Korean J. Environ. Biol. 30 (2) : 77~89 (2012)

〈Review paper〉

차세대 유전체 기술과 환경생물학 - 환경유전체학 시대를 맞이하여

송주연 · 김병권 · 권순경[1] · 곽민정[1] · 김지현*

연세대학교 생명시스템대학 시스템생물학과
[1]과학기술연합대학원대학교 이학부 시스템생명공학전공

Next-generation Sequencing for Environmental Biology - Full-fledged Environmental Genomics around the Corner

Ju Yeon Song, Byung Kwon Kim, Soon-Kyeong Kwon[1], Min-Jung Kwak[1] and Jihyun F. Kim*

Department of Systems Biology, Yonsei University, Seoul 120-749, Korea
[1]Biosystems and Bioengineering Program, University of Science and Technology, Daejeon 305-350, Korea

Abstract - With the advent of the genomics era powered by DNA sequencing technologies, life science is being transformed significantly and biological research and development have been accelerated. Environmental biology concerns the relationships among living organisms and their natural environment, which constitute the global biogeochemical cycle. As sustainability of the ecosystems depends on biodiversity, examining the structure and dynamics of the biotic constituents and fully grasping their genetic and metabolic capabilities are pivotal. The high-speed high-throughput next-generation sequencing can be applied to barcoding organisms either thriving or endangered and to decoding the whole genome information. Furthermore, diversity and the full gene complement of a microbial community can be elucidated and monitored through metagenomic approaches. With regard to human welfare, microbiomes of various human habitats such as gut, skin, mouth, stomach, and vagina, have been and are being scrutinized. To keep pace with the rapid increase of the sequencing capacity, various bioinformatic algorithms and software tools that even utilize supercomputers and cloud computing are being developed for processing and storage of massive data sets. Environmental genomics will be the major force in understanding the structure and function of ecosystems in nature as well as preserving, remediating, and bioprospecting them.

Key words : metagenomics, pyrosequencing, bridge amplification, single-molecule sequencing, nanopore DNA sequencing, sequence assembly, bioinformatics

* Corresponding author: Jihyun F. Kim, 02-2123-5561, Fax. 02-312-5657, E-mail. jfk1@yonsei.ac.kr

서 론

현대 생물학, 특히 유전체학 (genomics)의 발전으로 인간을 포함한 생명을 지닌 여러 유기체들과 그들이 상호작용하는 생물집단 및 자연환경 등 생명과 관련된 여러 연구에서 생물정보의 중요성을 간과할 수 없게 되었다. 생물체가 관여된 모든 곳에는 세포 내에 암호화되어 있는 유전물질 즉, DNA의 자취가 남겨지게 되므로 이를 해독하고 분석하는 것은 생물체의 유전, 형질, 생리, 대사 등 기초 연구뿐만 아니라, 의약, 농업 등 산업적 응용과 환경 및 생태 연구에 기본적인 자료가 될 만큼 중요하다. 따라서 인간을 비롯하여 모든 생물체를 연구하는 분야에서는 효율적인 유전정보 해독 및 분석 기술이 꾸준히 요구, 개발되어 왔다. 기술적 측면에서는 전통적인 염기서열 해독 방법인 생거 시퀀싱 (Sanger sequencing) 기술을 이용하여 대량의 서열을 생산하기 위한 분석기기들이 개발되었으며, 전략적 측면에서는 산탄총을 쏘듯 무작위로 염기서열 단편을 확보하고 후에 이를 조립하는 소위 "샷건 (shotgun) 전략"이 고안되어 효율적인 유전체 해독 방법들이 확립되면서 인간을 비롯한 여러 생명체들의 유전체 서열의 분석이 시작되었다 (Fleischmann *et al.* 1995; Fraser *et al.* 1995; Lander *et al.* 2001; Venter *et al.* 2001). 최근, 노동, 비용, 시간 면에서 획기적으로 경제적이고 효과적인 차세대 염기서열 분석 기술 (next-generation sequencing, NGS) 또는 제3세대 기술들이 개발됨으로써 단일 생물체 및 집단에 대한 유전체 해독이 급격히 증가하고 있으며 이와 더불어 방대한 양의 데이터를 분석하기 위한 생물정보학 (bioinformatics) 관련 기술이 급속히 발전하고 있다.

염기서열 해독 기술은 환경 미생물 및 집단의 유전체 연구에서도 활발하게 이용되고 있다 (Béjà *et al.* 2000; Tyson *et al.* 2004; Jeong *et al.* 2005; Hallam *et al.* 2006; Dinsdale *et al.* 2008). 환경 정화나 환경 지표 생물로써 연구되어 온 환경 미생물의 유전체 분석은 생물공학적 유용 인자 발굴을 통한 환경 정화에 공헌하였으며 환경 내 지표 유전인자들의 정보 제공을 통하여 환경 변화를 개체수준에서 모니터링 하는데 중요한 역할을 담당하였다. 또한 환경 시료로부터 직접 DNA를 채취하는 기술들이 개발되면서 배양 가능한 미생물뿐만 아니라 아직까지 배양이 불가능한 것으로 알려진 환경 내 대다수의 미생물들로부터 새롭고 다양한 유용 유전정보를 획득할 수 있게 되었다. 대용량 염기서열 해독이 가능한 차세대 유전체 분석 기술은 앞서 언급한 기술과 융합되어 환경 내 미생물의 군집 구조 정보 및 총 미생물체의 유전체 정보를 분석하는 염기서열 기반 환경유전체학 (environmental genomics, metagenomics) 분야를 한 단계 도약시키는 데 매우 큰 공헌을 하고 있다 (Galand *et al.* 2009; Hess *et al.* 2011).

미생물 유전체 및 메타지놈 (metagenome)의 연구 범위는 분석 기술의 발전에 따라 모델 균주나 표준 균주의 분석에 국한되지 않고 유전체 정보 비교를 위한 환경 분리 미생물이나 유연관계가 가까운 미생물의 분석으로 급속히 확장되고 있으며, 배양이 불가능한 미생물의 경우, BAC (Bacterial artificial chromosome) 및 fosmid를 이용한 단편서열분석으로부터 전 환경유전체 염기서열에서 개별 미생물 유전체를 완전하게 조립하거나 개별 세포에서 유전체를 획득하고 이로부터 유전체를 분석하는 수준에 이르기까지 기술이 진보되고 있다 (Quaiser *et al.* 2002; Hallam *et al.* 2006; Woyke *et al.* 2009). 또한 환경 보전과 자원의 발굴뿐만 아니라, 다양한 환경 및 생물간 상호작용 현상 관찰에 활용되어 일반 자연 환경부터 극한 환경, 식품, 공생관계에서의 유전체 및 메타지놈이 분석되고 있으며, 미지 환경의 탐색 및 보건 의료를 위한 연구에도 활용을 기대하고 있다 (Qin *et al.* 2010; Jung *et al.* 2011; Mackelprang *et al.* 2011). 유전체 서열 분석 기술은 현재까지 지속적으로 요구되고 발전되어왔던 것처럼 더 정확하고, 더 많은 정보를 생산하게 될 것이다. 따라서 여전히 진보중인 차세대 염기서열 분석 기술과 축적되어가는 생물정보의 활용 기술을 살펴보고자 한다.

