메타지놈 기반 지놈확보 및 분석 최신동향

아주대학교 응용화학생명공학과 이평천 교수
최보현 박사과정

1. 개요

합성생물학 연구 (특히, 대사회로 구축 및 최적화)를 위해서는 다양한 유전자 및 유전자 발현 조절 인자 등의 확보가 필요하다. 이를 위해 현재 다양한 미생물 유전체 분석 및 메타지놈 분석이 수행되고 있다. 특히 미생물 순수분리가 어려운 다양한 장소 (심해부터 장내 균총)에서 미생물 유전체를 직접 분리하고 분석하는 메타지놈(metagenome)에 대한 연구가 활성화 되고 있다[1]. 자연계에 존재하는 미생물 균총 (natural microbial community)은 바이러스를 포함한 다양한 미생물과 크로모좀 및 플라즈미드 등이 다양하게 존재하고 있다. 보통 메타지놈데이터를 분석하는 방법은 기존에 알려지는 레퍼런스 지놈 (reference genomes) 과 비교분석을 통하여 OTU (Operational Taxonomic Unit) 분석, 계통분석, 지놈 주석 (annotation)과 인실리코 대사회로 재구축 등이 수행 된다 (그림 1). 하지만 메타지놈의 레퍼런스 지놈과의 비교 분석은 다양한 환경에서 존재하는 미생물의 수 및 다양성이 현재 구축되고 있는 레퍼런스 지놈 수에 비하여 매우 크기 때문에 효율적이고 정확한 분석이 어렵다. 이러한 제한을 극복할 수 있는 방법으로 최근 상호 연관성 있는 유전자의 구간화 (binning)을 이용한 메타지놈 기반 지놈 (metagenome assembled genome, MAG) 유전체 분석이 소개되었고 이를 통한 다양한 메타지놈 기반 지놈이 발견되고 있다[2]. 본 글에서는 구간화 기법을 이용한 메타지놈 기반 지놈 제작에 대하여 간략히 살펴보고자 한다.

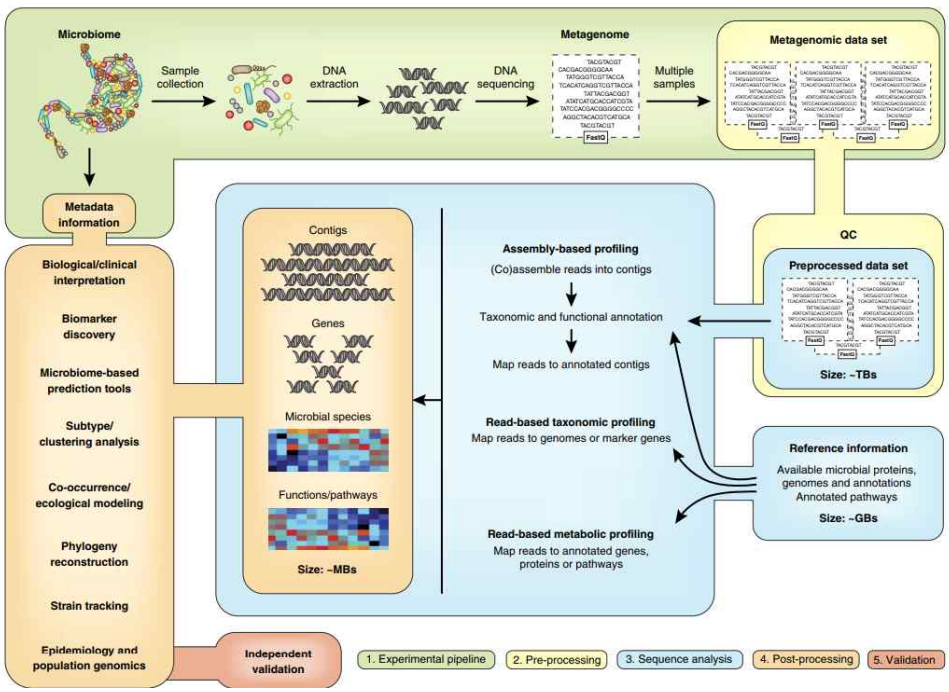


그림 1. 메타지놈 분석 단계 개요도. 다양한 샘플 확보부터 메타지놈 데이터 분석 (전처리 및 후처리 포함) 단계 (Quince et al. Nat Biotechnol 2017).

