Project Report

(Benchmarking of Variant Annotation Tools)

202202358 이나림, 202202665 이은서

1. Motivation and contributions

변이(variant) 분석은 유전 질환 연구에 있어 핵심적인 역할을 하며, 질병과 관련된 위험 변이를 식별함으로써 환자의 진단 및 치료 계획을 최적화할 수 있다. 이번 프로젝트에서는 환자군 데이터를 기반으로 변이를 분석하고, 이를 통해 환자 개개인이 보유한 위험 변이를 평가하였다. 이를 위해 SIFT, CADD, ANNOVAR 와 같은 유전적 변이 분석 도구를 활용하였으며, 각 환자별로 어떤 위험 변이를 가지고 있는지 확인하였다. 분석의 초점은 주로 해로운(deleterious) 변이에 맞추었으며, SIFT, CADD, ANNOVAR 를 통해 제공된 점수를 비교하였다. 특히, 각 도구에서 가장 낮은 점수를 보이는 변이가 어떤 변이인지 식별하는 데 중점을 두었다.

2. Methods

SIFT, CADD, ANNOVAR tool 을 이용해 환자군의 VCF 파일 필터링을 진행하였다. SIFT (Sorting Intolerant From Tolerant)는 단백질의 아미노산 서열에서 변이가 기능에 미치는 영향을 예측하는 도구이고, sequence homology 를 기반으로 amino acid 서열의 변이를 sort 해주는 도구이다. CADD(Combined Annotation Dependent Depletion)는 human genome 에서 single nucleotide variants(SNVs)가 유전자 기능에 미칠 수 있는 유해성을 평가하는 도구이다. 이것은 SIFT 뿐만 아니라 polyphen-2 점수도 확인이 가능하다. 마지막으로 ANNOVAR(ANNOtate VARiation)는 유전적 변이에 주석을 달기 위해 사용되는 소프트웨어 도구로써, Annotation(변이 파일에 상세 정보를 첨가하는 작업) tool 중 하나이다. 이것은 자체 input 방식 사용한다.

따라서 CADD 는 전체적인 중요성을 파악하는데 초점을 두어 전체적인 변이를 분석하는데 도움이 되고, sift 는 단백질 기능에 미치는 영향 예측에 특화된 툴로써 코딩영역만 분석가능한 툴이고, ANNOVAR 는 변이의 포괄적인 주석 및 기능적 평가에 초점을 두어 통합데이터를 제공한다는 장점이 있다.

Sift 와 ANNOVAR 사용 시, 사용한 데이터베이스는 다음과 같다.

먼저, sift 에서는 돌연변이가 있는 위치에서 단백질 서열의 진화적 보존성을 평가하기 위해 생성한 다중 서열 정렬(MSA: Multiple Sequence Alignment) data, 인간을 포함한 다양한 종의 표준 참조 유전체 및 단백질 서열을 제공하는 RefSeq, 돌연변이에 대한 정보를 포함해 주는 dbSNP (Single Nucleotide Polymorphism Database)를 사용하였다.

ANNOVAR 에서는 refGene, cytoband, exac03, avsnp147, dbnsfp30a 의 5 가지 데이터베이스를 사용하여 rsID 및 각 툴에 대한 정보를 제공하고 있다. 이러한 다양한 tool 들을 설치 후에 변이 분석을 진행하였다.

3. Approach and methodology

1) VCF 파일 수집 및 정리: 10 명의 환자로부터 받은 VCF 파일을 통해 변이 정보를 확인하였다. 각 변이의 염기서열 위치, 변이 유형, 변이 염기 정보 등의 필요한 정보만 추출하여 정리하였다.

2) 변이 정보 추출 후 분석 진행: 기존의 vcf 파일에서 변이정보만 추출하여 각 tool 에 대한 sift score 를 확인하였다. 기존 데이터베이스에서 보고된 변이 여부 등을 추가로 확보하여 각 사람 별 sift score 가 가장 낮은 변이의 이름을 확인한 후, 그에 대한 정보를 파악하였다.

3) 결과 비교 및 통합: SIFT, CADD, ANNOVAR 의 결과를 종합하여 각 도구가 변이를 어떻게 평가하는지 비교한다. 동일한 변이에 대해 서로 다른 평가를 내릴 수 있으므로, 각 도구의 결과를 상호 보완적으로 활용하여 각 사람별 deleterious 한 변이를 파악하였다.

