논문 초안 작성

연구 주제: **immune cell 조성에 영향을 미치는 miRNA를 알아보기**

Je-Keun Rhee, Seo-young Park

Department of Bioinformatics & Life Science, Soongsil University, Seoul, Republic of Korea

# Abstract

# Introduction

miRNA는 약 22nt의 크기를 가진 single-stranded, small, non-coding RNA로 target하는 RNA를 degradation 시키거나 translation을 block시켜 gene expression을 down-regulate하는 역할을 한다. 그렇게 함으로써 miRNA는 control virtually all cellular processes and playing a pivotal role in human diseases including cancer [2-8]. 또한, 최근 몇년간, the regulation of immunological key processes encompassing development, lineage commitment, activation, function and ageing of innate and adaptive immune cells와 연관된 a vast amount of miRNAs가 발견되어 왔다[9,39,40]. ~~또한, miRNA가 tumor suppressors or oncogenes, mediators of tumor immune escape로 작용할 수 있는 가능성은 지난 10년간 집중적인 연구 주제였다[119, 11].~~

Immune cell은 ~~(역할)이며 B cell, T cell, NK cell, ~~~ 들이 포함된다. 각각의 세포들은 각각의 생김새, 역할이 다르지만, ~~~를 한다 특히, cell composition의 변화는 metazoans와 그들의 complex tissue의 다양한 physiological state(e.g., cellular proliferation, cellular death)의 기저를 이루며, the proportion of different immune cell types in a tumor는 tumor growth, cancer progression, 치료에 대한 환자의 반응에 영향을 미칠 수 있다.

우리는 이를 random forest 모델을 이용해서 진행한다. Random Forest는 popular ensemble learning methods로, 데이터 마이닝 및 기계 학습에 매우 광범위하게 적용된다. The greedy nature of onestep-at-a-time node splitting enables trees (and hence forests) to impose regularization for effective analysis in “large p, small n” problems and the “grouping property” of trees [3] enables RF to adeptly deal with correlation and interaction among variables. RF can also be used to select and rank variables by taking advantage of variable importance measures. Thus, these properties of RF make it an appropriate tool for genomic data analysis and bioinformatics research.

이 paper에서 우리는 miRNA expression profile에서 immune cell composition을 예측하는 random forest 모델을 만든다. 그러고 그 모델의 예측에 영향을 미친 중요한 miRNAs가 무엇인지 알아내고, 그 miRNAs들이 어떤 것이길래 중요한지에 대해 분석을 진행했다.

우리가 진행한 이 연구는 ~~에 도움이 될 것으로 보인다.

# Result

## **miRNA expression profile로 immune cell composition 예측 수행**

우리는 “large p, small n" 문제에 효율적이며, 변수들 간의 correlation and interaction을 잘 다루고, feature selection에 유용하다는 특징을 가진 random forest 모델을 이용하여 miRNA expression profile data로부터 immune cell composition 예측을 진행했다. 여기서 우리는 mean squared error와 Pearson correlation coefficient를 평가지표로 삼았으며, 그것으로의 예측 성능은 Table과 Figure와 같다.

하나의 miRNA가 보통 200~400개의 gene을 target하고, 하나의 gene이 여러 miRNA로 target될 수 있기에 직접적으로 특정 phenotype과 연관된 miRNA를 알아내는 것은 쉽지 않다. 하지만 correlation이 있다는 것만으로도 특정 phenotype과 연관된 miRNA를 추릴 수 있다는 점에서 correlation이 0.5이상이다, 즉 이 결과가 random으로 나온 것이 아니라는 것은 충분히 유의미하다.

테이블이(가) 표시된 사진

자동 생성된 설명

Correlation coefficient로의 성능 plot (그래프에 들어갈 내용은,, 아래와 같이 각 cell type 별로 x축은 EPIC, y축은 우리의 모델에서의 결과로. Pearson correlation (R) and linear regression (dashed lines), p-value.)

