Benchmarking of Variant Annotation Tools

202202358 이나림, 202202665 이은서

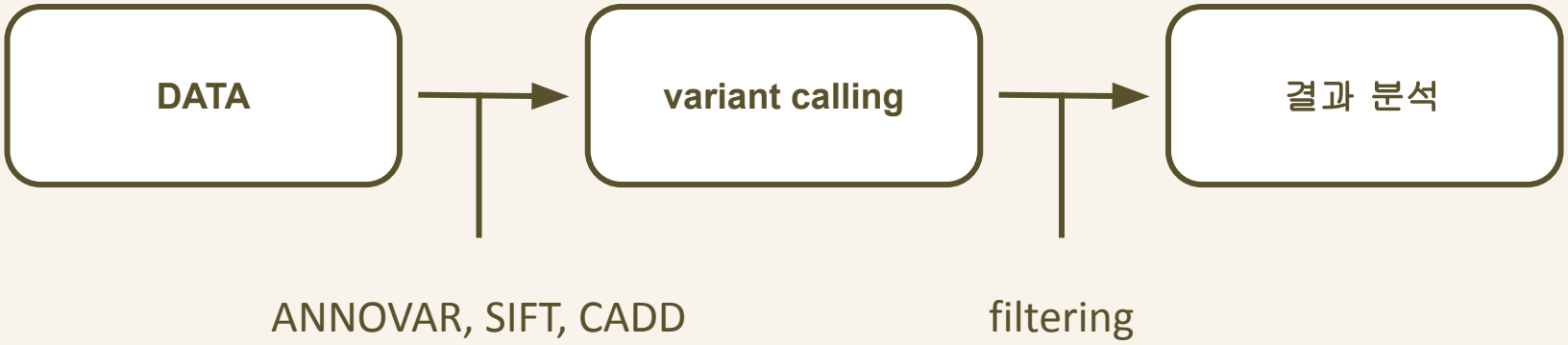
Contents

문 제

의생명 정보학 project

문제

Benchmarking of Variant Annotation Tools



접근 방식 및 방법론

의생명 정보학 project

접근 방식 및 방법론 _ SIFT

SIFT (Sorting Intolerant From Tolerant)

- 단백질의 아미노산 서열에서 변이가 기능에 미치는 영향 예측
- sequence homology를 기반으로 amino acid 서열의 변이를 sort해주는 도구
- Variation 분석에서 발견된 Nonsynonymous cSNP의 기능적 영향 평가 필요

- Nonsynonymous cSNP : 유전자의 코딩 영역 내에서 코딩된 단백질의 아미노산 서열을 변화시키는 유전적 변이체

접근 방식 및 방법론 _ SIFT

SIFT (Sorting Intolerant From Tolerant) SCORE

- 범위 : 0 ~ 1
- 낮은 점수일수록 변이가 단백질 기능에 유해할 가능성이 높다.
- **score 0.05 이하** : deleterious

  0.05 이상 : tolerated

접근 방식 및 방법론 _ CADD

CADD(Combined Annotation Dependent Depletion)

- human genome에서 single nucleotide variants(SNVs)가 유전자 기능에 미칠 수 있는 유해성을 평가하는 도구
- SIFT 뿐만 아니라 polyphen-2 점수도 확인 가능

접근 방식 및 방법론 _ ANNOVAR

ANNOVAR(ANNOtate VARiation)

- 유전적 변이에 주석을 달기 위해 사용되는 소프트웨어 도구
- Annotation(변이 파일에 상세 정보를 첨가하는 작업) tool 중 하나
- 자체 input 방식 사용 (각 필드가 tab으로 분리되어 있도록 하는 VCF 사용)

기존 논문에 따른 각각의 Tool의 차이 비교

특성	CADD	SIFT	ANNOVAR
초점	전체적인 중요성 파악	단백질 기능에 미치는 영향	변이의 포괄적 주석 및 기능적 평가
분석 범위	코딩 + 비코딩 변이	코딩 영역(비동의적 변이)	코딩 + 비코딩 변이
점수 체계	Phred score (1~99) 내부의 SIFT, polyphen-2 점수와 동일	SIFT 점수 (0~1, 낮을수록 유해)	데이터베이스 기반 점수 (CADD, SIFT 등)
사용 사례	질병 관련 변이 선별	단백질 기능 예측	종합적 변이 분석 및 주석
장점	비코딩 변이 포함, 머신러닝 기반 정밀성 전체 점수 포함 -> 비교 good	간단하고 단백질 기능 예측에 특화	통합 데이터 제공
한계	해석 복잡	비코딩 변이 미포함	데이터베이스 품질에 의존

