

Shotgun Metagenome 의 최근 동향과 분석 트렌드

미생물학자라면 시퀀싱 실험인 메타지놈(Metagenome) 또는 메타지노믹스(Metagenomics)라는 말을 한번쯤 들어봤을 것이다. 메타지놈은 과연 어떤 실험 분석일까? 메타지놈은 Meta 와 Genome 의 합성어인데 '초월한', '넘어서' 등의 의미를 가지는 접두어 'Meta'와 유전체를 뜻하는 'Genome'이 합쳐져 혼합된 미생물 군집의 유전체를 한꺼번에 분석하여 연구하는 기술을 말한다 [1]. 20 세기에는 PCR 로 16S rRNA 유전자만 증폭하거나 1 세대 시퀀싱 기술인 Sanger sequencing 을 이용해 전체 DNA 를 읽는 방식으로 시작된 미생물 유전체 분석은 2000 년대 초반 NGS, Next Generation Sequencing 의 등장으로 본격적인 발전을 이루기 시작했다 [2].

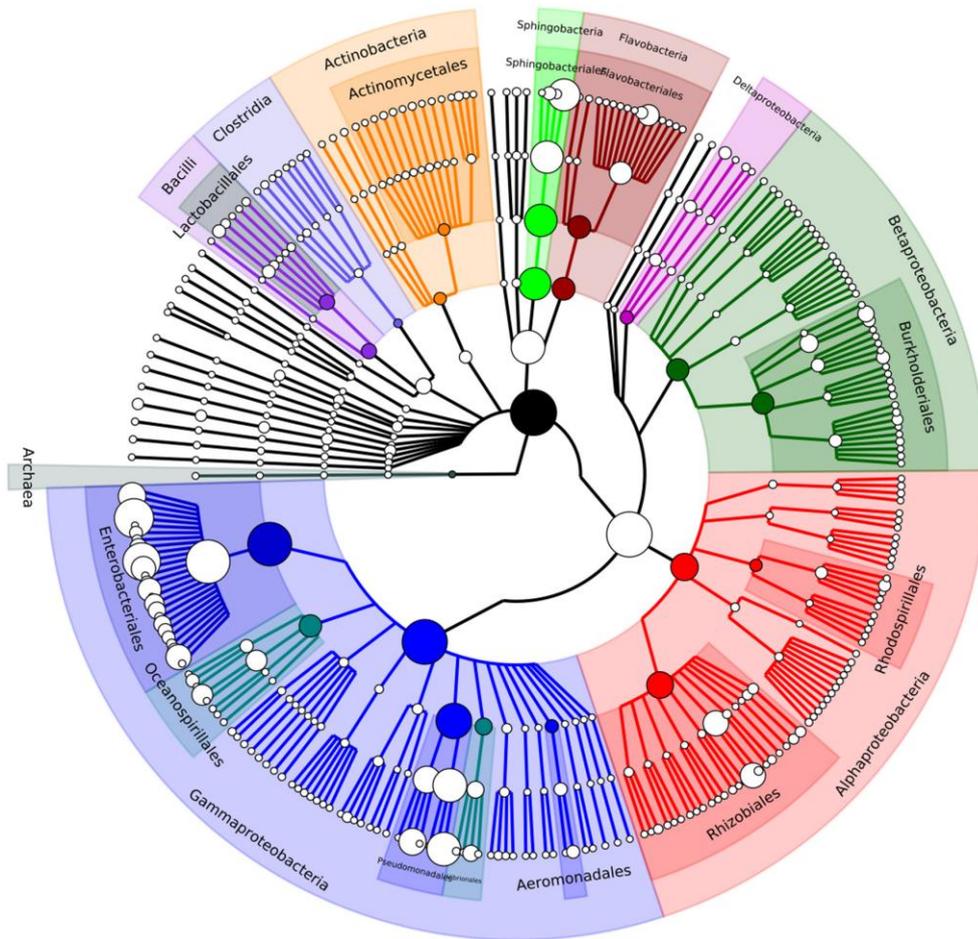


그림 1. Diversity of Gut Microbiome Phylogenetic Tree [3]

특히 2007 년에 'Human Microbiome Project'가 시작되었고 이 프로젝트에서 진행된 인체 서식 미생물 군집 연구가 성공적으로 마무리되면서 Metagenome sequencing 은 더욱 주목받았다 [4]. 이후 2010 년대에 시퀀싱 단가가 기하급수적으로 낮아지면서 일반 연구자들도 메타지놈을 활용하기 시작하였다. 현재까지 발전된 메타지놈 분석은 크게 두 가지로 나뉘며, 16S rRNA 유전자의 가변 영역만 분석하여 미생물 동정을 하는 Amplicon sequencing 과 전체 genomic DNA 를 읽는 Shotgun metagenome sequencing 으로 구분된다 [5]. 해당 기술자료에서는 Amplicon 과 Shotgun sequencing 의 차이를 설명하고 Shotgun metagenome sequencing 에 대해 자세히 살펴보고자 한다.

