# 인공지능을 이용한 암 패턴 분석

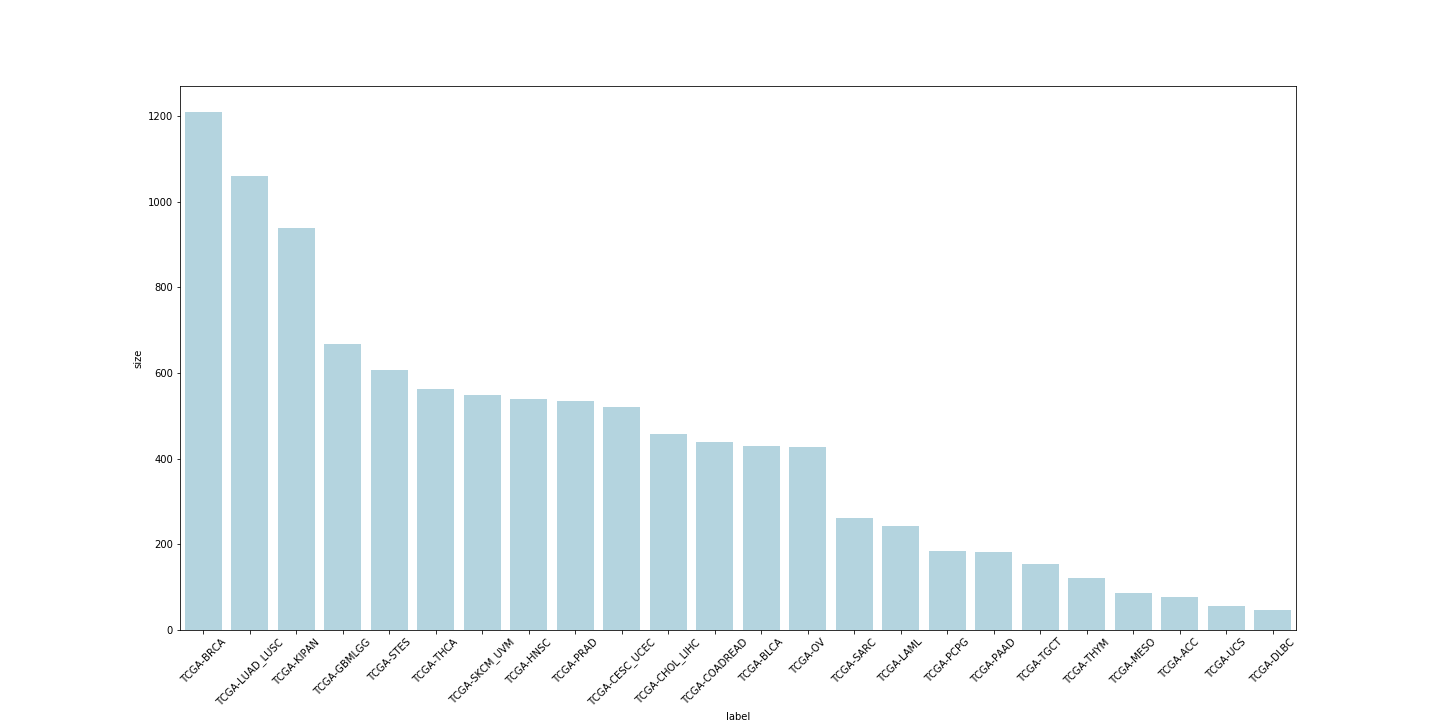
## 배경설명

유전자는 단백질을 암호화하고 단백질은 세포의 기능을 결정한다. 따라서 특정 세포에서 발현된 수많은 유전자는 그 세포가 무엇을 할지를 결정한다. 따라서유전자의 발현 데이터는 정상 세포와 암 세포를 구분할 수 있는 근거가 된다. 2006년에 시작한 The Cancer Genome Atlas (TCGA) 프로젝트는 대규모 유전체 시퀀싱 및 게놈분석 기술을 통해 암의 분자생물학적 기초에 대한 이해를 촉진하기 위해 표준화된 방법으로 33개 암종에서 유래한데이터를 연구자들에게 제공한다. 이 중 유전자 발현에 관련된 gene expression, DNA methylation, microRNA expression 데이터를 이용하여 인공지능 기술을 통하여 암과 정상을 구분하고, 암종의 발생 부위를 예측하는 기술을 개발하였다.