DNA 분석 기술의 발전과 NGS 시대의 도래

제1세대 염기서열 분석 방법인 생거 시퀀싱 기술은 1977년에 소개되어 수십 년간 기술적 진보를 거쳐 자동화된 염기서열 분석기에 적용되어 인간 유전체 서열 해독 프로젝트 (human genome project)에 크게 기여한 분석 방법이다 (Sanger *et al.* 1977; Lander *et al.* 2001; Venter *et al.* 2001). 생거 시퀀싱 기술에 기반을 둔 초고속 대용량 (high-throughput) 분석을 위한 여러 장비가 투입되었으나 인간 유전체 해독에 13년이라는 긴 시간과 약 30억 달러라는 막대한 비용과 많은 노동력, 시간이 투입되어 좀 더 효율적이고 경제적인 차세대 염기서열 분석기의 개발의 필요성이 제기되었다. 이를 위해 일명 대규모 병렬형 시퀀싱 (massively parallel sequencing) 즉, DNA 조각들을 수천 수백만 개의 나노 크기의 분리된 공간에 넣고 이로부터 증폭되어 나오는 대량의 신호들을 한꺼번에 읽을 수 있는 방법을 이용한 차세대 염기서열 분석

기들이 개발되기 시작되었다. 그 결과, 첫 번째 차세대 염기서열 분석기가 2005년 Jonathan Rothberg에 의해 창립된 454 Life Sciences (2007년 Roche에서 인수)에서 출시되었다. GS20이라는 이름으로 개발된 분석기는 이후, GS FLX, GS FLX titanium, 현재 GS FLX+ system까지 개발을 거듭하여 각 시퀀스 read의 길이와 해독 용량을 향상시켜왔다. GS 시리즈의 분석기기들은 단분자 액상 PCR (single-molecule emulsion PCR) 방법을 통해 증폭된 DNA에 대해 DNA 중합효소가 각 염기를 삽입시켜 새로운 가닥의 DNA를 만들 때 생성되는 pyrophosphate가 형광물질의 형광을 유도하고 이 형광 신호를 검출하는 pyrosequencing 기반의 분석 방법을 응용하여 염기서열을 생산하는 방식이 적용되었다 (Bult *et al.* 1996). 현재 GS FLX+ 버전은 최장 염기 천 개 길이의 read를 생산할 수 있으며 이러한 긴 read 길이를 바탕으로 유전체 전체에 대한 시퀀싱과 메타지놈 분석이 가능하고 그 외, 유전자 발현을 분석하기 위한 전사체 (transcriptome) 시퀀싱, 집단 내 유전적 다양성 분석을 수행하기 위한 PCR 산물 (amplicon) 시퀀싱, 단백질로 만들어지는 유전자 부위만 분석하는 엑솜 (exome) 시퀀싱 등을 수행할 수 있다. 그 이후, 2006년에 Solexa (Illumina사가 2007년 인수)가 Genome Analyzer를, 2007년에 Applied Biosystems (현재 Life Technologies로 통합)이 SOLiD system을 각각 출시하였다 (Bentley *et al.* 2008; Valouev *et al.* 2008). Illumina의 Genome Analyzer는 고체상에서 bridge amplification 방법을 통해 DNA를 증폭하여 cluster를 만든 후 (reversible terminator-based sequencing by synthesis), 염기서열을 각각의 형광 표지로 검출하는 방법을 이용한다 (Bentley *et al.* 2008). ABI의 SOLiD system은 합성을 통한 시퀀싱 (sequencing-by-synthesis) 방법을 차용한 위의 염기서열 분석기들과 달리 sequencing-by-ligation 방법을 염기서열 해독에 도입하였다 (Valouev *et al.* 2008). 두 가지 시스템으로부터 생산될 수 있는 read의 길이는 GS 시리즈의 길이보다 짧지만 그보다 훨씬 높은 용량과 정확도를 강점으로 내세워 인간을 비롯한 다양한 종류의 생물체들의 유전체 서열 전체 해독과 더불어 이미 완성된 표준 유전체 서열을 참조하는 SNP (single nucleotide polymorphism) 및 CNV (copy number variation) 등의 유전체 비교 분석과 유전자 조절 분석 등을 가능하게 하고 있다 (McKernan *et al.* 2009; Bickhart *et al.* 2012). 현재 위 세 가지 플랫폼들은 모두 초대용량 시퀀싱을 표방하고 있으나 read 길이, 용량, 정확도, 분석시간, 비용 등 각각의 특성에 차이가 있다. 즉, 세 가지 방법 중, GS FLX+ read의 평균 길이가 700 bp이고 한번 가동 시 700 Mb의 데이터를 생산할 수 있다. Illumina사의 Genome Analyzer는 150 bp까지 읽을 수 있고, 생산되는 데이터 양은 HiSeq 2000 기기에서 나오는 paired-end read를 기준으로 600 Gb까지 가능하다. 한편, Life Technologies의 SOLiD는 5500xl system을 기준으로 최대 75 bp의 read를 생산할 수 있고 한번 가동하면 두 인간 유전체의 평균 30x coverage 수준까지 염기서열이 생산된다 (Table 1). 따라서 유전체를 분석하려는 연구자들은 염기서열 분석 대상 및 용도를 고려하여 적합한 기기를 활용할 수 있다.

최근까지 위와 같은 염기서열 분석기가 상용화되어 생거 시퀀싱 시대에 요구되었던 cloning없이 경제적인 비용으로 대용량 데이터 생산에 대한 갈증이 해결되었다면 현재에는 차세대 염기서열 분석기들이 이용하고 있는 PCR 증폭에서의 bias를 극복하고 단일 DNA 분자를 바로 검출할 수 있으며 더불어 긴 read 길이와 높은 정확도를 갖는 염기서열 결과가 요구되고 있다. 이러한 필요에 힘입어 PCR없이 나노 스케일에서 분석 대상 DNA 분자를 읽는 단분자 염기서열 분석기 (single-molecule sequencer)들이 개발되었고 이를 이른바 제3세대 염기서열 분석기라고 한다 (Metzker 2010; Pareek *et al.* 2011).

Table 1. Features of NGS platforms

	Illumina HiSeq 2000	Roche GS FLX+	Life technologies 5500xl	Life technologies Ion Proton Sequencer	Pacific Bioscience PacBio RS
Template 생산 방식	Bridge amplification	Emulsion PCR	Emulsion PCR	Emulsion PCR	–
Sequencing 방법	Reversible terminator-based sequencing by synthesis	Pyrosequencing	Sequencing by ligation	Ion semiconductor sequencing by synthesis	Single molecule real time sequencing by synthesis
Read length	100 bp	700 bp	75 bp	200 bp	1300 bp
Throughput	600 Gb[1]	700 Mb	190 Gb/run[2]	10 Gb[4]	45 Mb/1 SMRT cell
Run time	11 days[1]	23 h	7 days[3]	2 h[5]	~2 hr

[1] 2 × 100 bp, dual flow cell
[2] 2 human genomes (30X average coverage), mate-paired 60 bp × 60 bp
[3] Mate-paired 60 bp × 60 bp, 1 human genome (4-5X average coverage)
[4] Ion Proton™ I Chip
[5] Read length 100 bp

Helicos Biosciences사의 HeliScope, Pacific Biosciences사의 SMRT 등이 상용화 되었고, Oxford Nanopore Technologies의 Nanopore DNA 시퀀서가 출시를 예고하고 있다. 이 외에 최근 합성을 통한 시퀀싱 방법으로서 DNA 중합반응 시 나오는 수소이온의 화학적 신호를 반도체 칩을 이용하여 검출하는 Ion Torrent 시퀀서가 Life Technologies사의 Ion PGM과 Ion Proton sequencer로 출시되어 좀 더 저렴한 비용으로 염기서열을 분석할 수 있게 되었다. 차세대 염기서열 분석기들이 여전히 단점을 극복하기 위해 시스템을 업그레이드하여 왔듯이 제3세대 분석기들 또한 새로운 기술이 접목된 만큼 안정화되고 향상된 질의 데이터를 생산하는 데에는 시간이 더 필요할 것으로 보인다.