Amplicon 과 Shotgun Metagenome 의 차이점

Amplicon sequencing 의 한 종류인 16S rRNA amplicon sequencing 은 박테리아(Bacteria)와 고균(Archaea)에서 발견되는 16S rRNA gene 영역을 target 으로 분석하고, 곰팡이(Fungi)에서는 ITS 영역, 원생생물(Protist)에서는 18S rRNA 영역을 분석한다 [6]. 이러한 Amplicon sequencing 은 target 영역을 증폭한 후 대량으로 병렬식 서열 분석을 수행하는 방식으로, 특정 미생물을 대표하는 유전자를 기준으로 분류하기 때문에 Marker gene analysis 라고도 한다 [7]. Classification 방식으로 OTU picking 방식과 ASV error read 방식으로 진행이 가능하다 [5, 9]. 장점으로는 빠르고 정량적이며 경제적인 비용으로 분석이 가능하다는 것이다. 또한 분류학적으로 의미 있는 16S rRNA V3-V4 영역을 선택적으로 증폭하여 분석하기 때문에 host DNA 의 contamination 을 최소화할 수 있으며, 분류학적으로는 속(genus) 단계까지 분석이 가능하다 [8]. 다만, 기존 Amplicon sequencing 은 16S rRNA 에서 V3-V4 영역만 분석하기 때문에 genus level 의 분류만 가능하다는 한계가 있으며, 이를 보완한 방법이 long read sequencing 방식의 16S rRNA full length sequencing 이다.

이와 달리, Shotgun Metagenomics 는 특정 유전자를 타겟팅 하지 않는 untargeted (shotgun) 방식의 미생물 군집 whole genome sequencing 이다. 즉, 16S, 18S, ITS1/2 처럼 특정 유전자 영역만을 분석하는 것이 아니라 샘플 내의 모든 genomic DNA 를 분석할 수 있다 [8]. Shotgun metagenome sequencing 방법은 조직, 체액, 변 등의 인체 샘플뿐만 아니라 토양, 하수, 채취샘플(swab, tape) 등 다양한 환경 샘플로 진행이 가능하다. DNA 를 추출한 후, 라이브러리 제작을 위해 무작위로 조각내어 단편화하며, 이후 키트로 adapter 와 index 를 붙인 뒤 시퀀싱 플랫폼을 이용하여 분석된다. 주로 Illumina 의 MiSeq 이나 NovaSeq 장비를 사용하여 시퀀싱 한다. Sequencing read 는 pre-processing 을 거쳐 여러 파이프라인으로 분석하는데 assembly-based profiling 이나 mapping-based profiling 로 분석한다 [8]. Shotgun metagenome 이 완료되면 종 분류와 상대적 분포도 (relative abundance) 분석뿐만 아니라 유전자의 유무 및 기능 분석과 같은 다양한 분석이 가능하다. 또한, Amplicon sequencing 분석과 달리 Shotgun sequencing 분석에서는 species 또는 strain level 의 분류학적 정보를 획득할 수 있다 [8].

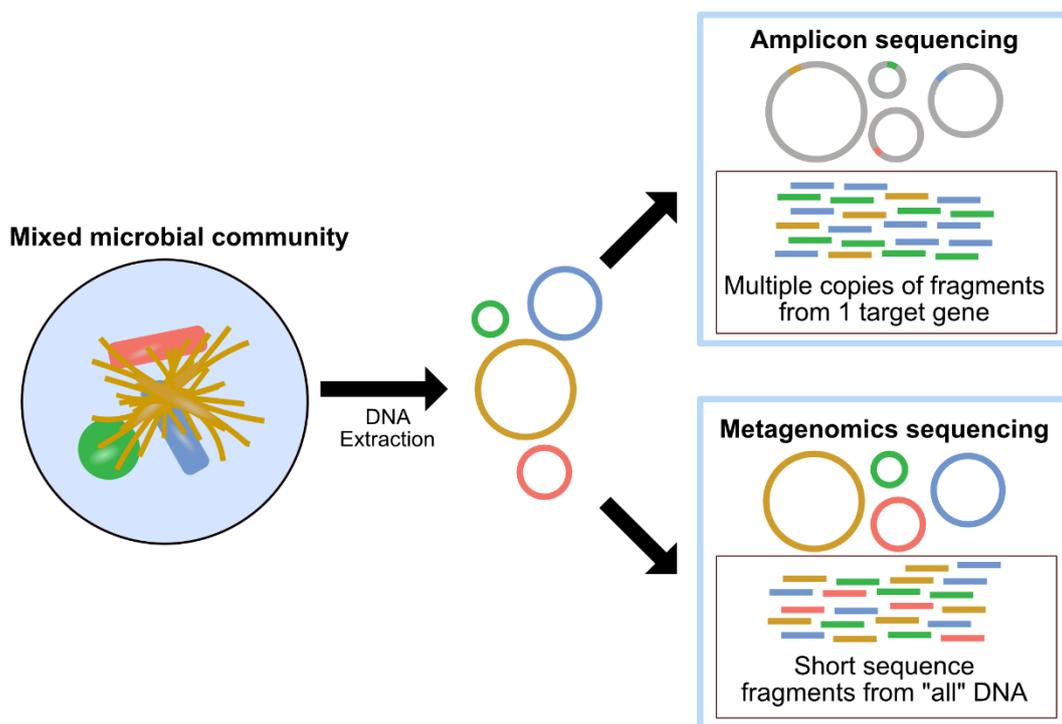


그림 2. Amplicon Sequencing vs Shotgun Sequencing in Metagenome [9]

Shotgun Metagenome 의 프로세스

Shotgun Metagenome 분석에는 여러가지 진행 방식이 있지만 가장 많이 사용되고 일반적으로 통용되는 진행 과정을 아래에 설명하고자 한다.