## 데이터

TCGA는 33개의 암종에 대해서 10,361명의 암 유전체 데이터를 제공하고 있다. 우리는 33개의 암종 중 같은 조직에서 유래한 암종을 병합하여 24개의 암종으로 분석을 진행하였다.

|  |  |  |  |
| --- | --- | --- | --- |
| 번호 | 암종 코드 | 암종 | 샘플 수 |
| 1 | TCGA-BRCA | 유방암 | 1,209 |
| 2 | TCGA-LUAD\_LUSC | 폐선암, 폐편평상피세포암 | 1,061 |
| 3 | TCGA-KIPAN | 신장암 | 938 |
| 4 | TCGA-GBMLGG | 신경교종 | 667 |
| 5 | TCGA-STES | 위-식도암 | 607 |
| 6 | TCGA-THCA | 갑상선암 | 562 |
| 7 | TCGA-SKCM\_UVM | 악성흑색종, 맥락막흑색종 | 549 |
| 8 | TCGA-HNSC | 두경부암 | 540 |
| 9 | TCGA-PRAD | 전립선암 | 535 |
| 10 | TCGA-CESC\_UCEC | 자궁경부암, 자궁내막암 | 521 |
| 11 | TCGA-CHOL\_LIHC | 담도암, 간암 | 458 |
| 12 | TCGA-COADREAD | 대장암 | 439 |
| 13 | TCGA-BLCA | 방광암 | 430 |
| 14 | TCGA-OV | 장액성 난소상피암 | 428 |
| 15 | TCGA-SARC | 육종 | 262 |
| 16 | TCGA-LAML | 급성골수성백혈병 | 244 |
| 17 | TCGA-PCPG | 부신암 | 185 |
| 18 | TCGA-PAAD | 췌장암 | 183 |
| 19 | TCGA-TGCT | 고환암 | 154 |
| 20 | TCGA-THYM | 흉선종 | 121 |
| 21 | TCGA-MESO | 악성중피종 | 87 |
| 22 | TCGA-ACC | 부신피질암 | 77 |
| 23 | TCGA-UCS | 자궁육종 | 57 |
| 24 | TCGA-DLBC | 미만성거대B세포림프종 | 47 |



데이터 내 각각의 클래스들이 차지하는 데이터의 비율이 균일하지 않고 한쪽으로 치우쳐져 있을 경우 클래스 불균형(class imbalance)이라고 한다. 우리는 이를 수치화 하기 위해 ID(imbalance degree)를 측정해 보았다[1]. 클래스 불균형 문제는 1) balanced problem, 2) unbalnaced problem showing multi-majority, 3) a multi-minority unbalanced problem 세 가지로 분류되는데[2,3], ID 값으로 어떠한 불균형에속하는지 알 수 있다. unbalanced problem showing multi-majority 와 a multi-minority unbalanced problem 중간 정도로 imbalance 한 것을 확인했다( ID : 11.56).

우리는 이러한 클래스 불균형에 의한 작은 샘플 수를 가진 클래스 예측 성능 저하 문제를 해결하기 위해서 data-level methods의 접근 방법을 선택했고, sampling이 아닌 여러 타입의 데이터로 feature selection 방법을 통해 모델의 예측 성능을 개선하는 전략을 취했다.

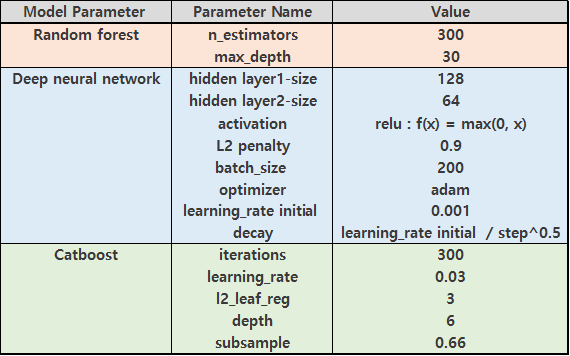
## 특징 선택 (Feature Selection)

RNA 발현량 값은 약 2만개(gene expression), methylation베타 값은 약 48만개(DNA methylation) 정도의 feature가 있다. 우리는 이 중 1,000개에서 5,000개까지 특징선택을 한 후에, 같은 숫자의 feature 수를 이용 할 때 RNA와 methylation 중 어떤 타입의 데이터가 예측하는 데 유리하며, 최적의 feature 수가 알아보려고 한다. 다음과 같은 선택 기준을 설정하고 weighted voting 방식으로 feature를 선정하였다.

feature의 선택 기준

* 각 feature별 variance.
* xgboost 및 lightgbm의 gradient boosting 모델의 트리 생성 시 사용되는 feature의 횟수.
* support vector machine에 L1 페널티를 적용하여 feature를 선정.

