FINDING COMMON MUTATIONS FOR CANCER TREATMENT

KAIMRC training Problem Solution

Shiekhah AL Binali
Shiekhah.albinali@gmail.com

Contents

Figure List

Introduction

Cancer is defined as an uncontrollable cellular proliferation; after scientists found the origin of the phenotypic expression is the DNA, research goals were focused to understand the involvement of gene expression and mutation in cancer development [1]. This disease has many contributed factors and untold clinical manifestations; thus, Genetic tools are used to evaluate cancer arise systematically; researchers have identified specific genes modification and determined how the gene changes can cause tumor growth in normal tissues. Cancer gene theory arranges a framework to interpret how hereditary and environmental factors contribute to cancers [1].

Germline mutation developed in the germ cells that construct sperm and oocytes and pass them vertically from generation to another[1]. If a change happened, a new variant is created. Small mutations can be detected using DNA sequencing while larger mutants need microscopy to be visualized[1]. Mutation can be categorized based on the amount of DNA sequence change, extensive mutations cause either insertion of a new DNA sequence or DNA sequence loss[1]. Mutation can change the gene function and affect protein structure; the change can happen directly by changing condons and change the amino acid[1]. Another way is to alter the RNA spliced way, where the splicing depends on the existence of short splice donor, splice acceptor and branch point[1].

The Concept of personalized medication has been used before to determine the suitable options for patients in case like determining the choice of antibiotics for bacterial pneumonia.[2] The goal for using personalized medicine on cancer patients is to increase efficiency of the treatment and decrease the toxicity by providing sufficient information to decide the appropriate time and type of the treatment.[2] Several techniques has been developed to define the genetic variability among patients and decode genetic signature to improve therapy.[2]

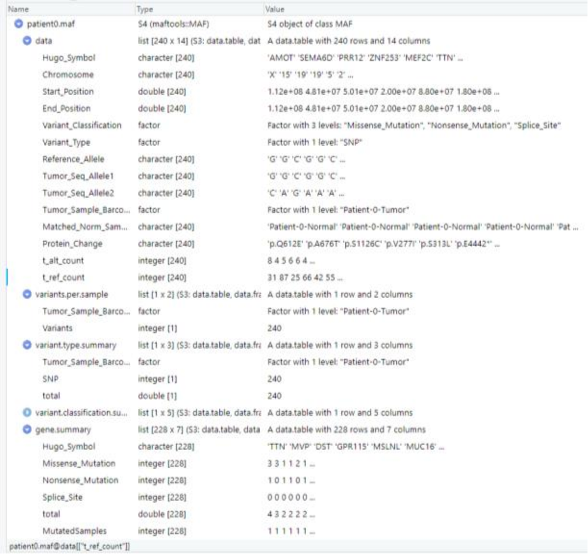
Having the data of patient and whether they respond or not to the treatment would contribute in enhancing the treatment, finding out the factors that are common in the responded group will give possible reasons about the reasons why the treatment did not benefit the non-respond group.

This report contains my attempt to solve a problem statement given by KAIMRC. The report starts with an introduction, followed by methodology section which describe the methods used to solve the problem statement, a discussion about the results is presented is also present. And finally, the conclusion section which has a brief description of the solution and the suggested future work.

Solution

To be able to solve the problems statement, which is examining genomic data for 50 treated patients with only half of the patient who respond to the treatment. A brief description of the data is required. The data contain 50 files for 50 patients and sample information file. The computer used for this problem statement has windows 10 operating system with an i7 core processor, 16 GB Ram, 1.5 TB storage and RStudio 4.0.3 is used.

At the begging, downloading, and reading all patients' files in RStudio to understand the data. Since all data is in MAF version, there is no need to convert the data or prepare it before the analysis. All files have main parts such as Data, Variants.per.sample which helps in handling the data. Figure 1 presents the content of one of the patients' files.

