BSPE17300-003-13

극지 메타지놈 염기서열 분석 파이프라인 구축

Developing a pipeline for analyzing polar metagenome sequences

2018. 03. 30

한 국 해 양 연 구 원
부 설 극 지 연 구 소

제 출 문

극지연구소장 귀하

본 보고서를 “ 극지 메타지놈 염기서열 분석 파이프라인 구축 에 관한 연구”과제의 (연차,최종)보고서로 제출합니다.

2018. 3 .30

연 구 책 임 자 : 김경모

참 여 연 구 원 : 황규인

보고서 초록

<table>
<tr><td>과제관리번호</td><td>PE17300</td><td>해당단계
연구기간</td><td>2017.4 - 2018.3</td><td>단계 구분</td><td>(1) / (1)</td></tr>
<tr><td rowspan="2"><u>연구사업명</u></td><td>중 사 업 명</td><td colspan="4">극지연구소 주요사업</td></tr>
<tr><td>세부사업명</td><td colspan="4">신진연구원 지원과제</td></tr>
<tr><td rowspan="2">연구과제명</td><td>중 과 제 명</td><td colspan="4"></td></tr>
<tr><td>세부(단위)과제명</td><td colspan="4">극지 메타지놈 염기서열 분석 파이프라인 구축</td></tr>
<tr><td>연구책임자</td><td>김경모</td><td>해당단계
참여연구원수</td><td>총 : 2 명
내부 : 1 명
외부 : 1 명</td><td>해당단계
연구비</td><td>정부: 30,000 천원
기업: 천원
계: 30,000 천원</td></tr>
<tr><td>연구기관명 및
소속부서명</td><td colspan="2">극지연구소 생명과학부</td><td colspan="2">참여기업명</td><td></td></tr>
<tr><td>국제공동연구</td><td colspan="2">상대국명 :</td><td colspan="3">상대국연구기관명 :</td></tr>
<tr><td>위 탁 연 구</td><td colspan="2">연구기관명 :</td><td colspan="3">연구책임자 :</td></tr>
<tr><td colspan="4">요약(연구결과를 중심으로 개조식 500자이내)</td><td>보고서
면수</td><td>42</td></tr>
<tr><td colspan="6">

- 대용량 메타지놈 염기서열 데이터 분석 시스템 최적화 수행
- 마이크로비옴 분석 파이프라인 구축
- 유전체 기반 메타지놈 염기서열 분석 파이프라인 구축
- 염기서열 시뮬레이션을 통해 데이터 분석 parameters, thresholds 최적화 수행
- 사용자 편의성을 고려한 분석 파이프라인 도입
- 기존 자체 개발 프로그램 및 관련 데이터베이스를 분석 파이프라인에 연동

</td></tr>
<tr><td rowspan="2">색 인 어
(각 5개 이상)</td><td>한 글</td><td colspan="4">메타지놈, 마이크로비옴, 파이프라인, 대용량, 최적화, 생물정보학</td></tr>
<tr><td>영 어</td><td colspan="4">metagenome, microbiome, pipeline, big data, optimization, bioinformatics</td></tr>
</table>

요 약 문

Ⅰ. 제 목

극지 메타지놈 염기서열 분석 파이프라인 구축

Ⅱ. 연구개발의 목적 및 필요성

- 메타지놈 데이터 분석은 다양한 극지 환경의 생리, 생태 이해를 위해 필수적
- NGS 기술 개발 이후, 메타지놈은 생물학계의 Big Data Science로 인식됨
- 전 세계적으로 메타지놈 염기서열 데이터 분석 인프라 구축이 걸음마 단계임
- 연구소 내외부 연구자들이 자유롭게 대용량 메타지놈 데이터를 분석할 수 있는 생물정보 시스템 제공이 필요

Ⅲ. 연구개발의 내용 및 범위

- Microbiome 염기서열 데이터를 분석할 수 있는 파이프라인 구축
- 신뢰도 있는 분석 결과를 바탕으로 남극 미생물시료의 군집구조를 보다 정확하게 해석할 수 있는 툴 제공
- Shotgun metagenome 염기서열 데이터를 분석할 수 있는 파이프라인 구축
- 남극 환경샘플의 미생물 물질대사능을 평가하는데 이용

Ⅳ. 연구개발결과

- 대용량 메타지놈 염기서열 데이터 분석 시스템 최적화 수행
- 마이크로비옴 분석 파이프라인 구축
- 유전체 기반 메타지놈 염기서열 분석 파이프라인 구축
- 염기서열 시뮬레이션을 통해 데이터 분석 parameters, thresholds 최적화 수행
- 사용자 편의성을 고려한 분석 파이프라인 도입
- 기존 자체 개발 프로그램 및 관련 데이터베이스를 분석 파이프라인에 연동

Ⅴ. 연구개발결과의 활용계획

- 구축된 생물정보 분석 파이프라인을 이용하여 극지에서 생산되는 환경유전체 데이터 분석 수행
- 추후 극지생물 유전체 및 비교유전체 분석 파이프라인과 연동하여, 유전체-메타유전체 수준에서의 극지 생물 생리 이해에 활용

S U M M A R Y

I. Title

Developing a pipeline for analyzing polar metagenome sequences

II. Research goals and background

- accurate analysis of metagenome sequence data is necessary for better understanding biological physiology and ecology of organisms in polar habitats
- On behalf of development of next-generation sequencing technology, metagenomics has been regarded as one of the big data sciences.
- Analyzing big sequence data like metagenomes is still immature and not familiar to most biologists.
- Need to provide a good bioinformatics pipeline for metagenome analysis and to be shared in KOPRI

III. Research stuffs and scope

- planning to establish a bioinformatics pipeline for analyzing microbiome sequence data
- planning to provide bioinformatics tools to analyze microbial community structures of polar samples
- going to prepare a bioinformatics pipeline for analyzing whole genome shotgun metagenome sequences
- All stuffs we are going to do can be utilized for evaluating metabolic potentials of polar microbial community

IV. Research results

- conducting bioinformatics pipeline optimization for metagenome sequence analysis
- preparing a microbiome analysis pipeline
- preparing a shotgun metagenome analysis pipeline
- preparing a user-friendly interface for metagenome analysis
- integrating previously developed cognate tools to the current version of the pipelines

V. Application plan

- Utilizing the prepared bioinformatics pipelines to analyze the big sequence data produced from polar samples
- Considering the upcomming development of single genome or comparative genome pipelines, the integrated system will be useful for more deeply understanding ecophysicology of polar organisms at genome and metagenome levels.

C O N T E N T S

목 차

제 1 장 서론

제 1-1절: 연구의 필요성 (종합)

○ 메타지놈 데이터 분석은 다양한 극지 환경의 생리, 생태 이해를 위해 필수적

○ NGS 기술 개발 이후, 메타지놈은 생물학계의 Big Data Science로 인식됨

○ 전 세계적으로 메타지놈 염기서열 데이터 분석 인프라 구축이 걸음마 단계임

○ 연구소 내외부 연구자들이 자유롭게 대용량 메타지놈 데이터를 분석할 수 있는 생물정보 시스템 제공이 필요

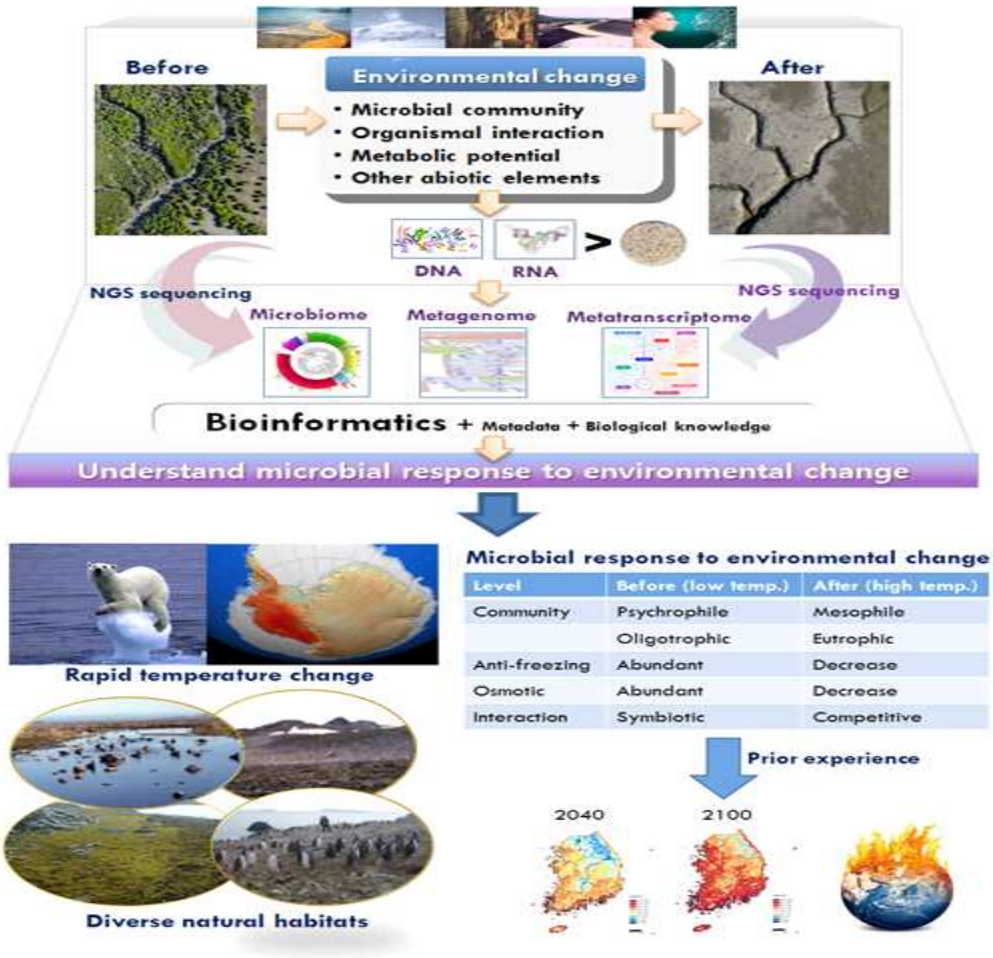


그림 1. 극지생물의 생태 및 생리 연구를 위한 메타지놈의 중요성

제 1-2절: 기술적 측면에서의 필요성

○ 극지 생태계는 물리, 화학, 생리적 조건하에 여러 생물들의 상호작용으로 유지됨. 미생물은 주변 생물들과 공존하며 진화해 왔기 때문에, 미생물 다양성 및 물질대사 능력의 평가는

극지 생태 이해를 위해 필수적임

○ 전 지구적으로 미생물(진핵미생물 제외)은 최대 3백50만여 종에 이를 것으로 추정됨. 하지만 현재까지 오직 1만2천여 종이 학계에 알려져 있으며, 보고되지 않는 대다수의 미생물들이 난배양성 또는 배양이 불가능할 것으로 예측됨

○ 모든 미생물 생리, 생태 연구는 배양이 첫걸음임. 하지만 미생물 Omics (유전체, 전사체, 단백체) 정보를 이용하여 관련 연구를 간접적으로 진행할 수 있음

○ 메타지놈은 미생물 배양 과정을 생략하고, 임의의 환경 샘플에 존재하는 전체 미생물의 유전체, 전사체를 직접 추출하여 미생물다양성 및 물질대사 능력 등을 평가하는 학문. 따라서, 배양이 불가능한 극지 미생물의 생태, 생리학적 역할을 평가하는데, 메타지놈 분석이 필수적임

○ 1985년 미국 콜로라도 대학의 Norman Pace는 환경 샘플에서 DNA를 추출하여, 16S PCR, cloning, 시퀀싱을 통해 미생물 microbiome연구를 세계 최초로 시작함. 하지만, gene cloning의 효율 및 생산성 문제로 인해 이 당시 microbiome 연구는 환경 내 몇몇 우점 미생물들을 screening 하는 수준에 그침

○ 2005년을 기점으로 Next Generation Sequencing (NGS) 기술이 개발되기 시작함. 이들 대용량염기서열결정법은 cloning단계 없이 DNA 추출 후 PCR 증폭 산물을 바로 시퀀싱 할 수 있게 함. 또한 기존 cloning기반 Sanger기술에 비해 저비용으로 수천~수만배 이상의 염기서열 생산성을 보여줌

○ 2000년대 후반부터 미생물다양성 및 메타지놈 연구도 본격적으로 NGS를 이용하기 시작함. Microbiome으로 규정되는 미생물다양성 연구의 경우 taxonomic resolution을 극대화하기 위해 Pyrosequencing 기반의 연구가 보편화 됨. 메타지놈의 또 다른 한 축은 환경 내 미생물이 담당하는 물질대사능력을 평가하고자 하는 ‘shotgun metagenome’ 임. 이는 환경 내 존재하는 모든 미생물 유전자 (또는 RNA)를 시퀀싱 하여, 이들 유전자의 기능을 in-silico로 규명하고, 물질대사 pathways에 맵핑하여 유전자 수준에서 biogeochemical cycles을 이해하고자 하는데 목적이 있음

○ 일반적으로 흙 (soil) 1 gram에 존재하는 미생물세포 수는 평균 4천만개. 미생물 유전체의 평균 크기를 4Mb라 할 때, 1 gram의 흙속에 존재하는 전체 미생물유전체를 시퀀싱하기 위해서는 산술적으로 160 Terabyte (160,000 Giga byte = 인간 유전체 크기의 5만 3천배)가 요구됨. 이는 현재까지 가장 큰 생산성을 보이는 HiSeq2000을 이용할 경우 약 300 runs을 가동시켜야 얻을 수 있는 DNA 양임.