사용한 Database _ SIFT

1. **다중 서열 정렬(MSA: Multiple Sequence Alignment) data**

   : 돌연변이가 있는 위치에서 단백질 서열의 진화적 보존성을 평가하기 위해 생성

2. **RefSeq** : 인간을 포함한 다양한 종의 표준 참조 유전체 및 단백질 서열 제공

3. **dbSNP (Single Nucleotide Polymorphism Database)** : 돌연변이에 대한 정보 포함

사용한 Database _ ANNOVAR

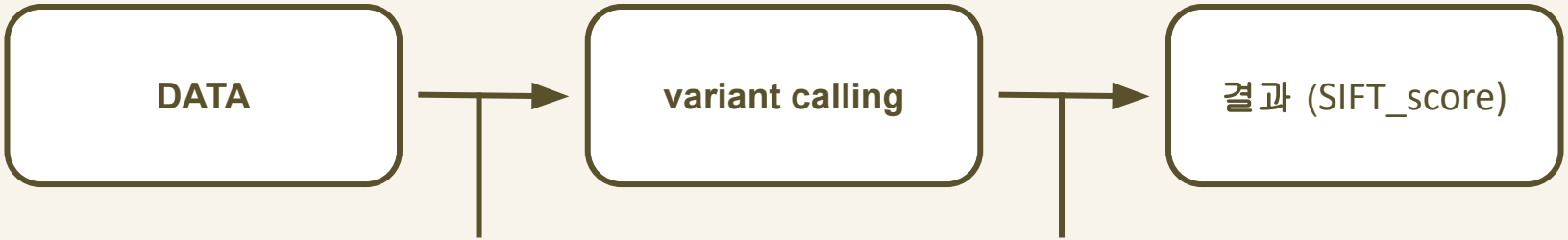
1. **refGene** : 변이가 유전자에 어떤 영향을 미치는지
2. **cytoBand** : 염색체 밴드 정보 제공
3. **exac03** : 엑솜 데이터베이스 (주어진 변이가 연구집단에서 얼마나 흔한지)
4. **avsnp147** : dbSNP 버전 147 데이터 -> rsID 제공
5. **dbnsfp30a** : sift, polyphen-2, cadd의 score 및 prediction 정보 제공

=> database 개수를 늘리면 결과의 변이 주석 개수도 늘어난다.

진행 방법

의생명 정보학 project

진행 방법



ANNOVAR, SIFT, (CADD)

