ExMEGA, 효과적인 Metagenome 분석 솔루션

1. 시퀀싱 기반 Metagenomics 의 개요

미생물 연구는 오랜 시간 동안 과학의 주요 분야 중 하나로 주목받아왔지만 우리가 알고 있는 미생물의 다양성은 전체의 일부에 불과했다. 미생물의 약 99%는 기존에 알려진 방법으로 배양하기 어려울 뿐만 아니라 배양 독립적인 기술로는 미생물 군집과 그 구성원에 대한 세부 정보를 파악하는 데 한계가 있었기 때문이다. 그러나 Next-Generation Sequencing(NGS) 기술의 도입과 whole-genome shotgun (WGS) sequencing 기술의 등장으로 다양한 환경에 포함된 미생물 군집의 유전체(Metagenome) 구조를 빠르게 분석할 수 있게 되면서 기존의 한계를 극복함과 동시에 시퀀싱 기반 연구에 혁명적인 변화를 일으켰다.[1] 특히, 미국 국립보건원의 Human Microbiome Project (HMP)와 Metagenomics of the Human Intestinal Tract(MetaHIT) 같은 성공적인 국제 마이크로바이옴 프로젝트들이 주목할 만한 성과를 거두면서 NGS 기반의 미생물 연구(Metagenomics)를 황금기로 이끌어주는 계기가 되었다. [2]

그림 1. 국제 마이크로바이옴 프로젝트 [3]

그러나, 기술의 급격한 발전에 따른 대량의 데이터와 NGS 기기의 특성으로 인한 다양한 오류는 metagenomics 분석에서 데이터 품질 관리의 중요성을 더욱 부각시켰다. 이 중 오류 수정(error correction)과 같은 방법이 특히 중요하게 여겨지면서 생물정보학 도구의 중요성이 점점 커지게 되었다.[1] Metagenome 데이터는 다양한 미생물들의 유전 정보를 포함하고 있기 때문에, 그 안에서 특정 미생물의 유전 정보를 분류하거나 분석하는 것은 굉장히 복잡하다. 최근에는 이러한 복잡한 과정을 지원하기 위한 다양한 생물정보학 도구들이 개발되었으며, Raw data 전처리부터 계통분류, 유전체 재구성 등 다양한 단계에서 활용되고 있다. 이러한 고도화된 도구들의 도입은 극한 환경의 미생물 군집을 연구하거나 메타게놈을 유전자나 대사 경로에 연결하는 등의 복잡한 연구를 가능하게 만들고 있다. 이런 복잡한 과정을 거치면서도, 메타게놈 데이터의 효율적 처리와 다양한 분석을 위해 계속해서 새로운 생물정보학 도구들이 개발되고 있다는 것은 주목할 만하다.

1-1. Shotgun Metagenome sequencing and 16s rRNA sequencing

Shotgun Metagenome sequencing 방법은 전체 genome 정보를 획득할 수 있기 때문에 종(species)정보 뿐만 아니라 strain 정보까지 분류학적 정보를 획득할 수 있다. 하나의 샘플에서 다양한 미생물, plasmid, phage, virus, 곰팡이와 같이 다양한 종류의 분석이 가능하며, 유전자의 유무나 기능에 대한 분석도 가능하다. 그러나 host DNA 를 제거하기 어려워 불필요한 sequencing 비용이 발생할 수 있다. 즉 Shotgun Metagenome sequencing 은 많은 분류학적 정보를 얻을 수 있으나, 분석하는데 비용과 시간이 많이 소요된다.

반면 16s rRNA sequencing 의 경우 분류학적으로 의미 있는 16s rRNA 영역을 선택적으로 증폭하여 분석한다. 미생물의 16s rRNA 유전자 영역만을 선택적으로 분석하기 때문에 host DNA 의 오염을 피할 수 있다. 16s rRNA sequencing 은 16s rRNA 영역의 일부만을 분석하기 때문에 분류학적으로는 속(genus) 단계까지 분석 가능하지만, Shotgun Metagenome sequencing 보다 빠르고 합리적인 비용으로 분석이 가능한 장점을 가지고 있다.