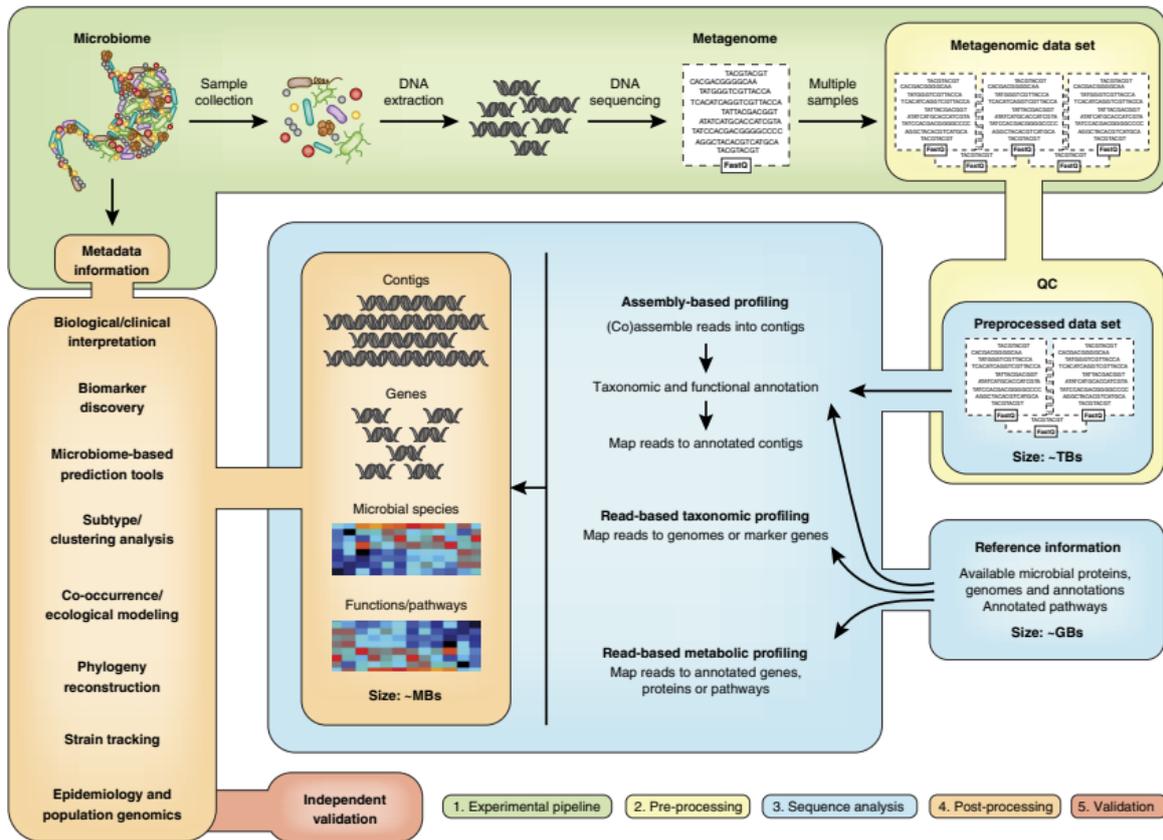


그림 3. Workflow of Shotgun Metagenomic processes [8]

1) Sample prep. & DNA extraction

시퀀싱에서 가장 중요한 요소는 샘플의 퀄리티이기 때문에 메타지놈 샘플의 채취 및 보관 프로토콜은 매우 중요하다. 동일 그룹 내 샘플을 동일한 조건으로 collection 하면 실험 변수를 최소화할 수 있으며, 샘플에 대한 충분한 양과 반복 수 (replicates)를 두는 것을 권장한다. 또한, 샘플 타입에 따라 알맞은 보관 온도와 DNA 추출키트를 사용하는 것도 중요하다. 특히, 메타지놈 샘플은 다양한 종이 혼재되어 있기 때문에 bead beating 과 같은 물리적인 분쇄도 자주 이용된다 [8].

2) Library prep. & Quality Control

DNA 추출 후 QC 는 간략하게 진행된다. 일반적으로 host DNA 와 미생물 DNA 가 혼재되어 있기 때문에 샘플 QC 에서의 Nanodrop 과 Qubit 을 통해서 샘플의 양과 질을 평가한다. QC 통과한 샘플로 라이브러리 제작을 하게 되는데 이바ิโอ젠에서는 NEB Ultra DNA Library kit 을 이용해 library 를 제작한다. 이후 Library QC 를 거치는데 Tape station 의 peak pattern 으로 확인한다 [10].

3) Sequencing

QC를 통과한 라이브러리가 준비되면 High-throughput NGS 시퀀싱을 진행하게 된다. Illumina, PacBio, Oxford 등 다양한 시퀀싱 플랫폼이 존재하지만 Shotgun metagenome 에서는 short read 방식인 Illumina 의 플랫폼으로 진행한다. 이바이오젠에서는 NovaSeq 장비를 이용하여 paired end 150bp 로 시퀀싱 하고 있다 [10].

4) Pre-processing of sequencing read data

시퀀싱 플랫폼을 통해 생산된 raw data 는 낮은 quality 의 read 를 필터링하고 trimming 하는 과정을 거친다. 주로 FastQC, Trimmomatic, Cutadapt 등의 툴이 이용된다. 이후, 높은 퀄리티로 필터링 된 read 들은 추가로 host DNA 오염을 제거하는데 DeconSeq 이나 KneadData 같은 툴을 이용한다 [8, 11].