4. Evaluation and result

데이터 세트

: 본 프로젝트에서 사용한 데이터셋은 환자 00 – 09의 raw data이다. 00번 환자는 SIFT, CADD, ANNOVAR 분석 tool, 01 – 09번 환자는 ANNOVAR, SHIFT 분석 도구를 사용하였다. 각 분석 기법

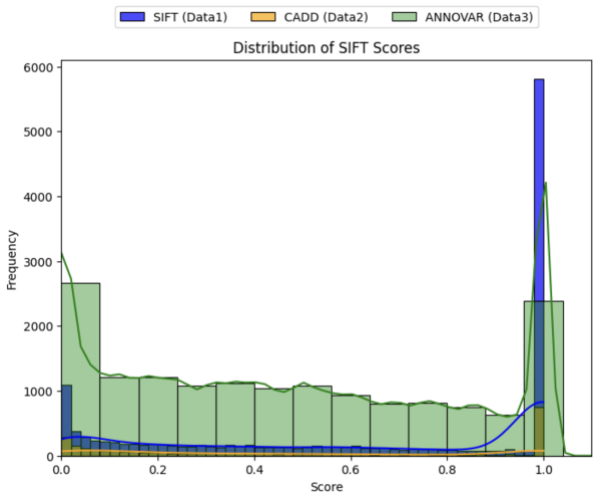
결과에서 추출해낸 SIFT 점수를 활용하였다. SIFT 점수는 낮을수록 단백질 기능에 유해한 변이를 나타내며, 0.05 이하는 deleterious로 본다.

결과 및 평가

세 가지 주요 분석 도구 SIFT, CADD, ANNOVAR 에서 추출된 SIFT 점수를 기준으로 사용하였고, 같은 데이터에 대하여 여러 분석 기법의 결과를 비교해보며 각 분석 기법의 특징을 분석하였다. 대표로 00, 01 환자의 결과를 활용해 각 분석 기법의 차이를 자세히 살펴보았다.

1) SIFT, CADD, ANNOVAR

00 번 환자에 대해 SIFT, CADD, ANNOVAR 세가지 분석 툴을 사용해 얻은 SIFT 점수 분포를 비교한 결과이다.



1) 낮은 점수 (유해한 변이) 구간 : Data1 과 Data3 는 0.0 ~ 0.05 구간에서 높은 빈도를 나타낸다. 그러나 Data2 는 동일 구간에서 상대적으로 낮은 빈도를 보인다.

2) 높은 점수 (무해한 변이) 구간 : Data3 은 0.8 ~ 1.0 구간에서 또 한번의 피크를 보인다. 이는 ANNOVAR 데이터에서 무해한 변이가 높은 비율로 나타냈음을 알 수 있다. Data1 과 Data2 는 높은 점수 구간에서 상대적으로 낮은 빈도를 보인다.

3) 점수 분포 패턴 차이: Data1 의 분포는 상대적으로 유해한 변이 구간 (0.0 ~ 0.05)에 집중되어 있다. Data2 의 분포는 다른 데이터에 비해 고르게 분포되어 있으며 낮은 빈도를 유지한다. Data3 의 분포가 가장 넓으며 유해한 변이, 무해한 변이 구간 모두에서 높은 빈도를 보인다.

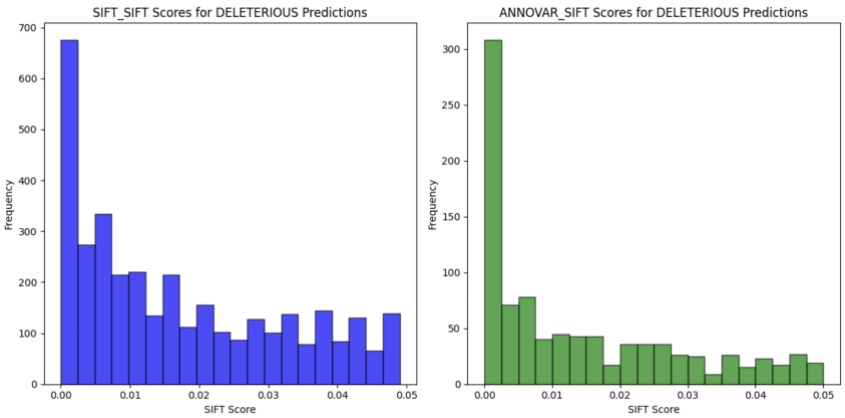
SIFT 와 ANNOVER 의 일치성 : Data1(SIFT)와 Data3(ANNOVAR)은 유해한 변이 구간(0.0 ~ 0.05)에서 유사한 패턴을 보이지만 ANNOVAR 는 무해한 변이(0.8 ~ 1.0)에서도 빈도가 높아 SIFT 점수와 분포 차이를 나타낸다.