2. 메타지노믹 콘티그 (metagenomic contig)의 구간화(binning)를 통한 지놈어셈블리

메타지노믹 어셈블러의 결과물로 다양한 길이의 콘티그(contigs)가 생성된다. 다음 단계는 생성된 다양한 콘티그를 클러스터링하여 미생물 종이나 하위종 수준 지놈으로 파티션 하거나 Taxonomic 분석을 통하여 계통분석을 수행한다[3] (그림 2).

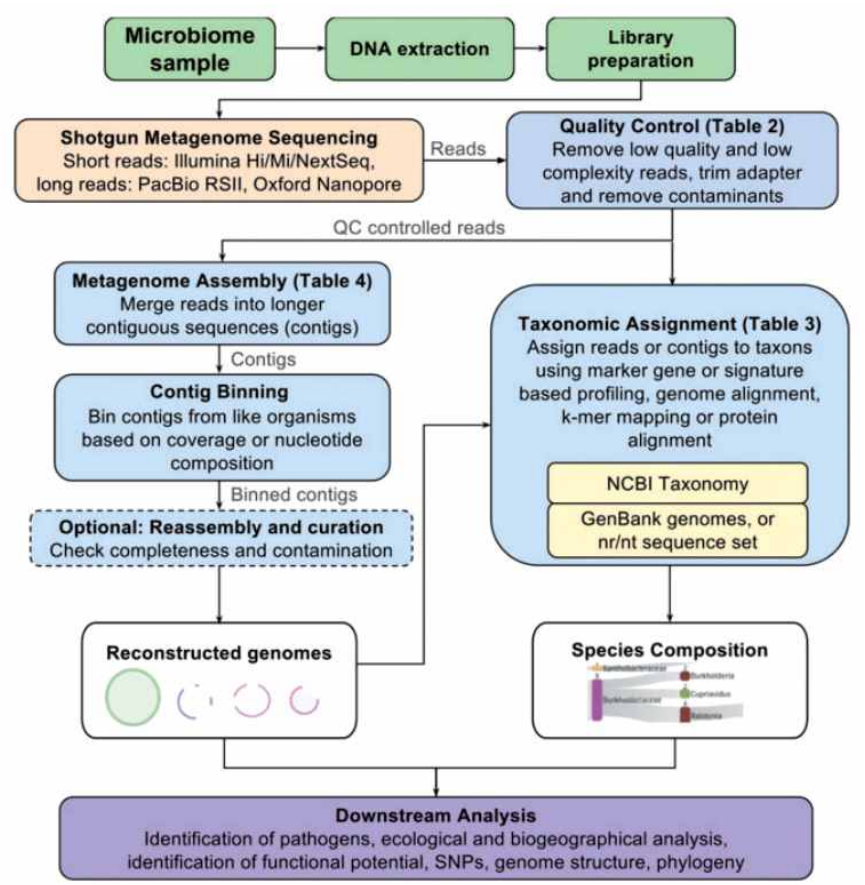
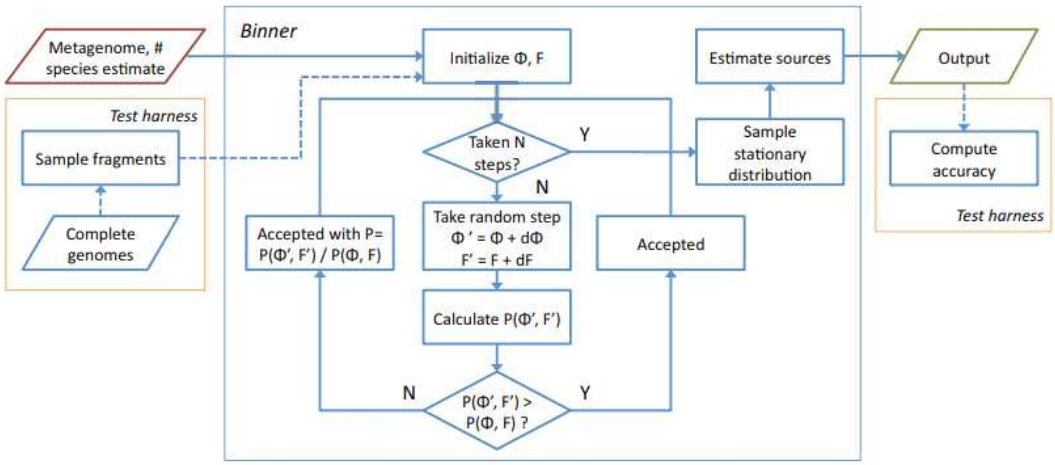