○ 현재까지 생산된 모든 유전체 및 microbiome데이터를 비교하면 (생산된 유전체의 bases 개수 기준), 세균 유전체가 1.8%, 고세균 유전체가 약 0.02%, 진핵생물 유전체 (인간, 고등동물 포함)가 0.7% 등이고, microbiome이 약 97.3%를 차지함

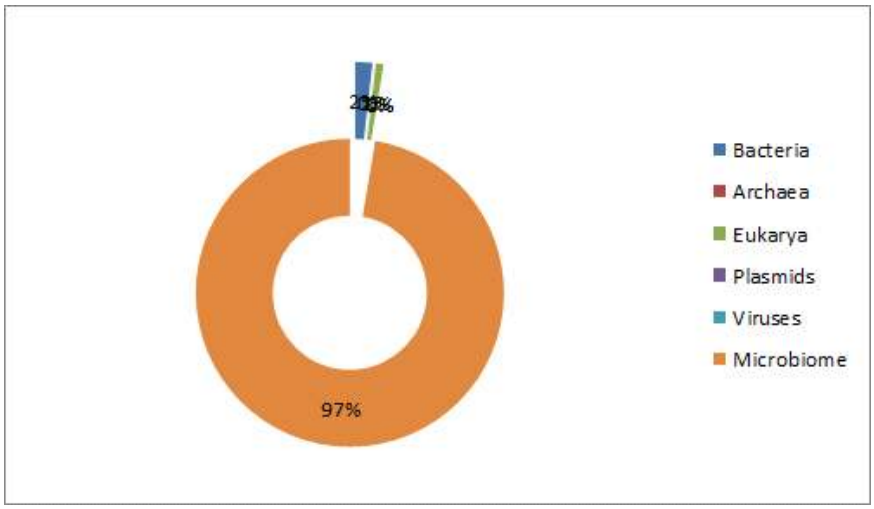


그림 2. 분류군별 전체 유전체와 마이크로비옴 nucleotide 개수 비교

○ 데이터 량 측면에서 메타지놈 (microbiome+shotgun metagenome)은 다른 omics분야와 비교 불가함. 더욱이 단일 organism과 달리 미생물은 환경 내에서 매우 복잡한 network으로 얽혀 있음. 따라서 메타지놈이 다른 omics 분야와 비교하여 진정 생물학계의 big data 분야라 할 수 있음

○ 엄청난 양의 데이터가 생산되고 있음에도 불구하고 메타지놈 염기서열 분석 프로그램 개발은 전 세계적으로 초보적인 수준임. 실제로 shotgun metagenome분석에 필요한 다양한 프로그램들은 데이터 성격이 다름에도 불구하고 개별 organism 유전체 데이터 분석에 사용했던 프로그램들을 차용해 와서 사용하고 있는 실정임

○ 하지만 극지 연구자들은 대용량 염기서열 데이터 기반 미생물다양성, 생태 및 메타지놈 연구 수행 필요성을 느끼고 있으며, 다양한 극지 시료로부터 대량의 염기서열 데이터가 축적되고 있음

○ 이에 본 연구과제를 통해, 기존에 개발된 Microbiome 및 shotgun metagenome 염기서열 분석 프로그램을 조합하여 파이프라인을 구축하고자 함

제 1-3절: 경제, 산업적 측면에서의 필요성

○ 미생물 메타지놈 분석 기술 자체에 대한 가치를 평가하기는 쉽지 않음. 다만, 해당 기술의 구현을 통해 2차적으로 경제 및 산업분야 파급 효과를 불러올 수 있음

○ 한 예로, 메타지놈 연구를 통해 고효율 또는 신규 효소를 발굴할 경우, 2020년 약 1000억 달러로 성장할 것으로 예상되는 바이오매스/바이오카탈리스트 시장의 매출 증대에 기여할 것임. 특히, 극지 환경에서는 상업성이 큰 저온활성효소 발굴이 용이함

제 1-4절: 과학적 측면에서의 필요성

○ 2007년 미국 NIH 의학 연구 로드맵 보고서에서는 메타지놈 연구가 올해의 ‘one of the new pathways to Discovery’로 선정됨

○ 2014년 세계 경제 포럼은 올해의 ‘Top 10 emerging technologies’의 하나로 메타지놈 분석 기술을 선정함

○ 이렇듯 메타지놈 연구의 중요성에도 불구하고 아직 전 세계적으로 메타지놈 분석 기술은 미성숙 단계이고, 누구든 연구를 통해 more efficient, more accurate, more useful한 기술을 창출해 낼 기회가 있는 분야임

○ 기존 454기반 메타지놈 데이터 생산이 Illumina 계열의 HiSeq 또는 MiSeq 이용 추세로 트렌드가 변화하고 있음. 따라서, 다양한 시퀀싱 프로토콜에 robust한 분석시스템 개발이 필요함

○ 앞으로 Nanopore sequencing technology가 상용화 될 것으로 예측됨. 따라서, 현재 메타지놈 데이터 분석의 가장 큰 bottleneck인 sequence assembly 문제가 해결되고, 대량의 환경 샘플간 comparative study가 강조될 것으로 예상됨. 이에 메타지놈 분석 프로그램 및 파이프라인을 클라우드 환경 속에서 구축하여 Big metagenome data era를 대비할 필요가 있음

제 2 장 국내외 기술개발 현황

제 2-1절: 염기서열 전처리기 프로그램 개발

○ 메타지놈 염기서열 생산은 다양한 NGS machines에 의해 지원됨

○ NGS machine별로 sequencing library prep 방법이 상이하여 다양한 염기서열 전처리 기술이 필요함

○ 염기서열 전처리를 위해 많은 프로그램들이 개발됨. 하지만 대부분 리눅스 환경에서 구동되어 일반 생물학자들이 사용하기 힘든 단점이 있음

○ 또한 microbiome 연구 특성상 다수의 환경샘플 염기서열 데이터를 분석해야 하지만, 기존 프로그램들은 단일샘플별 분석기능만 제공하는 한계가 있음

○ 메타지놈 염기서열 전처리시 다양한 옵션제공 (homopolymer에러 검증, 서열 quality 조사등)이 필요함

○ 대용량 메타지놈 염기서열 reads의 tag information을 제거하는 프로그램을 구현함. 본 프로그램은 tag을 구성하는 barcode, linker, primer서열을 제거해 주어, downstream 분석의 정확도를 높여줌

○ Pyrosequencing은 homopolymer, polynucleotide error, substitutions등 다양한 시퀀싱 에러를 발생시킴. 본 프로그램은 이들 에러를 교정하고, 시퀀싱 reads의 품질을 향상시켜줌

○ 전처리된 서열들을 사용자의 프로젝트 및 샘플 별로 분류하여 관리할 수 있는 기능을 탑재함

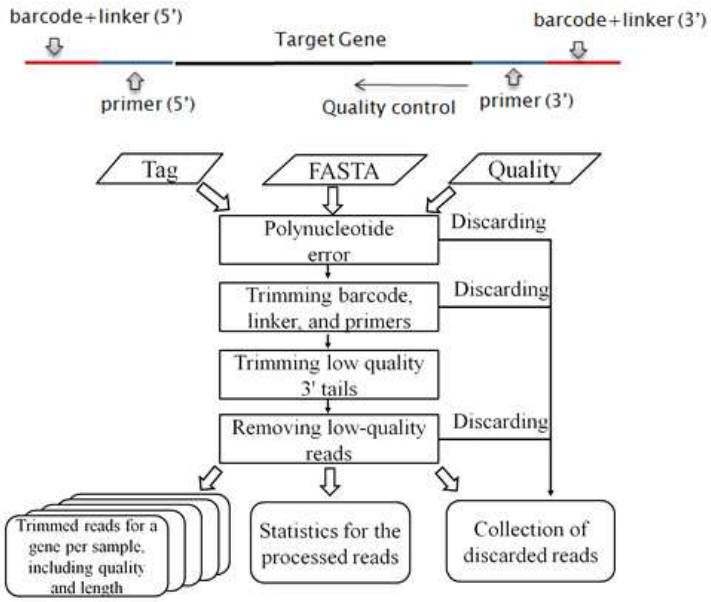


그림 3. PyroTrimmer (염기서열 전처리기) workflow

○ 메타지놈 염기서열 전처리 기능을 가진 다양한 프로그램들이 존재함
○ 하지만, 모든 프로그램들이 웹기반 또는 리눅스 환경 기반이어서 사용자 편이성이 떨어짐
○ 본 프로젝트에서는 JAVA기반 graphic user interface를 탑재한 프로그램을 개발하여 일반 사용자 이용성을 높임
○ 프로그램은 http://pyrotrimmer.kribb.re.kr에서 다운받아 사용 가능함

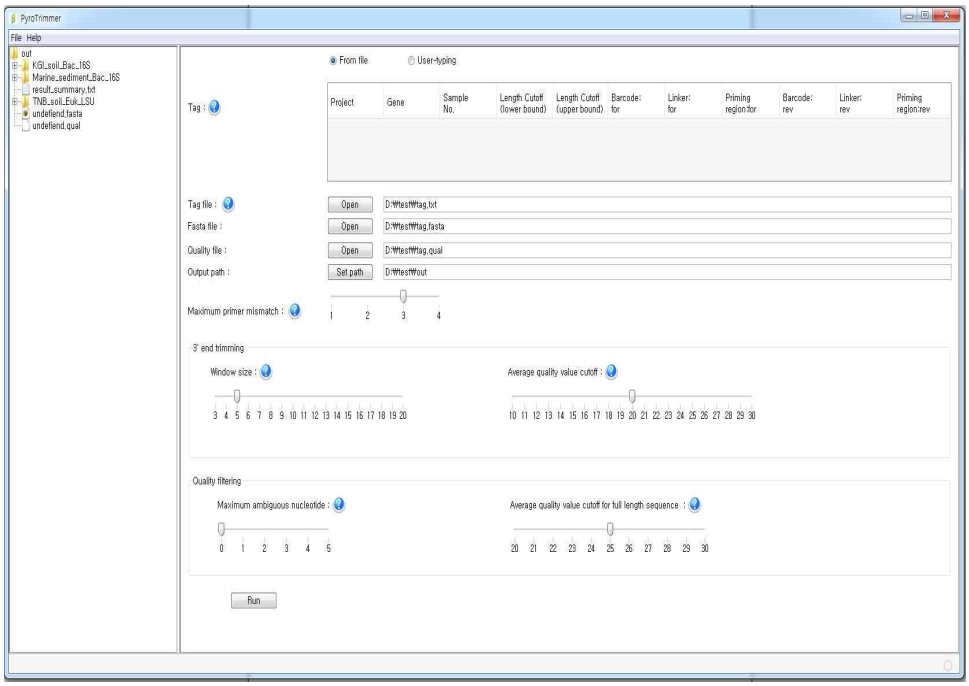


그림 4. Graphic User Interface of PyroTrimmer

제 2-2절: 염기서열 클러스터링 프로그램 개발

○ 염기서열 클러스터링은 메타지놈 데이터 분석의 핵심 단계. 클러스터링을 통해 operational taxonomic units이 결정되고, 이 수치가 모든 다양성 지수 계산에 이용됨

○ 기존 프로그램들을 통해 매우 다양한 클러스터링 알고리즘이 제공됨 (예: greedy algorithm, hierarchical clustering algorithms등)

○ 기존 프로그램들은 염기서열 유사도에 근거하여 클러스터링을 수행. 계산 정확도가 떨어짐

○ 미생물 종 분화 원리 (speciation)를 이용하여 신규 염기서열 클러스터링 알고리즘 구현 및 프로그램 개발미생물 종 분화 원리 (speciation)은 ecotype model을 따름

○ 진화적으로 유사한 미생물간에는 homologous recombination이 많이 일어나고, 진화적으로 먼 미생물간에는 recombination 빈도가 줄어들게 됨. 이와 같은 생물학적 원리에 따라 염기서열 유사도를 축으로 한 3D공간에서 각 taxon (예: species)은 고유의 bell-shape의 분포를 가지게 됨

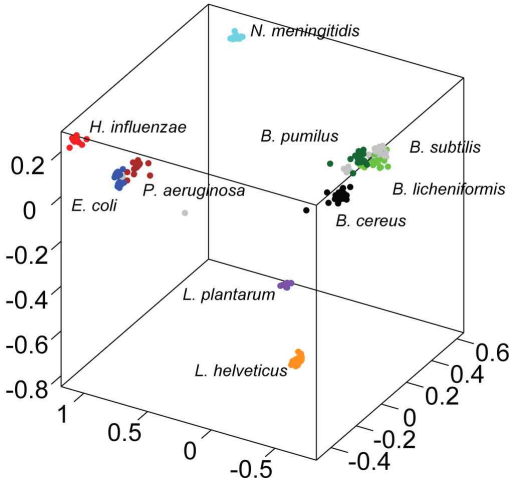


그림 5. 세균 종 염기서열 분포

○ 각 taxon의 염기서열 분포의 중심을 기점으로 user’ s defined distance cutoff를 반경으로 염기서열 클러스터링을 수행

