1. Data collection and processing

TCGA RNA expression data를 획득하기 위해 UCSC Xena를 이용하여 data를 수집하였다. 우리는 count 값을 사용하기 위하여 TCGA-PANCAN project의 TOIL RSEM - expected count data를 다운로드 받았다. 처음 Xena에서 제공하는 gene expression data는 log transformed 형태이기 때문에 raw count 값으로 converting 및 missing data에 대한 filtering을 거친 후 총 10,530명에 대한 60,498개의 gene expression data를 추후 분석에 이용하였다.

1. Feature generation

위의 gene expression data를 토대로 각 sample에 대한 feature 생성을 진행하였다. “Interpretable systems biomarkers predict response to immune-checkpoint inhibitors” 논문을 참고하여 feature 생성을 진행하였으며 총 9개의 category ( Gold Standard; Immune Cells quantification; Pathway activity; Transcription factors activity; Ligand-receptor pairs; Cell-cell pairs )로 나뉘어진 model과 gene expression data를 이용하여 각 sample에 대하여 총 1,134개의 feature data를 생성하였다. 이때 feature 생성을 위하여 R program의 “easieR” package를 이용하여 분석을 진행하였다.

1. Decoding feature data & Creating subgroup models

이렇게 뽑은 feature들을 이용하여 subgroup model을 생성하는 데에 있어서 data processing의 용이함을 위하여 unsupervised learning에 주로 쓰이는 neural-network based model인 AutoEncoder를 사용하여 decoding을 실시하였다. AutoEncoder를 통해 1134개의 feature를 50개의 새로운 feature로 decoding 진행 후 위 새로운 feature들을 이용하여 K-Means clustering 및 t-SNE visualization을 통하여 5000개의 subgroup model 생성. 이때 model evaluation 및 추후 filtering을 위하여 추론된 subgroup 기반으로 support vector machine (SVM), silhouette algorithm 적용하였으며 각 algorithm 별 score를 계산하였다. 각각의 model들에 대해서 k-Means clustering plot, group, model data, validation loss plot을 함께 도출하여 저장하였다.

1. Filtering & Model selection

생성된 모델 5000개 중 subgrouping의 신뢰도가 떨어지는 model을 filter하기 위해, Silhouette score > 0.5 / SVM\_anova score > 0.9 / SVM\_random\_forest score > 0.9, 또한 각 조건에 대하여 overfitting을 피하기 위해 각 score가 1인 model 또한 exclusion criteria에 포함시켰다.

Filtering 단계 이후, 실제로 DEA\_subgroup에 사용할 model들을 선정하기 위하여 실제 사용한 immune-related feature들이 survival rate에 영향을 미치는 model을 select하는 과정을 진행한다. Survival 관련 데이터는 UCSC Xena (<https://xenabrowser.net/>)를 통해 얻었으며, 분석은 R program의 “survival” package를 통해 진행하였다. 각 그룹 별로 Kaplan-Meier 생존 분석을 진행하여 “ggsurvplot” function을 통하여 생존 결과 분석에 대한 시각화 자료를 저장하며, 생존 분석을 통해 나온 p-value 값을 통하여 OS에 차이가 있는지 없는지 판별한다. (p-value < 0.05). 위 과정을 통해 선택된 group model들은 이후 DEA\_subgroup 분석에 이용된다.

1. DEA\_subgroup

Significant different OS model에서 실제로 OS에 영향을 줄 것으로 예상되는 candidate Immune-related target gene을 찾기 위하여 위의 model들을 이용하여 DEA를 진행한다. DEA과정에서 gene expression data는 UCSC Xena에서 제공하는 expected count data를 이용하였으며, group data는 select한 model들의 group data를 이용하였다. Group은 binary group을 사용하며, case group은 OS validation 과정에서 OS가 낮은 그룹, control group은 OS가 높은 그룹이다. 즉 DEA\_subgroup (Differential expression analysis)을 통하여 immune-related하게 OS에 영향을 미치는 candidate target을 찾는 작업을 진행한다. 사용하는 gene은 처음 60,498개의 gene에서 filtering 과정을 거친다. Exclusion criteria; 1) 모든 sample들에서의 count 총 합이 10 미만인 gene들, 2) HGNC gene symbol이 존재하지 않는 gene들, 3) sample들 중 20% 이상의 sample이 gene count가 0인 gene들. 이러한 filtering을 거친 gene들에 대하여 DEA\_subgroup을 진행하며 DEA는 R program의 “DESeq2” package를 이용하여 진행한다. 모든 OS significant model에 대하여 DEA\_subgroup을 진행하였다. 다음 스텝으로overlapping 과정에서 모든 DEA result를 이용하여 overlapping 되는 gene들을 도출하였다. 각 result에서 p-value를 확인하여 모든 DEA result에서 100% overlapping significant한 gene을 얻어 이를 primary candidate target으로 삼았다.