그림 2. 메타지놈 분석을 통한 지놈분석과 미생물 다양성 분석 단계 개요도. 메타지놈으로부터 어셈블리된 지놈 제작 및 미생물 다양성 분석 흐름 (Breitwieser et al. Briefings in Bioinformatics 2017).

기존의 알고리즘은 supervised learning 기법이 주로 이루어 졌으나 최근에는 unsupervised learning 기법을 활용한 구간화를 이용하는 방법이 소개되었다. LikelyBin 패키지[4]는 shotgun sequence을 sequence의 조성 (composition)을 기준으로 unsupervised Markov Chain Monte Carlo 기법을 이용하여 성공적으로 low-complexity 샘플로부터 다양한 지노믹 콘티그를 클러스터링하였다 (그림 3). 이와 같이 read나 콘티그 sequence의 조성 기준으로 메타지노믹 콘티그를 구간화하는 프로그램으로는 MetaWatt[5], CompostBin[6], SCIMM[7] 등이 있다. 앞선 4개의 프로그램과 달리 2014년에 소개된 CONCOCT[8]는 메타지노믹 콘티그를 sequence의 조성 (composition)과 커버러지 (coverage)를 이용하여 클러스터링하는 프로그램으로 다양성이 높은 미생물균총의 메타지놈으로 부터 성공적으로 높은 수준의 완성도를 갖는 지놈을 확보하였다. 커버러지가 낮은 메타지놈을 분석하기 위한 추가적인 구간



.

그림 3. LilelyBin의 binning pathway. LilelyBin 패기지의 메타지노믹 콘티그의 구간화의 흐름과 Markov Chain Monte Carlo의 iteration loop 모식도 (Kislyuk et al. BMC Bioinformatics 2009)

화하는 프로그램이 개발 중이다.

최근에 상기한 구간화 기법을 활용하여 기존의 공개된 메타지놈 데이터 (Sequence Read Archive, SRA)로부터 약 8000개의 메타지놈 기반 지놈을 확보한 연구가 보고되었다[9]. 논문 연구진에 의하면 약 1500개의 공개된 메타지놈 데이터로부터 7,903개의 박테리아와 고세균의 지놈을 확보하였으며, 확보된 메타지놈 기반 지놈은 50% 이상의 완성도를 갖고 있으며 이중 50%이상의 지놈은 90% 이상의 완성도를 갖는다고 보고하였다 (그림 4). 확보된 메타지놈 기반 지놈 분석을 통하여 17개의 새로운 박테리아 문 (phyla)과 3개의 고세균 문을 발견하였다 (그림 5).