텍스트, 도표, 라인, 폰트이(가) 표시된 사진

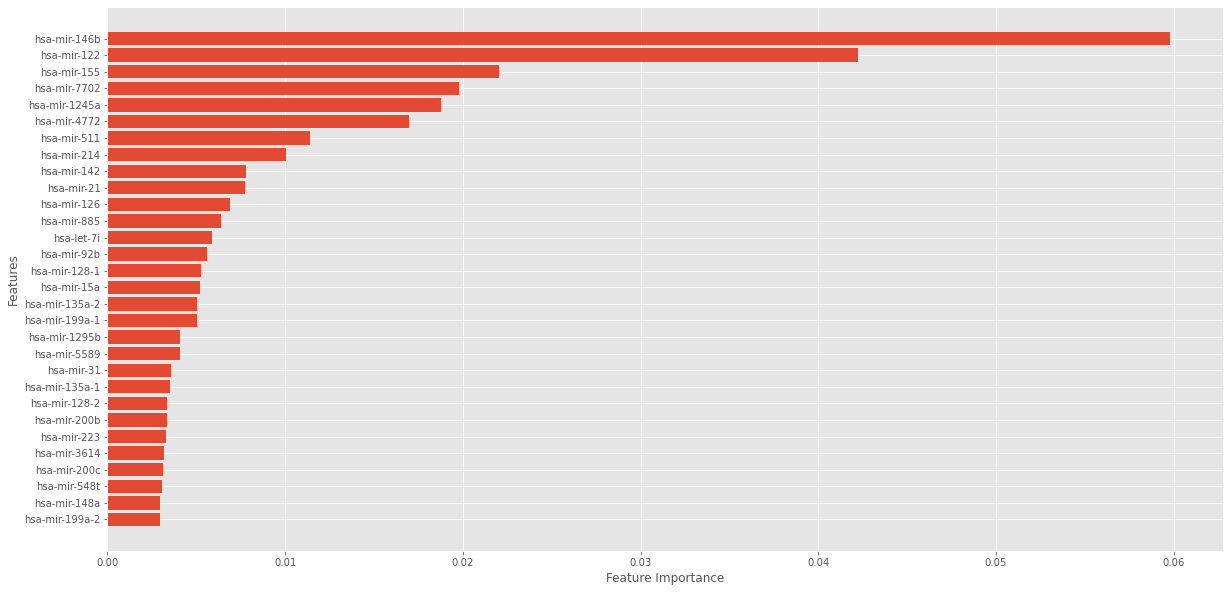
자동 생성된 설명텍스트, 영수증, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

## **immune cell composition에 주요한 영향을 끼치는 miRNA selection**

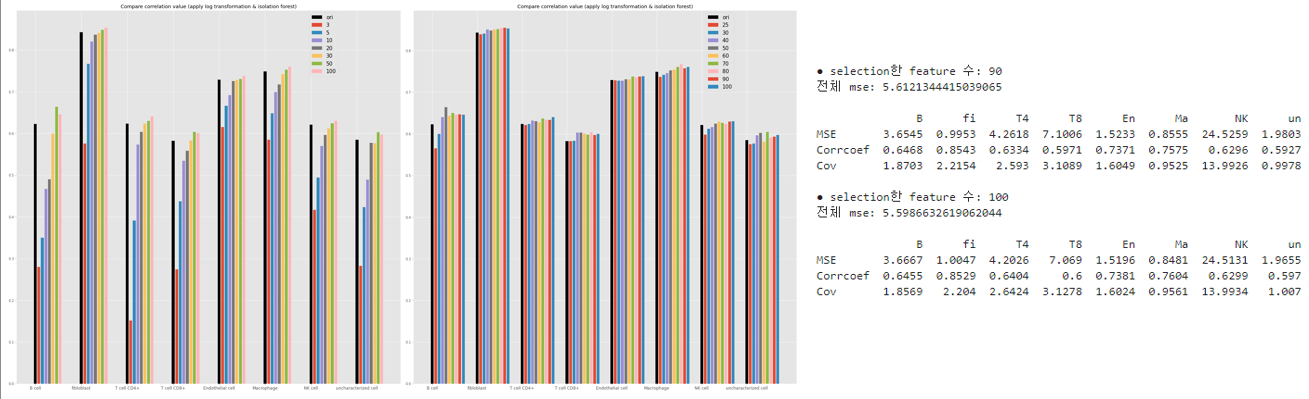
random forest모델로 feature importance 계산이 가능하다. 이 feature importance는 ~~(계산 방법)로 얻을 수 있으며, ~~의 의미를 가진다. 따라서 우리는 random forest 모델이 계산하는 feature importance 값 중 상위의 것을 기준으로 모델의 예측에 유의미한 miRNA를 선택하였다. Figure를 보면 feature importance score가 상위 6개의 feature까지는 ~%의 급격한 변화를 보이지만, 그 이상의 feature들은 feature importance score가 그만큼의 변화 없이 서로서로 비슷함을 알 수 있다. 그리고 그 선택한 miRNA 개수에 따른 immune cell composition 예측 통해 그 miRNA들 만으로 해당 예측 성능이 어느정도 나오는지 알아냄으로써 한 번 더 해당 miRNA들의 중요성을 확인하였다. 그 결과 성능이 매우 뛰어나지는 않지만 모델이 random이 아닌 유의미한 값을 냄을 알 수 있었다.

