Cancer Research 2020

# Two Data files

1. **Clinical\_sample\_data**
   1. Description of the samples (i.e. patients).
      1. Count of patients: 444
   2. Each SAMPLE\_ID (string) corresponds to a unique PATIENT\_ID (number).
   3. Meanwhile, we focus only on the TISSUE\_SITE information of a patient.
   4. TISSUE\_SITE is where a cancer sample is collected, which can be considered as the ground truth that a diagnostic test aims to match.
      1. Possible values (and their count)
         1. **LN (167), i.e. lymph node**
         2. **Bone (160)**
         3. **Liver (64)**
         4. **Prostate (12)**
         5. **Lung (7)**
         6. Other Soft tissue (29)
         7. Adrenal (2)
         8. Brain (1)
         9. Unknown (2)
      2. In this study, let’s focus ONLY on the top 5 types of cancer.
2. **data\_mRNA**
   1. Columns
      1. SAMPLE\_IDs (string) are as given in the clinical sample data above.
      2. 270 SAMPLE\_IDs (i.e. from B to JK in Excel file)
         1. \*\*\*SZ: please check whether all of these 270 are included in the 444 patients in the clinical sample data?
         2. \*\*\*SZ: find out how many of these 270 patients are of the top 5 types of cancer? Let’s denote this number as NP
         3. \*\*\*SZ: please **create a new, simplified Clinical\_sample-data file** that contains only the NP patients
            1. where, you replace each SAMPLE\_ID with a new reference number (from 0 to NP-1) so that the first group of patients are LN cancer patients, the next group are bone cancer, … (such that the 5 groups are in the order as listed above).
   2. Rows
      1. Hugo\_Symbol is the name of an mRNA
      2. 19,293 Hugo\_Symbols (i.e. from 2 to 19,294 in Excel file)
   3. Values
      1. Measured values of mRNAs of a patient
      2. Note that “0” means the measured value is zero.
      3. \*\*\* SZ: to simplify the calculation, please **create a new mRNA file**, which only contains the NP patients and each mRNA value is cut short to 3 digits after the decimal point.
3. Research plan: two stages
   1. Stage ONE
      1. Calculate Pearson coefficient between patients
         1. \*\*\* SZ: please do all the followings in this section
         2. To calculate the pearson coefficient of the mRNA values between every pair of patients.
         3. Number of pairs: NP\*(NP-1)/2
         4. To shorten the coefficient value to two digits after the decimal point (i.e. each value is a % value).
         5. Distribution for all patients
            1. To plot the distribution function of the coefficient values

X: -1.00 -> 1.00

Y: % of the pairs with the value X

* + - * 1. To plot the cumulative distribution function of the coefficient values.

X: -1.00 -> 1.00

Y: % of the pairs with values <= X

* + - * 1. 3-D distribution

X: patient reference number 0 -> (NP-1)

Please draw lines to separate different cancer types

Y: patient reference number 0 -> (NP-1)

Please draw lines to separate different cancer types

Z: heatmap where red colour is Pearson coefficient 1, white colour is value 0 and blue is -1.

* + - * 1. Table - summary of the 3D distribution

X: 5 different types of cancer

Y: 5 different types of cancer

Content: average of pearson coefficient of pairs of patients, one of which is of cancer type X and the other is of the Type Y.

* + - 1. \*\*\* SZ: your work should stop here because our next step depends on your results above.
      2. Other thoughts
         1. To normalise values of a mRNA?

E.g. x=(x - x\_min)/(x\_max - x\_min)?

* + - * 1. To replace values of a mRNA with ranks of the values?

E.g. x’ = 100 if x = x\_max; and x’ = 1 if x = x\_min (=0) ?

* + 1. Create patient network
       1. Based on the above distributions, we need to choose a few thresholds and then create corresponding networks showing the connections among patients based on their Pearson coefficient values.
    2. Analyse the community structure of patient network
       1. We need to analyse their community structures, and then compare against the ground truth to see whether each community corresponds to a particular cancer type.
       2. This can be done by Gephi
       3. Visualisation is important for this analysis
       4. \*\*\* SZ: please study the definition of network communities and the concept of modularity using the slides of COMP0123.
       5. \*\*\* SZ: please study how to use Gephi to do community detection in a network.
  1. Stage TWO - network graph among mRNA
     1. Similar procedure as above, but with much more complicated implications. To be thought above later.