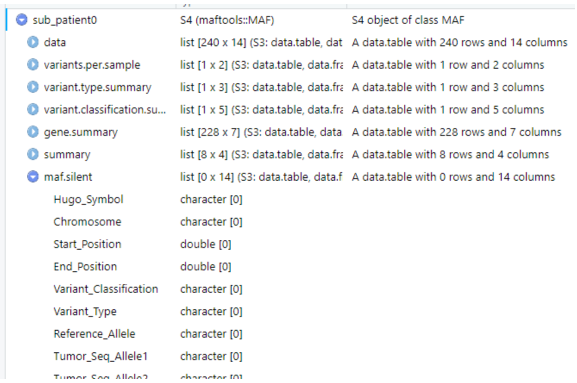


Name	Type	Value
patient0.maf	S4 (maftools::MAF)	S4 object of class MAF
data	list [240 x 14] (S3: data.table, dat	A data.table with 240 rows and 14 columns
Hugo_Symbol	character [240]	'AMOT' 'SEMA6D' 'PRR12' 'ZNF253' 'MEF2C' 'TTN' ...
Chromosome	character [240]	'X' '15' '19' '19' '5' '2' ...
Start_Position	double [240]	1.12e+08 4.81e+07 5.01e+07 2.00e+07 8.80e+07 1.80e+08 ...
End_Position	double [240]	1.12e+08 4.81e+07 5.01e+07 2.00e+07 8.80e+07 1.80e+08 ...
Variant_Classification	factor	Factor with 3 levels: "Missense_Mutation", "Nonsense_Mutation", "Splice_Site"
Variant_Type	factor	Factor with 1 level: "SNP"
Reference_Allele	character [240]	'G' 'G' 'C' 'G' 'G' 'C' ...
Tumor_Seq_Allele1	character [240]	'G' 'G' 'C' 'G' 'G' 'C' ...
Tumor_Seq_Allele2	character [240]	'C' 'A' 'G' 'A' 'A' 'A' ...
Tumor_Sample_Barco...	factor	Factor with 1 level: "Patient-0-Tumor"
Matched_Norm_Sam...	character [240]	'Patient-0-Normal' 'Patient-0-Normal' 'Patient-0-Normal' 'Patient-0-Normal' 'Pat ...
Protein_Change	character [240]	'p.Q612E' 'p.A676T' 'p.S1126C' 'p.V277I' 'p.S313L' 'p.E4442*' ...
t_alt_count	integer [240]	8 4 5 6 6 4 ...
t_ref_count	integer [240]	31 87 25 66 42 55 ...
variants.per.sample	list [1 x 2] (S3: data.table, data.fr	A data.table with 1 row and 2 columns
Tumor_Sample_Barco...	factor	Factor with 1 level: "Patient-0-Tumor"
Variants	integer [1]	240
variant.type.summary	list [1 x 3] (S3: data.table, data.fr	A data.table with 1 row and 3 columns
Tumor_Sample_Barco...	factor	Factor with 1 level: "Patient-0-Tumor"
SNP	integer [1]	240
total	double [1]	240
variant.classification.su...	list [1 x 5] (S3: data.table, data.fr	A data.table with 1 row and 5 columns
gene.summary	list [228 x 7] (S3: data.table, data	A data.table with 228 rows and 7 columns
Hugo_Symbol	character [228]	'TTN' 'MVP' 'DST' 'GPR115' 'MSLNL' 'MUC16' ...
Missense_Mutation	integer [228]	3 3 1 1 2 1 ...
Nonsense_Mutation	integer [228]	1 0 1 1 0 1 ...
Splice_Site	integer [228]	0 0 0 0 0 0 ...
total	double [228]	4 3 2 2 2 2 ...
MutatedSamples	integer [228]	1 1 1 1 1 1 ...

patient0.maf@data[["t_ref_count"]]

Figure 1 Content of patient 0 file

After exploring the patients' files, to focus on the "Nonsynonymous Mutation" type, subsetting the "Silent" type from each of the original patients' files. Which results in having 50 altered files.



sub_patient0	S4 (maftools::MAF)	S4 object of class MAF
data	list [240 x 14] (S3: data.table, dat	A data.table with 240 rows and 14 columns
variants.per.sample	list [1 x 2] (S3: data.table, data.fr	A data.table with 1 row and 2 columns
variant.type.summary	list [1 x 3] (S3: data.table, data.fr	A data.table with 1 row and 3 columns
variant.classification.su...	list [1 x 5] (S3: data.table, data.fr	A data.table with 1 row and 5 columns
gene.summary	list [228 x 7] (S3: data.table, data	A data.table with 228 rows and 7 columns
summary	list [8 x 4] (S3: data.table, data.fr	A data.table with 8 rows and 4 columns
maf.silent	list [0 x 14] (S3: data.table, data.f	A data.table with 0 rows and 14 columns
Hugo_Symbol	character [0]	
Chromosome	character [0]	
Start_Position	double [0]	
End_Position	double [0]	
Variant_Classification	character [0]	
Variant_Type	character [0]	
Reference_Allele	character [0]	
Tumor_Seq_Allele1	character [0]	
Tumor_Seq_Allele2	character [0]	

Figure 2 The data after subsetting the silent variants

After that, to find the 15 most frequent mutation, two possible solutions were applied. The first one is examining each patient file to find the top 15 mutation. And The other is determining the 15 frequent mutations for patients. For both ways. the frequent mutations were plotted and the plot for all patients is more accurate, thus the approach for each patient is discarded. Figure 3 shows the frequent mutation for a patient's data, and Figure 4 Shows the plot for frequent mutations for all patients.

