Prognosis prediction on pediatric neuroblastoma

Submit your answer (in notebook or PDF) along with your code (in notebook or scripts) in a zip file and submit to Coursework by Nov 7th, 23:59 PM.

Neuroblastoma (NBL) is a cancer that arises in immature nerve cells of the sympathetic nervous system, primarily affecting infants and children. In this datathon, we will analyze the genomic data of NBL samples from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) project to identify predictive mutation / gene expression features that could predict risk of patients, potentially informing the treatment strategy.

I have splitted the original data set into a training set and a test set. We will perform exploratory analyses on the training set, build a machine learning model, and test its performance on the test set.

PART 1: Identify risk-associated mutations (13 pts)

You can find the TARGET NBL somatic mutation data in MAF format on our share drive:

TARGET_NBL_WXS_somatic_verified.maf.txt

This is similar to what we analyzed in the class. You can find the patient ID in the `Case_USI` column, e.g. `TARGET-30-PANYGR`. Use this data table to answer the following questions:

1.1. What is the most mutated gene in the data set? The most mutated gene is defined as the gene that has at least one mutation in the most number of patients. (3 pts)

1.2. What is the most prevalent mutation? Similarly, this is defined as the mutation (not gene, but a unique mutation) observed in the most number of patients. (2 pts)

Hint: A unique mutation is defined the same as a unique variant. We try to match identical variants in HW1.

Next, we would like to investigate if the mutations in pediatric NBL are observed in other cancer dataset. Cancer hotspots is a database collecting frequently observed mutations from public cancer datasets. You can load the database as a file here:

cancerhotspots.v2.light.maf.gz

Note the alternate alleles in the MAF file can generally be found in the column `Tumor_Seq_Allele2`.

Now let's check how many mutations in TARGET NBL datasets are also hotspots observed in other datasets.

1.3. How many unique mutations in the TARGET NBL mutation file have been recorded in the cancer hotspots database? (3 pts)

Next, let's investigate the mutation information with the clinical information. You can find the clinical data of the data set here:

nbl_target_clinical.txt

You can find the patient ID in the `` `TARGET USI` `` column. There is a `` `5yr_efs` `` indicating whether the patient belongs to the `` `high risk` `` group (progression, relapse, or death within the 5 years of the genomic profiling) or the `` `low risk` `` (no event for more than 5 years). There is also a `` `train_test` `` column indicating this patient belongs to the training set (`` `train` ``) or the test set (`` `test` ``).

You will also notice there are `in_rnaseq` and `in_maf` columns indicating whether the patient has rnaseq data or has mutation data. You'll find the number of patients having both rnaseq and mutation data is very low (N=53). Hence we can only build a model using rnaseq data only later in the analysis, instead of both mutation and rnaseq.

For now, let's focus on the full set of mutation data. Combining the `` `5yr_efs` `` information of the patients, we'll investigate if the most mutated gene has association with high risk of disease development events.

Out of the 103 unique patients in the mutation MAF file (try to count it yourself!), subset the clinical information table to contain these 103 patients only. Create an additional column to indicate whether the patient has at least one mutation in the top-mutated gene (let's name it TMG here, you'll use the real symbol when you do it) in 1.1. Then create a count table as the following by combining the mutation info with the info from the `` `5yr_efs` `` column:

	High risk	Low risk
TMG mutated	c11	c12
TMG not mutated	c21	c22

(c11, c12, c21, c22) are patient counts that satisfy each condition. The sum of c11 + c12 + c21 + c22 should be 103 (we will not consider the train / test split for this part of the analysis). Once you have this table, you should also perform a statistical test to see if the TMG mutation is associated with risk. You can use either Fisher's exact test or the chi-square test for this.

1.4. Produce a patient count table as described above, perform a statistical test to check if the top-mutated gene mutation has significant association (P value < 0.05) with 5-year disease event risk (5 pts).

PART 2: NBL gene expression profiles difference between high-risk and low-risk patients (10 pts)

We are now moving into gene expression raw count data. You can find the training set RNASeq raw count data here:

nbl_target_counts.train.txt

Note that in the TARGET RNASeq table, like TCGA, each column is a sample name, whose first 16-character prefix indicates the patient ID. For example, the first sample in the table has an ID `TARGET-30-PAPVRN-01A`, meaning it's the tumor sample from patient `TARGET-30-PAPVRN`. You can use the prefixes to find whether the patient belongs to the high risk group or the low risk group.

2.1. Perform unsupervised learning on the expression data and visualize the result (*e.g.* MDS, PCA, hierarchical clustering, choose your favorite, but note if using PCA or hierarchical clustering, the expression value should be normalized), color the samples by risk group (high risk vs low risk). Describe if you see samples in different risk group can be separated in your plot. (2 pts)

2.2. Perform differential expression analysis between high risk group vs low risk group. List the top 10 genes over-expressed in the high risk group, and do a box plot showing the distribution of top over-expressed gene between two groups (5 pts)

2.3. Perform reactome pathway analysis between high risk group versus the low risk group. Report your result. (3 pts)

Now you should export all the significantly differentially expressed genes in your analysis in 2.2. (everything with FDR < 0.05) to be used in PART 3 to build a model. Also, you should perform the same normalization as you did on training set on the test set RNASeq data, which can be found here:

nbl_target_counts.test.txt

You can either perform trimmed mean of M-values (TMM) normalization on concatenated training and test set together, or you can separately normalize the test set.

Save the TMM-normalized Log2 CPM values of RNASeq training set, RNASeq test set, and the top differentially expressed genes (found using ONLY TRAINING SET) as files to be used by Colaboratory in PART 3.

PART 3: Predictive modeling using RNAseq for disease risk prediction (13 pts)

3.1. Build a machine learning model using RNASeq training set. Only input the significantly differentially expressed genes you found in 2.2. as features. Report what the top model is based on cross validation by autoML and what is its validation score (5 pts).

3.2. Evaluate your trained model on the TMM-normalized RNASeq test set (4 pts). Report accuracy, balanced accuracy, and F1 score (2 pts). Also produce a confusion matrix (2 pts).

BONUS (4 pts)

This question involves looking into cell line data in Module 3.

We are going back to the mutation data. Now we look at the cell lines from CCLE:

sample_info.csv
CCLE_mutations.csv
CRISPR_gene_effect.csv

We would like to find the neuroblastoma cell lines. You can identify the cancer type of the cell line in the `primary_disease` column of the sample info table. Once we identify the neuroblastoma cell lines, try to create a waterfall plot like we did in 2022-10-24's class using the CRISPR data targeting the TMG gene in 1.1. Highlight the cell lines that have TMG mutation, and comment if you see the TMG-mutated cell lines are dependent on TMG, meaning if most TMG-mutated neuroblastoma cell lines are concentrated to the lower-score end of the plot.

Identify neuroblastoma cell lines in CCLE. Create a waterfall plot as in the 2022-10-24 lecture using the NBL cell lines. Highlight the TMG (top mutated gene from 1.1.) mutated cell lines and comment if they have high dependency on TMG (4 pts).