clinVar EDA

CLINVAR VCF EDA

Table of Contents

Introduction

clinVar

NCBI의 clinVar 데이터베이스는 유전적 변이와 인간 질병 간의 관계를 파악하는 데 필수적인 자원이다.

하지만 데이터가 방대한 VCF(Variant Call Format) 파일 형태로 제공되어 직관적인 파악이 어렵고, 주기적인 업데이트로 인해 최신 경향성을 분석할 필요가 있다.

clinVar의 VCF파일을 분석하기 전에 INFO 컬럼을 딕셔너리로 파싱한 다음, 데이터프레임으로 변환하고 csv파일로 저장할 것이다.

Project information

Data resource: clinVar

Libraries

- Numpy
- Polars
- Plotly
- gzip

Visualiation tool

- Tableau

Project Information

1 → Data acquisition

2 → Data transformation

3 → Preprocessing

4 → EDA

5 → Tableau dashboard

Data acquisition&Transformation

1

Data acquisition

clinVar VCF파일 입수

GRCh37, 20260404

2

Data Transformation

VCF파일을 CSV파일로 변환

INFO칼럼을 딕셔너리로 파싱한 다음 내장된 데이터를 데이터프레임으로 변환

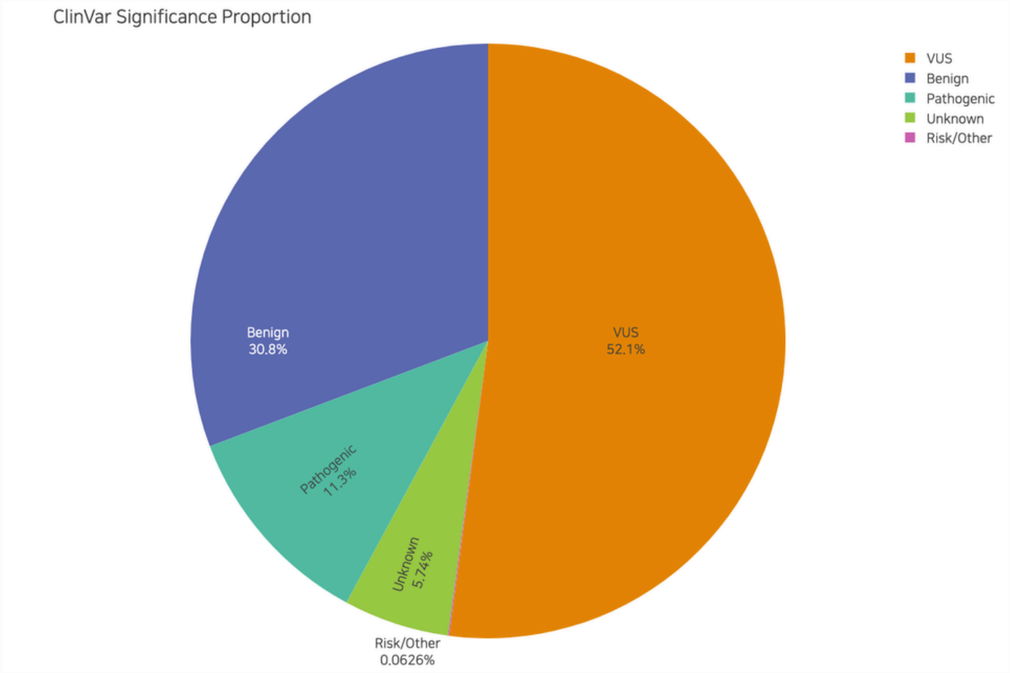
1. 결측값 처리
2. 사용할 칼럼만 정제
3. CLNSIG 범주화
4. 염색체 범주화(상염색체, 성염색체, 미토콘드리아)
5. Tableau로 시각화하기 위한 CSV파일로 저장

EDA: CLNSIG Group

CLNCIG Group

각 변이의 임상적 의미 비율

1. 전체 4,150,688개의 변이 중 VUS의 비율은 52.1%, Benign의 비율은 30.8%로 82.9%를 차지함.
2. Pathogenic한 변이의 비율은 약 11.3%에 불과함.