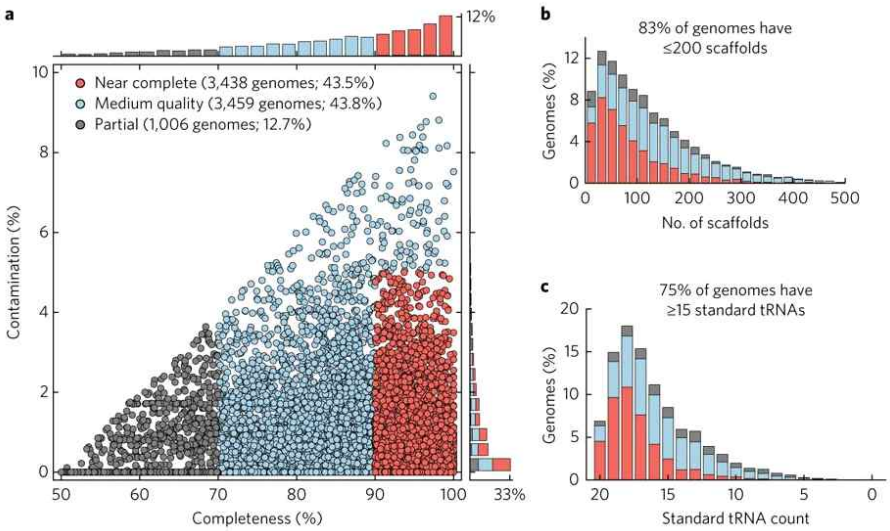


그림 4. 공개된 메타지놈으로부터 확보된 7,903 지놈. 메타지놈으로부터 완성된 (a) 지놈의 완성도, (b) 스카폴드 크기 분포, (c) tRNA 개수 (Parks,et al. Nat Microbiol 2017)

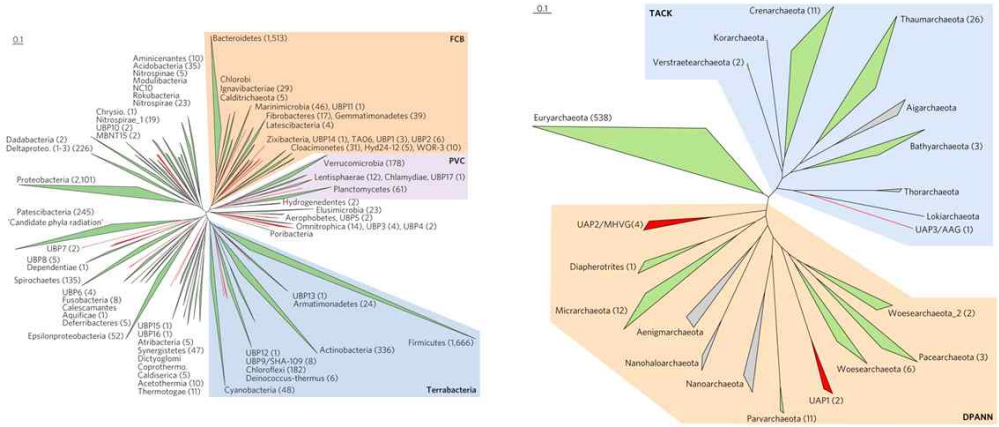


그림 5. 메타지놈으로부터 확보된 지놈의 계통분석. 왼쪽 박테리아, 오른쪽 고세균 (Parks,et al. Nat Microbiol 2017)

가장 최근에는 러시아의 캄차카 반도의 화산 부근의 온천으로 부터 메타지놈 기반 지놈을 확보하여 미생물의 다양성을 보고한 논문이 발표되었다[10]. 구간화 기법을 활용하여 36 개의 메타지놈 기반 지놈을 신규로 확보하였다 (그림 6).

