최종보고서(제출용) 작성용 표지 서식

(뒷면) (옆면) (앞면)

3 cm

발간등록 번호

4㎝

포스트게놈 다부처유전체사업 R&D Report

(견 고 딕 25p)

NGS를 활용한 미생물 유전체 및 전사체 분석 소프트웨어 및 시스템 개발 최 종 보 고 서

(견고딕 14p)

2018

(견고딕13p)

5㎝

농림축산식품부

(견고딕 17p)

3 cm

발 간 등 록 번 호
11-1543000-002347-01

(견고딕31p) 5㎝

NGS를 활용한 미생물 유전체 및 전사체 분석 소프트웨어 및 시스템 개발 최종보고서

(0.1㎝)

2018. 11. 06.

0.15㎝ (견고딕15p)

(별색바탕 : C50, M20, Y59, K0)

주관연구기관 / ㈜천랩

협동연구기관 / 테라젠이텍스 바이오연구소 2㎝

(견고딕 15.5p)

(백색바탕)

농 림 축 산 식 품 부

(견고딕 20p)

<제출문>

제 출 문

농림축산식품부 장관 귀하

본 보고서를 "NGS를 활용한 미생물 유전체 및 전사체 분석 소프트웨어 및 시스템 개발"(개발기간 : 2014. 08. 23 ~ 2018. 08. 22)과제의 최종보고서로 제출합니다.

2018. 11. 06.

주관연구기관명 : ㈜천랩 (대표자) 천 종 식 (인)
협동연구기관명 : 테라젠이텍스 (대표자) 고 진 업 (인)
참여기관명 : (대표자) (인)

주관연구책임자 : 천 종 식
협동연구책임자 : 홍 창 표

국가연구개발사업의 관리 등에 관한 규정 제18조에 따라 보고서 열람에 동의합니다.

<보고서 요약서>

보고서 요약서

<table>
<tr><td>과제고유번호</td><td>914008-04</td><td>해 당 단 계
연 구 기 간</td><td>2014.8.23
~
2018.8.22</td><td>단 계 구 분</td><td>1단계/1단계</td></tr>
<tr><td rowspan="2">연 구 사 업 명</td><td>단 위 사 업</td><td colspan="4">농식품기술개발사업</td></tr>
<tr><td>사 업 명</td><td colspan="4">포스트게놈 다부처유전체사업</td></tr>
<tr><td rowspan="2">연 구 과 제 명</td><td>대 과 제 명</td><td colspan="4">NGS를 활용한 미생물 유전체 및 전사체 분석 소프트웨어 및 시스템 개발</td></tr>
<tr><td>세부 과제명</td><td colspan="4">NGS를 활용한 미생물 유전체 및 전사체 분석 소프트웨어 및 시스템 개발</td></tr>
<tr><td rowspan="2">연 구 책 임 자</td><td rowspan="2">천종식</td><td>해당단계
참여연구원
수</td><td>총: 34명
내부: 34명
외부: 명</td><td>해당단계
연구개발비</td><td>정부: 800,000천원
민간: 266,672천원
계: 1,066,672천원</td></tr>
<tr><td>총 연구기간
참여연구원
수</td><td>총: 34명
내부: 34명
외부: 명</td><td>총 연구개발비</td><td>정부: 800,000천원
민간: 266,672천원
계: 1,066,672천원</td></tr>
<tr><td>연구기관명 및
소 속 부 서 명</td><td colspan="3">㈜천랩 생물정보연구소
테라젠이텍스 연구기획부</td><td colspan="2">참여기업명
1. ㈜천랩
2. 테라젠이텍스</td></tr>
<tr><td>국제공동연구</td><td colspan="3">상대국명:</td><td colspan="2">상대국 연구기관명:</td></tr>
<tr><td>위 탁 연 구</td><td colspan="3">연구기관명:</td><td colspan="2">연구책임자:</td></tr>
</table>

※ 국내외의 기술개발 현황은 연구개발계획서에 기재한 내용으로 갈음

연구개발성과의 보안등급 및 사유	보안등급 일반

9대 성과 등록·기탁번호

구분	논문	특허	보고서 원문	연구시설·장비	기술요약 정보	소프트웨어	화합물	생명자원		신품종	
								생명정보	생물자원	정보	실물
등록·기탁번호	2					1		56			

국가과학기술종합정보시스템에 등록한 연구시설·장비 현황

구입기관	연구시설·장비명	규격 (모델명)	수량	구입연월일	구입가격 (천원)	구입처 (전화)	비고 (설치장소)	NTIS 등록번호

연구개발성과요약

보고서 면수

○ 본 연구과제를 통해 **NGS를 활용한 미생물 유전체 통합 분석 system 개발 및 데이터베이스를 구축**하여 농식품 미생물 유전체 연구 활성화에 기여함.

- 제 1세부에서는 미생물중 원핵생물 (prokaryote)의 유전체 분석 및 전사체 분석을 위해 생물정보학적 기술을 활용해서 통합분석시스템과 데이터베이스를 구축.
- 제 1협동에서는 미생물 중 진핵생물 (eukaryote)에 해당하는 진균류의 유전 정보 분석을 위한 참조 유전체 조립 파이프라인 및 유전체 발굴 시스템 개발, 진균류 유전자 기능 연구 및 유용 유전자 발굴을 위한 통합 분석 시스템을 개발.

○ **세균 genome의 NGS 분석 파이프라인 개발**

- NGS data로부터 assemble, gene prediction, annotation, analysis, genome comparison 하는 분석 파이프라인을 구축 및 업데이트함.
- Visualization 모듈 개발 및 웹상에서 구현
- KEGG database와 연동하여 유전자 정보를 살펴 볼 수 있도록 구현.

○ **웹 방식을 통한 신규 비교유전체 통합분석시스템 구축**

(http://agri.ezbiocloud.net)

- 비교유전체 셋트 구축 및 업데이트
(44,048 genome/ 1, 945 Pan-genome set)
- 주요미생물 비교유전체 set 구축: 동식물 병원균, 유산균 및 식품위해균 등

○ **Genome database 구축 및 업데이트**

- 표준균주 (type strain) 유전체 데이터 생산 및 DB 구축
- EzBioCloud를 이용하여 주요 균주의 type strain에 대해서 genome sequencing 분석 수행.
- 농축산업에 유용한 유산균의 genome reference DB 구축을 위해 유산균 중에서 표준균주의 유전체 분석이 안 된 균주를 분석함.

○ **세균 전사체 분석 파이프라인 고도화 및 분석 모듈 개발 업데이트**

- 여러 정규화 방법을 이용한 전사체 발현량 제시
- 라이브러리 크기, 유전자 길이 등 발현량에 영향을 줄 수 있는 요인들을 고려하여 정규화 방법(normallization)을 도입함.
- 이를 위해 일반적으로 사용한 RPKM 이외에 RLE (Relative Log Expression), TMM (Trimmed Mean of M-value) 방법을 RNA-Seq 분석 결과를 제시하는 천랩이 본 과제를 통해서 개발한 CLRNASeq software에 적용.

○ **진균 참조 유전체 조립 및 유전자 예측 시스템 개발:** Long- 및 short-reads을 활용한 하이브리드 방법 기반의 참조유전체 조립 파이프라인 개발하여 국내 누룩 유전체 3종 서열 조립 지원함. 유전자 구조 예측을 위해 evidence-based prediction 시스템을 개발했고, 관련 유전자 정확도 예측 평가가 확보됨. 또한 taxonomic profiling 및 상동성 검색 기반의 유전자 예측 웹서버인 TaF를 개발함

○ **진균 전사체 분석을 위한 파이프라인 개발:** 턱시도 프로토콜 방식의 진균류 전사체 분석을 위한 파이프라인을 개발하였고, 시계열 전사체 데이터 및 KEGG core DB 적용 기능을 추가함. 이를 토대로 표고버섯 갈변화 관련 유전자 후보군 분석에 활용됨

○ **진균류 참조 유전체 정보 전체 정보 활용을 위한 데이터베이스 및 웹사이트 개발:** 10종 진균류 유전체 포함한 진균 유전체 데이터베이스 구축하였고, 가계도-기반의 형질 관련 유전변이 탐색 파이프라인 개발함. 또한 진균류 분석을 위한 통합 분석을 위한 웹사이트 구축함

○ **프로모터 및 전사인자 분석 파이프라인 개발:** 전사인자 및 히스톤 변형 분석을 위한 ChIP-Seq 분석 파이프라인 개발하였고, 효모에서 향 및 대사관련 전사인자들을 연구중에 있음. 또한 파이프라인에 효모 유전체 정보 기반의 TF 모티프 분석 모듈 개발 추가함 (orthologous gene cluster 분석 모듈 결합됨)

○ **미생물유전체 전략연구사업단 내 타과제 연구지원**
- 구축된 유전체 분석 및 전사체 분석 시스템을 활용하여 미생물유전체사업단 내의 타분야 과제의 연구 분석을 지원함.

<요약문>

연구의 목적 및 내용	**<연구의 목적>** ○ NGS를 활용한 미생물 유전체 통합 분석 system 개발 및 데이터베이스 구축 ○ 유전체 연구 경쟁력 제고 및 목적 지향적 바이오산업에 활용 **<연구의 내용>** ○ 1세부에서는 미생물중 원핵생물 (prokaryote)의 유전체 분석 및 전사체 분석을 위해 생물정보학적 기술을 활용해서 통합분석시스템과 데이터베이스를 구축함. ○ 1협동에서는 미생물 중 진핵생물 (eukaryote)에 해당하는 진균류의 유전 정보 분석을 위한 참조 유전체 조립 파이프라인 및 유전체 발굴 시스템 개발, 진균류 유전자 기능 연구 및 유용 유전자 발굴을 위한 통합 분석 시스템을 개발함.
연구개발성과	○ **세균 genome의 NGS 분석 파이프라인 개발** - NGS data로부터 assemble, gene prediction, annotation, analysis, genome comparison 하는 분석 파이프라인을 구축 및 업데이트함. - Visualization 모듈 개발 및 웹 상에서 구현 - KEGG database와 연동하여 유전자 정보를 살펴 볼 수 있도록 구현. ○ **웹 방식을 통한 신규 비교유전체 통합분석시스템 구축** (http://agri.ezbiocloud.net) - 비교유전체 셋트 구축 및 업데이트 (44,048 genome/ 1, 945 Pan-genome set) - 주요미생물 비교유전체 set 구축: 동식물 병원균, 유산균 및 식품위해균 등 ○ **Genome database 구축 및 업데이트** - 표준균주 (type strain) 유전체 데이터 생산 및 DB 구축 (유산균 80종 포함, 총 세균 188종에 대한 유전체 데이터생산) - EzBioCloud를 이용하여 주요 균주의 type strain에 대해서 genome sequencing 분석 수행. - 농축산업에 유용한 유산균의 genome reference DB 구축을 위해 유산균 중에서 표준균주의 유전체 분석이 안 된 균주를 분석함. ○ **세균 전사체 분석 파이프라인 고도화 및 분석 모듈 개발 업데이트** - 여러 정규화 방법을 이용한 전사체 발현량 제시 - 라이브러리 크기, 유전자 길이 등 발현량에 영향을 줄 수 있는 요인들을 고려하여 정규화 방법(normallization)을 도입함. - 이를 위해 일반적으로 사용한 RPKM 이외에 RLE (Relative Log Expression), TMM (Trimmed Mean of M-value) 방법을 RNA-Seq 분석 결과를 제시하는 천랩이 본 과제를 통해서 개발한 CLRNASeq software 에 적용.

	○ **진균 참조 유전체 조립 및 유전자 예측 시스템 개발:** Long- 및 short-reads을 활용한 하이브리드 방법 기반의 참조유전체 조립 파이프라인 개발하여 국내 누룩 유전체 3종 서열 조립 지원함. 유전자 구조 예측을 위해 evidence-based prediction 시스템을 개발했고, 관련 유전자 정확도 예측 평가가 확보됨. 또한 taxonomic profiling 및 상동성 검색 기반의 유전자 예측 웹서버인 TaF를 개발함 ○ **진균 전사체 분석을 위한 파이프라인 개발:** 턱시도 프로토콜 방식의 진균류 전사체 분석을 위한 파이프라인을 개발하였고, 시계열 전사체 데이터 및 KEGG core DB 적용 기능을 추가함. 이를 토대로 표고버섯 갈변화 관련 유전자 후보군 분석에 활용됨 ○ **진균류 참조 유전체 정보 전체 정보 활용을 위한 데이터베이스 및 웹사이트 개발:** 10종 진균류 유전체 포함한 진균 유전체 데이터베이스 구축하였고, 가계도-기반의 형질 관련 유전변이 탐색 파이프라인 개발함. 또한 진균류 분석을 위한 통합 분석을 위한 웹사이트 구축함 ○ **프로모터 및 전사인자 분석 파이프라인 개발:** 전사인자 및 히스톤 변형 분석을 위한 ChIP-Seq 분석 파이프라인 개발하였고, 효모에서 향 및 대사관련 전사인자들을 연구중에 있음. 또한 파이프라인에 효모 유전체 정보 기반의 TF 모티프 분석 모듈 개발 추가함 (orthologous gene cluster 분석 모듈 결합됨) ○ **미생물유전체 전략연구사업단 내 타과제 연구지원** - 구축된 유전체 분석 및 전사체 분석 시스템을 활용하여 미생물유전체사업단 내의 타분야 과제의 연구 분석을 지원함.
연구개발성과의 활용계획 (기대효과)	○ 개발한 미생물 유전체 분석 기술 및 시스템을 활용하여 사업단 및 관련 연구자들의 생물정보학적 분석을 지원함. ○ 개발된 유전체 분석 기술을 활용하여 유용 미생물 유전자원 탐색 및 사업화 계획 ○ 미생물 유전체 교육에 활용 및 지원

국문핵심어 (5개 이내)	유전체	농업미생물	통합시스템	비교유전체	생명정보분석
영문핵심어 (5개 이내)	Genome	Agricultural microbe	Total system	Comparative genomics	Bioinformatics

※ 국문으로 작성(영문 핵심어 제외)

<본문목차>

〈 목 차 〉

1. 연구개발과제의 개요 ····· 7
2. 연구수행 내용 및 결과 ····· 16
3. 목표 달성도 및 관련 분야 기여도 ····· 59
4. 연구결과의 활용 계획 등 ····· 63
붙임. 참고 문헌 ····· 65

<별첨1> 연구개발보고서 초록
<별첨2> 주관연구기관의 자체평가의견서
<별첨3> 연구성과 활용계획서

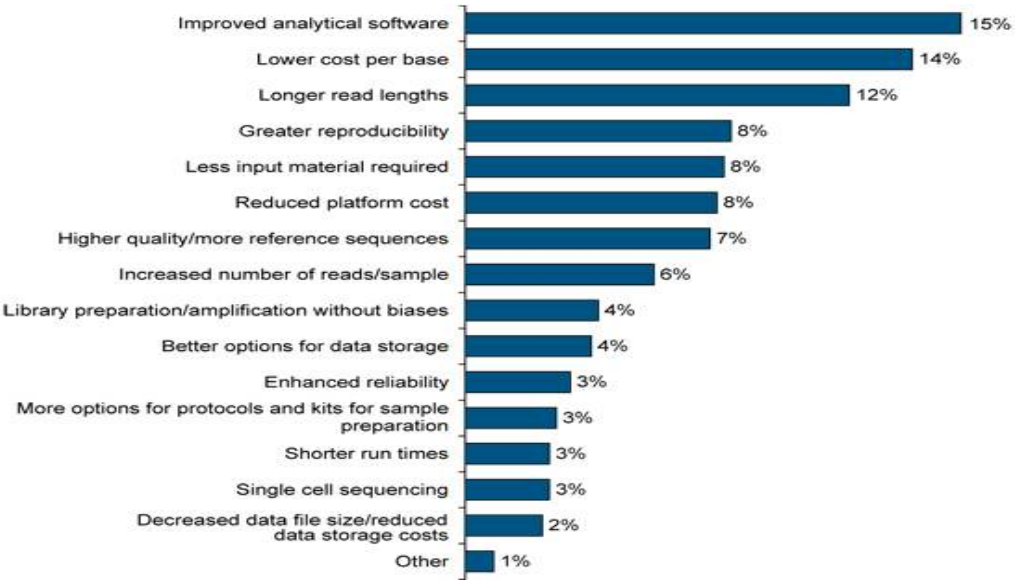
1. 연구개발과제의 개요

1-1. 연구개발 목적

1) 연구 배경

○ 차세대 유전체서열 해독기술(NGS)의 발달과 생명정보기술의 중요성 증대

- NGS (Next-Generation Sequencing) 기술의 보급과 생명정보기술(bioinformatics)의 발달로 유전체 정보에 대한 접근이 과거 어느 때보다 용이해졌음. 짧은 기간 동안 유전체 해독에 소요되는 비용을 획기적으로 감소시켰으며, 누구든지 유전체나 전사체 데이터를 쉽게 생산할 수 있게 되었음.
- 이러한 현상은 NGS라는 장비에 절대적으로 의존하는 서비스 시장이 형성되어 성장하는 긍정적인 효과를 보이기도 하였으나, 서열 정보 생산 이후의 분석과 이를 토대로 한 생명정보학 기술이 중요성이 매우 커졌음. 일선 연구자들이 NGS 데이터 활용 과정에서 데이터의 양이나 질이 아니라 쉽게 활용 가능한 분석용 소프트웨어와 분석 플랫폼임.
- 따라서 유용 미생물 관련 유전체·오믹스 정보의 심층 분석을 위한 생명정보처리 기술 및 빅데이터 처리 기술 개발로 생명정보 서비스 기반을 제공하는 것이 매우 중요함.



[그림1. NGS 데이터 작업에서 가장 시급한 개선 요망 사항]
출처: The Globa Outlook for Next Generation Sequencing: Usage, Platform Drivers & Workflow, Bioinformatics, LLC, (2011)

○ 긴 서열분석 NGS (Long read sequencing) 기술의 발전

- Short read 위주의 NGS가 대중화하면서 완성도가 낮은 초안 수준의 미생물 유전체 성과물이 지나치게 양산되는 문제가 있었으나, long read 기법이 출현하면서 활용성이 매우 높은 nearly finished genome 정보 생산의 새로운 기회가 열리고 있음.
- PacBio (Pacific Biosciences)사의 RS II나 Sequel 장비, Illumina의 TruSeq synthetic long read 등 10 kb를 훨씬 상회하는 길이의 시퀀싱 read가 생산되고 있음

- Oxford Nanopore사의 MinION 장비는 단백질 미소 채널로 DNA가 통과하면서 나타나는 전류의 변화를 측정하여 염기서열 정보로 전환하는 휴대 가능한 초소형 장비로 새로운 NGS 장비기술로 등장함.

○ 전장 유전체의 완전해독 중요성 심화

- 전 세계적으로 막대한 양의 미생물 유전체 정보가 공개되고 있으나, 염색체 수준까지 완성된 유전체(complete genome)의 비율은 획기적으로 증가하고 있지 않음
- 공개된 유전체 정보의 절대 다수를 차지하는 원핵 미생물의 경우 2009년에 이미 1천건을 돌파함. GOLD (Genomes Online Database)의 집계에 따르면 2014년 기준, 초안(draft) 수준으로 확정된 유전체 정보는 15,653건인데 비하여 완성본 유전체는 2,031건에 불과함
- NGS 기술에 의하여 유전체 초안을 만드는데 소요되는 시간과 비용은 Sanger chemistry에 의존하던 과거에 비하여 1천분의 1 수준으로 줄어들었으나, finishing 과정을 단축할 수 있는 획기적인 기술혁신은 여전히 필요한 실정임.
- 유전체 조립을 어렵게 하는 반복 서열 단위보다 짧은 read length의 서열 단편을 만들어내는 NGS 기술로는 근본적인 한계가 있음.

○ 표준 및 참조 유전체 조립 기술의 발전

- 참조 유전체 정보는 생물의 유전적 특성을 밝히는데 있어서 가장 핵심적인 기술임. 유전체 해독 기술은 과거 오랜시간과 비용이 드는 일이었지만, 차세대 서열해독 기술(NGS)의 등장 이후 저렴한 비용으로 단시간 내에 선도게놈 서열을 얻을 수 있는 조립기술이 등장하였으며, 최근 몇 년간 선도게놈 조립에서 광범위하게 사용되어 그 유용성과 기술의 완결성을 증명함.
- 다양한 선도게놈 조립기술 중 가장 대표적인 것으로 Illumina HiSeq 장비를 이용해서 생산된 short read에 기반한 contig assembly 기술과 long-mate read 를 이용한 scaffolding 기술을 이용한 유전체 조립 기술이 활용되고 있음. Short read 를 이용한 조립 기술은 다양한 종에 대한 금증이 완료 되었지만, 기술적인 한계로 긴 길이의 유전체 서열을 만드는 것이 어려움. 이를 극복하기 위해서 Long-read 를 생산하는 NGS 기술인 PacBio 혹은 Moleculo 기술을 이용함. 앞의 short read 와 long read를 이용한 Hybrid assembly 방법은 보다 품질 좋은 scaffold 를 생산할 수 있는 것으로 알려져 있음.
- NGS 기술에 기반한 조립 기술로 scaffold를 생산한 이후 조립 기술의 오류 수정이나 실제 chromosome에 가까운 유전체 완성을 위하여 Physical mapping 기술인 Bionano Irys 등의 기술을 이용함. 이 기술은 DNA에 특정위치에 형광 물질을 달아서 그 패턴을 실시간으로 인식하고, 이를 조립된 서열과 비교하여 조립 오류를 수정하하여, 실제 유전체에 가까운 super scaffold 를 만들어준다고 알려져 있음.

○ 유전자 발굴을 위한 예측 시스템 연구 중요성 증가

- 조립된 유전체에서 유용한 유전자 정보를 얻기 위한 사용하는 시스템을 유전자 예측 시스템으로 정의함. 이 시스템은 유전자의 특성을 이용하여 유전자의 위치를 예측하는 *ab initio* 방법과 근접종의 유전자 정보를 이용하여 유전자를 찾는 homology 기반의 방법으로 나눠지며, 정확한 유전자 예측을 위하여 두 가지 방법이 모두 사용되고 있음

- RNA-seq 분석 기술의 발달이후 RNA-seq 데이터를 이용한 유전자 예측이 활발하게 이루어지고 있으며, *ab inito* 방법과 homology 방법의 결과물과 합쳐서 더 정확한 결과를 보여주는 것으로 알려져 있음

- 예측된 유전자의 기능을 알기 위해서 유전자 기능 분석 파이프라인을 구성한다. 일반적으로 이런 시스템은 크게 DNA기반과 단백질 기반으로 나눠지며, 이 분석 결과로 이후의 다양한 분석이 진행되고 있음

○ 전사체 분석 기술의 중요성 증대

- 참조 유전체 기반의 유전자 발현량 측정 방법은 NGS기술의 발달과 함께 최근 각광 받고 있으며, 기존의 방법인 Microarray 분석 기법과 비교하여 정밀하며, 사전 유전자 정보가 없어도 발현량 분석이 가능하다는 장점 때문에 전사체 분석을 위해서 사용되고 있음
- 진균류의 유전자 발현 정보와 환경 변화에 따른 유전자 발현 특성을 알기 위해서 전사체 분석 파이프라인을 개발함. 전사체 분석 파이프라인은 참조 유전체 정보를 이용하여 전사체 분석을 진행하며, 진균류의 특성을 반영하여 유전자 발현량 예측을 최적화함

2) 미생물 유전체분석 기술의 국내외 현황

○ 국내 미생물 유전체 해독 및 활용연구 수준

- 전문적인 장비를 갖춘 유전체 정보의 생산 거점은 출연연과 기업을 중심으로 일정 수준의 국제적인 경쟁력을 갖추었고, 정부 과제의 지원에 힘입어 대학과 출연연을 중심으로 한 미생물 유전체 해독 및 활용 연구가 지속됨.
- 미생물 유전체 정보 분석을 위한 전산 알고리즘 개발, 많은 비용이 소요되는 오믹스 데이터베이스 구축 및 분석 소프트웨어 개발 등은 선진국의 기술력에 비해 약한 편이며, 이러한 분석을 위한 전산 인프라도 외국에 비해서 비교가 되지 않음.

○ 국내에서 생산하는 유전체 정보는 아직 국외의 증가속도를 따라가지 못하고 있어, 여전히 99% 이상의 정보를 외국에 의존해야하는 심각한 상황임.

- 마크로젠, 테라젠이텍스, 디엔에이링크 등 민간 유전정보 회사들이 있지만 국내 시장규모는 아직 1000억 원 대에 불과하며, 특히 유전체 분석의 핵심장비인 DNA 염기서열 분석기(Sequencer)의 경우 미국 Illumina社 등 외산에 전적으로 의존하고 있음.
- 마크로젠은 2014년 세계김치연구소와 김치 미생물 유전체 관련 연구개발 및 정보교류 등을 위한 업무협약을 체결하여 김치에 존재하는 다양한 미생물의 유전체를 연구하고 김치산업 발전에 활용하고자 함.
- 천랩에서는 메타지노믹스를 활용하여 미생물의 유전정보를 이용하여 새로운 항생제 및 신약 후보물질을 찾는 연구를 수행하고 있으며 미생물 신종을 발굴하고 다양한 미생물 DNA 데이터베이스를 구축하였음(EzBioCloud, EzGenome 등)

1-2. 연구개발의 필요성

○ 방대한 유전체, 전사체 데이터생산

- 2000년대 중반 이후의 NGS 기술의 급격한 발전으로 인한 유전체/전사체 데이터 생산의 가격 하락과 유전체, 전사체 시장의 급격한 팽창으로 많은 데이터가 생산 되고 있으나, 이를 분석할 수 있는 생물정보 tool 과 시스템은 이에 못 미치는 실정임.
- 단백체와 대사체를 유전체/전사체에 통합 분석하는 생물정보 분석 시스템은 매우 제한적임.

○ 생물 정보학 전공자가 아닌 일반 연구자가 사용할 수 있는 분석 툴 개발의 필요성 증가

- NGS를 통한 유전체, 전사체 연구를 하기 위해서는 생물정보학 지식이 필요하고 이에 따른 대규모 전산 인프라도 필요함.
- 생물정보 비전공자는 분석 인프라와 소프트웨어 활용성 때문에 연구에 어려움이 있기에 비전공자도 쉽게 사용할 수 있는 툴과 데이터베이스를 포함한 시스템 개발이 필요함.

○ 생물정보 데이터를 비교 분석하는 툴과 가공된 2차 데이터베이스의 필요성 증가

- 방대한 유전체 데이터와 전사체 데이터를 서로 비교 분석해야 좋은 결과를 생산 할 수 있음.
- 개별 sample을 비교해서 분석하기 위한 tool을 개발하고 각 데이터를 가공된 2차 데이터베이스화 하여 연구자들이 쉽게 비교 분석할 수 있게 개발할 필요성이 있음. 또한 유전체, 전사체, 군집 분석의 데이터를 연계해서 분석할 수 있는 시스템 개발이 필요함.

○ 빅데이터 처리 기술을 이용한 미생물의 표준 유전체 조립 기술 및 유용 유전자 발굴 기술 개발 필요성 증가

- 참조 유전체 정보는 생물의 유전적 특성을 밝히는데 있어서 가장 핵심적인 기술임
- 원핵미생물(prokaryotes)의 유전체 분석기술과 진핵미생물 (eukaryotes)의 유전체 분석 기법 및 전사체 분석은 차이가 있음.
- 또한 기존에 개발된 large size의 eukaryotes의 참조 유전체 기술과는 다르게 진균류의 유전체 조립 기술 및 분석 기술은 다르게 적용할 필요가 있음. 진균류에 적합한 유전체 조립기술 및 분석기술과 전사체 분석 시스템 개발이 시급함.

○ 진균의 체계적인 유전체 정보 분석과 유용유전자 발굴 및 유전적 특성 연구를 위한 기반 시스템이 필요함

- 진균에 최적화된 유전체 조립 파이프라인과 유전자 예측 시스템이 필요
- 진균의 유전자 발현연구를 위해 진균에 맞춰진 전사체 분석 파이프라인이 필요
- 다양한 정보의 확인 및 이용을 진균 유전체 정보를 통합한 통합 데이터베이스 필요

1-3. 연구개발 범위

1) 최종 목표

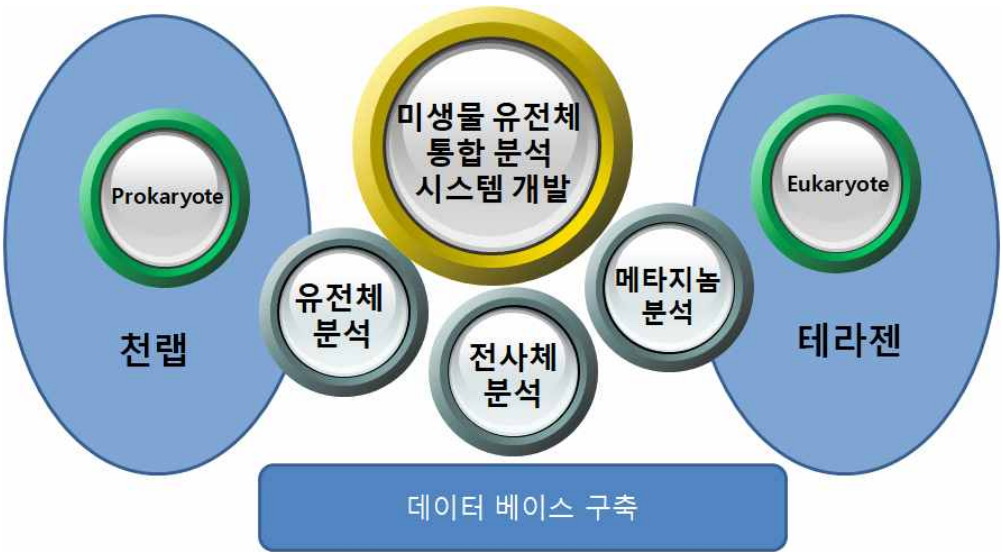
○ NGS를 활용하여 미생물 유전체 통합 분석 system 개발 및 데이터베이스를 구축하여 유전체 연구 경쟁력 제고 및 목적 지향적 바이오산업에 응용 가능하도록 개발함.

○ 제 1세부과제에서는 미생물중 원핵생물 (prokaryote) 의 유전체 분석, metagenome 분석, transcriptome 분석을 연구하며, 제1협동에서는 미생물중 진핵생물 (eukaryote)에 해당하는 진균류의 유전 정보 분석을 위한 참조 유전체 조립 파이프라인 및 유전체 발굴 시스템 개발, 진균류 유전자 기능 연구 및 유용 유전자 발굴을 위한 통합 분석 시스템을 개발하는 것이 목표임.