CLNSIG 그룹 비율

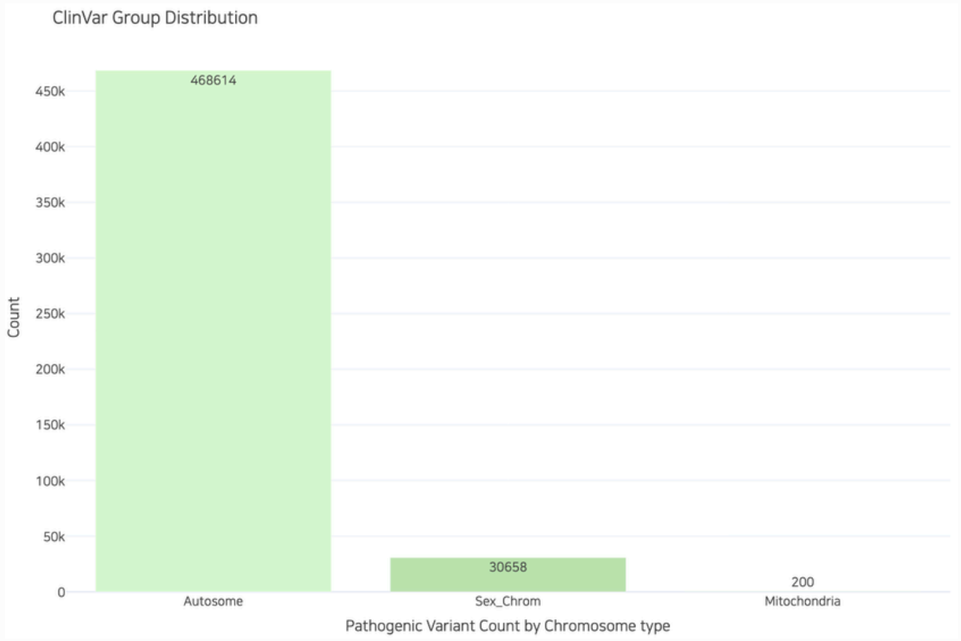
VUS: 임상적 의미 불명/Benign: 무해함/Pathogenic: 병원성

EDA: Pathogenic

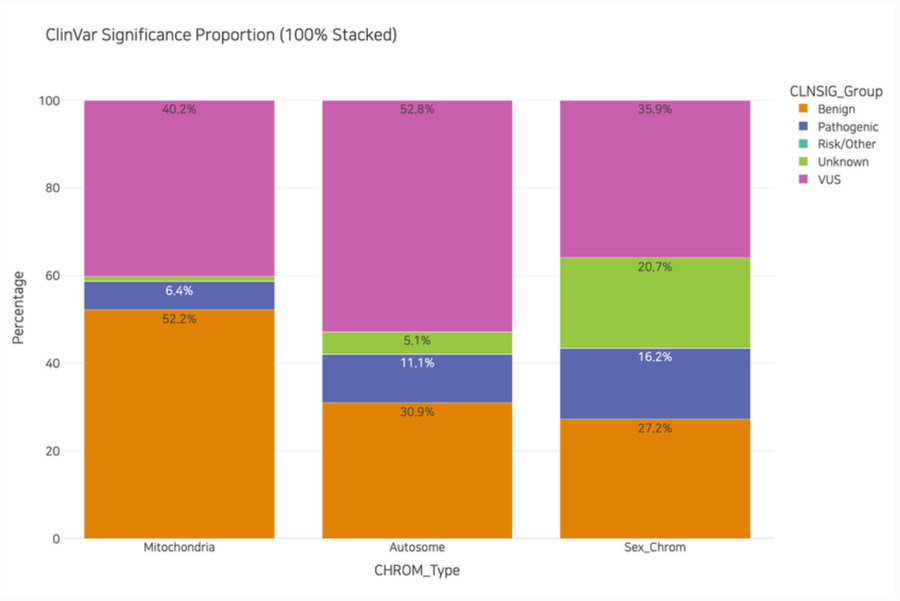
Pathogenic ratio

염색체별 Pathogenic 변이 비율

1. 염색체 그룹별로 Pathogenic한 변이의 분포를 확인해 본 결과, **상염색체가 가장 많았음.**
2. 염색체 그룹별로 전체 비율을 확인해 본 결과, **성염색체**에서 Pathogenic의 비율이 **16.2%로 가장 높았음.**