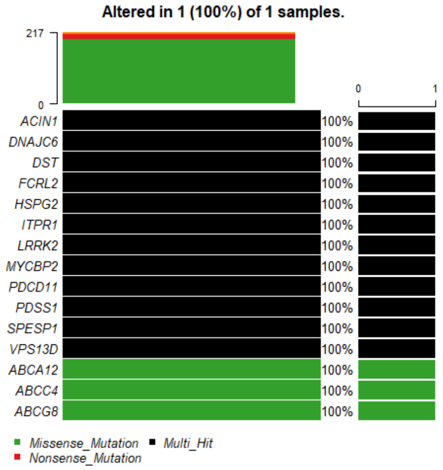


Figure 3 Frequent mutations for one patient

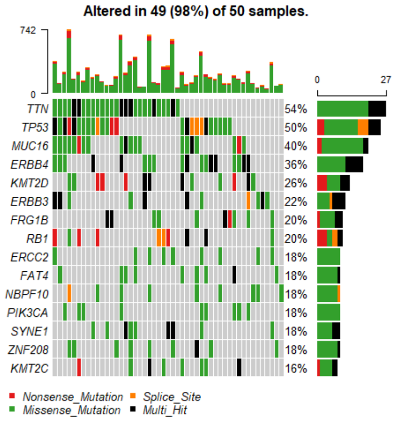
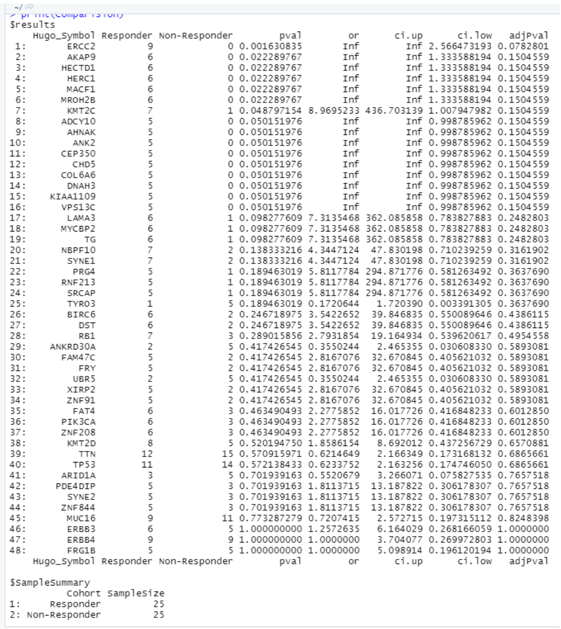


