유전체 분석 기술: Whole Genome Sequencing (WGS)과 Metagenomics

Genomic Analysis Technologies: Whole Genome Sequencing (WGS) and Metagenomics

이예원[1], 성미선[2], 박예진[2], 양은령[2], 박예지[2], 조정현[2], 윤요한[1,2], 오혜민[1,*]
(Yewon Lee[1], Miseon Sung[2], Yejin Park[2], Eunryeong Yang[2], Yeji Park[2], Jeonghyun Cho[2], Yohan Yoon[1,2] and Hyemin Oh[1,*])
[1]숙명여자대학교 위해분석연구센터, [2]숙명여자대학교 식품영양학과
[1]Risk Analysis Research Center, Sookmyung Women's University, Seoul 04310, Korea
[2]Department of Food and Nutrition, Sookmyung Women's University, Seoul 04310, Korea

I. 서론

유전체(genome)는 암호화된 생물의 고유 정보이다. 이런 유전체를 읽어내는 기술은 1977년 Sanger 염기서열 분석(Sanger sequencing)의 개발로 시작되었으나, 한동안 발전하지 못하였다. 그러나 2000년대 초에 pyrosequencing 기술을 기반으로 한 Roche사의 'Roche 454' 개발을 필두로 Solexa sequencing, ion torrent sequencing, PacBio sequencing 등 다양한 차세대 염기서열 분석(next generation sequencing; NGS) 기술이 빠르게 발전하면서 인류 역사상 가장 빠른 발전이 이루어진 분야로 꼽힌다. 현재까지도 다양한 분야에서 Sanger sequencing 방법이 사용되지만, 유전체와 같은 거대한 DNA의 염기서열을 분석하는 것은 시간과 비용이 많이 소요되기 때문에 적합하지 않다. 반면에 차세대 염기서열 분석은 유전체를 무수히 많은 조각으로 나누어 증폭시킨 뒤 조각을 병렬로 조합하는 대규모 병렬 시퀀싱(massive parallel sequencing) 방식을 사용하여 보다 신속하고 정확하게 분석이 가능하다(Ballard et al.,2020). 또한 개발 초기에 유전체 분석을 위한 비용으로 약 30억 달러가 발생했으나, 현재는 약 200달러의 적은 비용으로 빠른 속도의 유전체 분석이 가능하기 때문에 차세대 염기서열 분석법이 널리 사용된다(우영춘 외, 2016). 차세대 염기서열 분석을 통해 한 생물의 유전체를 온전히 읽어낸 전장 유전체(whole genome)와 한 환경에 존재하는 미생물 군집의 유전체 총합인 군유전체(metagenome)의 서열을 분석할 수 있다(Jagadeesan et al., 2019). 그러나 염기서열 분석을 통해 읽어낸 유전체는 여전히 암호화되어 있기 때문에 유전체 정보를 활용하려면 읽어낸 서열을 해독 및 분석하여 생명현상에 어떻게 작용하는지 이해해야 한다. 이를 위해 발생한 학문이 생물 정보학(bioinformatics)이다. 생물 정보학은 응용수학, 통계학,

*Corresponding author: Hyemin Oh
Risk Analysis Research Center, Sookmyung Women's University, Seoul 04310, Korea.
Tel: +82-2-2077-7585
Email: odry0731@naver.com

컴퓨터 과학 등을 이용하여 분자 수준의 생물학적 정보를 수집, 분석한다. 유전자의 기능과 조절 메커니즘을 연구하는 기능 유전체학, 생물 간의 유전체를 비교하여 공통점과 차이점을 분석해 진화적 관계를 연구하는 비교 유전체학, 특정 환경에 존재하는 다수의 미생물 유전체 정보로 군집의 다양성과 분포를 연구하는 군유전체학, 유전자 발현과 RNA에 관해 연구하는 전사체학 등이 생물 정보학에 포함된다(Luscombe et al., 2001).

이런 유전체 분석과 생물 정보학의 발전은 보건, 의학, 환경 분야뿐만 아니라, 식품 제조부터 안전까지 식품 산업 분야에도 굉장한 영향을 미치고 있다. 미국에서는 식중독 발생 시 원인균의 발생 경로를 파악하기 위해 전장 유전체 분석을 활용한다(Brown et al., 2019). 국내에서도 마찬가지로 식품안전 강화를 위해 유전체 분석 기술을 활용하기 위해 노력하고 있다. 식중독균의 전장 유전체 정보를 활용한 상동성 분석, 특성 분석, 추적조사를 수행하는 국가 식중독균 유전체 정보망(national genome information network for foodborne pathogen; NGIN-F)이 그 예이다(식품의약품안전처, 2023). 그뿐만 아니라 발효 소시지, 발효 음료류, 장류와 같은 발효식품 내 균유전체를 확인함으로써 품질관리, 표시사항 준수 여부 및 제품개발에서도 유전체 분석을 폭넓게 활용하고 있다(Cha and Seo, 2017). 본 글에서는 차세대 염기서열 분석방법의 기초적인 원리부터, 전장 유전체와 군유전체 분석 결과를 활용할 수 있는 생물 정보학적 분석방법을 소개하고, 이를 바탕으로 진행한 숙명여자대학교 위해분석연구센터의 유전체 분석 연구를 소개하고자 한다.

II. 전장 유전체 분석 (Whole Genome Sequencing; WGS)

생명체가 가지고 있는 모든 유전적 정보의 전체를 genome이라고 하며, DNA 염기서열로 나타낸다. 이러한 유전체의 염기서열을 확인하는 방법으로는 전장 엑솜 분석(whole exome sequencing)과 전장 유전체 분석(WGS)이 있다. 전장 엑솜 분석은 DNA 중 단백질 비암호화 서열인 intron 영역을 제외한 exome 영역만 sequencing 하는 방법이다. 최근 주목받고 있는 전장 유전체 분석은 DNA의 모든 부위를 분석하여 구조적 변이나 유전자 발현 조절 부위의 변이까지 확인할 수 있다.

1. WGS 방법

WGS의 흐름은 크게 유전체 준비와 WGS 분석으로 나눌 수 있다. 유전체 준비 단계에서는 분석하고자 하는 DNA를 준비하기 위해 미생물 또는 생물 조직으로부터 DNA를 추출하고, DNA 조각을 무작위적으로 분절화하여 5′에서 3′ 방향의 adaptor(NGS 장비가 인식할 수 있도록 고유한 염기서열을 가진 oligonucleotide)를 접합 및 증폭시켜서 서열 분석에 필요한 library를 구축한다(Wheeler et al., 2008). NGS 장비에서 분석이 가능한 크기로 절단된 DNA 조각들인 reads의 overlap 정보를 이용하여 2개 이상의 reads가 연결된 contig로 조립되고, 이를 바탕으로 contig가 2개 이상 연결된 scaffold로 조립되는 *de novo* assembly를 통해 전체 염기서열을 구성한다(Ayling et al., 2020). *De novo* assembly는 reference없이 reads의 염기서열 정보를 재조합하여 전체 염기서열로 재구성하는 것으로, 이미 서열이 알려진 reference 서열에 reads를 정렬하여 길게 재조합하는 reference assembly와 차이가 있다(Noune, 2017).

