Sonia Savur

QBIO 490: Directed Research – Multi-Omic Analysis

Spring 2024

26 February 2024

R Review Project

Part 1: Review Questions

General Concepts

1. What is TCGA and why is it important?

TCGA is the Cancer Genome Atlas, a publicly available multi-omic dataset that can be used to explore the genomic data from a large sample of cancer patients. Its importance lies in its ability to associate clinical, proteomic, and/or genomic data with the presence of disease.

2. What are some strengths and weaknesses of TCGA?

Strengths of TCGA include its large sample size (data collected from ~20,000 patients), ease of use (i.e. clinical data can be analyzed as dataframes in R), diverse set of variables/data, and its real-world applications. Weaknesses of TCGA include the size of the package and how long it takes to download.

Coding Skills

1. What commands are used to save a file to your GitHub repository?

“git add <file>” tells git to track all local modifications done to <file>
“git push” pushes modified files in your local repository to your GitHub repository

2. What command(s) must be run in order to use a package in R?

install.packages() and library(), with the desired package passed as an argument in both functions.

3. What command(s) must be run in order to use a *Bioconductor* package in R?

```
install.packages("BiocManager")
library(BiocManager)
```

4. What is boolean indexing? What are some applications of it?

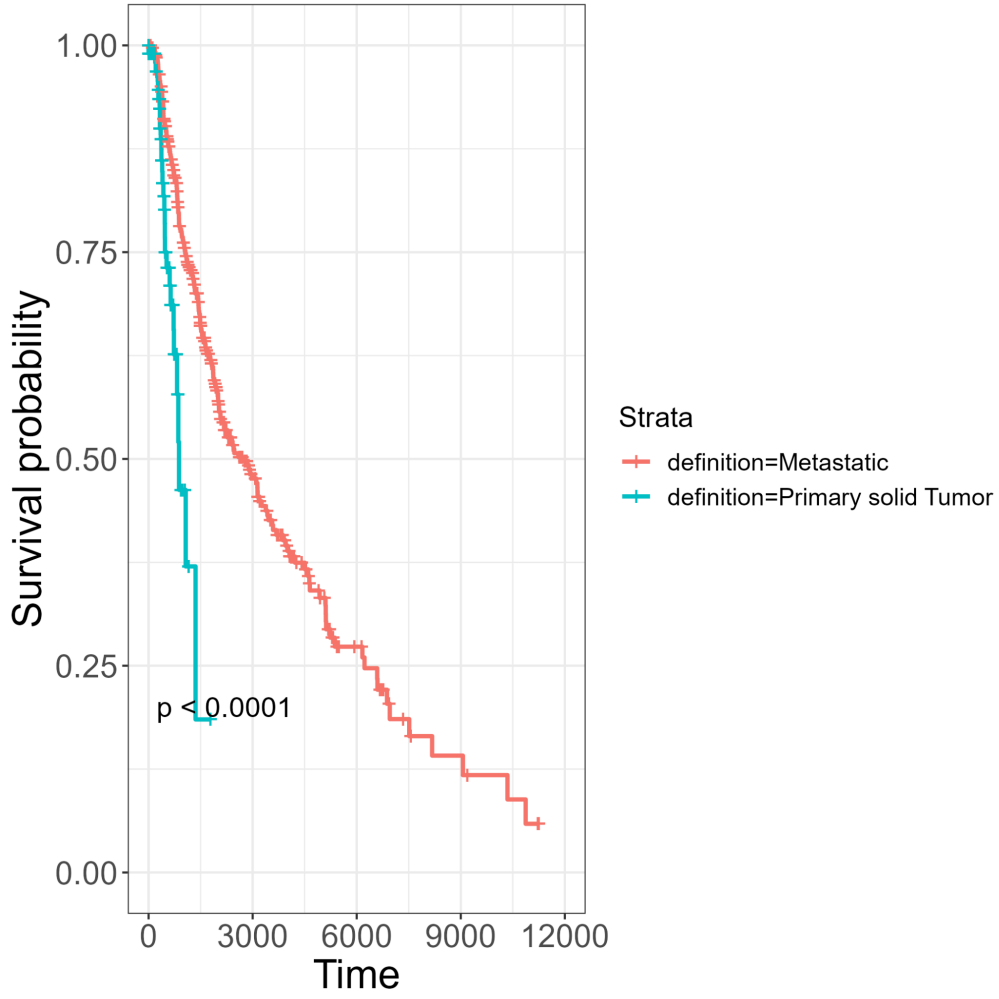
Boolean indexing is the process of applying a vector of boolean values (i.e. TRUE/FALSE) to a column/row in a dataframe. This helps with data dimensions reduction: the column/row originally could be holding 0+ different values, but with boolean indexing the column/row can be viewed as a set of either TRUEs (select the data) or FALSEs (ignore the data). Its biggest application is with data subsetting. Boolean indexing could be used to create boolean masks to effectively subset for a group of interest or to clean data of null values (i.e. NAs).