2. ExMEGA (Excel based MetaGenome Analysis) Tools

㈜이바이오젠에서 개발 및 보유하고 있는 ExMEGA 는 엑셀 기반으로 실행되며, 다양한 환경내 존재하는 미생물 군집의 생태학적 의미를 분석하는 도구로 활용되고 있다. 또한, Metagenome 데이터에 대한 다양한 분석 기능과 데이터 시각화를 용이하게 해주는 프로그램으로, 데이터 분석과 엑셀 사용에 익숙하지 못한 연구자들도 쉽게 활용할 수 있다. ㈜이바이오젠은 연구자의 요구사항을 지속적으로 반영하여 프로그램을 업데이트하고 있다. 본문에서는 ExMEGA 의 분석 기능과 용어, 기타 분석에 대해 간략히 설명하고자 한다.

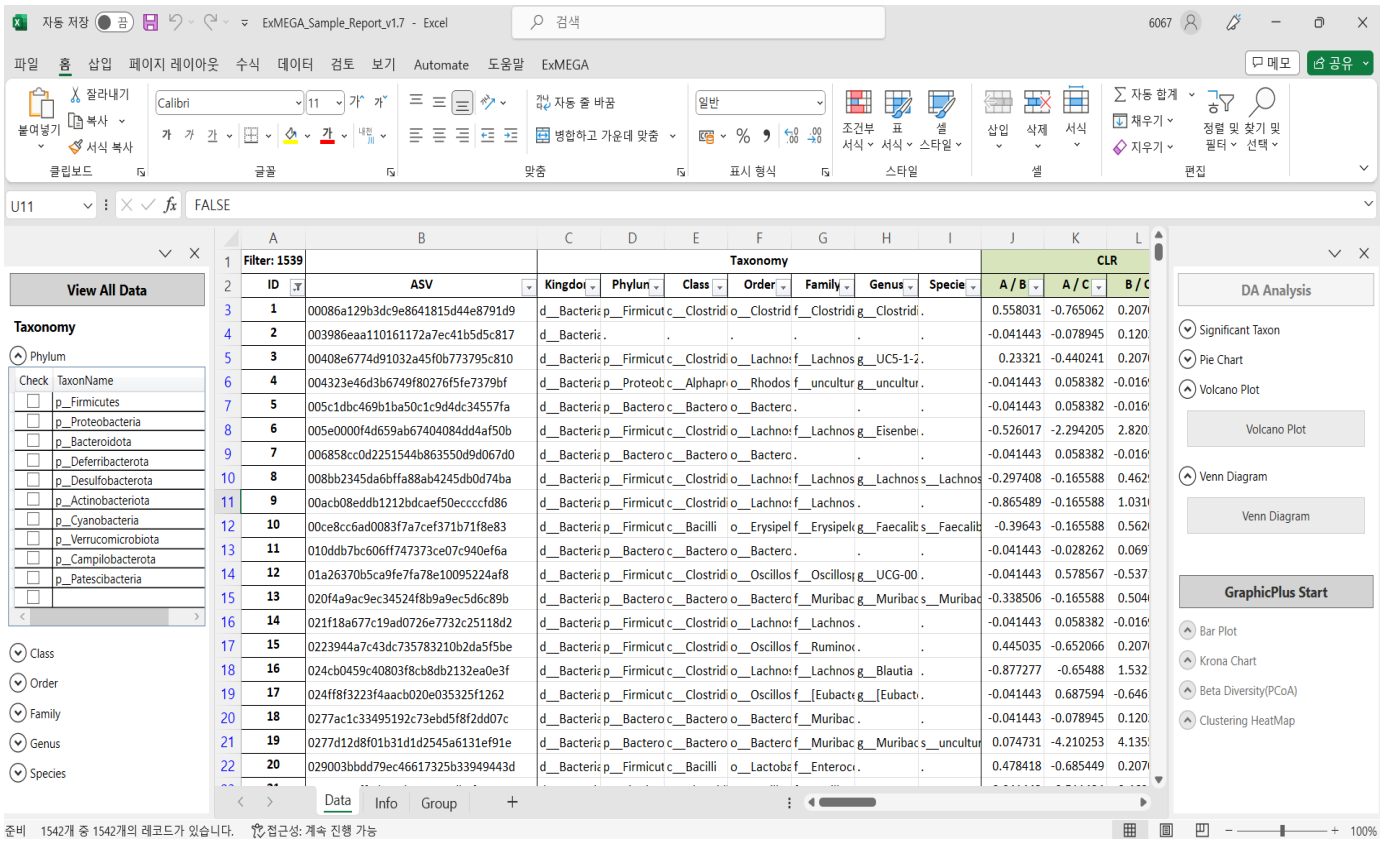


그림 2. 엑셀 기반 ExMEGA 분석 프로그램

2-1. 엑셀내 용어 설명

ASV(Amplicon Sequence Variant)	유사한 DNA 서열 그룹 단위. 즉, Taxon을 구성하는 최소단위
OTU(Operational Taxonomic Unit)	운영 분류 단위: 97% 유사성을 갖는 DNA 서열 그룹
Taxonomy	각 ASV에 대한 7가지 Taxon 정보
CLR(Central Log Ratio)	ANCOM 결과로 나온 중심로그비율
W	ANCOM결과에 따른 통계적 유의성을 나타내는 수치이며, 값이 높을수록 유의미함.
SigW (Significant W)	W결과에 따른 유의미한 수준에 따라 TRUE/FALSE로 반환