## 모델

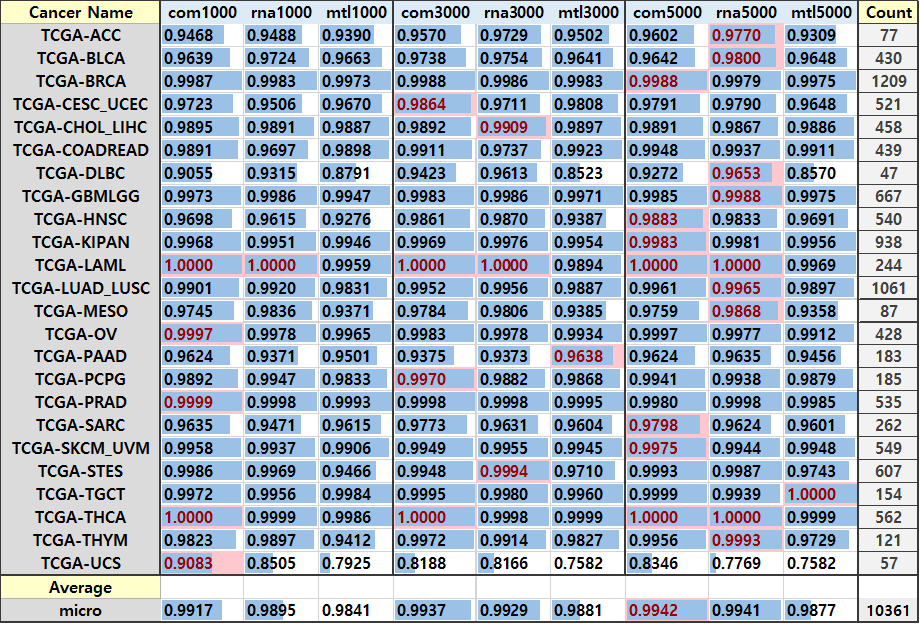
불균형 데이터에는 bagging 및 boosting 기반의 알고리즘이 잘 동작한다고 알려져 있다[5,6,7]. 따라서 bagging 방식의 random forest, boosting 방식의 catboost, deep neural network를 가지고 실험해 보았다. 각각의 모델의 하이퍼 파라미터는 다음과 같다.

## 검증방법

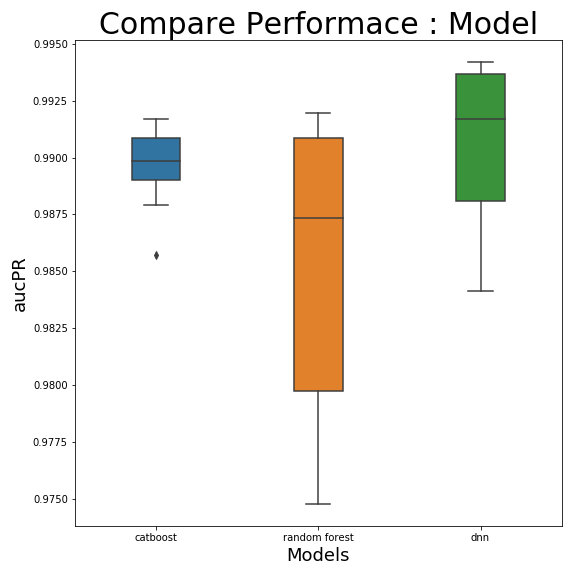
우리가 사용한 데이터는 클래스 불균형이므로 일반적인 accuracy로는 적은 샘플의 클래스에 대한 성능 측정이 어렵기 때문에, 주 성능 지표로는 aucPR(Area Under the Precision-Recall Curve)를 사용했다. (micro-average 사용) 그리고 데이터의 샘플 수가 많지 않기 때문에 5-fold cross validation으로 성능을 측정했다.