전장 유전체는 sequencing depth를 통해 그 신뢰도를 판단하는데, 이는 reads의 정렬 횟수를 의미하며, 정렬된 reads의 개수에 '×' 기호를 붙여서 표시한다. Sequencing depth가 높을수록 genome이 더 많이 판독되어 정확도와 신뢰도가 높다고 판단하는데, 적정 depth에 대해서는 다양하게 제시되고 있다. 한 연구에서는 40× 이상이 되어야 독성 유전자가 검출된다고 하는 반면, 다른 연구에서는 4× 정도의 depth도 적당하다고 제시되기도 한다(Lambert et al., 2015; Lindsey et al., 2016). WGS를 마치면 FASTA와 FASTQ를 얻을 수 있는데, FASTA는 genome의 염기서열 또는 단백질의 아미노산 서열을 저장할 때 사용되는 형식으로 이를 활용하여

다양한 상용 프로그램으로 유전체의 추가적인 특성 분석을 할 수 있다. FASTQ는 FASTA에서 확장된 형식으로 염기서열뿐만 아니라, quality score도 함께 제시된다.

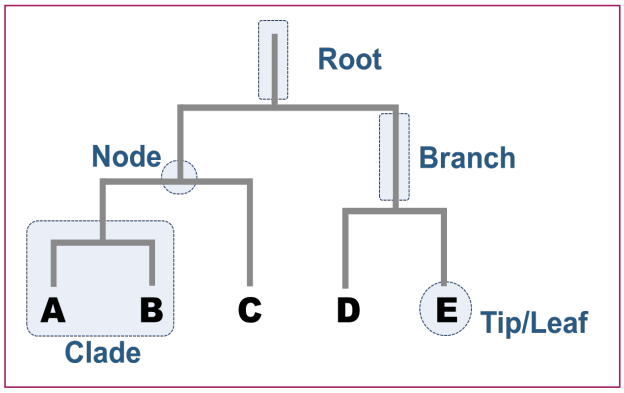
2. 전장 유전체와 유사한 유전체 탐색 및 계통수 분석

전장 유전체의 염기서열 정보를 획득한 후에는 이를 이용하여 유전체의 특성을 확인할 수 있다. 획득한 유전체와 유사한 reference 유전체의 정보를 비교하고, 유사 유전체와의 계통수(생물의 진화에 따라 여러 종의 유전적 특징의 유사성과 차이를 바탕으로 관계를 수형도로 나타낸 다이어그램) 및 유사도를 나타내는 average nucleotide identity (ANI)를 확인할 수 있으며, 염기서열의 차이를 확인하는 variant 분석, 유전적 특성(유전자의 기능, 위치 및 구조적 특성)을 확인하는 gene annotation 분석 등을 실시한다.

Reference 유전체와 염기서열을 비교하기 위해 basic local alignment search tool (BLAST)을 이용할 수 있다. 분석하고자 하는 염기서열을 query sequence라고 하며, BLAST는 query sequence와 database의 유사 염기서열 간에 구조적, 기능적 및 진화론적 관련성을 추정해낼 수 있다(Lobo, 2008). BLAST는 단백질의 아미노산 서열이나 DNA 및 RNA 서열의 nucleotide와 같은 주요 생물학적 서열 정보를 비교하기 위한 algorithm 및 program으로, query sequence를 library 또는 염기서열 database와 비교하고, 특정 임계값 이상에서 유사한 database 서열을 식별할 수 있다(Ye et al., 2006). NCBI에 등록된 database를 바탕으로 query sequence와 유사한 reference 서열을 비교 분석하는 것이 일반적이며, NCBI website, CLC Genomics Workbench 등을 활용하여 BLAST를 이용할 수 있다.

BLAST를 이용하여 나타난 유사한 유전체와의 계통수 분석을 통해 유사도를 확인할 수 있다. 계통수의 형태는 그림 1과 같이 나타날 수 있는데, root는 공동 조상이며, 아래의 tip/leaf 방향으로 갈수록 후손의 계통을 의미한다. Branch는 한 생물의 계통을 의미하여 동일 branch

그림 1. 계통수 형태 모식도



는 동일 생물의 계통이라고 볼 수 있다. Branch에서 나아가다가 갈라지는 분지점인 node는 진화가 일어난 부분으로 계통의 분화 역사를 나타낸다. Tip/leaf는 terminal nodes라고도 불리며, 현존하여 관찰되는 생물군을 뜻하고, clade는 node 이후에 갈라져 나온 모든 계통을 하나로 묶은 것이다. 계통수는 사선 방향, 원형 방향 등 다양한 형태로 표현될 수 있다. 이러한 계통수의 종류는 분석하고자 하는 유전체 간 관계를 어떤 것에 초점을 두는가에 따라 크게 3가지로 구분될 수 있다. 진화 관계만을 나타내는 cladogram, branch 길이에 따라 유전적 관계를 구분 짓는 phylogram, branch 길이가 taxa(다른 생물들과 구분되는 동질의 생물군) 사이에 시간의 흐름이 눈에 보이도록 시간적 요소가 가중치된 ultrametric tree로 나눌 수 있다(Bleidorn and Bleidorn, 2017). 이러한 계통수를 분석하는 algorithm은 거리기반 방법과 특성기반 예측방법으로 나뉜다. 거리기반 방법은 각각의 염기 치환을 계산식을 통해 분류군 간에 차이를 계통수에서의 길이로 표현한 것으로 유전자 간의 상대적 차이를 거리로 나타낸다. 이는 계산이 빠르고 데이터의 빠른 이해를 돕지만, 유의성과 오류의 가능성이 크다. Neighbor-joining (NJ), unweighted pair group method with arithmetic mean (UPGMA)가 거리기반 방법의 대표적인 예이다. 반면, 특성기반 예측방법은 많은 정보를 가진 이산집단의 data를 기반으로 서열 정렬에서 각 유전자의 위치에 대해 각각의 계통수를 만들어내고, 주어진 서열 내의 특정요소에 대한 진화를 추적하는 방식으로, 진

화적 역사를 잘 예측하지만 긴 시간이 소요된다. 이를 반영한 것이 maximum parsimony (MP), maximum likelihood (ML), Bayesian inference (BI) 등에 해당한다(Horner et al., 2004). 산출된 계통수를 토대로 분석하고자 하는 유전체 간에 계통 관계를 추론할 수 있다.

3. 전장 유전체의 Genome Annotation 분석

전장 유전체 서열은 genome annotation을 통해 genetic elements에 대한 구조적 정보(structural annotation) 및 유전자의 기능(functional annotation)을 확인할 수 있다. Genome annotation은 대부분 Prokka program을 이용한 1차 분석을 하는데, 이는 prokayote genome에 대해 유전자 annotation 및 coding sequence를 확인하는 program이다. 유전체 내에 단백질로 번역될 가능성이 있는 coding 영역 내 DNA 서열인 open reading frame (ORF)을 식별하고 그 기능성을 확인한다. Gene ontology는 다양한 유기체에 걸쳐 유전자 및 유전자 산물의 functional annotation을 위해 널리 사용되는 표준화된 system으로, 유전체의 생물학적 과정(biological process), 분자 기능(molecule function), 세포 구성 요소(cellular component)에 따라 구조화된 모델로 제시된다(GOC, 2017). 이를 분석하기 위해 상용화된 program으로는 GO (AmiGO2), EggNOG, InterPro 등이 있으며, 분석을 원하는 유전체의 FASTA 또는 protein sequence 등을 입력하면 GO ID, ontology, definition 등이 산출되어 기능을 확인할 수 있다. 이외에도 유전체의 안전성과 관련된 독성 인자, 항생제 내성 인자, 플라스미드, 파지에 대한 분석과 기능성과 관련된 2차 대사산물 분석이 가능한 프로그램도 다양하게 상용화되고 있다(표 1).