Relative Abundance	Feature Count 로 계산된 샘플/그룹 별 비율.
Absolute abundance	생태계의 단위 부피에서 관찰할 수 없는 실제 분류군의 풍부함.
Feature Count	각 ASV에 매핑 된 서열 수
BlastAnnotation	각 ASV의 시퀀스에 대한 Blast 정보

ASV 는 dada2 알고리즘으로 생성되며, 기존 OTU 방법보다 뛰어난 해상도를 보여준다. 더불어, 최근에는 Relative abundance 를 통한 분석만으로는 Metagenome 생태계를 비교하기 어려운 문제로 다양한 통계방식(ANCOM 등)을 이용하여 Absolute abundance 를 추정하고 있다.[4]

2-2. Taxonomic classification

ExMEGA 는 사용자가 원하는 Taxon(분류군)에 따라 필터링하여 데이터를 확인할 수 있는 기능을 제공한다. 이 기능을 통해 Phylum, Class, Order, Family, Genus, Species 와 같은 다양한 분류 수준을 이해하고 분석할 수 있다. 뿐만 아니라, 필터링된 데이터를 Bar plot 과 pie chart 형태로 시각화 하여 정보를 더 명확하게 전달할 수 있도록 지원한다. Metagenome 의 Clustering 최소 단위인 ASV(Amplicon Sequence Variant)는 정해진 분류군을 직접 적으로 볼 수 없기 때문에 실제 Taxon 별로 합산한 상대적 비율을 확인하는 것이 중요하다.