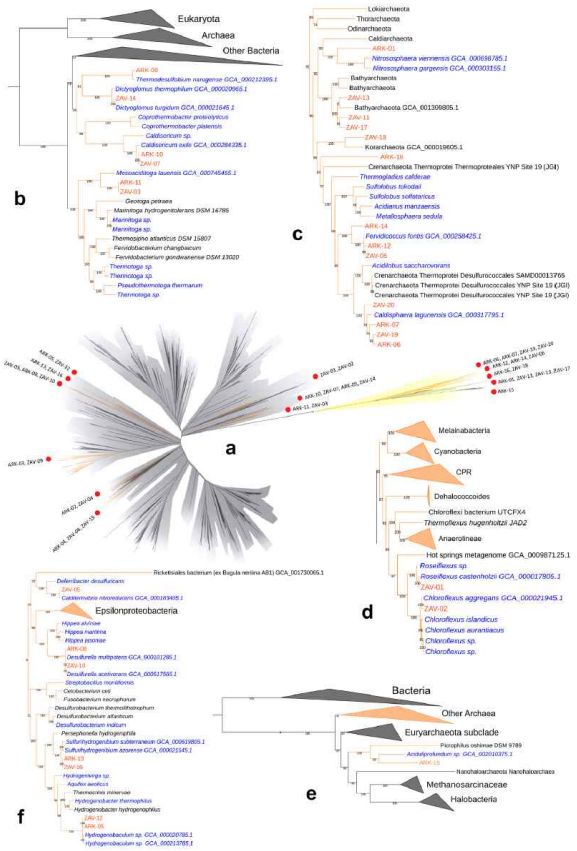


그림 6. 메타지놈으로부터 확보된 지놈의 계통분석. (Wilkins, et al. Scientific Report 2017)

3. 맺음말

합성생물학에서 다양하게 활용할 수 있는 요소부품의 확보는 필수적이다. 이러한 요소부품은 현재 공개된 광범위한 지놈정보 뿐만 아니라 다양한 환경 (극한 환경)에 존재하는 미생물 균총으로부터 확보될 수 있다. 특히, 연관성 있는 유전자/콘티그의 구간화를 이용한 메타지놈 기반 지놈의 확보는 앞으로 다양한 미생물의 발견을 촉진시킬 것으로 판단된다. 환경에 존재하는 바이러스를 포함한 다양한 미생물과 플라즈미드와 같은 extra-chromosomal DNA는 생명체의 다양성 분석뿐만 아니라 재조합 기술의 발전 등에도 크게 기여할 것으로 판단된다. 또한 메타지놈으로부터 확보된 지놈은 완성도에 따라서 인실리코 대사회로 재구축을 수행할 수 있으며, 이를 통하여 새로운 대사회로를 발견하고 합성생물학적 활용에 넓게 이용될 수 있을 것으로 판단된다.

4. 참고문헌

1) Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N. Nat Biotechnol 35, 833 (2017).

2) Ayling, M., Clark, M.D. & Leggett, R.M. Briefing Bioinformatics 10, 1 (2019).

3) Breitwieser, F.P., Lu, J. & Salzberg, S.L. Briefing Bioinformatics 10, 1 (2017).

4) Kislyuk, A., Bhatnagar, S., Dushoff, J. & Weitz, J.S. BMC Bioinformatics 10, 316 (2009).

5) Strous, M., Kraft, B., Bisdorf, R. & Tegetmeyer, H. Front. Microbiol. 3, 410 (2012).

6) Chatterji, S., Yamazaki, I., Bai, Z. & Eisen, J.A. in Res. Comput. Mol. Biol. (eds. Vingron, M. & Wong, L.) 17-28 (Springer, 2008).

7) Kelley, D.R. & Salzberg, S.L. BMC Bioinformatics 11, 544 (2010).

8) Alneberg J, Bjarnason BS, de Bruijn I, et al. Nat Methods 11, 1144 (2014)

9) Parks, D.H., Rinke, C., Chuvochina, M. et al. Nat Microbiol 2, 1533 (2017).

10) Wilkins, L, Ettinger, C., Jospin, G. et al. Scientific Report 9, 3059 (2019).