하지만, 분석된 특성이 실질적으로 표현되는지는 추가적인 실험을 통해서 검증하는 절차도 필요하다. Foudraine et al. (2021)의 연구에 따르면, *Klepsiella pneumoniae*에 대해 WGS를 시행하고, Prokka를 이용한 genome annotation을 실시하여 항생제 내성 관련 유전자(AMR gene)를 확인하고, 이를 LC-MS/MS를 통해 분석된 AMR proteome과 일치도를 비교하였다. 그 결과, DHA gene은 WGS를 통해 5개의 genome이 있다고 분석되었지만, proteome에서는 실제로 2개가 나타나서, 40%의 일치도를 보였다. 이는 유전체 분석을 통한 결과가 실제로 표현형으로 나타나는지 검증 실험을 통해 확인하는 것이 필요할 것으로 사료된다.

표 1. 유전체의 안전성 및 기능성 인자 탐색 tool

Gene/feature	Tools/database	Reference
AMR genes	ResFinder	Zankariet al., 2012
	CARD	McArthur et al., 2013
Virulence factor	VFDB	Chen et al., 2005
	VirulenceFinder2.0	-
CRISPR-Cas system	CRISPRCasFinder	Grissa et al., 2007
Bacteriocin	BAGEL4	Van Heel et al., 2018
Plasmid	Plasmidspades	Antipov et al., 2016
	PLSBD	Galata et al., 2018
Phage	PHAST	Zhou et al., 2011
	PHASTER	Arndt et al., 2016
Insertion sequence	ISFinder	Siguier et al., 2006
Second metabolites	antiSMASH	-

4. 전장 유전체의 Variant 분석

이러한 주요 인자들에 대한 분석을 진행할 경우, 모든 서열이 100% 일치할 수도 있으나, 일부 서열에서 변이가 발생하기도 하는데, 이를 확인하기 위해 variant 분석을 한다. 분석 대상의 유전체 서열과 reference 서열들의 한쪽 말단체에서부터 다른 쪽의 말단체까지 정확한 순서를 좌표화 하여 만든 genome 서열의 database인 gene mapping을 실시하고, reference 서열을 바탕으로 분석하고자 하는 유전체 서열에서 몇 개의 변이가 발생했는지 수치화하여 single nucleotide variant (SNV) 또는 single nucleotide polymorphism (SNP)으로 제시한다. SNV는 변이의 빈도수 제한 없이 일반적인 모든 변이를 포함하는 반면, SNP는 전체 서열의 1% 이상의 변이만을 포함한다(Haraksingh and Snyder, 2013). 이러한 variant 분석의 결과는 주로 안전성 관련 인자와 분석 대상의 유전체의 SNV 또는 SNP의 개수 혹은 frequency 등을 제시하는 방법으로 사용된다.

본 연구팀은 대장염에 효능이 있는 *Limosilactobacillus fermentum*을 분리하고, WGS의 variant 분석을 통해 신종 균주임을 제시하였다(Ann et al., 2023). BLAST를 이용하여 유사도가 가장 높은 *L. fermenteum* LDTM7301을 선정하고, 이와 variant 분석을 하여 주요 아미노산 서열의 변이를 확인한 결과, 엑손 부분에서 7.55%의 SNP가 나타나서 두 균주 간 차이를 제시하였다(그림 2). 또한, 본 연구팀은 high-risk *Listeria monocytogens*와 low-risk *L. monocytogenes* 분리 균주를 선정하고, WGS를 분석하여 이 두 균주의 genetic variation을 비교하는 연구를 수행하여, *L. monocytogenes*의 주요 병원성 유전자를 선정 및 variant 분석을 하고 SNV를 확인하여 *L. monocytogenes*의 risk에 따라 서열 변이와의 상관성을 제시하였다(Ryu et al., 2023; 그림 3).

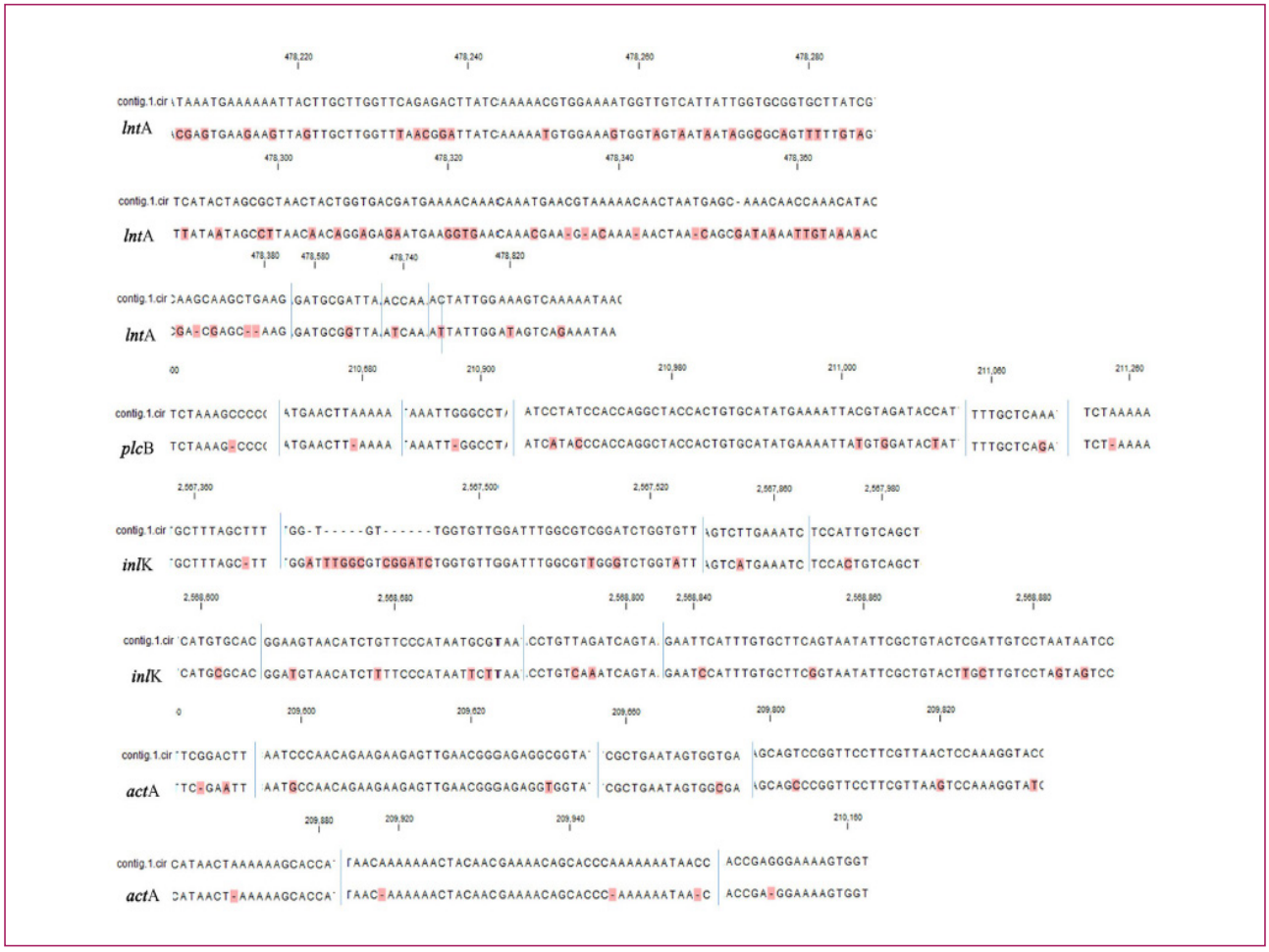
5. 전장 유전체의 비교 분석

유전체의 특성을 확인한 후에는 다양한 유전체 개체들의 특징을 비교하는 comparative genomics를 실시하

그림 2. 분리한 젖산균과 유사도가 높은 젖산균의 전장 유전체 특성 비교(Ann et al., 2023)