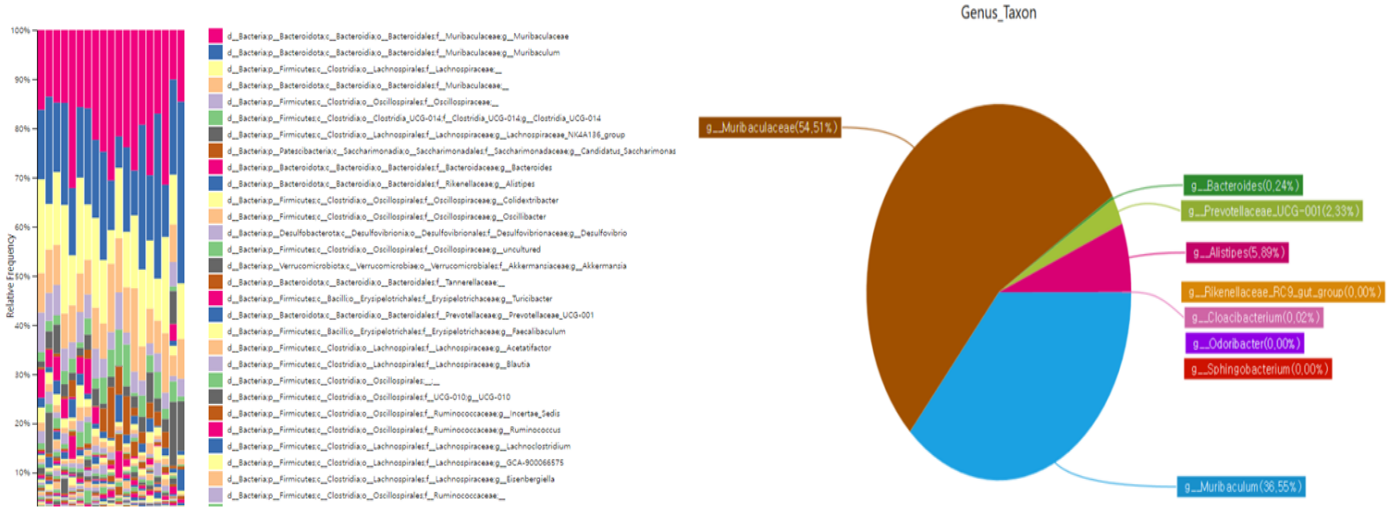


그림 3. ExMEGA Bar plot 과 Pie chart [4]

2-3. DA (Differential abundance) analysis

DA analysis 는 주요 통계분석 중 하나로 어떤 특정 변수와 함께 변동하는 미생물 분류군의 풍부도를 찾아내는 것을 목표로 한다. 이렇게 찾아낸 미생물 분류군은 질병 메커니즘에 대한 생물학적 통찰력을 제공할 수 있으며, 질병 예방, 진단, 치료를 위한 바이오 마커로 활용할 수 있다. [5] 해당 분석 또한 ExMEGA 를 통해 분석이 가능하며, Volcano plot, Venn Diagram 등 다양한 시각화 이미지를 제공한다.

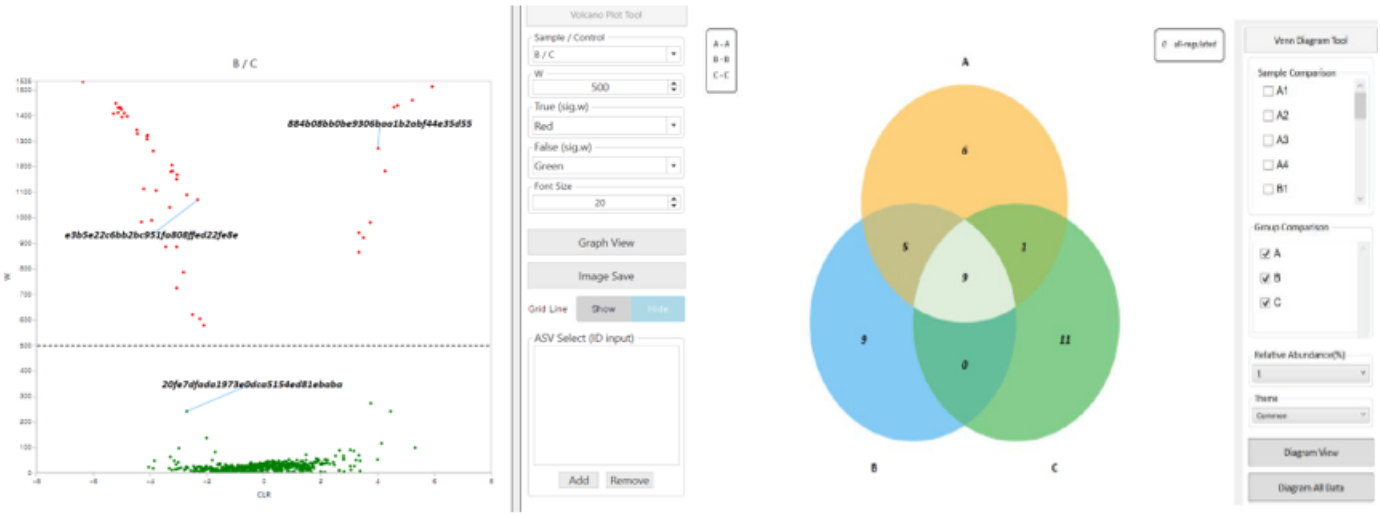


그림 4. ExMEGA Volcano plot, Venn Diagram [4]

