Analysis Report 설명서
(FLX - 16s rDNA metagenome)

분석방법

1. Raw data 생성

Sequencing을 통해 얻은 raw images는 454에서 제공하는 소프트웨어를 이용하여 data processing (image processing, signal processing)을 거쳐 raw data (SFF - Standard Flowgram Format)를 생성합니다.

Image processing은 raw images의 normalization 단계를 거쳐 각 flow 별 raw signals를 생성하는 단계이고 Signal processing은 raw signal의 correction, filtering, trimming 과정을 거친 후 basecalling을 하게 됩니다.

* Flowgram : 각 well에 대한 light intensity를 나타내는 signal로, signal의 강도는 nucleotide 수에 비례합니다.

2. 데이터 분석

위와 같은 과정을 거쳐 생성된 raw data를 barcode sequence를 이용하여 각 샘플별로 분류하고 각각의 리드를 RDP 데이타베이스와 비교(local alignment 수행)하여 최상의 매치 정보를 찾아 미생물 동정을 합니다.

또한 리드간의 시퀀스 유사성을 검사하여 OTU를 구하고 샘플별 종의 풍부도,균등도, 우점도 및 주성분 분석, rarefaction 분석, Tree 작성 등의 통계분석을 진행합니다.

데이터 다운로드

1. Raw data 다운로드

보고서의 '5.Data Download → 5.1. Raw data'를 보시면 SFF 파일과 Read sequence/quality 파일이 링크되어 있습니다. 결과 데이타는 압축된 상태로 제공되오니 알집이나 winzip 등을 이용하셔서 압축을 해제하시면 됩니다.

SFF file은 trace data가 들어 있는 파일로 common header section/read header section/read data section으로 이루어져 있습니다. SFF파일은 바이너리 형식이므로 바이너리를 볼 수 있는 SFF viewer 프로그램이 필요합니다. 여러 툴이 있지만 그 중 하나를 링크해 드리오니 참고하시기 바랍니다.

- SFF Workbench : http://www.dnabaser.com/download/SFF tools/

Sequence/quality file은 각각 리드의 시퀀스와, quality 정보를 담고 있습니다.
Sequence/quality 파일은 텍스트 형식이므로 일반 메모장이나 워드, 에디트 파일을 이용해서 열 수 있지만 파일 사이즈가 너무 클 경우에는 리눅스를 이용하시는 것이 수월합니다.

2. 분석결과 다운로드

보고서의 '5.Data Download → 5.2. Results of Analysis'를 보시면 샘플별 sequence/quality, 미생물 동정, OTU분석 등 모든 분석결과 데이타가 하나의 파일로 압축되어 있습니다.

결과 데이터

1. 결과 파일

링크를 통해 다운로드하신 파일은 아래와 같이 구성되어 있습니다.

A. samplename~.fasta

Barcode trim 한 read의 sequence의 내용을 포함하는 파일입니다.

Primer 시퀀스가 제공되었을 경우 primer 시퀀스도 함께 제거됩니다.

B. samplename~.fasta.qual

Barcode trim 한 read의 quality의 내용을 포함하는 파일입니다.

C. ~summary.xlsx

Barcode sorting,trim 후의 read의 상태를 시트 한 장에 정리하여 담은 파일입니다.

샘플별로 분류된 리드 개수와 길이, 평균 길이와 같은 정보를 포함합니다.

D. samplename~RDP.xlsx

미생물 동정 결과를 나타내는 엑셀 파일로 총 두 장의 시트로 구성되어 있습니다.

a. ~Summary_Genus sheet: Accession 별 hit 정보를 정리하였습니다.

b. ~Alignment sheet: alignment, taxonomic assignment 결과를 담고 있습니다.

<column 설명 : Sample_Summary 시트>

Accession	Description	Query#	Query length	Hit	Nohit	Total Query

Accession	RDP ID 정보입니다.
Description	Query의 organism 정보를 나타냅니다.
Query#	몇 개의 query가 해당 Accession ID에 매치됐는지 보여줍니다.
Query length	매치된 query의 길이를 나타냅니다.
Hit	Alignment를 통해 Hit 정보를 찾은 query 개수입니다
Nohit	Alignment를 통해 Hit 정보를 찾지 못한 query 개수입니다.

<column 설명 : Sample_Alignment 시트>

Query					Subject					Score		Identities		Gaps		Strand		Taxonomy						
Name	Length	Start	End	Coverage	Accession	Length	Start	End	Coverage	Bit	E-Value	Match/Total	Pct.(%)	Match/Total	Pct.(%)	Query	Subject	Organism	superkingdome	phylum	class	order	family	genus