대용량 유전체 정보 분석 기술

NGS 시스템으로부터 확보된 유전체를 해독하기 위해서는 사전에 read 또는 라이브러리 종류를 고려하여 해독 전략을 세워 single 또는 paired-end read를 획득하게 된다. Paired-end 시퀀싱의 경우에는 다양한 길이를 가진 여러 종류의 라이브러리를 사용하여 분석을 하는 것이 정확한 염기서열 획득과 해독 효율을 높이는 결과를 가져오기도 한다. 생거 시퀀싱을 이용한 미생물 유전체 해독 시대에는 전통적으로 가장 보편적으로 이용된 Phrap을 비롯하여 TIGR assembler, CAP3와 같은 overlap-lay-out-consensus 방식을 이용한 서열 조립 프로그램이 사용되었다면 짧은 길이의 대용량의 염기서열 조합을 위해서 de Bruijin graph를 이용한 Velvet, SOAPdenovo, ABySS, CLC Genomics Workbench *de novo* assembler 등의 다양한 서열 합체 프로그램이 개발되어 이용되고 있다(Zerbino and Birney 2008; Simpson *et al.* 2009; Li *et al.* 2010). 그러나 참조할 유전체 서열이 많아지고 비교유전체 분석의 빈도가 높아지면서 유전체를 해독하는 방법은 이러한 *de novo* assembly 외에도 reference-guided assembly 방법을 통해 도출된 공통(consensus) 염기서열을 활용하는 방법이 고안되었다. Reference mapping은 유전체 분석에서 주로 SNP나 DIP 분석을 할 때 접근하는 방법이지만 MAQ, SSAHA, BWA (Ning *et al.* 2001; Li *et al.* 2008; Li and Durbin 2009), CLC Genomics Workbench reference mapping와 같은 서열 맞춤 도구들을 이용, 근연 관계의 미생물 유전체 해독 시 consensus 염기서열을 불러와 *de novo* as-sembly와 접목시켜 효율적인 서열 합체 결과를 얻게 해준다(Nishito *et al*. 2010). 유전체 완성을 위해서는 일차 서열 조립을 통해 얻어진 contig들의 염기서열을 가지고 scaffold를 구성을 선행하게 된다. Scaffold의 구성은 paired-end sequence를 가지고 있을 때 가능하며, 다른 크기의 삽입 DNA를 갖는 라이브러리들로부터 나온 여러 종류의 mate 정보를 이용하면 scaffold를 효과적으로 구성할 수 있다. SSPACE나 SOPRA 등의 프로그램의 도움을 받아 scaffold를 만들거나, 서열 조립 프로그램이 자체적으로 scaffold를 구성하여 염기서열 초안을 구성하기도 한다(Dayarian *et al.* 2010; Boetzer *et al.* 2011). 일반적으로 유전체 완성을 위해서는 IMAGE, CLoG 등과 같은 프로그램을 이용하여 *in silico* gap filling을 수행하거나(Tsai *et al.* 2010; Xing *et al.* 2011), physical gap 처리를 위한 PCR이나 ACP 기술들을 활용하여 조립을 수행하기도 한다. 완성된 서열 내에서 유전자 정보를 분석하기 위해 유전체 주석화를 수행하게 되는데 일차적으로 Glimmer, GeneMark 등의 유전자 구조 예측 프로그램을 통해 structural annotation을 수행한 후(Delcher *et al.* 1999; Besemer *et al.* 2001), 예측한 유전자들을 가지고 UniProt Knowledge-Base, KEGG genes, COG, NCBI NR, Pfam, TIGRfam, InterPro 등의 데이터베이스를 대상으로 BLAST 및 HMM search를 수행하여 기능을 유추한다. 이러한 일련의 유전체 해독 과정 후에 기능적, 생태학적, 또는 진화적 의미를 찾을 수 있는 유전체 분석을 개시할 수 있다.

NGS 개발 이후로 환경을 포괄하기에 충분한 양의 유전정보를 얻을 수 있게 되면서 환경유전체학 분야도 제2의 전성기를 맞이하게 되었다. 특히 염기서열 기반 메타지놈 연구는 심도 있는 미생물 다양성 분석 및 기능유전체 분석이 가능하게 되어 미생물 군집과 기능성 메타지놈을 거시적인 관점에서 해석할 수 있게 되었다. 미생물 및 유전자 다양성 분석을 수행하기 위해서 특정 유전자 단편을 PCR 방법을 활용하여 증폭된 amplicon을 준비한 후, NGS 시스템(read 길이 및 예상되는 다양성을 계산하여 선택된 최적 플랫폼; 예를 들어 계통학적 미생물 다양성을 분석하기 위해 선택된 16S rRNA 유전자가 대상일 경우, amplicon의 길이가 길수록 많은 정보를 이용하여 자세한 정보를 얻을 수 있기에 최근에 454/Roche GS FLX Titanium을 가장 많이 사용하고 있음)을 이용하여 이들의 염기서열을 대량으로 획득하고, 특화된 또는 GenBank 데이터베이스를 활용하여 BLAST 방법 및 k-mer 패턴분석 등을 통해 각각의 read를 계통학적 정보 및 그룹정보를 할당하게 된다. 대량의 reads에 할당된 정보는 MEGAN과 같은 프로그램을 활용하여 정리된 형태로 볼 수 있으며(Mitra *et al.* 2011), 16S rRNA 유전자의 경우 Ribosomal Database Project 웹에 구축되어 있는 RDP classifier 등의 웹 기반 프로그램을 통하여 계통학적 군

집 구조를 파악할 수 있다(Lan *et al.* 2012). 또한, 군집의 다양성 (Simpson's index), 풍부함 (Chao, abundance-based coverage estimator), 균등성 (Shannon's index) 등과 같은 알파 다양성 수치를 얻기 위해서는 reads를 정렬하고 이로부터 각 reads 간의 진화적 거리 또는 유사도를 근거로 각 그룹을 정의하여 시료 내에 존재하는 총 operational taxonomic unit (OTU) 및 개체 수에 대한 정보를 획득한 후, 추정된 수를 이용하여 계산해야 한다. 하지만, NGS에서 생산된 대량의 reads를 전통적 프로그램들을 활용하여 정렬하고 이로부터 그룹을 정의하는 것은 시간이 매우 오래 걸리며, 아주 많은 메모리를 요구하게 된다. 최근 이러한 문제를 극복하기 위하여 많은 생물정보학 도구들이 개발되고 있다. T-Coffee, MAFFT, Clustal Omega 등은 대량의 염기서열을 빠르게 정렬하기 위해 만들어진 프로그램이며, CD-HIT, Esprit-tree 등은 염기서열을 각 OTU로 binning하여 주는 프로그램이다(Notredame *et al.* 2000; Li and Godzik 2006; Katoh and Toh 2010; Sievers *et al.* 2011; Cai and Sun 2011). 또한, Mothur나 Qiime 프로그램의 경우, 앞서 언급한 파이프라인을 하나의 package 프로그램으로 묶어 사용자가 대용량 데이터를 쉽게 정리할 수 있도록 도와주기도 하며, 리눅스 및 생물정보학을 처음 접하는 사용자의 경우, Ribosomal Database Project 웹 페이지에 구축된 pyrosequencing pipeline 페이지에서 순차적으로 웹 기반 프로그램을 실행시켜 손쉽게 알파다양성 정보를 획득할 수도 있다(Schloss and Handelsman 2005; Schloss *et al.* 2009; Caporaso *et al.* 2010). 이외에도 Fast UniFrac 프로그램은 계통수 기반으로 각 미생물 집단간 비교분석 (Clustering, 주성분분석 등)을 실시하여 군집 간의 차이를 시각화할 수 있게 해준다(Hamady *et al.* 2010). 한편, 미생물 집단 전체의 유전정보 비교분석을 위해서 환경시료로부터 총체적인 DNA를 수집하여 획득한 대용량의 염기서열은 MetaVelvet, MAP 등의 metagenomics를 위한 시퀀스 조립 프로그램을 활용하여 contig으로 조립할 수 있으며 k-mer 패턴에 따라 각 contig들을 분류군에 따라 정렬할 수도 있다(McHardy *et al.* 2007; Pignatelli and Moya 2011; Lai *et al.* 2012). 조립된 염기서열로부터 GenBank NR, KEGG, subsystem 등의 데이터베이스와 비교하여 환경 시료 내 유전자 목록을 작성할 수 있으며, 각 유전자에 따른 빈도 또한 계산할 수 있다. 이러한 유전자 정보를 바탕으로 메타게놈 내 대사 네트워크 및 생리 · 생태학적 기능을 유추하게 된다.

NGS 개발은 지금까지 접근하지 못했던 미생물 유전체 및 환경까지 탐구 가능하게 한다. 또한 더 많은 데이터의 생산과 함께 더 높은 컴퓨터 분석 기술이 요구되고 있고, 연구자들이 접근 가능할 수 있는 프로그램들의 수요도 함께 증가하고 있으므로 메타지놈 분석을 위한 기술 및 생물정보학적 도구 개발의 필요성이 대두되고 있다. 더불어 단순히 대용량 정보의 획득에 그치지 않고 미생물 유전체 및 환경으로부터 생물, 생태학적 의미를 찾기 위한 심화된 연구가 필요한 시점이다.

미생물 유전체 및 메타지놈 연구 동향

454/Roche Genome Sequencer와 Solexa/Illumina Genome Analyzer 장비와 같은 차세대 염기서열 분석 기술의 진보와 함께 그로부터 생산되는 유전체 서열 정보의 생산량도 기하급수적으로 증가하기 시작하였다. 염기서열 분석 시장은 신규 유전체 전체 시퀀싱을 넘어, 참조 유전체를 활용한 시퀀싱을 통한 유전체 비교 분석뿐 아니라, microarray가 자리하던 전사체 분석 영역을 대체하고, 이전에는 막대한 비용과 용량의 한계로 인해 불가능하게 여겨지던 메타지놈 분석이 가능하게 되면서 시장이 크게 확장되고 있다.