5) Metagenome assembly & Profiling

필터링된 높은 퀄리티의 read 는 assembler 를 통해서 contig 형태로 정리한다. Contig 는 contiguous sequence 의 약자로 연속적인 DNA 시퀀스가 모여 세트를 이루어 하나의 DNA 구간을 나타내는 것을 말한다. 주로 MEGAHIT 나 SPAdes assembler 툴로 genome 을 재구성하게 된다. 그후 assembly 가 끝난 수천개의 contig 들을 특정 종으로 그룹핑하는 것을 contig binning 이라 하고 MetaWatt 나 SCIMM 의 알고리즘을 통해서 clustering 된다. Assembly 와 binning 이 완료되면 profiling 이 가능해지는데 종에 따른 상대적 분포도를 분석하는 taxonomic profiling 은 Kraken2, MetaPhlAn 툴으로 classification 이 가능하고 functional annotation 과 pathway analysis 는 KEGG 데이터베이스 기반 또는 HUMAnN 파이프라인으로 분석이 가능하다 [8, 12, 13].

6) Post-processing & Statistical Analysis

메타지놈 분석에는 다양한 툴을 사용할 수 있으며, 적용한 툴과 관계없이 분석 데이터 output 에는 샘플 대비 feature(species, taxa, genes)의 결과가 포함된다 [8]. 분석된 결과는 여러 bioinformatics 툴을 통해 통계처리 후 visualization 이 가능하다. Clustering heatmap 이나 분포도를 나타내는 phylogenetic tree, volcano plot, bar plot, Krona chart 등의 그래프로 표현이 가능하다.

7) Validation

분석 결과에 대한 통계적 유의성으로도 데이터 신뢰도를 확인할 수 있지만 추가적인 검증을 통해서 확인하는 방법도 있다. Shotgun metagenome 분석 진행 후에 16S rRNA amplicon-seq 이나 qRT-PCR 을 통해 교차검증이 가능하고 동일 샘플에서 단일 균주로 배양하여 순수분리하는 방법으로도 검증할 수 있다 [8].

Metagenome 데이터의 Multi-Omics

NGS 기술의 발달과 시퀀싱 비용 감소로 인해 메타지놈 연구가 활발히 진행되면서 관련 논문도 지속적으로 증가하고 있다. 기본적인 데이터베이스가 쌓이면서 일반적인 메타지놈보다 다른 Omics 와 연계된 통합 분석으로 연구가 확장되고 있다 [4]. 멀티오믹스 분석에서는 다른 플랫폼 간의 연계성이 필요하기 때문에, marker gene 으로만 분석하는 16S amplicon sequencing 보다는 더 많은 정보를 제공하는 shotgun metagenome 이 적합하다고 할 수 있다 [7]. 그 중 가장 많이 진행되고 있는 것이 Metatranscriptome 과 Metabolomics 이다.

Metatranscriptome 은 전체 microbial community 에 대한 유전자 발현 프로파일링 분석이다. Metagenome 과 달리 Metatranscriptome 은 발현된 RNA 를 타겟하여 캡처하는 방식으로 진행된다. Shotgun metagenome 에서는 전체 genomic DNA 에 대한 정보를 읽기 때문에 각 taxon 에 대한 gene 정보도 포함 되어있어 Metatranscriptome 과의 통합분석이 더욱 용이하다 [4]. Metatranscriptome 분석방법은 rRNA depletion 및 라이브러리 제작을 통해 복잡한 microbial community 의 유전자 발현을 분석하게 된다. Taxon 별 분석의 경우, *de novo* assembly 를 통해 종별 contig 로 분류한 뒤에 유전자 분석을 진행하게 된다 [14]. Metatranscriptome 데이터는 Trinity 와 RSEM 과 같은 전용 툴을 사용하여 주로 분석한다. Metagenome 과 Metatranscriptome 의 연관성을 integration 하는 것은 아직까지 연구자들이 맞서고 있는 한계점이다. 두 개 연구의 통합 분석은 주로 한가지 종이나 종 그룹에 대한 functional annotation 위주로 활용되며 genomic DNA 와 RNA 샘플에서의 기능적 유사성을 확인할 수 있다 [14].

Metabolomics 는 대사과정에서 발생하는 downstream metabolite 들을 고성능 질량분석기로 분석하는 방식으로 GC-MS, LC-MS, CE-MS, NMR 등으로 분석한다 [14]. 주로 많이 사용되는 분석법은 LC-MS/MS 방식이다. 현재까지 알려진 Metabolomics 데이터베이스 (ex. ChemSpider, mzCloud, MassBank)는 종별 구분을 하지 않기 때문에 여러 종이 혼합된 샘플에서도 진행이 가능하다. 이러한 이유로 metagenomics 와 분석이 수월하며, 주로 KEGG 를 통한 pathway 분석과 연동하여 분석을 진행한다.