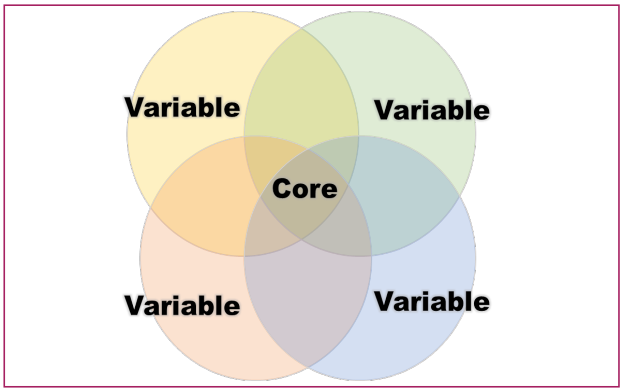
Effect	Position	Reference Allel	Alternative Allele	Depth	Codon	Amino Acid	Gene	Transcript
Frameshift_variant	16,086	C	CG	369	c.635dupG	p.Val213fs	Gene_gene17	gene17
	135,033	C	CG	391	c.96dupG	p.Pro33fs	Gene_gene151	gene151
	157,379	T	TG	319	c.516dupG	p.Arg173fs	Gene_gene175	gene175
	327,600	A	AC	216	c.538dupC	p.Arg180fs	BGV76_RS01760	Transcript_gene350
	401,859	C	CA	716	c.524dupT	p.Ile176fs	BGV76_RS02115	Transcript_gene421
	402,254	A	ATCGC	580	c.129_130insGCGA	p.Tyr44fs	BGV76_RS02115	Transcript_gene421
	402,280	GA	G	638	c.103delT	p.Ser35fs	BGV76_RS02115	Transcript_gene421
	531,259	CT	C	308	c.209delA	p.Glu70fs	Gene_gene559	gene559
	579,296	G	GC	288	c.126dupC	p.Ser43fs	Gene_gene613	gene613
	606,522	C	CG	321	c.968dupG	p.Cys324fs	Gene_gene637	gene637
	619,737	A	AAT	59	c.4_5insAT	p.Ser2fs	Gene_gene646	gene646
	664,789	T	TCCTAAATTGC AAGATTAAGTG AGCCACCCG GCCACGGGAG	93	c.410_411insCTCCCG TGGCCGGGTGGCT CACTTAATCTT GCAATTTAGG	p.Gln137fs	tnpA	Transcript_gene696
	727,563	T	TC	274	c.557dupG	p.Val187fs	Gene_gene763	gene763
	881,163	G	GC	311	c.698dupG	p.Lys234fs	Gene_gene905	gene905
	1,039,055	AG	A	427	c.1167delC	p.Ser389fs	BGV76_RS05400	Transcript_gene1078
	1,076,406	CG	C	408	c.383delG	p.Gly128fs	BGV76_RS05580	Transcript_gene1114
	1,076,406	CG	C	408	c.19delG	p.Ala7fs	BGV76_RS05585	Transcript_gene1115
	1,094,586	CG	C	259	c.75delC	p.Cys25fs	BGV76_RS05690	Transcript_gene1136
	1,094,590	CA	C	264	c.71delT	p.Val24fs	BGV76_RS05690	Transcript_gene1136
	1,212,247	AACAAAGAAAT	A	63	c.242_251delACAAAGAAAT	p.Asn81fs	Gene_gene1265	gene1265
	1,212,259	CTATT	C	50	c.254_257delTATT	p.Leu85fs	Gene_gene1265	gene1265
	1,224,178	C	CT	423	c.1164dupA	p.Ala389fs	Gene_gene1275	gene1275
	1,245,304	TA	T	387	c.204delT	p.Phe68fs	BGV76_RS06470	Transcript_gene1292
	1,252,008	A	AG	351	c.377dupG	p.Val127fs	BGV76_RS06515	Transcript_gene1301
	1,408,360	CG	C	400	c.1306delC	p.Arg436fs	Gene_gene1448	gene1448
	1,472,109	CT	C	425	c.55delA	p.Ser19fs	Gene_gene1508	gene1508
	1,494,271	AT	A	581	c.355delA	p.Ile119fs	Gene_gene1538	gene1538
	1,496,486	T	TTA	673	c.220_221insAT	p.Ser74fs	Gene_gene1541	gene1541
	1,567,158	C	CG	445	c.968dupG	p.Arg324fs	Gene_gene1609	gene1609
	1,570,436	C	CG	448	c.230dupG	p.Ser78fs	Gene_gene1612	gene1612
	1,578,725	C	CG	438	c.749dupG	p.His251fs	Gene_gene1621	gene1621
	1,589,275	C	CG	457	c.108dupC	p.Asp37fs	BGV76_RS08160	Transcript_gene1630

그림 3. 주요 병원성 유전자에 대한 *Listeria monocytogenes*의 variant 분석 결과(Ryu et al., 2023)



여 도식화할 수 있다. Pan-genome(모든 계통의 전체 유전자 집합)은 생물의 개체가 보유하고 있는 유전자 집합을 도식화하여 나타내며, core genome과 variable genome을 포함한다(그림 4; Medini et al., 2005). Core genome은 핵심 유전체로서 임의의 한 생물 분류군에 속하는 모든 개체가 지니고 있는 유전자의 집합이며, variable genome은 모든 계통의 개체에서 발현되지 않는 유전자를 의미한다. 이를 바탕으로 비교하고자 하는 유전체의 개체 간 unique genome이 무엇인지 확인할 수 있고, 이에 대한 특성 분석이 가능하며, 더 나아가 특이 유전자마커 발굴로 이어질 수 있다. Pan genome 분석은 Roary, panX, PGAP (PGAweb) 등의 상용화된 프로그램을 통해 분석이 가능하다. 최근에는 새롭게 분리한 미생물의 신종 균주임을 판단하기 위해, BLAST를 이

그림 4. Pan genome의 형태



용하여 유사도가 높은 미생물들의 염기서열들과 pan-genome을 분석하고 분리한 미생물의 unique genome을 제시하는 방법이 사용되고 있다.

Ⅲ. Metagenome 분석

Metagenome은 'complex'를 뜻하는 그리스어인 'meta'와 유전체를 뜻하는 'genome'의 합성어로, 특정 환경(공기, 물, 토양, 인체 등)에 존재하는 모든 미생물의 유전체 집합을 의미한다(Handelsman et al., 1998). 또한, 환경 시료에 존재하는 미생물의 유전체 정보를 유기적으로 연구하고 비교하는 분야를 metagenomics 또는 environmental genomics, community genomics라고 한다. Metagenome 연구는 환경에서 수집된 모든 미생물의 유전체를 배양하지 않고 직접 추출하여 분석하므로, 99% 이상을 차지하는 난배양성 미생물을 포함하여 미생물 군집의 유전적, 대사적 다양성을 분석할 수 있다(Kim et al., 2009). 장내 microbiome 분석에서도 기존의 배양법이 갖는 한계를 극복하고, 시료에 존재하는 미생물 군집의 유전체를 직접 추출하여 분류학적, 기능적 특성을 파악하는 분자생물학적 접근법이 활용된다. 특히 2000년대 이후에 NGS 분석이 보편화되면서 metagenome 관련 연구가 매우 활발하게 진행되었다(Mardis, 2008).