미생물 유전체 분석 분야도 이와 함께 급물살을 타게 되어 현재 총 약 2,000종의 세균 유전체가 완성되었고 초안으로 발표되는 유전체 수도 증가되고 있으며 이에 비례하여 미생물 유전체 염기서열 정보량 또한 상당량 축적되고 있다(Fig. 1). *Haemophilus influenzae* Rd (KW20)가 1995년 완전 해독된 첫 번째 세균 유전체로 발표된 이후(Fleischmann *et al.* 1995), 여러 동식물 병원성 세균들을 비롯하여 모델 미생물이나 산업적 유용 미생물들의 유전체가 해독되었다. 유전체 해독 기술은 인간, 동식

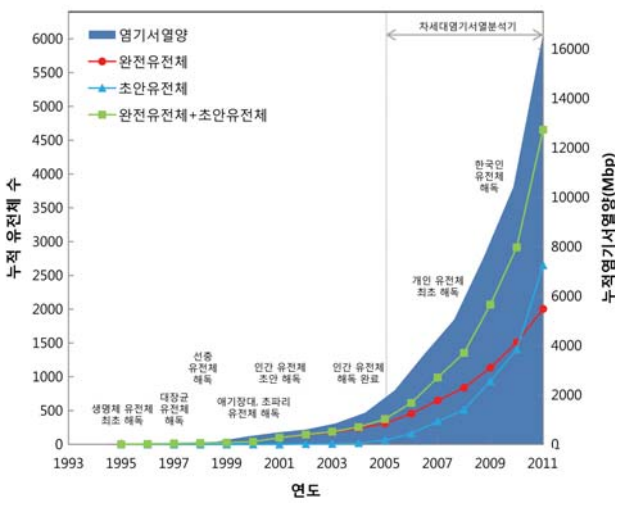

Fig. 1. Cumulative number of microbial genomes and consequent accumulation of sequence information. See GenBank statistic in June 2012.

물의 생명건강과 직접적인 영향을 주는 병원성 세균 연구에 큰 기여를 할 것이라고 기대를 모았다. 그 결과 *Helicobacter pylori, Staphylococcus aureus, Mycobacterium tuberculosis*를 비롯하여 (Tomb *et al.* 1997; Kuroda *et al.* 2001; Fleischmann *et al.* 2002), 해독된 유전체들은 병원성 인자 및 병원성 유전체로서의 특징을 찾기 위한 pathogenomics에 기여하고 있다. 대표적인 모델 세균인 *E. coli*의 유전체는 유전학 연구 및 단백질 생산 세포공장 개발에 활용하는 데 좋은 재료로서 그 유전체가 분석되었고 한편에서는 병원성 대장균의 유전체 해독을 통해 병원성 인자를 규명하기 위해 유전체 분석이 꾸준히 진행되어 왔다 (Blattner *et al.* 1997; Ohnishi *et al.* 2000; Jeong *et al.* 2009). 또한 NGS가 발전하면서 변이를 염기서열 수준에서 관찰할 수 있는 진화 연구도 더불어 활기를 띠고 있는데, 모델 세균인 대장균을 대상으로 진화 연구들이 수행되고 있으며 이러한 해독 결과들은 비교 유전체 분석에 활용되어 생명 현상의 기본 원리를 설명하는데 큰 뒷받침이 되고 있다 (Barrick *et al.* 2009). 자연계 물질대사 순환에 관여하는 rhizobium, cyanobacteria를 비롯 (Freiberg *et al*. 1997; Kaneko *et al*. 2003), *Azotobacter, Methanococcus* (Setubal *et al.* 2009), 난배양성 미생물 그룹인 mesophilic Crenarchaeota (Walker *et al.* 2010; Kim *et al.* 2011) 등의 유전체가 해독되면서 그들에 의한 생태계 내의 대사순환을 설명할 수 있는 유전적 단서를 제공하였다. 또한 *Burkholderia, Aromatoleum, Shewanella, Geobacter* 등 환경으로부터 분리된 균주는 방향족 화합물의 분해 능력이나 금속환원성 등의 오염 및 독성을 가진 환경을 정화할 수 있는 능력을 가지고 있어 생물분해, 환경정화에 의미 있게 활용될 가능성이 높으므로 이들 유전체 및 대사회로 분석이 시도되고 있다 (Kwak *et al.* in press, Rabus *et al.* 2002; Heidelberg *et al.* 2002; Methe *et al*. 2003). 농업, 어업 분야에서는 친환경적인 농작물 생산을 위해 미생물들을 종종 이용하고 있는데, 항균물질이나 식물생장촉진인자를 생산하는 *Paenibacillus polymyxa*의 유전체가 해독되어 식물유용인자를 생산하는 유전자들을 규명하고 산업적 생산까지 가능하게 되었다 (Kim *et al.* 2010). 또한 *Hahella chejuensis*과 같은 미생물들의 유전체가 해독을 통해 살적조 물질 생합성 및 조절 유전자를 찾음으로써 환경 생태계 유지 및 산업적 생산성을 높이는데 활용 가능성을 보여주었다 (Jeong *et al.* 2005). 새로운 의약을 개발하는데 있어서도 미생물 유전체 분석이 기초 자료를 제시해주고 있는데, 방선균과 같이 다양한 항생물질과 생리활성물질을 생산하는 균주로부터 이차대사산물 생합성 관련 유전정보들을 분석함으로써 이들을 의약 개발 및 생산에 적용하기 위한 시도들이 이뤄지고 있다 (Bentley *et al.* 2002, Song *et al.* 2010). 한편 생명건강과 관련하여 인간의 건강에 유익한 역할을 담당하는 여러 bifidobacteria나 lactobacilli의 유전체 서열도 해독되어왔다 (Altermann *et al.* 2005; Kim *et al.* 2009). 장내에서 분리된 유산균들의 유전체 분석을 통해 대사 및 유용 물질 생산을 분석하여 그 유용성을 입증하고 장내에 미치는 기능을 유추하여 probiotic으로써의 가치를 재발견하기도 한다.

유전체 해독을 위해서는 염기서열 분석기에 적용 가능한 수준의 DNA를 취득해야 하므로 대부분의 미생물 유전체 해독은 순수 배양이 가능한 미생물을 대상으로 이뤄져 왔다. 그러나 자연 환경에 존재하는 미생물 중 난배양성 미생물이 99%를 차지하고 있으며, 이들을 배양할 수 있는 대사 조건을 찾거나 유용 유전자를 탐색하기 위해 오히려 유전체 서열이 필요했다. 이러한 난제를 해결하기 위해 배양에 의존하지 않고 단일 세포 단위의 미생물을 가지고 유전체를 해독할 수 있는 single-cell genome sequencing이 고안되었다 (Woyke *et al.* 2009; Kalisky *et al.* 2011). 즉, microfluidic flow나 flow cytometry를 이용하여 단일 세포를 물리적으로 분리한 후, phi29 DNA polymerase를 이용한 multiple displacement amplification (MDA)를 수행하여 전 유전체를 증폭하게 되는데, 이렇게 증폭하여 얻은 MDA 산물을 가지고 유전체 해독을 수행할 수 있게 된 것이다 (Kalisky *et al.* 2011). 이 방법을 이용하여 환경의 탄소, 질소 순환에 영향을 미치는 암모니아 산화 고세균 중, *Nitrosoarchaeum limnia*을 해독하였고 (Blainey *et al.* 2011), 소 반추위의 메타지놈으로부터 난배양성 미생물 유전체들이 밝혀지게 되었다 (Hess *et al.* 2011). 이러한 single-cell genome sequencing은 NGS 기술의 개발에 힘입어 다룰 수 없는 미생물의 유전 정보를 보여주는 중요한 방법이 되었고, 99%의 미지의 미생물들에 대해 그들의 진화적 위치, 환경 내 대사 및 역할, 집단내 미생물 간의 상호작용, 유용 유전자 등을 연구할 수 있는 무궁무진한 개발 가능성을 보여주고 있다.