2) 세부 목표

○ 원핵 미생물의 Genomics, Metagenomics를 위한 분석 플랫폼 및 분석 소프트웨어의 개발 및 고도화
○ 세균의 비교 유전체 분석 모듈 및 Pathway 비교 분석 모듈 개발
○ 세균의 유전체, 전사체의 2차 데이터베이스 구축
○ 진균류 분석을 위한 유전체 조립 및 전사체 분석 파이프라인 개발
○ 진균류 유전체 정보 활용을 위한 기반 시스템 개발 및 기능분석을 위한 파이프라인 개발
○ 진균류 전사체 데이터 활용을 위한 시스템 고도화 및 프로모터 분석 파이프라인 개발
○ 진균류 유전체 분석을 위한 통합 분석 시스템 개발 및 고급 분석 파이프라인 개발
○ 미생물 유전체/전사체 연구 지원 분석

3) 연구개발 추진 모식도



1-4. 연차별 연구개발의 목표 및 내용

구분	연도	연구개발의 목표	연구개발의 내용
1차 년도	2014	1세부: 분석 플랫폼 및 분석 소프트웨어 개발 및 업데이트	**1) 미생물 유전체 분석 파이프라인 개발** 가) short-read 및 hybrid NGS data NGS 분석 파이프라인 개발 및 업데이트 나) 유전자 예측 모델 및 annotation 분석 파이프라인 개발 및 업데이트 :COG, SEED, Ref Seq database 등을 localization 하여 유전자의 annotation 시에 사용할 예정임. 다) 비교 유전체 분석 파이프라인 및 분석 모듈 개발 : 기존에 개발된 비교 유전체 분석 모듈을 고도화 하고, 비교유전체 데이터베이스를 구축하여 비교 유전체 분석을 보다 빠르고 정확하게 구현할 예정임 **2) Metagenomics 분석 파이프라인 개발** 가) NGS를 활용한 metagenomics 분석 파이프라인 개발 나) 다양한 NGS machine 에서 나오는 sequencing data를 처리할 수 있도록 개발 다) 유전자 기능 예측 및 annotation 분석 파이프라인 개발 : **3) 유전체 분석 소프트웨어 개발 및 업데이트** 가) 기존에 개발된 CLgenomics 프로그램을 바탕으로 다양한 분석 기능이 포함된 유전체 데이터를 분석할 수 있는 소프트웨어를 업데이트함. 나) 비교 유전체 분석 모듈을 고도화 하고, 데이터베이스화 할 예정임. 유전자간들의 similarity 값을 계산하여, database화 시켜서, 비교유전체 분석시 속도를 향상시킬 예정임
		1협동: 진균류 분석을 위한 유전체 조립 및 전사체 분석 파이프라인 개발	**1) 진균류 참조 유전체 조립 파이프라인 개발** 가) long-read (Molecule, PacBio 등)를 이용한 참조 유전체 조립 파이프라인 개발 및 shot-read 기반의 분석 파이프라인과 결합 나) Physical Mapping 기술 (Bionano Irys 등)을 이용한 다) Super scaffold 조립 기술 개발 **2) 진균류 유전체 발굴을 위한 유전자 예측 및 기능 예측 파이프라인 개발** 가) 효모 균주를 위한 유전자 예측 모델 개발 나) 유전자 발굴을 위한 유전자 예측 파이프라인 개발 다) 유전자 기능 예측 파이프라인 개발 **3) 진균류 유전체정보 기반의 전사체 분석 파이프라인 개발** 가) 참조유전체 데이터에 기반한 진균류에 특화된 전사체분석 파이프라인 구축 **4) 진균류 유전체 및 전사체 분석을 위한 데이터 생산 및 유전체 발굴** 가) 사업단내 타과제와 연계를 통한 산업적 유용 종 선택 나) 진균류 유전체 및 전사체 해독

구분	연도	연구개발의 목표	연구개발의 내용
2차 년도	2015	1세부: Pathway analysis, 비교유전체 분석 등 생물 정보 분석 모듈 개발 및 Database 구축	**1) Pathway 분석 모듈 개발 및 pathway 비교 분석 모듈 개발** 가) 유전체 데이터를 바탕으로 KEGG database를 활용하여 Pathway 분석을 수행하는 분석 모듈을 개발 나) 두 개의 sample에서 pathway를 비교 분석하는 모듈 개발 다) 다양한 분석 알고리즘 및 통계 분석을 활용하여 사용자가 쉽게 pathway를 비교할 수 있도록 개발 **2) RNA seq data 분석 모듈 및 pathway 분석 모듈 개발** 가) 미생물의 전사체 분석 파이프라인 업데이트 및 다양한 분석 모듈 개발 나) Pathway 분석 모듈을 추가 업데이트 하여 유전체 pathway 데이터와 비교하여 분석 하는 모듈 개발 및 업데이트 **3) 비교 유전체 분석 모듈 업데이트** 가) 비교 유전체 분석 모듈을 고도화 하여, 기존에 분석 된 다양한 유전체 데이터와 모든 유전자들 간의 homology 분석을 수행할 있도록 고도화 나) 비교 유전체 데이터베이스를 구축 **4) 유전체 데이터베이스 구축** 가) 농업에 중요한 미생물의 유전체 데이터베이스를 구축함. 나) 본연구팀이 개발한 소프트웨어에서 활용 가능하도록 구축된 분석 파이프라인에서 나온 결과물로 2차 데이터베이스를 구축함.
		1협동: 진균류 유전체 정보 활용을 위한 기반 시스템 개발 및 기능 분석을 위한 파이프라인 개발	**1) 진균류 기능 분석을 위한 참조데이터베이스 구축** 가) 10종 이상의 진균류 참조 데이터베이스 구축 **2) 참조 유전체 조립 및 유전자 예측 파이프라인 고도화** 가) 유전체 조립 품질 개선 및 최적화 나) 유전자 예측 품질 향상을 위한 시스템 고도화 **3) 참조 유전체 정보 활용을 위한 데이터베이스 및 웹사이트 개발** 가) 참조 유전체 정보 활용을 위한 데이터베이스 개발 나) 참조 유전체 정보의 효과적인 활용을 위한 웹페이지 개발 **4) 시계열 전사체 데이터 분석을 위한 파이프라인 개발** 가) 시계열 데이터 분석을 위한 파이프라인 개발 **5) 유전체 기능 분석을 위한 연계 분석 파이프라인 개발** 가) 전사체와 LC-MS 데이터 연계분석 파이프라인 개발 **6) 진균류 유전체 및 전사체 분석을 위한 데이터 생산 및 발굴** 가) 시계열 전사체 데이터 생산

구분	연도	연구개발의 목표	연구개발의 내용
3차 년도	2016	1 세부: Database 업데이트, 분석 소프트웨어 및 생물 정보 분석 모듈 고도화	**1) 농업에 중요한 미생물 및 식품 관련 미생물, 병원성 미생물의 유전체 데이터베이스를 구축** 가) comparative genomics 및 metagenome의 분류학적 군집 비교 및 기능 비교 등이 포함된 2차적인 데이터베이스 나) 본 연구팀이 개발한 CLgenomics 프로그램에서 분석할 수 있도록 2차 데이터베이스를 구축 다) 비교 유전체 분석을 통하여 모든 유전자들의 비교 유전체 분석 결과를 DB화 할 예정이고, 이를 통하여 비교 유전체 결과를 빠르게 분석하고, 다양한 비교 유전체 조합에 따라 분석 프로그램에서 쉽게 살펴볼 수 있도록 개발 라) RNA seq data와 연계하여 통합하여 분석하는 모듈도 개발 마) 미생물과 관련하여, 분리 지역, 날짜, 병원성 여부, 감염 경로, molecular type 등의 정보를 포함하는 메타 데이터 베이스를 구축할 예정임 바) 계통 분석과 epidemiology 분석에 활용하여 미생물의 진화, phylogeny 분석, 전파 경로 등에 활용할 예정임 **2) 식품 관련 및 병원성 미생물 종에 대한 생물정보학적 분석 고도화** 가) RBH (Reciprocal Blast Hit) 방법을 이용하여 비교 유전체 분석을 통한 각 유전체의 homologous gene group을 설정하고, core genome과 pan genome 분석을 수행할 예정임 나) pathway 분석과 연계하여 각 미생물의 특이적인 metabolism을 phenotype 및 식품 미생물의 역할과 연계하여 분석할 예정임 다) 식품 관련 미생물 종에 대한 이해 증진 및 모니터링 기술 등에 활용함.
		1 협동: 진균류 전사체 데이터 활용을 위한 시스템 고도화 및 프로모터 분석 파이프라인 개발	**1) 진균류 기능 분석을 위한 참조데이터베이스 추가 구축** 가) 10종 이상의 진균류 참조 데이터베이스 추가 구축 **2) 진균류 참조 유전체 정보 활용을 위한 데이터베이스 및 웹사이트 고도화** 가) 유전자 발현데이터 추가 나) 유전자 발현데이터 추가를 위한 사용자 환경 개발 다) 시계열 데이터를 위한 사용자 환경 개발 **3) 프로모터 및 전사 인자 분석 파이프라인 개발** 가) 전사인자 결합 부위 데이터베이스 구축 나) 프로모터 분석을 위한 전사인자 결합부위 (Transcription 다) Factor Binding Site) 예측 파이프라인 개발 **4) 유전체 기능 분석을 위한 연계 분석 추가 개발 및 파이프라인 고도화** 가) RNA-seq 분석과 전사인자 분석 결과 연계 파이프라인 개발 **5) 진균류 유전체 및 전사체 분석을 위한 데이터 생산 및 발굴** 가) 유전체 해독 데이터 생산 나) 전사체 해독 데이터 생산

구분	연도	연구개발의 목표	연구개발의 내용
4차 년도	2017	1세부: 미생물 genomics, metagenomics, transcriptomics 통합 분석 시스템 구축	**1) 미생물 genomics, metagenomics, transcriptomics 통합 분석 시스템 구축** 가) 동정부터 genome분석, phylogeny 분석, 비교 유전체 분석, pathway 분석 등을 수행할 수 있는 NGS data 통합 분석 system 구축함. 나) 기존에 구축한 분석 시스템 및 database를 고도화 하여 업데이트 다) 비교 유전체 분석 모듈 고도화 및 비교 유전체 databae 구축 및 업데이트 라) 식품 및 기능성 미생물에 활용 가능하도록 업데이트 마) Pathway 분석, pathway 비교 분석 등의 모듈 업데이트 바) RNA seq data도 genome data와 연계하여 분석 할 수 있도록 개발 **2) 통합 분석 시스템 및 소프트웨어를 통합 상업화** 가) 기존의 상업화를 확대 하여, 본 과제에서 개발된 통합 분석 시스템을 도입함 나) user workshop을 8회이상 개최하여 사용자 교육 및 사용자 확보함 다) 해외 학회 참가 및 해외 홍보활동을 통하여 해외 상업화를 촉진시킴
		1협동: 진균류 유전체 분석을 위한 통합 분석 시스템 개발 및 고급 분석 파이프라인 개발	**1) 진균류 참조 유전체 정보 활용을 위한 데이터베이스 및 웹사이트 고도화** 가) 프로모터 분석 결과 시각화 모듈 개발 나) 네트워크 분석 결과 시각화 모듈 개발 **2) 유전자 상호작용 네트워크 분석 파이프라인 개발** 가) 특이 발현 유전자 연구를 위한 상호작용 네트워크 분석 파이프라인 개발 나) 네트워크 분석 파이프라인 시각화 사용자 환경 개발. **3) 진균류 유전 정보 분석을 위한 통합 분석 시스템 개발** 가) 유전체, 전사체, 프로모터, 네트워크 분석 파이프라인 통합 나) 연계 분석 모듈 통합 다) 비교 유전체 분석 자동화 시스템 개발 라) 통합 분석 데이터베이스 구축

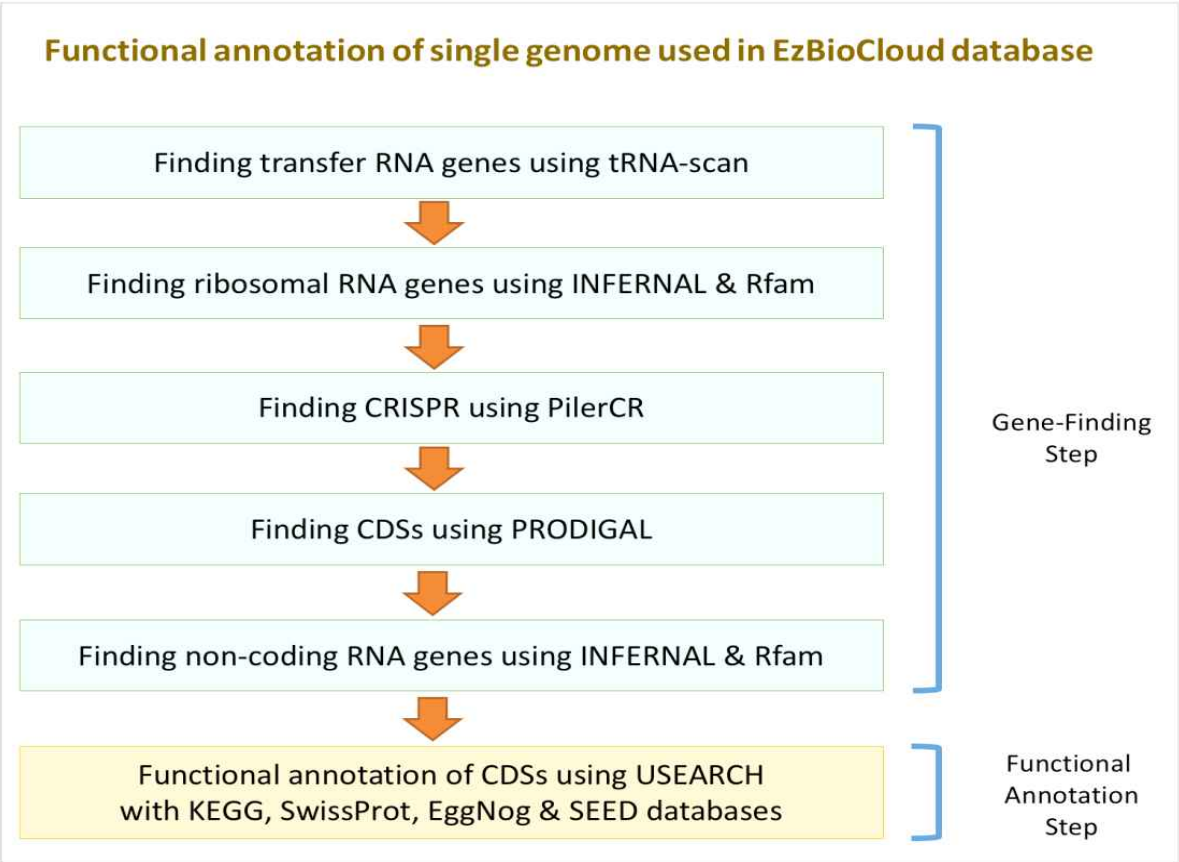
2. 연구수행 내용 및 결과

2-1 제 1세부 연구수행 내역 : ㈜ 천랩

(1) 미생물 유전체 분석 파이프라인 고도화

○ 미생물 genome의 NGS 분석 파이프라인 개발

- NGS data로부터 assemble, gene prediction, annotation, analysis, genome comparison 하는 분석 파이프라인을 구축 및 업데이트함. Glimmer, GeneMark, Prodigal, Metagene 등을 이용하여 유전자를 detection 하고, tRNA-sca-SE 으로 tRNA prediction, RNAmmer를 이용하여 rRNA을 prediction 하고, Rfam을 이용하여 다른 non-coding gene을 prediction 하여 Database에 parsing 한 후, Refseq, COG, SEED, eggNog , KEGG database를 localization 한 DB에 blast 하여 annotation 한 data를 합치는 annotation 분석 파이프라인을 구축하고 각 구성 요소를 모듈화 시키고 업데이트함. 추후 각 모듈이 업데이트 될 때마다 적용할 수 있도록 구현함.



[그림2. EzBioCloud에서 유전체 annotation 파이프라인 절차]

1) tRNA 유전자 찾기

- 프로그램: tRNA-scan 버전 1.3.1
- 실행 파라메터: tRNA-scan-SE -bact [Fasta File]

2) rRNA 유전자 찾기

- 프로그램: INFERNAL 버전 1.0.2 (cmsearch)
- 데이터베이스: rfam 12.0
- 실행 파라메터: -E 1.0E-5 -Z 700 –noali rfam12.0/rRNA_bact.cm [Fasta File]

3) CRISPR 찾기

- 프로그램: PilerCR 버전 1.06
- 실행 파라메터: pilercr -in [Fasta File] -out [Output File]
- 프로그램: CRT 버전 1.2
- 실행 파라메터: java -cp CRT1.2-CLI.jar crt [Input Fasta File]

4) ncRNA 찾기

- 프로그램:INFERNAL 버전 1.0.2 (cmsearch)
- 데이터베이스: Rfam 12.0
- 실행 파라메터: cmsearch -E 1.0E-5 -Z 700 –noali rfam12.0/RNase_bact.cm [Fasta File]
- 실행 파라메터: -E 1.0E-5 -Z 700 –noali rfam12.0/Gene_bact.cm [Fasta File]

5) CDS 찾기

- 프로그램: PRODIGAL 버전 2.6.2
- 실행 파라메터: -i [Input Fasta File] -o [Output GFF File] -f gff -m -c -g 11 -a [Output Protein Fasta File]

6) Functional annotation

- 프로그램: usearch 64bit 버전 8.0.1517
- 데이터베이스:
  KEGG 버전(2015.12.10일자)
  eggnog 버전 4.1
  swissprot(2015.12.10일자)
  SEED subsystems (2015.12.10일자)
- 실행 파라메터: -ublast [Input Fasta File] -db [DB File] -maxaccepts 1 -evalue 1.0E-5 -accel 1.0 -ka_dbsize 700000000 -alnout [Output File]

(2) Pathway 분석 모듈 업데이트 및 비교 분석 모듈 개발

○ Pathway Database 구축 및 분석 파이프라인 확립

- KEGG, eggNOG, GO database를 localization 하여 Database화시킴.
- 이를 통하여 genome annotation 시에 gene function 정보를 추가 입력하는 pipeline을 구축함.

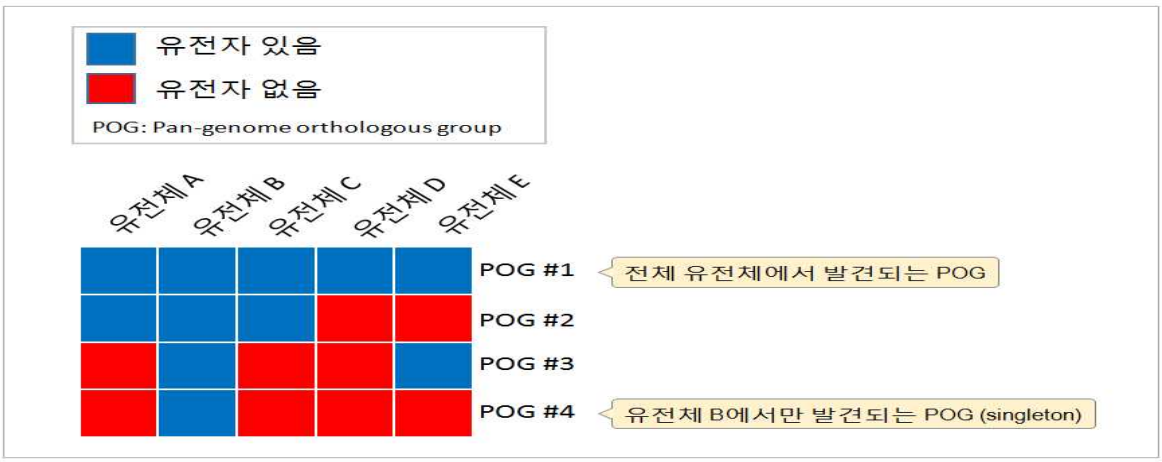
○ Visualization 모듈 개발 및 웹 상에서 구현

- pipeline을 통하여 분석된 결과는 CLG파일 format 으로 결과를 압축함.
- 이를 웨상에서 분석할 수 있도록 분석 모듈을 개발함. 최대 20개의 유전체를 동시에 비교 분석할 수 있도록 interactive 하게 개발함.
- KEGG database 와 연동하여 유전자 정보를 살펴 볼 수 있도록 구현함.

○ 유전자 유무를 이용한 분석 (Gene presence/absence analysis)

(3) 알고리즘

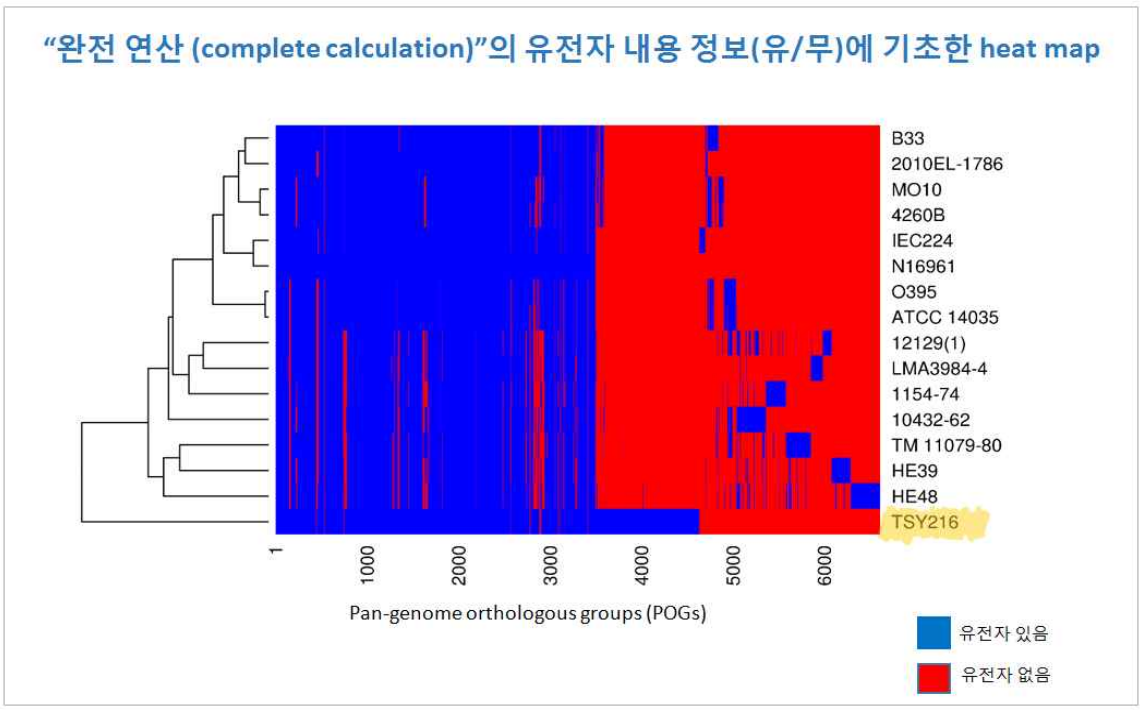
- 다수의 유전체로 범유전체 (Pan-genome)을 생성하고 나면, 모든 CDS는 pan-genome rthologous group (POG) 으로 클러스터링 구성.
- POG는 적어도 한 개 이상의 CDS를 포함.
- 보존이 잘 된 POG는 모든 유전체에서 발견되며, 이것이 *core-genome을 구성



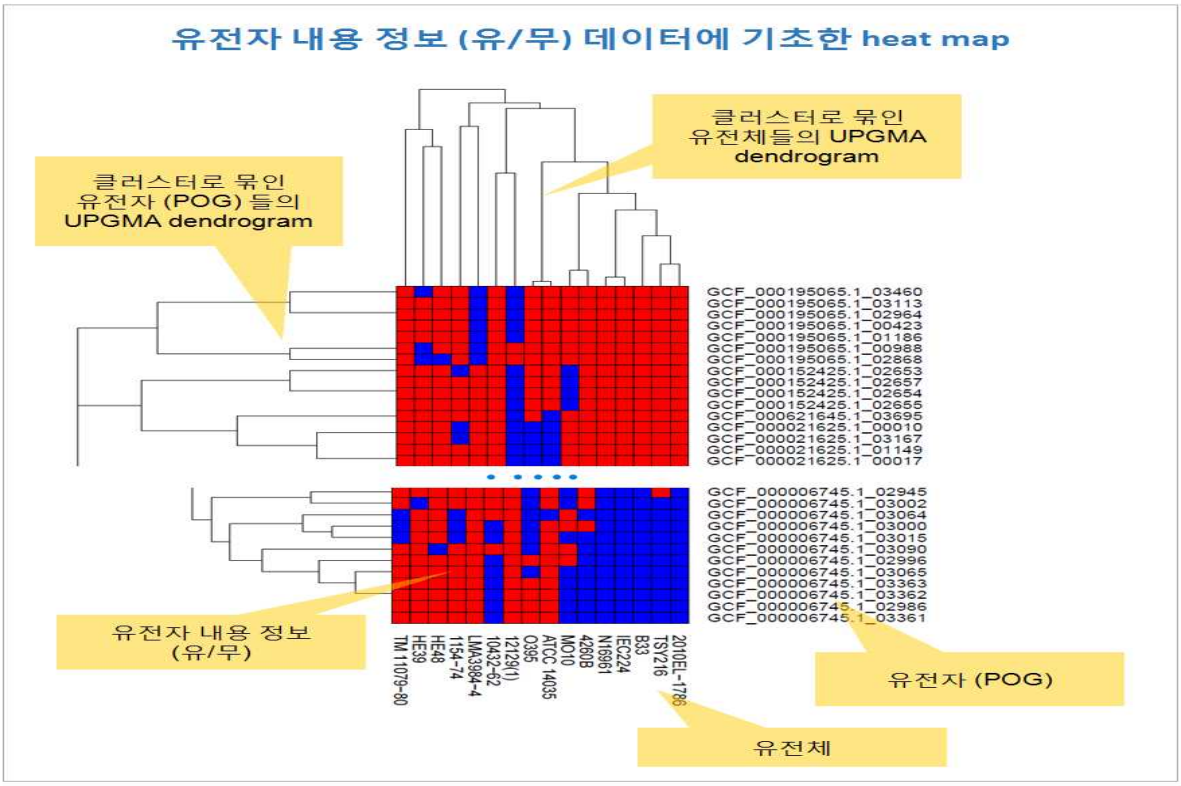
[그림3. 유전자 유무를 이용한 비교 유전체 분석]

(4) EzBioCloud에서 제공하는 2가지의 다른 연산방법

- 완전 연산 (complete calculation): 모든 POG에 기초한 분석
- 차등 연산 (differential calculation): 모든 유전체에 존재하는 POG (core-genome) 와 단일 유전체에만 존재하는 POG (singleton) 가 제외된 POG에 기초한 분석.



[그림4. 완전 연산법을 이용한 유전자 유무 heatmap 작성]



[그림5. 차등 연산법을 이용한 유전자 유무 heatmap 작성]

(5) 전사체 분석 파이프라인 고도화 및 분석 모듈 개발 업데이트

○ 여러 정규화 방법을 이용한 전사체 발현량 제시

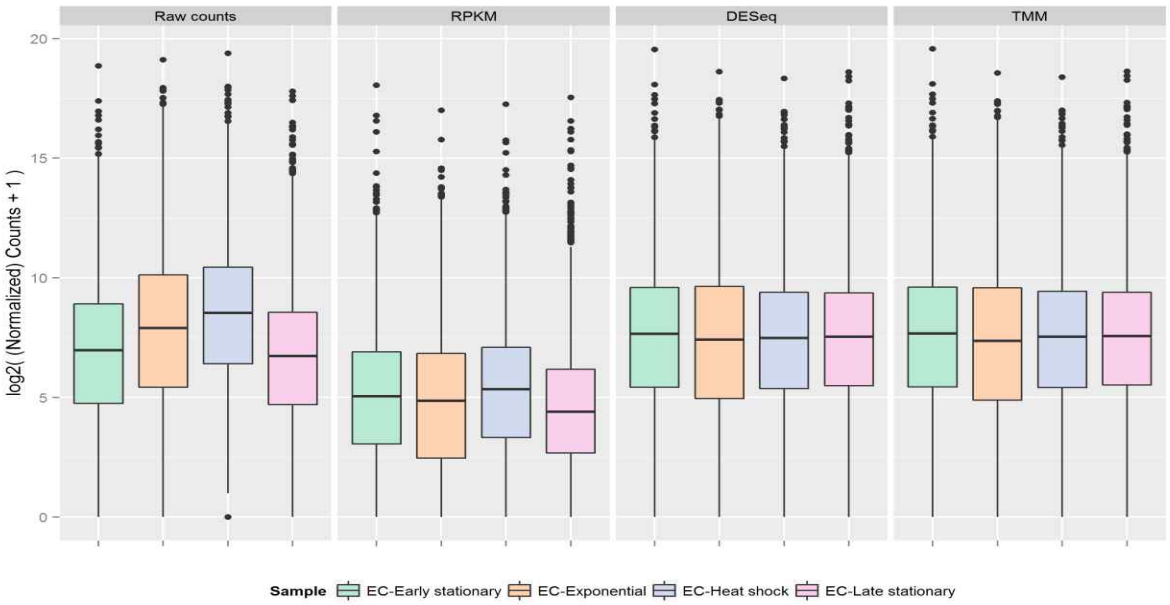
- NGS 시퀀싱 결과를 이용하여 유전자에 맵핑된 read를 발현량으로 계산할 시에 시퀀싱 방법에 따라서 다르게 나타타는 발현량을 보정할 필요가 있음.
- 즉, 라이브러리 크기, 유전자 길이 등 발현량에 영향을 줄 수 있는 요인들을 고려하여 정규화 방법(normallization)을 도입함.
- 이를 위해 일반적으로 사용한 RPKM 이외에 RLE (Relative Log Expression), TMM (Trimmed Mean of M-value) 방법을 RNA-Seq 분석결과를 제시하는 천랩이 본 과제를 통해서 개발한 CLRNASeq software 에 적용함.

※ Relative Log Expression (RLE)

각 유전자에 대해서 Geometric mean(기하평균)을 구하여 read count와의 ratio를 계산한 후, 중앙값(median)을 계산하여 scaling factor로 정의하여 계산함.

※ Trimmed Mean of M-value (TMM)

두 샘플에 대해서 log ratio를 구하여 log ratio가 큰 유전자들, expression이 큰 유전자들을 제외한 후 Weighted mean을 계산함. 이것을 normalization factor라고 정의 했을 때, library size을 반영한 re-scale factor를 다시 계산함.



[그림6. RPKM, RLE, TMM 방법 비교 분석]

Summary of comparison results for the three normalization methods under consideration[1]

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
RLE	++	++	++	++	++
TMM	++	++	++	++	++
RPKM	-	+	+	-	-

[표1. RPKM, RLE, TMM 방법 비교]

○ 통계적 접근을 통한 발현량 차이 유전자 탐색

- NGS 분석을 통해 생산된 read 수를 이용하여 유전체에 Mapping시에 통계적인 알고리즘을 통해 차별 발현된 유전자를 탐색: DESeq2, edgeR 등의 통계적 검정 방법을 CLRNASeq에 도입

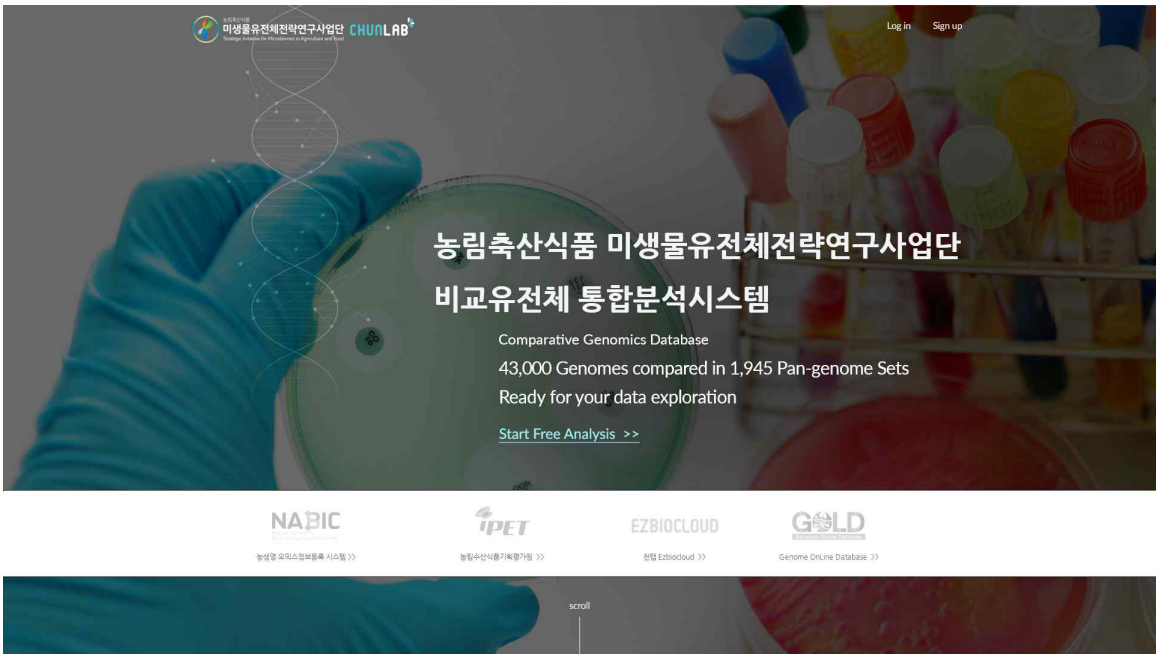
○ 차별 발현 유전자 (DEG)를 이용한 Functional annotation 유의성 검정 알고리즘 개발

- 유전자 기능 분석을 위한 데이터베이스 (KEGG, GO, eggNOG) 등을 프로그램내 도입.
- 차별 발현 유전자(DEG)가 작용하는 특정 기능에 대한 유의성 검정 알고리즘 (Fisher exact test, EASE Score)을 프로그램내 도입.

