```

그림 9. R의 wilcox.test function을 이용한 wilcoxon rank sum test 결과.

제 4 장 연구개발목표 달성도 및 대외기여도

제 4-1절: 2016 3차년도

연구목표	연구내용	달성도	대외기여도
환경유전체의 metabolic potential 분석 시스템 개발	메타지놈 내 특이 유전자 클러스터링 분석 및 빈도 비교 분석 파이프라인 구축	100%	- RNA-Seq 데이터에서 사용하는 FPKM (fragment per kilobase of transcript per million fragments)의 개념을 차용하여, *recA*, *gyrB*, *rpoB* 및 EF-Tu 등 유전체 내 1 copy만을 지니는 유전자들과 관심있는 특이 유전자들의 reference 유전자 염기서열의 단위 길이당 match되는 read 수를 파악하여 해당 시료 내 특이 유전자를 지니는 미생물의 빈도수를 계산함 - 임의로 단편화되어 align이 되지 않는 메타지놈 reads를 clustering하기 위하여 해당되는 각 read의 염기서열에 표준화된 정보값을 부여하고 이를 character value로 인지하여 clustering을 수행할 수 있는 체계를 구현함. 이로부터 시료 내 유전자 그룹별 빈도를 파악할 수 있는 자동화 분석 시스템 개발함 - 메타지놈 서열로부터 유전자의 빈도수 계산을 위한 프로그램을 개발하고, 일련의 과정들을 서로 연동하여 standalone 수준에서 분석 파이프라인을 구축함
	metabolic pathway 관련 유전자 빈도 비교 분석 파이프라인 구축	100%	- 확장형 KEGG database의 염기서열과 메타지놈 염기서열 함께 CLUSTOM 알고리즘을 이용하여 clustering한 후, 각 cluster에서 KEGG ID를 추출하고 *rpoB*, *recA* 유전자에 match된 수 및 각 reference 유전자 그룹의 size로 정

			규화하여 메타지놈 내 각 metabolism에 관여하는 유전자들의 빈도수 계산 방법을 구축함 - 구축된 방법을 이용하여 KEGG ID에 따른 빈도수를 자동으로 계산하는 프로그램을 구현하고 table 형식으로 도출할 수 있도록 설계함 - Hypergeometric distribution, t-test 등의 통계적 방법을 이용하여 샘플별 enriched된 세부 metabolic pathway 정보 표현 - KEGG ID 및 EC number를 선택하여 해당하는 metagenome reads를 다운로드할 수 있도록 설계함

제 5 장 연구개발결과의 활용계획

제 5-1절: 추가연구의 필요성

- 세균에 비해 균류의 경우 표준염기서열 DB가 없어 다양성 연구에 어려움이 많음. 따라서 본 과제 수행을 통해 확충 및 업데이트된 LSU 표준염기서열 DB은 균류 다양성 연구를 위한 국제적 표준으로 자리 잡을 것으로 기대함
- 메타지놈 염기서열 분석은 데이터의 방대함으로 인해 고성능의 컴퓨터 자원이 없는 경우 일반 연구자가 접근하기 힘듬. CLOUD 기반의 메타지놈 분석 프로그램 개발은 시스템 메모리 부족을 해결할 수 있는 좋은 학문적 모델이 될 것으로 기대함
- 잘 정립되어 있는 미생물다양성 염기서열 분석 방법에 비해 shotgun 메타지놈 분석은 아직까지 분석 protocol에 대한 consensus가 없는 실정임. 본 과제 수행을 통해 속도 및 정확도를 고려하여 구축될 메타지놈 분석 파이프라인이 일반 생물학자들에게 데이터 분석 표준을 제시할 것으로 기대함
- 물질대사 측면에서 서로다른 환경유전체 샘플을 비교하는 생물정보학적 도구가 부족한게 현실임. 본 과제를 통해 구축될 'web system for evaluating metabolic potential'은 계층구조적 물질대사경로 별 서로 다른 환경 샘플의 metabolic potential를 비교하고, 특정 환경에 specific 또는 enrich된 유전자를 결정해 줄것으로 기대함

제 5-2절: 기업화 추진방안

- 최근 NGS 기술의 발전으로 유전체학 및 생태학 분야에서 대량의 데이터가 생산되고 있음. 데이터 크기로 인해 단일 연구실에서 분석이 불가능함. 따라서, 본 과제에서 구축 및 개발된 프로그램 및 pipelines을 발전시켜 상용 분석 서비스 제공이 가능함
- 균류 rDNA 레퍼런스 DB가 구축은 정확한 균류 동정 및 환경 내 균류 다양성 정보를 제공하게 되어, 식물병원균류/환경부후균류/생리활성물질생성균류 연구에 기반이 될 것임. 또한, 향후 농업, 환경, 의료, 보건 등의 산업분야에서의 높은 활용도를 가질 것으로 예상됨
- 구축될 메타지놈 분석 프로그램과 물질대사 평가 시스템은 급변하는 전 지구적 기후변화 패턴을 미생물 수준에서 monitoring 하기 위한 기본 도구로 이용될 수 있음

제 6 장 연구개발과정에서 수집한 해외과학기술정보

제 6-1절: Microbiome

○ 2007년부터 2015년까지 진행되는 미국 NSF Human microbiome project는 인간의 다양한 조직 (장내, 피부, 구강 등)에 공생하는 미생물 군집을 조사하고, 인간 질병과 미생물간의 인과관계 및 상호작용을 이해하고자 함

○ 이에 앞서 유럽과 중국을 중심으로 구성된 MetaHit consortium은 Danish와 Chinese를 대상으로 인간공생 장내미생물 군집 및 메타지놈 연구를 수행하고, 장내 미생물 및 미생물 유전자 카탈로그를 제작함

○ Microbiome데이터 분석 기술 쪽에서는 University of Michigan의 P. Schloss 교수와 University of Colorado의 R. Knight 교수 등에 의해 다양한 분석 프로그램들이 개발됨 (Mothur, Qiime, Unifrac)