○ 세부적으로 연산속도 향상을 위해 1) 대용량 메타지놈 서열에서 random sampling을 통해 일부 서열을 추출; 2) k-mer distance와 pairwise distance간의 관계를 고려하여 염기서열 cutoff 수준 결정; 3) network기반 서열 간 초기 clustering을 수행; 4) 초기 clustering 패턴 기반 대표서열 간 clustering 재 수행; 5) 최종 clustering 결정 단계로 알고리즘이 구성되어 있음

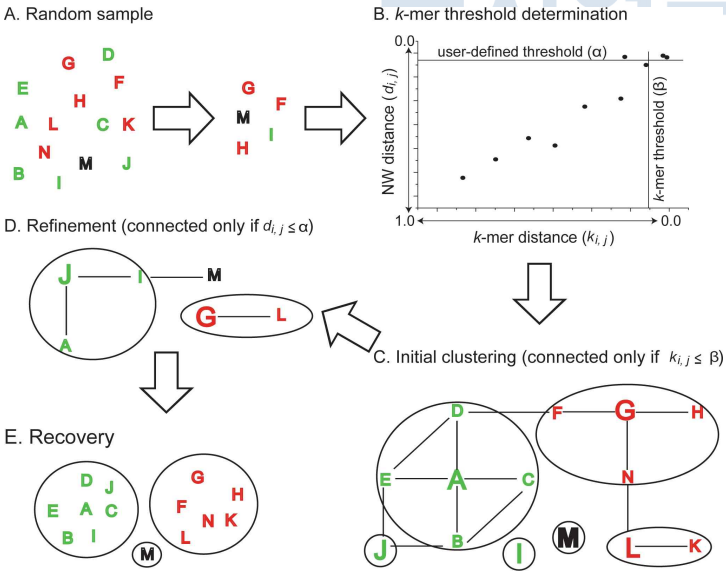


그림 6. CLUSTOM 알고리즘

○ 대용량 염기서열 처리 속도를 극대화하기 위해 내부적으로 C programming language

를 이용함

○ 서열 clustering 과정에서 발생하는 대규모 연산작업은 모두 분산처리 되도록 설계함

○ 대용량 서열 계산에 요구되는 메모리량을 최소화하기 위해 프로그램 design을 최적화함 본 연구팀에서 개발한 클러스터링 알고리즘 (=CLUSTOM) 정확도를 대표적 기존 프로그램인 DOTUR, ESPRIT-TREE, Mothur등과 비교함

○ Species, Genus 수준에서 clustering한 결과 본 프로그램이 DOTUR 다음으로 정확함을 증명함

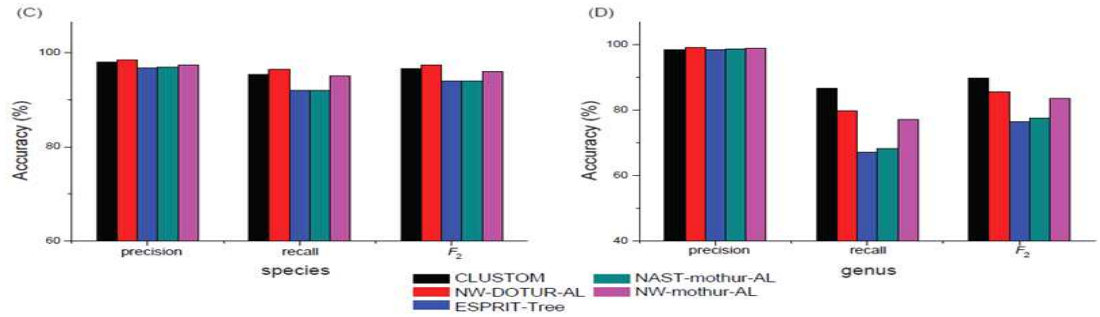


그림 7. 염기서열 클러스터링 결과 비교

○ DOTUR는 비록 가장 높은 정확도를 보였지만, 연산 속도가 매우 느려 대용량 메타지놈 염기서열 클러스터링에 사용 불가. 하지만, 본 연구팀이 개발한 CLUSTOM의 경우 몇 시간 안에 수십만 염기서열들을 클러스터링 할 수 있음. 따라서, 정확도 및 분석 속도 - 이 두 가지를 모두 고려하였을 때 현존하는 메타지놈 염기서열 클러스터링 프

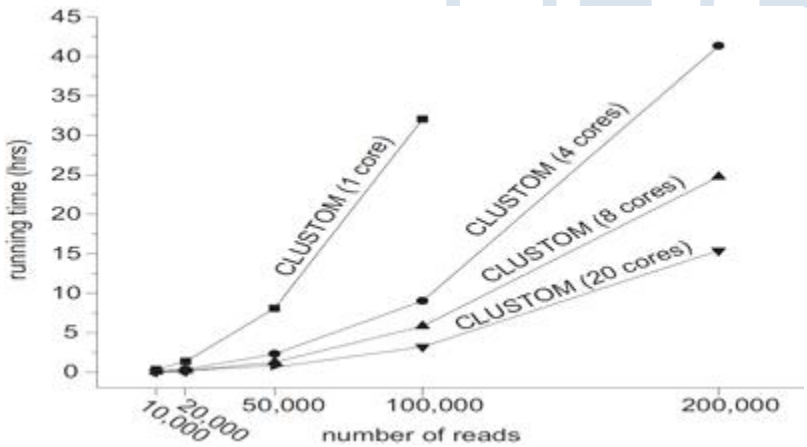


그림 8. CLUSTOM 계산 속도 측정

로그램 중 CLUSTOM이 가장 좋음

○ standalone version으로 C source code를 compile하여 사용할 수 있음

○ 일반 생물학자들의 경우 web에 직접 염기서열 데이터를 upload하여, 염기서열 유사도

cutoff값을 입력하면, Operational taxonomic units의 개수, 크기, 대표 OTU 서열, OTU별 염기서열 결과 등을 받아 볼 수 있음

그림 9. CLUSTOM web site

제 2-3절: 클라우드 기반 클러스터링 프로그램 개발

- 현재까지 차세대 염기서열 결정법에 의해 생산된 대용량 염기서열 데이터를 클러스터링 할 수 있는 프로그램은 CD-HIT과 HPC-CLUST 등이 있음
- CD-HIT은 greedy heuristic clustering 기법을 사용하여 처리 속도가 매우 빠르나, 단일 노드에서의 병렬처리만을 지원하기 때문에 프로그램의 성능 및 컴퓨터의 자원 임계치를 넘어서는 대용량의 데이터에 대해서는 처리하지 못함. 무엇보다 hierarchical clustering 알고리즘에 비해 정확도가 매우 떨어지는 단점이 존재
- HPC-CLUST 경우 hierarchical clustering 알고리즘을 사용하여 정확성이 높고 MPI를 활용한 병렬 및 분산처리를 지원하기 때문에 속도도 빠른 장점이 있으나, 클러스터링 시 필요한 Similarity matrix데이터를 프로그램 내에서 연산을 통해 생성하는 것이 아닌 외부에서 입력받아 실행되기 때문에 단독으로 수행될 수 없음. 무엇보다 서열 유사도를 기반으로 Similarity matrix데이터를 생성하는데 엄청난 시간이 들기 때문에 대용량 데이터에 대한 클러스터링을 수행하는 것이 실질적으로 매우 어려움

○ 본 연구팀은 이러한 기존의 서열 클러스터링 프로그램의 한계를 극복하기 위해 대용량 염기서열 클러스터링을 분산환경 및 클라우드 환경에서 실행 가능한, 메모리 기술 기반 클러스터링 분산처리 시스템을 개발하였음

○ 본 프로그램은 기존에 본 연구팀에서 고안한 정확한 서열 클러스터링을 위한 CLUSTOM(Hwang et al, 2013) 알고리즘을 기반으로 정확한 서열 클러스터링이 가능

○ 그러나 실행 안정성이 떨어지고 분산처리를 위한 여러 가지 설정이 필요하기 때문에 사용하기 어려운 단점이 있음

가. Cluster computer가 갖춰진 분산 환경이나 아마존 EC2와 같은 클라우드 환경에서 편리하고 빠르게 구동할 수 있도록 기존의 설정파일을 간소화 함

나. 구조화된 xml설정 파일에서 클라우드 환경에서 구동할 때 필요한 옵션 값만 입력하면 됨. 만약 클라우드 환경이 아닌 일반 클러스터 환경에서는 설정할 필요 없이 간편하게 사용가능하도록 함

다. 일반적으로 분산처리 프로그램을 구동 시 여러 설정으로 인해 사용에 어려움이 있음에 따라 손쉽게 구동할 수 있도록 구동에 필요한 설정 및 명령어를 한번에 실행하는 쉘 스크립트 파일을 만들어 사용자들이 손쉽게 사용할 수 있도록 함.

라. 클라우드 환경에서 본 연구결과물의 안정성과 확장성을 확인하기 위해 아마존 EC2 클라우드 환경에서 실험을 진행함. 실험 환경은 A*pplication* 노드는 아마존 EC2의 High-CPU Extra Large Instance 로서 2.8 GHz 32 코어 CPU와 60GB 메모리를 탑재하였으며, *Cluster* 노드는 EC2 High-Memory Extra Large Instance로서 2.5 GHz 4 코어 CPU와 30.5 GB 메모리를 탑재함. 실험 데이터는 Human Microbiome Project 에서 랜덤으로 100만개의 서열을 추출하였음. 실험 결과 클라우드 환경에서 대용량 서열을 빠르고 안정적으로 클러스터링 가능함을 확인

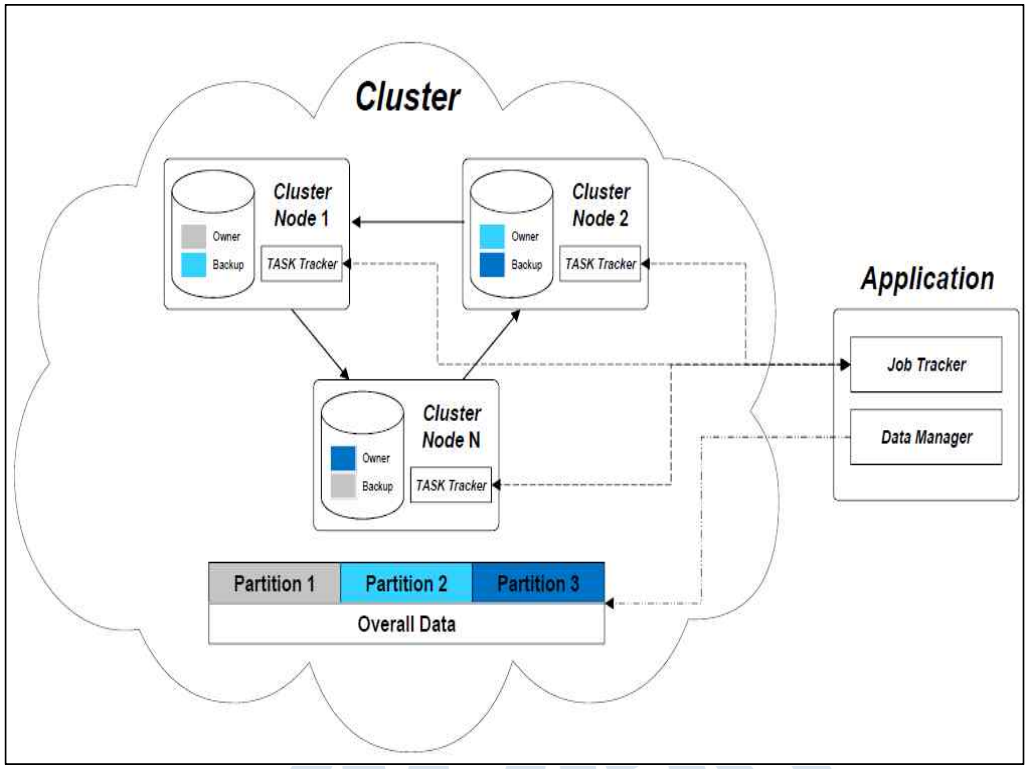


그림 10. CLUSTOM 클라우드 환경 아키텍처

- 일반적으로 차세대 시퀀싱(NGS)을 통해 나온 대용량 메타지놈 서열에는 추출한 샘플의 미생물 다양성의 특성에 따라 많은 중복서열이 발생함. 이러한 중복서열은 클러스터링 과정에서 서열간의 거리 계산 및 유사도 계산의 시간에 많은 영향을 미침. 따라서 중복서열의 제거 기능을 구현하여 전체 클러스터링 실행시간을 단축하도록 함. 데이터의 중복 개수에 따라 최대 2배 정도 시간 단축 효과를 보임
- 입력 데이터의 미생물 다양성 complexity 에 따라 서열의 중복개수에 차이의 정도가 다르며, complexity가 낮을수록(다양성이 낮을수록) 중복개수가 많아짐. 따라서 미생물 다양성이 낮은 샘플일수록 시간 단축의 효과가 더 발휘됨
- CLUSTOM 알고리즘에서는 이러한 중복서열도 네트웍 상의 클러스터의 중심을 선택하는데 필요한 중요한 요소임. 따라서 서열간의 거리 및 유사도 계산이 끝난 후 클러스터링을 결정할 시 중복서열의 개수도 참조 할 수 있도록 함. 중복된 서열들은 최종 OTU 결과에서 복구되며 최종 군집화 결과는 기존의 오리지널 CLUSTOM 버전과 차이가 없도록 구현