Query Name	Query sequence를 구분하는 ID입니다.
Query Length	Query sequence의 길이를 나타냅니다.
Query Start	Query가 Subject(DB)에 매치되는 start position을 의미합니다.
Query End	Query가 Subject(DB)에 매치되는 end position을 의미합니다.
Query Coverage	전체 Query 중 Subject(DB)에 매치된 부분을 백분율로 나타냅니다.
Subject Accession	Query가 매치된 Subject(DB)의 Accession ID입니다.
Subject Length	Subject의 길이를 의미합니다.
Subject Start	Query가 매치된 Subject의 start position을 의미합니다.
Subject End	Query가 매치된 Subject의 end position을 의미합니다.
Subject Coverage	전체 Subject 중 Query와 매치되는 부분을 백분율로 나타냅니다.
Score Bit	Query와 Subject의 유사성의 정도를 나타내는 raw score를 정규화시킨 값입니다.
Score E-Value	우연히(진화적 상관성이 없어도) 서로 매치될 수 있는 확률을 의미합니다.
Identities Match/Total	매치되는 전체 서열 수를 의미하는 것으로, variation을 포함합니다.
Identities Pct.(%)	매치되는 전체 서열 중에서 정확히 매치되는 서열 수를 비율로 나타낸 수치로, 단순 일치성을 의미합니다.
Taxonomy Organism	Query의 organism 정보를 나타냅니다.
Taxonomy superkingdom	Query의 organism 정보를 단계별로 분류한 것으로 계의 분류에 해당합니다.
Taxonomy phylum	Query의 organism 정보를 단계별로 분류한 것으로 문의 분류에 해당합니다.
Taxonomy class	Query의 organism 정보를 단계별로 분류한 것으로 강의 분류에 해당합니다.
Taxonomy order	Query의 organism 정보를 단계별로 분류한 것으로 목의 분류에 해당합니다.
Taxonomy family	Query의 organism 정보를 단계별로 분류한 것으로 과의 분류에 해당합니다.
Taxonomy genus	Query의 organism 정보를 단계별로 분류한 것으로 속의 분류에 해당합니다.

E. ~OTUAnalysis.xlsx

Sequence similarity를 기반으로 하여 측정한 종의 수 (OTU-Operational Taxonomic Unit)의 분석결과를 담고 있는 엑셀 파일로 총 세 개의 시트로 구성되어 있습니다.

Similarity로 나눌 때 보통 Species는 97%, genus 94%, family 90%, order 85%, class 80%, phylum 75%의 기준을 사용합니다.

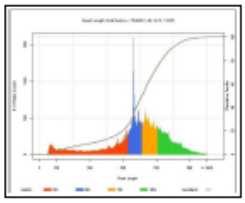
a. Similarity_97 sheet : Species 레벨에서의 OTU 개수와 종 분포도(shannon, simpson index 이용) 정보를 가지고 있습니다.

b. Similarity_94 sheet: Genus 레벨에서의 OTU 정보를 가지고 있습니다.

c. Similarity_90 sheet: Family 레벨에서의 OTU 정보를 가지고 있습니다.

<column 설명 : Similarity 시트>	
Sample	Sample 이름을 의미합니다.
RawData	해당 샘플로 분류된 리드 개수를 의미합니다.
OTU	분석결과 생성된 OTU 개수를 의미합니다.
Shannon	종이 얼마나 풍부하고 각 종에 속하는 개체수가 얼마나 고르게 분포하는가를 나타내는 척도인 다양성 지수입니다.
Shannon_lci	Shannon index에 대한 95% 신뢰구간의 최소값을 의미합니다.
Shannon_hci	Shannon index에 대한 95% 신뢰구간의 최대값을 의미합니다.
Simpson	종 분포가 얼마나 고른가를 알 수 있는 척도로 종 개체수가에 대한 각 종이 차지하는 비율을 나타내는 우점도 지수입니다.
Simpson_lci	Simpson index에 대한 95% 신뢰구간의 최소값을 의미합니다.
Simpson_hci	Simpson index에 대한 95% 신뢰구간의 최대값을 의미합니다.

2. 그래프 및 결과 폴더 내 파일

A . Raw data 결과 그래프	
	리드 개수, 전체 bases, 평균 리드 길이를 테이블로 제공하고 길이 분포도를 그래프로 표현합니다. 그래프의 가로축은 리드 길이이고, 세로축은 리드 개수, 리드 분포의 각 색상은 4분위수를 나타내며 갈색 곡선은 리드 개수의 누적값을 의미합니다.

B. Analysis 결과 그래프

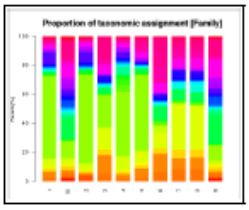
a. Reads status

MID 시퀀스를 이용하여 샘플별로 분류된 결과를 테이블로 정리하고 리드 상태를 파이 그래프로 표현합니다.

정상적으로 분류되는 sorted, 양쪽에 서로 다른 MID 가 붙어 분류가 안되는 ERR, 양 끝에 MID 정보가 없어 분류할 수 없는 경우가 NA, MID와 Primer를 trim하고 나면 남는 base가 전혀 없는 LowQual이 있습니다.

b. Reads count

가로축은 샘플명, 세로축은 리드 개수를 의미하는 것으로 각 샘플별로 몇 개의 리드가 분류되었는지를 표현하였습니다.



c. Taxonomic assignment

샘플간 abundance (각각의 종을 구성하는 개체수를 고려한 종의 풍부도) 차이를 superkingdom에서 genus까지 레벨별로 비교할 수 있는 그래프입니다. 가로축은 샘플명이고, 세로축은 리드 개수의 %를 나타내는 것으로 샘플별로 존재하는 종의 차이 및 개체수의 차이를 비교할 수 있습니다.