많은 유전체가 해독되어 왔음에도 불구하고 일각에서는 현재까지 해독된 미생물의 유전체들이 생리적 특성에 의해 임의적으로 선정된 것이 대부분이기 때문에 계통수 상에 존재하는 대표적인 미생물들을 대상으로 고르게 유전체 서열을 확보해야 할 필요성이 제기되었다. 이에 따라 미국의 Joint Genome Institute에서는 독일 DSMZ과 함께 Genomic Encyclopedia of Bacteria and Archaea (GEBA) 프로젝트라는 이름으로 진화적 연관성에 따라 대표적인 미생물들을 선정하여 유전체를 해독하고 있다. 이러한 대형 유전체 프로젝트는 대량의 시퀀스를 생산할 수 있는 차세대 염기서열 분석 기술의 진

보를 설명하는 단적인 예로 해석될 수 있다. GEBA 프로젝트를 통해 56종의 원핵 미생물 유전체가 일차적으로 발표되었고, 미생물 유전체 해독 면에서 계통분류학적인 간극을 채우는 것은 물론, 효과적인 새로운 유전자 발견이 가능하다는 것을 보여주었다. 결과적으로는 유전체 해독 및 분석으로부터 생물간의 유전체를 비교하고 진화적으로 설명하는 등 생물학적 의미를 이해하는 학문적인 성장이 분석 기술의 진보를 바탕으로 이루어진 대표적인 예라고 할 수 있다. 생명공학, 환경, 의학 분야에서 응용될 수 있는 산업적 유용 유전자 및 효소를 발굴하는 경쟁에서도 이러한 염기서열분석기술의 활용은 증가하는 추세이며, 더불어 방대한 양의 데이터 처리 및 보관, 그리고 정보를 해석할 수 있는 생물정보학적 분석 능력과 프로그램 및 기술 개발이 중요시 되고 있다.

메타지놈은 1998년 Jo Handelsman에 의해 정의된 것으로 환경 내 존재하는 미생물들로부터 수집된 유전체를 말하며, 이를 활용한 연구 접근 방법을 메타지노믹스, 환경생물학이라 한다 (Handelsman *et al.* 1998). 특정 환경의 DNA를 수집하여 직접 목적하는 유전물질을 획득, 연구하거나, 환경 및 생태를 고려한 유전적 다양성을 연구하는 것이 그 목적이다. 초기의 전형적인 메타지놈 프로젝트들은 주로 단백질 및 천연물의 발굴을 위하여 환경샘플의 DNA를 대장균에서 발현하여 클론 라이브러리 (clone library)를 제작한 후, 염기서열을 분석하고 유전자 기능 연구가 수반되는 기능적 접근 방식이었다. 근래에는 이와 더불어, 16s rRNA 서열과 같은 생물학적 유연관계를 설명할 수 있는 유전자나 특정 생물학적 관심이 있는 유전자 서열 기반 하에 클론을 선별하고 해당 미생물과 그 유전체를 목적으로 하는 염기서열 기반 접근 방식이 이루어지고 있다. 생거 시퀀싱 시대에는 비용 대비 낮은 용량이 염기서열 기반의 메타지놈 연구에 큰 걸림돌이었으나, 최근 NGS 및 생물정보학적 분석 기술이 고도화됨에 따라 한 환경 내 존재하는 미생물 커뮤니티의 생물 종 분포와 양상을 해석하는 것에서 대규모 시퀀싱과 정보 분석을 통해 커뮤니티 전체의 기능을 이해하려는 시도로 발전하였다. 이러한 연구는 생물학적 다양성 분석뿐 아니라, 대용량의 NGS 기술에 힘입어 유전 정보들을 대량으로 획득할 수 있으므로 전반적인 유전자 구성을 분석하여 대사 흐름을 유추하고, 전사체 분석을 통해 특정 환경 내에서 일어나는 생물 집합체 및 유전체 변이, 적응 등을 관찰할 수 있다. 또한 환경마다 물리화학적 조건에 따라 일어나는 변이 분석은 비교메타유전체 분석을 가능하게 하여 특정 환경의 진단과 해석, 예측에 활용하거나 특정 환경 내 발현 조절 및 기능성 유전자원의 발굴이 가능하게 되었다. 그 예로, 특정 물리적 자연 환경, 즉 추위에 적응한 미생물들을 메타지노믹스 접근 방법으로 해석하였을 때 저온 환경에서 생존하기 위한 분자 진화 및 적응 방법을 염기서열을 통해 이해하고, 저온 적응과 관련된 새로운 유전자를 탐색할 수 있다 (Casanueva *et al.* 2010). 또한 염기서열 기반의 특정 미생물 유전자를 발굴하고 탐색하여 실제 응용하려는 시도도 기능성 메타지노믹스와 서열 기반의 메타지노믹스를 함께 접목하여 이루어지고 있다. 즉, 메타지놈 염기서열 분석을 통해 신규 생물정보 자원의 획득과 동시에 기능적으로 우수한 유전 자원을 탐색 및 확보할 수 있게 된 시대가 된 것이다. Béjà 등은 해양에서 얻은 메타지놈으로부터 γ-Proteobacteria 세균 유래의 130 kb의 유전체 단편을 얻었고, 그 단편에서 빛을 에너지로 전환시킬 수 있는 로돕신을 가지고 있음을 처음으로 발견하게 되었다 (Béjà *et al.* 2000). 이 연구는 세균의 새로운 대사와 기능을 이해하게 된 계기가 되었을 뿐 아니라, 메타지놈 연구를 통해서 배양 불가능한 미생물의 새로운 유전자를 발견하고 산업적으로 이용할 수 있는 유전자원 획득의 좋은 예가 되었다.

다양한 미생물이 존재하는 공간인 해양과 토양에서는 지구상의 화학적 반응이 지대하고 빠르게 일어나고 있다는 것이 여러 메타지놈 연구를 통해 알려지고 있다. 2003년에 Craig Venter가 주도하여 해양 미생물 다양성을 연구하고자 시작했던 Global Ocean Sampling Expedition (GOS) 프로젝트의 시발점이 된 버뮤다 해역 (Sargasso Sea) 프로젝트가 큰 반향을 일으킨 적이 있다. 생거시퀀싱 방법을 이용하여 대량의 해양 메타지놈 염기서열을 해독하고 유전자 목록 작성과 다양성 조사를 수행하면서 새로운 미생물 그룹과 새로운 유전자들을 찾아내었으며, 그 중 새로운 로돕신 유사 광수용체 유형들을 다수 발견하여 해양 내 탄소 순환과 밀접한 다수의 로돕신 유사 단백질들의 존재 가능성을 확인하였다 (Venter *et al.* 2004). 이러한 미생물 다양성을 위한 해양 탐사는 Sorcerer II GOS 탐사로 계속 이어져 태평양과 대서양에 이르는 대형 메타지놈 염기서열을 확보 및 분석하면서 샘플을 채취한 지역마다의 메타지놈 정보 및 새롭게 확장된 protein family를 보여주고, 메타지놈 샘플 간 비교분석 방법을 제시하였다 (Rusch *et al.* 2007; Yooseph *et al.* 2007). 이러한 GOS 탐사를 통한 메타지놈 데이터들은 공개되어 호기성 비산소발생광합성 (Aerobic Anoxygenic Photosynthetic) (Yutin *et al.* 2007)이나 미생물 집단의 철 대사 및 흡수 유전자 분포 (Toulza *et al.* 2012)가 연구되는 등 지속적으로 활용되어 해양내의 미생물들의 대사, 생태, 진화, 환경에서의 기능을 통찰할 수 있는 보고들이 계속 되고 있다. 지구에서 큰 비율을 차지하는