1. DEA\_NT/TP

DEA\_subgroup을 통해 나온 primary candidate target들에 대하여 DEA를 진행하였다. 여기에서의 DEA는 TPM 형태의 TCGA gene expression을 이용하여 실제로 암 환자군에서 정상군과 비교해 많은 발현량을 가지는지 확인하는 1차 filtering 작업이다. 암 환자군 내에서 비록 OS를 떨어뜨리는 gene일지라도 실제로 암 환자군과 정상군에서 발현량의 차이가 거의 없는 gene은 우리 연구에서의 candidate target의 조건에 부합하지 않는다. 따라서 DEA\_NT/TP를 통하여 암환자 그룹에서 실제로 더 많이 발현하는 gene들을 filter한다. DEA\_NT/TP의 경우 R의 “TCGABiolinks”와 “DESeq2” package를 이용하여 분석 진행하였으며, “TCGABiolinks”로 solid tissue normal sample과 primary tumor sample의 gene expression 값을 각각 획득하였으며, “DESeq2”로 DEA를 진행하였다. DEA 이후 significant 기준은 p\_adj < 0.1이며, tumor sample에서 높은 gene만 획득하기 위하여 log2FoldChange값이 0 이상인 gene들만 filter하였다.

1. Candidate target validation

1차 filtering까지 마친 candidate target들에 대하여 신뢰성을 획득하기 위하여 여러 validation 작업을 진행한다. 첫 번째 validation으로는 gene localization 확인이다. 추후의 실험을 위해서 candidate target gene들은 membrane에 존재해야 한다는 조건이 있으며, 이를 확인하기 위하여 genecard를 이용하여 (<https://www.genecards.org/>), Uniprot/ HPA (Human Protein Atlas) / GO (Gene Ontology)/ COMPARTMENTS 등의 gene localization과 연관된 database에 접근하여 gene localization을 확인하였다. 만약 gene localization이 membrane에 존재하지 않는다면, filtering을 통해 candidate에서 제거된다.

Gene localization을 확인한 후 다음으로는 gene abundance 값을 확인한다. 이때 gene abundance란 TCGA sample에서 tumor sample들에 대하여 해당 gene의 TPM값을 확인한 후 각 gene의 TPM이 0.5가 넘는 sample들의 비율을 말한다. (해당 gene TPM > 0.5 samples / 전체 samples) 해당 기준은 Expression atlas (<https://www.ebi.ac.uk/gxa/home>)에서의 기준을 이용하였으며, gene expression data는 UCSC XENA에서 각 암종에 대한 TOIL-RSEM TPM data를 수집하여 이용하였다. 각 암종에서 candidate target gene들의 gene expression data를 다운받아 log2 transformed 되어있는 형태를 raw TPM 형태로 바꾸어 준 후 gene abundance를 확인 및 visualization을 진행한다.

3번째 validation 작업으로는 text-mining 작업을 진행한다. 여기서 말하는 textmining이란 Pubtator (<https://www.ncbi.nlm.nih.gov/research/pubtator/>)의 API를 이용하여 candidate target gene과 해당 분석 암종 또는 전암종의 keyword를 이용하여 두개의 조합이 pubmed에 올라와 있는 논문들을 검색하여 confidence, support, lift, count등을 계산하였으며 filtering 기준은 count를 기준으로 5 이상의 gene들을 선택하였다. 마지막으로 DGIdb (<https://www.dgidb.org>)를 이용하여 실제로 위 gene을 target 또는 biomarker로써 개발된 약물이 있는지 확인하여 너무 많은 약이 개발된 candidate target에서 exclude 되었다.