**우수학부생 연구학점제 연구결과보고서**

**우수학부생 연구학점제 연구결과보고서**

**○ 연구 제목 :** NGS data에 대한 ML기술의 적용

**○ 연구 기간 :** 2022.12.12 ~2023.01.25

**○ 연구결과 요약**

|  |
| --- |
| **『 Abstract 』**  sequencing 기술이 발전해가며 germline mutation뿐만 아니라 somatic mutation에 대한 연구도 활발하게 이루어지고 있다. 하지만 somatic mutation의 경우 모든 cell에서 mutation이 일어난 것이 아니라 일부 세포에서 mutation이 일어난 것이기 때문에 germline mutation calling과 대비해 비교적 낮은 VAF로 인해 SNV들이 제대로 검출되지 않고 있다. 이에 따라 이번 겨울 방학 URP에서는 여러가지 ML기술들을 활용하여 somatic mutation variant calling 결과의 sensitivity를 복잡한 parameter tunning과정을 거치지 않고 precision을 유지한 상태로 높이는 것을 목표로 하여 machine modeling을 실시해보았다.  **『 Introduction 』**  기술 발전에 따라 생물학 연구에서 점점 data의 양은 방대해지고 컴퓨터를 통한 계산의 필요성이 늘어나며 bioinformatics의 중요성이 강조되고 있다. 그 중에서도 NGS data를 활용하여 환자의 샘플에서 다양한 variant를 찾아내는 것 역시 활발하게 연구가 이루어지고 있는 분야 중 하나이며 최근까지도 끊임없이 기술적 발전을 보여주고 있다. 하지만 somatic mutation calling의 경우 낮은 VAF로 인해 SNP를 찾아내는 것에 있어서 대부분의 somatic mutation이 검출되지 않고 있으며 caller의 sensitivity를 높이면 실제 somatic mutation과 variant calling 과정에서 나온 false positive를 구분할 수 없어 실제 분석에 사용할 수 없는 precision이 낮은 결과물이 나오게 된다. 이 때문에 somatic mutation calling의 결과에서 실제 somatic mutation과 variant calling 과정에서 나온 false positive를 구분하는 것은 somatic mutation 연구에 있어서 중요한 토픽 중 하나이다. 이번 URP에서는 이러한 문제를 spike-in data를 활용한 variant calling 결과를 통해 살펴보고 여러가지 ML기법들을 활용하여 이를 해결할 수 있는 해결방안을 제시하는 것을 목표로 하고 있다.  **『 Result 』**  **Spike-in data**  이번 URP에서는 김준호 교수님께서 발표하셨던 Replow 논문의 혈액 샘플 data 중 hybrid capture로 sequencing한 data를 사용하였다. 해당 데이터는 두가지 서로 다른 혈액 샘플을 각각 다른 비율로 섞어 A( VAF : 0.5% ), B( VAF : 1% ), C( VAF : 5% ), D( VAF : 10% )로 VAF를 조정하였고 두가지 혈액 샘플에 대해 각각 Whole Exome Sequencing을 실시하여 두 샘플 간 비교를 통해 true mutation을 정의하였다. 본 보고서에서는 여기서 나온 true mutation을 answer set으로 variant calling 결과에 대한 confusion matrix를 살펴보았다.    Figure 1 : Replow 논문의 spike-in data 자료, A( VAF : 0.5% ), B( VAF : 1% ), C( VAF : 5% ), D( VAF : 10% ) 를 각각 library와 replication을 통해 11,12,21,31와 같이 나누었다.  **Mutect2를 이용한 somatic mutation calling**  먼저 spike-in data를 활용하여 낮은 VAF에서 Mutect2의 variant calling 결과를 확인해 보았다. 그 결과 VAF가 낮은 A( VAF : 0.5% ), B( VAF : 1% )샘플의 경우 True positive가 거의 존재하지 않는 것을 확인할 수 있었고 이에 따라 PPV(precision)와 TPR(sensitivity)가 실제 somatic mutation의 VAF와 가까운 VAF 1% 이하의 샘플에서 눈에 띄게 낮은 것을 확인할 수 있었다. 