염색체 그룹별 Pathogenic 변이 수



염색체 그룹별 분포

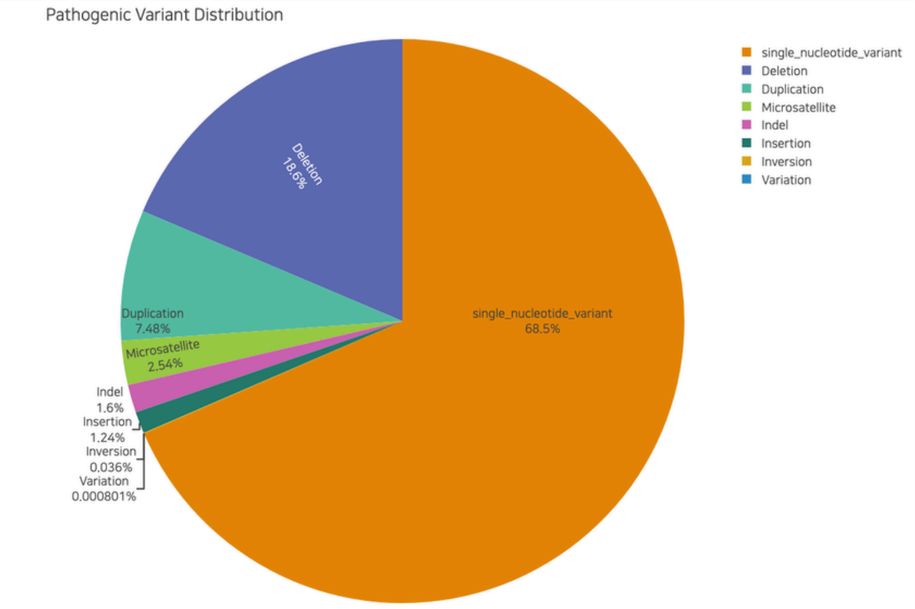
Autosome: 상염색체/Sex chromosome: 성염색체/Mitochondria: 미토콘드리아

EDA: Pathogenic

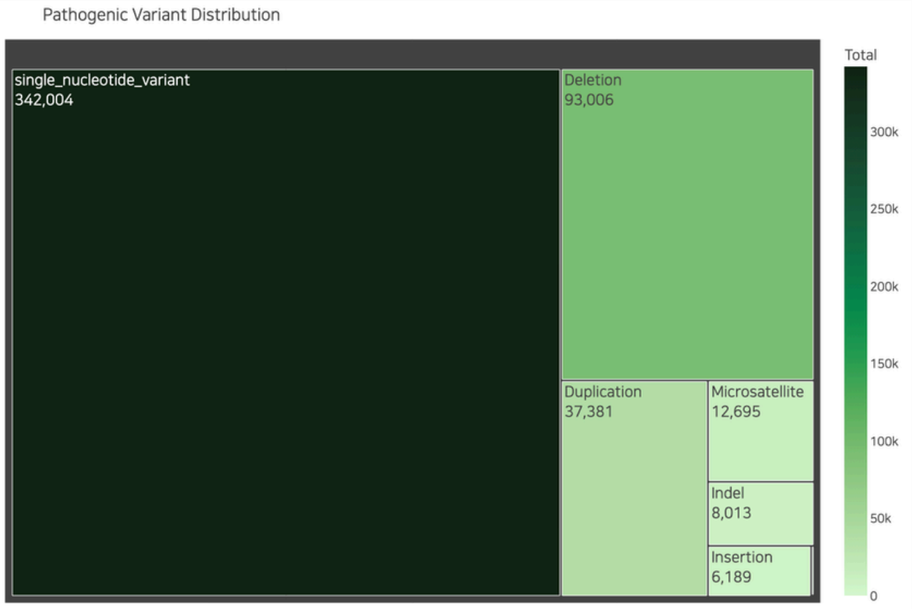
CLNVC

Pathogenic한 변이 내 CLNVC 유형

1. Pathogenic한 변이 중에서도 가장 많은 비중을 차지하는 변이 유형은 SNV(single nucleotide variant)로, 342,004건에 달함. (전체 중 68.5%)
2. 두번째로 많은 변이 유형인 Deletion은 93,006건에 달하고, 전체 중 18.6%를 차지함.