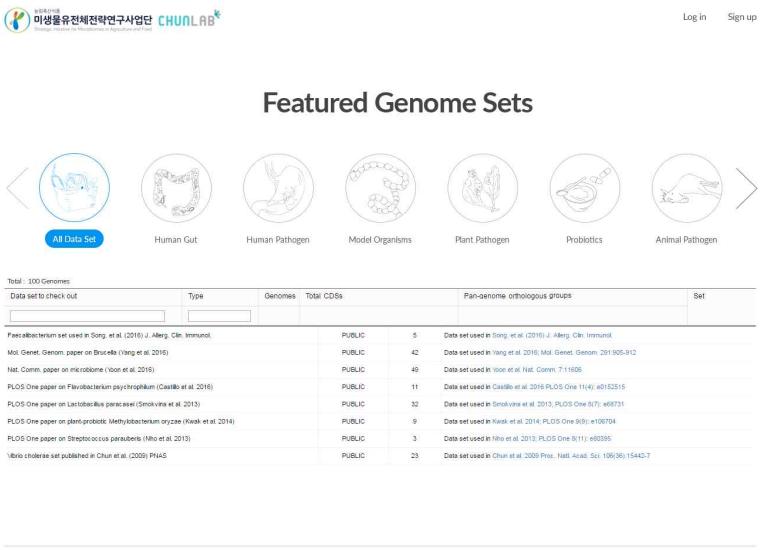
(6) Genome database 구축 및 업데이트

○ 웹 방식을 통한 신규 비교유전체 통합분석시스템 구축 (agri.ezbiocloud.net)

- 비교유전체 셋트 구축 및 업데이트 (44,048 genome/ 1, 945 Pan-genome set)
- 주요 미생물 비교 유전체 set 구축: 동식물 병원균, 유산균 유전체 및 식품위해균 등



[그림7. 비교유전체 통합분석 시스템 홈페이지1]

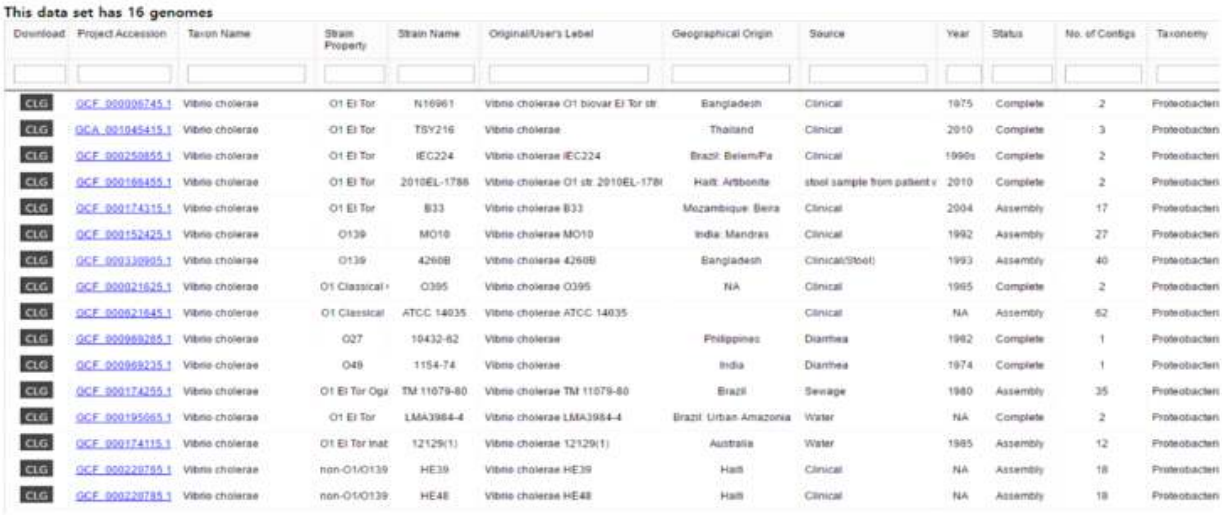


[그림8. 비교유전체 통합분석 시스템 유전체 set]

○ 웹 방식의 비교유전체 분석 파이프라인

- [DATA SET]→[Genomes in This Data Set] 을 선택해서 각 유전체에 대한 정보 (=Metadata)를 확인가능함. EzBioCloud가 보유한 각 유전체에 대한 metadata들은 공개 데이터베이스와 논문들로부터 확보한 것이며 정교한 분석을 할 수 있도록 가능한 많은 metadata를 확보하였음.
- CLG 파일은 개별 유전체를 분석한 결과인 유전자 구성 및 염기서열 등의 정보를 저장하고 있으며

사용자 컴퓨터에 다운로드 받아 저장 할 수 있음. 저장된 CLG 파일은 천랩 소프트웨어인 CLgenomics를 이용해서 열어 볼 수 있는데, 여러 개의 유전체에 해당하는 CLG 파일들을 동시에 열어서 유전체간 차이점 및 공통점 등을 확인 가능함.



This data set has 16 genomes

Download	Project Accession	Taxon Name	Strain Property	Strain Name	Original/User's Label	Geographical Origin	Source	Year	Status	No. of Contigs	Taxonomy
CLG	GCF_000006745.1	Vibrio cholerae	O1 El Tor	N16961	Vibrio cholerae O1 biovar El Tor str	Bangladesh	Clinical	1975	Complete	2	Proteobacteri
CLG	GCA_001045415.1	Vibrio cholerae	O1 El Tor	TSY216	Vibrio cholerae	Thailand	Clinical	2010	Complete	3	Proteobacteri
CLG	GCF_000250855.1	Vibrio cholerae	O1 El Tor	IEC224	Vibrio cholerae IEC224	Brazil: Belem/Pa	Clinical	1990s	Complete	2	Proteobacteri
CLG	GCF_000166455.1	Vibrio cholerae	O1 El Tor	2010EL-1786	Vibrio cholerae O1 str. 2010EL-1786	Haiti: Artibonite	stool sample from patient v	2010	Complete	2	Proteobacteri
CLG	GCF_000174315.1	Vibrio cholerae	O1 El Tor	B33	Vibrio cholerae B33	Mozambique: Beira	Clinical	2004	Assembly	17	Proteobacteri
CLG	GCF_000152425.1	Vibrio cholerae	O139	MO10	Vibrio cholerae MO10	India: Mandras	Clinical	1992	Assembly	27	Proteobacteri
CLG	GCF_000330905.1	Vibrio cholerae	O139	4260B	Vibrio cholerae 4260B	Bangladesh	Clinical/Stool	1993	Assembly	40	Proteobacteri
CLG	GCF_000021625.1	Vibrio cholerae	O1 Classical	O395	Vibrio cholerae O395	NA	Clinical	1965	Complete	2	Proteobacteri
CLG	GCF_000621845.1	Vibrio cholerae	O1 Classical	ATCC 14035	Vibrio cholerae ATCC 14035		Clinical	NA	Assembly	62	Proteobacteri
CLG	GCF_000969285.1	Vibrio cholerae	O27	10432-62	Vibrio cholerae	Philippines	Diarrhea	1962	Complete	1	Proteobacteri
CLG	GCF_000969235.1	Vibrio cholerae	O49	1154-74	Vibrio cholerae	India	Diarrhea	1974	Complete	1	Proteobacteri
CLG	GCF_000174255.1	Vibrio cholerae	O1 El Tor Ogi	TM 11079-80	Vibrio cholerae TM 11079-80	Brazil	Sewage	1980	Assembly	35	Proteobacteri
CLG	GCF_000195065.1	Vibrio cholerae	O1 El Tor	LMA3984-4	Vibrio cholerae LMA3984-4	Brazil: Urban Amazonia	Water	NA	Complete	2	Proteobacteri
CLG	GCF_000174115.1	Vibrio cholerae	O1 El Tor Inat	12129(1)	Vibrio cholerae 12129(1)	Australia	Water	1985	Assembly	12	Proteobacteri
CLG	GCF_000220795.1	Vibrio cholerae	non-O1/O139	HE39	Vibrio cholerae HE39	Haiti	Clinical	NA	Assembly	18	Proteobacteri
CLG	GCF_000220785.1	Vibrio cholerae	non-O1/O139	HE48	Vibrio cholerae HE48	Haiti	Clinical	NA	Assembly	18	Proteobacteri

[그림9. 16개의 *Vibrio vulnificus* 16개 genome의 data set]

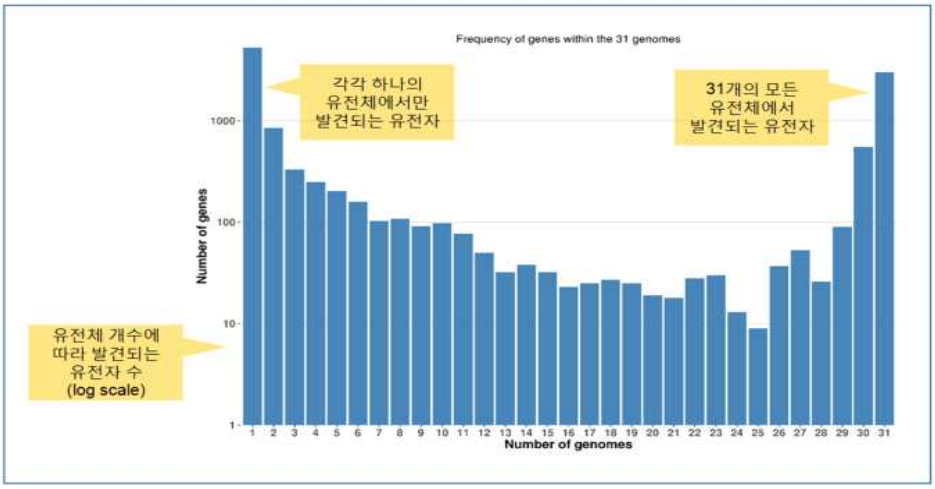
○ 주요 미생물 비교 유전체 data set 결과

[표2. 비교 유전체 분석이 완료된 주요 유전체 목록]

Type	Species	Genomes	Total CDSs	Pan-genome orthologous groups
Human gut	*Faecalibacterium* set	5	14,804	5,322
Human pathogens	*Bacillus anthracis* and related species set	12	68,251	11,615
Soil model organism	*Streptomyces* model set	5	40,314	18,146
Probiotics	commercial probiotics strains	12	25,457	7,403
	Lactobacillus paracasei	32	99,311	5,691
Animal Pathogens	*Flavobacterium psychrophilum* set	11	27,227	3,213
	Streptococcus parauberis set	3	6,618	2,652
Food poison	*Escherichia coli* set	42	201,615	13,710
	Escherichia coli European outbreak	7	34,594	7,388
Total		129	518,191	75,140

(7) Pan-genome의 유전자 빈도 그래프 작성

- 모든 단백질 코딩 유전자(CDS)들, 이 중에서 상동 유전자 후보군 (potentially orthologous genes)은 “Pan-genome Orthologous Group (POG)”을 생성하는 pan-genome 연산 후에 비중복 유전자 세트 (non-redundant gene set)로 묶이게 됨.
- “유전자 빈도 그래프 (gene frequency plot)”는 전체 유전체 세트가 나타내는 일반적인 유전자 빈도를 보여줌. 전형적인 그래프는 U 형태로 그려지며, 전체 유전체에서든 단일 유전체에서든 대부분의 유전자가 여기서 발견됨.
- 아래 예시는 31개의 *Vibrio vulnificus* 유전체를 분석한 결과를 보여줌. 이 유전체들은 총 144,931개의 유전자를 가지고 있으며, 이 중 비중복 유전자 (non-redundant gene)의 개수는 13,220개임을 확인 할 수 있음 (그림에서는 log-scale로 유전자 개수를 나타냄).

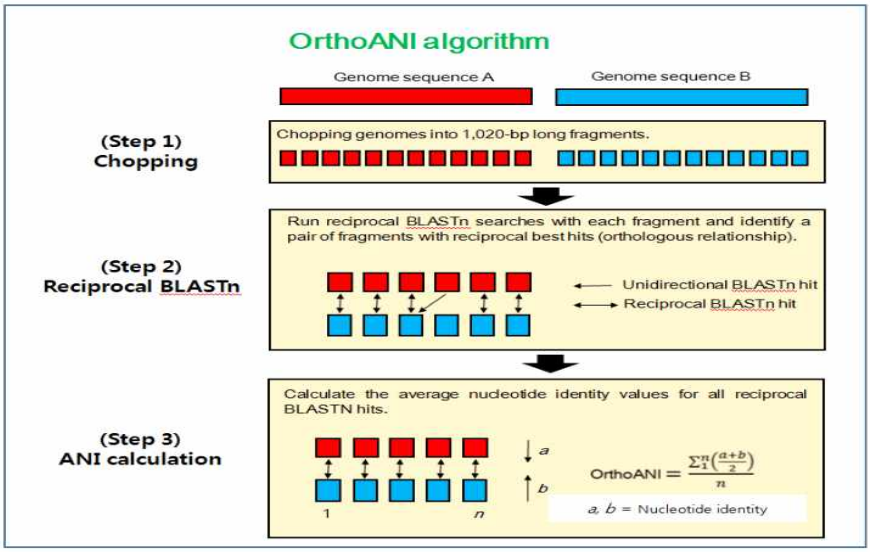


[그림10. Vibrio vulnificus 종의 Pan-genome의 유전자 빈도 그래프]

(8) 농식품 미생물 및 병원성 미생물 유전체 분석 고도화

○ OrthoANI 유전체 유사도 알고리즘 개발

- OrthoANI (Orthologous Average Nucleotide Identity) 값은 두 유전체 염기서열 간의 유사도를 나타나는 값의 한 종류로 기존의 ANI (Average Nucleotide Identity)를 개선하였음.
- OrthoANI는 미생물 분류와 동정에 사용될 수 있으며, 종을 구분하기 위해 제시된 경계값은 약 95%임. 이를 설명하는 알고리즘은 Lee et al. (2015)에 의해 발표되었으며, EzBioCloud 데이터베이스를 구축하기 위해 사용된 표준 알고리즘으로 소프트웨어는 이곳에서 이용 가능함.
- 기존 ANI와 OrthoANI의 주요한 차이점: 기존 ANI의 경우, 사용자는 상호값 (i.e., A->B&B->A)을 얻어야만 했으며 분류학적으로 이용하기 위해서는 이들의 평균값을 사용했음. 이와 달리 OrthoANI는 단일값 (A<->B)을 제공하며, OthoANI는 ANI보다 계산 속도가 빠름.

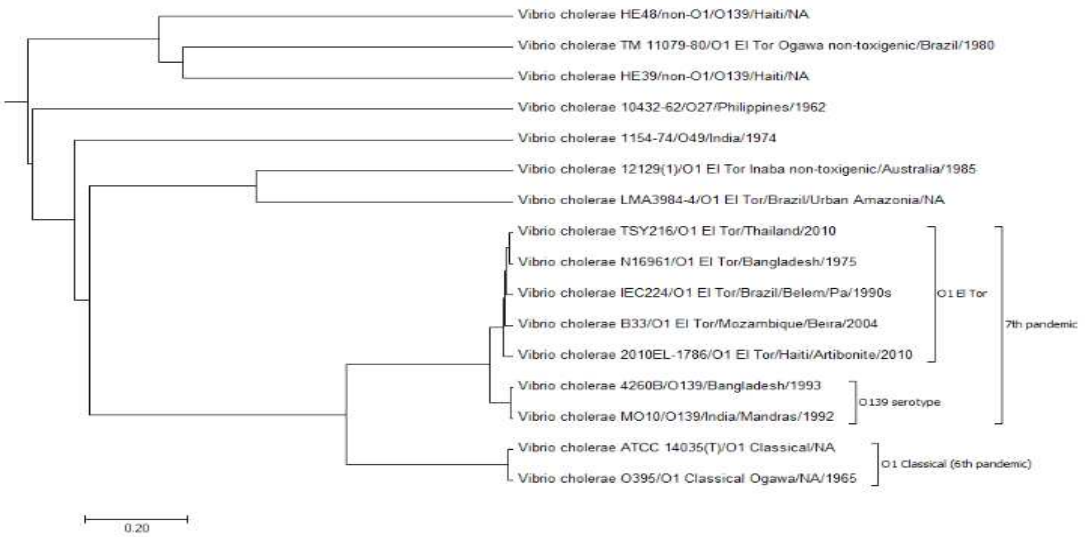


[그림11. OrthoANI 유전체 유사도 알고리즘 모식도]

○ 유전체 기반 계통 분석 (Phylogenomic analysis)

- [PHYLOGENOMICS]→[OrthoANI] 기능: 여러 개의 유전체/균주들 간의 계통학적 유연관계를 추론하기 위한 가장 좋은 방법이 OrthoANI 값을 이용한 계층적 클러스터링 (일반적으로 UPGMA)임. MEGA program을 이용하면 아래 그림에 더 많은 정보를 추가할 수 있음. 지난 수십 년간 무수한 사람들을 사망에 이르게 만든 *Vibrio cholerae* 6기와 7기 유행성 균주가 OrthoANI를 기반으로 분석했을 때 하나의 클러스터에 포함되는 것을 확인 할 수 있음.

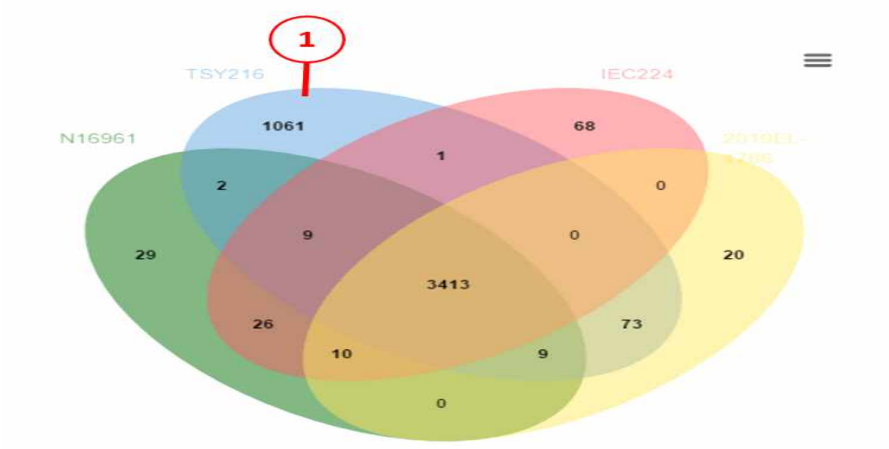
- [PHYLOGENOMICS]→[Tetra-nucleotide] 기능: Tetra-nucleotide 분석 (TNA) 은 실제 염기서열 정렬이 아닌 네 개의 뉴클레오타이드 염기서열의 조합을 이용하는 것이기 때문에 정확히는 계통분석이 아니지만 유전체들이 어떻게 연관되어 있는지에 대해 유용한 추가 정보를 제공함.



[그림12. 16개의 *Vibrio cholerae* 16개 genome의 phylogenomics tree]

○ 핵심유전체 (core genome)과 범유전체 (Pan-genome) 분석

- 비교 유전체 내에서 핵심 유전체와 범유전체의 양상은 [Pan-genome]→[Venn Diagram] 을 통해 확인해 볼 수 있음. 예를 들어, 4개의 유전체 (N16961, TSY216, IEC224, 2010EL-1786)를 비교시 각자의 유전체 데이터를 선택하고 [Draw] 버튼을 클릭하면, 개별 유전체 데이터의 비교 수치를 구할 수 있음.



[그림13. *Vibrio cholerae* 4개 genome의 Venn diagram 모식도]

○ 수평적 유전자 이동(Horizontal genetrasnfer) 추적

- 유전자 클러스터(gene cluster) 을 찾는 유용한 방법은 pairwise gene content matrix를 생성하는 것이므로, 각각을 1대 1로 (16 x 16)으로 비교하여 ortholog를 찾아낼 수 있음.

- [Pairwise analysis]→[Browse Pairwise Ortholog Matrix] 로 이동하여, “N16961”을 기준 유전체(reference genome)로 선택한 후, [RBH(Protein)] type으로 결과를 확인함.

[그림14. 16개의 *Vibrio cholerae* genome의 개별 유전자 비교표]

○ 주요 농식품 병원균의 genome data를 EzBiocloud DB에 업데이트 (http://www.ezbiocloud.net)

[표3. 주요 병원균 genome 업데이트 수]

No	Species	genome 수
1	*Acinetobacter baumannii*	2,293
2	*Campylobacter jejuni*	1,081
3	*Escherichia coli*	8,999
4	*Klebsiella pneumoniae*	2,816
5	*Listeria monocytogenes*	1,095
6	*Mycobacterium abscessus*	1,376
7	*Mycobacterium tuberculosis*	5,450
8	*Pseudomonas aeruginosa*	2,569
9	*Salmonella enterica*	7,495
10	*Staphylococcus aureus*	8,146
11	*Streptococcus pneumoniae*	8,053
	합 계	49,373

(2018. 6. 15. 현재)

(9) 표준균주 (type strain) 유전체 데이터 생산 및 DB 구축

○ 표준균주 유전체 데이터 생산 및 분석/ DB 구축

- 농식품 주요세균 (동식물병원균, 유산균, 식중독균, 토양미생물 등)의 표준균주 (type strain) 들 중에서 유전체 분석 (genome sequencing) 이 필요한 종들을 선정하여 총 188 종에 대하여 유전체 분석을 수행함.
- Genome sequencing 은 일루미나 MiSeq 플랫폼을 이용하였고, 모두 coverage 50 이상으로 수행하였음.
- 분석한 결과는 자체 구축한 분석 프로그램인 CLgenomics 프로그램과 생명정보 플랫폼인 BIOiPLUG를 통하여 분석하여 SQL DB로 구축함.

○ 유산균 80종에 대한 유전체 데이터 생산 및 분석 / DB 구축

- 농식품 산업에 유용한 유산균의 표준균주의 유전체 분석이 안 된 균주를 선정함.
- 생명연 KCTC와 농진청 KACC 등의 미생물 자원센터에서 균주를 분양받아 genome 분석을 수행.
- 주요 유산균 (*Lactobacillus, Lactococcus, Leuconostoc, Streptococcus, Weissella*) 등의 표준균주 80 종을 분석함.

[표4. 시퀀싱된 주요세균 genome 목록]

	TAXON NAME	STRAIN	Genome Size	G+C ratio	CDS	Contigs	Coverage
1	*Achromobacter spanius*	CCUG 47062	6,370,283	0.64	5,717	62	117
2	*Acinetobacter baumannii*	CCUG 19096	3,967,924	0.39	3,734	69	102
3	*Acinetobacter courvalinii*	CCUG 67960	3,957,728	0.43	3,620	72	161
4	*Acinetobacter gandensis*	CCUG 68482	3,183,002	0.40	2,988	70	164
5	*Adlercreutzia equolifaciens*	KCTC 15235	3,028,603	0.62	2,347	329	408
6	Anaerotignum lactatifermentans	KCTC 15066	3,612,416	0.45	3,465	247	287
7	*Arcobacter skirrowii*	CCUG 10374	1,953,172	0.28	1,984	30	170
8	*Arthrobacter agilis*	KCTC 3200	3,492,146	0.69	3,170	25	146
9	*Arthrobacter humicola*	NRRL B-24479	4,613,800	0.67	4,147	55	134
10	*Arthrobacter oryzae*	NRRL B-24478	4,320,069	0.67	3,869	29	124
11	*Bacillus cihuensis*	DSM 25969	5,471,976	0.37	5,048	125	152
12	*Bacillus firmus*	KACC 10897	4,521,364	0.42	4,561	272	98
13	*Bacillus flexus*	KACC 10893	3,993,180	0.38	4,148	120	222
14	*Blastomonas natatoria*	KCTC 2886	3,868,225	0.63	3,656	92	108
15	*Blautia glucerasea*	KCTC 15131	2,980,405	0.42	2,757	29	332
16	*Bordetella bronchiseptica*	KACC 11941	5,135,085	0.68	4,797	46	94
17	*Bordetella parapertussis*	KACC 11942	4,727,047	0.68	4,502	61	217
18	*Brachybacterium paraconglomeratum*	KCTC 9916	3,829,697	0.69	3,370	38	96
19	*Burkholderia latens*	CCUG 54555	7,103,502	0.67	6,290	291	200
20	*Burkholderia stagnalis*	CCUG 65686	8,127,870	0.67	7,069	401	231
21	*Campylobacter coli*	CCUG 11283	1,914,845	0.31	2,012	27	187
22	*Campylobacter hyointestinalis subsp. lawsonii*	CCUG 34538	1,796,518	0.33	1,806	36	175
23	*Campylobacter jejuni subsp. jejuni*	CCUG 11284	1,749,738	0.30	1,812	42	221
24	*Campylobacter lari subsp. concheus*	CCUG 55786	1,465,880	0.30	1,469	36	489
25	*Campylobacter sputorum subsp. sputorum*	CCUG 9728	1,728,151	0.30	1,747	26	190
26	*Campylobacter volucris*	CCUG 57498	1,520,196	0.29	1,517	31	319
27	*Castellaniella defragrans*	CCUG 39790	3,962,237	0.69	3,528	345	419
28	*Clostridium citroniae*	KCTC 5743	6,206,595	0.49	5,609	89	180
29	*Clostridium paraputrificum*	DSM 2630	3,653,827	0.31	3,378	255	309
30	*Comamonas kerstersii*	CCUG 15333	3,510,398	0.60	3,213	37	84
31	*Corynebacterium ulcerans*	DSM 46325	2,457,240	0.53	2,197	39	678
32	*Cupriavidus gilardii*	CCUG 38401	5,793,931	0.67	5,173	119	315
33	*Cupriavidus pauculus*	CCUG 12507	6,416,057	0.64	5,850	112	110
34	*Curtobacterium citreum*	KCTC 9100	3,593,933	0.72	3,399	65	88
35	*Edwardsiella anguillarum*	CCUG 64215	4,340,292	0.58	3,844	102	185
36	*Eisenbergiella tayi*	KCTC 15433	7,610,967	0.47	6,468	126	89
37	*Enterobacter cancerogenus*	KACC 10528	4,789,395	0.56	4,446	42	120
38	*Enterobacter cloacae subsp. dissolvens*	KACC 13002	4,871,778	0.55	4,512	75	114
39	*Erythrobacter citreus*	KCTC 12214	2,975,298	0.64	2,846	31	123
40	*Fructobacillus pseudoficulneus*	DSM 15468	1,417,479	0.44	1,328	18	482
41	*Fusobacterium necrophorum subsp. funduliforme*	CCUG 42162	2,130,453	0.35	2,025	27	190
42	*Halobacillus trueperi*	KCTC 3686	4,112,483	0.44	4,103	66	71
43	*Helicobacter pullorum*	CCUG 33837	1,990,340	0.36	1,899	314	188
44	*Hungatella effluvii*	KCTC 15431	6,881,844	0.49	5,901	70	70
45	*Ideonella dechloratans*	CCUG 30977	4,510,912	0.69	4,123	158	346

46	*Kerstersia gyiorum*	CCUG 47000	3,986,919	0.62	3,467	22	144
47	*Klebsiella michiganensis*	CCUG 66515	6,157,837	0.56	5,692	236	137
48	*Klebsiella oxytoca*	KACC 11934	5,813,674	0.55	5,311	87	70
49	*Kocuria rosea*	KCTC 3137	3,915,004	0.73	3,562	143	139
50	*Lactobacillus acetotolerans*	JCM 3825	1,608,615	0.36	1,545	123	367
51	*Lactobacillus acidifarinae*	JCM 15949	2,947,326	0.52	2,752	69	339
52	*Lactobacillus acidipiscis*	JCM 10692	2,559,098	0.39	2,356	496	343
53	*Lactobacillus amylophilus*	KCTC 3161	1,564,051	0.44	1,568	40	409
54	*Lactobacillus amylovorus*	DSM 20531	2,326,498	0.38	2,407	5	552
55	*Lactobacillus amylovorus*	KCTC 3597	2,149,148	0.38	2,176	3	386
56	*Lactobacillus amylovorus*	KCTC 3597	2,149,148	0.38	2,176	3	386
57	*Lactobacillus animalis*	KCTC 3501	1,914,183	0.41	1,837	81	189
58	*Lactobacillus apodemi*	JCM 16172	2,196,188	0.39	2,086	223	366
59	*Lactobacillus aviarius subsp. araffinosus*	JCM 5667	1,509,666	0.38	1,425	84	365
60	*Lactobacillus aviarius subsp. aviarius*	KCTC 5063	1,694,651	0.40	1,597	49	333
61	*Lactobacillus bobalius*	KACC 16343	2,885,890	0.35	2,803	29	140
62	*Lactobacillus bombicola*	KACC 19374	1,799,336	0.35	1,639	251	326
63	*Lactobacillus brevis*	KCTC 3498	2,484,326	0.46	2,429	22	164
64	*Lactobacillus buchneri*	KCTC 5064	2,482,540	0.44	2,365	113	250
65	*Lactobacillus ceti*	JCM 15609	1,444,872	0.34	1,322	35	630
66	*Lactobacillus coleohominis*	KCTC 21007	1,889,911	0.41	1,986	66	159
67	*Lactobacillus collinoides*	KCTC 5050	3,767,101	0.46	3,380	245	181
68	*Lactobacillus coryniformis subsp. coryniformis*	KCTC 3167	2,951,508	0.43	2,807	1	348
69	*Lactobacillus crispatus*	KCTC 5054	2,086,436	0.37	2,038	195	362
70	*Lactobacillus crustorum*	KACC 16344	2,242,351	0.35	2,173	92	164
71	*Lactobacillus delbrueckii subsp. lactis*	DSM 20072	2,633,895	0.49	2,615	16	401
72	*Lactobacillus delbrueckii subsp. sunkii*	JCM 17838	2,004,337	0.50	1,851	1	297
73	*Lactobacillus equicursoris*	JCM 14600	2,180,521	0.47	1,951	278	425
74	*Lactobacillus farciminis*	DSM 20184	2,557,666	0.36	2,485	1	588
75	*Lactobacillus fermentum*	KCTC 3112	1,835,472	0.53	1,782	117	316
76	*Lactobacillus formosensis*	KACC 18721	2,543,509	0.36	2,445	105	63
77	*Lactobacillus fructivorans*	KCTC 3543	1,375,168	0.39	1,337	10	442
78	*Lactobacillus fructivorans*	JCM 1198	1,397,321	0.39	1,384	24	628
79	*Lactobacillus frumenti*	JCM 11122	1,804,088	0.43	1,684	188	379
80	*Lactobacillus fuchuensis*	KCTC 3797	2,141,008	0.42	2,031	81	186
81	*Lactobacillus futsaii*	KACC 18747	2,555,449	0.36	2,510	134	194
82	*Lactobacillus gasseri*	KCTC 3163	1,836,679	0.35	1,754	21	182
83	*Lactobacillus graminis*	KCTC 3542	1,864,485	0.40	1,768	83	242
84	*Lactobacillus hordei*	JCM 16179	2,442,019	0.35	2,329	383	358
85	*Lactobacillus kimchiensis*	KACC 15533	2,717,139	0.35	2,586	66	74
86	*Lactobacillus kimchii*	KACC 12383	2,754,256	0.35	2,630	37	313
87	*Lactobacillus kisonensis*	JCM 15041	3,048,108	0.42	2,761	183	125
88	*Lactobacillus koreensis*	KCTC 13530	2,940,525	0.49	2,660	116	126
89	*Lactobacillus malefermentans*	KCTC 3548	2,065,638	0.41	2,016	126	189
90	*Lactobacillus mali*	KCTC 3596	2,687,026	0.36	2,620	117	275
91	*Lactobacillus manihotivorans*	KCTC 21010	3,379,545	0.48	3,242	448	163
92	*Lactobacillus mucosae*	KCTC 21011	2,290,210	0.46	2,041	94	198
93	*Lactobacillus mudanjiangensis*	KCTC 21026	3,380,381	0.43	3,322	25	94
94	*Lactobacillus murinus*	JCM 1717	2,223,603	0.40	2,071	126	249
95	*Lactobacillus nantensis*	KACC 12408	2,944,487	0.36	2,780	75	268