## 실험 결과

우리는 실험에서 1. 어떤 모델이 최적 인 지, 2. 최적 모델에 대해서 어떤 데이터세트가 가장 성능이 좋은지를 알아보려고 합니다. 하지만 모든 암 종에 대해서 최적인 모델과 데이터세트는 존재 하지 않았고, 따라서 우리는 평균 성능으로 평가를 했습니다. 예시로서 다음의 테이블은 Deep neural network모델의 각 암 종별 예측 성능 테이블 입니다. 여기서 우리는 암 종 별로 최적의 데이터세트가 다르다는 것을 알 수 있습니다. (**com = RNA + methylation , rna = RNA , mtl = methylation** 을 나타내고, 뒤의 숫자는 feature 수를 나타냅니다. 파란색 음영의 길이는 성능과 비례하고, 빨간색 숫자는 각 암 종에서 aucPR 값이 가장 높은 데이터 셋을 나타냅니다. 마지막 열 **Count**는 샘플 수를 나타냅니다.)



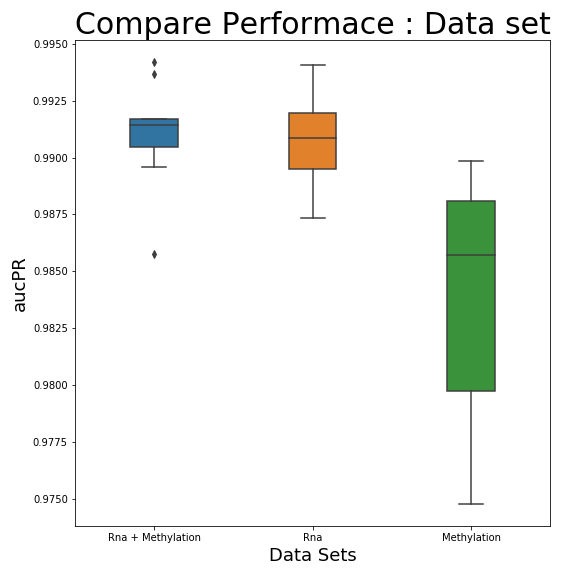
먼저 아래의 결과는 1. 어떤 모델이 최적 인 지를 알아보기 위해서, 각각의 암 종과 데이터 세트에 대한 평균 성능을 나타낸 box-plot입니다.



아래는 통계적으로 유 의미 한지 살펴보기 위해 실행한 T-test 결과 입니다.

3.0435128920709293, pvalue=0.0030829854152844122

T-test 의 결과를 보면 p-value는 0.05 이하로, 성능 비교는 통계적으로 유의미한 것을 알 수 있다. 따라서 DNN(Deep Neural Net)이 통계적으로 유의하게 가장 좋은 성능을 보이는 것을 알 수 있었습니다.

다음은 2. 최적 모델에 대해서 어떤 데이터세트가 가장 성능이 좋은지를 알아보기 위해, 데이터 조합 기준으로 그린 box-plot입니다. 

아래는 통계적으로 유 의미 한지 살펴보기 위해 실행한 T-test 결과 입니다.

3.0435128920709293, pvalue=0.0030829854152844122

T-test 의 결과를 보면 p-value는 0.05 이하로, 성능 비교는 통계적으로 유의미한 것을 알 수 있다. 따라서 RNA+Methylation조합이 통계적으로 유의하게 가장 좋은 성능을 보이는 것을 알 수 있었습니다.

## 

## 결론

이번 실험을 통해 TCGA의 24종 암 종 데이터의 클래스 불균형 문제에서 RNA와 methylation 데이터를 조합하여 사용시, 어떠한 효과가 있는지 확인해 보았습니다. 암 종에 따라서 최적의 모델과 데이터 조합은 다르지만, 평균적인 성능을 aucPR지표로 측정 시 DNN모델, RNA+Methylation조합이 통계적으로 유 의미 하게 최적의 성능을 보여주는 것을 확인 했습니다.

## Appendix A : Feature Importance Result

모든 값은 0~1 사이로 스케일링 했다.