Pathogenic한 변이 내 CLNVC 분포



염색체별 변이 수(전체 CLNVC 합계)

Single nucleotide variant: 단일염기변이/Deletion: 삭제/Duplication: 중복/insertion: 삽입
Indel: 삽입+삭제/Inversion: 역위/Microsatellite: 마이크로새틀라이트(짧은 염기가 반복되는 변이)

EDA: TOP 3 Gene

TOP 3 Gene

Pathogenic한 변이가 가장 많은 유전자

1. 유형별로 세분화하지 않고 합산했을 때 Pathogenic한 변이가 가장 많은 것은 BRCA2, TTN, BRCA1이었음.
2. 돌연변이 유형별로 세분화한 후 집계했을때도 BRCA2, TTN, BRCA1이 가장 많았음.

Top 3 Mutated Genes Highlighted

Chromosome	Top Gene Symbol	Variant Count
13	BRCA2	11,249
2	TTN	9,859
17	BRCA1	6,963
11	ATM	5,211
15	FBN1	4,254
5	APC	3,307
16	TSC2	2,988
X	DMD	2,826
1	USH2A	2,556
3	MLH1	2,407

각 염색체별 변이수가 많은 유전자들

Top 3 Mutated Genes Highlighted

Chromosome	CLNVC	Top Gene Symbol	Variant Count
13	single_nucleotide_variant	BRCA2	6,891
2	single_nucleotide_variant	TTN	6,206
17	single_nucleotide_variant	BRCA1	3,864
15	single_nucleotide_variant	FBN1	3,032
11	single_nucleotide_variant	ATM	2,741
16	single_nucleotide_variant	TSC2	2,201
X	single_nucleotide_variant	DMD	1,880
1	single_nucleotide_variant	USH2A	1,648
5	single_nucleotide_variant	APC	1,482
19	single_nucleotide_variant	LDLR	1,388

각 염색체별 변이수가 많은 유전자들 (변이 유형별 집계)

이미지가 세로로 길어서 일부만 첨부했습니다.

Tableau Dashboard

각 염색체별 비율, 그리고 유전자별 비율

1. Tableau 대시보드를 통해 각 염색체별로 **변이 수가 가장 많은 상위 25개 유전자 및 임상적 의미, 변이 유형 비율**을 시각화
2. 2페이지에서는 **각 염색체별 Pathogenic한 변이**들에 대한 비율을 따로 확인할 수 있음
3. 유전자를 검색해서 **해당 유전자에 대한 변이 분포**를 확인할 수 있음