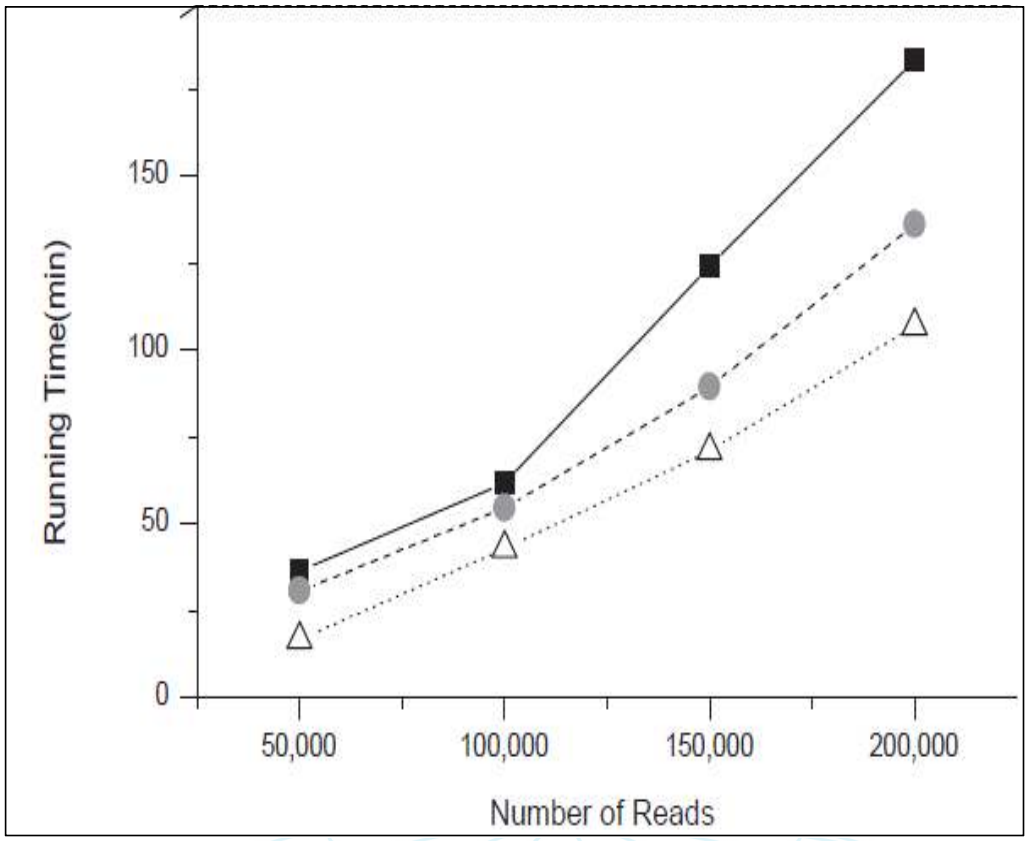


그림 11. CLUSTOM-CLOUD의 계산 속도

○ CLUSTOM-CLOUD의 실행파일 및 사용 매뉴얼은 http://clustomcloud.kopri.re.kr에서 제공. local end-user를 위해, standalone JAVA version도 동일 web에서 제공

○ 비록 유료이기는 하지만, 대용량 메타지놈 서열 클러스터링을 위해서는 아마존EC2에서 CLUSTOM-CLOUD를 실행할 수 있음

그림 12. CLUSTOM-CLOUD Website

제 2-4절: 메타지놈 분석 파이프라인 초안 개발

○ 분석 파이프라인은 Microbiome 용, shotgun metagenome 용 - 두 개로 분리해서 구축

됨 (shotgun metagenome 용은 현재 지속적으로 개발 중)

○ Microbiome 데이터 분석 파이프라인은 서열 전처리 (trimming), chimer 에러 검증, 염기서열 에러 교정, 염기서열 클러스터링, 다양성 지수 계산 (alpha- and beta-diversity), 군집 구조 계산 (community structure)로 구성됨

○ Shotgun metagenome 파이프라인은 서열 전처리 (trimming), 염기서열 short read assembly (de novo assembly), ORF region finding (structural annotation), 유전자 기능 주석 (functional annotation), 물질대사경로 mapping (metabolic potential) 단계로 구성됨

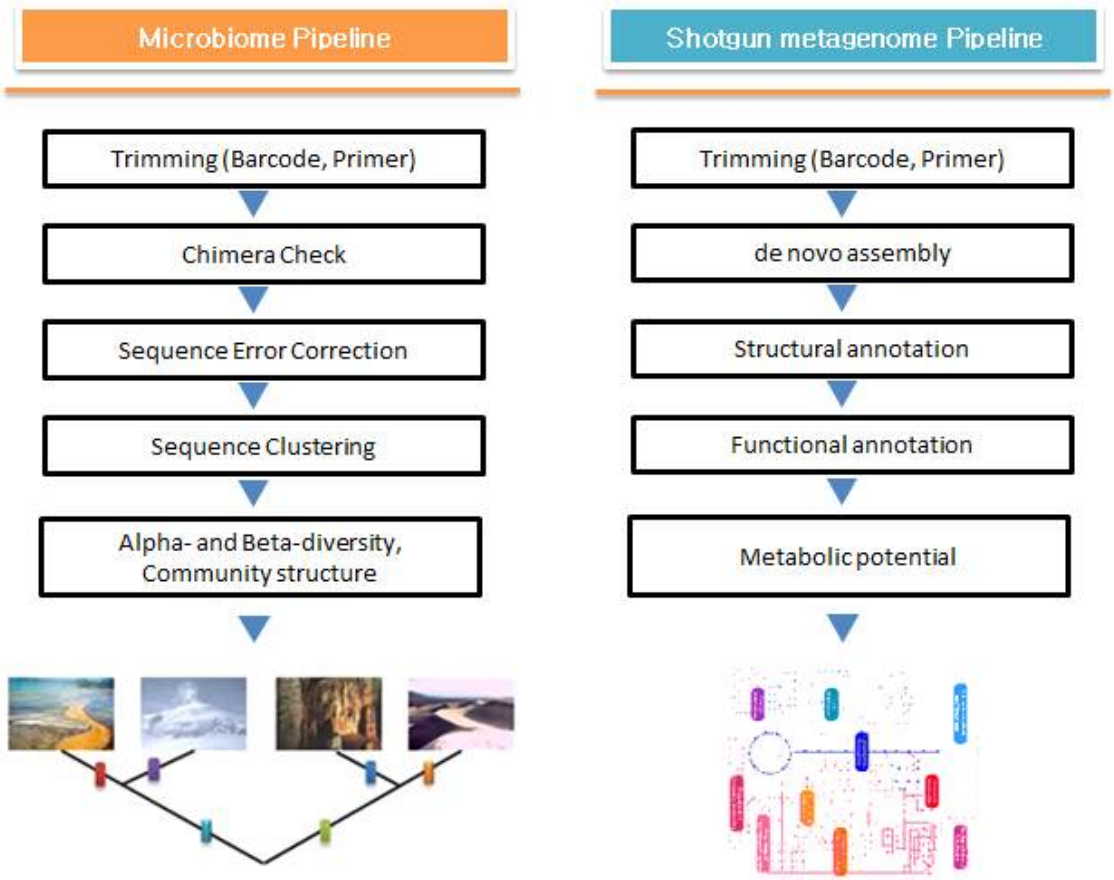


그림 13. Microbiome, Shotgun metagenome 데이터 분석 파이프라인

○ Microbiome용 분석 파이프라인을 세부적으로 살펴보면, 1) 시퀀싱 이미지 파일로부터 text format의 서열 추출; 2) multiplex barcodes의 decoding; 3) 염기서열 error correction (Pyronoise 프로그램 탑재); 4) 서열 전처리 (본 연구팀 개발 프로그램인 PyroTrimmer 이용); 5) PCR 과정에서 발생하는 chimera 에러 검증 (Uchimie 프로그램 탑재); 6) 환경샘플 간 염기서열 개수 rarefying (R module 이용); 7) 염기서열 클러스터링 수행 (본 연구팀에서 개발한 CLUSTOM 이용); 8) OTU 개수 및 OTU sizes 값을 이용하여 alpha-diversity (Shannon, Chao1, Simpson indexes 등) 계산; 9) BLAST-NW

및 RDP classifier를 이용하여 beta-diversity 계산 (taxonomic assignment to query sequences); 10) 분산처리가 지원되는 Clustal Omega를 이용하거나, NAST 프로그램을 이용해서 multiple sequence alignment 생성; 11) 계통도 작성 (PAUP 또는 FastTree이용); 12) UniFrac을 이용하여 환경 샘플 간 Community structure 추정 - 으로 구성되어 있다. 이 모든 분석과정이 하나의 Pipeline으로 연결되어 있어 일반 생물학자들이 손쉽게 microbiome 염기서열 데이터를 분서할 수 있도록 개발되어 있음

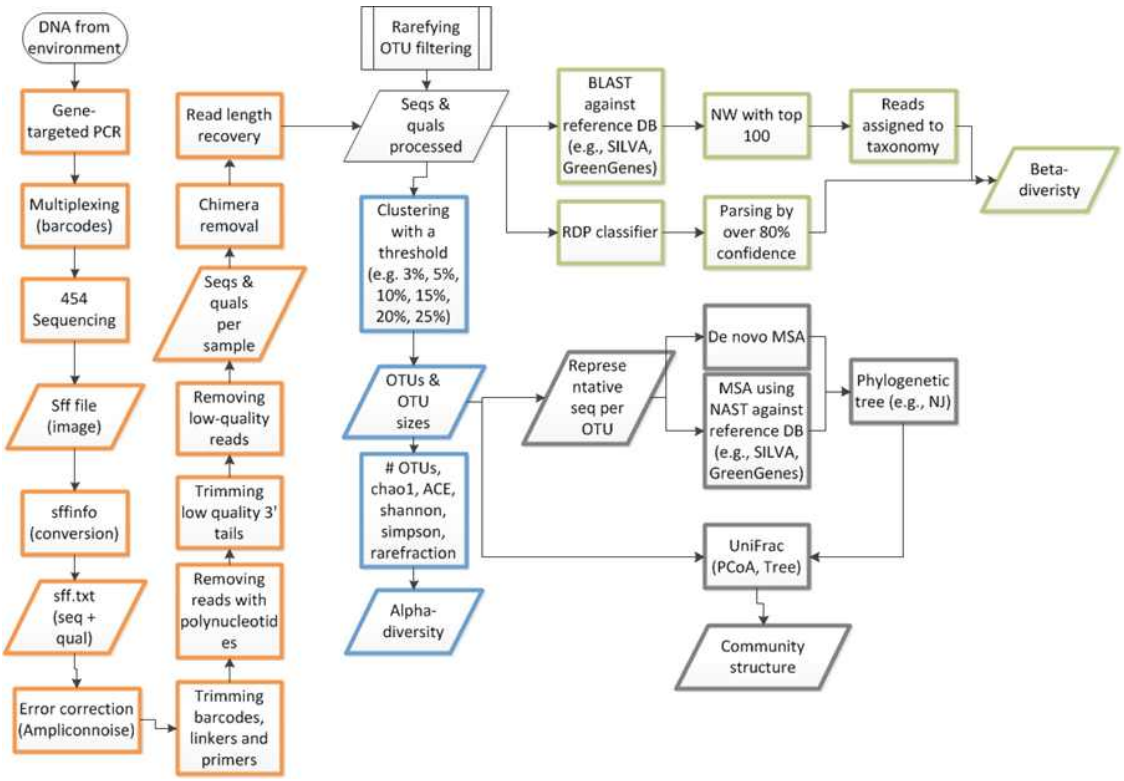


그림 14. Microbiome 염기서열 데이터 분석 세부 workflow

○ shotgun metagenome용 파이프라인은 현재 구축 중에 있음. Microbiome용 데이터와 마찬가지로 multiplex tag제거 및 염기서열 에러 교정 작업을 거친 뒤, MetaVelvet 등을 이용하여 short reads를 assembly하여 contigs을 생산함. 이 후 ORF finding 기능의 프로그램들(예: MetaGenMark)을 사용하여 structural annotation을 수행하고, Seed subsystem, COG, gene ontology databases등을 이용하여 functional annotation을 수행함. 마지막으로 KEGG에 contigs를 염기서열 유사도에 따라 비교하여 알려진 metabolic pathways에 환경 유래 유전자들을 mapping함