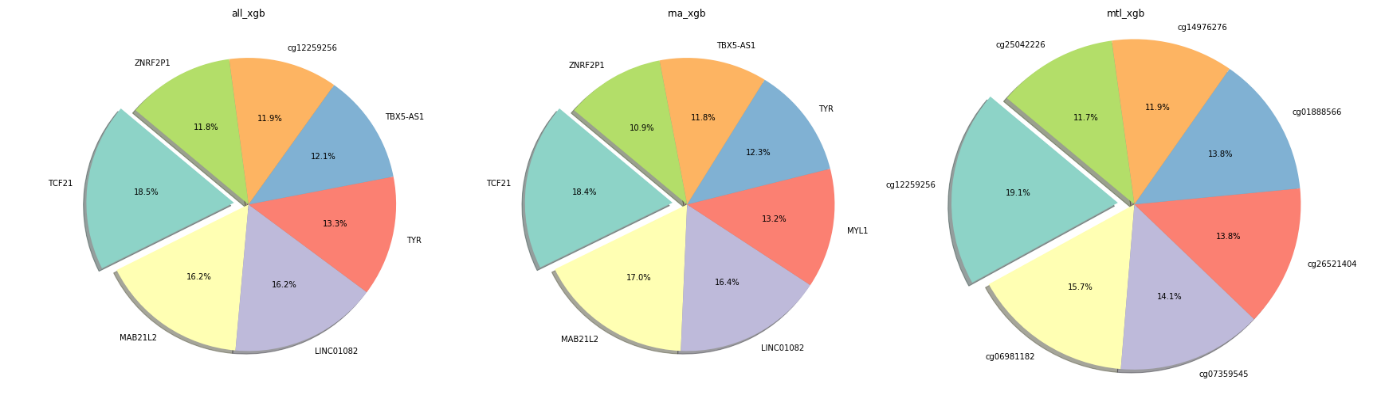


Figure A.1: Top7 Feature from lightgbm

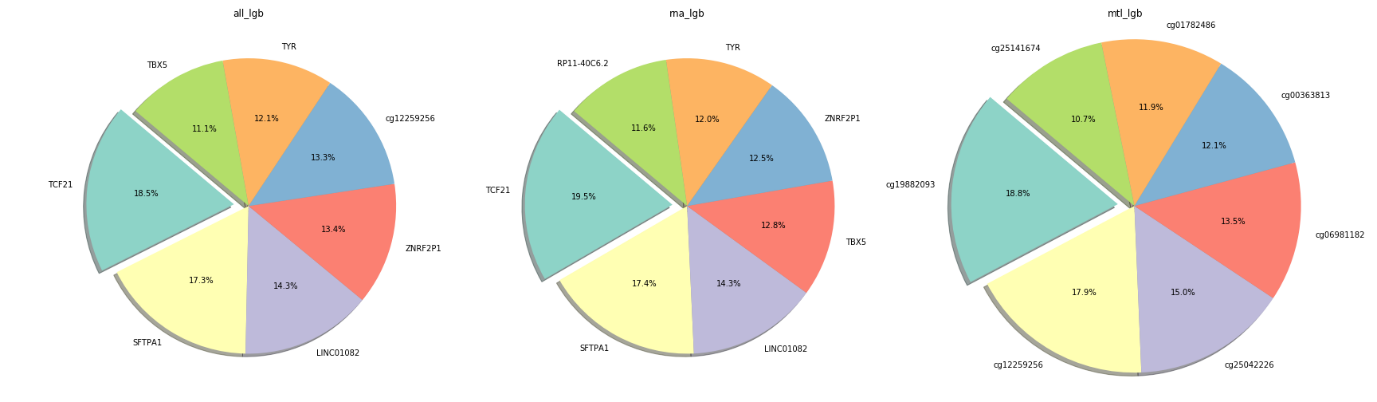