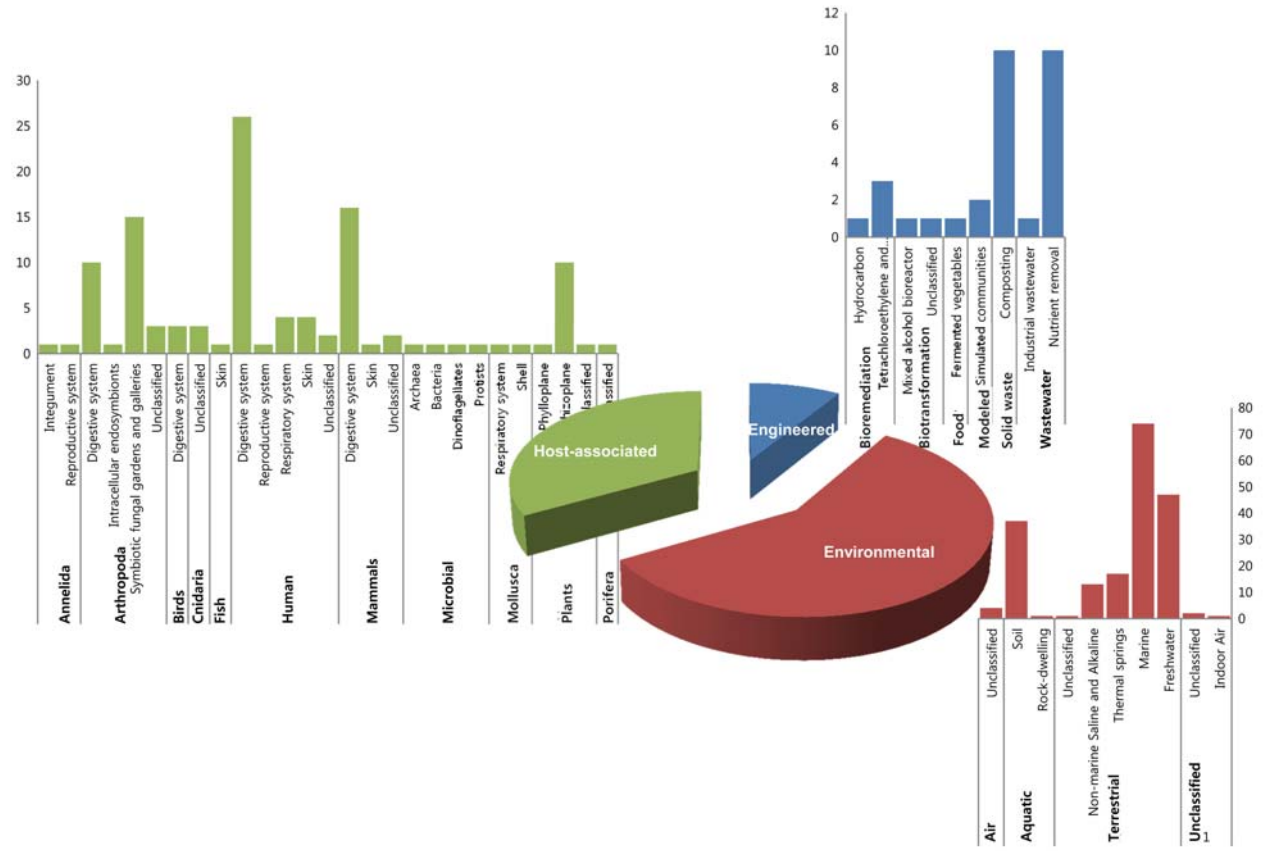

Fig. 2. Classification of metagenome analysis projects based on ecosystem and the project type. See the GOLD statistic in June 2012.

해양은 온도, 압력, 생물자원의 종류 등 여러 가지 환경 조건을 갖추고 있어 다양한 미생물들의 방대한 터전이며 새로운 종과 유전체 자원의 보고이다. 그리고 그들의 물리적, 생화학적 반응들이 유기적인 생태계를 구성하고 있다. 따라서 지구 생태계에 큰 역할을 하고 있는 해양은 메타지놈 연구에서 큰 관심사 중의 하나로 GOS 프로젝트 외에도 많은 해양 메타지놈 연구가 이뤄지고 있다(Fig. 2).

한편 육지 환경의 토양은 다양한 물리적, 화학적 특성에 따라 다양하게 나누어지고 이러한 성질에 따라 4×10^7 내지 $2\times10^9\,g^{-1}$의 원핵생물을 포함하면서 다양한 생물자원이 풍부하게 공존하여 메타지놈 연구에 주요한 관심 대상이다. 토양 내 식물의 생장 및 면역력은 미생물 유전체 분석을 통해서도 알려져 왔듯이 식물 뿌리와 잎에 서식하는 미생물과 밀접하게 연관되어 있다. 최근 질병억제 토양들을 대상으로 PhyloChip을 이용한 메타지놈 분석을 한 보고에서는 질병억제 토양들이 곰팡이에 의해 발병하는 작물의 질병을 억제와 관련된 미생물 그룹이 토양에서 확인되었고, 그 중 γ-Proteobacteria들의 nonribosomal peptide synthetases에 의해 합성되는 항생물질에 의해 억제 효과가 나타난다고 한다(Mendes *et al.* 2011). 대기 중의 질소와 탄소, 수소의 순환에도 환경과 미생물은 밀접한 상관관계를 가지고 있다. 북극 알래스카 지대의 동토 토양의 샘플을 대상으로 메타지놈을 분석한 결과, 지구 온난화에 따라 동토 토양이 녹으면서 메탄과 질소의 빠른 순환이 미생물 집단의 변화와 활성화되는 유전자들과 관련 있다고 보고된 바 있다(Mackelprang *et al.* 2011). 이는 특정 물리적 환경을 가지고 있는 토양 내 미생물들의 작용을 통해 토양 환경의 생화학적 변화 및 전 지구적인 물질 순환의 관계를 지적한 내용이다.

토양은 그 물리화학적 특성 때문에 DNA를 획득하는 단계부터 어려움이 있어 전반적인 연구 진전은 빠르게 이루어지지 않고 있음에도 불구하고, 동식물, 미생물이 상호작용을 하는 환경으로 인간의 삶과 건강에 직접적인 영향을 끼치는 공간이므로, Timothy M. Vogel 등을 중심으로 국제 컨소시엄, TerraGenome을 형성하기에 이르렀고, 토양 메타게놈의 연구의 어려움을 해소하고 연구의 표준화를 구축하기 위해 노력하고 있다.

현재 메타지노믹스의 대상은 자연 환경뿐 아니라 식

물, 동물 그리고 인간에 이르기까지 다양하게 확대되어 가고 있다. NGS의 계발에 따라 메타지놈 분석을 통해 인간의 건강과 직결된 공생 미생물들에 대한 궁금증도 해소될 수 있는 방법이 모색되면서 질병과 인체 메타지놈의 연관성 연구가 많이 수행되고 있다. 인체 장내 세균의 종류에 따라 비만이 유도된다는 보고는 메타지놈 분석을 통해 질병 치료법이 제시될 수 있음을 크게 시사하였고, 비알콜성 간염 (Henao-Mejia *et al.* 2012) 이나 염증성 장질환 (Elinav *et al.* 2011) 등 대사질환, 자가면역성 질환이 장내 미생물과 밀접한 연관이 있음을 보여주고 있다. 인체의 미생물들을 통해 질병의 예측, 진단, 치료의 가능성을 보고 여러 나라의 연구진들이 데이터를 공유하고 국제적으로 상호 협력하기 위한 International Human Microbiome Consortium이 구성되었고 HMP, MetaHIT 등의 프로그램을 통해 인체 장내 미생물 유전자 카탈로그를 만들고 (Qin *et al.* 2010), 장내 뿐 아니라 인체 곳곳의 미생물 다양성을 조사하고 카탈로그를 작성하는 등 기초적인 microbiome 유전 정보들을 수집하고 건강 및 질병과 관련된 인간과 미생물간의 상관관계를 연구 중이다.

인간, 동물, 식물, 미생물 사이의 상호작용의 이해를 돕는 데에도 메타지놈 분석 방법이 이용되고 있다. 즉, 공생, 기생, 병원성 등의 상호작용을 염기서열 분석을 통해 모니터링할 뿐 아니라, 질병, 에너지, 환경 및 생태 복원 등에 해결책을 제시할 수 있는 방법들을 메타지놈 분석을 통해 찾고 있다.

응용과 앞으로 나아가야 할 방향

차세대 염기서열 분석기의 지속적인 개발과 새로운 분석 기술의 출현이 시사하고 있듯이 앞으로 환경 및 생물 연구에 있어서 생물정보와 관련 기술은 지속적으로 진보될 것이다. 또한 생물체의 유전, 형질, 생리, 대사 등을 비롯하여 의약, 농업 등 산업적 응용 및 개발 뿐 아니라, 환경 및 생태 등이 앞으로 현 세기의 세대들이 해결해야 할 문제라면 환경 및 생물을 다루는 연구자들에게 있어서도 미생물 유전체와 메타지놈 연구 및 생물정보학은 간과할 수 없는 분야가 될 것이다. GEBA와 같은 대규모의 프로젝트들이 이미 제시하였듯이 미생물 유전체의 커다란 진화적 기본 유전체 골격을 만드는 것은 새로운 자원 발견 및 유전적 이해를 돕게 될 것이고, 한편 메타지놈은 유전체 정보들을 기반으로 환경에 존재하는 집단에 대한 이해와 및 상호작용을 이해하는 큰 그림을 그리는 데에 일조하게 될 것이다.