따라서 우리는 6개의 miRNA를 유의미한 feature로 삼고 downstream analysis를 진행하였다.

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

feature importance에 대한 그래프 (top 30개에 대한 것)



Feature가 어느정도 했을 때 성능이 어느정도 나오는지.

(feature 수가 5개로 하여 같은 random forest 모델을 수행한 것에서는 correlation coefficient가 0.6보다 큰 것 3개, 0.4<x<0.5인 것 3개, x<0.4인 것 2개)

## **miRNA의 target gene에 대해서 조사**

우리는 선택한 miRNA 각각의 target gene prediction을 TargetScan[]과 miRDB[]를 이용해서 진행하였다. 또한 target gene들이 어떤 역할을 하는지 알기 위해서 enrichr을 이용하여 gene set enrichment analysis를 수행하였다. 그 수행 및 분석 결과는 위의 table/figure와 같다.

이를 통해 우리는 miRNA의 target gene들(그들 중 일부)이 ~~~한 역할을 하는 것을 알 수 있다. 또한 이로 인해 immune cell composition에 중요한 역할을 한다고 생각할 수 있는 것을 예상할 수 있다.

## **Validation of our model with external data**

이 결과가 해당 TCGA miRNA expression profile만이 아닌 다른 데서도 유의미하다

# Discussion

우리는 immune cell composition에 영향을 미치는 miRNA를 알아보았다. 이를 위해 random forest 알고리즘을 이용해서 miRNA expression profile로부터 immune cell composition을 잘 예측하는 모델을 만들고, 그 모델의 feature importance를 바탕으로 유의미한 miRNA를 뽑았다. 그리고 그 miRNA가 gene expression에 어떤 영향을 주는지 알기 위해 해당 miRNA의 target gene analysis를 진행했다.

그 결과 (무슨 gene/miRNA) 이 유의미함을 알 수 있었고, 그들은 ~~~한 역할을 함을 알 수 있었다. 이들을 ~~~의 이유로 immune cell composition에서 중요한 역할을 함 또한 추론할 수 있다.

또한, 모델 구축에 사용되지 않는 데이터를 테스트 세트로 사용하여 모델과 기능을 추가로 검증하였을 때, 이 n개의 miRNA를 이용한 모델은 충분한성능을 내었다.

우리의 연구는 ~~과 ~~에 유용히 쓸 수 있을 것이다. 또한 이를 이용해 ~~도 가능할 것이다. *현재 immunotherapy에의 활용, 진단 마커로의 활용을 생각하고 있지만, 그것이 맞는지 조사 필요. 그리고 구체적으로 어떻게 활용될 수 있는지 또한 조사 필요.*

하지만 우리의 연구에 몇 가지 한계점이 있다.

우선, immune cell composition data가 직접적인 측정을 통해 나온 결과가 아닌 EPIC이라는 tool로 mRNA expression profile을 바탕으로 예측한 것이라는 것이다. 이로 인해 EPIC에서 중요시하는 mRNA를 target으로 하는 miRNA가 많이 선택되었다. 하지만, 우리의 연구는 miRNA를 이용한 immune cell composition 예측에 대한 방법론적인 기반이 되어, 이를 바탕으로 이 연구를 재현하여 모델을 구현 및 target 분석을 진행한다면, 충분히 유의미한 것을 찾아내는 것이 가능할 것이다.

또한, 우리는 random forest 모델만을 사용하는 모델로서 택했지만, 다른 대안이 있을지 모른다. 우리의 모델보다 성능이 좋고 feature importance와 같이 모델 예측에 중요한 feature를 뽑아낼 수 있으면 더욱 유의미한 miRNA를 뽑아낼 수 있을 것이다.

후속 연구로 위의 한계를 보완하는 연구가 가능할 것이다.

# Method

## **Data access**

MiRNA expression quantification (stem loop expression) data were downloaded from TCGA Genomic Data Commons (GDC) and it covers 11,020 patients across 33 cancer types. These datasets were represented as 1881 miRNA expression values (log2(RPM+1)) per sample.

We downloaded immune infiltration estimations by EPIC[] algorithms from TIMER[] web server, a comprehensive resource for systematical analysis of immune infiltrates across diverse cancer types.

## **Data preprocessing**

It must be ensured that both the input data and the output data share the same sample. Therefore, first, both datasets are limited to contain samples that are available in both datasets. Finally, 10,648 sample were used in our experiment.

Next, we performed two processing procedure on the output data.

We performed log-transformation to deal with skewed data and removed outliers with isolation forest [] using the IsolationForest class from the Sklearn ensemble module.