96	*Lactobacillus parabuchneri*	KCTC 3503	2,591,857	0.43	2,395	55	170
97	*Lactobacillus paracasei supsp. tolerans*	KACC 12427	2,411,151	0.46	2,396	375	125
98	*Lactobacillus paraplantarum*	KCTC 5045	3,455,312	0.44	3,269	192	177
99	*Lactobacillus pasteurii*	JCM 18989	1,905,038	0.39	1,810	63	558
100	*Lactobacillus perolens*	JCM 8646	3,132,982	0.49	3,040	175	292
101	*Lactobacillus plantarum*	KCTC 3108	3,226,491	0.44	3,017	26	147
102	*Lactobacillus plantarum subsp. plantarum*	KCTC 3108	3,226,491	0.44	3,017	26	147
103	*Lactobacillus pobuzihii*	KCTC 13174	2,430,804	0.37	2,176	202	127
104	*Lactobacillus porcinae*	KCTC 21027	2,864,444	0.47	2,693	77	67
105	*Lactobacillus rapi*	JCM 15042	2,873,848	0.43	2,639	172	123
106	*Lactobacillus rhamnosus*	KACC 11953	2,944,953	0.47	2,729	41	106
107	*Lactobacillus rossiae*	JCM 16176	2,963,677	0.43	2,772	217	299
108	*Lactobacillus sakei subsp. carnosus*	KCTC 3802	1,993,539	0.41	1,997	54	425
109	*Lactobacillus sakei subsp. sakei*	KCTC 3603	1,917,787	0.41	1,891	34	371
110	*Lactobacillus salivarius*	JCM 1231	2,012,368	0.33	1,938	40	293
111	*Lactobacillus salivarius subsp. salicinius*	KCTC 3600	1,999,025	0.33	1,946	62	373
112	*Lactobacillus uvarum*	JCM 16870	2,742,877	0.37	2,576	176	262
113	*Lactobacillus vaginalis*	KCTC 3515	1,844,407	0.40	1,777	145	226
114	*Lactobacillus versmoldensis*	KCTC 3814	2,431,179	0.38	2,358	153	343
115	*Lactobacillus zeae*	KCTC 3804	3,133,564	0.48	2,982	71	148
116	*Lactococcus fujiensis*	JCM 16395	2,130,437	0.37	2,079	84	329
117	*Lactococcus lactis*	KCTC 3769	2,586,834	0.35	2,557	138	211
118	*Lactococcus piscium*	KCTC 3639	2,451,990	0.39	2,383	83	226
119	*Leclercia adecarboxylata*	KCTC 1036	5,245,695	0.55	4,732	472	98
120	*Leuconostoc carnosum*	KCTC 3525	1,825,401	0.37	1,789	52	216
121	*Leuconostoc citreum*	KCTC 3526	1,739,383	0.39	1,709	22	256
122	*Leuconostoc fallax*	KCTC 3537	1,643,579	0.38	1,618	10	551
123	*Leuconostoc gelidum subsp. gelidum*	KCTC 3527	1,983,727	0.37	1,918	52	258
124	*Leuconostoc kimchii*	IH25	2,111,413	0.38	2,096	66	246
125	*Leuconostoc lactis*	DSM 20202	1,663,178	0.43	1,643	69	270
126	*Leuconostoc mesenteroides subsp. suionicum*	KACC 17730	2,019,080	0.37	1,978	28	180
127	*Leuconostoc pseudomesenteroides*	JCM 9696	2,151,861	0.39	2,089	31	228
128	*Nocardia harenae*	NRRL B-24459	6,150,566	0.72	5,649	64	122
129	*Nocardia sienata*	IFM 10088	6,874,984	0.68	6,261	97	62
130	*Ochrobactrum pituitosum*	CCUG 50899	5,170,436	0.53	4,960	60	60
131	*Paenibacillus amylolyticus*	KACC 11263	7,146,360	0.46	6,291	65	73
132	*Paenibacillus hordei*	KACC 15511	5,198,384	0.39	4,495	25	159
133	*Paenibacillus jamilae*	KACC 10925	5,632,380	0.45	4,911	89	106
134	*Paenibacillus lautus*	KCTC 3456	7,094,103	0.50	6,387	57	68
135	*Paenibacillus nicotianae*	KACC 18746	5,226,249	0.39	4,568	53	82
136	*Pantoea agglomerans*	KCTC 2564	4,659,933	0.55	4,261	33	104
137	*Pantoea stewartii subsp. stewartii*	CCUG 26359	4,916,637	0.54	4,986	352	80
138	*Planomicrobium okeanokoites*	KCTC 3672	3,318,407	0.46	3,311	74	118
139	*Pseudoalteromonas issachenkonii*	KCTC 12958	4,098,137	0.40	3,621	28	189
140	*Pseudochrobactrum saccharolyticum*	CCUG 33852	3,762,651	0.51	3,370	12	88
141	*Pseudomonas aeruginosa*	CCUG 551	6,423,673	0.66	5,904	75	113
142	*Pseudomonas argentinensis*	CCUG 50743	5,156,427	0.64	4,653	101	222

143	*Pseudomonas brassicacearum subsp. brassicacearum*	CCUG 51508	6,733,367	0.61	5,937	61	117
144	*Pseudomonas brenneri*	CCUG 51514	5,998,537	0.60	5,397	57	75
145	*Pseudomonas chlororaphis subsp. aureofaciens*	CCUG 712	7,011,928	0.63	6,265	102	69
146	*Pseudomonas cissicola*	CCUG 18839	5,353,485	0.64	4,580	315	103
147	*Pseudomonas extremorientalis*	CCUG 51517	6,351,674	0.61	5,663	66	93
148	*Pseudomonas fragi*	NRRL B-727	5,039,202	0.59	4,540	46	62
149	*Pseudomonas koreensis*	CCUG 51519	6,084,671	0.61	5,465	33	60
150	*Pseudomonas lini*	CCUG 51522	6,501,740	0.59	5,822	45	105
151	*Pseudomonas moorei*	CCUG 53114	6,472,558	0.60	5,960	59	94
152	*Pseudomonas reinekei*	CCUG 53116	6,256,747	0.59	5,654	72	66
153	*Pseudomonas rhodesiae*	KACC 10811	6,330,905	0.60	5,763	104	62
154	*Pseudomonas tolaasii*	CCUG 23369	6,744,628	0.61	6,043	55	58
155	*Pseudomonas vancouverensis*	CCUG 49675	6,498,322	0.60	5,841	215	95
156	*Ralstonia insidiosa*	CCUG 46789	5,723,831	0.64	5,301	15	117
157	*Rhizobium leguminosarum*	KACC 10640	7,888,479	0.61	7,469	161	52
158	*Rhizobium radiobacter*	CCUG 3354	5,500,217	0.59	5,217	27	53
159	*Rhodococcus aetherivorans*	DSM 44752	6,448,087	0.70	5,949	137	120
160	*Rhodococcus erythropolis*	KCTC 1062	6,617,571	0.62	6,214	91	71
161	*Rhodococcus globerulus*	KACC 20816	6,743,939	0.62	6,294	34	52
162	*Rhodococcus gordoniae*	KACC 20860	4,847,028	0.68	4,410	22	72
163	*Rhodococcus pyridinivorans*	KACC 14478	5,211,959	0.68	4,795	60	60
164	*Rhodococcus qingshengii*	KCTC 19205	7,334,709	0.62	6,970	131	55
165	*Romboutsia lituseburensis*	KCTC 5843	3,941,444	0.28	3,634	301	276
166	*Salinicola salarius*	KCTC 12664	4,001,911	0.62	3,572	26	80
167	*Serratia marcescens subsp. sakuensis*	KCTC 42172	5,068,642	0.60	4,671	36	79
168	*Shewanella algidipiscicola*	KCTC 22879	4,208,887	0.47	3,699	63	115
169	*Shewanella aquimarina*	KCTC 22430	4,422,112	0.53	3,786	91	78
170	*Shewanella marinintestina*	KCTC 22440	4,884,448	0.43	4,166	60	80
171	*Staphylococcus capitis subsp. capitis*	NRRL B-14752	2,441,108	0.33	2,353	146	135
172	*Staphylococcus caprae*	KCTC 3583	2,606,927	0.34	2,475	25	140
173	*Staphylococcus epidermidis*	KACC 13234	2,438,306	0.32	2,218	43	224
174	*Staphylococcus haemolyticus*	KCTC 3341	2,459,930	0.33	2,356	75	137
175	*Staphylococcus xylosus*	KCTC 3342	2,748,712	0.33	2,519	37	219
176	*Streptococcus gallolyticus subsp. gallolyticus*	KACC 13794	2,464,121	0.38	2,412	22	257
177	*Streptococcus gallolyticus subsp. macedonicus*	KACC 13851	2,155,104	0.37	2,231	93	160
178	*Streptococcus salivarius subsp. thermophilus*	KACC 11857	2,031,901	0.39	2,158	96	126
179	*Streptococcus thermophilus*	KACC 11857	2,031,901	0.39	2,158	96	126
180	*Streptomyces acidiscabies*	DSM 41668	11,095,879	0.70	9,720	551	427
181	*Sulfitobacter pontiacus*	KCTC 32185	3,768,175	0.60	3,681	33	92
182	*Veillonella rogosae*	KCTC 5967	2,192,914	0.39	1,961	108	256
183	*Vibrio chagasii*	CCUG 48643	5,363,722	0.44	4,866	194	65
184	*Vibrio fluvialis*	KCTC 2473	4,767,334	0.50	4,372	53	133
185	*Vibrio kanaloae*	CCUG 56968	4,585,638	0.44	3,995	74	113
186	*Vibrio xuii*	KCTC 12703	5,016,111	0.47	4,526	415	114
187	*Weissella minor*	JCM 1168	1,772,829	0.39	1,774	28	341
188	*Weissella soli*	KCTC 3789	1,683,184	0.44	1,553	1	439

○ 미생물 유전체 농생명정보센터(NABIC)에 등록

- 유산균을 포함한 농식품 주요균주 중 7개 균주의 유전체 결과를 아래와 같이 농생명정보센터(NABIC) 에 등록함.

[표5. 농생명정보센터(NABIC)에 등록한 세균유전체 목록]

No.	Species	Strain name	Genome size	G+C ratio (%)	CDS	Contig number	Coverage
1	*Bacillus vietnamensis*	B-23890	4,457,232	43.74	4,524	104	110
2	*Arthrobacter oryzae*	B-24478	4,323,954	67.11	3,876	41	366
3	*Arthrobacter oxydans*	KCTC 3383	4,860,016	65.67	4,504	72	342
4	*Arthrobacter nitroguajacolicus*	KCTC 9902	2,116,740	71.61	1,805	46	1089
5	*Lactobacillus fructivorans*	KCTC 3543	1,375,168	39.45	1,337	10	442
6	*Lactobacillus brevis*	KCTC 3498	2,484,326	46.31	2,429	22	164
7	*Lactobacillus parabuchneri*	KCTC 3503	2,591,857	43.42	2,395	55	170

(10) 미생물 전략유전체 사업단 내 타과제 지원 분석 (세균류 유전체 분석지원)

* 본 연구팀이 개발하는 내용은 미생물의 유전체, 전사체 분석 파이프라인, 소프트웨어, 데이터베이스를 포함한 전체 시스템임. 기존의 공개된 프로그램은 리눅스 기반의 프로그램들로 일반 연구자들이 사용하기 어려운 점이 많았음. 일반 연구자들이 사용할 수 있도록 최적의 분석 파이프라인을 set up 할 뿐만 아니라 GUI 기반의 분석 소프트웨어를 개발하여 생물정보 비전공자인 일반 연구자도 쉽게 유전체, 전사체를 분석 및 비교 분석할 수 있도록 하는 것임. 또한 데이터베이스를 구축하여 공개된 유전체 데이터, 또는 기존에 연구한 데이터와 쉽게 비교 할 수 있도록 시스템화 하는 것임.

(1차년도)

○ 유전체 분석지원 내용 (2014. 8~ 2015. 6): 4개 과제분야 총 52건 (시료별) 분석

- 생물비료 분야 충북대 이이 교수님 (유전체 분석: 4건)
- 생물비료 분야 충북대 사동민 교수님 (메타지놈 분석: 37건)
- 동물병원성 분야 경상대 김석 교수님 (전사체 분석: 8건)
- 메타유전체 연세대 송주연 교수님 (유전체 분석: 3건)

(2차년도)

○ 유전체 분석지원 내용 (2015. 8~ 2016. 6): 4개 과제분야 총 53건 (시료별) 분석

- 생물비료 분야 충북대 이이 교수님 (비교유전체 분석: 18건)
- 생물비료 분야 충북대 사동민 교수님 (메타지놈 분석: 25건)
- 동물병원성 분야 경상대 김석 교수님 (전사체 분석: 8건)
- 김치분야 중앙대 전체옥 교수님 (유전체 분석: 2건)

(3차년도)

○ 유전체 분석지원 내용 (2016. 8~ 2017. 6): 3개 과제분야 총 16건 (시료별) 분석

- 생물비료 분야 충북대 이이 교수님 (비교유전체 분석: 3건)
- 동물병원성 분야 경상대 김석 교수님 (전사체 분석: 6건)
- 표준유전체분야 경북대 이동우 교수님 (군집분석 4건, 유전체 분석: 3건)

(4차년도)

○ 유전체 분석지원 내용 (2017. 8~ 2018. 6): 2개 과제분야 총 2건 (시료별) 분석

- 생물비료 분야 충북대 이이 교수님 (유전체 분석: 1건)
- 표준유전체분야 고려대 이하나 교수님 (유전체 분석: 1건)

(11) 산업기술인력 양성교육

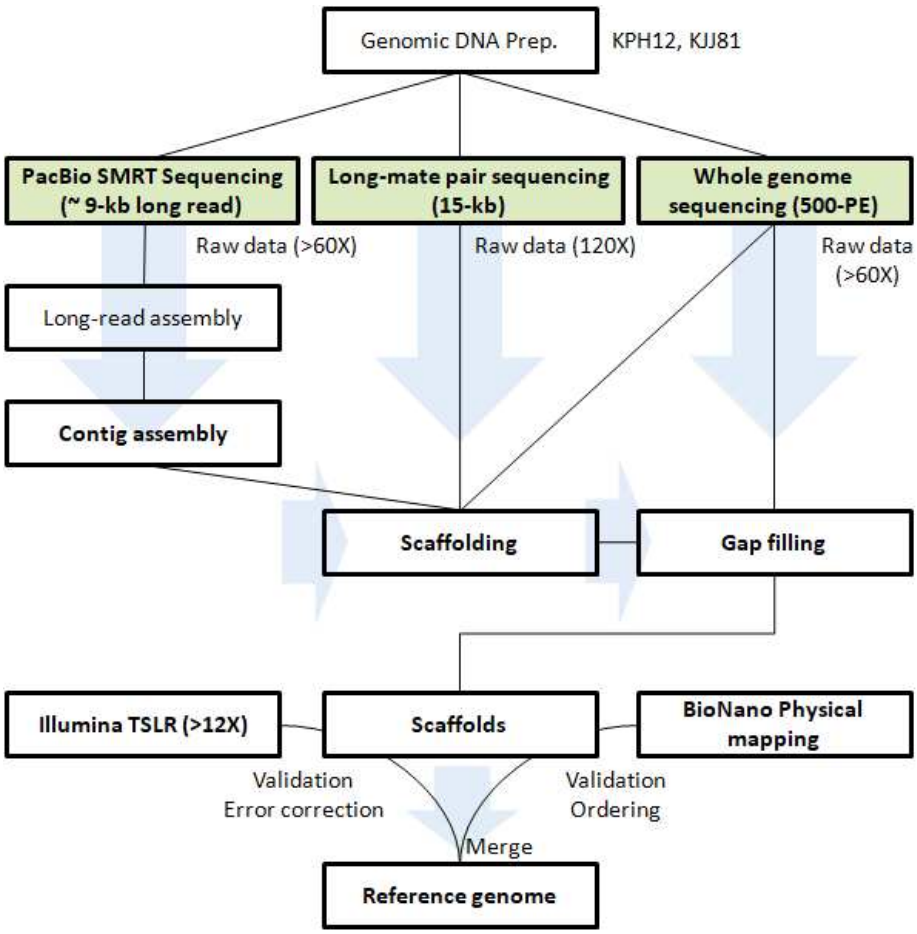
No	프로그램명	프로그램 내용	교육기관	교육 개최회수	총 교육시간	총 교육인원
1	Advanced RNA-Seq Workshop	-Introduction of NGS -Introduction to RNA-seq experiment : sample & Library preparation -Primary data analysis : Bioinformatics pipeline -Secondary data analysis : Data Interpretation -Case study 1 : Prokaryote (Bacteria) -Case study 2 : Eukaryote (Human)	천랩	**3회(1차년도)** (2015.03.20) (2015.04.30) (2015.06.02.)	18시간 (6시간* 3회)	90명
2	1-day NGS workshop -서울대	-Introduction of NGS -Microbiome 연구 방법 및 동향 -CLcommunity 소프트웨어를 이용한 분석 실습 -Microbial genomes 연구 방법 및 동향 -CLgenomics 소프트웨어를 이용한 분석 실습 -Microbial Transcriptomes 연구 방법 및 동향 -CLRNASeq 소프트웨어를 이용한 분석 실습	천랩	**2회(1차년도)** (2015.04.07) (2015.05.19.)	12시간 (6시간 *2회)	66명
3	1-day NGS workshop	-Introduction of NGS -Microbiome 연구 방법 및 동향 -CLcommunity 소프트웨어를 이용한 분석 실습 -Microbial genomes 연구 방법 및 동향 -CLgenomics 소프트웨어를 이용한 분석 실습 -Microbial Transcriptomes 연구 방법 및 동향 -CLRNASeq 소프트웨어를 이용한 분석 실습	천랩	**5회(2차년도)** (2015.9.17) (2015.10.29) (2016.1.27) (2016.3.29) (2016.4.26)	30시간 (6시간* 5회)	198명
4	Advanced Genomics Workshop	- Introduction of NGS - Bacterial Genome 연구 방법 및 동향 - 유전체 기반의 Bacterial Identification - Web-based comparative genomics - CLgenomics 소프트웨어를 이용한 분석 실습 - NCBI Genome submission	천랩	**1회(2차년도)** (2016.5.24)	6시간	39명
5	1-day NGS workshop -서울대	-Introduction of NGS -Microbiome 연구 방법 및 동향 -CLcommunity 소프트웨어를 이용한 분석 실습 -Microbial genomes 연구 방법 및 동향 -CLgenomics 소프트웨어를 이용한 분석 실습 -Microbial Transcriptomes 연구 방법 및 동향 -CLRNASeq 소프트웨어를 이용한 분석 실습	천랩	**1회(3차년도)** (2016.09.27)	8시간	37명
6	Microbiome 교육 Workshop	-Microbiome 연구 방법 및 동향 -CLcommunity 소프트웨어를 이용한 분석 실습 -Microbial genomes 연구 방법 및 동향 -CLgenomics 소프트웨어를 이용한 분석 실습	천랩	**1회(3차년도)** (2017.05.30)	4시간	31명
7	Microbiota 교육 Workshop -서울대	-Microbiota 연구 방법 및 동향 -BIOiPLUG를 이용한 미생물군집 분석 실습 -Microbial genomes 연구 방법 및 동향 -BIOiPLUG를 이용한 비교유전체 분석 실습	천랩	**1회(4차년도)** (2018.01.30)	5시간	38명

2-2. 제 1 협동 연구수행 내역 : 테라젠이텍스

(1) 진균류 참조유전체 조립 파이프라인 개발 및 고도화

1) 진균류 참조유전체 조립 파이프라인

○ Long-reads을 중심으로 한 진균류 참조유전체 조립 파이프라인 workflow



[그림15. Long-reads, short-reads, long-mate pair reads을 이용한 진균류 참조유전체 서열 조립 및 평가 방법]

- **PacBio SMRT sequencing 데이터** 50x (분석 대상 유전체 크기의 50배에 상응하는 데이터 생산량)를 기준으로 contig assembly를 수행함 (HGAP3/FALCON 이용). Scaffolds의 완성도를 높이기 위해 100x 이상의 **long-mate pair reads 데이터** (SSPACE 이용)와, gap filling을 위해 50x 이상의 **whole-genome short-insert reads 데이터** (GapCloser 이용)를 활용함. 생성된 Scaffold 데이터는 whole-genome short-insert reads 데이터에 의해 error correction을 수행함

- 참조유전체 서열 조립 시 정확도 확보를 위해 **BioNano physical mapping 데이터**를 활용하여, 스캐폴드 데이터와 physical mapping 데이터 사이의 비교를 하여 어셈블리 에러를 확인할 수 있음

2) 누룩 참조유전체의 서열 조립 정확도 확보

○ 2종의 누룩 참조유전체 (*Saccharomycopsis fibuligera* KPH12와 KJJ81) 서열 조립을 위해 위에서 언급된 workflow 적용함 (**Choo and Hong et al. Biotechnology for Biofuels. 2016. 9:246**)

[표6. 참조유전체 서열 조립을 위해 사용된 유전체 서열데이터양]

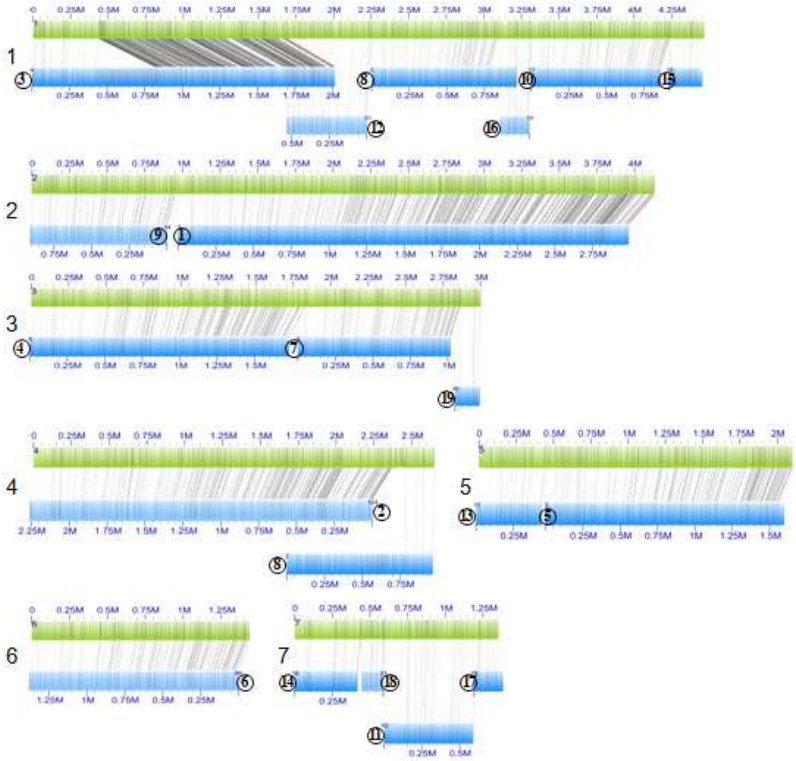
	KPH12	KJJ81
Short-insert paired-end reads (500 bp)		
Raw reads (No.)	24,132,834	22,897,290
Total bases (coverage, X)	122.91	59.34
Q20 bases (%)	88.86	86.23
Long-mate paired-end reads (15 kb)		
Raw reads (No.)	44,904,808	54,748,720
Total bases (coverage, X)	228.91	141.89
Q20 bases (%)	87.14	87.75
TSLRs		
Long reads (No.)	94,757	80,922
Average length of reads (bp)	4,806	5,016
Total bases (coverage, X)	23.19	10.52
Long SMRT sequencing reads		
Long reads (No.)	207,691	388,982
Average length of leads (bp)	8,955	9,226
Total bases (coverage, X)	94.73	94.44

[표7. 파이프라인 적용 어셈블리 결과]

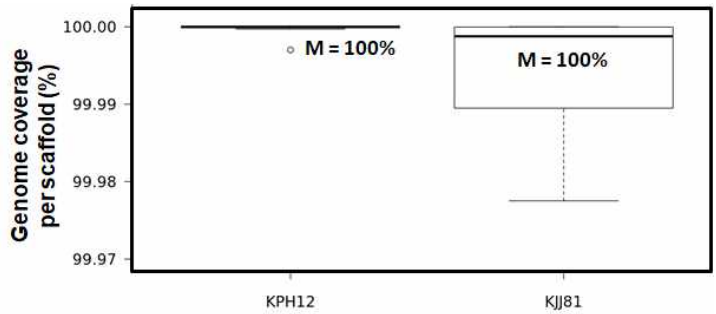
	KPH12	KJJ81
Contig assembly by long reads		
Contig number	18	45
Total length (bp)	19,193,493	37,774,778
N50	3,000,789	1,903,357
N90	1,353,387	721,347
Longest	4,132,216	3,922,141
Scaffolding and gap filling		
Scaffold number	7	14
Total length (bp)	19,567,216	38,516,460

○ **누룩 참조유전체의 서열 조립 정확도 확보:** 기본적으로 PacBio SMRT sequencing 데이터를 토대로 scaffolds을 만들었고, BioNano physcal mapping을 통해 어셈블리 ordering을 확인함. 또한 다른 technology platform (Illumina TSLR)을 이용하여 서열의 정확도를 비교함 (**Choo and Hong et al. Biotechnology for Biofuels. 2016. 9:246**)

- **BioNano physcal mapping**을 통해 어셈블리의 ordering이 상당히 정확함을 확인함. 아래 그림에서 녹색바는 scaffolds을, 파란색바는 BioNano physical mapping 데이터를 나타냄



- Scaffolds에 얼라인먼트된 short-insert reads 데이터의 genome coverage: Replicates의 중간값이 genome coverage가 100%로 평가됨



- Scaffolds 정확도 평가: PacBio SMRT sequencing 데이터 어셈블리 결과를 기준으로 하여 Illumina TSLR 결과와 비교됨. QUAST를 이용하여 평가를 수행함. TSLR 보다 PacBio long-reads 데이터를 이용한 어셈블리 정확도가 뛰어남 (contig 및 scaffolds 개수 및 길이, coverage 등)

[표8. QUAST를 이용한 PacBio SMRT sequencing 데이터 어셈블리 결과를 토대로 한 Illumina TSLR 어셈블리 결과와의 비교]

Metric	Value	
	KPH12	KJJ81
No. of contigs (≥ 1 kb)	292	1,025
Total length (bp)	19,483,087	37,860,684
Physical coverage (%)	99.62	98.37
Longest contig length (bp)	614,411	446,381

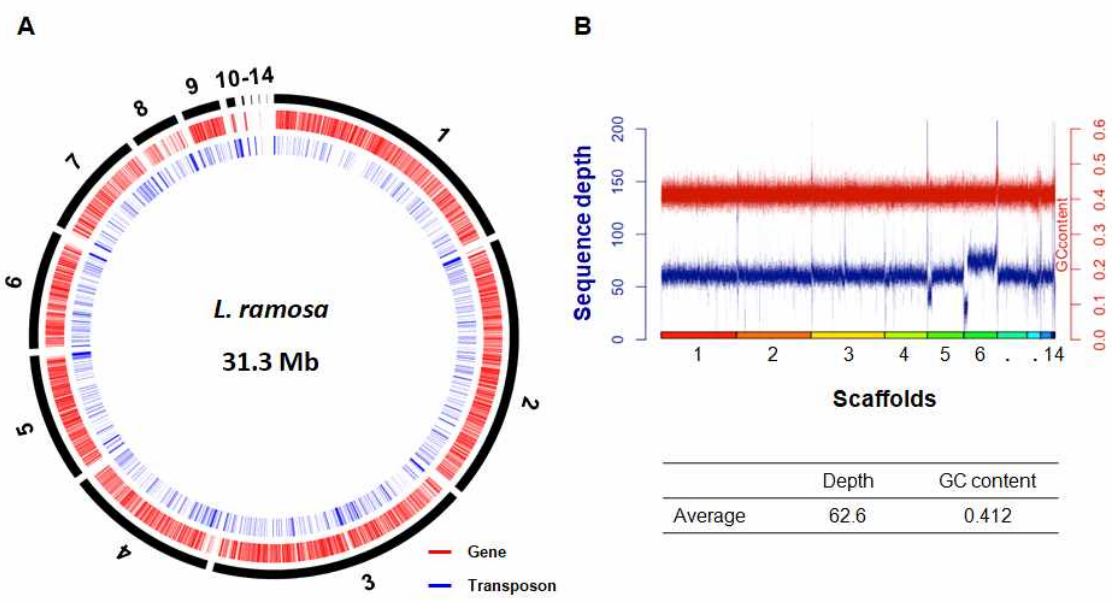
N50 contig length (bp)	157,627	71,542
N90 contig length (bp)	43,980	19,473
Contig GC content (%)	38.07	38.41
Genome fraction (%)	96.368	96.559
Duplication ratio	1.031	1.019
NA50	156,180	70,317
LA50	42	170
Mismatches per 100 kb	6.05	12.2
InDels per 100 kb	15.45	14.74
Ns per 100 kb	0.05	0.05
Fully unaligned contigs (No.)	1	4
Fully unaligned length (bp)	1,002	2,373
Partially unaligned contigs (No.)	11	6
Fully unaligned length (bp)	89,551	74,561

○ Merged whole-genome assembly 파이프라인 개발

- 해독 데이터의 다양성과 각 데이터의 특성 차이를 효과적으로 조합하여 최적의 선도 게놈을 완성하는 merged assembly 파이프라인 개발함. *Lichtheimia ramosa* 유전체 서열 조립을 위해 파이프라인 적용 조립 결과 contig 숫자가 PacBio 대비 70% 정도 감소하여 14래로 줄어들었으며, 했으며, N50이 약 70%증가하였음

[표9. Hybrid assembly에 의한 *L. ramosa* 유전체 서열 조립 결과]

	No. of sequences	Total bases (bp)	N50 (bp)	N90 (bp)	Longest (bp)
PacBio Assembly					
FALCON (PacBio)	41	31,311,778	3,276,235	2,226,143	5,998,895
TSLR Assembly					
CA (TSLR)	257	32,199,549	588,350	150,645	3,190,457
CA + SSPACE (TSLR + PE + MP)	122	32,213,623	1,732,988	596,778	3,721,959
Merge Assembly					
HaploMerger	23	31,251,992	3,202,655	2,385,506	6,009,921
SSPACE (PacBio + TSLR + PE + MP)	**14**	**31,271,552**	**5,850,932**	**2,385,506**	**6,009,921**



[그림 16. *L. ramosa* 유전체 서열의 구조. (A) 유전자 및 transposons의 밀집도. (B) Scaffolds의 sequence depth와 GC 함량]