CADD 에서의 SIFT 점수 : Data2(CADD)는 다른 두 데이터에 비해 유해한 변이 빈도가 낮고, 분포가 고르게 퍼져 있어 다른 데이터 세트와의 결과 차이를 보인다. 이는 CADD 에서 변이를 해석하고 점수를 추출하는 방식이 SIFT 및 ANNOVER 와 다를 수 있음을 나타낸다.

2) SIFT, ANNOVAR

아래는 01 번 환자에 대해 SIFT, ANNOVAR 두가지 분석 툴을 사용해 얻은 SIFT 점수 분포를 비교한 결과이다.

2-1) SIFT 와 ANNOVAR_SIFT 의 변이 분포를 시각화한 히스토그램



1) 점수 0.00 의 빈도 : SIFT 는 약 700 개 이상의 변이가 해당 점수에 집중되고, ANNOVAR_SIFT 는 약 300 개의 변이가 점수 0.00 에 집중된다. SIFT 데이터에서 점수 0.00 에 변이가 더 많이 집중됨을 확인할 수 있다.

2) 점수 0.01~0.05 구간의 분포 : SIFT 는 점수가 증가할수록 빈도가 서서히 감소하는 경향을 보이고 0.01~0.05 사이에서도 비교적 고른 분포를 가진다. ANNOVAR_SIFT 는 점수가 증가할수록 급격히 빈도가 감소하며, 0.01~0.05 구간에서 변이의 개수가 훨씬 적다.

3) 분포의 형태 차이 : SIFT 데이터는 점수 0.00 외에도 0.01~0.05 구간에서 변이가 비교적 고르게 분포한다. ANNOVAR_SIFT 데이터는 점수 0.00 에서 대부분의 변이가 집중되고, 나머지 구간에서는 급격한 감소를 보인다.

SIFT, ANNOVAR_SIFT 데이터 비교

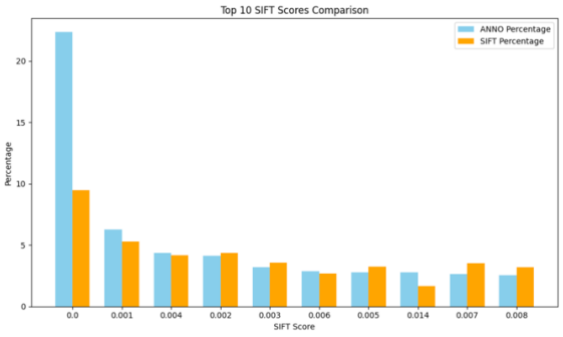
- SIFT 데이터 : 점수 0.00 에서 매우 높은 빈도를 보이지만, 다른 구간에서도 변이가 꾸준히 나타나고, 이를 통해 SIFT 점수 계산 방식에서 "deleterious" 변이를 더 다양하게 평가했을 가능성을 시사한다.

- ANNOVAR_SIFT 데이터 : 점수 0.00 에서 대부분의 변이가 집중되며, 나머지 구간에서는 변이 빈도가 급격히 감소한다. ANNOVAR 가 "deleterious" 변이를 더 엄격하게 정의했거나, 점수 0.00 에 더 높은 중요성을 부여했을 가능성이 있다.

2-2) SIFT 와 ANNOVAR_SIFT 의 변이 비율

상위 10개 SIFT_score 비교:

SIFT_score	count_ANNO	ANNO_percentage	count_SIFT	SIFT_percentage
0.000	210	22.340426	335.0	9.495465
0.001	59	6.276596	186.0	5.272109
0.004	41	4.361702	147.0	4.166667
0.002	39	4.148936	154.0	4.365079
0.003	30	3.191489	126.0	3.571429
0.006	27	2.872340	94.0	2.664399
0.005	26	2.765957	115.0	3.259637
0.014	26	2.765957	58.0	1.643991
0.007	25	2.659574	125.0	3.543084
0.008	24	2.553191	112.0	3.174603



1) 점수 0.000 : ANNO 데이터와 SIFT 데이터 모두에서 가장 큰 비율을 차지하지만, SIFT 데이터에서는 상대적으로 낮은 비율을 나타낸다.