Isolation forest는 high dimensional data set에서 outlier를 탐지할 때 효과적인 decision tree 계열의 비지도학습 알고리즘이다. 이는 임의로 feature를 선택하고 선택된 feature의 최댓값과 최솟값 사이의 임의의 분할값(split value)선택함으로써 observations을 분리시킴으로써 데이터셋을 결정 트리 형태로 표현해, 정상 데이터를 분리하기 위해서는 트리의 깊이가 깊어지고, 반대로 이상치는 트리의 상단에서 분리할 수 있다는 개념을 이용하여 outlier를 제거한다.

## **Model implementation and selection**

All datasets from our database were divided into training and validation datasets (80% of all datasets and split it into 80% and 20% for each) and holdout datasets (test only, 20% of all datasets) for final evaluation of algorithm performance.

We used random forest model to perform prediction of immune cell composition from miRNA expression data. We implemented it using RandomForestRegressor class from the Sklearn ensemble module and tuned its parameter using grid search with 10-fold cross-validation using GridSearchCV class from the Sklearn model\_selection module.

이에 사용한 parameter 목록과 최종적으로 선택된 model의 parameter는 아래의 표와 같다.

Table . Considered parameters and values for grid search cross-validation and selected value for training the models.

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Considered value** | **Selected value** |
| n\_estimators | 10, 30, 50, 100, 200 | 200 |
| max\_depth | None, 5, 7, 10, 20, 30 | None |
| max\_features | 0.5, 0.7, 0.8 | 0.5 |

또한, 몇 개의 miRNA가 모델에서 유용하게 쓰였는지 알기 위해, random forest가 계산하는 값 중 feature importance의 상위 n개 feature(miRNA)를 대상의 것으로 삼아 같은 모델에 대해 해당 조사를 진행함.

model selection 및 평가의 성능지표로 Pearson correlation coefficient을 사용했다.

## **Analysis of the target genes of miRNAs**

여기서 ~~~의 이유로 유의미한 miRNA를 ~~의 6개를 택하였고 그들 각각에 대해 target gene prediction을 했다.

우리는 target gene prediction을 위해 현존하는 공개된 miRNA target gene prediction tool 중 가장 성능이 좋다고 알려진 TargetScan[]과 그 외의 target gene prediction에 많이 쓰이는 miRDB[]를 이용했다.

구체적으로는, TargetScan과 miRDB에 miRNA name을 넣어 prediction된 target gene을 대상으로 두 DB의 교집합을 이용해서 믿을 수 있는 miRNA를 대상으로 하여 진행함. TargetScan에서 나오는 score는 ~ 의미가 있어서 >80의 조건을 적용했고, 조사한 target gene이 모두 괜찮은지 한 번 더 확인을 위해서 miRDB의 결과와 겹치는 것을 사용하였다.

그 후, 그 target gene set이 어떤 특성을 갖는지 알기 위한 enrichment analysis 분석을 위해 enrichr을 사용했다. 여기서 p-value가 0.005 이하인 것을 대상으로 분석을 진행했다. ~~~

## **Validation of our model with external data**

validation을 위해 어떤 data를 썼다. 그를 위해 어떤 preprocessing 과정을 거쳤고, 동일한 모델을 썼다.

# Reference

Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32

Julien RacleKaat de JongePetra BaumgaertnerDaniel E SpeiserDavid Gfeller (2017) Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data eLife 6:e26476.

Pedregosa, F.; Varoquaux, G.; Gramfort, A. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 2011, 12, 2825–2830

[1] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on.

[2] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation-based anomaly detection.” ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012): 3.

Taiwen Li, Jingxin Fu, Zexian Zeng, David Cohen, Jing Li, Qianming Chen, Bo Li, and X. Shirley Liu. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Research 2020

Taiwen Li, Jingyu Fan, Binbin Wang, Nicole Traugh, Qianming Chen, Jun S. Liu, Bo Li, X. Shirley Liu. TIMER: A web server for comprehensive analysis of tumor-infiltrating immune cells. Cancer Research. 2017;77(21):e108-e110.

Bo Li, Eric Severson, Jean-Christophe Pignon, Haoquan Zhao, Taiwen Li, Jesse Novak, Peng Jiang, Hui Shen, Jon C. Aster, Scott Rodig, Sabina Signoretti, Jun S. Liu, X. Shirley Liu. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biology. 2016;17(1):174.

Lewis, B.P.; Burge, C.B.; Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005, 120, 15–20.

Chen, Y.; Wang, X. miRDB: An online database for prediction of functional microRNA targets. Nucleic Acids Res. 2020, 48, D127–D131.