이것은 앞서 교수님의 Replow 논문에서 언급되었던 사실과도 일치하는 결과이며 현재 보편적으로 많이 사용되고 있는 Mutect2의 variant calling만으로는 somatic mutation에 대한 분석이 어려울 수 있다는 것을 시사한다. 이에 따라 먼저 Mutect2의 parameter tunning을 통해 이 문제를 해결할 수 있는지를 확인하기 위해 Mutect2의 force calling과 mitochondria mode등 sensitivity를 높일 수 있는 몇몇 옵션을 이용해 sensitivity를 높여 보았다. 그 결과 True positive는 증가하였지만 False positive가 폭발적으로 증가하면서 실질적으로 데이터 분석에 사용할 수 없는 결과물 밖에 얻을 수 없었다. 이는 다른 옵션을 사용하여 variant calling을 진행하였을 때도 크게 다르지 않았으며 hard filter와 같은 필터링들을 거쳐도 여전히 해결되지 않았다. 이에 따라 이번 프로젝트에서는 SNP에 초점을 맞추어 mutect2의 parameter tunning을 통해 우선적으로 precision을 고려하지 않고 낮은 VAF에서의 sensitivity를 높인 후 그 결과물을 ML를 이용한 추가적인 classification을 통해 False positive를 제거하는 것으로 방향을 설정하였다.  Figure 2 : VAF에 따른 mutect2 기본 옵션 variant calling 결과물의 TP, FP, FN 갯수    Figure 3 : VAF에 따른 mutect2 기본 옵션 variant calling 결과물의 PPV, TPR  Figure 4 : VAF에 따른 mutect2 parameter tunning을 마친 variant calling 결과물의 TP, FP, FN 갯수    Figure 5 : VAF에 따른 mutect2 parameter tunning을 마친 variant calling 결과물의 PPV, TPR  **Data preprocessing**  Machine modeling에 앞서 먼저 Mutect2를 통해 나온 결과물의 preprocessing을 진행하였다. 이 때 사용한 Mutect2의 결과물은 실제 variant calling의 환경과 최대한 유사하게 맞추기 위해 VAF 1%의 B21 샘플을 사용하였고 normal sample의 경우 depth200으로 down sampling한 것, disease sample의 경우 depth1000으로 down sampling한 것을 사용, confusion matrix로 보았을 때 True Positive : 491, False Positive : 2512, False Negative : 49, PPV (precision) : 0.16, TPR (sensitivity) : 0.909까지 sensitivity를 높인 후 데이터를 추출하여 spike-in data뿐만 아니라 실제 calling의 결과에 최대한 적합한 모델을 만들 수 있도록 유도하였다. 이번 프로젝트에서 변수로는 vcf파일에서 기본적으로 제공하는 정보들 중 해당 샘플에 특이적으로 작용할 수 있는 POS나 CHORM과 같은 정보를 제외한 모든 요소를 추출하여 사용하였다. bcftools query기능과 파이썬을 사용하여 추출을 진행하였고 Info와 Format정보를 모두 추출하여 csv형태로 저장하였다. 이때 한가지 position에 여러 SNP를 가지는 경우는 각각 다른 SNP로 분리하여 저장하였다.  **Machine Learning methods**  이번 프로젝트에서는 Logistic Regression, Gaussian Naïve Bayes Classification, Decision Tree Classifier, Random Forest Classifier로 총 4가지 ML기법을 활용하여 Machine modeling을 진행해보았다. Modeling은 파이썬의 scikit-learn과 앞서 언급한 과정을 통해 얻어진 B21 data를 사용하여 수행하였다. 각각의 모델은 K–fold cross validation을 거친 평균값으로 이번 프로젝트에서 사용할 모델 평가 지표를 얻었으며 K-value = 5이다. Gaussian Naïve Bayes Classification, Logistic Regression의 경우 scikit-learn의 SelectKbest를 이용, 상관관계 분석을 실시하여 상위 7개 변수만을 추출하여 modeling에 사용하였다.  