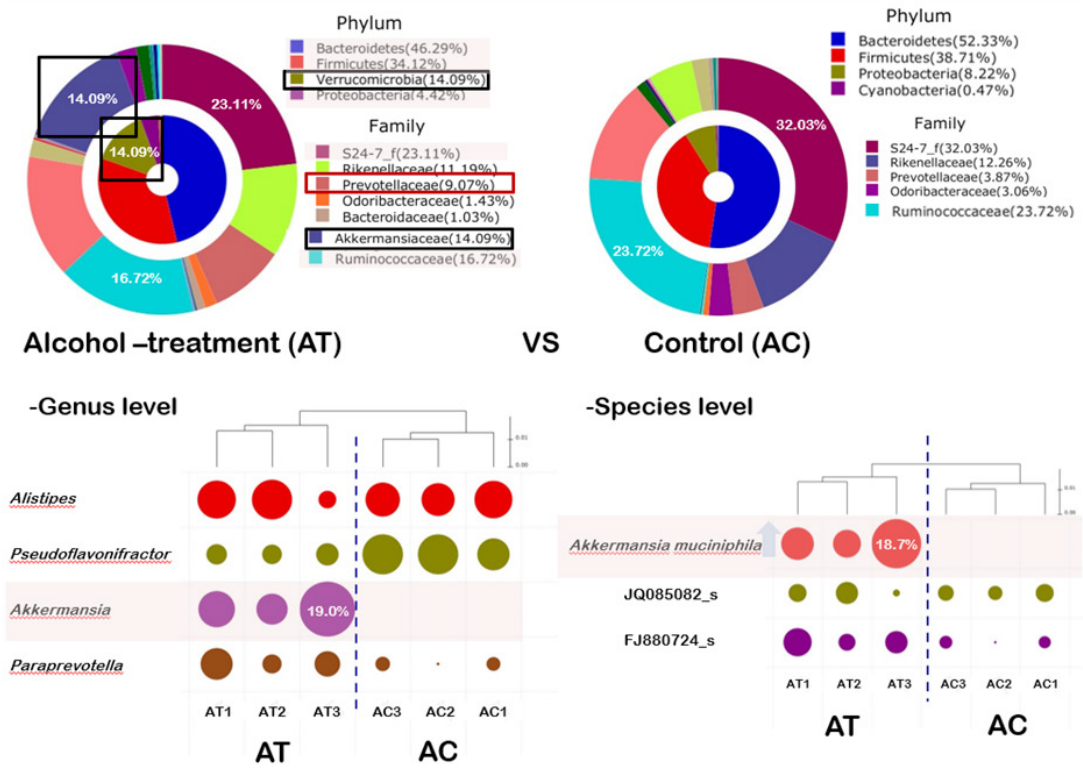
1. Metagenome 분석을 위한 Sequencing 방법

장내 microbiome 분석의 주요 목적은 1) 장내 미생물 군집의 분류학적 구성을 비교하고, 2) 각 군집의 기능적 특성 및 유전적 차이에 대한 확인이며, 이를 위해 amplicon(증폭 또는 복제된 DNA 또는 RNA 조각) sequencing과 shotgun metagenomics sequencing의 분석 방법을 활용한다(Durazzi et al., 2021).

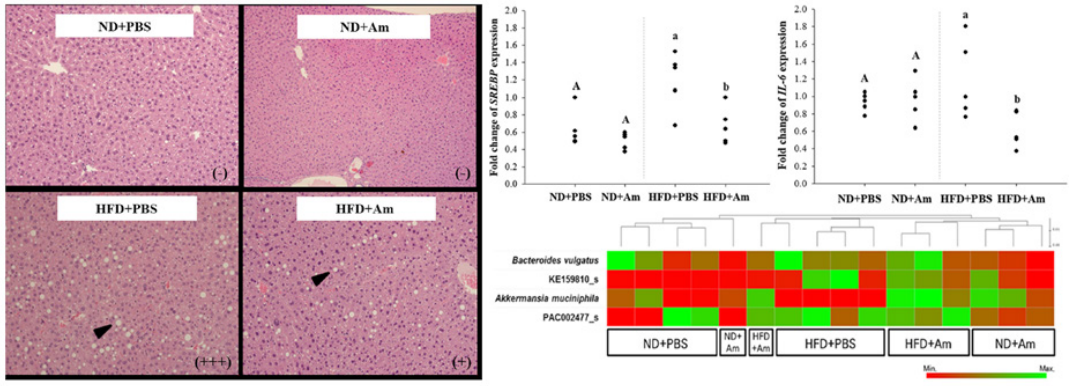
미생물 군집 분석을 통해 어떤 미생물이 존재하는지 분류학적 구성을 확인하고자 할 때는 종 동정을 위한 특정 표지 유전자(marker gene)만을 선택적으로 증폭하고, 그 증폭된 sequence를 분석하는 amplicon sequencing 방법을 통해 분석 비용과 시간을 줄일 수 있다(Liu et al., 2021). Amplicon sequencing에 사용되는 주요 표지 유전자에는 16s rRNA(원핵생물), 18s rRNA 및 ITS(진핵생물)가 있다. 이 중 세균의 군집 분석을 위해서는 16s rRNA 유전자를 표지 유전자로 사용하며, 이 유전자는 모든 세균에 존재하고 보존 영역(conserved region)과 변이 영역(variable region; V1-V9)을 적절히 포함하고 있어 계통 분석과 생태 연구에 적합하다(Lopez-Aladid et al., 2023). 16s rRNA amplicon sequencing 분석은 barcode sequence(생물 종을 구분하는 식별 코드)가 들어있는 primer를 이용하여 16s rRNA 유전자를 PCR(polymerase chain reaction)로 증폭하고, 증폭된 각각의 sequence는 16s rRNA database인 Greengenes(http://greengenes.lbl.gov), SILVA(http://www.arb-silva.de/) 또는 RDP(http://rdp.cme.msu.edu/) 등에 대조하여 그에 대한 미생물 종을 확인할 수 있다(DeSantis et al., 2006; Quast et al., 2013; Cole et al., 2014). 본 연구팀은 16s rRNA amplicon sequencing 분석을 통하여 알코올 섭취에 따른 장내 microbiome의 구성을 확인하였고, 알코올 투여군에서 *Akkermansia muciniphila*의 비율이 증가하는 것을 확인하였다(그림 5A). 이를 바탕으로 고지방 식이 마우스에 *A. muciniphila*를 투여한 결과, *A. muciniphila* 투여군에서 혈중 중성지방이 감소하였고, 간에서는 지방 합성 및 염증 관련 지표의 발현량이 감소하는 것을 확인하였으며, 이를 통해 지방간이 개선되는 것을 확인하였다(Kim et al., 2020)(그림 5B).

미생물 군집 분석을 통한 분류학적 구성 확인과 함께 군집 내에 존재하는 전체적인 기능성 유전자를 식별하기 위해서는 shotgun metagenomic sequencing 분석이 적합하다. WGS가 단일 종 또는 개체의 전체 유전체를 분석하는 방법이라면, shotgun metagenomic sequencing은 환경 시료에 존재하는 다양한 종의 유전체를 분석하여 분류학적 분석뿐만 아니라, 다양한 미생물의 기능 및 대사 경로 등을 분석하는 방법이다. Shotgun metagenomic sequencing은 시료에 존재하는 모든 DNA를 무작위로 절단하여 많은 수의 짧은 reads를 확보하여 sequencing를 용이하게 한다. 크기가 큰 DNA의 경우 전체 서열을 sequencing하는 것은 시간과 비용이 많이 필요하기 때문이다. 다양한 platform (Illumina, PacBio, Nanopore, Ion Torrent 등)을 활용하여 모든 유기체의 유전자를 종합적으로 확인할 수 있으며, gene

그림 5. 16s rRNA amplicon sequencing을 통한 장내균총 변화 확인 및 장내균총에 의한 질환 개선 효과 연구(Kim et al., 2020)



A. 알코올 섭취에 따른 장내 microbiome 비교



B. *Akkermansia muciniphila*의 지방간 개선 효능 확인

prediction과 functional annotation 과정을 통해 시료 내에 존재하는 미생물의 구성뿐만 아니라, 기능 및 대사 경로 등의 예측이 가능하다. Shotgun metagenomic sequencing 분석 범위는 단순히 세균의 군집을 분석하는 16s rRNA amplicon sequencing과 달리 plasmid, phages, 바이러스 및 곰팡이 등 시료에 존재하는 다양

한 미생물 종을 포함한다. 또한, 기능성 유전자들의 구성을 포함하여 전체 metagenome을 확인할 수 있다는 장점이 있으나, 아직은 많은 비용과 시간이 소요되어 반복 실험이 어렵고, 종 다양성이 크거나 소수의 종이 주로 존재하는 시료의 경우에는 정확한 동정에 어려움이 있다는 단점이 있다. Amplicon sequencing과 shotgun metagenomic sequencing에 대한 비교는 표 2와 같다 (Srinivas et al., 2022; Liu et al., 2021).