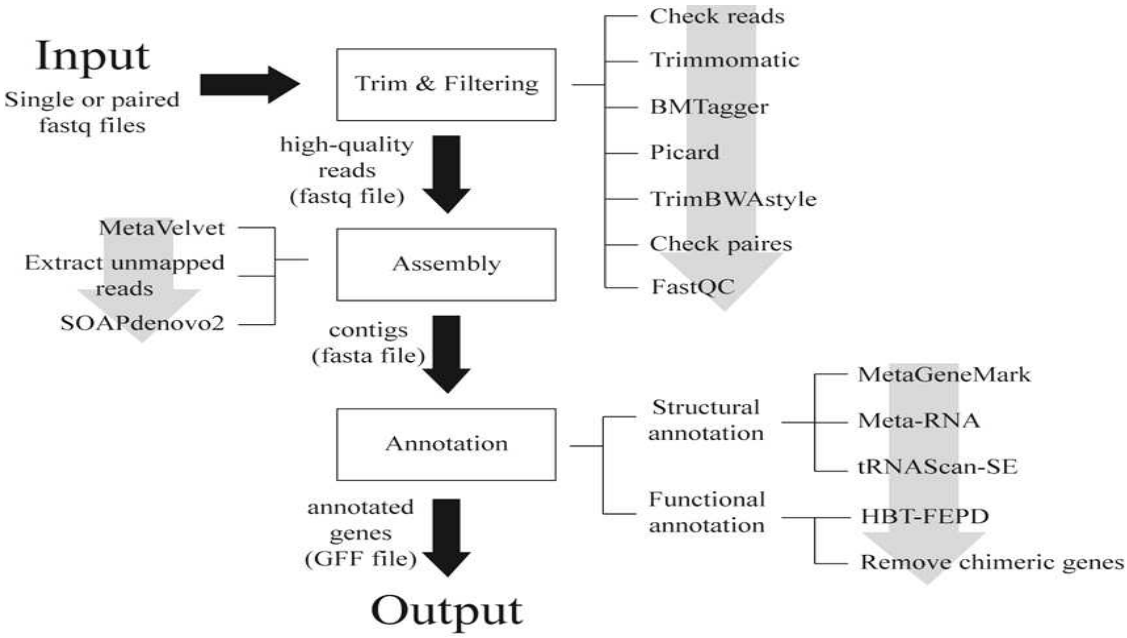


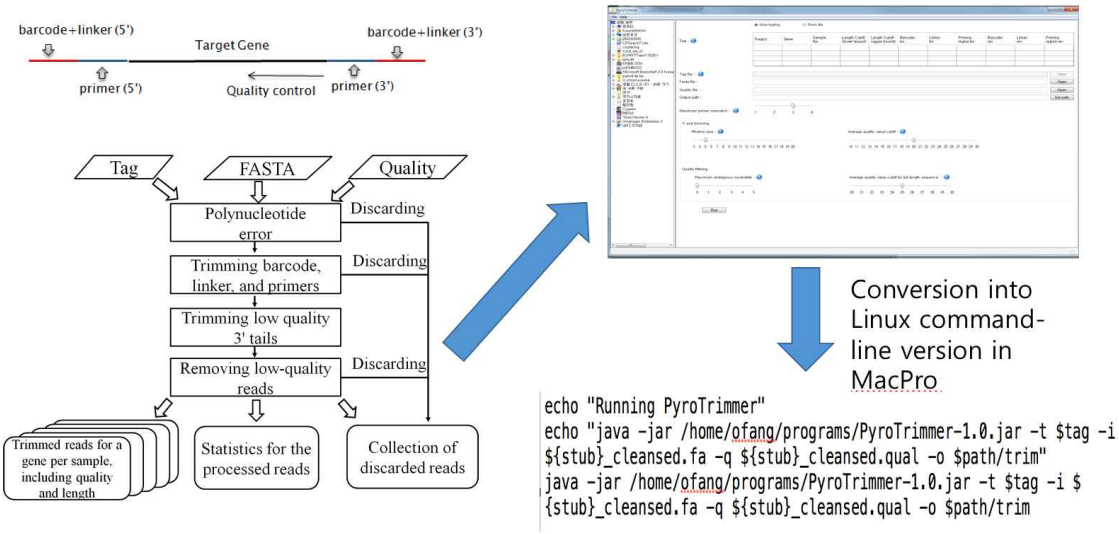
그림 15. Shotgun metagenome 데이터 분석 개념도

제 3 장 연구개발수행 내용 및 결과

제 3-1절: 마이크로비옴 염기서열 분석 파이프라인 구축

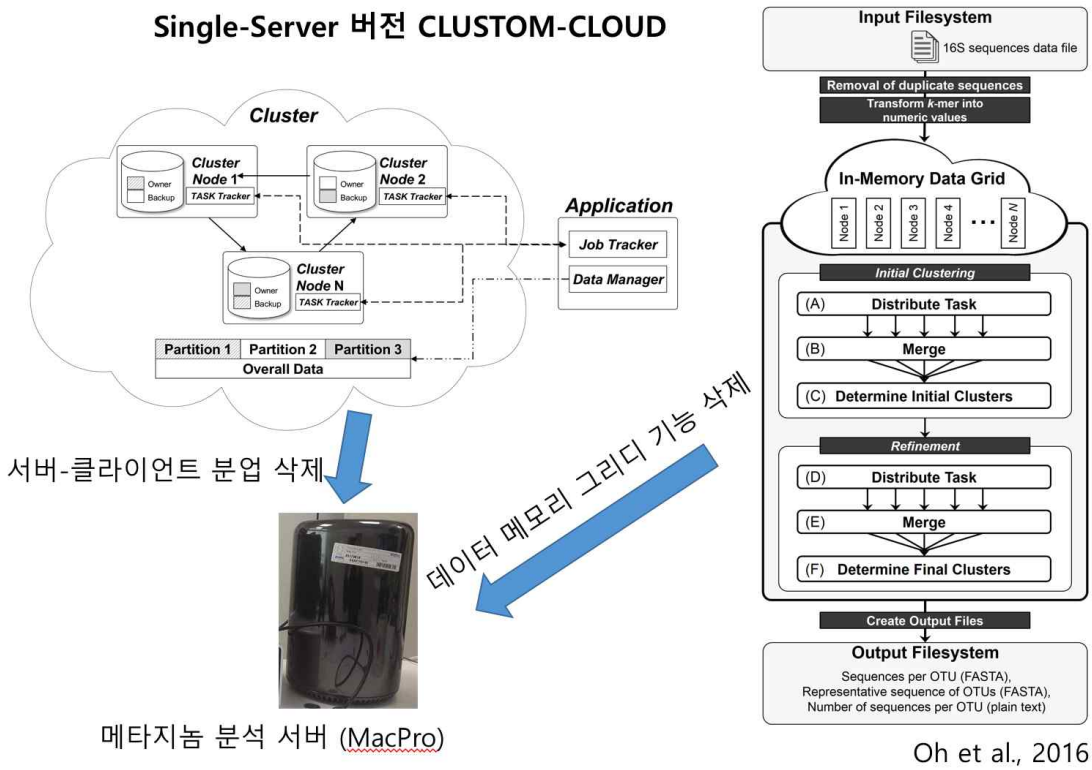
기존 자체 개발 프로그램 및 데이터베이스를 마이크로비옴(메타지놈) 분석 파이프라인에 탑재하는 것을 목표로 함. 이를 위해, 자체 개발 프로그램 중, PyroTrimmer, CLUSTOM, CLUSTOM-CLOUD와 데이터베이스 MycoDE를 분석 파이프라인에 탑재함

PyroTrimmer. 대용량 염기서열 전처리 프로그램. 차세대염기서열결정법을 이용해 얻어진 raw sequence reads에서 PCR시 사용되는 primer 서열, 환경샘플 identifiers (i.e. barcodes), 서열 3‘ 말단 low-sequencing quality 부분을 제거하여 신뢰할 수 있는 온전한 유전자 서열을 얻게 해주는 프로그램. 2012년 개발이 완료됨(Oh et al. 2012). 그 당시 개발 버전은 사용자 편의성을 고려하여 graphic user interface (GUI)를 제공하여, standalone으로 단독 사용이 가능하도록 디자인. 하지만, 본 과제에서 구축할 분석 파이프라인은 Linux system in MacPro에서 구동 예정. 따라서, 기존 PyroTrimmer에서 GUI를 걷어 내고, linux command line에서 실행하는 한 신규 버전 JAVA 실행파일(PyroTrimmer-1.0.jar)개발



PyroTrimmer 알고리즘 (Oh et al., 2012)　　마이크로비옴 분석파이프라인에 탑재

CLUSTOM-CLOUD. CLUSTOM-CLOUD는 대용량 염기서열들을 서열유사도에 따라 묶는 기능. 염기서열 기반 생물다양성 분석에 핵심이 되는 기능. 2013년 신규 알고리즘을 고안하여, C 프로그래밍 언어를 이용하여 프로그램을 구현하고, CLUSTOM (*CLUST*ering by *O*verlap *M*inimization)으로 명명 (Hwang et al. 2013). CLUSTOM은 QIIME (Caporaso et al. 2010), Mothur (Schloss et al. 2009) 등에 탑재된 기존 clustering 알고리즘과 비교하여 최고의 정확도를 보임. 하지만, 대용량 염기서열 데이터를 분석하는데 긴 시간이 걸리고, 많은 시스템 메모리량을 요구하는 단점이 있었음. 이를 보완하기 위해, cloud computing 환경에서 구동될 수 있는 CLUSTOM-CLOUD를 개발함 (Oh et al. 2016). 본 과제에서는 마이크로비옴 분석 파이프라인에 필요한 clustering 프로그램을 CLUSTOM-CLOUD로 탑재함. 분석 파이프라인이 설치된 컴퓨팅 환경이 현재는 single-server system (MacPro). 따라서, 'CLOUD' 분산 계산 기능이 현재 작동할 수 없는 상태임. 따라서, 기존 CLUSTOM-CLOUD에서 'Server'-'Client' 통신 기능을 빼고, 단일 계산서버의 command-line에서 실행할 수 있는 JAVA 버전 개발하고, 마이크로비옴 분석 파이프라인에 탑재함




```
echo "Running PyroTrimmer"
echo "java -jar /home/ofang/programs/PyroTrimmer-1.0.jar -t $tag -i ${stub}
_cleansed.fa -q ${stub}_cleansed.qual -o $path/trim"
java -jar /home/ofang/programs/PyroTrimmer-1.0.jar -t $tag -i ${stub}
_cleansed.fa -q ${stub}_cleansed.qual -o $path/trim
```

• • •

```
echo "Running CLUSTOM-CLOUD, for single-server"
echo "java -jar /home/ofang/programs/clustom_cloud.jar -p $input -c 0.03"
```

• • •

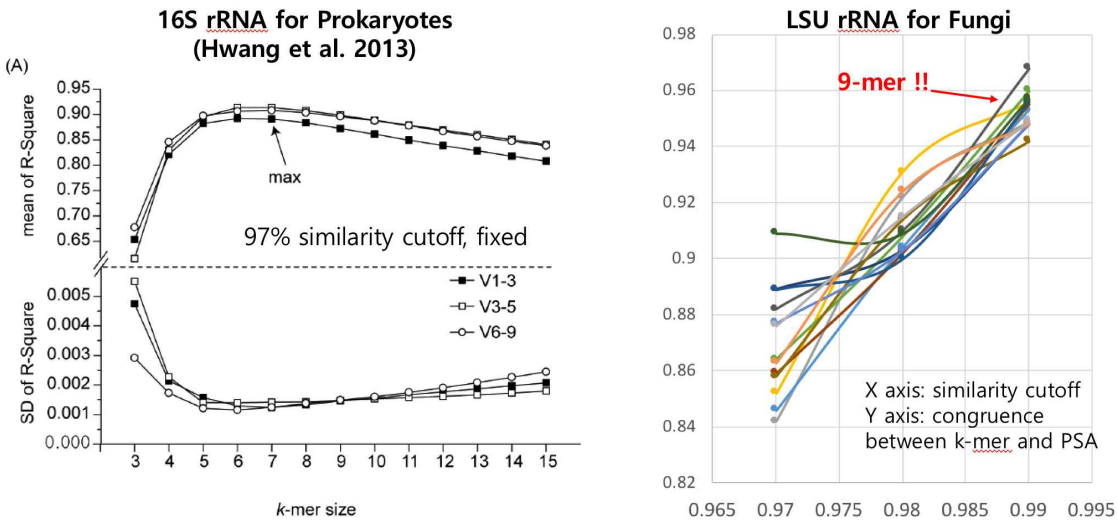
```
#echo "Extracting represent sequence"
#echo "java -jar /home/ofang/programs/ExtractRepresent.jar -p $clusterpath"
#java -jar /home/ofang/programs/ExtractRepresent.jar -p $clusterpath
```