Figure A.2: Top7 Feature from xgboost

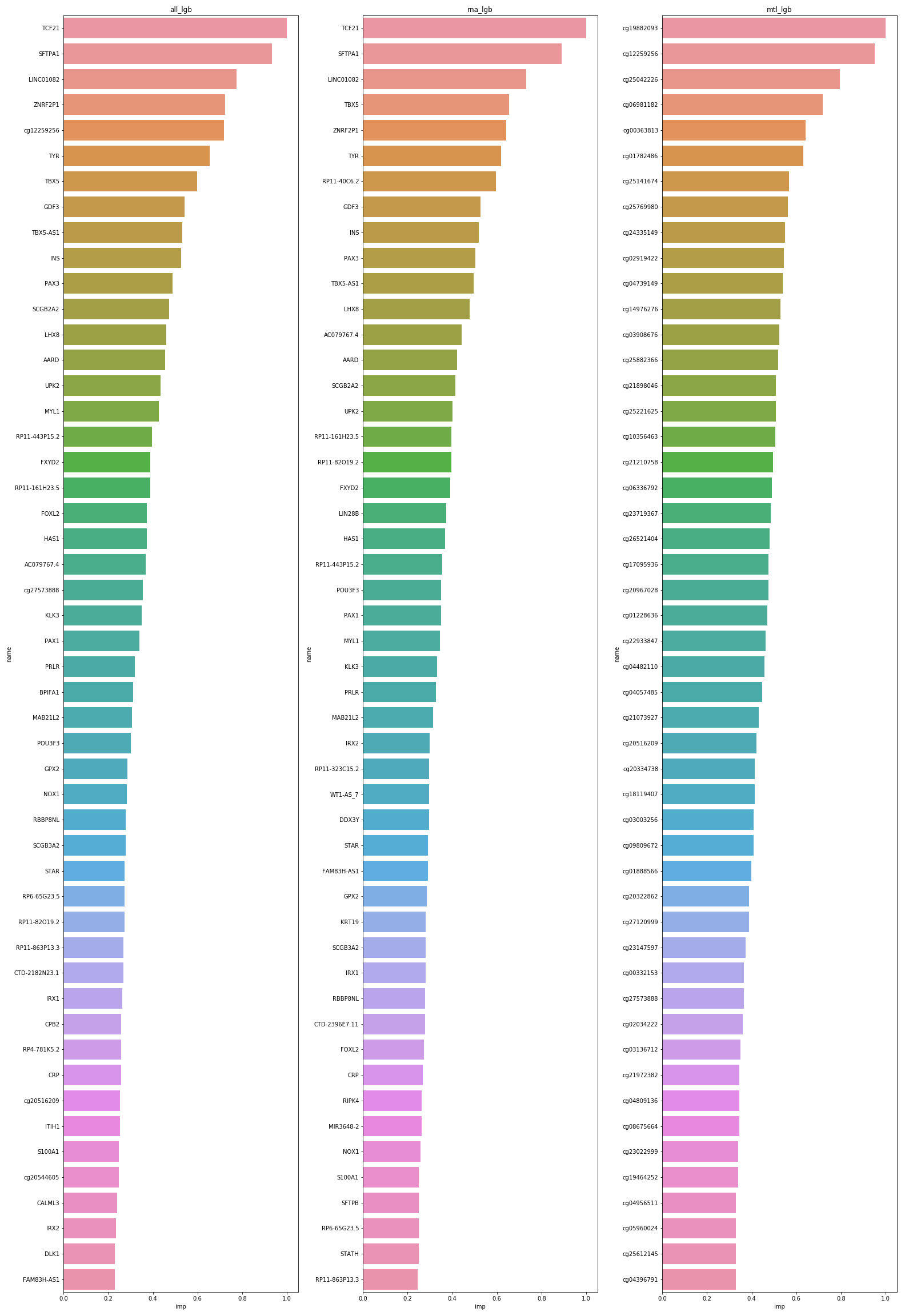