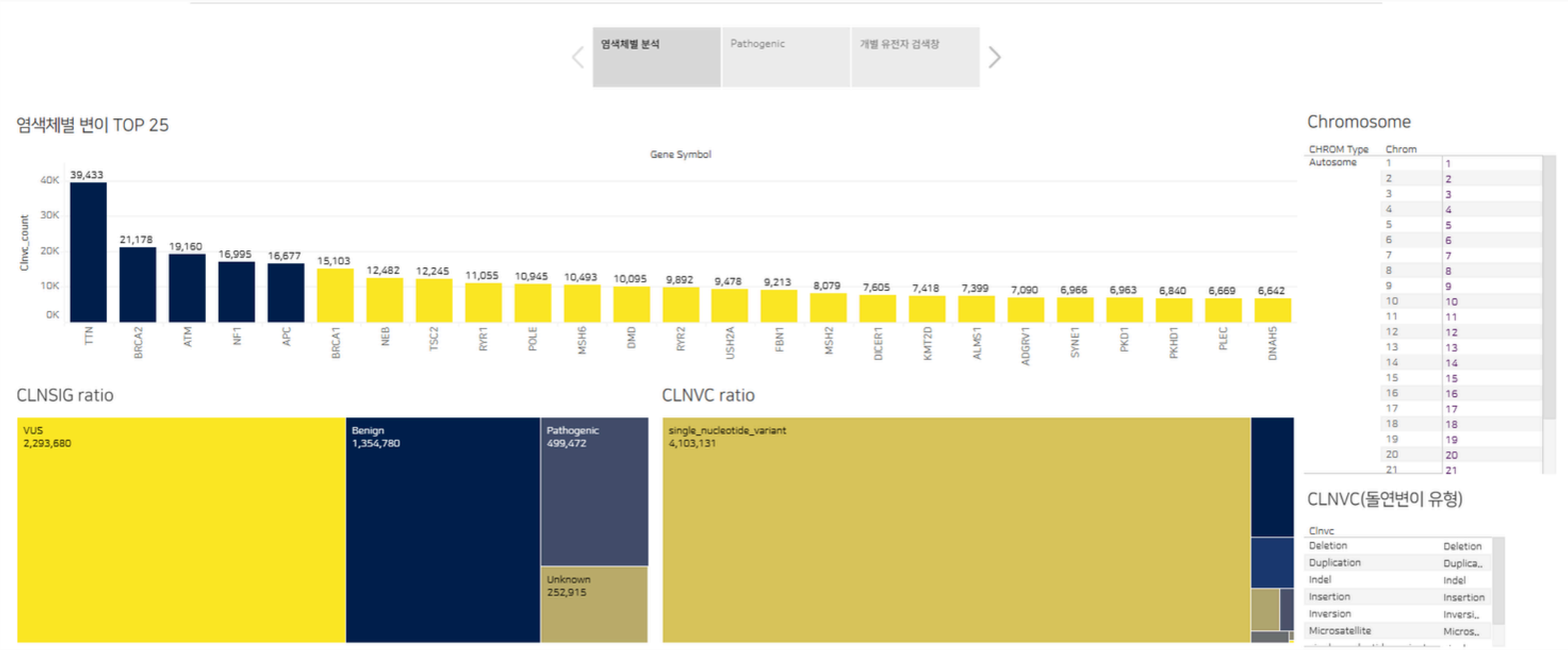
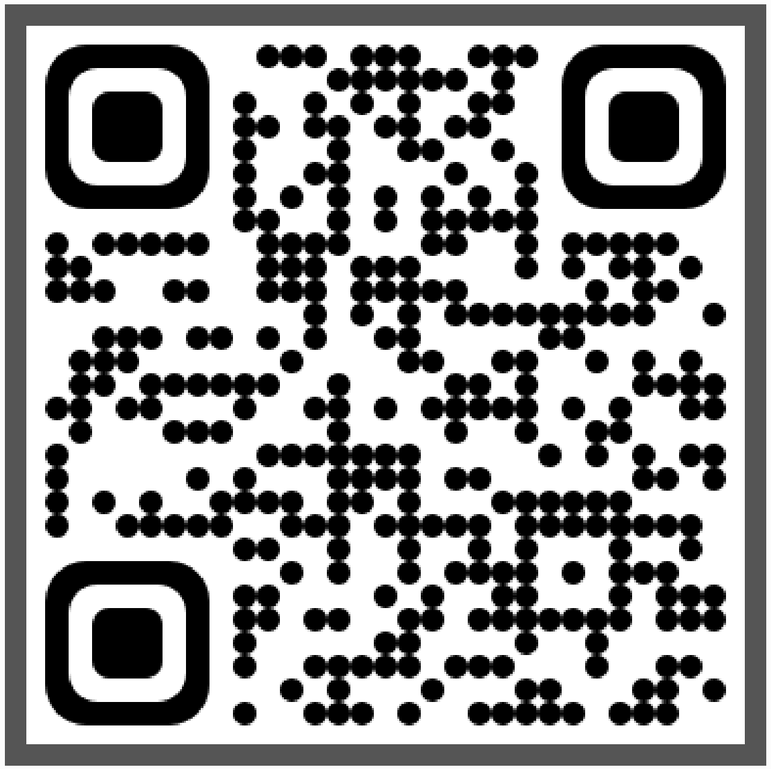


Tableau dashboard (염색체별 분석 페이지)

Contact



Mail

pokemonms@naver.com
blackholekun@gmail.com

Cellular

+82-10-5027-0328

Github

https://github.com/koreanraichu

Blog

https://koreanraichu.tistory.com/