Figure 4 Frequent mutations for all patients

Next, the sample information is filtered based on response which leads to having two equally separate groups. Then based on the data for both groups, two MAF file are created for respond and non-respond groups. After that, mutation comparison between these two groups and the results is shown in figure 5.



```
~/
> print(comparison)
$results
    Hugo_Symbol Responder Non-Responder        pval       or      ci.up     ci.low   adjPval
 1:       ERCC2         9             0 0.001630835      Inf        Inf 2.566473193 0.0782801
 2:       AKAP9         6             0 0.022289767      Inf        Inf 1.333588194 0.1504559
 3:      HECTD1         6             0 0.022289767      Inf        Inf 1.333588194 0.1504559
 4:       HERC1         6             0 0.022289767      Inf        Inf 1.333588194 0.1504559
 5:       MACF1         6             0 0.022289767      Inf        Inf 1.333588194 0.1504559
 6:      MROH2B         6             0 0.022289767      Inf        Inf 1.333588194 0.1504559
 7:       KMT2C         7             1 0.048797154 8.9695233 436.703139 1.007947982 0.1504559
 8:      ADCY10         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
 9:       AHNAK         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
10:        ANK2         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
11:      CEP350         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
12:        CHD5         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
13:      COL6A6         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
14:       DNAH3         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
15:    KIAA1109         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
16:      VPS13C         5             0 0.050151976      Inf        Inf 0.998785962 0.1504559
17:       LAMA3         6             1 0.098277609 7.3135468 362.085858 0.783827883 0.2482803
18:      MYCBP2         6             1 0.098277609 7.3135468 362.085858 0.783827883 0.2482803
19:          TG         6             1 0.098277609 7.3135468 362.085858 0.783827883 0.2482803
20:      NBPF10         7             2 0.138333216 4.3447124  47.830198 0.710239259 0.3161902
21:       SYNE1         7             2 0.138333216 4.3447124  47.830198 0.710239259 0.3161902
22:        PRG4         5             1 0.189463019 5.8117784 294.871776 0.581263492 0.3637690
23:      RNF213         5             1 0.189463019 5.8117784 294.871776 0.581263492 0.3637690
24:       SRCAP         5             1 0.189463019 5.8117784 294.871776 0.581263492 0.3637690
25:       TYRO3         1             5 0.189463019 0.1720644   1.720390 0.003391305 0.3637690
26:       BIRC6         6             2 0.246718975 3.5422652  39.846835 0.550089646 0.4386115
27:         DST         6             2 0.246718975 3.5422652  39.846835 0.550089646 0.4386115
28:         RB1         7             3 0.289015856 2.7931854  19.164934 0.539620617 0.4954558
29:    ANKRD30A         2             5 0.417426545 0.3550244   2.465355 0.030608330 0.5893081
30:      FAM47C         5             2 0.417426545 2.8167076  32.670845 0.405621032 0.5893081
31:         FRY         5             2 0.417426545 2.8167076  32.670845 0.405621032 0.5893081
32:        UBR5         2             5 0.417426545 0.3550244   2.465355 0.030608330 0.5893081
33:       XIRP2         5             2 0.417426545 2.8167076  32.670845 0.405621032 0.5893081
34:       ZNF91         5             2 0.417426545 2.8167076  32.670845 0.405621032 0.5893081
35:        FAT4         6             3 0.463490493 2.2775852  16.017726 0.416848233 0.6012850
36:      PIK3CA         6             3 0.463490493 2.2775852  16.017726 0.416848233 0.6012850
37:      ZNF208         6             3 0.463490493 2.2775852  16.017726 0.416848233 0.6012850
38:       KMT2D         8             5 0.520194750 1.8586154   8.692012 0.437256729 0.6570881
39:         TTN        12            15 0.570915971 0.6214649   2.166349 0.173168132 0.6865661
40:        TP53        11            14 0.572138433 0.6233752   2.163256 0.174746050 0.6865661
41:      ARID1A         3             5 0.701939163 0.5520679   3.266071 0.075827535 0.7657518
42:     PDE4DIP         5             3 0.701939163 1.8113715  13.187822 0.306178307 0.7657518
43:       SYNE2         5             3 0.701939163 1.8113715  13.187822 0.306178307 0.7657518
44:      ZNF844         5             3 0.701939163 1.8113715  13.187822 0.306178307 0.7657518
45:       MUC16         9            11 0.773287279 0.7207415   2.572715 0.197315112 0.8248398
46:       ERBB3         6             5 1.000000000 1.2572635   6.164029 0.268166059 1.0000000
47:       ERBB4         9             9 1.000000000 1.0000000   3.704077 0.269972803 1.0000000
48:       FRG1B         5             5 1.000000000 1.0000000   5.098914 0.196120194 1.0000000
    Hugo_Symbol Responder Non-Responder        pval       or      ci.up     ci.low   adjPval

$SampleSummary
          Cohort SampleSize
1:     Responder         25
2: Non-Responder         25
```

Figure 5 The mutant comparison results for Respond/ Non-respond groups

After storing the comparison in a new file, presenting the comparison in a plot that shows the comparison results and the percent of cases. Figure 6 shows the comparison visualization.

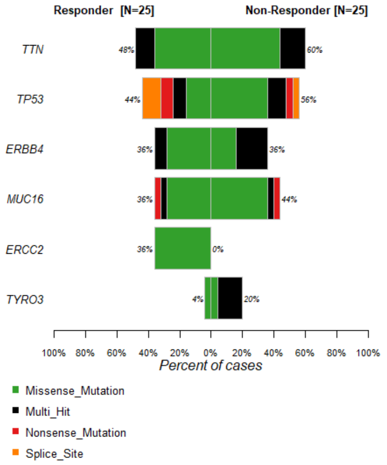


Figure 6 Mutation results in respond group compared to non-respond group

To differentiate between wild type and mutant samples, subsetting the data when the variant classification is “Missense Mutation” as mutant type and otherwise as wild type. After subsetting the samples, A plot is required to compare the results between the sample groups.

Conclusion

This report presents an attempt to solve a problem statement to find mutations associated with treatment response for 50 patients. Genetic information is helpful in developing precision medication which handles the starting period and the appropriate type of treatment most suitable for each patient. This field has improved due to contributions of previous research and having patients' data. After solving the problem statement, The Responded group has more mutation samples than non-responded group which leads to considering this discovery as one of the potential reasons for the non-respond group. For possible future work, analyzing the non-responded group and comparing the non-respond mutant with the responded group and observe the differences, find a way to decrease the number of required mutation samples to detect the cancer subtype since it is a possible cause for non-responding to the treatment. Another possible future work is focusing on a specific variant classification and observe the results for both responding groups.

References

[1] F. Bunz, *Principles of cancer genetics*. 2008. doi: 10.1007/978-1-4020-6784-6.

[2] *Bioinformatics in Cancer and Cancer Therapy*. 2009. doi: 10.1007/978-1-59745-576-3.