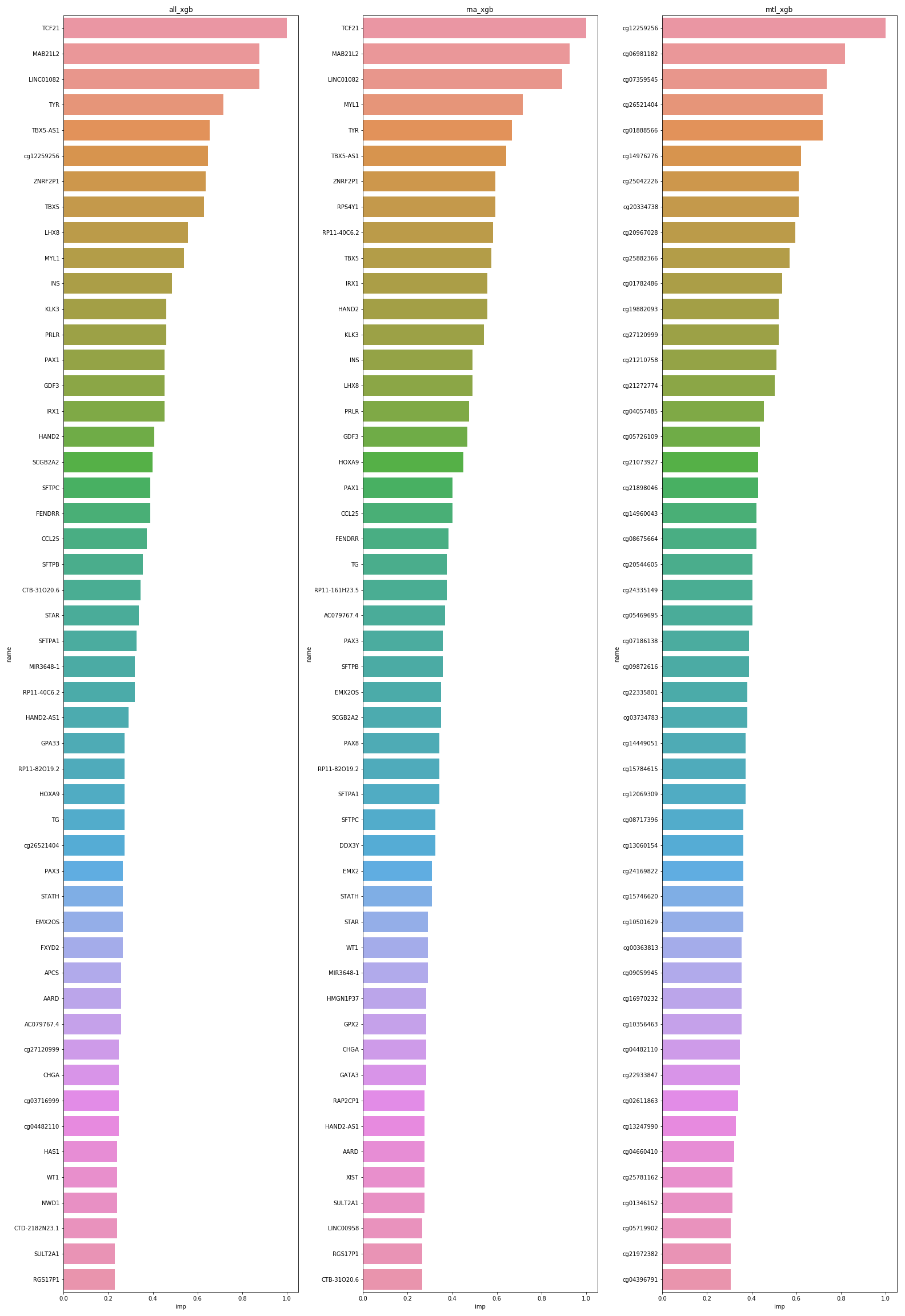
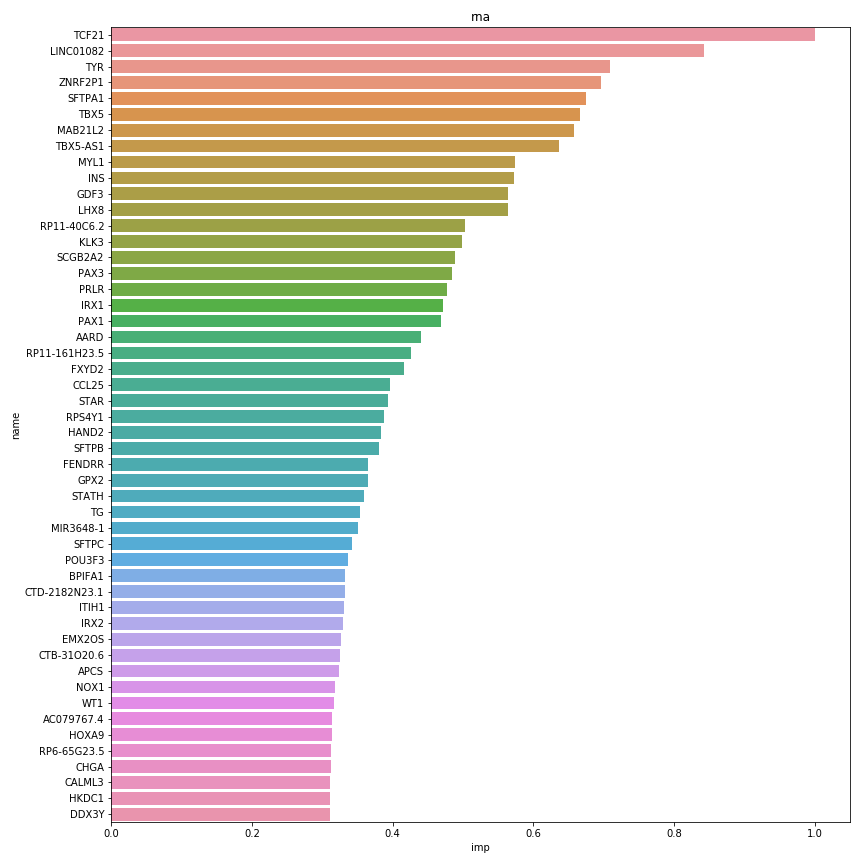
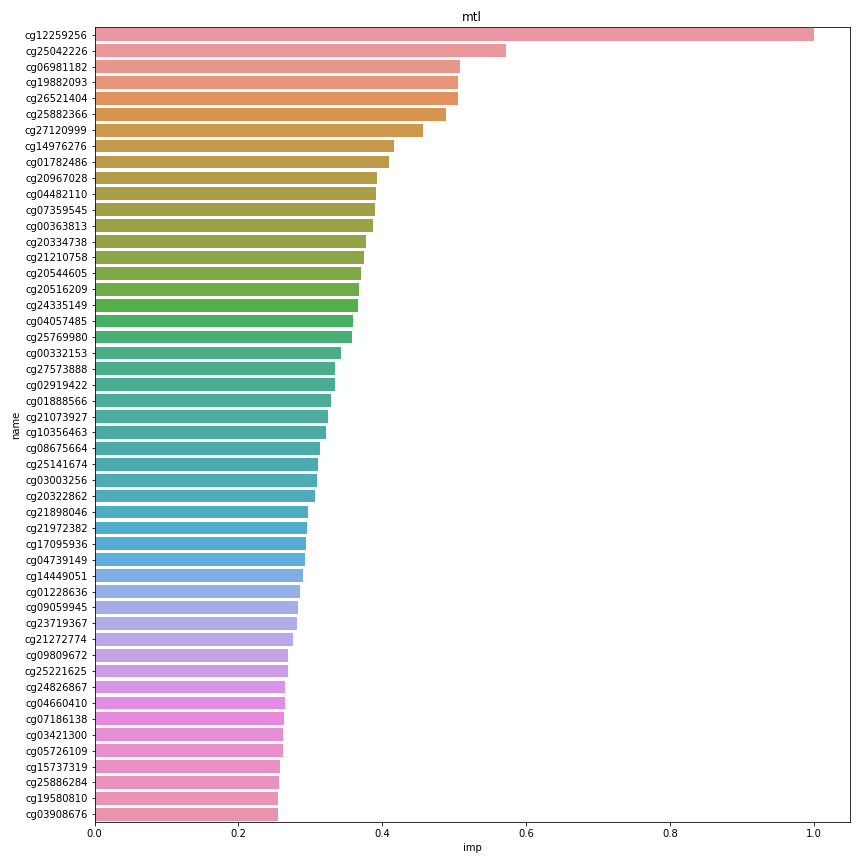