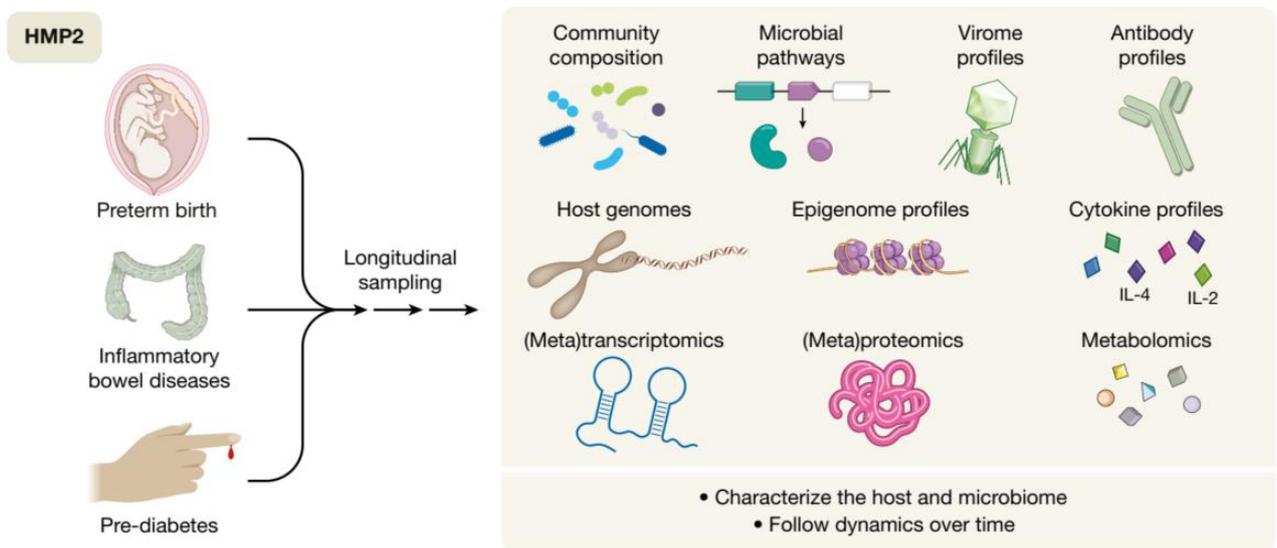


그림 4. Human microbiome multi-omics with various platforms [4]

ExMEGA (Excel based MEtaGenome Analysis)

이바이오젠에서 개발한 ExMEGA 는 엑셀기반의 프로그램으로, 엑셀 내에서 ExMEGA 툴바가 추가되어 사용이 가능하다. 다양한 환경내 존재하는 미생물 군집의 생태학적 의미를 분석하는 도구로 활용 가능하다. 또한, Metagenome 데이터에 대한 다양한 분석 기능과 데이터 시각화를 용이하게 해주며, 데이터 분석과 엑셀 사용에 익숙하지 않은 연구자들도 쉽게 활용할 수 있다. ExMEGA 에서 사용가능한 기능으로는 Taxon 별 filtering 과 Relative abundance, Pie chart, DA analysis, Volcano Plot, Venn Diagram, Significant Taxon 이 있다. 이와 더불어서 ExMEGA GraphicPlus 에서 사용가능한 기능으로는 메타지놈에서 핵심적인 Krona Chart 를 비롯하여 PCoA 분석, Bar plot, Clustering Heatmap 이 제작 가능하다 [10].