평가 지표로는 모델이 실행한 분류의 정확도를 보여주는accuracy, 전체 mutation중 검출된 mutation의 비율을 보여주는 sensitivity, 모델이 positive로 분류한 요소들의 true/false여부를 비율로 나타낸 precision, precision과 sensitivity의 조화 평균인 F1-score로 4가지를 선정하였다. 특히 현재 제작 중인 모델의 목적이 filter이기 때문에 False Positive의 비율이 적을 수록 그만큼 유효한 모델이라고 할 수 있다. 이에 따라 4가지 평가 지표 중 precision과 F1-score에 좀 더 중점을 두고 모델의 성능을 평가하였다. 앞서 소개한 지표들의 결과로는 accuracy를 기준으로는 Random Forest Classifier이 0.930, Logistic Regression이 0.916으로 상대적으로 높은 점수를 얻었고 sensitivity를 기준으로는 Logistic Regression이 0.577로 상대적으로 낮은 점수를 얻었고 나머지는 0.65±0.02의 비슷한 점수를 얻었다. precision은 Random Forest가 0.859로 가장 높은 점수를 얻었고 Logistic Regression에서 0.819로 다음으로 높은 점수를 얻을 수 있었다. 마지막 F1-score에서는 나머지 3개의 모델에서 0.7이하를 보여주는 가운데 Random Forest는 0.736으로 상대적으로 높은 점수를 보여주였다. 이에 따라 해당 프로젝트에서는 4가지의 machine learning기법 중 해당 모델의 목적에 가장 부합하는 precision과 F1-score에서 가장 높은 점수를 받았고 accuracy, sensitivity에서도 높은 점수를 보여준 Random Forest를 somatic mutation의 filtering을 위한 모델에서 최적의 machine learning 기법으로 가정하고 modeling을 해보았다.  Figure 6 : Machine Learning 알고리즘 별 accuracy의 K-fold CV평균값  Figure 7 : Machine Learning 알고리즘 별 sensitivity K-fold CV평균값  Figure 8 : Machine Learning 알고리즘 별 precision K-fold CV평균값  Figure 9 : Machine Learning 알고리즘 별 F1-score K-fold CV평균값  **Machine modeling**  앞선 과정을 통해 선택한 Random Forest Classifier와 B31 sample을 이용하여 machine modeling을 실시하였다. 그 결과 accuracy 0.942, sensitivity 0.686 precision 0.88, F1-score 0.77의 모델을 얻을 수 있었다. 모델의 성능 확인에 앞서 해당 모델의 feature importance 확인을 진행하였다. feature importance에서 상위를 차지한 주요 요소들은 disease sample의 Count of fragments supporting each allele (FAD), allele fraction(AF), allele depth(AD), strand bias(SB)등과 TLOD score였다. 모두 실제 somatic mutation을 찾아내는 것에 유효할 것이라 생각되는 항목들이므로 별도의 feature selection을 추가로 진행하지 않고 해당 모델의 유효성을 검증하기 위한 테스트를 진행하였다. test sample로는 같은 spike-in data 중 동일한 VAF를 가진 library replicate를 통해 만들어진 B21을 사용하였다. 그 결과 그냥 mutect2를 돌렸을 때 보다 precision은 거의 비슷하게 나오면서 sensitivity를 6.55배 이상 높일 수 있었다. 또한 mutect2의 sensitivity를 높인 후 hard filter를 거친 결과물과 비교하여도 sensitivity는 30%정도 감소하였지만 precision이 5.02배 상승한 것을 관찰할 수 있다. 이것은 mutect2만을 사용하여 somatic variant calling을 진행 할 때보다 이번 프로젝트에서 만든 Random Forest model을 사용한 filtering을 거쳤을 때 precision은 유지하면서 조금 더 높은 sensitivity의 결과물을 얻을 수 있다는 것을 의미한다. 이와 같이 동일한 precision 대비 높아진 sensitivity는 somatic mutation을 찾아내는 것에 있어서 시간적 물적 자원의 소모를 줄여주고 좀 더 정확한 sequencing 결과 분석을 할 수 있도록 도와줄 수 있다.    