Figure A.3: Top50 Feature of RNA+Methylation , RNA , Methylation from lightgbm

Figure A.4: Top50 Feature of RNA+Methylation , RNA , Methylation from xgboost



Top50 Feature of mRNA from weighted voting



Top50 Feature of Methylation from weighted voting

[1]Ortigosa-Hernandez, J., Inza, I., Lozano, J.A., 2017. Measuring the class- ´ imbalance extent of multi-class problems. Pattern Recognition Letters 98, 32–38. [2] He, H., and Garcia, E. A. Learning from imbalanced data. IEEE Trans. on Knowledge and Data Engineering 21, 9 (Sep. 2009), 1263–1284. [3] Wang, S., and Yao, X. Multi-class imbalance problems: Analysis and potential solutions. IEEE Trans. on Systems, Man and Cybernetics - Part B 42, 4 (Aug. 2012), 1119–1130. [4]Blaszczynski, J., Stefanowski, J.: Neighbourhood sampling in bagging for imbalanced data. Neurocomputing 150, 529–542 (2015) [5]Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C 42(4), 463–484 (2012) [6]Krawczyk, B., Wo´ zniak, M., Schaefer, G.: Cost-sensitive decision tree ensembles for effective imbalanced classification. Appl. Soft Comput. 14, 554–562 (2014) [7] https://en.wikipedia.org/wiki/TCF21\_(gene)#Role\_in\_development