**EgGutPro NGS 알고리즘**

----------------------------------EgGutPro NGS MIrROR 분석 파이프라인-------------------------------

1. **새로운 Sample sequencing data Input**
2. **Basecalling/Read correction**
3. **De-multiplexing/Adapter trimming**
4. **MIrROR process**
   1. Long read mapping
   2. Classification (Read count)
   3. Output = Mapping output file(paf 파일), Read count classification file(.txt, .std, mpa 파일)
5. **질환 별 MRS 지수 도출을 위한 taxonomy proportion 계산**

---------------------------------질환 별 MRS 계산 논리 및 프로세스-------------------------------

**참고:** [https://github.com/Gyungbu/EgGutPro](https://github.com/Gyungbu/EgGutProt)

**EGgutPro\_percentile\_rank.py 파일**

1. **파일 load**
   1. /input/EGgutPro\_mircobe\_list.xlsx – 유익균, 유해균 리스트, 프로바이오 리스트, 각 질환 별 미생물 리스트가 저장된 파일
   2. /input/EGgutPro\_healthy\_person\_profile\_v2.xlsx – 건강한 샘플들의 proportion 이 merge 된 파일
   3. /input/EGgutPro\_db\_abundance.xlsx – 1144개 샘플에 대한 Relative Abundance 값이 저장된 파일
   4. /input/EGgutPro\_mrs\_db.xlsx – 1144개 샘플에 대한 각 질환 별 MRS(Microbiome Risk Score), DysbiosisHarmful, DysbiosisBeneficial, HealthyDistance 값이 저장된 파일
   5. /input/EGgutPro\_percentile\_rank\_db.csv – 1144개 샘플에 대한 각 질환 별 MRS(Microbiome Risk Score), DysbiosisHarmful, DysbiosisBeneficial, HealthyDistance, Diversity 값의 백분위 값이 저장된 파일
2. **MRS 계산**

: Microbiome Risk Score

: Relative abundance of microbiome associated with specific diseases

: The number of microbiome associated with a specific disease

: Health sign +1 for 유해/증가, -1 for 유익/감소

1. **Dysbiosis 계산**

: Relative abundance of harmful microbiome

: Relative abundance of beneficial microbiome

: The number of harmful microbiome

: The number of beneficial microbiome

1. **Healthy Distance 계산**

: Healthy Distance

: Relative abundance vector of healthy sample.

: Relative abundance vector of sample.

: The number of healthy samples.

: Centered log ratio transformation.

1. **최종 Percentile Rank 계산**

앞서 구한 샘플의 MRS, DysbiosisHarmful, DysbiosisBeneficial, Healthy Distance, Diversity를 기존 /input/EGgutPro\_mrs\_db.xlsx 파일을 기준으로 Percentile Rank 를 계산

scipy.stats.percentileofscore를 사용.

* 1. 백분위 값이 95 이상의 outliers 95로 처리, 5 이하의 outliers 5로 처리
  2. Main category의 백분위 값은 세부 질환 백분위 점수의 평균으로 계산

1. **평가**
   1. 각 질환 별 , , , 값을 기준으로 [95,100]: ‘좋음’, (80,95]: ‘보통’, (50,80]: ‘주의’, (20,50]: ‘나쁨’, [0,20]: ‘아주 나쁨’의 5등급으로 나눔.
   2. 각 질환 별 , 값을 기준으로 [80,100]: ‘아주 나쁨’, [50,80]: ‘나쁨’, [20,50]: ‘주의’, (5,20): ‘보통’, [0,5]: ‘좋음’의 5등급으로 나눔.
2. **유익균/유해균 total abundance 계산**
   1. /input/EGgutPro\_mircobe\_list.xlsx – 유익균, 유해균 리스트를 기준으로 샘플의 유익균, 유해균 abundance[%] 값을 계산

: Beneficial bacteria ratio

: Relative abundance of beneficial microbiome

: The number of beneficial microbiome

: Harmful bacteria ratio

: Relative abundance of harmful microbiome

: The number of harmful microbiome

* 1. 1144 개 샘플 데이터를 바탕으로 유익균, 유해균의 평균 abundance[%] 값을 계산
* num\_total\_species: 해당 샘플 proportion 파일의 observed 값
* harmful\_mean\_abundance[%]: 1144 개 샘플의 평균값
* beneficial\_mean\_abundance [%]: 1144 개 샘플의 평균값

1. **유해균 개별 abundance 계산**
   1. /input/EGgutPro\_mircobe\_list.xlsx – 유해균 리스트를 기준으로 샘플의 개별 유해균의 abundance 값을 계산
   2. /input/EGgutPro\_mircobe\_list.xlsx – 유해균 리스트를 기준으로 1144개 샘플의 개별 유해균에 대한 평균 abundance 값을 계산
   3. 유해균 평균 abundance값을 기준으로 내림차순 정렬
   4. EGgutPro\_harmful.csv : 유해균 Top 10 만 표시, EGgutPro\_harmful\_tot.csv : 모든 유해균 리스트 표시
2. **유익균 개별 abundance 계산**
   1. /input/EGgutPro\_mircobe\_list.xlsx – 유익균 리스트를 기준으로 샘플의 개별 유익균의 abundance 값을 계산
   2. /input/EGgutPro\_mircobe\_list.xlsx – 유익균 리스트를 기준으로 1144개 샘플의 개별 유익균에 대한 평균 abundance 값을 계산
   3. 유익균 평균 abundance값을 기준으로 내림차순 정렬
   4. EGgutPro\_beneficial.csv: 모든 유익균 리스트 표시
3. **프로바이오 total 및 개별 abundance 계산**
   1. /input/EGgutPro\_mircobe\_list.xlsx – 프로바이오 리스트를 기준으로 샘플의 total 프로바이오 abundance 값을 계산

: Total probio ratio

: Relative abundance of probio

: The number of probio

* 1. probio\_abundance\_mean[%]: 1144 개 샘플의 평균값
  2. observed\_mean : 1144 개 샘플의 num\_total\_species 평균값
  3. /input/EGgutPro\_mircobe\_list.xlsx – 프로바이오 리스트를 기준으로 샘플의 개별 프로바이오 abundance 값을 계산
  4. 샘플의 probio abundance값을 기준으로 내림차순 정렬
  5. EGgutPro\_probio\_tot.csv: 모든 프로바이오 리스트 표시