사 사

이 논문은 한국연구재단 이공분야 기초연구사업 중견연구자지원사업 (도약연구-도전연구; 과제번호: 2011-0017670)과 선도연구센터지원사업 (SRC) 곰팡이병원성 연구센터 및 농촌진흥청 차세대바이오그린21사업 차세대유전체연구사업 (과제번호: PJ008201012011)의 지원으로 작성되었습니다.

참 고 문 헌

Altermann E, WM Russell, MA Azcarate-Peril, R Barrangou, BL Buck, O McAuliffe, N Souther, A Dobson, T Duong, M Callanan, S Lick, A Hamrick, R Cano and TR Klaenhammer. 2005. Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. Proc. Natl. Acad. Sci. U.S.A. 102:3906-3912.

Béjà O, L Aravind, EV Koonin, MT Suzuki, A Hadd, LP Nguyen, SB Jovanovich, CM Gates, RA Feldman, JL Spudich, EN Spudich and EF DeLong. 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289:1902-1906.

Barrick JE, DS Yu, SH Yoon, H Jeong, TK Oh, D Schneider, RE Lenski and JF Kim. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461:1243-1247.

Bentley DR, S Balasubramanian, HP Swerdlow, GP Smith, J Milton *et al.* 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.

Bentley SD, KF Chater, AM Cerdeno-Tarraga, GL Challis, NR Thomson *et al.* 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417:141-147.

Besemer J, A Lomsadze and M Borodovsky. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic. Acids. Res. 29:2607-2618.

Bickhart DM, Y Hou, SG Schroeder, C Alkan, MF Cardone, LK Matukumalli, J Song, RD Schnabel, M Ventura, JF Taylor, JF Garcia, CP Van Tassell, TS Sonstegard, EE Eichler and GE Liu. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res. 22:778-790.

Blainey PC, AC Mosier, A Potanina, CA Francis and SR Quake. 2011. Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis,

PLos One 6.

Blattner FR, G Plunkett, 3rd, CA Bloch, NT Perna, V Burland, M Riley, J Collado-Vides, JD Glasner, CK Rode, GF Mayhew, J Gregor, NW Davis, HA Kirkpatrick, MA Goeden, DJ Rose, B Mau and Y Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277:1453-1462.

Boetzer M, CV Henkel, HJ Jansen, D Butler and W Pirovano. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578-579.

Bult CJ, O White, GJ Olsen, L Zhou, RD Fleischmann, GG Sutton, JA Blake, LM FitzGerald, RA Clayton, JD Gocayne, AR Kerlavage, BA Dougherty, JF Tomb, MD Adams, CI Reich, R Overbeek, EF Kirkness, KG Weinstock, JM Merrick, A Glodek, JL Scott, NS Geoghagen and JC Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058-1073.

Cai Y and Y Sun. 2011. ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasi-linear computational time. Nucleic. Acids. Res. 39:e95.

Caporaso JG, J Kuczynski, J Stombaugh, K Bittinger, FD Bushman *et al.* 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7:335-336.

Casanueva A, M Tuffin, C Cary and DA Cowan. 2010. Molecular adaptations to psychrophily: the impact of 'mic' technologies. Trends Microbiol. 18:374-381.

Dayarian A, TP Michael and AM Sengupta. 2010. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. BMC Bioinformatics 11.

Delcher AL, D Harmon, S Kasif, O White and SL Salzberg. 1999. Improved microbial gene identification with GLIMMER. Nucleic. Acids. Res. 27:4636-4641.

Dinsdale EA, RA Edwards, D Hall, F Angly, M Breitbart, JM Brulc, M Furlan, C Desnues, M Haynes, L Li, L McDaniel, MA Moran, KE Nelson, C Nilsson, R Olson, J Paul, BR Brito, Y Ruan, BK Swan, R Stevens, DL Valentine, RV Thurber, L Wegley, BA White and F Rohwer. 2008. Functional metagenomic profiling of nine biomes. Nature 452: 629-632.

Elinav E, T Strowig, AL Kau, J Henao-Mejia, CA Thaiss, CJ Booth, DR Peaper, J Bertin, SC Eisenbarth, JI Gordon and RA Flavell. 2011. NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. Cell 145:745-757.

Fleischmann RD, D Alland, JA Eisen, L Carpenter, O White, J Peterson, R DeBoy, R Dodson, M Gwinn, D Haft, E Hickey, JF Kolonay, WC Nelson, LA Umayam, M Ermolaeva, SL Salzberg, A Delcher, T Utterback, J Weidman, H Khouri, J Gill, A Mikula, W Bishai, WR Jacobs Jr, Jr., JC Venter and CM Fraser. 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J. Bacteriol. 184:5479-5490.

Fleischmann RD, MD Adams, O White, RA Clayton, EF Kirkness *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-512.

Fraser CM, JD Gocayne, O White, MD Adams, RA Clayton *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397-403.

Freiberg C, R Fellay, A Bairoch, WJ Broughton, A Rosenthal and X Perret. 1997. Molecular basis of symbiosis between Rhizobium and legumes. Nature 387:394-401.

Galand PE, EO Casamayor, DL Kirchman and C Lovejoy. 2009. Ecology of the rare microbial biosphere of the Arctic Ocean. Proc. Natl. Acad. Sci. U.S.A. 106:22427-22432.

Hallam SJ, KT Konstantinidis, N Putnam, C Schleper, Y Watanabe, J Sugahara, C Preston, J de la Torre, PM Richardson and EF DeLong. 2006. Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*, Proc. Natl. Acad. Sci. U.S.A. 103:18296-18301.

Hamady M, C Lozupone and R Knight. 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J. 4:17-27.

Handelsman J, MR Rondon, SF Brady, J Clardy and RM Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5:R245-249.

Heidelberg JF, IT Paulsen, KE Nelson, EJ Gaidos, WC Nelson *et al.* 2002. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. Nat. Biotechnol. 20:1118-1123.

Henao-Mejia J, E Elinav, C Jin, L Hao, WZ Mehal, T Strowig, CA Thaiss, AL Kau, SC Eisenbarth, MJ Jurczak, JP Camporez, GI Shulman, JI Gordon, HM Hoffman and RA Flavell. 2012. Inflammasome-mediated dysbiosis regulates progression of NAFLD and obesity. Nature 482:179-185.

Hess M, A Sczyrba, R Egan, TW Kim, H Chokhawala, G Schroth, S Luo, DS Clark, F Chen, T Zhang, RI Mackie, LA Pennacchio, SG Tringe, A Visel, T Woyke, Z Wang and EM Rubin. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331:463-467.

Jeong H, JH Yim, C Lee, SH Choi, YK Park, SH Yoon, CG Hur, HY Kang, D Kim, HH Lee, KH Park, SH Park, HS Park, HK Lee, TK Oh and JF Kim. 2005. Genomic blueprint of *Hahella chejuensis*, a marine microbe producing an algicidal agent. Nucleic. Acids. Res. 33:7066-7073.

Jeong H, V Barbe, CH Lee, D Vallenet, DS Yu, SH Choi, A Couloux, SW Lee, SH Yoon, L Cattolico, CG Hur, HS Park, B Segurens, SC Kim, TK Oh, RE Lenski, FW Studier, P

Daegelen and JF Kim. 2009. Genome sequences of *Escherichia coli* B strains REL606 and BL21 (DE3). J. Mol. Biol. 394: 644-652.

Jung JY, SH Lee, JM Kim, MS Park, JW Bae, Y Hahn, EL Madsen and CO Jeon. 2011. Metagenomic analysis of kimchi, a traditional Korean fermented food. Appl. Environ. Microbiol. 77:2264-2274.

Kalisky T, P Blainey and SR Quake. 2011. Genomic analysis at the single-cell level. Annual Review Genetics. 45:431-445.

Kaneko T, Y Nakamura, S Sasamoto, A Watanabe, M Kohara, M Matsumoto, S Shimpo, M Yamada and S Tabata. 2003. Structural analysis of four large plasmids harboring in a unicellular cyanobacterium, *Synechocystis* sp. PCC 6803. DNA Res. 10:221-228.

Katoh K and H Toh. 2010. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 26:1899-1900.

Kim BK, MY Jung, DS Yu, SJ Park, TK Oh, SK Rhee and JF Kim. 2011. Genome sequence of an ammonia-oxidizing soil archaeon, "*Candidatus* Nitrosoarchaeum koreensis" MY1. J. Bacteriol. 193:5539-5540.