Metagenome 분석은 특정 환경에 존재하는 미생물 유전체 정보를 통해 분류학적 및 기능적 분석 등을 진행하는 과정으로, sequencing 과정에서 발생할 수 있는 오류를 제거하여 정확성을 높인 후 분석을 진행해야 한다. Sequencing을 통해 발생할 수 있는 오류는 주로 기술적 한계와 실험적 원인 및 bias에 의한 특정 sequence의 과대 또는 과소 표현 등이 있으며, 이러한 오류를 줄이기 위해 sequencing 결과에서 추정되는 오류 발생률을 수치화하는 Phred quality score를 통해 정확도를 확인하여 분석을 진행한다(Zhang et al., 2017). 일반적으로 Phred quality score는 $Q_{phred} = -10 \log_{10}P(error)$로 계산되며, $P(error)$는 오류 발생률이다(Goswami and Sanan-Mishra, 2022)(표 3). Q_{phred}는 일반적으로 0에서 41 사이의 수치를 나타내며, 이 수치가 클수록 sequencing 결과의 정확도는 높고, 오류 발생률은 낮다.

표 3. Phred quality score에 따른 오류 발생률 및 정확도

Phred quality score	Error	Accuracy (1-error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%

예를 들어, Q_{phred}=20의 오류 발생률은 1%를 나타내며, 이는 sequencing 결과의 정확도가 99%를 갖는 것을 의미한다. 주로 Q_{phred}=20-30일 때 정확도가 높다고 판단하며, Q_{phred}를 높이기 위해 sequencing 과정에서 사용되는 adaptor sequence 및 정확도가 낮은 sequence 부분을 제거하는 trimming 과정을 진행한다.

2. 분류학적 분석(Taxonomic Analysis)

환경 시료에 존재하는 미생물 군집의 분류학적 구성을 파악하기 위하여 사용하는 전통적인 분류 방법에는 operational taxonomic unit (OTU)을 통한 군집화 방법이 있으며, 최근에는 amplicon sequence variant (ASV)의 사용이 선호되고 있다. OTU에 의한 군집화는 서열의 유사성을 기반으로 서열을 그룹화하는 방법으로, 일반적으로 분석된 유전체 서열을 97% 수준(종, species)의 유사도를 기준으로 묶어 하나의 cluster로

표 2. 메타지노믹스 분석법 비교

	Amplicon sequencing	Shotgun metagenomic sequencing
비용	-저렴한 분석 비용 -빠른 분석 시간	-많은 비용과 시간 소요 -분석 수준에 따른 비용 차이 발생
해상도 (정확도)	-속(genus) 수준에 대한 미생물 군집 프로파일링	-종(species)과 균주(strain) 수준에 대한 미생물 군집 프로파일링
분류학적 범위	-사용하는 primer에 따른 제한 16s rRNA: 박테리아, 고세균 18s rRNA 및 ITS: 곰팡이, 효모 등	-다양한 미생물 종에 대한 분석 가능
기능적 분석	-기능적 정보 관련 예측 프로그램은 있으나, 완전한 정보를 파악할 수 없음.	-미생물 유전체의 기능적 정보 확인
분석 용이성	-분석 프로그램이 다양하며, visualization 용이	-다양한 데이터 분석이 가능 -분석 방법이 복잡하고 시간 소요
Reference DB	-다양한 DB 구축 및 활용 가능 예: Greengenes, SILVA, UNITE 등	-Whole genome 관련 정보가 아직은 부족한 실정

표현하는 방법이다(Park et al., 2014). 즉, 유사도가 높은 비슷한 종을 하나의 단위인 OTU로 그룹화하여 군집을 나눌 수 있지만, 일부 낮은 비율의 종을 정확하게 감지하기 어려워 구체적인 종 분류에는 한계점을 갖는다. 반면, ASV에 의한 군집화는 sequencing 결과에서 확인되는 단일 nucleotide 차이를 기준으로 cluster를 표현하기 때문에, OTU에 의한 군집화 방법에 비해 고해상도의 미생물 분류(종 수준)가 가능하다(Eren et al., 2013). OTU에 의한 군집화 방법의 경우, 97% 유사도를 기준으로 그룹화되어 유전적 다양성(유전적 변이 또는 혼합)의 일부 정보가 누락될 수 있으나, ASV에 의한 방법은 미세한 유전적 다양성을 인식할 수 있어 미생물 구성을 종 수준까지 더 정확하게 분류될 수 있다. 그러나, ASV에 의한 방법은 sequencing 오류 및 바이어스에 의한 차이도 민감하게 처리하므로 실제로는 동일한 cluster임에도 다르게 분류될 수 있다. 최근에는 OTU에 의한 clustering 결과와 ASV에 의한 clustering 결과를 함께 비교하여 제시하는 연구 논문이 많이 게재되고 있다(Jeske & Gallert, 2022; Tipton et al., 2022).

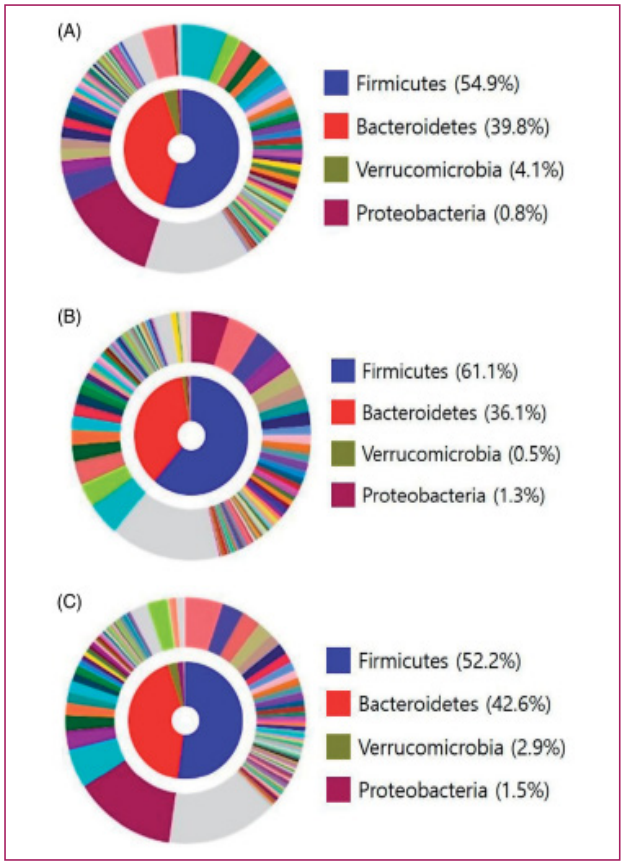
OTU와 ASV에 의한 미생물 군집화 과정에는 크게 세 가지 방법이 있다. 먼저, Greengenes, SILVA, RDP 등의 database에 전체 sequence reads를 대조(mapping)해 보는 방식인 Closed-reference clustering 방법이 있으며, 유사도 기준에 적합한 reads만 reference sequence에 mapping되고, 그 외의 reads는 제거된다. 이 방법은 상대적으로 빠르고 재현성은 있으나, database의 오류에 취약하고, database에 존재하지 않는 새로운 종이나 군집이 누락될 수 있는 제한점을 갖는다. 반면에 database 없이 분석된 sequence reads를 유사도를 기준으로 묶는 방법인 *de novo* clustering 방법이 있다. 이 방법은 정보의 누락 없이 새로운 종 또는 군집을 발견하는데 유용하다는 장점이 있으나, 결과의 재현성이 높지 않아 시료 간 또는 다른 연구와의 비교가 불가능하다는 제한점이 있다. 이 두 가지 방법의 제한점을 보완한 방법으로 open-reference clustering 방법이 있으며, database에 먼저 reads를 mapping하고, mapping되지 않은 reads에 한하여 *de novo* clustering을 진행하는 방법이다. 이 방법은 시료 특성에 따라 달라질 수 있는데, database에 이미 다양한 연구 결과를 통해 보고된 데이터가 확보되어 있는 장내미생물 시료의 분석일 경우에는 closed-reference 방법과 유사하게 수행될 수 있다.