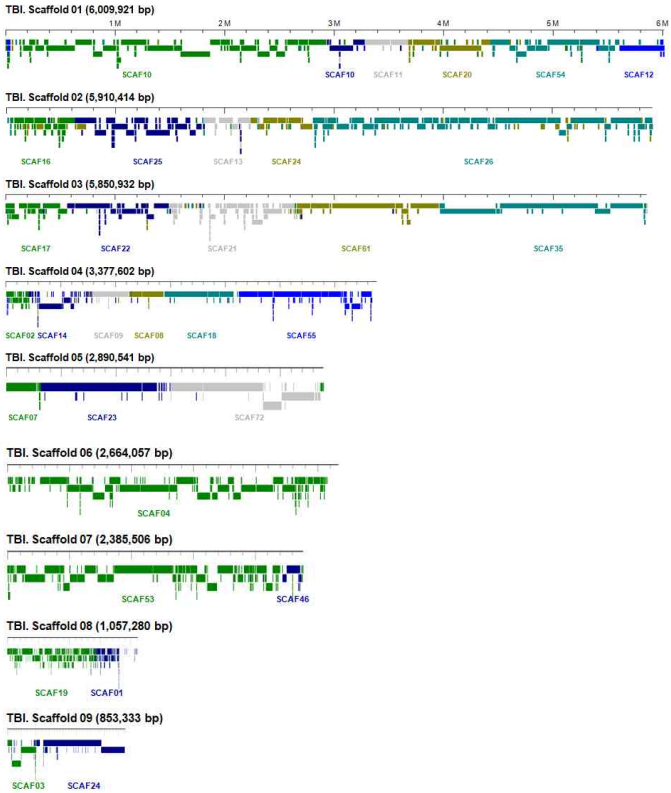
- HaploMerger의 적용의 경우, heterozygous nucleotides의 correction을 수행함으로써, n+n 또는 이형접합성으로 인한 어셈블리 에러를 방지해주는 효과를 확인함

- 서술된 Merged whole-genome assembly의 경우, (1) physical mapping 부재, (2) n+n 또는 이형접합성 발생, (3) 염색체 수준의 scaffolds 생성의 어려움에 직면했을 경우 매우 효과적임

- 기존에 보고된 *L. ramosa* 유전체 스캐폴드 데이터 (Linde *et al.* (2014))와 퀄리티 비교: Linde et al. (2014)에 의해 보고된 스캐폴드는 Illumina sequencing 데이터에 의해 생성된 것으로 총 74개의 스캐폴드들 (N50: 1.22 Mb)로 구성되어 있음. Linde et al. 결과들과 본 연구 결과와 비교되었을 때, Linde et al.에 의한 대부분의 긴 스캐폴드들이 본 연구를 통해 생성된 9개의 스캐폴드들 (아래 그림에서 'TBI'에 해당하는 스캐폴드)에 anchoring됨 (드노보 어셈블리 퀄리티 향상 결과)

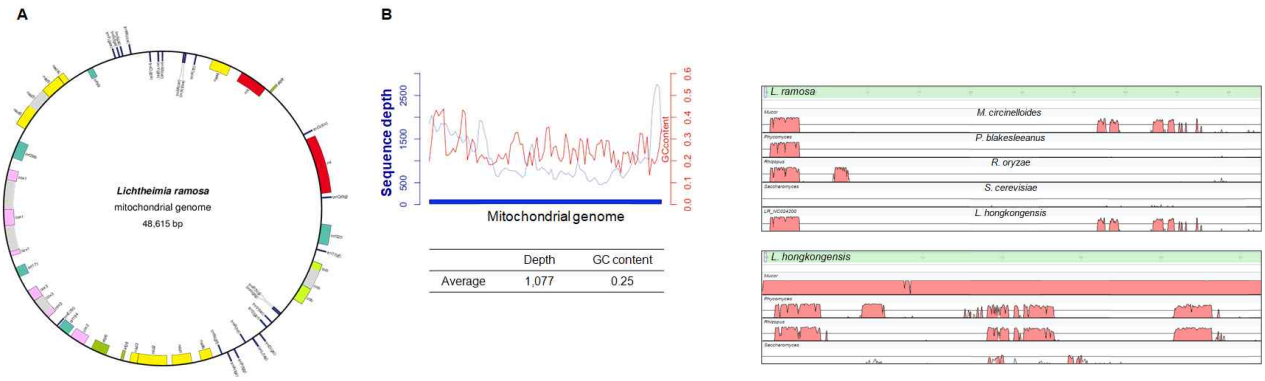
[표10. *L. ramosa* 유전체 스캐폴드 데이터의 퀄리티 개선]

본과제팀	HK (Linde et al. 2014)
- 14 scaffolds	-74 scaffolds
- 31.27 Mb	- 30.71 Mb
- N50: 5.85 Mb	- N50: 1.22 Mb
- 12,827 genes	- 11,510 genes
* 14 vs. 74 : Genome coverage 89% * Conserved blocks: 98.45% identity	
- PacBio/Illumina (TSLR, short reads) 플랫폼 데이터 사용	- 454/Illumina (short reads) 플랫폼 데이터 사용



[그림17. 과제를 통해 조립된 서열과 Linde *et al.*에 의해 조립된 서열과의 비교]

- Mitochondrial genome 서열 결정 (48,615 bp; 아래 그림에서 첫 번째) 및 기존에 보고된 *Lichtheimia* 종 (*L. hongkongensis* (NCBI GenBank Accession no. NC_024200; 31,830 bp)의 서열 교정 (교체 수준임; 아래 그림에서 세 번째)



[그림18. Mitochondrial genome 서열의 구조. (A) Mitochondrial genome 내 유전자 구조. (B) 조립된 서열에 대한 sequence depth와 GC 함량. (C) 보고된 근연종과의 비교 및 서열 교정]

(2) 진균류 유전체 발굴을 위한 유전자 예측 및 기능 예측 파이프라인 개발 및 고도화

1) 유전자 구조(모델) 및 기능 예측 파이프라인 개발

○ 진균류 (진핵생물 대상) 유전체 유전자 구조(모델) 예측을 위한 분석파이프라인 개발:

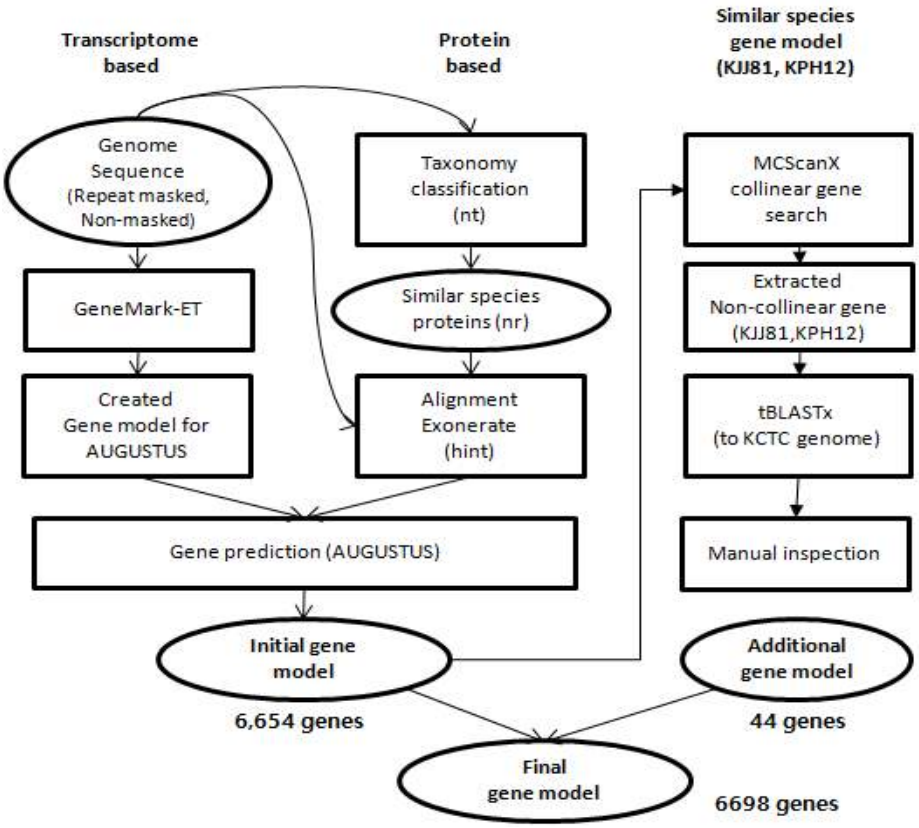
Evidence-driven gene prediction을 기반으로 개발됨

- Transcriptome-based prediction: 게놈 분석종의 RNA-Seq 데이터를 이용하여 유전체 서열 데이터 내 유전자의 엑손-인트론 경계 부위를 확인하는 hint 데이터를 생성함. 우선적으로 GenMark-ET를 이용하여 하나의 training dataset을 생성함 (AUGUSTUS를 이용한 evidence-driven prediction을 위한 데이터셋 생성)

- Protein-based prediction: 상동성이 높은 근연종의 게놈의 단백질 서열을 확보하여 유사 유전자의 엑손-인트론 경계 부위를 확인하는 hint 데이터를 생성함 (Exonerate 이용)

- AUGUSTUS를 이용한 transcriptome- 및 protein-based prediction (Evidence-based prediction) 수행함. 두 종류의 evidence 이외에 AUGUSTUS에 의한 *ab initio* prediction 결과 또한 생성됨

- Manual correction: Tblastx 및 MCScanX를 이용하여 codon usage의 변화로 인한 missing genes을 탐색함. 또한 repeat-masking 부분에서의 gene models 또한 탐색함 (non-repeat masked genomic sequence 데이터 활용함)



[그림19. 진균류의 유전체 서열에서 evidence-driven gene prediction 방법]

○ 누룩 참조유전체의 유전자 구조 예측

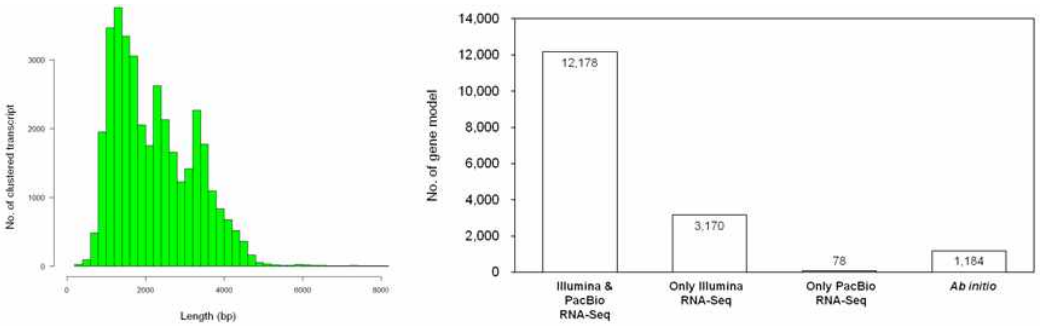
- Evidence-driven gene prediction 방법을 이용한 누룩 유전자 구조 예측 (**Choo and Hong et al. Biotechnology for Biofuels. 2016. 9:246**)

[표11. Evidence-driven gene prediction 방법을 이용한 누룩 참조유전체의 유전자 구조 예측 결과 요약]

	KPH12	KJJ81
Protein-coding gene models (No.)	6,155	12,185
Unique gene models (No.)	6,028	11,966
Genes with isoforms (No.)	127	219
Supported by RNA-Seq (No.)	6,154	12,184
Annotated (No.)	5,435	10,810
Average gene length (bp)	1,788	1,782
Total length of gene models (Mb)	11.00	21.72
Exons		
No. of exons	7,710	15,067
No. of average exons per gene	1.25	1.23
Average exon length (bp)	1,384	1,402
Introns		
No. of introns	1,555	2,882
No. of average introns per gene	0.25	0.23
Average intron length (bp)	213	205

○ Iso-Seq을 이용한 유전자 구조의 정확도 향상 (**Park et al. Data in Brief. 2017. 15:454-458**)

- PacBio transcriptome long-reads 데이터는 **full-length transcripts 생성 가능성**을 높여 줌. 따라서 유전자 구조, 특히 엑손-인트론 구조 예측 시 그 경계를 보다 더 명확히 해주기 때문에 유전자 구조 예측 시 매우 효과적이고, full-length transcripts를 이용한 정확도 검증 결과를 제시해 줄 수 있음. 그러나 유전체 내에서 작은 사이즈의 유전자들에 대한 gene coverage가 부족해질 수 있기 때문에 RNA-Seq과 함께 활용할 것을 추천함. 아래 그림에서 왼쪽은 PacBio transcriptome long-reads를 이용한 unigenes의 크기 분포이고, 오른쪽은 유전자 구조 예측 후 전사체 데이터에 의한 gene model coverage를 나타냄. 또한 아래의 표는 PacBio transcriptome long-reads와 Illumina short-reads를 이용한 유전자 예측 시 correction된 유전자 모델 유형을 나타냄



[그림20. PacBio Iso-Seq 데이터를 이용한 full-length transcripts의 길이 분포 및 유전자 구조 예측 시 evidence의 support]

[표12. 표고버섯에서 correction된 유전자 모델의 개수]

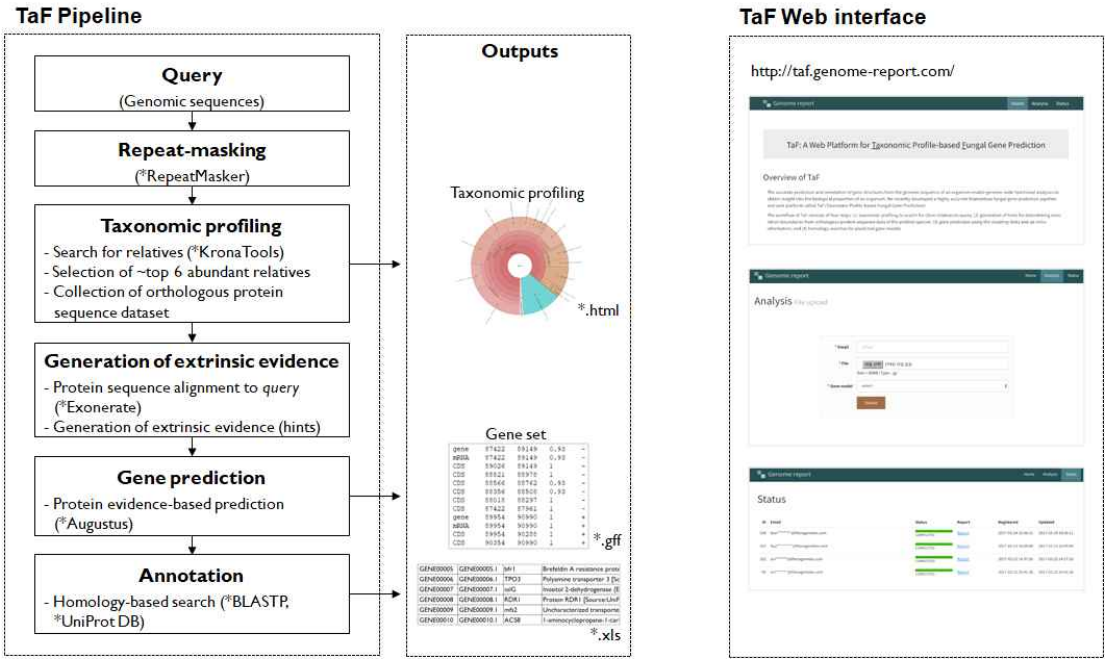
	No. of gene models
Exactly overlapped	7,889
Split into ≥ two gene models	4,742
Fused with ≥ two gene models	343
Structurally re-predicted	261
Newly found	1,344
Predicted in the only previous study	2,031

- 그러나 좋은 전사체 데이터에 의한 gene model coverage가 좋다고 할지라도, full-length transcripts에 의해 cover되는 유전자는 42.2%임 (유전자 구조의 정확도를 나타내는 지표)

2) TaF 서버를 통한 진균류 유전자 모델 예측

○ TaF (Web Platform for Taxonomic Profile-based Fungal Gene Prediction) 서버 개발 (**Park et al. 2018. Submitted**)

- TaF는 **taxanomic profiling**된 종의 orthologous protein 서열로부터 엑손-인트론 경계점을 결정하기 위한 **extrinsic evidence hint들을 생성**시켜 진화적으로 먼 orthologous 데이터의 사용으로 인한 유전자 구조 예측 시 false-positive prediction을 방지해 주는 기능을 함. 따라서 TaF는 새롭게 해독된 또는 전혀 분석되지 않은 fungal genomes로부터 유전자 예측을 위해 매우 효과적임. 또한 분석 대상종의 전사체 데이터가 없을 경우 비교유전체학을 위해서도 효과적임



[그림21. TaF의 workflow 및 web interface]

- TaF의 workflow (4가지 단계로 구성됨)

▪Query genome에 대한 근연종 탐색을 위한 **taxonomic profiling**

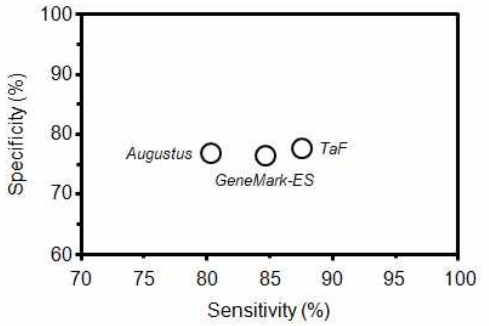
▪프로파일된 종의 orthologous 데이터로부터 엑손-인트론 경계점들을 결정하기 위한 **extrinsic**

evidence hint들을 생성

▪Hint와 ab initio 정보를 이용한 **유전자 구조 예측**

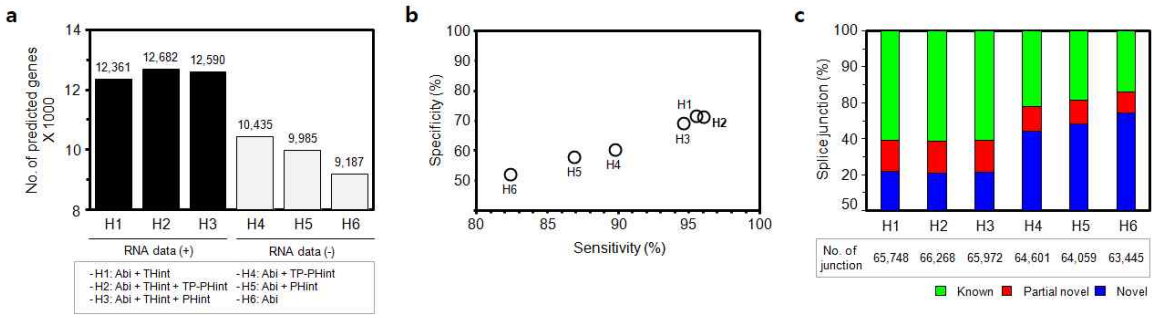
▪**상동성 검색** 기반의 gene models에 대한 기능 예측

- TaF의 정확도 평가 (1): TaF, Augustus, GeneMark-ES의 민감도 (Sensitivity (Sn))와 특이도 (specificity (Sp)를 계산 후 비교 평가함. Sn와 Sp는 Augustus는 80.33%와 76.86%, GeneMark-ES는 84.67%와 76.41%, TaF는 87.60%와 77.58%를 보임



[그림22. aF, Augustus, GeneMark-ES의 sensitivity와 specificity의 비교 평가]

- TaF의 정확도 평가 (2): 서로 다른 방법들 {ab initio (Abi), transcriptome-based prediction (THint), taxonomic profiling을 적용한 homologous protein-based prediction (TP-PHint), taxonomic profiling을 적용하지 않은homologous protein-based prediction (PHint)} 적용 시 Sn 및 Sp 비교함. 즉 (1) Hint (H)1: Abi+ THint, (2) H2: Abi + THint + TP-PHint, (3) H3: Abi + THint + PHint, (4) H4: Abi + TP-PHint, (5) H5: Abi + PHint ,(6) H6: Abi 등의 방법이 비교됨. TaF의 성능은 전사체 데이터 활용했을 경우, 또는 전사체 데이터 없이 상동성-기반의 예측했을 경우 정확도가 가장 높게 평가됨 (아래 그림 a: 유전자 예측 개수, b: Sn 및 Sp 정도, c: 유전자의 known splice site의 개수를 나타냄)



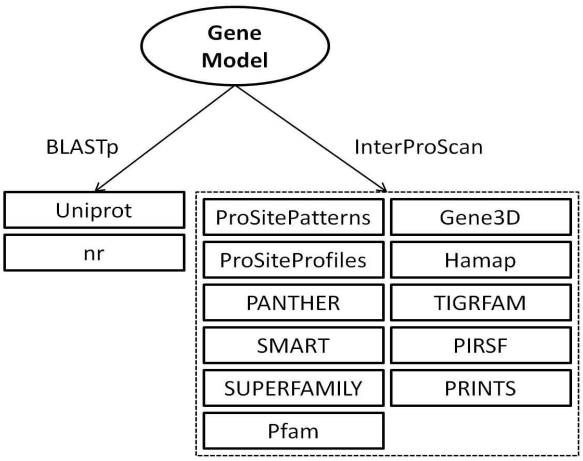
[그림23. Evidence 적용 방식에 따른 TaF의 정확도 평가 결과. (A) 예측된 유전자 모델의 개수. (B) Sensitivity와 specificity의 비교 평가. (C) 예측된 splice sites의 평가]

○ 추후 개발 방향: Evidence의 강화를 위해 분석 대상종의 transcriptome data (i.e. RNA-Seq data) 적용을 통해 gene prediction의 예측율 및 정확도를 더 높일 수 있는 분석과정을 추가할 계획임

3) 유전자 기능 예측 파이프라인

○ **상동성 검색 기반**의 유전자 기능 예측 파이프라인 개발

- 예측된 유전자 모델들은 크게 BLAST 및 InterProScan을 토대로 검색됨. BLASTP (단백질 서열을 이용한 상동성 검색)는 UniProt과 NCBI NR 데이터베이스를 검색함. InterProScan은 sequence retrieval system (SRS) 방식을 통해 다양한 데이터베이스로부터 단백질 도메인 검색 수행함. 특히 UniProt 데이터베이스 검색을 통해 known homologous **gene names**이 예측되고, InterProScan 결과는 **GO annotation과 link**되어 있음



[그림24. 상동성 검색 기반의 기능 예측]

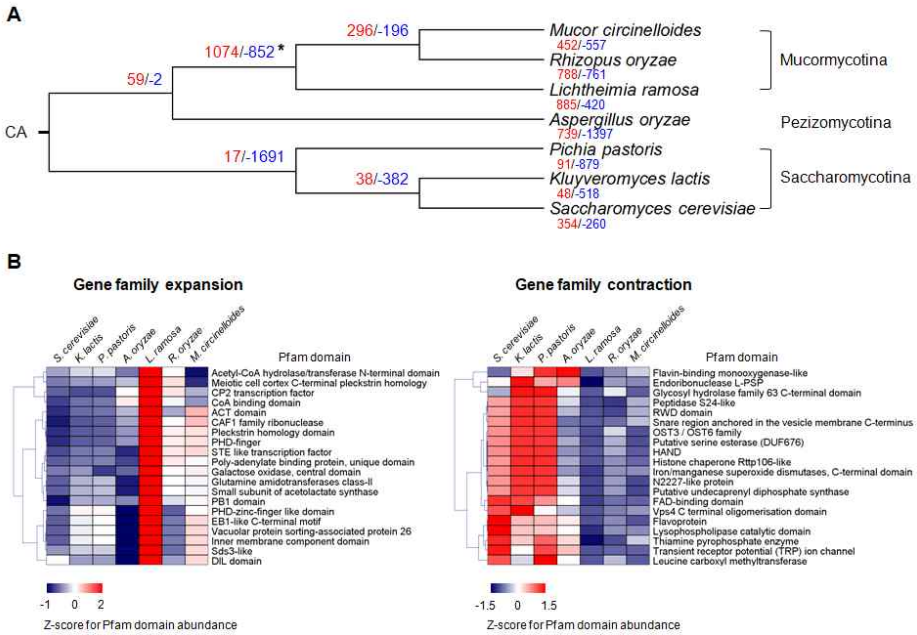
4) 해당 분석 파이프라인을 활용한 *Lichtheimiaramosa* genome annotation

○ *L. ramosa*의 genome annotation

[표13. 본 과제를 통해 개발된 유전자 구조 및 기능 예측 파이프라인에 의한 *L. ramosa* genome의 annotation]

Feature	Value
Protein-coding gene models (No.)	12,827
Supported by RNA-Seq (No.)	12,622
Annotated (No.)	12,615
Average gene length (bp)	1,469
Total length of gene models (Mb)	19.49
Exons	
No. of exons	68,489
No. of average exons per gene	5.16
Average exon length (bp)	227
Introns	
No. of introns	55,221
No. of average introns per gene	4.16
Average intron length (bp)	71

○ 다른 진균류 유전체와 비교를 통한 *L. ramosa*의 evolutionary relationship (A)과 gene family expansion 및 contraction



[그림25. *L. ramosa*의 gene family expansion 및 contraction]

- Gene family expansion: **아세테이트 관련 효소의 풍부성**

Carbohydrate metabolism (acetyl-CoA hydrolase/transferase, CoA binding domain, ACT domain, small subunit of acetolactate synthase, galactose oxidase, and glutamine amidotransferase class II), cellular signaling (pleckstrin homology domain, PB1 domain, and inner membrane component domain, EB1-like C-terminal motif), binding (PHD-finger, STE domain, sds3-like, poly(A)-binding protein), membrane morphogenesis during sporulation (vacuolar protein sorting-associated protein 26)

- Gene family contraction: **Flavonoid 관련 효소들의 억제**

Flavonoid-related proteins (flavin-binding monoxygenase, FAD-binding domain, flavoprotein), endoribonuclease L-PSP, peptidase S24-like, glycosyl hydrolase family 63 C-terminal domain, RWD domain, snare region anchored in the vesicle membrane C-terminus, OST3/OST6 family, serine esterase, HAND, histone chaperone Rttp 106-like, iron/maganese superoxide dismutases, N2227-like protein, undecaprenyl diphosphate synthase, vps4 C terminal oligmerisation domain, lysophospholipase catalytic domain, thiamine pyrophosphate enzyme, TRP ion channel, leucine carboxyl methyltransferase

5) Orthologous gene 클러스터링 파이프라인 개발

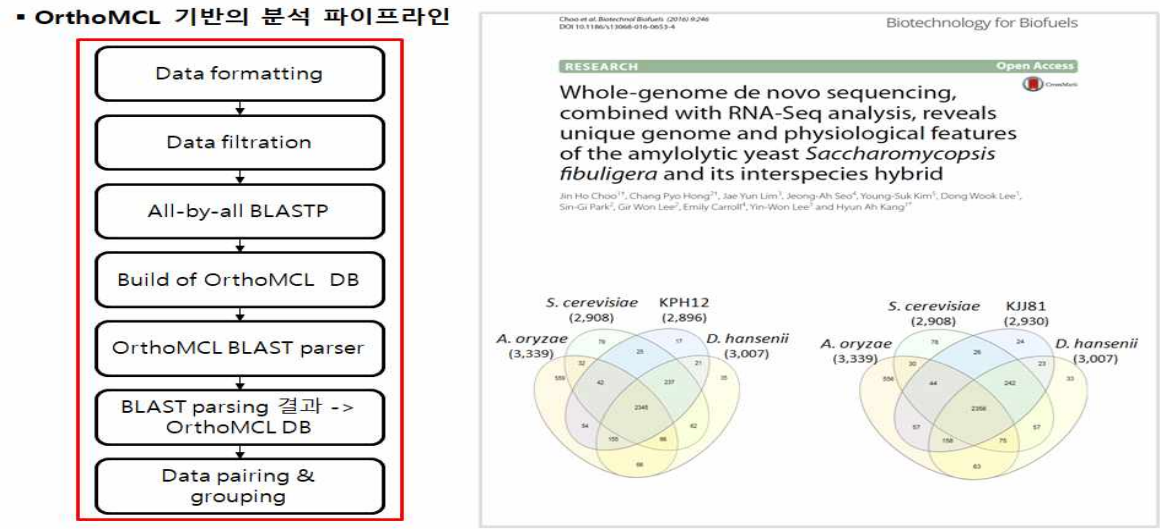
○ Orthologous gene 클러스터링 파이프라인 구성 개요

- (1) OrthoMCL input data의 포맷팅, (2) Input data의 filtration, (3) All-by-all BLASTP search, (4) OrthoMCL DB 생성, (5) OrthoMCL BLAST parser, (6) 생성된 DB에 BLAST parsing 결과를 import시킴, (7) 결과 데이터 페어링 및 그룹핑

- Orthologous gene들을 효과적으로 탐색하기 위해 이전에 수행된 '진균류 근연종 프로파일링' 방

법을 적용시켰을 수 있음

○ 개발된 파이프라인을 이용하여 아래와 같이 *S. cerevisiae, Saccharomycopsis fibuligera, A. oryzae, D. hansenii*에서 orthologous gene cluster 수행함 (Choo, Hong et al. 2016. Biotechnol Biofuels. 9:246)



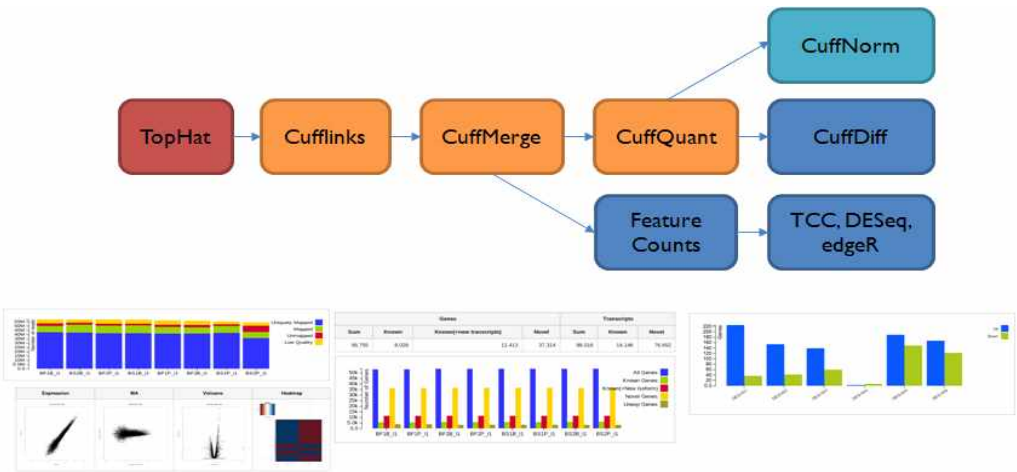
[그림26. Orthologous gene 클러스터링 파이프라인의 workflow 및 실제 분석 사용 사례]

(3) 진균류 전사체 분석을 위한 유전체 정보 기반의 전사체 분석 파이프라인 개발

1) 유전체 정보 기반의 전사체 분석 파이프라인 개발

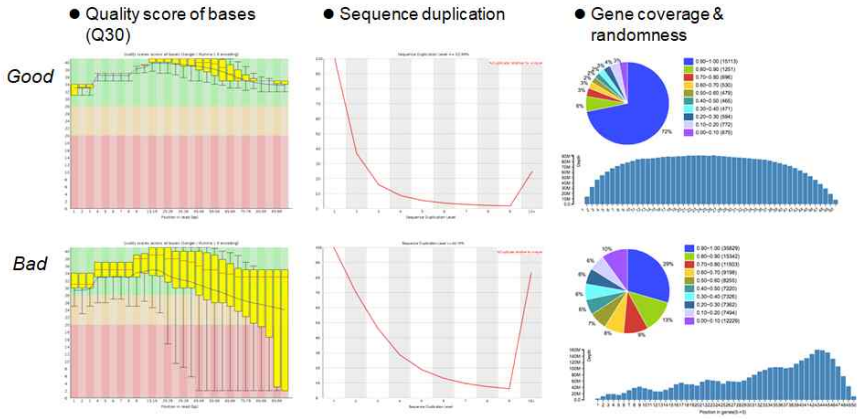
○ Tuxedo Protocol을 이용한 유전체 정보 기반의 전사체 분석파이프라인 개발