<마이크로비옴 분석 파이프라인 script에서 PyroTrimmer - CLUSTOM-CLOUD - OTU 대표서열 추출 부분>

CLUSTOM for Fungi + MycoDE database. 2013년 개발된 CLUSTOM과 2016년 개발된 CLUSTOM-CLOUD (Hwang et al. 2013; Oh et al. 2016)은 모두 원핵생물의 small subunit of ribosomal RNA 서열을 클러스터링 하기 위해 제작됨. 따라서, 원핵생물 종을 구분할 수 있는 conventional distance cuffoff 3%이상만을 프로그램이 지원함. 즉, 서열 차이가 3% 미만인 경우, 기존 CLUSTOM, CLUSTOM-CLOUD 프로그램을 이용할 수 없다. 하지만, 곰팡이를 포함한 균류의 경우, 세균과 달리 large subunit of ribosomal gene (LSU) 염기서열을 종 동정 및 분류에 사용됨 (Schoch et al. 2012). 최근, 생물생

태분야에서, 환경샘플에서 원핵생물 뿐만 아니라, 진핵미생물 서열을 얻고자하는 수요가 증가. 따라서, 대용량염기서열결정법으로부터 얻은 균류 LSU서열을 정확하게 클러스터링 하여, 환경 샘플의 균류 다양성 지수를 계산할 필요가 증대. 이를 위해, 2013년에 개발된 CLUSTOM C언어 소스코드를 수정하여, clustering sequence distance cutoff의 3% 제한 설정을 해제함. 수정 구현된, 프로그램 "CLUSTOM for Fungi"는 cutoff 제한이 없이, LSU서열 클러스터링이 가능. 계산속도를 빠르게 하기 위해, pairwise sequence alignment using the Needle algorithm을 대체하기 위해 k-mer distance를 도입. k-mer distance와 pairwise sequence alignment간 congruence를 최대화하는 k-mer size를 결정하기 위해, 환경샘플 유래 균류 NGS 서열을 100반복, random sampling하여 (sample size = 1,000 LSU sequences), k-mer size를 3부터 15까지 바꾸면서 pairwise sequence alignment와 k-mer distance간 congruence를 계산. 그 결과, 균류 LSU 서열의 경우, 원핵생물의 16S rRNA 서열과 달리 (7-mer in prokaryotes), 9-mer 일 때 가장 좋은 결과를 보여줌. 그래서, 9-mer를 k-mer distance계산에 최적화 값으로 이용

Parameters & threshold optimization (k-mer size)



시뮬레이션으로 결정된 최적화된 parameter k-mer size를 결정한 후, CLUSTOM 프로그래밍 코드를 변경하고, conventional distance cutoff값인 1%로 설정하여, 균류 대용량 염기서열을 성공적으로 클러스터링 할 수 있었음. 또한, 균류는 세균과 달리 클러스터링 된 operational taxonomic units (OTUs)에 대한 taxonomic assignment를 수행할 마땅한 sequence database가 없는 실정. 원핵미생물의 경우, 유럽의 SILVA database, 미국의 ribosomal database project (RDP) & Greengenes 등 다양한 '신뢰도있는' DBs가 있다. 하지만, 균류의 경우, RDP for fungi – DB가 있지만, 정확도가 떨어지는 것으로 알려짐. 이에, 본 연구팀에서는 MycoDE라 서열 데이터베이스를 자체 개발하였고, 이를 마이크로

비옴 파이프라인에 탑재함 (MycoDE는 기존 RDP에 비해 정확함. 관련 데이터 제시는 진행 중인 프로젝트라, 생략). “CLUSTOM for Fungi’를 통해 얻어진, 각 OTU의 대표서열을 MycoDE DB에 BLAST하여, 정확한 taxonomic assignment 정보를 얻을 수 있었음

CLUSTOM for Fungi 프로그램

Clustom.c*

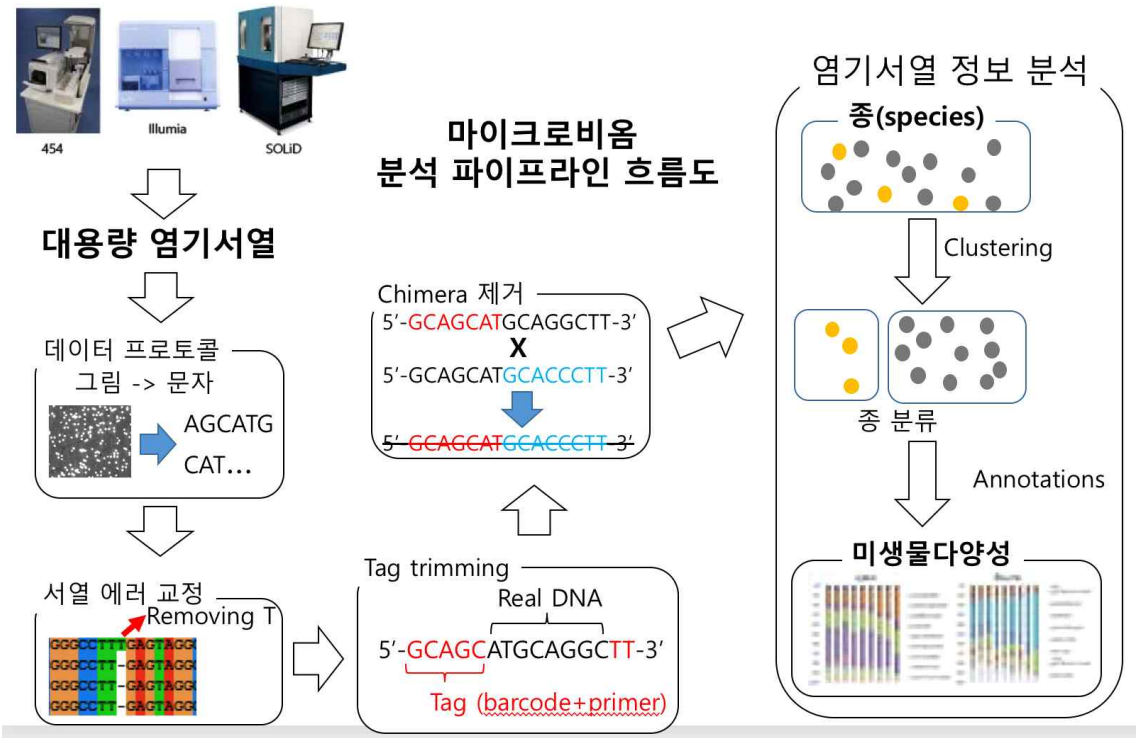
```
119   start_cutoff = 1 - atof(argv[cnt+3]);
120   end_cutoff = 1 - atof(argv[cnt+4]);
121   initial_cutoff = 0.99;
122 /*
123   initial_cutoff = 0.97;
124   if(start_cutoff > 0.971 || start_cutoff < end_cutoff)
125   {
126
127       printf("Supporting only 3%~10% distance\n");
128       printf(help);
129       exit(0);
130   }
131 */
132   if(end_cutoff < 0.899)
133   {
134       printf("Supporting only 0%~10% distance\n");
135       printf(help);
136       exit(0);
137   }
138   sprintf(cmd,"mkdir %s",tfolder);
139
```

Taxonomic assignment from MycoDE

Cluster	Species	Accession
CL1981_IOAIYA001DHDU5	Cladonia chlorophaea_s1	EF489927
CL1981_IOAIYA001DHDU5	Cladonia pyxidata	EF489926
CL1981_IOAIYA001DHDU5	Cladonia peziziformis	AY756320
CL1981_IOAIYA001DHDU5	Cladonia chlorophaea_s2	EF489928
CL1981_IOAIYA001DHDU5	Stereocaulon alpinum	EF489952
CL1982_IOAIYA001DXBAI	Usnea hirta	AJ457151
CL1982_IOAIYA001DXBAI	Melanelia hepatizon	DQ923667
CL1982_IOAIYA001DXBAI	Usnea ceratina	AJ457142
CL1982_IOAIYA001DXBAI	Cladonia peziziformis	AY756320
CL1982_IOAIYA001DXBAI	Cetrelia chicitae	DQ923658
CL1983_IOAIYA001DJA6Q	Exophiala dermatitidis	DQ823100
CL1983_IOAIYA001DJA6Q	Phaeomoniella capensis	FJ372408
CL1983_IOAIYA001DJA6Q	Thelidium incavatum	EF643780
CL1983_IOAIYA001DJA6Q	Verrucaria hochstetteri	EF643795
CL1983_IOAIYA001DJA6Q	Verrucaria lecideoides	EF643798

마이크로비옴 파이프라인 구축. 기존 자체 개발 프로그램 및 데이터베이스를 메타지놈 분석 파이프라인에 탑재. 위에 열거한 내용 대로, 1) 기존 graphic user interface를 갖추어 구현된 대용량염기서열전처리 프로그램 PyroTrimmer는 MacPro에서 linux command-line version용으로 재개발되어 신규 마이크로비옴 분석 파이프라인에 탑재, 2) 현재 준비된 computing resource가 단일 서버(MacPro)인 관계로, 기존 CLUSTOM-CLOUD는 server-client 통신 기능, 물리적으로 분리된 physical memories 간 data grid 형성 기능을 뺀 버전이 새로 개발되어, 분석 파이프라인에 탑재, 3) 메타지놈 연구에서 원핵생물 뿐만 아니라, 균류도 분석을 할 수 있게 하기 위해서, 균류 서열 분석을 위한 CLUSTOM를 개발하고, 균류 전용 염기서열 데이터베이스인 MycoDE를 파이프라인에 연결 – 1), 2), 3)을 통해, 마이크로비옴 분석 파이프라인을 고도화 함

마이크로비옴 분석 파이프라인 workflow. 새로 구축된 파이프라인은 1) NGS (대용량염기서열) raw data file 변환, 2) 염기서열 에러 교정, 3) PyroTrimmer의 linux command-line version을 이용한 서열 전처리, 4) CLUSTOM-CLOUD를 이용한 서열 클러스터링 (세균과 균류서열 분석용 분리), 5) MycoDE를 이용한 균류 OTUs의 종 동정 – 이 5개 step으로 구성되어 있다. OTUs 결정 이후, 여러 downstream 분석 (예: community structure, alpha-diversity indeces계산 등)은 R packages, QIIME, Mothur 등을 이용가능



Step 1. NGS (대용량염기서열) raw data file 변환

```
if [ ! -f ${stub}.dat ]; then
    echo "FlowsFA200KeyDistance.pl $primer $stub $key $barcode< ${sfffile}"
    #FlowsFA200Key.pl $primer $stub $key < ${sfffile}
    #FlowsFA200KeyDistance.pl $primer $stub $key $barcode < ${sfffile}
    #perl
/home/ofang/programs/AmpliconNoiseV1.25/Scripts/FlowsFA250KeyDistance.pl
$primer $stub $key $barcode < ${sfffile}
     perl
/home/ofang/programs/AmpliconNoiseV1.25/Scripts/FlowsFA200KeyDistance.pl
$primer $stub $key $barcode < ${sfffile}
fi
```

Step 2. 염기서열 에러 교정

```
if [ ! -f ${stub}_U.fa ]; then
    echo "Getting unique sequences-FastaUnique -in ${stub}.fa > ${stub}_U.fa"
    FastaUnique -in ${stub}.fa > ${stub}_U.fa
fi
if [ ! -f ${stub}_U_I.ndist ]; then
    echo "mpirun -np $np NDist -i -in ${stub}_U.fa > ${stub}_U_I.ndist"
    mpirun -np $np NDist -i -in ${stub}_U.fa > ${stub}_U_I.ndist
```

```
fi
if [ ! -f ${stub}_U_I.list ]; then
    echo "Cluster sequences..-FCluster -a -w -in ${stub}_U_I.ndist -out ${stub}_U_I > ${stub}_U_I.fcout";
    FCluster -a -w -in ${stub}_U_I.ndist -out ${stub}_U_I > ${stub}_U_I.fcout
fi
SplitClusterEven -din ${stub}.dat -min ${stub}.map -tin ${stub}_U_I.tree -s 5000 -m 1000 > ${stub}_split.stats

for c in C*
do
        if [ -d $c ] ; then
                 echo "mpirun -np $np PyroDist -in ${c}/${c}.dat -out ${c}/${c} > ${c}/${c}.fout";
                mpirun -np $np PyroDist -in ${c}/${c}.dat -out ${c}/${c} > ${c}/${c}.fout
        fi
done

for c in C*
do
        if [ -d $c ] ; then
                FCluster -in ${c}/${c}.fdist -out ${c}/${c}_X > ${c}/${c}.fout
                rm ${c}/${c}.fdist
        fi
done

for dir in C*
do
        if [ -d $dir ] ; then
                mpirun -np $np PyroNoiseM -din ${dir}/${dir}.dat -out ${dir}/${dir}_s60_c01 -lin ${dir}/${dir}_X.list -s 60.0 -c 0.01 > ${dir}/${dir}_s60_c01.pout
        fi
done

mpirun -np $np SeqDist -in All_s60_c01_T220.fa > All_s60_c01_T220.seqdist

FCluster -in All_s60_c01_T220.seqdist -out All_s60_c01_T220_S >
```

```
All_s60_c01_T220.fcout

mpirun -np $np SeqNoise -in All_s60_c01_T220.fa -din All_s60_c01_T220.seqdist -lin All_s60_c01_T220_S.list -out All_s60_c01_T220_s30_c08 -s 30.0 -c 0.08 -min All_s60_c01.mapping > All_s60_c01_T220_s30_c08.snout

if [ -z $barcode ] ; then
        java -jar /home/ofang/programs/MakeCleansedFasta.jar -p $path -i ${stub}.fa -s ${sfffile}
else
        java -jar /home/ofang/programs/MakeCleansedFasta.jar -p $path -i ${stub}.fa -s ${sfffile} -b $barcode
fi
```