본 연구팀에서는 vitamin E의 섭취와 장내 microbiome의 상관성을 확인하기 위한 연구를 진행하였으며(Choi et al., 2020), 16s rRNA amplicon sequencing을 통해 OTU로 군집화하여 비교하였다. 그 결과, 고농도의 vitamin E 투여군에서 Firmicutes (Bacillota)와 Bacteroidota의 비율을 감소시키는 것을 확인하였다. 즉, 고농도의 vitamin E 투여군에서 Firmicutes의 비율이 감소하였고 이를 통해 vitamin E 섭취가 장내 microbiome 구성에 영향을 주는 것을 확인하였다(그림 6). 또한, 본 연구팀은 다양한 식이에 따른 장내 microbiome의 변화를 파악하기 위하여 16s rRNA amplicon sequencing을 통해 OTU로 군집화하여 비교하였고(Oh, 2020), 식이에 따라 Bacteroidota와 Firmicutes의 비율 차이 및 분변 이식을 통해 장내 microbiome 구성에 따른 발효유의 항콜레스테롤 효능 차이를 확인하는 연구를 수행하였다.

3. 미생물 다양성 분석(Diversity Analysis)

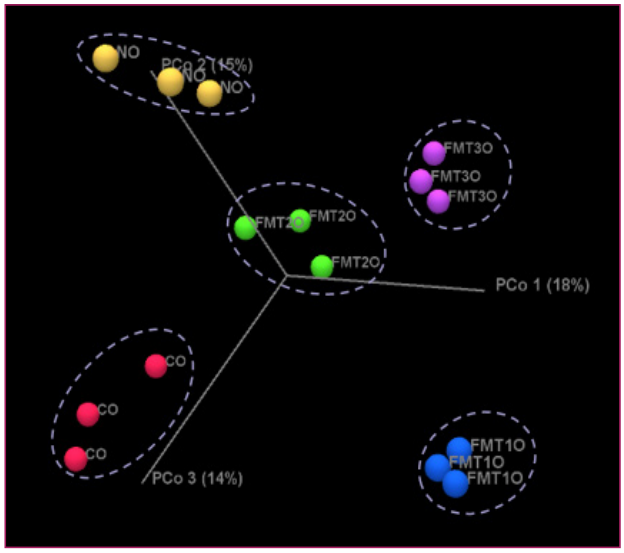
미생물 군집의 구성을 파악했다면 구성하고 있는 미생물 군집의 종 다양성이 서로 어떤 차이를 나타내는지를 비교하는 다양성 분석(diversity analysis) 과정을 수행한다. 시료 내 다양성을 나타내는 alpha diversity와 시료 간 다양성을 나타내는 beta diversity가 있다. 예를 들어, 어떤 특정 시료의 종 다양성이 높고 낮음을 alpha diversity를 통해 확인할 수 있고, 두 시료 간 미생물 구성이 얼마나 다른지 beta diversity를 통해 확인할 수 있다(Andermann et al., 2022). 다양성 분석의 계산에는 다양한 종류의 계산법이 사용된다. Alpha diversity 분석에는 Chao1, Observed features, Shannon index, Simpson's index 등을 사용할 수 있

그림 6. OTU에 의한 군집화를 사용한 분류학적 분석 예시 (Choi et al., 2020)



다. 이때, Chao1과 Observed features는 종 수의 다양성(species richness)를 나타내고, Shannon's index와 Simpson's index는 종간 비율의 다양성(species evenness)를 나타낸다. Shannon's index는 종의 풍부도와 고르게 분포된 정도를 고려하여 계산하는 방법으로 값이 클수록 다양성이 높음을 나타낸다. Simpson's index는 특정한 종이 얼마나 풍부하게 존재하는지와 그 종의 상대적 중요성을 고려하여 계산하는 방법으로 값이 작을수록 다양성이 낮음을 나타낸다. Beta diversity 분석에는 Jaccard distance, Bray-Curtis distance, unweighted UniFrac distance, weighted UniFrac distance 등이 있다. Jaccard distance는 시료 간 종의 존재 유무를 계산하여 비교하는 방법으로 유사도와 다양성을 비교하기 위한 통계법이다. 시료 간의 차이를 종이 있는지, 없는지 비교하고, 풍부도에 대한 고려는 하지 않으며, 0부터 1까지 범위로 나타내되 1에 가까울수록 두 시료가 유사함을 의미한다. Unweighted UniFrac distance는 Jaccard distance와 유사하게 특정 종의 존재 유무만 고려하여 종 구성 차이를 측정한다. Jaccard distance와 unweighted UniFrac distance는 새로운 종 또는 드문 종의 구성 차이를 중점적으로 파악할 때 유용하다. 반면 Bray-Curtis distance는 시료 간 종의 풍부도를 계산하여 비교하는 방법으로 두 군간 미생물 구성의 차이를 정량화하기 위한 통계법이다. 0부터 1까지 범위로 나타내며, 0에 가까울수록 두 시료가 유사함을 의미한다. Weighted UniFrac distance는 sequence의 양과 중요도를 고려하며, 시료 간 특정 종 또는 풍부도 차이를 고려하여 유사도를 측정한다. Bray-Curtis distance와 weighted UniFrac distance는 종의 풍부도와 상대적 중요성을 고려하여 유사도를 파악할 때 유용하다. 다양성 분석은 연구 목적과 확인하고자 하는 가설에 따라 사용하는 방법에 차이가 있다. 이러한 시료 간 유사성을 시각화하기 위해 2, 3차원의 좌표에 PCoA (Principal coordinates analysis) plot으로 표현할 수 있다(그림 7). 좌표상의 거리가 가까울수록 시료 간 다양성이 유사하다는 것을 나타낸다.

그림 7. 그룹 간 유사성을 가시화하기 위한 PCoA 분석 예시



Ⅳ. 결론

유전체 분석 기술은 꾸준히 발전하고 있으며, 연구의 목적에 따른 적절한 분석 방법을 선택하여 진행함으로써 원하고자 하는 결과를 도출할 수 있다. 유전체 분석 기술로서 WGS 및 metagenome 관련 연구는 유전체의 특성을 확인하거나, 특정 환경에서의 미생물 군집 및 다양성을 파악하고, 신규 미생물 또는 유전체를 발견하는데 의의가 있다. WGS 분석을 통해 신종 유전체에 대한 database를 확보하거나, 기존 유전체와의 비교를 통한 특성 분석으로의 연구 확장이 가능하다. 또한, 장내 환경과 질병 발생 간의 상관성 규명연구에서 장내 미생물에 대한 metagenome 연구가 활용될 수 있으며, 기능성 물질의 효능을 높이는 장내 미생물 구성을 파악하고, 그에 따른 특정 미생물 발굴 및 신규 유전자 marker의 발굴에도 활용될 수 있을 것으로 사료된다.