2-4. Significant Taxon

아래 스크린샷과 같이 ExMEGA 에서는 그룹 별로 비교한 결과에서 유의하게 발현 차이가 나는 분류군을 필터링 할 수 있는 Significant Taxon 도구를 사용할 수 있다. 연구자는 Relative Abundance(상대 풍부도)와 W 값(통계적 유의성)을 기준으로 필터링하여 각 샘플 및 그룹에 대한 조건을 설정하여 분석 결과를 확인할 수 있다. W 값의 경우 값이 높을수록 해당 분류군의 발현차이가 통계적으로 유의미하다는 것을 나타낸다.

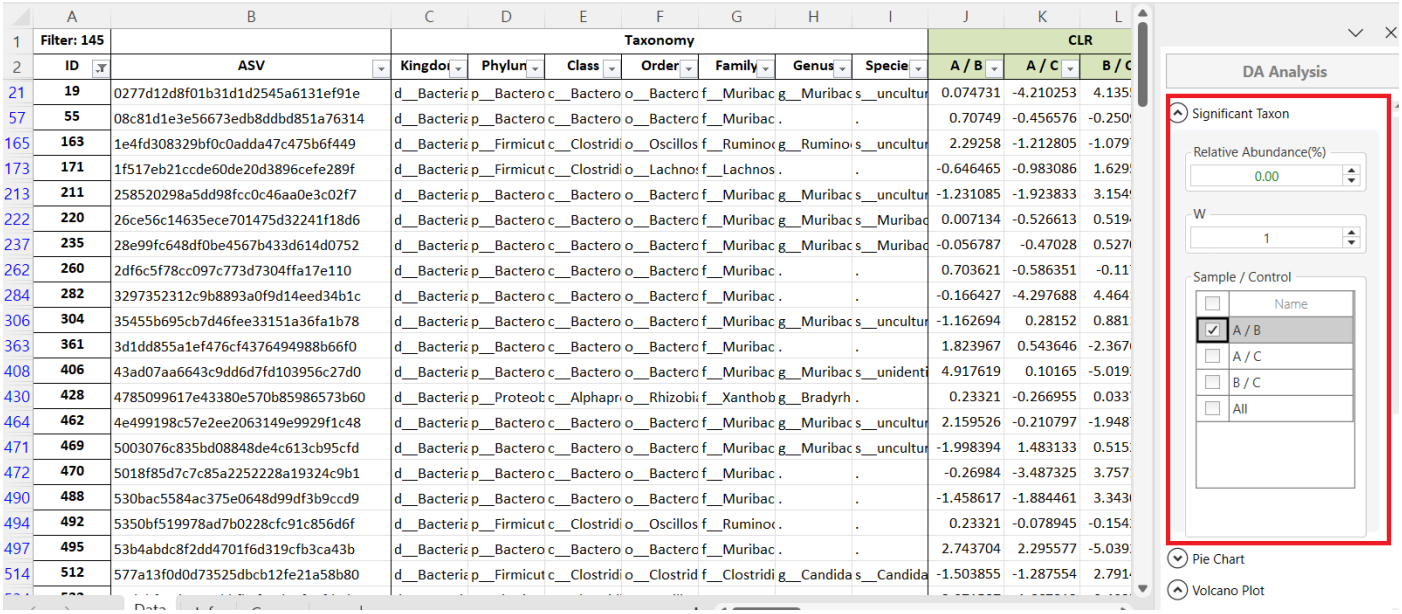


그림 5. ExMEGA Significant Taxon screenshot

3. ExMEGA GraphicPlus

위에서 설명한 시각화 이미지 외에도 ExMEGA GraphicPlus 를 통한 추가적인 분석이 가능하다. 대표적으로 Krona Chart, PCoA Principal Coordinates Analysis), Clustering Heatmap 으로 표현 가능하다. 모든 그래프는 단일 샘플 비교(Single Data)와 그룹 비교 (Group Data) 가 모두 가능하며 W 값을 적용하여 분석된다. 만약, 통계분석을 원하지 않을 경우에는 Relative Abundance 만으로도 분석 가능하다.