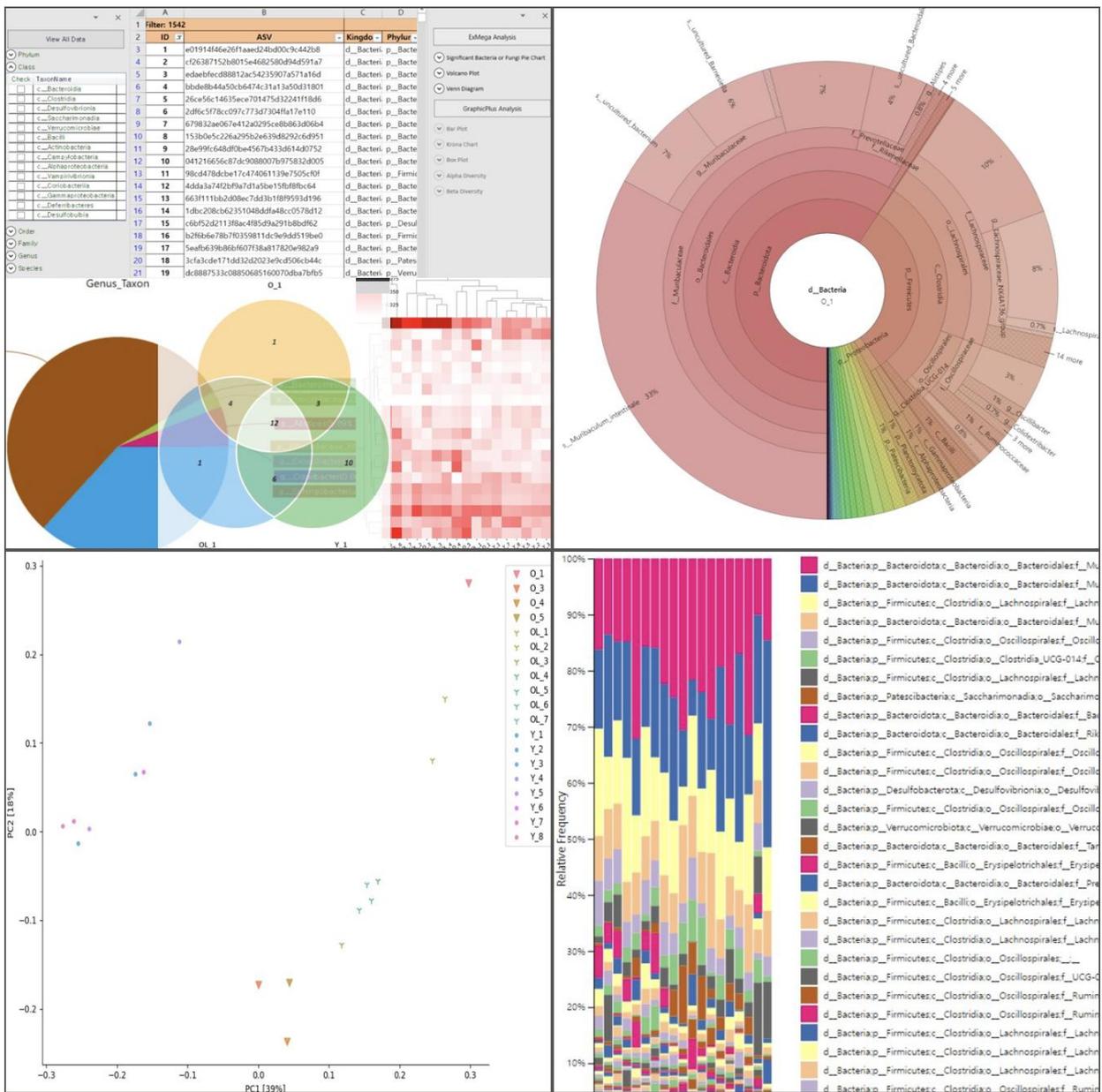


그림 5. ExMEGA 프로그램과 그 기능을 사용하여 제작 가능한 그래프 모음 [10]

Shotgun Metagenome 의 연구동향

1) Shotgun metagenome 을 이용한 Gut Microbiome 특성 분석

Shotgun metagenome 기반으로 장내미생물(gut microbiome)을 분석한 논문을 소개하고자 한다. 2019 년 발표된 논문으로 26 명의 우울 장애 환자(MDD)와 29 명의 정상 환자(HC)의 stool 샘플으로 진행했다. DNA 는 StoolGen DNA 키트를 사용하여 추출 후 영하 80 도에 균일하게 보관되었고 라이브러리는 Illumina 의 TruSeq 키트 시퀀싱 플랫폼은 HiSeq 2500 장비로 샘플 당 최소 6G 의 데이터생산을 하였다. 정상군과 우울증 환자군을 비교하였을 때 유의미한 차이로 장내 미생물 분포도 변화 및 Tryptophan pathway 변화를 보였다 [15].

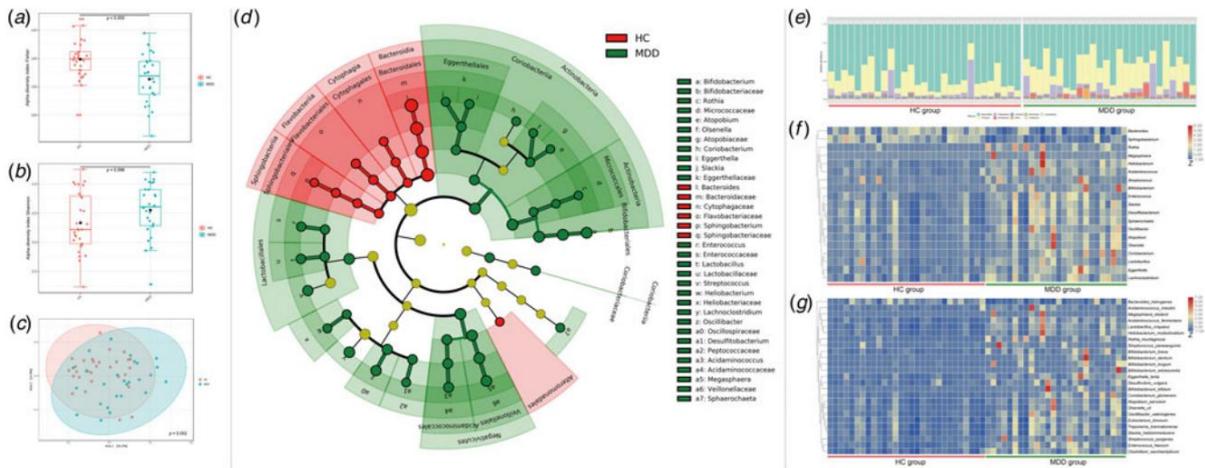


그림 5. 정상군과 우울증 환자군 그룹별 결과의 diversity 와 taxonomy [15]

메타지놈 결과는 Alpha & Beta diversity 와 Cladogram 등의 다양한 그래프로 visualization 했다. 전체 결과를 보았을 때 나이, 성별, BMI에 따른 변화는 관측되지 않았다. 분류학적으로 보았을 때 MDD 그룹의 phylum level 에서 *Actinobacteria* 가 유의미하게 증가하였고 genus level 에서도 *Actinobacteria* 계열이 증가했다. 기능 분석에서도 Tryptophan metabolism 과 관련된 KO (KEGG Orthology) 중 MDD 그룹의 3 개 KO 는 유의미하게 증가하였고 1 개 KO 는 감소했다 [15].