Figure 10 : Random Forest model의 feature importance  Figure 11 : RF model을 적용한 결과와 기존 사용하던 somatic mutation calling의 결과  **『 Discussion 』**  이번 프로젝트에서는 사용할 수 있는 샘플의 제한으로 machine modeling에는 sample B31을 사용하였고 model의 성능 검증에서는 library replicate를 통해 만들어진 sample B21을 사용하였다. 따라서 해당 두 샘플 사이 연관성이 전혀 없다고 보기는 힘들다. 이에 따라 해당 model의 성능을 정확히 확인하기 위해서는 다른 해당 샘플과는 연관성이 없는 별개의 샘플을 활용하여 추가적인 검증이 필요할 것으로 보인다. 또한, 해당 model같은 경우 spike-in data를 통해 modeling 및 검증을 실시하였는데 실제 somatic mutation에서도 spike-in data에서 수행한 결과와 같이 정확한 성능을 보여줄 수 있는지는 추가적인 연구가 필요한 부분이다.  이와 더불어 이번 프로젝트에서는 modeling에 사용되는 sample로 여러가지 다양한 sample을 사용해보았는데 그 중 B31 sample의 down sampling 결과물 중 가장 depth가 높았던 disease sample depth 1900을 3가지 다른 normal sample depth의 bam file을 활용하여 variant calling을 실시한 후 그렇게 얻은 세가지 결과물을 합친 후 해당 vcf file을 활용하여 machine modeling을 실시해보았다. 그 결과 precision을 포함한 모든 성능 지표가 0.98이상의 높은 성능을 보여주는 model을 만들 수 있었다. 해당 model을 이용하여 B21 sample을 통한 성능 검증을 실시하였을 경우 앞선 Result에서 언급한 결과물보다 높은 precision 0.96이라는 유의미한 결과를 보여준 바 있다. 하지만 Disease sample을 같은 sample로 3번을 사용해 variant calling을 진행하였기 때문에 중복학습으로 인한 overfitting의 가능성을 염두에 두고 해당 프로젝트의 결과물에서는 제외하였다.  추가적으로 mutect2의 parameter optimization의 경우 이번 프로젝트 내에서도 가장 많은 시간이 들여 여러 시도를 해보았던 부분이지만 mutect2에 대한 이해도가 더 높은 연구자일 경우 필자가 찾아내지 못한 최적의 parameter를 찾아냄으로써 해당 프로젝트에서 도출된 mutect2의 결과물보다 더 높은 수행능력을 보여줄 수 있다. 이에 따라 이번 프로젝트를 통해 만들어진 RF model의 precision과 sensitivity 숫자보다는 해당 방법을 통한 filtering의 발전 가능성에 대한 이야기를 하고싶다. 앞서 언급했던 것처럼 이번 프로젝트에서 가장 많은 시간을 할애했던 부분이 mutect2의 parameter tunning이었다. 해당 프로그램에 대한 숙련도가 낮은 경우 대부분의 연구자들이 mutect2를 사용하며 이와 같은 경험이 있을 것이라고 생각한다. 하지만 RF model이 해당 샘플 뿐 아니라 다른 샘플에서도 일반적으로 기능을 할 수 있다면 아직 숙련도가 부족한 연구자들이나 해당 분야와 약간의 거리가 있는 분야를 연구하고 있는 연구자들 역시 해당 툴을 사용함으로써 parameter tunning에 소모해야 하는 시간을 획기적으로 줄일 수 있을 것이다.  **『 Reference 』**  Kim, J., Kim, D., Lim, J.S. *et al.* The use of technical replication for detection of low-level somatic mutations in next-generation sequencing. *Nat Commun* **10**, 1047 (2019). https://doi.org/10.1038/s41467-019-09026-y  Li J, Jew B, Zhan L, Hwang S, Coppola G, Freimer NB, et al. (2019) ForestQC: Quality control on genetic variants from next-generation sequencing data using random forest. PLoS Comput Biol 15(12): e1007556. https://doi.org/10.1371/journal.pcbi.1007556 |

2023년 01월 25일

제출자 : 박성부 (인)

일반대학원장 귀하