○ Microbiome 데이터 분석의 핵심단계라 할 수 있는 염기서열 클러스터링*을 위해 University of Florida의 Sun 교수 그룹에서 ESPRIT-Tree 프로그램을 개발함(* 염기서열 클러스터링은 환경샘플 내 operational taxonomic units의 개수 및 크기를 결정. 이 두 parameters만을 가지고 모든 alpha-diveristy 지수 계산이 가능해짐)

○ Microbiome 데이터 서열 전처리부터 다양성 지수계산까지의 일련의 분석을 한번에 수행할 수 있도록, University of Florida의 E. Triplett 교수는 Pangea라는 분석 파이프라인을 개발함

제 6-2절: Shotgun metagenome

○ Microbiome에 비해 shotgun metagenome 분석 기술은 다루어야 할 염기서열 데이터 양이 수천배~수만배 큼. 이와 같은 대용량 데이터에 대한 메모리 관리가 쉽지 않아 shotgun metagenome분석용 프로그램 개발이 더딘 상태임

○ 이미 NGS기반 메타지놈 데이터 축적 속도는 컴퓨터 resource 확장 속도를 추월한 상태임. 따라서, 클라우드 등 신개념의 데이터 처리 기법을 이용하여 데이터 분석시 발생하는 메모리 부족 문제를 해결해야 함

○ 대부분의 경우 개별 생물종 유전체 분석에 사용되는 프로그램들을 차용하고 있는 수준임. shotgun metagenome에 특화된 프로그램으로는 일본 연구진에 의해서 개발된 MetaVelvet이 있음

○ Microbiome 데이터 분석시 환경샘플간 community structure를 비교하거나, shotgun metagenome 데이터 분석시 gene clustering을 수행할 경우 계통도 작성이 필수불가결함. 하지만, 데이터 양이 커서 전통적인 phylogenetic tools를 이용하는데에는 한계가 있음

○ 대용량 염기서열 데이터에 대한 계통도를 구성하기 위해서 University of

California, Berkeley 연구진들에 의해 FastTree 프로그램이 개발되어, 수백만 염기서열을 다룰 수 있게 됨

제 7 장 참고문헌

1 Schloss, P. D.*et al.*Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*75, 7537-7541 (2009).

2 Kuczynski, J.*et al.*Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Current protocols in microbiology*, 1E. 5.1-1E. 5.20 (2012).

3 Truong, D. T.*et al.*MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods*12, 902-903 (2015).

4 Mitchell, A.*et al.*EBI metagenomics in 2016-an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, gkv1195 (2015).

5 Wilke, A.*et al.*The MG-RAST metagenomics database and portal in 2015. *Nucleic acids research*44, D590-D594 (2016).

6 Markowitz, V. M.*et al.*IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*42, D568-D573 (2014).

7 Huson, D. H., Buchfink, B. & Ruscheweyh, H.-J. An Efficient Pipeline for Microbiome Analysis. (2015).

8 Lee, J.-H., Yi, H. & Chun, J. rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *The Journal of Microbiology*49, 689-691 (2011).

9 DeSantis, T. Z.*et al.*Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*72, 5069-5072 (2006).

10 Jones, P.*et al.*InterProScan 5: genome-scale protein function classification. *Bioinformatics*30, 1236-1240 (2014).

11 Mitchell, A.*et al.*The InterPro protein families database: the classification resource after 15 years. *Nucleic acids research*, gku1243 (2014).

12 Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic acids research*, gku1223 (2014).

13 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, gkv1070 (2015).

14 Consortium, G. O. Gene ontology consortium: going forward. *Nucleic acids research***43**, D1049-D1056 (2015).

15 Overbeek, R.*et al.*The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research***42**, D206-D214 (2014).

16 Finn, R. D.*et al.*The Pfam protein families database: towards a more sustainable future. *Nucleic acids research***44**, D279-D285 (2016).

