GDC lung (2024.08.07)

1. 파일 다운로드

1. GDC 포털에서 TCGA-LUAD 프로젝트 데이터만 다운로드

- 샘플 수가 많아서 총 142개만 선택
- 총 142개: Normal 59 / oct 샘플 83개

tcga 전처리

데이터

각 폴더에 gene counts 만 다운로드하여 추출하여 정리

- gene name과 FPKM 값만 사용
- data 폴더 참고

샘플정보

sample_sheet 참고

gdc_manifest.2024-08-06.txt 참고

분석

TCGA 데이터

데이터 전처리

- 낮은 값 가진 유전자 제거

significant

P(pvalue), logFC(log2 fold change)를 유전자별로 확인 가능

- P(pvalue): t-test 결과
- Fold change: 발현 차이
- 선별 기준은 logFC 2배 pvalue 0.05이하

- significant.csv 참고

Volcano plot

- P(pvalue), logFC(log2 fold change)에 선별된 유전자 확인 가능
- Volcano_plot.pdf 참고

유저 데이터

- 유저 데이터는 1개의 샘플 비교라 단순 FC 값만 가지고 분석
- User_sig.csv 참고

벤다이어그램

- up, down 유전자를 TCGA 데이터와 유저데이터를 비교하고자 하였음
- Venn Diagram.pdf 참고