Step 3. PyroTrimmer의 linux command-line version을 이용한 서열 전처리

```
java -jar /home/ofang/programs/PyroTrimmer-1.0.jar -t $tag -i ${stub}_cleansed.fa -q ${stub}_cleansed.qual -o $path/trim
```

Step 4. CLUSTOM-CLOUD를 이용한 서열 클러스터링 (세균과 균류서열 분석용 분리)

```
echo "Running CLUSTOM-CLOUD, for single-server"
echo "java -jar /home/ofang/programs/clustom_cloud.jar -p $input -c 0.03“
echo "Running CLUSTOM-CLOUD, for Fungi"
echo "java -jar /home/ofang/programs/clustom_fungi.jar -p $input -c 0.01"
```

Step 5. MycoDE를 이용한 균류 OTUs의 종 동정

```
#echo "Extracting represent sequence"
echo "java -jar /home/ofang/programs/ExtractRepresent.jar -p $clusterpath"
java -jar /home/ofang/programs/ExtractRepresent.jar -p $clusterpath
echo “blastn -query representative.fa -db /home/ofang/DB/mycoDE -out rep_fungi.blast -evalue 0.01 -num_threads 10”
```

제 3-2절: 전장유전체 메타지놈 염기서열 분석 파이프라인 구축

Sthogun metagenome (전장 유전체 메타지놈)은 환경샘플 안에 존재하는 모든 유전자의 염기서열을 시퀀싱하는 것을 목표로 한다. 따라서, 단일 유전자의 종 다양성을 분석하는 ‘마이크로비옴 파이프라인’에 비해, 차세대염기서열결정법에 의해 생산되는 서열 데이터가 수천배 크다. 또한, 단일 유전자 기반 마이크로비옴 분석과 달리, 매우 짧은 염기서열

로부터 온전한 유전자 서열을 복원하기 위해, 서열조립(sequence assembly)과정이 필요. 이는 메타지놈 분석 시, 대량의 computing resources (CPU, physical memory)가 필요함을 의미함. 결국, 전반적으로 마이크로비옴 분석에 비해, 분석 과정이 복잡하고, 고성능 서버가 필요함

메타지놈 염기서열 분석 전략 차이

	Assembly-based	Mapping-based
Contig	Longer	Shorter (reads)
Accuracy	Better	Worse
Running time	Longer	shorter
RAM memory	More	Less

위 표에서 볼 수 있듯이, 'assembly-based' 접근법은 Illumina short reads를 조합하여, 상대적으로 길이가 긴 contigs를 만들어 낸다. 이렇게 긴 contigs는 short reads에 비해 염기서열 정보량이 많아, 각 sequence에 대한 functional annotation 정확도를 높일 수 있다. 반면, contig를 만들어 내기 위해서는 생산된 모든 Illumina short reads를 컴퓨터 메모리에 올려야 하고, assembly 과정에서 reads간 connection graph정보를 생산하는데에도 대량의 메모리가 필요하게 된다. 실제, NCBI SRA SRX3549446 (short read archive)에서 Hawaiian river soil에 대한 shotgun metagenome sequence data를 내려받아 (1 Illumina MisEq run: 3.7 Gb, 250 bp paired end), assembly-based 접근법을 현실적으로 사용할 수 있는지 테스타하였다. 일단, 외부의 36 cores intel CPU, 1 Terabyte memory의 고성능 서버에, 현재까지 가장 정확하다고 알려진 metagenome short read assembly program인 ray meta (Boisvert et al. 2012)를 설치하고, public test sequence dataset인 SRX3549446를 돌려보았다.

Table 1 Comparison of assemblies produced by MetaVelvet and Ray Meta

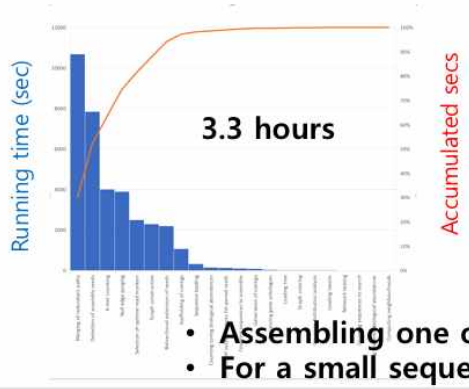
	MetaVelvet	Ray Meta	Shared
ERS006494			
Reads	372,147,956		
Scaffolds[a]	50,136	56,363	
Total length (nt)	150,904,880	156,075,852	130,979,321
Average length (nt)	3,009	2,769	
N50 length (nt)	6,141	12,117	
Longest length (nt)	146,549	570,359	

(Boisvert et al. 2012)

그 결과, 서열 assembly를 끝내는데 약 3.3시간 (12,000 secs)이 소요되었고, 약 120

Gb의 RAM memory를 필요로 하였다. SRX3549446이 3.7 Gb이기 때문에, 3.7 Mb 크기의 미생물 1,000개 정도 시퀀싱한 양이다. 즉, soil의 natural microbial complexity (수백~수천만 cells)을 고려하였을 때, SRX3549446은 매우 작은 shotgum metagenome data라 할 수 있다. 그럼에도 불구하고, 120 Gb의 RAM memory를 요구하는 것으로 보아, 보통 크기의 shotgun metagenome sequence data를 분석하기 위해서는 1 Tera급 이상의 고성능 서버자원이 필요함. 그래서, 현재 구비되어 있는 MacPro (12 Cores CPU + 64 Gb RAM memory)에서 assembly-based approach를 이용할 수 없다는 결론에 이르렀고, 대안으로 정확도는 떨어지지만 소량의 computing 자원을 요구하고 계산속도도 빠른 'mapping-based' 접근법을 선택함. 이 접근법을 지원하면서, 계산 정확도가 검증되었고, 사용자 편의성을 극대화한 Biobakery package (Lloyd-Price et al. 2017)를 MacPro에 구축하기로 결정함

시뮬레이션: sequence assembly of SRX3549446 (3.7 Gb soil data)




```
contig: 7828000003 vertices: 111 averageCover
Rank 0 objectName: 7750000014 => peakCoverage
contig: 7645000018 vertices: 112 averageCover
contig: 7458000024 vertices: 114 averageCover
Rank 7: gathering scaffold links [8001/18051]
Rank 0 ineligibleObjects: 0/111 (0.00%)
Rank 7: assembler memory usage: 1311396 KiB
contig: 8000000007 vertices: 111 averageCover
contig: 7953000021 vertices: 109 averageCover
contig: 7908000012 vertices: 112 averageCover
```

- Assembling one of contigs, 1.3 Gb required
- For a small sequence data 3.7 Gb, 125 Gb RAM required

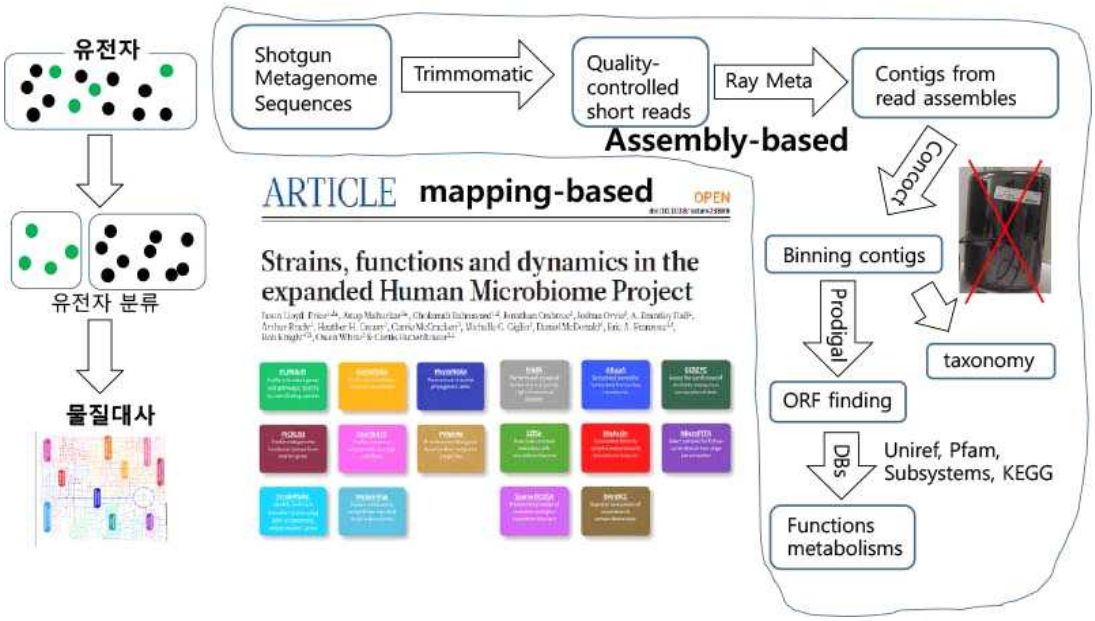
메타지놈 염기서열 분석 전략 차이

	Assembly-based	Mapping-based
Contig	Longer	Shorter (reads)
Accuracy	Better	Worse
Running time	Longer	shorter
RAM memory	More	Less

Biobakery. 대용량 메타지놈 염기서열 분석을 위해, 본 연구팀에서 기 구축해 놓은 assembly-based approach를 이용하여 파이프라인 개발을 할 계획이었음. 하지만, 프로그램 ray meta을 돌려 Illumina short reads를 assembly하기 위해서는 대용량 RAM 메

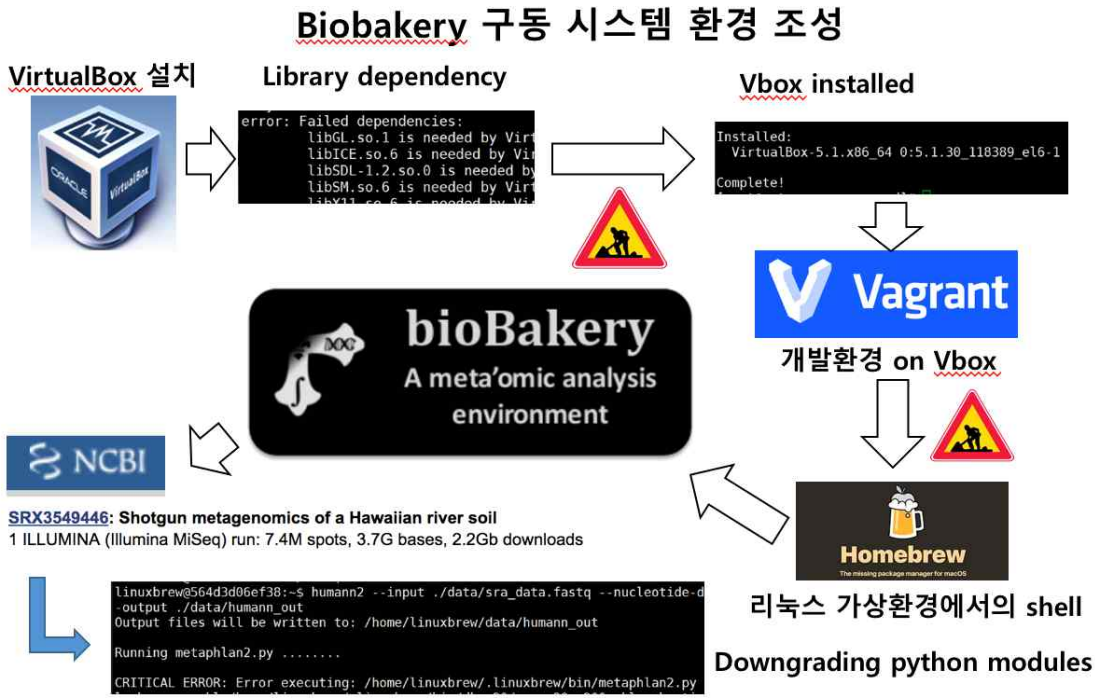
모리 (예: RAM 1 terabytes)를 탑재한 계산서버가 필요한데, 과제를 통해 구입한 MacPro (RAM 64 Gb)로는 계산이 불가능. 이에, assembly-based 접근법을 버리고, 상대적으로 RAM을 적게 소비하는 mapping-based 접근법을 채택함. mapping-based approach의 경우에도, assembly-based와 마찬가지로, 서열 전처리, contig binning for taxonomy, ORF finding, functional annotation은 동일하게 필요함. 다만, sequence assembly 대신, previously published prokaryotic reference genomes에 Illumina short reads를 mapping하는 부분이 필요. 관련된 여러 sortware를 연결하여 분석 파이프라인을 구축할 수 있지만, 2017년 Nature지 Article로 소개된 Biobakery package (Lloyd-Price et al. 2017)을 설치하기로 결정. 이는 본 과제의 연구 목표 중 하나인, 'microbiome, shotgun metgenome 분석 파이프라인 통합 구축'과 '사용자 편의성을 고려한 분석파이프라인 구축'을 동시에 만족시킬 수 있는 선택임. bioBakery는 아래 그림에서도 볼 수 있듯이, HUMAnN, MetaPhlAn, PhyloPhLAN, PICRUSt 등을 대표해 약 10개 안팎의 세부 프로그램으로 구성됨. 즉, bioBakery를 설치하면, microbiome, shotgun metagenome 분석에 필요한 대부분의 프로그램을 이용할 수 있게 됨