참고문헌

1. 우영춘, 김영우, 김학영, 배승조, 김홍연, 정호열, 최완. 2016. 유전체분석용 슈퍼컴퓨팅 시스템 기술. 정보과학회지, 34(2), 43-56.
2. 식품의약품안전처. 2023.09. 국가 식중독균 유전체 정보망. https://nginf.nifds.go.kr/cm/main.do
3. Andermann T, Antonelli A, Barrett RL, Silvestro D. 2022. Estimating alpha, beta, and gamma diversity through deep learning. Front. Plant Sci. 13.
4. Ann S, Choi Y, Yoon Y. 2023. Comparative genomic analysis and physiological properties of *Limosilactobacillus fermentum* SMFM2017-NK2 with ability to inflammatory Bowel Disease. Microorganisms 11(3): 547.
5. Ayling M, Clark MD, Leggett RM. 2020. New approaches for metagenome assembly with short reads. Briefings. Bioinf. 21(2): 584-594.
6. Ballard D, Winkler-Galicki J, Wesoły J. 2020. Massive parallel sequencing in forensics: Advantages, issues, technicalities, and prospects. Int. J. Legal Med. 134(4): 1291-1303.
7. Bleidorn C, Bleidorn C. 2017. Phylogenetic analyses. Phylogenomics: An Introduction, 143-172.
8. Brown E, Dessai U, McGarry S, Gerner-Smidt P. 2019. Use of whole-genome sequencing for food safety and public health in the United States. Foodborne Pathog. Dis. 16(7): 441-450.
9. Cha ITA, Seo MJ. 2017. Analysis techniques for fermented foods microbiome. Food Sci. Ind. 50(1): 2-10.
10. Choi Y, Lee S, Kim S, Lee J, Ha J, Oh H, Lee Y, Kim Y, Yoon Y. 2020. Vitamin E (α-tocopherol) consumption influences gut microbiota composition. Int. J. Food Sci. Nut. 71: 221-225.
11. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Titus Brown C, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42(Database issue), D633-D642.
12. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and work-bench compatible with ARB. Appl. Environ. Microbiol. 72: 5069-72.

13. Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, Cesare AD. 2021. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. Sci. Rep. 11: 3030.
14. Elbrecht V, Vamos EE, Steinke D, Leese F. 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ. 6: e4644.
15. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML, Freckleton R. 2013. Oligotyping: differentiating between closely related microbial taxa using 16s rRNA gene data. Methods Ecol. Evol. 4: 1111–1119.
16. Goswami K, Sanan-Mishra N. 2022. Chapter 7. RNA-seq for revealing the function of the transcriptome. Bioinformatics: Methods Appl. 105–129.
17. Haraksingh RR, Snyder MP. 2013. Impacts of variation in the human genome on gene regulation. J. Mol. Biol. 425(21): 3970–3977.
18. Horner DS, Pesole G. 2004. Phylogenetic analyses: a brief introduction to methods and their application. Expert Rev. Mol. Diagn. 4: 339–350.
19. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, Chaffron S, Vossen JVD, Tang S, Katase M, McClure P, Kimura B, Chai LC, Chapman J, Grant K. 2019. The use of next generation sequencing for improving food safety: Translation into practice. Food Microbiol. 79: 96–115.
20. Jeske JT, Gallert C. 2022. Microbiome analysis via OTU and ASV-based pipelines-A comparative interpretation of ecological data in WWTP systems. Bioengineering (Basel) 9: 146.
21. Kim JM, Song S, Jeon CO. 2009. Deciphering functions of uncultured microorganisms. Korean J. Microbiol. 45: 1–9.
22. Kim S, Lee Y, Kim Y, Seo Y, Lee H, Ha J, Lee J, Choi Y, Oh,H, Yoon Y. 2020. *Akkermansia muciniphila* prevents fatty liver disease, decreases serum triglycerides, and maintains gut homeostasis. App. Environ. Microbiol. 86: e03004–19.
23. Lambert D, Carrillo CD, Koziol AG, Manninger P, Blais BW. 2015. Genesippr: A rapid whole-genome approach for the identification and characterization of foodborne pathogens such as priority Shiga toxigenic *Escherichia coli* . PLoS One. 10: 1–19.
24. Lindsey RL, Pouseele H, Chen JC, Strockbine NA, and Carleton HA. 2016. Implementation of whole genome sequencing (WGS) for identification and characterization of Shiga toxin-producing *Escherichia coli* (STEC) in the United States. Front Microbiol. 7: 1–9.
25. Liu YX, Qin Y, Chen T, Lu M, Qian X, Guo X, Bai Y. 2021. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell 12: 315–330.
26. Lobo I. 2008. Basic local alignment search tool (BLAST). Nature Education 1(1).
27. Lopez-Aladid R, Fernandez-Barat L, Alcaraz-Serrano V, Bueno-Freire L, Vazquez N, Pastor-Ibanez R, Palomeque, A, Oscanoa P, Torres A. 2023. Determining the most accurate 16S rRNA hypervariable region for taxonomic identification from respiratory samples. Sci. Rep. 13: 3974.
28. Luscombe NM, Greenbaum D, Gerstein M. 2001. What is bioinformatics? An introduction and overview. Yearbook of Medical Informatics 10(01): 83–100.

29. Mardis ER. 2008. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9: 387-402.
30. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Curr. Opin. Genet. Dev. 15: 589-594.
31. Noune C. 2017. Dynamics, diversity and evolution of Baculoviruses (Doctoral dissertation, Queensland University of Technology).
32. Oh H. 2020. Correlation between gut microbiota composition and effect of fermented milk on anti-cholesterol, identification of enhancing bacteria. [Doctoral dissertation, Sookmyung Women's University].
33. Oh H, Lee HJ, Lee J, Jo C, Yoon Y. 2019. Identification of microorganisms associated with the quality improvement of dry-aged beef through microbiome analysis and DNA sequencing, and evaluation of their effects on beef quality. J. Food Sci. 84: 2944-2954.
34. Park SH, Choe HS, Gwon SY, Yun SR. 2014. Comparative performance evaluation of OTU binning methods for metagenomic sequence analysis. KIISE 32: 46-53.
35. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41(Database issue): D590-D596.
36. Ryu J, Choi Y, Yoon Y. 2023. Comparison of genetic variations between high-and low-risk *Listeria monocytogenes* isolates using whole-genome de novo sequencing. Front. Microbiol. 14.
37. Srinivas M, O'sullivan O, Cotter PD, van Sinderen D, Kenny JG. 2022. The application of metagenomics to study microbial communities and develop desirable traits in fermented foods. Foods 11: 3297.
38. The Gene Ontology Consortium. 2017. Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res. 45(D1): D331-8.
39. Tipton L, Zahn GL, Darcy JL, Amend AS, Hynson NA. 2022. Hawaiian fungal amplicon sequence variants reveal otherwise hidden biogeography. Microb. Ecol. 83: 48-57.
40. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189): 872-876.
41. Ye J, McGinnis S, Madden TL. 2006. BLAST: Improvements for better sequence analysis. Nucleic Acids Res. 34(suppl_2): W6-W9.
42. Zhang S, Wang B, Wan L, Li LM. 2017. Estimating phred scores of Illumina base calls by logistic regression and sparse modeling. BMC Bioinf. 18: 335.