- ANNO 데이터 : 변이 210 개(22.34%) / SIFT 데이터 : 변이 335 개(9.50%)

2) 점수 0.001 : SIFT 데이터에서 이 점수 구간의 비율이 ANNO 데이터 대비 조금 낮다.

- ANNO 데이터 : 변이 59 개(6.77%) / SIFT 데이터 : 변이 186 개(5.77%)

3) 점수 0.002 ~ 0.008 : ANNO 와 SIFT 데이터 모두에서 이 점수 구간의 변이 비율은 점진적으로 감소하고, ANNO 데이터에서의 비율 감소폭이 SIFT 데이터 대비 작다.

ANNOVAR 데이터와 SIFT 데이터의 차이점 : ANNOVAR 데이터는 점수 0.000 에서 22.34%를 차지하며 가장 집중된 변이를 보이는 반면, SIFT 데이터는 9.50%로 점수 0.000 에 대한 집중도가 낮다. 점수 0.001 ~ 0.008 의 구간에서는 두 데이터 세트가 비슷한 경향을 보이지만 SIFT 데이터에서의 변이 비율이 ANNO 데이터에 비해 다소 낮게 나타난다.

일관된 경향 : 두 데이터 세트 모두 점수가 증가할수록 변이 개수와 비율이 감소하는 패턴을 보인다. 이는 낮은 SIFT 점수가 더 많은 변이를 포괄하고 높은 점수로 갈수록 변이가 희소해지는 일반적인 경향과 일치한다.

SIFT 점수 0.000 구간에서 ANNO 와 SIFT 데이터 간 큰 비율 차이가 존재하며, 이는 두 데이터 세트 간 변이의 분포 특성 차이를 반영할 가능성이 있다. 전체적으로 두 데이터 세트는 점수가 증가함에 따라 변이 비율이 감소하는 유사한 경향을 보인다.

각 분석 기법의 활용 방안

1. **SIFT**
   - 단백질 기능 손실과 밀접하게 연관되어 있어 기능 손실 가능성이 높은 변이를 탐지하는 데 매우 유용
   - 초기 필터링 단계에서 SIFT 도구의 결과를 우선적으로 사용하는 것이 효과적

2. **CADD 에서의 SIFT 점수**
   - 단백질 기능 유해성 뿐만 아니라, 생물학적 맥락과 중요성까지 고려된 결과로 해석 가능

- SIFT 결과에 대한 추가적인 통합적 맥락 분석이 필요할 때 활용

3. ANNOVAR 에서의 SIFT 점수

- 유해 변이와 무해 변이를 명확히 구분하는 데 적합하며 대규모 변이 데이터의 포괄적 필터링 및 주석 작업에 유리
- 최종적으로 결과를 정리하고 추가 분석을 위한 후보 변이를 선정하는 단계에서 효과적

결론

SIFT, CADD, ANNOVAR 도구에서 추출된 SIFT 점수는 유전체 변이 분석에서 상호 보완적인 정보를 제공한다. 단백질 기능 손실 가능성에 중점을 둔 SIFT 도구, 상대적 중요성을 평가하는 CADD, 그리고 다양한 데이터 통합 분석을 제공하는 ANNOVAR 은 각각의 강점이 뚜렷하다. 따라서 각 도구의 SIFT 점수를 조합적으로 활용하면 변이의 유해성과 생물학적 중요성을 보다 정확하게 평가하고, 연구 및 임상적 응용에서 최적의 결과를 얻을 수 있다.

5. Conclusion and future work

본 프로젝트는 SIFT, CADD, ANNOVAR 도구를 활용하여 환자군에서 주요 해로운 변이를 식별하였다. 이 분석은 환자별 어떠한 변이가 있을 수 있는지 확인이 가능하며 가지고 있는 변이를 이용해 진단과 치료 계획 수립에 기여를 할 수 있는 기초 데이터가 될 수 있다. 식별된 변이들은 나아가 맞춤형 치료 전략 개발에 중요한 정보를 제공할 수 있다. 이를 통해 정밀의학 발전에 기여할 수 있다. 본 프로젝트에서 식별된 해로운 변이를 통해 정밀의학 기반의 진단 및 치료 전략 수립에 중요한 자료로 활용될 수 있다. 이를 통해 맞춤형 치료 계획을 수립할 수 있다. 또한, 인공지능 및 머신러닝을 활용한 알고리즘을 통해, 변이 데이터를 학습하고, 학습된 데이터로 변이의 결과예측이 가능하도록 할 수 있는 시스템에도 도움이 될 수 있을 것이다.