- Trapnell et al. (2012)에 의해 개발된 tuxedo protocol을 진균류 유전체에 최적화시켜 파이프라인을 개발함



[그림27. Tuxedo Protocol을 이용한 RNA-Seq 분석파이프라인 및 output 예제]

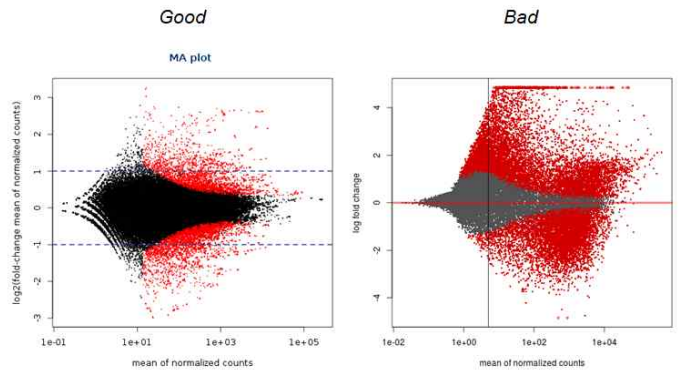
- [1] **Data QC**: RNA-Seq reads의 base quality (Q30 이상), sequence duplication (cDNA 및 시퀀싱 라이브러리 제작 시 과도한 PCR amplification 유무), reads의 genome coverage (유전자 coverage 포함), reads의 무작위 정도 (유전자 부위에 걸쳐 일정하게 분포되어 있는지 유무) 등이 체크됨



[그림28. RNA-Seq data quality control (QC) output 예제]

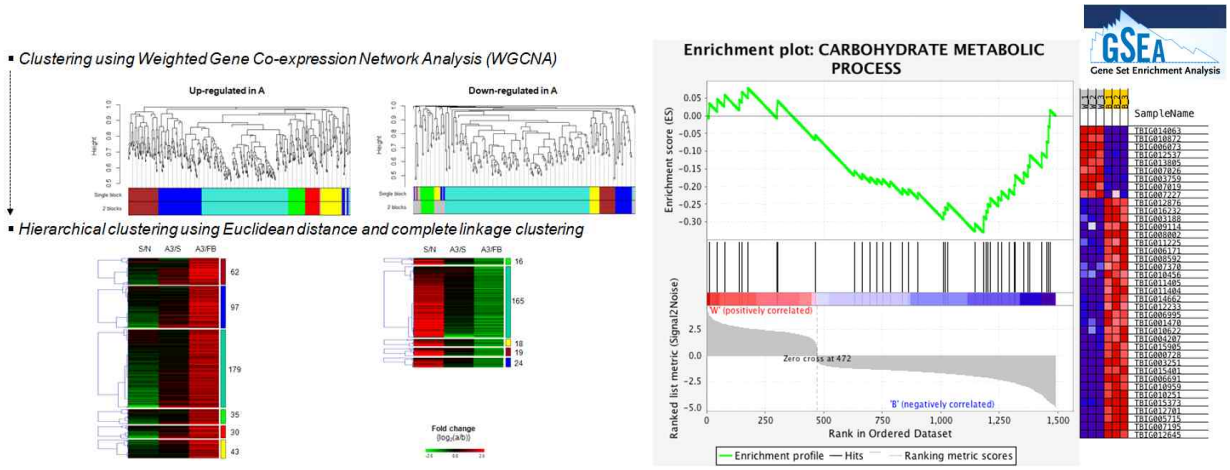
- [2] TopHat을 통해 RNA-Seq reads을 레퍼런스에 맵핑시키고, Cufflinks을 이용하여 전사체를 어셈블리 후 유전체 및 유전자에 대한 전사체 구조를 정립시킴. 이를 토대로 **유전자의 발현량**이 결정됨. FPKM을 이용한 발현량의 normalization

- [3] 차등발현 유전자 탐색 (**Differentially expressed genes 탐색**): CuffDiff를 기본적으로 이용하여 DEGs을 분석함. 아래 그림의 왼쪽처럼 MA plo에서 up-/down-regulation된 유전자들의 균형이 조화롭게 되는지 판단하고 DEGs에 대한 foldchange 및 p-value를 계산함



[그림29. DEG 평가를 위한 MA plot 예제]

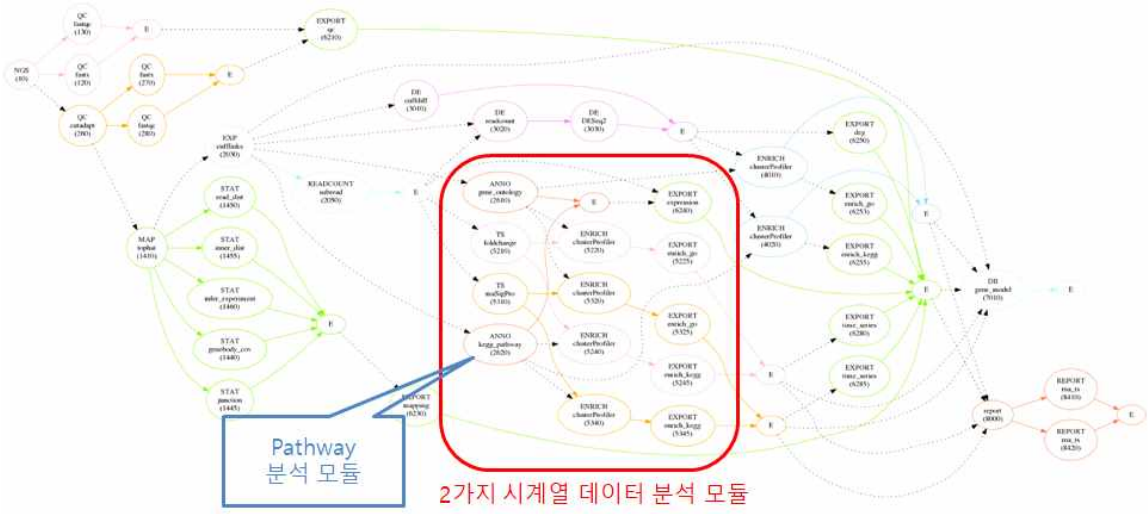
- [4] 분석된 DGEs에 대한 **clustering 및 GO enrichment**를 수행함: WGCNA를 이용한 유전자 발현에 대한 클러스터링 (파이프라인 이외 분석모듈; 아래 그림에서 위쪽)과 GO enrichment (파이프라인 이외 분석모듈; 아래 그림에서 아래쪽). GO enrichment 분석을 통해 input DEGs이 특정 유전자 기능 카테고리에서의 단순 hit율(또는 풍부성)을 체크하는 것이 아니라 그 카테고리를 대표하는 특정 유전자와의 hit율 및 확률을 나타냄



(WGCNA를 이용한 유전자 발현에 대한 클러스터링) (GO enrichment 분석)

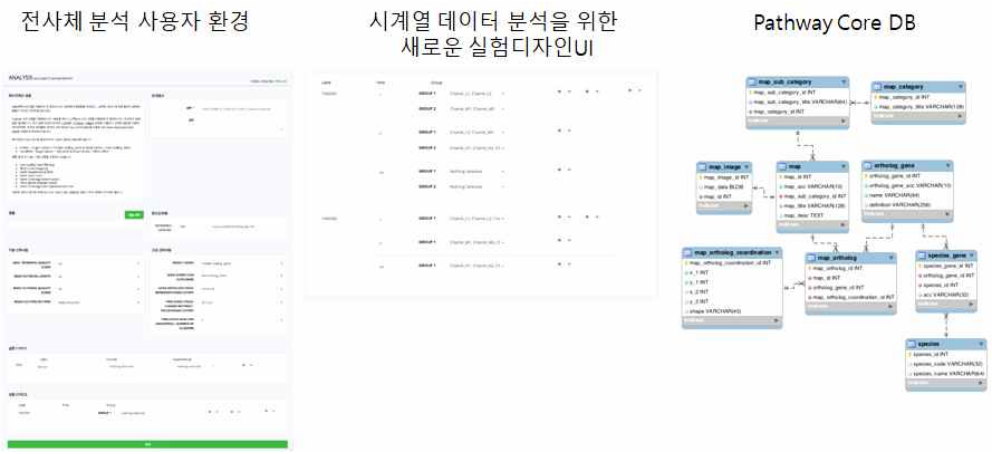
[그림30. 분석된 DEGs에 대한 클러스터링 및 GO enrichment 분석 결과 예제]

○ 전사체 분석 파이프라인 강화: KEGG pathway core DB 검색과 시계열 분석이 추가되었고 사용자 환경이 개선됨



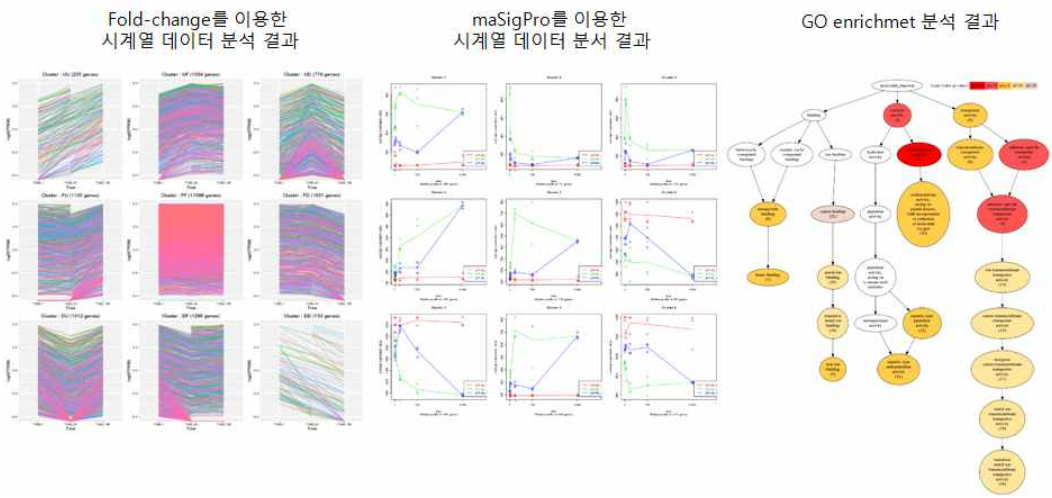
[그림31. 개발된 RNA-Seq 파이프라인에 KEGG pathways 및 시계열 분석 추가]

- 사용자 환경 개선과 시계열 분석 및 KEGG pathway core DB 구축



[그림32. KEGG pathways 및 시계열 분석을 위한 사용자 환경 개발]

- 시계열 분석방법과 결과 예시



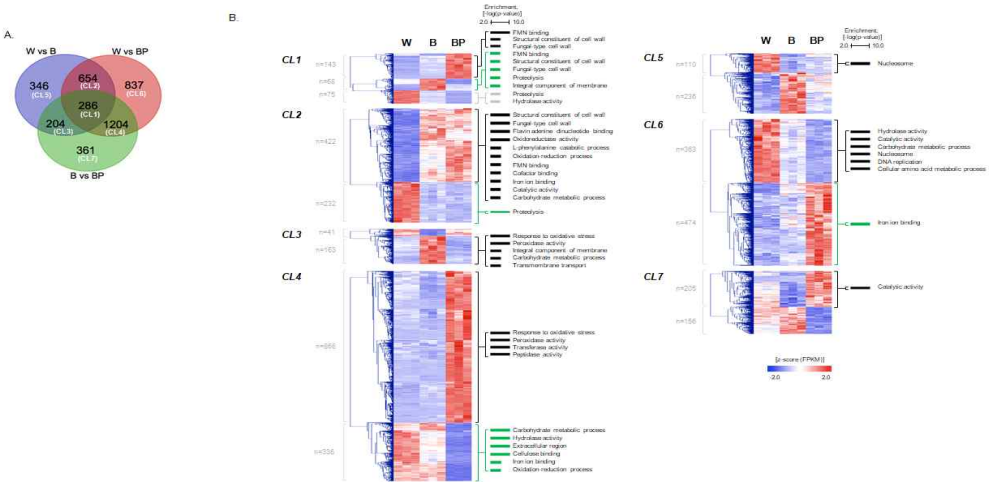
[그림33. Fold-change 및 maSigPro를 이용한 시계열 분석방법과 결과 유전자들에 대한 GO enrichment 분석 결과]

2) 표고버섯의 균사체의 갈변화 관련 전사체 분석

○ **표고버섯 균사체 갈변화(browning)**는 균사체의 영양생장에서 생식생장으로 발달하기 위한 중요한 표현형질 중의 하나로, 빛에 의한 조절이 필수적임. 표고버섯 균사체 갈변화 전사체 분석을 통해 본 과제에서 개발된 파이프라인을 이용하여 **진균류 분석을 위한 파이프라인 최적화**시키고자 함. 표현형은 갈변화전 (W), 갈변화 (B), 빛과 암상태의 임의적 노출로 인한 변이체 (BP)를 이용함. 현재 본 연구결과는 논문에 투고 및 심사중에 있음 (Yoo et al. 2018. Submitted)

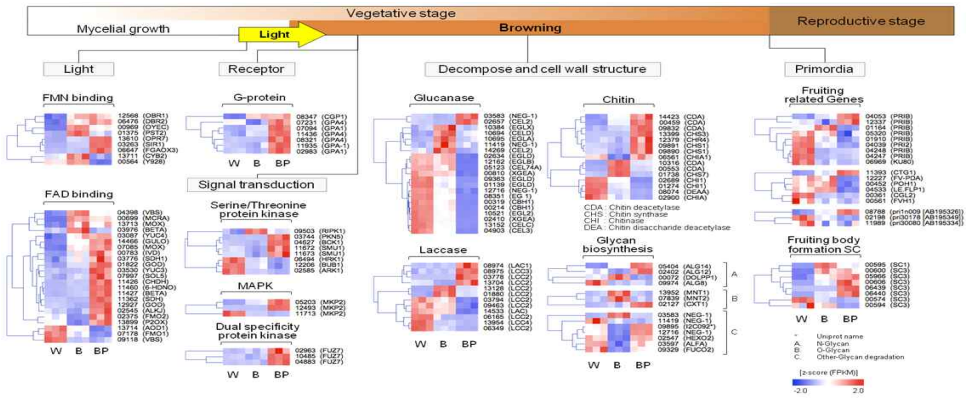
○ W, B, BP 사이의 DEG 분석 결과 및 표현형 특이적 발현 그룹 확인: W vs. B (갈변 관련 변화), B vs. BP (갈변화 변이 관련), W vs. BP 사이의 벤다이어그램 분석을 통해 표현형 특이적 발현 유전자들을 선발 함. 총 7개의 클러스터들에 대한 GO enrichment 분석을 통해 각 클러스터들의 기능적 특성을 분석함. 예를 들어, CL1은 표현형 특이적 발현군, CL2는 갈변화 관련 공통적으로 나타나는 발현군, CL3는 B 표현형 특이적 발현군, CL4는 BP 표현형 특이적 발현군을 나타냄. 각

특이적 발현군의 기능에서 light-sensing, G-protein receptor, Cell wall degradation 및 구조화, Primordia 형성 개시 등과 관련된 유전자들이 확인되었고, B 또는 BP 특이적 발현군들이 잘 나타남



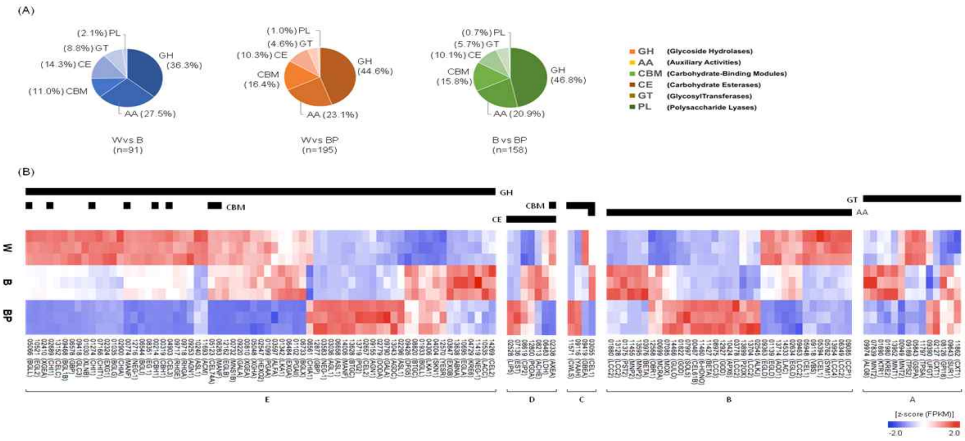
[그림34. 표고버섯에서 갈변화 관련 특이적 발현 유전자 클러스터 및 기능 확인]

○ 전사체 분석을 통해 예측된 균사체 갈변화에 관여하는 요인들 및 유전자 후보군 탐색: 표고버섯의 균사체에서 갈변화와 관련된 light-sensing, G-protein receptor, Cell wall degradation 및 구조화, Primordia 형성 개시 등과 관련된 유전자들을 확인함



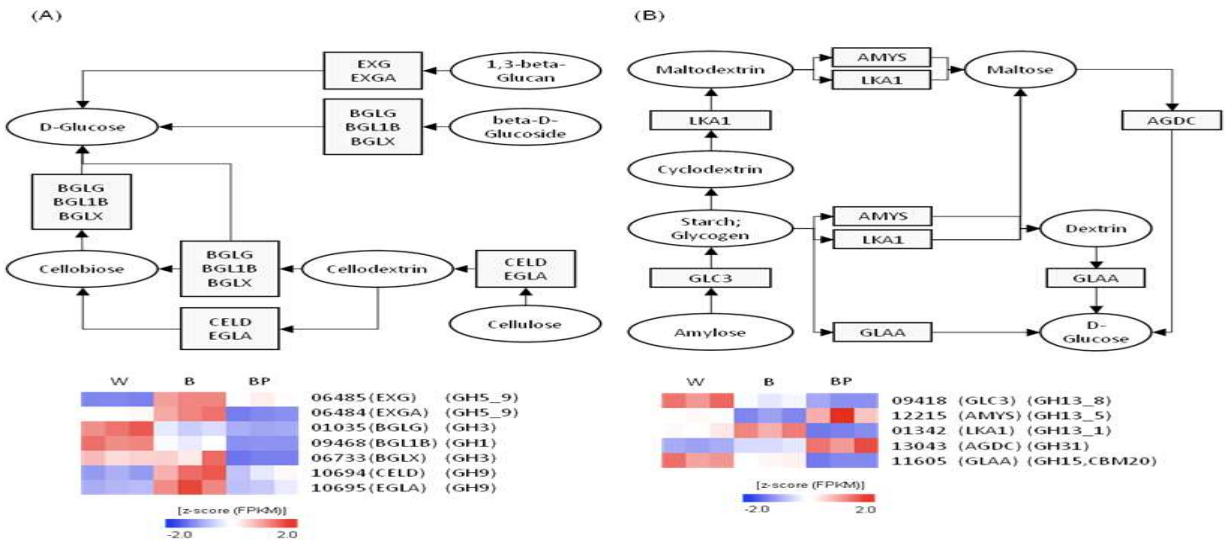
[그림35. 균사체 갈변화에 관여하는 요인들 및 해당 유전자 후보군의 확인]

○ **CAZyme 분석을 통한 GH 그룹 관련 유전자들의 발현 변화 확인**: CAZyme 분석은 glycosidic bond들을 분해, 변형, 생성하는 structurally-related catalytic 및 carbohydrate-binding module들의 효소들 검색해 줌. DEGs 중에서 CAZyme class에 해당하는 유전자들이 유무를 확인함. GH domain (cell wall polysaccharide degradation에 중요한 역할을 함)을 가진 DEGs이 우세적으로 확인됨



[그림36. CAZyme 분석을 통한 GH 그룹 관련 유전자들의 발현 변화 확인]

○ Cell wall degradation 대사경로 관련 유전자 발현의 변화 확인: CAZyme 분석을 통한 GH 그룹 관련 유전자들 중에서 KEGG 패스웨이에서 starch 및 sucrose 대사 관련 (cellulose 및 glucanase 등) 유전자들이 갈변화 시 발현 증가 패턴을 보이는 것으로 확인됨



[그림37. KEGG 패스웨이에서 cell wall degradation 대사경로 관련 유전자 발현의 변화 확인]

○ **진균류에 최적화된 전사체 분석파이프라인**은 표고버섯에서 갈변화 관련 유전자 후보군 탐색 및 분석에 매우 효과적인 결과들을 보여줌

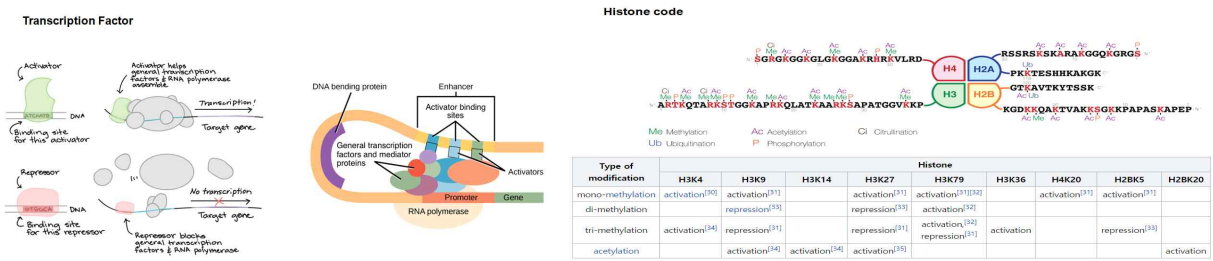
(4) 프로모터 및 전사 인자 분석 파이프라인 개발

1) 프로모터 및 전사 인자 분석을 위한 ChIP-Seq 파이프라인 개발

○ 분석 파이프라인 개요

- 전사인자 (transcription factor, TF) 및 히스톤 변형 (histone modifications)에 의한 발현조절: 크로마틴의 열림 또는 닫힘 구조에 따라 activators 또는 repressors가 프로모터 및 인헨서 부위

에 결합하여 RNA polymerase 유전자의 발현을 조절함. 이때 히스톤들의 변형되어 크로마틴의 구조 변화에 작용함. 히스톤 변형 역시 activation 또는 repression 코드가 정해져 있음

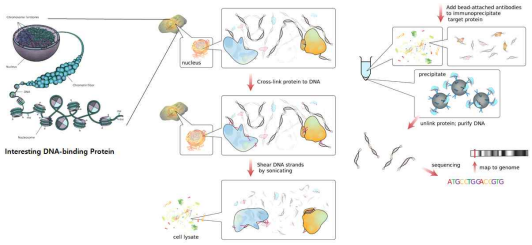


Type of modification	Histone								
	H3K4	H3K9	H3K14	H3K27	H3K79	H3K36	H4K20	H2BK5	H2BK20
mono-methylation	activation[30]	activation[31]		activation[31]	activation[31][32]		activation[31]	activation[31]	
di-methylation		repression[33]		repression[33]	activation[32]				
tri-methylation	activation[34]	repression[31]		repression[31]	activation[32] repression[31]	activation		repression[33]	
acetylation		activation[34]	activation[34]	activation[35]					activation

[그림38. 전사인자 (transcription factor) 및 히스톤 변형 (histone modification)에 의한 유전자 발현 조절]

- 분석 개요

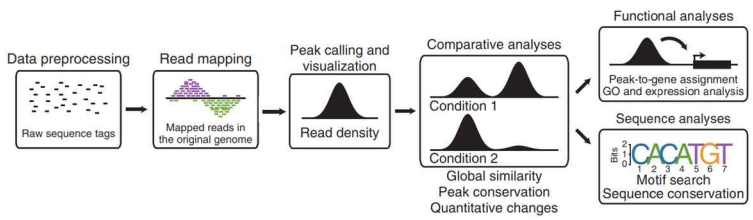
(1) TFs 또는 히스톤 변형 부위에 대한 ChIP 및 시퀀싱 (ChIP-Seq)



[그림39. ChIP-Seq 실험 과정의 개요]

- Cell crosslink (단백질-DNA 복합체)
- Sonication에 의한 DNA 사슬 절단 (약 300-bp)
- TFs 또는 히스톤 변형 부위에 대한 항체를 이용한 immuno-precipitation 수행
- DNA 분리
- 시퀀싱 라이브러리 제작 및 일루미나 HiSeq을 이용한 시퀀싱

(2) ChIP-Seq 분석: ChIP-Seq read 데이터 맵핑 - peak calling - peak annotation - 양적 변화 부위 탐색 (비교 분석) - 바인딩 부위 관련 functional motifs 탐색



[그림40. ChIP-Seq 데이터 분석 과정 개요]

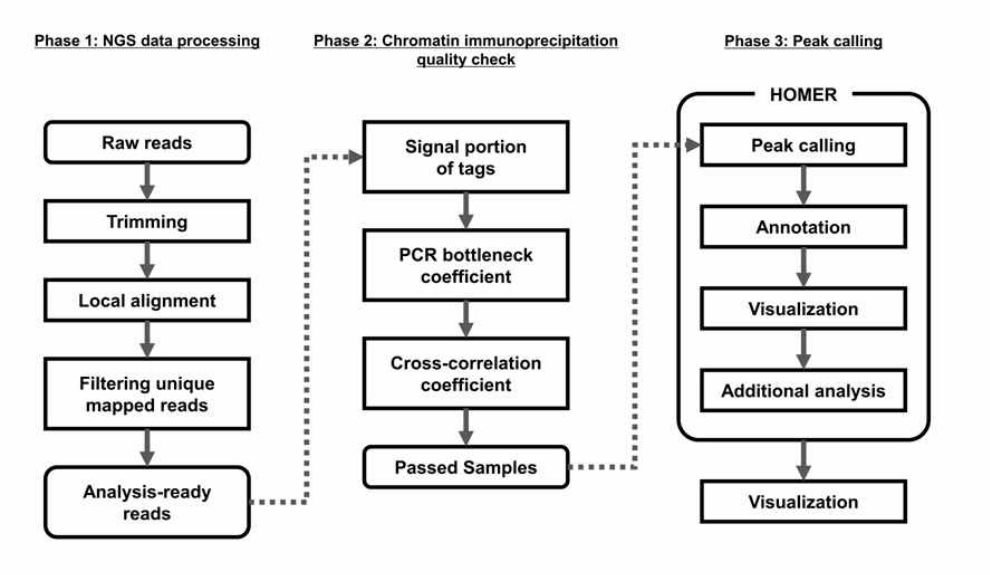
- Preprocessing (data quality 체크)
- ChIP-Seq read mapping
- Peak calling
- 샘플들 사이의 peak 비교를 통해 정확도 및 정량적 평가 수행
- Functional annotation

○ ChIP-Seq 분석 파이프라인 workflow 및 결과

▣ 분석 파이프라인 workflow

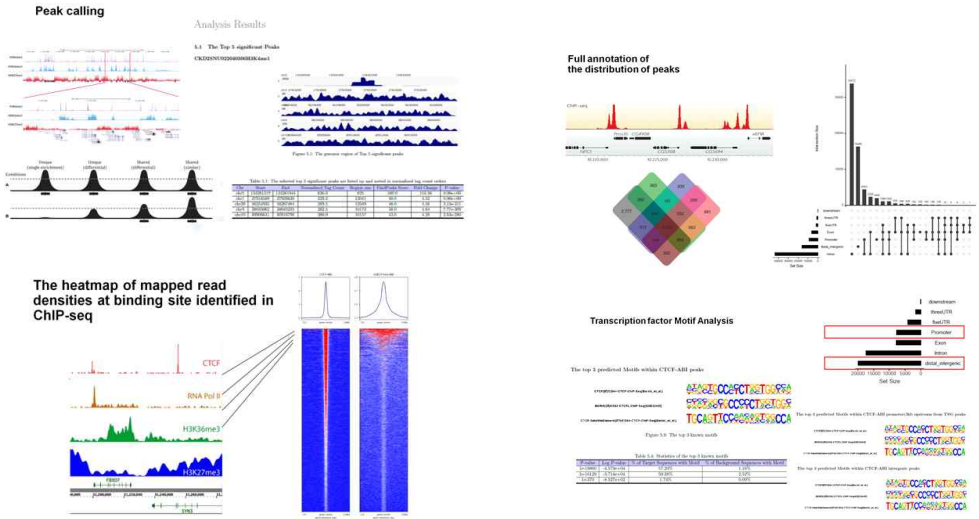
- ChIP-Seq 분석 파이프라인을 개발하기 위해, NCBI SRA 데이터베이스에 저장된 yeast 대상으로 분석된 H3K4me3 및 H3K27me3의 히스톤변형 데이터와 PolII TF 데이터를 이용하여 개발함
- 현재 Gcr1 (glycolysis regulator1으로 해당과정을 조절하는 전사조절인자임)을 대상으로 한 ChIP-Seq 분석을 통해 개발된 파이프라인의 sensitivity 및 specificity를 평가하고 최적화 시킬 예정임
- 단계1은 NGS data processing (데이터 퀄리티 체크, 필터, 맵핑으로 구성), 단계2는 ChIP의 퀄리티 체크를 수행, 3단계는 HOMER 툴을 이용하여 peak calling, annotation, 데이터 시각화 처리

순으로 분석됨



[그림41. ChIP-Seq 분석 파이프라인 workflow]

▣ 분석 파이프라인 레포트 내 결과물



[그림42. ChIP-Seq 분석 파이프라인에 의한 output 예제]

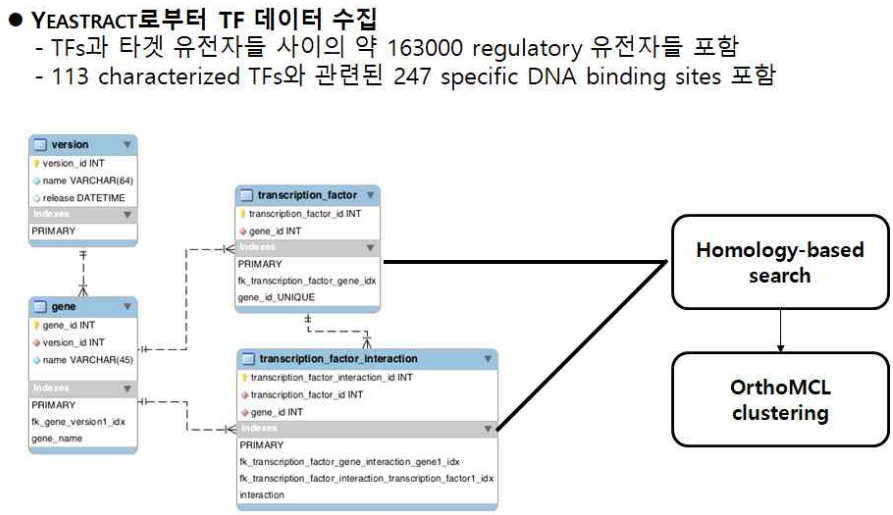
- 레포트 구성은 시퀀싱 결과 요약, peak calling (게놈 브라우저 연동을 통한 peak calling 상태 확인),annotation (유전자 위치를 상대로 peak들의 주요 분포도), 서로 다른 샘플들 사이의 peak overlaps, conserved sequence logo 분석 등을 포함

○ TF 모티프 분석 모듈 개발: Peak 부위들에 대해 sequence conservation 정도를 평가 후, JASPAR (transcription factor binding profiles을 위한 database)를 이용하여 TF binding sites 에서 known functional motifs 또는 *de novo* (unknown) motifs을 검색함. 진균류에서 known motifs을 정의하기 위해, 효모 데이터베이스를 이용한 상동성 검색을 기반으로 함