Mapping 기반 shotgun 메타지놈 분석 파이프라인 구축



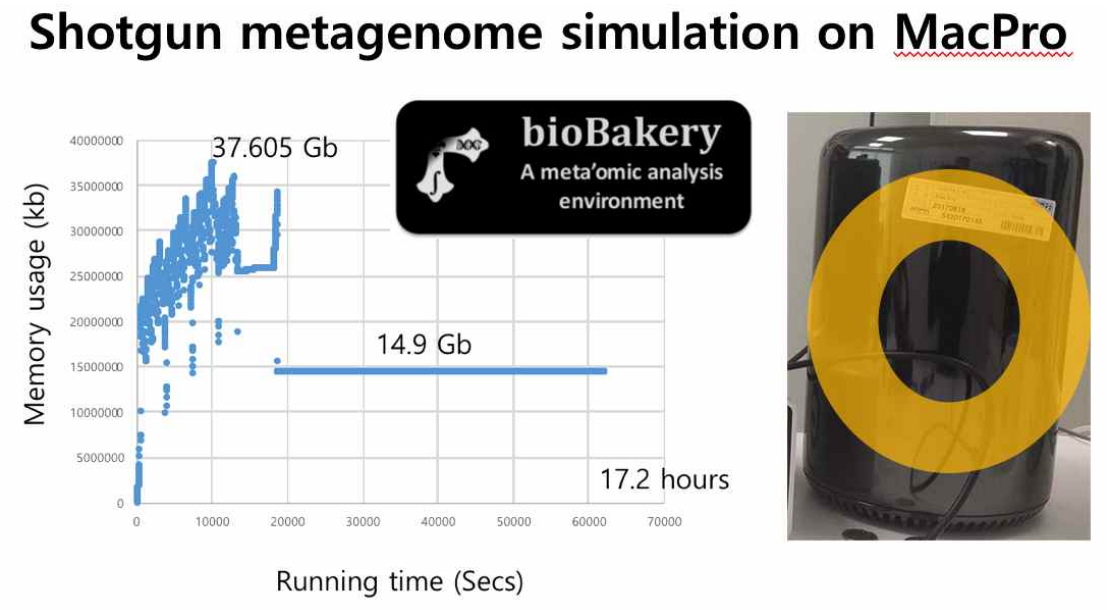
bioBakery는 MacPro내 linux환경 위에 설치된 가상환경에서 구동됨. 이에, 먼저 VirtualBox rpm을 다운로드 받아서, MacPro에 설치함. BioBakery가 개발 될 당시의 system 환경, 그리고 선호 linux의 차이 (bioBakery는 Ubuntu에서 개발), 현재 Mac의 linux system version 등, linux 종류, version 차이로 인해, 여러 system libraries에 충돌이 일어남. VirtualBox 설치, linux system과 virtualbox간 system library dependency 해결, VirtualBox 위에서 구동되는 개발환경 Vagrant 설치, 가상환경에서의 shell 환경인 Homebrew 설치를 여러 시행착오를 거쳐 완료함. 이 후, shotgun metagemome 염기서열

분석에 필요한 여러 소프트웨어를 묶어 놓은 BioBakery package를 시스템에 설치. Public sequence database인 미국 NCBI에서 shotgun metagenome sequence data SRX3549446 (MiSeq 1 run 분량, paired end, 250 bp)를 다운로드 받아, bioBakery를 실행. 그 결과, bioBakery 프로그램 내부의 python scripts에서 critical errors를 발생. 원인 분석 결과, bioBakery에 이용된 python modules (함수)은 package 개발 당시의 옛날 버전인데, 현재 MacPro에 있는 python modules은 최신 버전. 그래서, bioBakery python codes을 조사해서, 옛 pyhon modules이 최근 신규 modules로 대치 된 경우, 옛 version으로 downgrading해 줌. 이 후, bioBakery의 모든 기능이 정상적으로 돌아감



BioBakery 시뮬레이션 및 테스트. Illumina MiSeq single run에 해당하는 SRX3549446 dataset를 bioBakery를 MacPro에서 돌림. 총 running time은 약 17.2 hours이고, 아래 그래프에서 볼 수 있듯이, 평균 약 15 Gb (max = 37.6 Gb)의 RAM을 소비함. SRX3549446는 3.7 Gb size로 3.7 Mb의 미생물 유전체를 1000배 읽은 양. 즉, 메타지놈 환경샘플에 있는 미생물 세포 1,000개를 읽은 분량. MacPro의 RAM이 64 Gb이기 때문에, MiSeq 1판에 해당하는 shotgun metagenome data는 충분히 돌아 갈 수 있음을 의미. 만약, 더 큰 data의 경우, hard disk의 swap 늘려서 해결할 수 있음. 분석의 기본적인 결과로 gene family abundance와 pathway coverage spreadsheet 결과를 얻을 수 있음. Gene family abundance sheet은 실제 환경 샘플에 존재하는 유전자의 종류와 더불어, 각 유전자가 존재하는 상대적 양적 차이 정보를 제공해 준다. SRX3549446의 경우, 세균의 바이러스 감염과 관련된 유전자들 (예: protein C packing viral DNA into phage procapsid in the late stage of viral infection; spike proteins detecting types of

bacterial LPS and determining phage host-range, etc)이 매우 높은 빈도로 나타남. 또한 pathway 분석 결과를 보면, SRX3549446 (Hawaiian river soil)의 경우, 주요 primary metabolisms이 모두 나타남. 특히, cytochrome이 많이 나타나, 아마 환경 샘플이 산소가 많이 함유된 soil sample임을 의미. 또한, Pyruvate에서 주로 생합성이 일어나는 branched-chain amino acids (예: leucine, isoleucine, valine) 합성에 관련된 pathway가 강하게 나타남. 또한, TCA cycle II가 있어, soil내 식물 뿌리 associated 균근이 존재함을 의미. 그리고, aerobic respiration II는 yeast등 eukaryotic microorgnisms에 의한 산소호흡도, 해당 soil sample에서 활발함을 의미. Incomplete reductive TCA cycle이 있어, reverse TCA에 의한 CO2 (carbon dioxide) fixation이 일어 날 수도 있음을 의미. 이는 soil sample내 autotrophs이 존재할 가능성을 의미. Sulfur oxidation이 돌아가서, hydrogen sulfide가 sulfite, sulfate로 산화되서, 주변 미생물이 sulfate를 electron acceptor로 혐기호흡에 사용할 수도 있고, ROS 특히 hydrogen peroxide 항산화 (H2O2 + 2H+ -> 2H2O)에 수소를 제공할 수도 있음



Shotgun metagenome분석 결과 (Hawaiian river soil)

Abundance of gene families

# Gene Family	sra_data_edited_Abundance-RPKs
UNMAPPED	14695635
UniRef90_d	2899.25489
UniRef90_W0Z7D8|unclassified	2899.25489
UniRef90_T3IRN4	2407.30122
UniRef90_T3IRN4|unclassified	2407.30122
UniRef90_A0A024DGV5	2215.47191
UniRef90_A0A024DGV5|unclassified	2215.47191
UniRef90_P69172	1634.40343
UniRef90_P69172|unclassified	1634.40343
UniRef90_T3IKD6	1395.94646
UniRef90_T3IKD6|unclassified	1395.94646

⋮

Pathway coverage

Pathway	Coverage
TCA cycle (prokaryotic)	0.80935223
aerobic respiration I (cytochrome c)	0.88330321
L-isoleucine biosynthesis I (from threonine)	0.83434502
pyruvate fermentation to isobutanol	0.83434502
L-valine biosynthesis	0.83434502
TCA cycle II (plants and fungi)	0.7164802
adenosine deoxyribonucleotides biosynthesis	0.66028945
guanosine deoxyribonucleotides biosynthesis	0.66028945
aerobic respiration II (yeast)	0.68009783
pyrimidine deoxyribonucleotide phosphorylation	0.62969263
L-isoleucine biosynthesis	0.2919801
branched amino acid biosynthesis	0.29182674
gondoate biosynthesis (anaerobic)	0.17748072
incomplete reductive TCA cycle	0.02853405
pyrimidine deoxyribonucleotides biosynthesis	0.03316819
adenosine ribonucleotides biosynthesis	0.01999559
5-aminoimidazole ribonucleotide biosynthesis	0.0289504
sulfur oxidation	0.00024085

제 4 장 연구개발목표 달성도 및 대외기여도

연구목표	연구내용	달성도	대외기여도
극지 메타지놈 염기서열 분석 파이프라인 구축	Microbiome 염기서열 데이터 분석 파이프라인 구축	100%	- 기존 자체 개발 프로그램 및 데이터베이스 (예: Pyrotrimmer, Clustom, Clustom-cloud, MycoDE)를 마이크로비옴 분석 파이프라인에 연동 - 서열 시뮬레이션을 통해, 마이크로비옴 분석 parameters & threshold 최적화
	Shotgun metagenome 염기서열 데이터 분석 파이프라인 구축	100%	- shotgun metagenome 분석 파이프라인 및, microbiome+ shotgun metagenome 파이프라인 통합 구축 - 사용자 편의성을 고려한 분석 파이프라인 구축

제 5 장 연구개발결과의 활용계획

제 5-1절: 기술적 측면

○ NGS 기반 메타지놈 분석 기술을 선진국 수준으로 향상시킴으로서, 수준 높은 극지 미생물생태학 연구를 수행할 수 있을 것으로 기대됨

○ 균류 표준염기서열 DB (MycoDE) 구축 완료시 원핵생물 표준염기서열 DB인 Eztaxon과 더불어 국산 microbiome 분석 DB기술이 전세계 학계를 선도해 나갈 것으로 기대됨

○ 현 수준에서의 분석 파이프라인을 고도화하면 Microbiome, shotgun 메타지놈 분석 프로그램 수준이 선진국 대비 100% 다다를 것임

○ 새로운 개념의 In-memory Data Grid기반 클라우드 컴퓨팅 기술을 메타지놈 서열 분석에 적극적으로 활용함으로서, 기존의 Hadoop기반 Mapreduce, MPI기반의 클라우드 시장과 경쟁할 수 있을 것으로 기대됨

○ 물질대사 측면에서 서로 다른 환경유전체 샘플을 비교하는 생물정보학적 도구가 부족한 게 현실임. 본 과제를 통해 구축될 물질대사 DB을 이용하여 특정 환경에 specific 또는 enrich된 유전자를 결정이 가능함. 이는 다양한 환경오염 (예: 녹조, 적도, 수질 오염) 등에 대한 유전자수준에서의 지표를 제공해 줄 것으로 기대함

제 5-2절: 경제, 산업적 측면

○ 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

○ 메타지놈 정보로부터 기존 생촉매보다 효율이 월등히 높거나 새로운 기능을 가지는 생촉매 개발이 가능하며, 이는 장기적으로 석유 기반 화학산업을 바이오매스 및 생촉매 기반의 Green chemistry로 전환시키는데 기여할 것임

○ 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

○ 대용량 염기서열 데이터에 대한 계통도를 구성하기 위해서 University of California, Berkeley 연구진들에 의해 FastTree 프로그램이 개발되어, 수백만 염기서열을 다룰 수 있게 됨

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

○ Microbiome 데이터 분석의 핵심단계라 할 수 있는 염기서열 클러스터링*을 위해 University of Florida의 Sun 교수 그룹에서 ESPRIT-Tree 프로그램을 개발함

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

제 7 장 참고문헌

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., &Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome biology*, *13*(12), R122.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... &Huttley, G. A. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, *7*(5), 335.

Hwang, K., Oh, J., Kim, T. K., Kim, B. K., Yu, D. S., Hou, B. K., ... &Kim, K. M. (2013). CLUSTOM: a novel method for clustering 16S rRNA next generation sequences by overlap minimization. *PloS one*, *8*(5), e62623.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., ... &McDonald, D. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, *550*(7674), 61.

Oh, J., Kim, B. K., Cho, W. S., Hong, S. G., &Kim, K. M. (2012). PyroTrimmer: a software with GUI for pre-processing 454 amplicon sequences. *Journal of Microbiology*, *50*(5), 766-769.

Oh, J., Choi, C. H., Park, M. K., Kim, B. K., Hwang, K., Lee, S. H., ... &Kim, K. M. (2016). Clustom-cloud: In-memory data grid-based software for clustering 16s rrna sequence data in the cloud environment. *PloS one*, *11*(3), e0151064

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B.,

... &Sahl, J. W. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, *75*(23), 7537-7541.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., ... &Miller, A. N. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, *109*(16), 6241-6246.

주 의

1. 이 보고서는 극지연구소에서 수행한 기본연구사업의 연구결과보고서입니다.

2. 이 보고서 내용을 발표할 때에는 반드시 극지연구소에서 수행한 기본연구사업의 연구결과임을 밝혀야 합니다.

3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 안 됩니다.