2) Shotgun metagenome 을 이용한 Maize Rhizosphere 특성 분석

다음으로는 토양에서의 미생물 군집 분포도를 Shotgun metagenome 으로 분석한 논문이다. 해당 논문에서는 옥수수(maize)를 배양하는 근권 샘플(뿌리에서 3mm 이내 토양)에서 천연비료와 화학비료 (organic & inorganic fertilizer) 사용여부에 따라 달라지는 미생물 패턴을 분석했다. 토양 샘플은 모두 균일한 방법으로 채취했고, 모두 남아프리카공화국의 노스웨스트 대학의 농업 연구 농장인 Molelwane 농장에서 채취했다. 총 5 개의 그룹으로 채취했고 그룹당 9 개의 sub-sample 으로 구성하여 collection 했다. DNA 추출은 토양 전용 PowerSoil Kit 를 사용하고 Qubit 을 통해 DNA 농도를 측정했다. 시퀀싱 플랫폼은 NovaSeq 6000 장비가 사용되었고 pair end 로 300 cycle 으로 진행했다 [16].

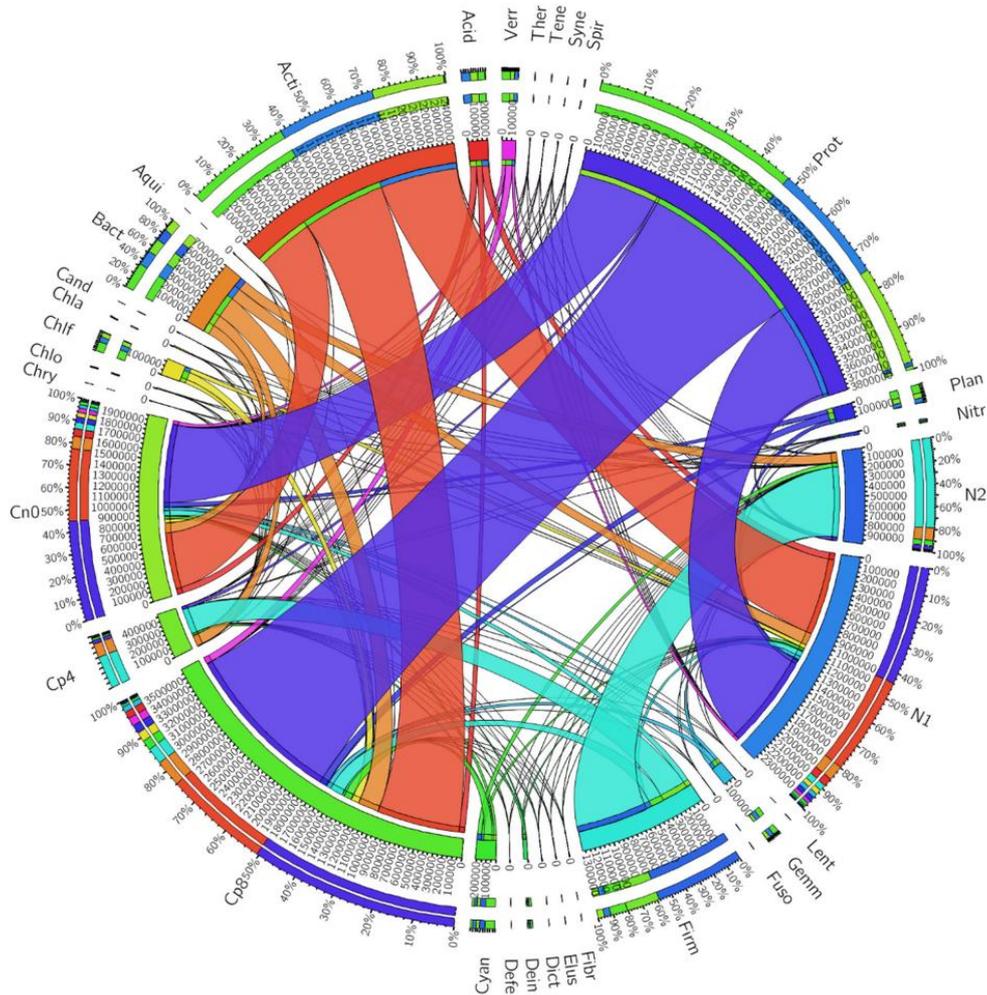


그림 6. 천연비료와 화학비료 처리에 따른 토양 샘플 미생물 군집 변화 [16]

메타지놈 분석 데이터로 PCA 와 PCoA 로 비교하였을 때 5 개 그룹이 4 가지의 cluster 를 이루는 것을 확인했고 그중 Control 그룹과 저농도 화학비료 처리 그룹이 유사하게 cluster 를 이루었다. 천연비료를 사용하였을 때는 극명한 미생물 군집 변화가 일어났고 비료의 투여량 (dose dependent)에 따라서도 극명한 변화가 나타났다. 또한 분류학적 프로파일링을 진행하였을 때 토양에서 가장 많은 종은 Bacteria(98.2%)였고 eukaryote 과, virus, archaea 는 소량의 퍼센티지로 존재하였다. 비료에 성분에 따라 토양의 microbial diversity 가 달라지고 유용균의 분포도가 달라지기 때문에 논문에서는 Shotgun metagenome 결과를 바탕으로 고농도의 천연비료(퇴비)와 저농도의 화학비료를 혼용하는 것이 좋다고 결론지었다 [16].