- Yeast genome으로부터 연구된 TFs에 대한 수집 및 DB 구축 (데이터 출처: Yeastract

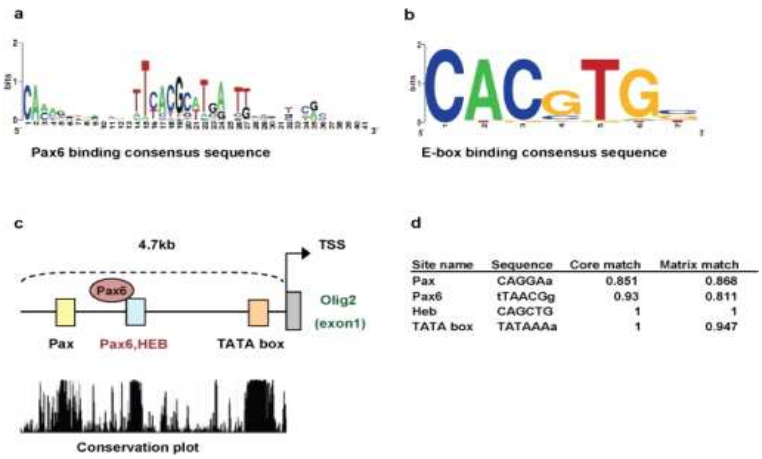
(http://www.yeastract.com/)): Yeast에서 TFs과 타겟 유전자들 사이의 약 163000 regulatory 유전자들 포함. 또한 113 characterized TFs와 관련된 247 specific DNA binding sites들도 포함됨

- 다음과 같은 구조와 함께 Local Yeast TF DB 구축함



[그림43. Yeast TFs 데이터 수집 및 DB 구축]

- 또한 다른 종에서 TFs들을 분석하기 위해, Yeast TFs을 상대로 homology-based TFs을 검색할 수 있는 모듈을 개발함

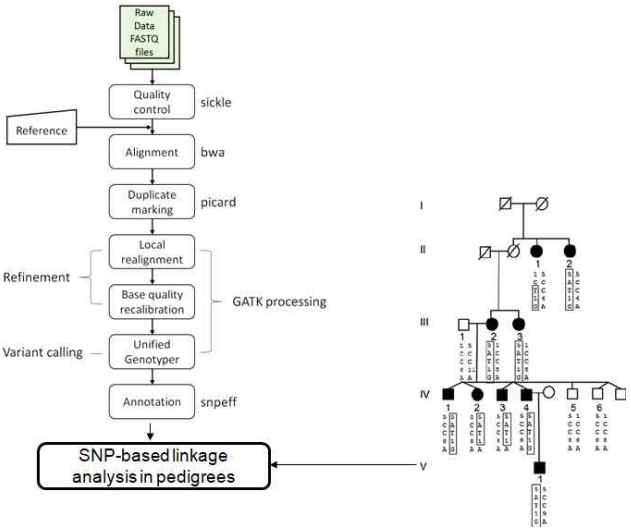


[그림44. Yeast TFs homology-based 검색 결과]

(5) 진균류 유전체 변이 분석을 위한 형질 관련 SNP 탐색 파이프라인 개발

○ 진균류 형질 관련 SNP 탐색 파이프라인 개발 개요

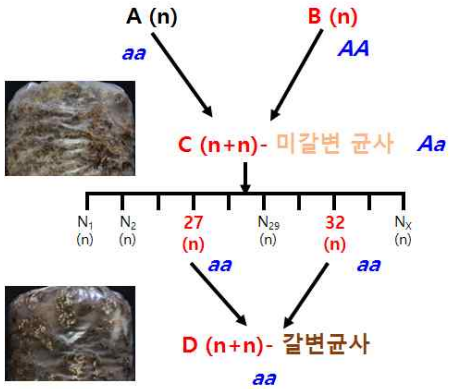
- 특정 표현형에서 우성 및 열성 형질을 나타내는 가계도 형식의 진균류 샘플에서 형질 관련 특이적 SNP를 탐색하기 위한 파이프라인을 개발함. GATK 처리 과정을 통해 분석된 SNPs은 주어진 가계도를 근거로 형질 특이적인 공통성을 가진 SNPs을 탐색하고 유전자 연관성을 확인함 (특히 missense 변이 위주)



[그림45. 가계도 기반의 SNP 분석법]

○ 분석 파이프라인 성능 검증

- 표고버섯 샘플에서 갈변(aa) 및 미갈변(AA, Aa) 형질을 나타내는 가계도 형식의 샘플로부터 DNA 분리 후, P1(#B), F1(#C), F1 단포자체(#27, #32), F2(#D) 샘플에 대해 리시퀀싱(50X 기준) 수행함

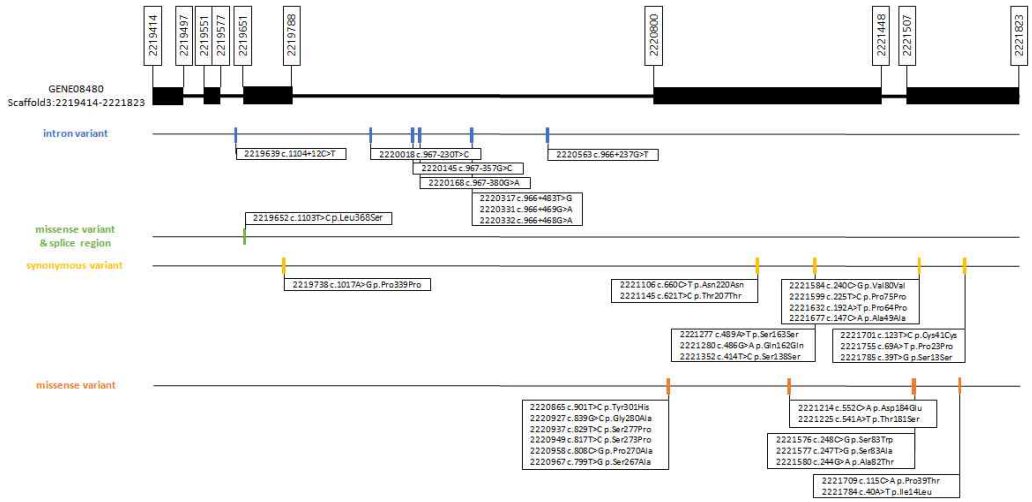


[그림46. 표고버섯에서 갈변(aa) 및 미갈변(AA, Aa) 형질 가계도]

- 우선적으로 갈변(#27, #32, #D) 및 미갈변(#B)을 나타내는 샘플에 대해 서로 다른 homozygous SNPs을 선발 후, 선발 SNP에 대해 #C를 갈변/미갈변 homozygous SNPs을 나타내는 샘플들과 비교하여 heterozygous SNPs을 나타내는 최종적으로 선발함 (80개의 SNP 선발)

- 흥미로운 결과는 80개의 선발된 변이들 중, 표고버섯 스캐폴드 3번에 위치한 유전자 GENE08480에서 35개의 SNPs이 집중적으로 분포함 (특히, 14개의 missense variants(아미노산 서열을 변형시키는 변이들)이 확인됨)

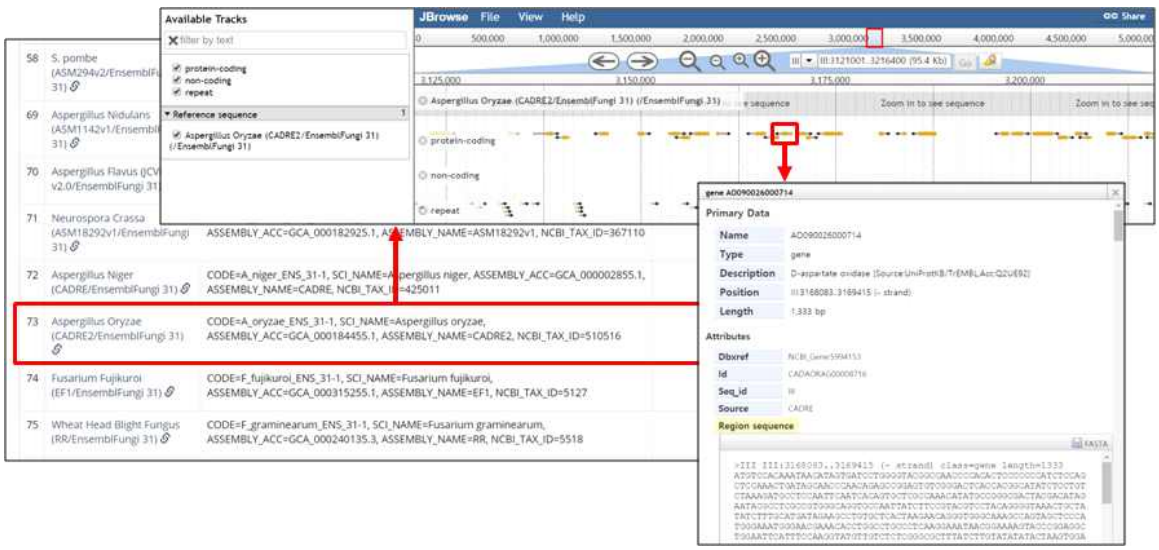
- 19개의 unrelated samples (5개 비정상 갈변 샘플 포함)에 대한 홀지놈 리시퀀싱 데이터 추가 분석하여 분석된 SNP 후보들에 대해 검증할 예정임



[그림47. 표고버섯에서 갈변화 관련 유전자(GENE08480) 내 변이의 집중적 분포]

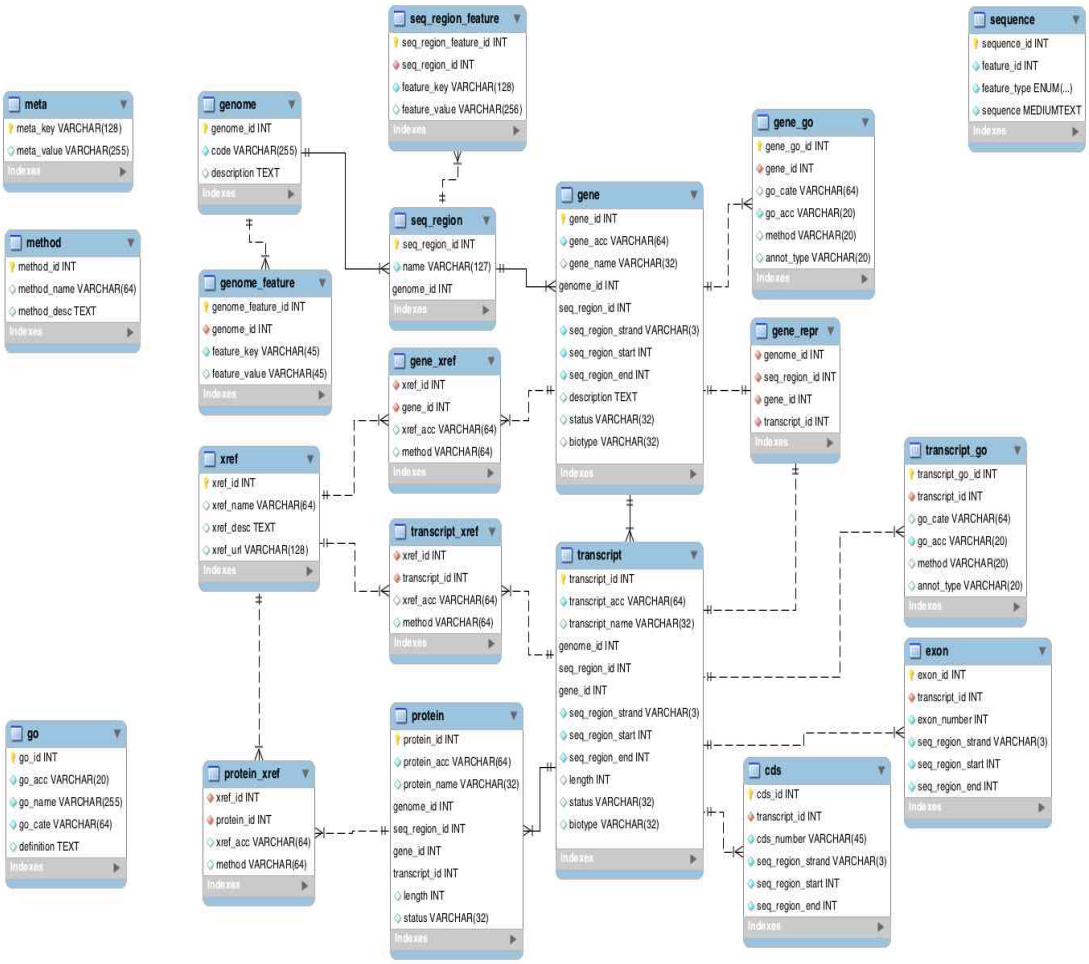
(6) 진균류 기능 분석을 위한 참조데이터베이스 구축

○ Genome 정보 저장을 위한 데이터베이스 구조: *Saccharomyces cerevisiae, Schizosaccharomyces pombe, Aspergillus, Fusarium , Neurospora, Magnaporthe* 등 진균류 유전체 10종에 대한 데이터베이스 구축 (https://omics.genome-report.com/genomes)



[그림48. 진균류 유전체 정보 저장을 위한 데이터베이스 구조]

▣ 데이터베이스의 구조 (아래 그림 참조): 구축된 데이터베이스에는 크게 (1) 유전체 서열 (염색체 또는 스캐폴드(컨티그 포함) 수준으로 구분), 유전자 영역 (transcripts 및 proteins 구조, 유전자 서열 방향, 서열 정보, 기능 및 GO 정보) 등을 포함함



[그림49. 진균류 참조유전체 데이터베이스 구조]

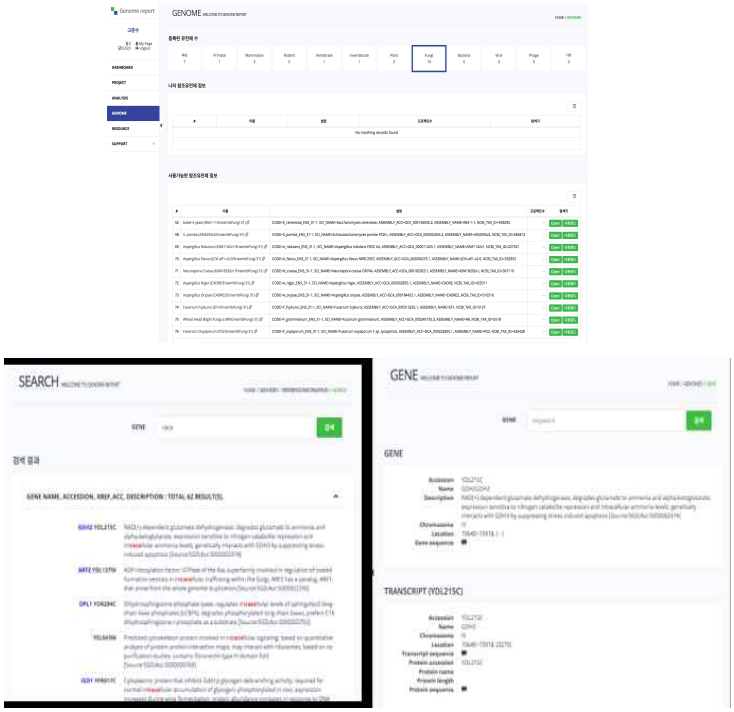
[표14. 데이터베이스 구축에 사용된 진균류 목록]

번호	학명	크기	Contig	Chromsome 수
1	*Saccharomyces cerevisiae S288c*	12,272,600	17	17
2	*Schizosaccharomyces pombe 972h-*	12,821,472	6	6
3	*Aspergillus nidulans FGSC A4*	29,925,798	8	8
4	*Aspergillus flavus NRRL3357*	39,974,047	2,761	
5	*Aspergillus niger*	34,405,768	8	
6	*Aspergillus oryzae*	37,952,189	8	
7	*Fusarium fujikuroi*	43,847,810	12	
8	*Fusarium graminearum*	38,060,440	5	
9	*Fusarium oxysporum f. sp. lycopersici*	61,470,697	15	
10	*Neurospora crassa OR74A*	39,225,835	251	

(7) 진균류 연구 지원을 위한 웹사이트 구축

○ 진균류 유전체 연구를 위해 구축된 데이터베이스 활용을 위한 웹사이트 구축 완료

○ 웹사이트 개발은 NGS분석을 위하 기개발된 genome-report 플랫폼을 이용하여 개발되었으며, 다양한 진균류 유전체를 등록하고, 이용하기 위해서 시스템 개편을 하는 형식으로 진행되었음



[그림50. 구축된 진균류 데이터베이스를 개발된 웹사이트에서 검색하는 예제 (특정 유전체 내 유전자 검색 예제)]

(8) 미생물 전략유전체 사업단 내 타과제 지원 분석 결과

(1차년도)

○ 참조유전체 과제팀(중앙대 및 숭실대)의 유전체서열 조립 지원

- 누룩유전체 3종 (*Saccharomycopsis fibuligera* 2종, *Lichtheimia ramosa* 1종)에 대해 TSLR, PacBio 등의 long-read 기술과 short-insert, long mate-pair를 통합한 조립 기술 지원

(2차년도)

○ 국순당 참조유전체 서열 조립 지원

○ “벼와 고추 침해 주요 공기전반 병원성 곰팡이의 발병유전체 분석 및 기능연구”팀: 유전체 분석 기술 지원 (1건), 전사체 시계열 데이터 분석 지원 (2건)

(3차년도)

○ 사업단 내 진균류 유전체 분석 대상 과제(참조유전체 연구팀- ‘숭실대 서정아 교수팀’) 전사체 분석 지원: *Saccharomycopsis fibuligera* RNA-Seq 분석 지원 (4건)

(4차년도)

○ 사업단 내 진균류 유전체 분석 대상 과제(참조유전체 연구팀- ‘중앙대 강현아 교수팀’) 전사체 분석 지원: *Saccharomycopsis fibuligera* RNA-Seq 분석 지원 (16건)

○ 향미 관련 전사인자 분석을 위한 ChIP-Seq 및 RNA-Seq 분석 지원 (참조유전체 연구팀- ‘중앙대 강현아 교수팀’) (16건)

3. 목표 달성도 및 관련 분야 기여도

3-1. 목표

○ 세균 genome의 NGS 분석 파이프라인 개발
○ 웹 방식을 통한 신규 비교유전체 통합분석시스템 구축
○ Genome database 구축 및 업데이트
○ 세균 전사체 분석 파이프라인 고도화 및 분석 모듈 개발 업데이트
○ 진균류 분석을 위한 유전체 조립 및 전사체 분석 파이프라인 개발
○ 진균류 유전체 정보 활용을 위한 기반 시스템 개발 및 기능분석을 위한 파이프라인 개발
○ 진균류 전사체 데이터 활용을 위한 시스템 고도화 및 프로모터 분석 파이프라인 개발
○ 진균류 유전체 분석을 위한 통합 분석 시스템 개발 및 고급 분석 파이프라인 개발

3-2. 목표 달성여부

(1) 정성적 목표

개발 목표	목표 달성 내용
○ 세균 genome의 NGS 분석 파이프라인 개발	○ NGS data로부터 assemble, gene prediction, annotation, analysis, genome comparison 하는 분석 파이프라인을 구축 및 업데이트함. ○ Visualization 모듈 개발 및 웹 상에서 구현 ○ KEGG database와 연동하여 유전자 정보를 살펴 볼 수 있도록 구현. * **발표논문 1건:** Yoon et al. (2017) A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie van Leeuwenhoek. 2017 Oct;110(10):1281-1286.
○ 웹 방식을 통한 신규 비교유전체 통합분석시스템 구축	○ 비교유전체 셋트 구축 및 업데이트 (http://agri.ezbiocloud.net) (44,048 genome/ 1, 945 Pan-genome set) ○ 주요미생물 비교유전체 set 구축: 동식물 병원균, 유산균 및 식품위해균 등
○ Genome database 구축 및 업데이트	○ 표준균주 (type strain) 유전체 데이터 생산 및 DB 구축 - 총 188종 (유산균 80종) 에 대한 유전체 데이터 생산 ○ EzBioCloud를 이용하여 주요 균주의 type strain에 대해서 genome sequencing 분석 수행. ○ 농축산업에 유용한 유산균의 genome reference DB 구축을 위해 유산균 중에서 표준균주의 유전체 분석이 안된 균주를 분석함. * **발표논문 1건:** Yang et al. (2017) Rejection of

	reclassification of *Lactobacillus kimchii* and *Lactobacillus bobalius* as later subjective synonyms of *Lactobacillus paralimentarius* using comparative genomics. Int J Syst Evol Microbiol. 2017 Nov;67(11):4515-4517.
○ 세균 전사체 분석 파이프라인 고도화 및 분석 모듈 개발 업데이트	○ 여러 정규화 방법을 이용한 전사체 발현량 제시 ○ 라이브러리 크기, 유전자 길이 등 발현량에 영향을 줄 수 있는 요인들을 고려하여 정규화 방법(normallization)을 도입함. ○ 이를 위해 일반적으로 사용한 RPKM 이외에 RLE (Relative Log Expression), TMM (Trimmed Mean of M-value) 방법을 RNA-Seq 분석결과를 제시하는 천랩이 본 과제를 통해서 개발한 CLRNASeq software 에 적용.
○ 진균 참조 유전체 조립 및 유전자 예측 시스템 개발	○ Long-/short-reads 하이브리드 방법 기반의 참조유전체 조립 파이프라인 개발 (1건) (*누룩 3종 유전체 서열 조립 지원) ○ Evidence(전사체/단백질데이터)-기반의 유전자 구조 예측 파이프라인 개발 (1건) (*진균류 유전자 정확도 예측 평가 시스템 포함) ○ TaF - Taxonomic profiling 및 상동성 검색 기반의 유전자 예측 웹서버 개발 (1건) ○ Orthologous gene cluster 분석 파이프라인 개발 (1건)
○ 진균 전사체 분석을 위한 파이프라인 개발	○ 턱시도 프로토콜 방식의 진균 전사체 분석 파이프라인 개발 (1건) (*표고버섯 갈변화 관련 유전자 후보군 분석에 적용됨) ○ 시계열 전사체 데이터 분석 기능 추가 ○ KEGG core DB 구축 및 기능 추가
○ 진균류 참조 유전체 정보 전체 정보 활용을 위한 데이터 베이스 및 웹사이트 개발	○ 진균 유전체 데이터베이스 구축 (1건) (10종 진균류 유전체 포함) ○ 가계도-기반의 형질 관련 유전변이 탐색 파이프라인 개발 (1건) ○ 진균류 분석을 위한 통합 분석을 위한 웹사이트 구축
○ 프로모터 및 전사인자 분석 파이프라인 개발	○ 전사인자(TFs) 및 히스톤 변형 분석을 위한 ChIP-Seq 분석 파이프라인 개발 (1건) (*효모에서 향 및 대사관련 전사인자 ChIP-Seq 분석) ○ 효모 유전체 정보 기반의 TF 모티프 분석 모듈 개발 추가

(2) 정량적 목표

성과목표	전략 미생물 해독	유용 유전자원 확보	사업화·실용화	표준 유전체 해독	메타지놈 분석	유전체 분석기술개발	NABIC 등록	병원성 미생물 진단마커개발	병원성 미생물 정보완성	미생물 병발생 기작 규명
최종목표						2	6			
연구기간 내 달성실적						9	56			
연구종료 후 성과창출 계획										

성과목표	사업화지표											연구기반지표								
	지식재산권			기술실시(이전)		사업화					기술인증	학술성과				교육지도	인력양성	정책활용·홍보		기타(타연구활용등)
												논문		논문평균IF	학술발표					
	특허출원	특허등록	품종등록	건수	기술료	제품화	매출액	수출액	고용창출	투자유치		SCI	비SCI					정책활용	홍보전시	
단위	건	건	건	건	백만원	건	백만원	백만원	명	백만원	건	건	건		건		명			
가중치																				
최종목표		3				2						5				4				80
연구기간 내 달성실적		1		2	11	2	199					2	1		3	15				165
연구종료 후 성과창출 계획		2										4								

3-3. 목표 미달성 시 원인(사유) 및 차후대책(후속연구의 필요성 등)

○ **특허 2건 미달성 원인**:

- 비교유전체(천랩) 와 참조유전체 조립(테라젠) 시스템의 업그레이드를 위해서 프로그램 등록이 미루어졌으며, 과제 종료시 완결되는 바, 이에 대한 등록을 준비하고 있음.

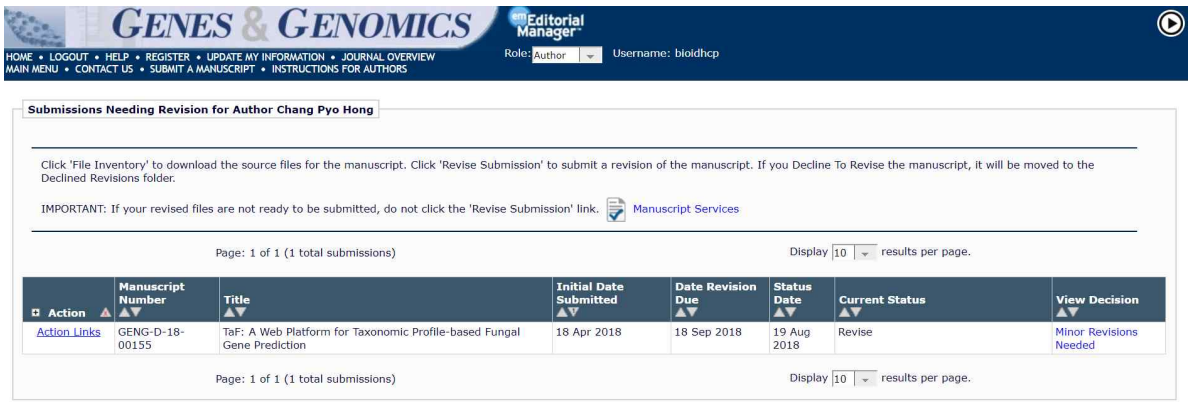
○ **SCI급 논문 3건 미달성 원인**: 현재 투고된 논문 4편이 심사 중에 있음

1) 천랩: 논문 2건 제출 및 심사중

- Genome-based reclassification of *Weissella jogaejeotgali* as a later heterotypic synonym of

Weissella thailandensis (Int J Syst Evol Microbiol.에 투고/심사; 과제번호 사사)
- Genome-based reclassification of *Marinobacter adhaerens* as a later heterotypic synonym of *Marinobacter flavimaris* (Int J Syst Evol Microbiol.에 투고/심사; 과제번호 사사)

2) 테라젠: 논문 2건 제출 및 심사중
- TaF: A Web Platform for Taxonomic Profile-based Fungal Gene Prediction (Genes and Genomics (IF 0.566)에 투고 / 현재 final acceptance 받음 (2018/08/20); 과제번호 사사)


GENES & GENOMICS
Editorial Manager
HOME • LOGOUT • HELP • REGISTER • UPDATE MY INFORMATION • JOURNAL OVERVIEW
MAIN MENU • CONTACT US • SUBMIT A MANUSCRIPT • INSTRUCTIONS FOR AUTHORS
Roles: Author Username: bioidhcp

Submissions Needing Revision for Author Chang Pyo Hong

Click 'File Inventory' to download the source files for the manuscript. Click 'Revise Submission' to submit a revision of the manuscript. If you Decline To Revise the manuscript, it will be moved to the Declined Revisions folder.

IMPORTANT: If your revised files are not ready to be submitted, do not click the 'Revise Submission' link. Manuscript Services

Page: 1 of 1 (1 total submissions) Display 10 results per page.

Action	Manuscript Number	Title	Initial Date Submitted	Date Revision Due	Status Date	Current Status	View Decision
Action Links	GENG-D-18-00155	TaF: A Web Platform for Taxonomic Profile-based Fungal Gene Prediction	18 Apr 2018	18 Sep 2018	19 Aug 2018	Revise	Minor Revisions Needed

Page: 1 of 1 (1 total submissions) Display 10 results per page.

- Comparative transcriptome analysis identified candidate genes involved in browning of mycelium in *Lentinula edodes* (BMC Genomics에 투고/심사중; 과제번호 사사; 2018년까지 SCI(E)급 논문에 게재 계획)

○ 추후 대책:
(1) 추가 특허 관련: 2건의 프로그램을 추가로 등록할 계획임.
- 천랩: 세균의 비교유전체 분석플랫폼에 대해서 프로그램 등록을 준비하고 있음
- 테라젠: 참조 유전체 조립 및 유전자 예측 시스템에 대해서 프로그램 등록을 준비하고 있음 (2018까지 SW 국가 R&D 성과물 등록 예정)

(2) 투고 예정인 논문 3편(천랩 2건, 테라젠 1건)
- 천랩: 유전체 정보에 기반한 세균 표준균주의 재분류에 관련된 논문을 2018년 9월 중으로 미생물분류학회지 관련 저널에 투고할 예정임
- 테라젠: 가계도-기반의 형질 관련 유전변이 탐색 파이프라인 개발 관련 논문을 과제종료 후 1년 내에 SCI(E)급 논문에 게재 계획

4. 연구결과의 활용 계획 등

○ 본 과제의 목적은 생명정보 비전공자도 쉽게 사용할 수 있는 미생물 유전체, 전사체 분석 시스템 구축과 이를 활용한 타 과제 지원 분석임. 이를 위하여 본 과제팀은 유전체 분석 및 전사체 분석 파이프라인 개발 및 최적화할 뿐만 아니라, 국내외 연구자가 활용할 수 있는 유전체 분석 플랫폼과 소프트웨어를 개발하였음.

○ 생물정보 비전공자를 위한 유전체 및 전사체 분석 시스템은 국내 유일일 뿐만 아니라, 해외 사례에서도 찾아볼 수 없는 독자적인 통합 시스템이라 할 수 있음. 본 연구팀이 구축한 분석 시스템을 이용하여 이미 사업화를 진행하고 있으며, 충분히 국내외적으로 유전체 사업화에 경쟁력 있는 시스템으로 인정받고 있음. 본 과제뿐만 아니라 유전체 분석 서비스를 통해서 기업의 R&D 투자를 진행해왔고 현재도 상당한 부분을 투자하고 있는 부분임.

○ 본 연구팀(천랩과 테라젠)은 이와 같은 시스템을 구축 및 완성하기 위해 1,2,3 차년도에 분석 파이프라인 및 소프트웨어, DB등의 만들어 공개하였으며, 4차년도에 이를 고도화 및 최적하였음. 구축된 통합된 시스템을 활용하여, 개별 유전체 분석과 더불어 비교유전체 분석이 적합한 솔루션을 연구자들에게 제공할 예정임.

○ 본 과제의 최종적인 활용 기대 성과로는 농식품산업에 중요한 미생물의 유전체/전사체 데이터베이스를 구축함으로써 농식품산업에 활용할 수 있는 미생물의 유전체 연구 활성화에 기여할 수 있음

○ 통합분석시스템 및 소프트웨어를 통한 유전체 분석기술의 사업화 방안
- 기존의 사업화 솔루션을 확대하여, 2단계에서 좀더 발전된 통합분 시스템을 구축함.
- 활용워크숍을 지속적으로 수행하여 개발된 플랫폼에 대한 사용자 교육 및 분석 지원할 지속할 예정임
- 논문 출판과 국내외 학회 참가 및 홍보활동을 통하여 사업화 촉진

❖ 최종보고서 수정, 보완 사항

○ 중복된 과제와의 차별화된 점
- 본 과제의 목적은 원핵 및 진핵을 포함한 미생물의 유전체, 전사체 분석 시스템 구축과 이를 활용한 타 과제 지원 분석이 목적으로 기존에 수행했던 과제와는 내용과 목표에 있어서 중복되는 것이 없음. 특히 본 과제에서는 유전체사업단의 다른 분석을 165건의 지원 분석한 바가 있으며, 15건의 교육지도를 수행하였고, 56건의 NABIC등록을 하는 등 수행내용에 있어서도 기존 과제와 차별성이 있음.

○ 분석시스템에 대한 신뢰도
- 본 연구팀에서 개발한 미생물 유전체 및 전사체 분석 파이프라인에 사용되는 각각의 프로그램은 기존에 유전체 연구에 많이 사용되는 알고리즘을 이용한 것이 대부분이며, orthoANI-u 등 비교 유전체에 활용되는 새로운 알고리즘은 논문으로 발표한 바가 있음. 본 연구팀이 개발한 유전체, 전사체 분석의 장점은 클라우드 분석 시스템을 활용한 다양한 비교 분석에 있으며, 개발된 분석 시스템은 이미 전세계 연구자들이 활용하고 있고, 이를 활용한 논문은 100여편이 넘게 발표되고 있음.