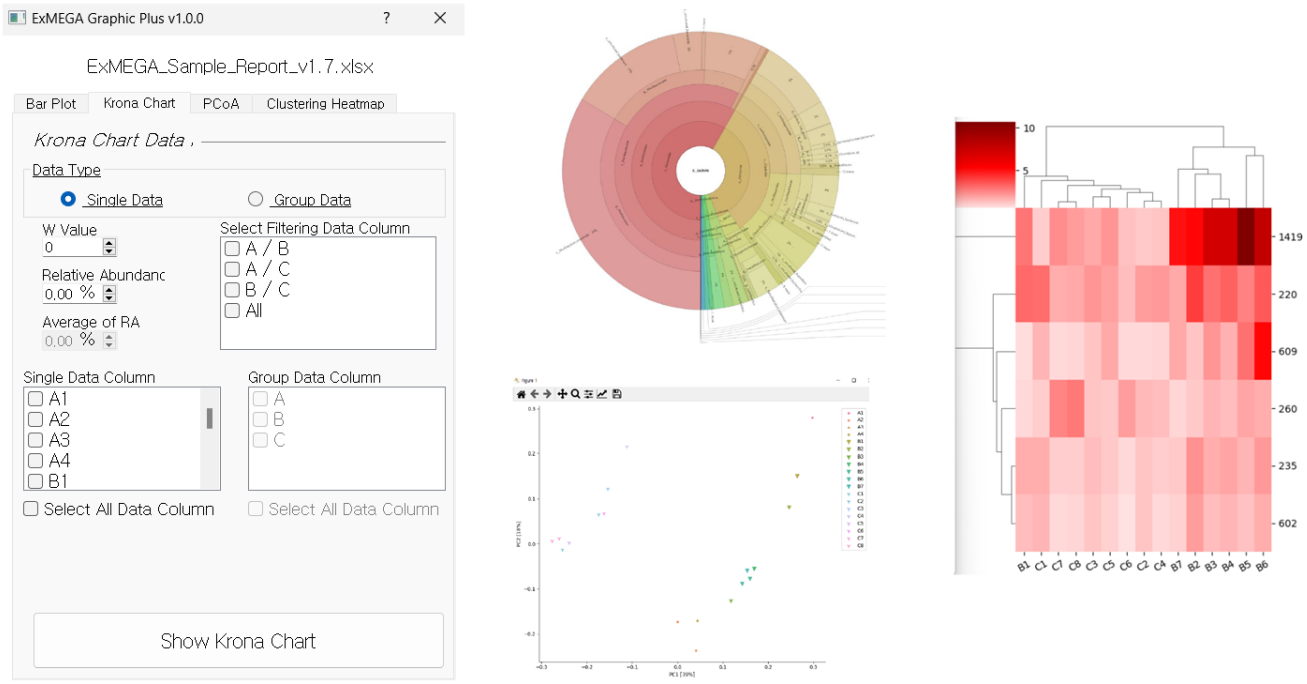


그림 6. ExMEGA GraphicPlus - Krona Chart, PCoA, Clustering Heatmap[4]

4. E-biogen's Metagenome Service

이바이오젠에서는 Bacteria 의 V3-V4 region, Fungi 의 ITS region 을 증폭하여 실험하는 16s rRNA amplicon sequencing 을 통해 토양, 공기, 물, 조직, 분변시료 등의 다양한 환경에서 채취되는 샘플 내에 어떠한 미생물 (Bacteria, Fungi 등) 들이 얼마만큼 존재하는지 분석할 수 있는 Metagenome & Microbiome 서비스를 제공하고 있다. 해당 서비스를 진행할 경우 위에서 언급한 ExMEGA&GraphicPlus 와 함께 분석 데이터를 제공한다.

