NetID: au198
Aizhan Uteubayeva

Analysis Logic
Homework 6: DNA copy number analysis using R
HIDS-7003-01

Overview:

- Clinical data on Hepatocellular Carcinoma (HCC/Liver cancer):
    - 198 patients in total // rows. Confirmed de-identification
    - 104 patients with HCC
    - 94 patients who are "NON_TUMOR_CIRRHOTIC" (pre-cancer stage)
- DNA copy number data in the form of CINdex values (Chromosome instability index values) on 197 patients. Processed and cleaned. Data is NOT in Log2 Scale.

Goal:

The goal is to perform a group comparison analysis on the copy number data to compare the patients that had Vascular Invasion vs. those that did not have Vascular Invasion. So that means VASCULAR INVASION = Yes (comparison group) vs. VASCULAR INVASION = No (baseline group)

Clinical data:

- 198 rows that correspond to 198 patients
- Unique ID = “BiospecimenID_copyNumber”
- The target attribute is the “SAMPLE_TYPE” = HEPATOCELLULAR_CARCINOMA
- Then we filter out the “HEPATOCELLULAR_CARCINOMA” = YES/NO
- The subset of clinical data should consist of 104 patients, and the patient IDs are unique

Gene expression file:

- There are 197 columns// patients. Processed and cleaned. Not log2 data.
- IDs match
- The cytobands file will be used to subset the patients of interest

Sanity check:

The data needs to match, be a numeric matrix and have no extra rows// junk.