○ 기존 유전체 분석 시스템과의 차별성

- 기존의 유전체 분석 시스템에서 많이 사용되는 CLCworkbench 와 본연구팀이 개발한 클라우드 웹 방식의 미생물 유전체 분석 시스템과의 차이를 살펴보면, CLCworkbench는 서버급의 고성능 컴퓨팅 사양을 필요로 하지만, 본 연구팀이 구축한 분석 시스템은 클라우드 웹 방식으로 사용자가 웹이 가능한 노트북 정도의 사양만 되어도 유전체 분석 및 비교 분석이 가능한 시스템임. 또한 본 연구팀은 전세계에 공개된 모든 미생물 유전체 정보를 데이터베이스화 하여 제공하고 있으며, 이를 통하여 다양한 비교 유전체 분석이 가능하나, CLCworkbench 에서는 이러한 비교 유전체 분석이 불가능함. 또한 생명정보 분석 비전공자가 사용하기 쉬운 사용성 측면에서도 본 연구팀이 구축한 미생물 유전체 분석/비교 분석 시스템이 훨씬 용이한 측면이 있음.

[첨부] 참고문헌

1. Agers-Loustau A, Petrillo M.The challenges of designing a benchmark strategy for bioinformatics
pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. F1000Res. 2018 Apr 13;7:459. doi: 10.12688/f1000research.14509.1. eCollection 2018.

2. Hornung B, Martins Dos Santos VAP, Smidt H, Schaap PJ. Studying microbial functionality within the gut ecosystem by systems biology. Genes Nutr. 2018 Mar 6;13:5. doi: 10.1186/s12263-018-0594-6. eCollection 2018. Review.

3. Yang Z, Mammel M, Papafragkou E, Hida K, Elkins CA, Kulka M. Application of next generation sequencing toward sensitive detection of enteric viruses isolated from celery samples as an example of produce. Int J Food Microbiol. 2017 Nov 16;261:73-81. doi: 10.1016/j.ijfoodmicro.2017.07.021.

4 Hücker SM, Ardern Z, Goldberg T, Schafferhans A, Bernhofer M, Vestergaard G, Nelson CW, Schloter M, Rost B, Scherer S, Neuhaus K. Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157:H7 Sakai genome. PLoS One. 2017 Sep 13;12(9):e0184119. doi: 10.1371/journal.pone.0184119. eCollection 2017.

5. Sharma TR, Devanna BN, Kiran K, Singh PK, Arora K, Jain P, Tiwari IM, Dubey H, Saklani B, Kumari M, Singh J, Jaswal R, Kapoor R, Pawar DV, Sinha S, Bisht DS, Solanke AU, Mondal TK. Status and Prospects of Next Generation Sequencing Technologies in Crop Plants. Curr Issues Mol Biol. 2018;27:1-36. doi: 10.21775/cimb.027.001. Epub 2017 Sep 8. Review.

6. Cao J, Yu Y, Huang J, Liu R, Chen Y, Li S, Liu J. Genome re-sequencing analysis uncovers pathogenecity-related genes undergoing positive selection in Magnaporthe oryzae. Sci China Life Sci. 2017 Aug;60(8):880-890. doi: 10.1007/s11427-017-9076-4. Epub 2017 Jul 25.

7. Zhao H, Sun W, Wang Z, Zhang T, Fan Y, Gu H, Li G. Mink (Mustela vison) Gut Microbial Communities from Northeast China and Its Internal Relationship with Gender and Food Additives. Curr Microbiol. 2017 Oct;74(10):1169-1177. doi: 10.1007/s00284-017-1301-3. Epub 2017 Jul 14.

8. Margos G, Hepner S, Mang C, Marosevic D, Reynolds SE, Krebs S, Sing A, Derdakova M, Reiter MA, Fingerle V. Lost in plasmids: next generation sequencing and the complex genome of the tick-borne pathogen Borrelia burgdorferi. BMC Genomics. 2017 May 30;18(1):422. doi: 10.1186/s12864-017-3804-5.

9. Furió-Tarí P, Conesa A, Tarazona S. RGmatch: matching genomic regions to proximal genes in omics data integration. BMC Bioinformatics. 2016 Nov 22;17(Suppl 15):427. doi: 10.1186/s12859-016-1293-1.

10. Mariano DC, Pereira FL, Aguiar EL, Oliveira LC, Benevides L, Guimarães LC, Folador EL, Sousa TJ, Ghosh P, Barh D, Figueiredo HC, Silva A, Ramos RT, Azevedo VA. SIMBA: a web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology. BMC Bioinformatics. 2016 Dec 15;17(Suppl 18):456.

11: Chan CHS, Octavia S, Sintchenko V, Lan R. SnpFilt: A pipeline for reference-free

assembly-based identification of SNPs in bacterial genomes. Comput Biol Chem. 2016 Dec;65:178-184. doi: 10.1016/j.compbiolchem.2016.09.004. Epub 2016 Sep 9.

12. Thao NP, Tran LS. Enhancement of Plant Productivity in the Post-Genomics Era. Curr Genomics. 2016 Aug;17(4):295-6. doi: 10.2174/1389202917041606071825O7.

13. Klosterman SJ, Rollins JR, Sudarshana MR, Vinatzer BA. Disease Management in the Genomics Era-Summaries of Focus Issue Papers. Phytopathology. 2016 Oct;106(10):1068-1070.

14. Clausen PT, Zankari E, Aarestrup FM, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. J Antimicrob Chemother. 2016 Sep;71(9):2484-8. doi: 10.1093/jac/dkw184.

15. Deng X, den Bakker HC, Hendriksen RS. Genomic Epidemiology: Whole-Genome-Sequencing-Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. Annu Rev Food Sci Technol. 2016;7:353-74. doi: 10.1146/annurev-food-041715-033259.

16. Pauletto M, Carraro L, Babbucci M, Lucchini R, Bargelloni L, Cardazzo B. Extending RAD tag analysis to microbial ecology: a comparison between MultiLocus Sequence Typing and 2b-RAD to investigate Listeria monocytogenes genetic structure. Mol Ecol Resour. 2016 May;16(3):823-35. doi: 10.1111/1755-0998.12495.

17. Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E, Hendriksen RS; Global Microbial Identifier initiative's Working Group 4 (GMI-WG4). Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. BMC Infect Dis. 2015 Apr 3;15:174. doi: 10.1186/s12879-015-0902-3.

18. Faison WJ, Rostovtsev A, Castro-Nallar E, Crandall KA, Chumakov K, Simonyan V, Mazumder R. Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes. Genomics. 2014 Jul;104(1):1-7. doi: 10.1016/j.ygeno.2014.06.001.

<별첨작성 양식>

[별첨 1]

연구개발보고서 초록

<table>
<tr><td rowspan="2">과 제 명</td><td colspan="5">(국문) NGS를 활용한 미생물 유전체 및 전사체 분석 소프트웨어 및 시스템 개발</td></tr>
<tr><td colspan="5">(영문) Development of total analyzing software and system for fungi genomes and transcriptomes using NGS data</td></tr>
<tr><td>주관연구기관</td><td colspan="2">㈜천랩</td><td rowspan="2">주 관 연 구 책 임 자</td><td colspan="2">(소속) 대표이사</td></tr>
<tr><td>참 여 기 업</td><td colspan="2">테라젠이텍스</td><td colspan="2">(성명) 천종식</td></tr>
<tr><td rowspan="4">총연구개발비

(1,066,672 천원)</td><td>계</td><td>1,066,672천원</td><td>총 연 구 기 간</td><td colspan="2">2014.8.23 ~ 2018.8.22 (4년)</td></tr>
<tr><td>정부출연 연구개발비</td><td>800,000천원</td><td rowspan="3">총 참 여 연 구 원 수</td><td>총 인 원</td><td>34 명</td></tr>
<tr><td>기업부담금</td><td>266,672천원</td><td>내부인원</td><td>34 명</td></tr>
<tr><td>연구기관부담금</td><td>-</td><td>외부인원</td><td>-</td></tr>
</table>

□ **연구개발 목표**

○ NGS를 활용한 미생물(진균류) 유전체 통합 분석 system 개발 및 데이터베이스 구축함으로써 유전체 연구 활성화에 기여하고 목적 지향적 바이오산업에 활용을 가능케 함

□ **연구의 내용**

○ 1세부에서는 미생물중 원핵생물 (prokaryote)의 유전체 분석 및 전사체 분석을 위해 생물정보학적 기술을 활용해서 통합분석시스템과 데이터베이스를 구축함

○ 1협동에서는 미생물 중 진핵생물 (eukaryote)에 해당하는 진균류의 유전 정보 분석을 위한 참조 유전체 조립 파이프라인 및 유전체 발굴 시스템 개발, 진균류 유전자 기능 연구 및 유용 유전자 발굴을 위한 통합 분석 시스템을 개발함

□ **연구의 내용 및 결과**

○ **세균 genome의 NGS 분석 파이프라인 개발**

- NGS data로부터 assemble, gene prediction, annotation, analysis, genome comparison 하는 분석 파이프라인을 구축 및 업데이트함.
- Visualization 모듈 개발 및 웹 상에서 구현
- KEGG database와 연동하여 유전자 정보를 살펴 볼 수 있도록 구현

○ **웹 방식을 통한 신규 비교유전체 통합분석시스템 구축**

(http://agri.ezbiocloud.net)

- 비교유전체 셋트 구축 및 업데이트
  (44,048 genome/ 1, 945 Pan-genome set)
- 주요미생물 비교유전체 set 구축: 동식물 병원균, 유산균 및 식품위해균 등

○ Genome database 구축 및 업데이트

- 표준균주 (type strain) 유전체 데이터 생산 및 DB 구축
(유산균 80종 포함, 총 세균 188종에 대한 유전체 데이터생산)
- EzBioCloud를 이용하여 주요 균주의 type strain에 대해서 genome sequencing 분석 수행.
- 농축산업에 유용한 유산균의 genome reference DB 구축을 위해 유산균 중에서 표준균주의 유전체 분석이 안 된 균주를 분석함

○ 세균 전사체 분석 파이프라인 고도화 및 분석 모듈 개발 업데이트

- 여러 정규화 방법을 이용한 전사체 발현량 제시
- 라이브러리 크기, 유전자 길이 등 발현량에 영향을 줄 수 있는 요인들을 고려하여 정규화 방법(normallization)을 도입함
- 이를 위해 일반적으로 사용한 RPKM 이외에 RLE (Relative Log Expression), TMM (Trimmed Mean of M-value) 방법을 RNA-Seq 분석결과를 제시하는 천랩이 본 과제를 통해서 개발한 CLRNASeq software 에 적용

○ 진균 참조 유전체 조립 및 유전자 예측 시스템 개발

- Long- 및 short-reads을 활용한 하이브리드 방법 기반의 참조유전체 조립 파이프라인 개발하여 국내 누룩 유전체 3종 서열 조립 지원함
- 유전자 구조 예측을 위해 evidence-based prediction 시스템을 개발했고, 관련 유전자 정확도 예측 평가가 확보됨. 또한 taxonomic profiling 및 상동성 검색 기반의 유전자 예측 웹서버인 TaF를 개발함

○ 진균 전사체 분석을 위한 파이프라인 개발

- 턱시도 프로토콜 방식의 진균류 전사체 분석을 위한 파이프라인을 개발하였고, 시계열 전사체 데이터 및 KEGG core DB 적용 기능을 추가함. 이를 토대로 표고버섯 갈변화 관련 유전자 후보군 분석에 활용됨

○ 진균류 참조 유전체 정보 전체 정보 활용을 위한 데이터베이스 및 웹사이트 개발

- 10종 진균류 유전체 포함한 진균 유전체 데이터베이스 구축하였고, 가계도-기반의 형질 관련 유전변이 탐색 파이프라인 개발함. 또한 진균류 분석을 위한 통합 분석을 위한 웹사이트 구축함

○ 프로모터 및 전사인자 분석 파이프라인 개발

- 전사인자 및 히스톤 변형 분석을 위한 ChIP-Seq 분석 파이프라인 개발하였고, 효모에서 향 및 대사관련 전사인자들을 연구중에 있음. 또한 파이프라인에 효모 유전체 정보 기반의 TF 모티프 분석 모듈 개발 추가함 (orthologous gene cluster 분석모듈 결합)

□ 연구성과 활용실적 및 계획

○ 개발한 유전체 분석 기술 및 시스템을 활용하여 사업단 및 관련 연구자들의 생물정보학적 분석을 지원할 계획임
○ 개발된 유전체 분석 기술을 활용하여 유용 미생물 유전자원 탐색 및 사업화 계획
○ 미생물 유전체 교육에 활용 및 지원

[별첨 2]

자체평가의견서

1. 과제현황

	과제번호			914008-04	
사업구분	농식품기술개발사업				
연구분야			과제구분	단위	
사 업 명	포스트게놈 다부처유전체사업			주관	
총괄과제	기재하지 않음		총괄책임자	기재하지 않음	
과 제 명	NGS를 활용한 미생물 유전체 및 전사체 분석 소프트웨어 및 시스템 개발		과제유형	(개발)	
연구기관	㈜천랩		연구책임자	천종식	
연구기간 연 구 비 (천원)	연차	기간	정부	민간	계
	1차년도		200,000	66,668	266,668
	2차년도		200,000	66,668	266,668
	3차년도		200,000	66,668	266,668
	4차년도		200,000	66,668	266,668
	계		800,000	266,672	1,066,672
참여기업	㈜천랩, 테라젠이텍스				
상 대 국		상대국연구기관			

※ 총 연구기간이 5차년도 이상인 경우 셀을 추가하여 작성 요망

2. 평가일 : 2018.11.06

3. 평가자(연구책임자) :

소속	직위	성명
㈜천랩	대표이사	천종식

4. 평가자(연구책임자) 확인 :

본인은 평가대상 과제에 대한 연구결과에 대하여 객관적으로 기술하였으며, 공정하게 평가하였음을 확약하며, 본 자료가 전문가 및 전문기관 평가 시에 기초자료로 활용되기를 바랍니다.

확 약	

Ⅰ. 연구개발실적

※ 다음 각 평가항목에 따라 자체평가한 등급 및 실적을 간략하게 기술(200자 이내)

1. 연구개발결과의 우수성/창의성

■ 등급 : (<u>아주우수</u>, 우수, 보통, 미흡, 불량)

○ **웹 방식을 통한 신규 비교유전체 통합분석시스템 구축**
- 비교유전체 셋트 구축 및 업데이트
  (44,048 genome/ 1, 945 Pan-genome set)
- 주요미생물 비교유전체 set 구축: 동식물 병원균, 유산균 및 식품위해균 등

○ 최근 기술 동향에 적합한 오믹스 분석을 위한 시스템을 개발
- Long- 및 short-reads을 활용한 정확도 높은 유전체 서열 조립 방법 고안
- 정확도 높은 유전자 구조 예측을 위한 evidence-based prediction 시스템 개발함. 특히 전사체 데이터 부재 시 정확도가 확보된 taxonomic profiling 및 상동성 검색 기반의 유전자 예측 웹서버(TaF) 최초 개발
- 진균류에서 전사인자 및 유전자 발현을 통합 분석 가능케하는 파이프라인 개발

2. 연구개발결과의 파급효과

■ 등급 : (<u>아주우수</u>, 우수, 보통, 미흡, 불량)

○ 본 과제의 결과물로 생물정보 비전공자도 쉽게 사용할 수 있는 미생물 유전체, 전사체 분석 시스템 구축과 이를 활용한 타 과제 지원 분석임. 이를 위하여 본 과제팀은 유전체 분석 및 전사체 분석 파이프라인 개발 및 최적화할 뿐만 아니라, 연구자가 활용할 수 있는 유전체 분석 플랫폼과 소프트웨어를 개발하였음.

○ 개발된 유전체 분석 기술 및 시스템들은 whole-genome *de novo* assembly, 신규 유전체내 유전자 구조 예측, 유용물질 합성 관련 대사경로에 관여하는 유전자 후보군 발굴, 미생물 내 중요 형질 조절 전사인자 탐색에 매우 유용함을 보여줌

3. 연구개발결과에 대한 활용가능성

■ 등급 : (아주우수, <u>우수</u>, 보통, 미흡, 불량)

○ 유전자 구조 예측 및 전사체 분석은 웹사이트를 통해 연구자가 분석을 직접 할 수 있도록 구현함. 그 이외의 시스템들은 공동연구를 통해 사업단 및 관련 분야 연구자들의 생물정보학적 분석을 지원 가능함
○ 개발된 유전체 분석 기술을 활용하여 유용 미생물 유전자원 탐색 및 사업화 계획
○ 미생물(세균류, 진균류) 유전체 교육에 활용 및 지원

4. 연구개발 수행노력의 성실도

■ 등급 : (**아주우수**, 우수, 보통, 미흡, 불량)

○ 계획서 대비 연구개발 목표를 충실히 달성함
○ 생물정보 비전공자를 위한 유전체 및 전사체 분석 시스템은 국내 유일할 뿐만 아니라, 해외 사례에서도 찾아볼 수 없는 독자적인 통합 시스템이라 할 수 있음. 본 연구팀이 구축한 분석 시스템을 이용하여 이미 사업화를 진행하고 있으며, 충분히 국내외적으로 유전체 사업화에 경쟁력 있는 시스템으로 인정받고 있음.
○ 특히, 사업화 및 교육지원, 분석지원은 당초 목표보다 초과달성하였음

5. 공개발표된 연구개발성과(논문, 지적소유권, 발표회 개최 등)

■ 등급 : (아주우수, 우수, 보통, 미흡, 불량)

○ 계획서 목표 연구논문 게재 편수(SCI 5편)에 대한 목표를 달성 못했지만, 2018년 12말까지 초과달성 가능할 것으로 판단함 (현재 2편 출판, 4편 투고심사 중에 있어 목표 달성시 최소 6편의 논문발표가 예상됨)

Ⅱ. 연구목표 달성도

세부연구목표 (연구계획서상의 목표)	비중 (%)	달성도 (%)	자체평가
○ 세균 genome의 NGS 분석 파이프라인 개발	15	100	-NGS data로부터 assemble, gene prediction, annotation, analysis, genome comparison 하는 분석 파이프라인을 구축 및 업데이트함. - Visualization 모듈개발 및 웹상에서 구현
○ 웹 방식을 통한 신규 비교유전체 통합분석시스템 구축	15	100	- 비교유전체 셋트 구축 및 업데이트 (44,048 genome/ 1, 945 Pan-genome set) - 주요미생물 비교유전체 set 구축: 동식물 병원균, 유산균 및 식품위해균 등
○ Genome database 구축 및 업데이트	10	100	- 표준균주 (type strain) 유전체 데이터 생산 및 DB 구축 (유산균 80종 포함, 총 세균 188종에 대한 유전체 데이터생산) - 농축산업에 유용한 유산균의 genome reference DB 구축을 위해 유산균 중에서 표준균주의 유전체 분석이 안 된 균주를 분석함.
○ 세균 전사체 분석 파이프라인 고도화 및 분석	10	100	- 라이브러리 크기, 유전자 길이 등 발현량에 영향을 줄 수 있는 요인들을 고려하여 정규화

모듈 개발 업데이트			방법(normallization)을 도입함. - 이를 위해 일반적으로 사용한 RPKM 이외에 RLE (Relative Log Expression), TMM (Trimmed Mean of M-value) 방법을 RNA-Seq 분석결과를 제시하는 천랩이 본 과제를 통해서 개발한 CLRNASeq software 에 적용.
○ 진균 참조 유전체 조립 및 유전자 예측 시스템 개발	15	100	- 관련 분석 파이프라인 총4건 개발 - 진균류 참조유전체 분석에 최적화됨 - 유전체 분석 기술 관련 당초 목표 5건 개발 대비 8건 개발함으로써 초과달성
○ 진균 전사체 분석을 위한 파이프라인 개발	15	100	- 관련 분석 파이프라인 총1건 개발. 웹을 통해 사용자 직접 분석 가능 - 대사경로 후보군 탐색 기능 추가
○ 진균류 참조 유전체 정보 전체 정보 활용을 위한 데이터베이스 및 웹사이트 개발	10	100	- 관련 분석 파이프라인 총1건과 데이터베이스 1건 개발 - 진균류 분석을 위한 통합 분석을 위한 웹사이트를 구축
○ 프로모터 및 전사인자 분석 파이프라인 개발	10	100	- 관련 분석 파이프라인 총4건 개발. 특히 전사인자 데이터 부족으로 인한 분석의 어려움을 극복하기 위해 효모의 데이터를 상동성 검색 기반의 분석법을 적용함
합계	100	100	

Ⅱ. 종합의견

1. 연구개발결과에 대한 종합의견

○ 최근 기술 동향에 적합한 미생물(세균류, 진균류) 오믹스 (유전체, 전사체, 전사인자) 분석을 위한 시스템을 개발함
○ 개발된 유전체 분석 기술 및 시스템들은 whole-genome *de novo* assembly, 신규 유전체내 유전자 구조 예측, 유용물질 합성 관련 대사경로에 관여하는 유전자 후보군 발굴, 미생물 내 중요 형질 조절 전사인자 탐색에 매우 유용함을 보여줌
○ 유전체 분석 기술 관련 당초 목표 5건 개발 대비 8건 개발함으로써 초과달성

2. 평가시 고려할 사항 또는 요구사항

○ 생물정보 비전공자를 위한 유전체 및 전사체 분석 시스템은 국내 유일할 뿐만 아니라, 해외 사례에서도 찾아볼 수 없는 독자적인 통합 시스템이라 할 수 있기에 독창적인 연구라 할 수 있음.
○ 본 연구팀이 구축한 분석 시스템을 이용하여 이미 사업화를 진행하고 있으며, 충분히 국내외적으로 유전체 사업화에 경쟁력 있는 시스템으로 인정받고 있음.
○ 특히, 사업화 및 교육지원, 분석지원은 당초 목표보다 초과달성하였음.

3. 연구결과의 활용방안 및 향후조치에 대한 의견

○ 공동연구 및 교육 지원을 통해 연구결과물을 이용하여 연구자들이 연구계획 및 목표 성과에 부합된 분석 지원을 지속적으로 해 나갈 계획임
○ 최근 기술 동향에 적합하도록 지속적인 분석기술의 업그레이드를 실시할 예정임
○ 논문의 경우 2018년 말까지 목표성과(SCI급 논문 5편)달성할 예정임 (현재 2편 출판완료, 4편 투고심사중, 3편 투고준비중)

Ⅳ. 보안성 검토

o 해당사항 없음.

※ 보안성이 필요하다고 판단되는 경우 작성함.

1. 연구책임자의 의견

○ 본 연구는 생물정보 비전공자를 위한 유전체 및 전사체 분석 시스템을 개발한 것으로 국내 연구자들에게 널리 활용되기를 희망함.

2. 연구기관 자체의 검토결과

○ 본 연구결과는 국내 농식품 미생물 유전체의 활용효과와 사업화 성과에도 중요하므로 많은 연구자들에게 폭 넓게 활용되기를 희망함.
○ 핵심연구 파이프라인과 데이터베이스는 다른 연구자들이 모방할 수 없으므로 특별한 보안성이 요구되지 않음.

[별첨 3]

연구성과 활용계획서

1. 연구과제 개요

사업추진형태	□자유응모과제 ■지정공모과제	분 야	생물정보학
연 구 과 제 명	NGS를 활용한 미생물 유전체 및 전사체 분석 소프트웨어 및 시스템 개발		
주관연구기관	㈜천랩	주관연구책임자	천종식
연 구 개 발 비	정부출연 연구개발비 / 기업부담금 / 연구기관부담금 / 총연구개발비		
	800,000 천원 / 266,672 천원 /		1,066,672 천원
연구개발기간	2014.8.23 ~ 2018.8.22 (4년)		
주요활용유형	□산업체이전 ■교육 및 지도 □정책자료 ■기타(공동연구 분석 지원) □미활용 (사유:)		

2. 연구목표 대비 결과

당초목표	당초연구목표 대비 연구결과
○ 세균 genome의 NGS 분석 파이프라인 개발	○ NGS data로부터 assemble, gene prediction, annotation, analysis, genome comparison 하는 분석 파이프라인을 구축 및 업데이트 완료. ○ Visualization 모듈개발 및 웹상에서 구현완료
○ 웹 방식을 통한 신규 비교유전체 통합분석시스템 구축	○ 비교유전체 셋트 구축 및 업데이트 (44,048 genome/ 1, 945 Pan-genome set) ○ 주요미생물 비교유전체 set 구축: 동식물 병원균, 유산균 및 식품위해균 등 완료
○ Genome database 구축 및 업데이트	○ 표준균주 (type strain) 유전체 데이터 생산 및 DB 구축완료 (유산균 80종 포함, 총 세균 188종에 대한 유전체 데이터생산)
○ 세균 전사체 분석 파이프라인 고도화 및 분석 모듈 개발 업데이트	○ 일반적으로 사용한 RPKM 이외에 RLE (Relative Log Expression), TMM (Trimmed Mean of M-value) 방법을 RNA-Seq 분석결과를 제시하는 천랩이 본과제를 통해서 개발한 CLRNASeq software 에 적용 완료
○ 진균 참조 유전체 조립 및 유전자 예측 시스템 개발	○ Long-/short-reads 하이브리드 방법 기반의 참조유전체 조립 파이프라인 개발 (1건) ○ Evidence(전사체/단백질데이터)-기반의 유전자 구조 예측 파이프라인 개발 (1건) ○ TaF - Taxonomic profiling 및 상동성 검색 기반의 유전자 예측 웹서버 개발 (1건) ○ Orthologous gene cluster 분석 파이프라인 개발 (1건)
○ 진균 전사체 분석을 위한 파이프라인 개발	○ 진균 유전체 데이터베이스 구축 (1건)

	○ 가계도-기반의 형질 관련 유전변이 탐색 파이프라인 개발 (1건) ○ 진균류 분석을 위한 통합 분석을 위한 웹사이트 구축
○ 진균류 참조 유전체 정보 전체 정보 활용을 위한 데이터베이스 및 웹사이트 개발	○ 진균 유전체 데이터베이스 구축 (1건) ○ 가계도-기반의 형질 관련 유전변이 탐색 파이프라인 개발 (1건) ○ 진균류 분석을 위한 통합 분석을 위한 웹사이트 구축
○ 프로모터 및 전사인자 분석 파이프라인 개발	○ 전사인자(TFs) 및 히스톤 변형 분석을 위한 ChIP-Seq 분석 파이프라인 개발 (1건). 효모 유전체 정보 기반의 TF 모티프 분석 모듈 개발 추가됨

* 결과에 대한 의견 첨부 가능

3. 연구목표 대비 성과

성과목표	전략 미생물 해독	유용 유전자원 확보	사업화·실용화	표준 유전체 해독	메타지놈 분석	유전체 분석기술개발	NABIC 등록	병원성 미생물 진단마커개발	병원성 미생물 정보완성	미생물 병발생 기작 규명
최종목표						2	6			
연구기간 내 달성실적						9	56			
달성율(%)						450%	933%			

성과목표	사업화지표											연구기반지표								
	지식재산권			기술실시(이전)		사업화					기술인증	학술성과				교육지도	인력양성	정책활용·홍보		기타(타연구활용 등)
												논문								
	특허출원	특허등록	품종등록	건수	기술료	제품화	매출액	수출액	고용창출	투자유치		SCI	비SCI	논문평균IF	학술발표			정책활용	홍보전시	
단위	건	건	건	건	백만원	백만원	백만원	백만원	명	백만원	건	건	건		건		명	건	건	
가중치																				
최종목표		3				2						5				4				80
연구기간 내 달성실적		1		2	11	2	199					2	1		3	15				165
달성율(%)		33%				100%						40%				375%				206%

4. 핵심기술

구분	핵 심 기 술 명
①	○ 세균 genome의 NGS 분석 파이프라인 개발
②	○ 웹 방식을 통한 신규 비교유전체 통합분석시스템 구축
③	○ Genome database 구축 및 업데이트
④	○ 세균 전사체 분석 파이프라인 고도화 및 분석 모듈 개발 업데이트
⑤	○ 진균 참조 유전체 조립 및 유전자 예측 시스템 개발
⑥	○ 진균 전사체 분석을 위한 파이프라인 개발
⑦	○ 진균류 참조 유전체 정보 전체 정보 활용을 위한 데이터베이스 및 웹사이트 개발
⑧	○ 프로모터 및 전사인자 분석 파이프라인 개발

5. 연구결과별 기술적 수준

구분	핵심기술 수준					기술의 활용유형(복수표기 가능)				
	세계 최초	국내 최초	외국기술 복 제	외국기술 소화·흡수	외국기술 개선·개량	특허 출원	산업체이전 (상품화)	현장애로 해 결	정책 자료	기타
①의 기술					√		√			
②의 기술	√						√			
③의 기술					√			√		
④의 기술		√					√			
⑤의 기술		√						√		
⑥의 기술		√						√		
⑦의 기술		√						√		
⑧의 기술					√			√		

6. 각 연구결과별 구체적 활용계획

핵심기술명	핵심기술별 연구결과 활용계획 및 기대효과
①의 기술	국내외 연구자가 활용할 수 있는 시스템 구현
②의 기술	국내외 연구자의 활용 지원 및 사업화 촉진 계획
③의 기술	국내외 연구자의 활용 지원 및 추가 업데이트 계획
④의 기술	국내외 연구자의 통합시스템 활용 지원
⑤의 기술	국내외 연구자의 통합시스템 활용 지원
⑥의 기술	국내외 연구자의 통합시스템 활용 지원
⑦의 기술	국내외 연구자의 통합시스템 활용 지원
⑧의 기술	국내외 연구자의 통합시스템 활용 지원

7. 연구종료 후 성과창출 계획

성과목표	전략 미생물 해독	유용 유전자원 확보	사업화·실용화	표준 유전체 해독	메타지놈 분석	유전체 분석기술개발	NABIC 등록	병원성 미생물 진단마커개발	병원성 미생물 정보완성	미생물 병발생 기작 규명
최종목표						2	6			

연구기간 내 달성실적						9	56			
연구종료 후 성과창출 계획										

성과목표	사업화지표											연구기반지표								
	지식재산권			기술실시(이전)		사업화					기술인증	학술성과				교육지도	인력양성	정책활용·홍보		기타(타연구활용등)
												논문		논문평균IF	학술발표					
	특허출원	특허등록	품종등록	건수	기술료	제품화	매출액	수출액	고용창출	투자유치		SCI	비SCI					정책활용	홍보전시	
단위	건	건	건	건	백만원	건	백만원	백만원	명	백만원	건	건	건		건		명			
가중치																				
최종목표		3				2						5	-			4				80
연구기간 내 달성실적		1		2	11	2	199					2	1		3	15				165
연구종료 후 성과창출 계획		2										4								

8. 연구결과의 기술이전조건(산업체이전 및 상품화연구결과에 한함)

핵심기술명[1]			
이전형태	□무상 □유상	기술료 예정액	천원
이전방식[2]	□소유권이전 □전용실시권 □통상실시권 □협의결정 □기타( )		
이전소요기간		실용화예상시기[3]	
기술이전시 선행조건[4]			

1) 핵심기술이 2개 이상일 경우에는 각 핵심기술별로 위의 표를 별도로 작성
2) 전용실시 : 특허권자가 그 발명에 대해 기간·장소 및 내용을 제한하여 다른 1인에게 독점적으로 허락한 권리
통상실시 : 특허권자가 그 발명에 대해 기간·장소 및 내용을 제한하여 제3자에게 중복적으로 허락한 권리
3) 실용화예상시기 : 상품화인 경우 상품의 최초 출시 시기, 공정개선인 경우 공정개선 완료시기 등
4) 기술 이전 시 선행요건 : 기술실시계약을 체결하기 위한 제반 사전협의사항(기술지도, 설비 및 장비 등 기술이전 전에 실시기업에서 갖추어야 할 조건을 기재)