4-1 Service info

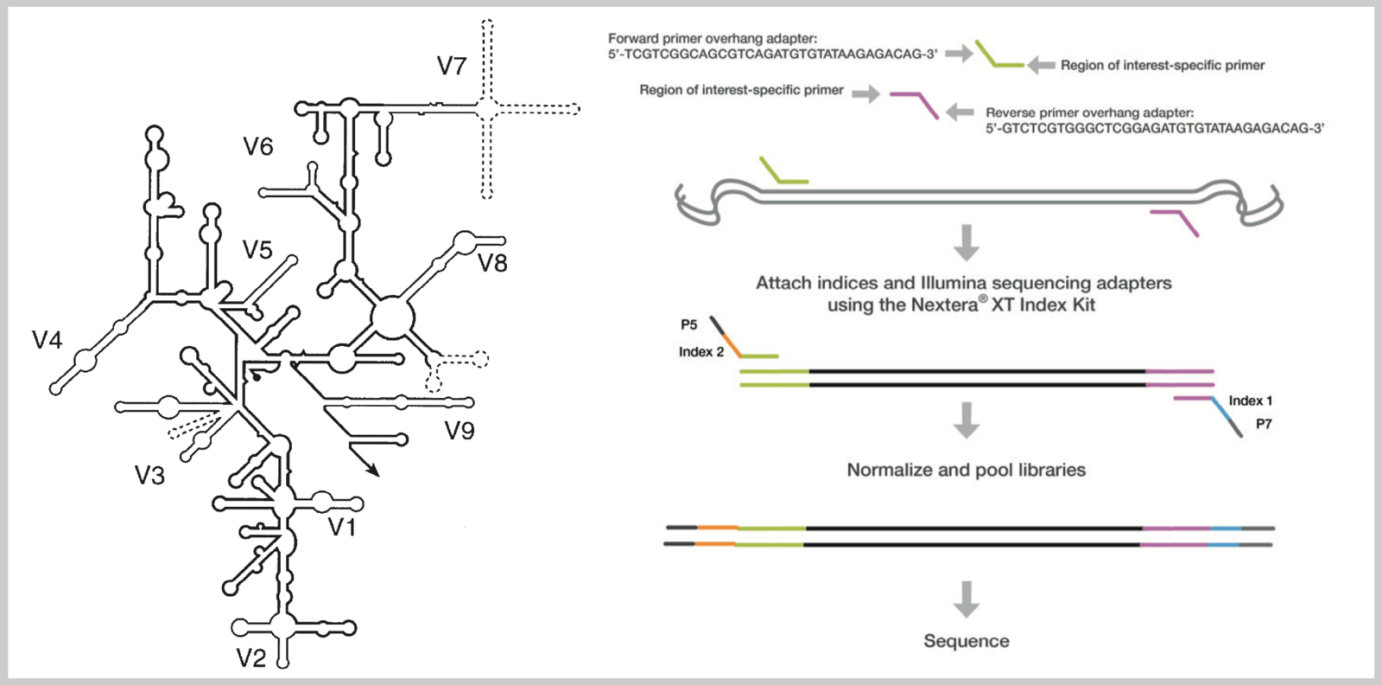
Sample requirement	>0.1ng gDNA
Library method	16S rRNA Amplicon Library for illumina
NGS run format	MiSeq, PE300
Data yield	~50,000 read/sample
Turnaround time	Library QC 통과 후 4 주 이내
Sample type	gDNA, fecal, soil, water, feedstuff, tissue, saliva 등

++ DNA prep.을 원하실 경우 담당자와 상의하여 진행이 가능합니다.

4-2 Metagenome Service Process

Library Prep → Library QC → NGS → Data Analysis → Report

4-3 Library Construction Workflow



<참고문헌>

1. Jünemann, S., Kleinbölting, N., Jaenicke, S., et al (2017). Bioinformatics for NGS-based metagenomics and the application to biogas research. Journal of Biotechnology, 261, 10–23.

2. Yong-Xin Liu., Yuan Qin., Yang Bai., et al (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell, 315–330.

3. 인체 마이크로바이옴 연구개발 동향, KHIDI 전문가 리포트, 4.

4. EBIOGEN, ExMEGA v.10 & Data Analysis User Manual

5. Lu Yang., Jun Chen. (2022). A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions, Microbiome 10, 130