Ebiogen's Metagenome Service

이바이오젠에서는 Bacteria의 V3-V4 영역, Fungi의 ITS 영역을 증폭하여 실험하는 16S rRNA amplicon sequencing, V1 영역부터 V9 영역까지 전체 16S rRNA를 읽는 Full length sequencing, 그리고 샘플 내 모든 genome data를 분석하는 Shotgun metagenome sequencing 까지 다양한 메타지놈 서비스를 제공한다. 메타지놈 서비스를 통해 토양, 공기, 물, 조직, 분변샘플 등의 다양한 환경에서 채취되는 샘플 내에 어떠한 미생물이 얼마만큼 존재하는지 확인할 수 있다. 해당 서비스를 진행할 경우 위에서 언급한 ExMEGA & GraphicPlus 와 함께 분석 데이터를 제공한다.

Shotgun Metagenome Seq.

Sample Requirement	>5µg gDNA, >50ng/µl gDNA, >4 DIN
Library Method	NEB Ultra DNA Library kit
NGS run format	NovaSeq, PE150bp
Data yield	~10Gb/sample
Turnaround time	~ 4 weeks after Library QC
Sample type	gDNA (prep 서비스도 제공)

Metagenome (16S V3-V4 Amplicon Seq.)

Sample Requirement	>0.1 ng/µl gDNA
Library Method	16S rRNA Amplicon Library(16S/ITS) for Illumina
NGS run format	MiSeq, PE300bp
Data yield	~50,000 reads/sample
Turnaround time	~ 4 weeks after Library QC
Sample type	gDNA (prep 서비스도 제공)

Metagenome (16S rRNA Full length Seq.)

Sample Requirement	>0.1 ng/μl gDNA
Library Method	SMRTbell prep kit for PacBio
NGS run format	PacBio Revio / Sequel I & Sequel II / RS II
Data yield	~50,000 reads/sample
Turnaround time	~ 4 weeks after Library QC
Sample type	gDNA (prep 서비스도 제공)

Ebiogen's Metabolomics Service

이바이오젠의 LC-MS/MS Metabolomics 서비스는 여러가지 샘플에서 존재하는 metabolite 의 동정(Identification) 및 정량(Quantification)을 목적으로 샘플에 포함되어 있는 대사산물을 검출할 수 있는 서비스이다. 주로 Untargeted 방식으로 서비스하고 있으며 샘플 전처리부터 LC-MS/MS 실험, 분석 및 데이터 분석까지 서비스 과정에 포함되어 있다. 최근에는 Biocrates 키트방식의 Targeted metabolomics 도 진행하고 있다.

Untargeted Metabolomics

Sample Type	Cell pellet, Tissue, Serum/Plasma, Bacteria, Fecal, Plant, etc.
Analysis Instrument	LC: Agilent 1290 MS: Thermo Q-Exactive MS (High Resolution MS)
Software	Compound Discoverer, MetaboAnalyst
Turnaround time	~ 7 weeks after sample arrival

< 참고 문헌 >

1. <https://en.wikipedia.org/wiki/Metagenomics>
2. Thomas T, Gilbert J, Meyer F. *Metagenomics - a guide from sampling to data analysis*. *Microb Inform Exp*. 2012 Feb 9;2(1):3. doi: 10.1186/2042-5783-2-3. PMID: 22587947; PMCID: PMC3351745.
3. Joynson, R., Pritchard, L., Osemwekha, E., & Ferry, N. (2017). *Metagenomic Analysis of the Gut Microbiome of the Common Black Slug Arion ater in Search of Novel Lignocellulose Degrading Enzymes*. *Frontiers in Microbiology*, 8. doi:10.3389/fmicb.2017.02181
4. The Integrative HMP (iHMP) Research Network Consortium. *The Integrative Human Microbiome Project*. *Nature* 569, 641–648 (2019). <https://doi.org/10.1038/s41586-019-1238-8>
5. Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). *Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing*. *Biochemical and biophysical research communications*, 469(4), 967-977.
6. https://www.e-biogen.com/support/tech-info?tpf=board/view&board_code=2&code=2757
7. Sharpton, T. J. (2014). *An introduction to the analysis of shotgun metagenomic data*. *Frontiers in plant science*, 5, 209.
8. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). *Shotgun metagenomics, from sampling to analysis*. *Nature biotechnology*, 35(9), 833-844.
9. https://astrobiomike.github.io/misc/amplicon_and_metagen
10. <https://www.e-biogen.com/service>
11. Bolger, A.M., Lohse, M. & Usadel, B. *Trimmomatic: a flexible trimmer for Illumina sequence data*. *Bioinformatics* 30, 2114–2120 (2014).
12. Strous, M., Kraft, B., Bisdorf, R., & Tegetmeyer, H. E. (2012). *The binning of metagenomic contigs for microbial physiology of mixed cultures*. *Frontiers in microbiology*, 3, 410.
13. Abubucker, S. et al. *Metabolic reconstruction for metagenomic data and its application to the human microbiome*. *PLoS Comput. Biol.* 8, e1002358 (2012).
14. Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., & Narasimhan, G. (2016). *Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data*. *Evolutionary bioinformatics*, 12, EBO-S36436.
15. Lai, W. T., Deng, W. F., Xu, S. X., Zhao, J., Xu, D., Liu, Y. H., ... & Rong, H. (2021). *Shotgun metagenomics reveals both taxonomic and tryptophan pathway differences of gut microbiota in major depressive disorder patients*. *Psychological medicine*, 51(1), 90-101.
16. Enebe, M. C., & Babalola, O. O. (2020). *Effects of inorganic and organic treatments on the microbial community of maize rhizosphere by a shotgun metagenomics approach*. *Annals of Microbiology*, 70, 1-10.