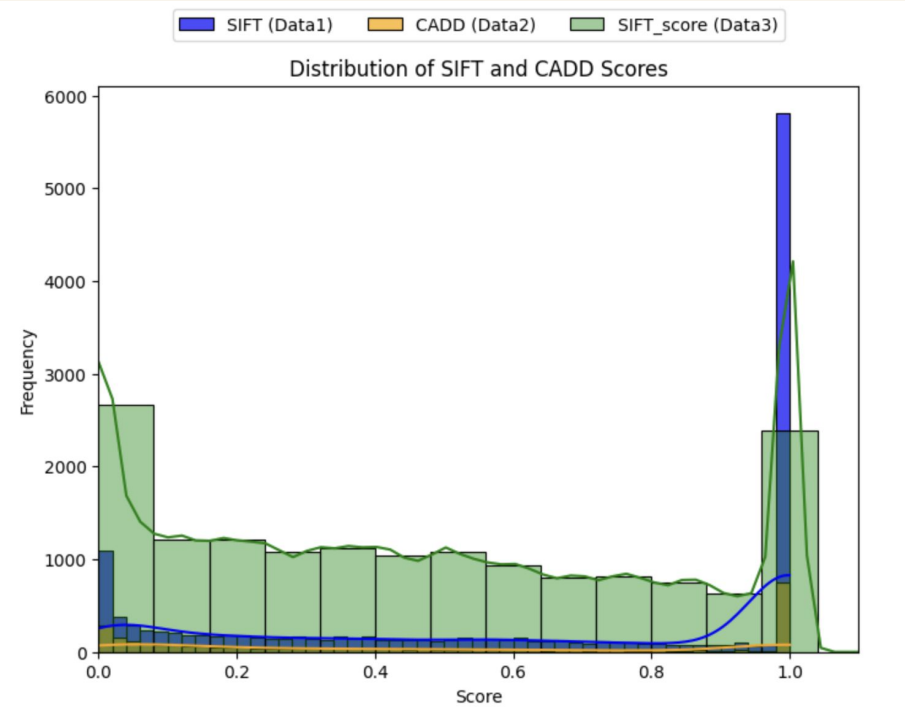
filtering - DELETERIOUS

00번 환자 : ANNOVAR, SIFT, CADD / **01 ~ 09번 환자** : ANNOVAR, SIFT

결 과

의생명 정보학 project

결과 _ SIFT vs ANNOVAR vs CADD



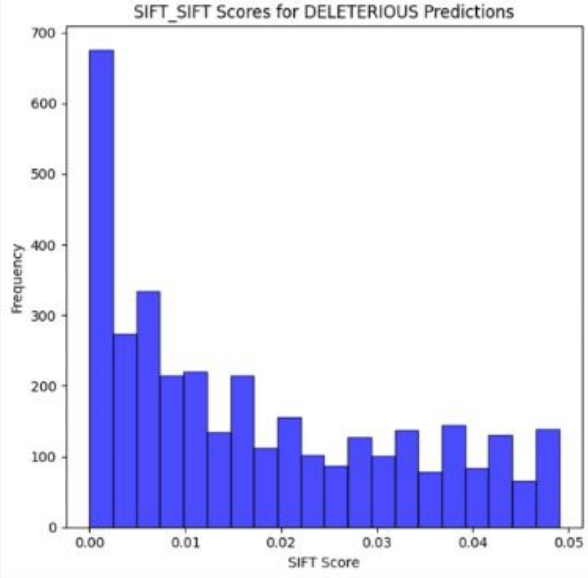
00번 환자

- DATA1 : SIFT
- DATA2 : CADD
- DATA3 : ANNOVAR

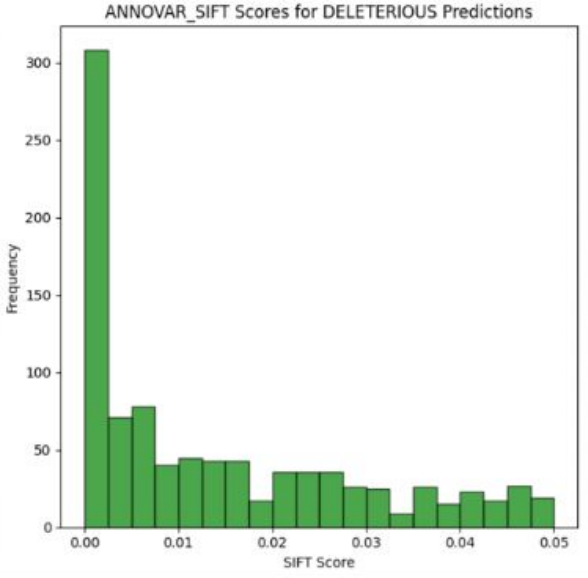
결과 _ SIFT vs ANNOVAR

01번 환자

SIFT



ANNOVAR



결과 _ SIFT vs ANNOVAR

ANNOVAR의 빈도수가 더 작은 이유

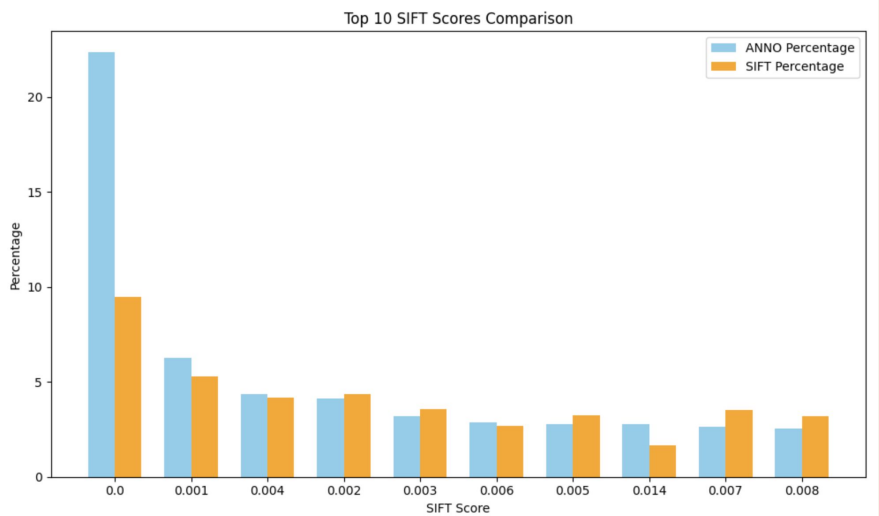
- SIFT 점수 제한
- Isoform 차이 : 한 유전자 locus 에서 primary RNA 가 전사된 뒤 alternative splicing 을 거쳐 나온 여러 형태의 mRNA 산물
- 데이터 통합 방식 차이

결과 _ SIFT vs ANNOVAR

01번 환자

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



References

1. **CADD** : Kircher et al., "A general framework for estimating the relative pathogenicity of human genetic variants," Nature Genetics (2014). https://cadd.gs.washington.edu

2. **SIFT** : Ng and Henikoff, "Predicting deleterious amino acid substitutions," Genome Research (2001). http://sift.bii.a-star.edu.sg

3. **ANNOVAR** : Wang et al., "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," Nucleic Acids Research (2010). http://annovar.openbioinformatics.org

감사합니다

의생명 정보학 project