17 Meyer, F., Overbeek, R. & Rodriguez, A. FIGfams: yet another set of protein families. *Nucleic acids research***37**, 6643-6654 (2009).

18 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods***12**, 59-60 (2015).

19 Chevreux, B. MIRA: an automated genome and EST assembler. (2007).

20 Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S. & Pop, M. Next generation sequence assembly with AMOS. *Current Protocols in Bioinformatics*, 11.18. 11-11.18. 18 (2011).

21 Treangen, T. J.*et al.*MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome biology***14**, 1 (2013).

22 Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome research***18**, 324-330 (2008).

23 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research***18**, 821-829 (2008).

24 Liu, C.-M.*et al.*SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics***28**, 878-879 (2012).

25 Simpson, J. T.*et al.*ABySS: a parallel assembler for short read sequence data. *Genome research***19**, 1117-1123 (2009).

26 Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics***27**, i94-i101 (2011).

27 Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research***40**, e155-e155 (2012).

28 Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics***28**, 1420-1428 (2012).

29 Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glöckner, F. O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC bioinformatics***5**, 1 (2004).

30 Chan, C.-K. K., Hsu, A. L., Halgamuge, S. K. & Tang, S.-L. Binning

sequences using very sparse labels within a metagenome. *BMC bioinformatics***9**, 1 (2008).

31 McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods***4**, 63-72 (2007).

32 Deng, D. & Kasabov, N. in *Proc. of IJCNN.* 3-8.

33 Pati, A., Heath, L. S., Kyrpides, N. C. & Ivanova, N. ClaMS: a classifier for metagenomic sequences. *Standards in genomic sciences***5**, 248 (2011).

34 Alneberg, J.*et al.*CONCOCT: clustering contigs on coverage and composition. *arXiv preprint arXiv:1312.4038*(2013).

35 Krause, L.*et al.*Phylogenetic classification of short environmental DNA fragments. *Nucleic acids research***36**, 2230-2239 (2008).

36 Liu, B., Gibbons, T., Ghodsi, M. & Pop, M. in *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on.* 95-100 (IEEE).

37 Haque, M. M., Ghosh, T. S., Komanduri, D. & Mande, S. S. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics***25**, 1722-1730 (2009).

38 Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic acids research***38**, e132-e132 (2010).

39 Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic acids research***34**, 5623-5630 (2006).

40 Hyatt, D.*et al.*Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics***11**, 119-119, doi:10.1186/1471-2105-11-119 (2010).

41 Hoff, K. J., Lingner, T., Meinicke, P. & Tech, M. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic acids research***37**, W101-W105 (2009).

42 Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research***38**, e191-e191 (2010).

43 Bland, C.*et al.*CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC bioinformatics***8**, 1 (2007).

44 Edgar, R. C. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC bioinformatics***8**, 1 (2007).

45 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids*

*research***25**, 955-964 (1997).

46 Quast, C.*et al.*The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research***41**, D590-D596 (2013).

47 Cole, J. R.*et al.*The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research***37**, D141-D145 (2009).

48 Huerta-Cepas, J.*et al.*eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic acids research*, gkv1248 (2015).

49 Haft, D. H.*et al.*TIGRFAMs and genome properties in 2013. *Nucleic acids research***41**, D387-D395 (2013).

50 Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics***16**, 1 (2015).

51 Davenport, C. F.*et al.*Genometa-a fast and accurate classifier for short metagenomic shotgun reads. *PloS one***7**, e41224 (2012).

52 Freitas, T. A. K., Li, P.-E., Scholz, M. B. & Chain, P. S. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic acids research*, gkv180 (2015).

53 Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology***15**, 1 (2014).

54 Ames, S. K.*et al.*Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics***29**, 2253-2260 (2013).

55 Dröge, J., Gregor, I. & McHardy, A. C. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics*, btu745 (2014).

56 Angly, F. E.*et al.*The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol***5**, e1000593 (2009).

57 Raes, J., Korbel, J. O., Lercher, M. J., Von Mering, C. & Bork, P. Prediction of effective genome size in metagenomic samples. *Genome biology***8**, 1 (2007).

58 Beszteri, B., Temperton, B., Frickenhaus, S. & Giovannoni, S. J. Average genome size: a potential source of bias in comparative metagenomics. *The ISME journal***4**, 1075-1077 (2010).

59 Manor, O. & Borenstein, E. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances

in the microbiome. *Genome biology***16**, 1 (2015).

주 의

1. 이 보고서는 극지연구소 위탁과제 연구결과보고서 입니다.
2. 이 보고서 내용을 발표할 때에는 반드시 극지연구소에서 위탁연구과제로 수행한 연구결과임을 밝혀야 합니다.
3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 안됩니다.