Kim JF, H Jeong, DS Yu, SH Choi, CG Hur, MS Park, SH Yoon, DW Kim, GE Ji, HS Park and TK Oh. 2009. Genome sequence of the probiotic bacterium *Bifidobacterium animalis* subsp. *lactis* AD011. J. Bacteriol. 191:678-679.

Kim JF, H Jeong, SY Park, SB Kim, YK Park, SK Choi, CM Ryu, CG Hur, SY Ghim, TK Oh, JJ Kim, CS Park and SH Park. 2010. Genome sequence of the polymyxin-producing plant-probiotic rhizobacterium *Paenibacillus polymyxa* E681. J. Bacteriol. 192:6103-6104.

Kuroda M, T Ohta, I Uchiyama, T Baba, H Yuzawa *et al*. 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 357: 1225-1240.

Kwak MJ, JY Song, H Jeong, SY Kim, SG Kang, BK Kim, SK Kwon, CH Lee, DS Yu, SH Park and JF Kim. in press. Genome Sequence of the Endophytic Bacterium *Burkholderia* sp. KJ006, J. Bacteriol.

Lai B, R Ding, Y Li, L Duan and H Zhu. 2012. A de novo metagenomic assembly program for shotgun DNA reads. Bioinformatics 28:1455-1462.

Lan Y, Q Wang, JR Cole and GL Rosen. 2012. Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. PLoS One 7:e32491.

Lander ES, LM Linton, B Birren, C Nusbaum, MC Zody *et al.* 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Li H and R Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li H, J Ruan and R Durbin. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18:1851-1858.

Li R, H Zhu, J Ruan, W Qian, X Fang, Z Shi, Y Li, S Li, G Shan, K Kristiansen, S Li, H Yang, J Wang and J Wang. 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20:265-272.

Li W and A Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.

Mackelprang R, MP Waldrop, KM DeAngelis, MM David, KL Chavarria, SJ Blazewicz, EM Rubin and JK Jansson. 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480:368-371.

McHardy AC, HG Martin, A Tsirigos, P Hugenholtz and I Rigoutsos. 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods 4:63-72.

McKernan KJ, HE Peckham, GL Costa, SF McLaughlin, Y Fu *et al.* 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 19:1527-1541.

Mendes R, M Kruijt, I de Bruijn, E Dekkers, M van der Voort, JH Schneider, YM Piceno, TZ DeSantis, GL Andersen, PA Bakker and JM Raaijmakers. 2011. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science 332:1097-1100.

Methe BA, KE Nelson, JA Eisen, IT Paulsen, W Nelson *et al.* 2003. Genome of Geobacter sulfurreducens: metal reduction in subsurface environments. Science 302:1967-1969.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11:31-46.

Mitra S, M Stark and DH Huson. 2011. Analysis of 16S rRNA environmental sequences using MEGAN. BMC Genomics 12 Suppl 3:S17.

Ning Z, AJ Cox and JC Mullikin. 2001. SSAHA: a fast search method for large DNA databases. Genome Res. 11:1725-1729.

Nishito Y, Y Osana, T Hachiya, K Popendorf, A Toyoda, A Fujiyama, M Itaya and Y Sakakibara. 2010. Whole genome assembly of a natto production strain Bacillus subtilis natto from very short read data. BMC Genomics 11:243.

Notredame C, DG Higgins and J Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205-217.

Ohnishi M, T Murata, K Nakayama, S Kuhara, M Hattori, K Kurokawa, T Yasunaga, K Yokoyama, K Makino, H Shinagawa and T Hayashi. 2000. Comparative analysis of the whole set of rRNA operons between an enterohemorrhagic *Escherichia coli* O157:H7 Sakai strain and an *Escherichia*

coli K-12 strain MG1655. Syst. Appl. Microbiol. 23:315-324.

Pareek CS, R Smoczynski and A Tretyn. 2011. Sequencing technologies and genome sequencing. J. Appl. Genet. 52: 413-435.

Pignatelli M and A Moya. 2011. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. PLoS One 6:e19984.

Qin J, R Li, J Raes, M Arumugam, KS Burgdorf *et al.* 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59-65.

Quaiser A, T Ochsenreiter, HP Klenk, A Kletzin, AH Treusch, G Meurer, J Eck, CW Sensen and C Schleper. 2002. First insight into the genome of an uncultivated crenarchaeote from soil. Environ. Microbiol. 4:603-611.

Rabus R, M Kube, A Beck, F Widdel and R Reinhardt. 2002. Genes involved in the anaerobic degradation of ethylbenzene in a denitrifying bacterium, strain EbN1. Arch. Microbiol. 178:506-516.

Rusch DB, AL Halpern, G Sutton, KB Heidelberg, S Williamson *et al.* 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5:e77.

Sanger F, S Nicklen and AR Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74:5463-5467.

Schloss PD and J Handelsman. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbio. 71:1501-1506.

Schloss PD, SL Westcott, T Ryabin, JR Hall, M Hartmann, EB Hollister, RA Lesniewski, BB Oakley, DH Parks, CJ Robinson, JW Sahl, B Stres, GG Thallinger, DJ Van Horn and CF Weber. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75:7537-7541.

Setubal JC, P dos Santos, BS Goldman, H Ertesvag, G Espin *et al.* 2009. Genome sequence of Azotobacter vinelandii, an obligate aerobe specialized to support diverse anaerobic metabolic processes. J. Bacteriol. 191:4534-4545.

Sievers F, A Wilm, D Dineen, TJ Gibson, K Karplus, W Li, R Lopez, H McWilliam, M Remmert, J Soding, JD Thompson and DG Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539.

Simpson JT, K Wong, SD Jackman, JE Schein, SJ Jones and I Birol. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117-1123.

Song JY, H Jeong, DS Yu, MA Fischbach, HS Park, JJ Kim, JS Seo, SE Jensen, TK Oh, KJ Lee and JF Kim. 2010. Draft genome sequence of *Streptomyces clavuligerus* NRRL 3585, a producer of diverse secondary metabolites. J. Bacteriol. 192:6317-6318.

Tomb JF, O White, AR Kerlavage, RA Clayton, GG Sutton *et al.* 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539-547.

Toulza E, A Tagliabue, S Blain and G Piganeau. 2012. Analysis of the global ocean sampling (GOS) project for trends in iron uptake by surface ocean microbes. PLos One 7:e30931.

Tsai IJ, TD Otto and M Berriman. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 11:R41.

Tyson GW, J Chapman, P Hugenholtz, EE Allen, RJ Ram, PM Richardson, VV Solovyev, EM Rubin, DS Rokhsar and JF Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43.

Valouev A, J Ichikawa, T Tonthat, J Stuart, S Ranade, H Peckham, K Zeng, JA Malek, G Costa, K McKernan, A Sidow, A Fire and SM Johnson. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. 18:1051-1063.

Venter JC, K Remington, JF Heidelberg, AL Halpern, D Rusch, JA Eisen, D Wu, I Paulsen, KE Nelson, W Nelson, DE Fouts, S Levy, AH Knap, MW Lomas, K Nealson, O White, J Peterson, J Hoffman, R Parsons, H Baden-Tillson, C Pfannkoch, YH Rogers and HO Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74.

Venter JC, MD Adams, EW Myers, PW Li, RJ Mural *et al.* 2001. The sequence of the human genome. Science 291: 1304-1351.

Walker CB, JR de la Torre, MG Klotz, H Urakawa, N Pinel *et al.* 2010. *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. Proc. Natl. Acad. Sci. U.S.A. 107:8818-8823.

Woyke T, G Xie, A Copeland, JM Gonzalez, C Han, H Kiss, JH Saw, P Senin, C Yang, S Chatterji, JF Cheng, JA Eisen, ME Sieracki and R Stepanauskas. 2009. Assembling the marine metagenome, one cell at a time. PLoS One 4:e5299.

Xing Y, D Medvin, G Narasimhan, D Yoder-Himes and S Lory. 2011. CloG: A pipeline for closing gaps in a draft assembly using short reads. In Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on, pp. 202-207.

Yooseph S, G Sutton, DB Rusch, AL Halpern, SJ Williamson *et al.* 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS

Biol. 5:e16.

Yutin N, MT Suzuki, H Teeling, M Weber, JC Venter, DB Rusch and O Béjà. 2007. Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. Environ. Microbiol. 9:1464-1475.

Zerbino DR and E Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

Received: 11 June 2012
Revised: 14 June 2012
Revision accepted: 19 June 2012