5. Draw a mock up (just a few rows and columns) of a sample dataframe. Show an example of the following and explain what each line of code does.
    a. an ifelse() statement
    b. boolean indexing

```{r part_1_dataframe}

# creating family member df
Name <- c("Sonia", "Sameer", "Amy", "Rishi")
Sex <- c("Female", "Male", "Female", "Male")
Age <- c(22, 60, 56, 29)
df <- data.frame(Name, Sex, Age)

# example of ifelse() statement and boolean indexing:

# create Boolean vector: TRUE if female, FALSE if not
female_mask <- ifelse(df$Sex == "Female", T, F)

# subset df by just female rows, i.e. indices for which female_mask is TRUE
female_df <- df[female_mask,]

```

Part 3: Results and Interpretations

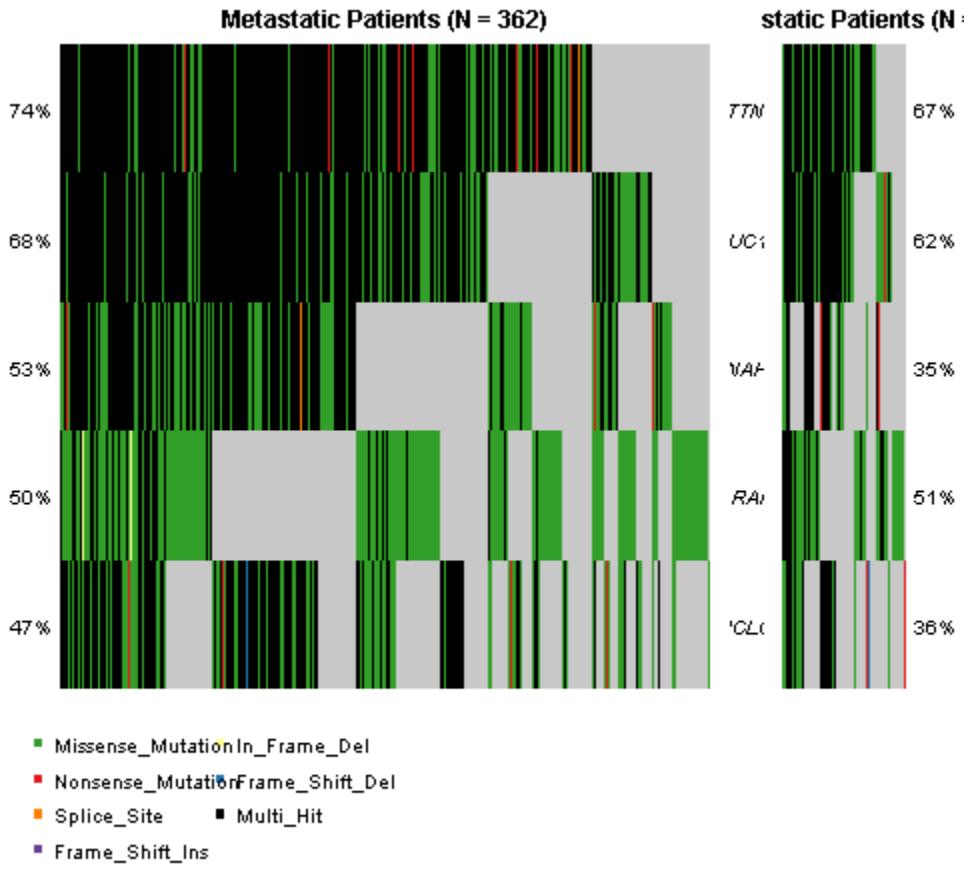
1) Difference in survival between metastatic and non-metastatic patients



The Kaplan-Meier plot (KM plot) above shows the probability of survival over time between metastatic and non-metastatic (i.e. primary solid tumor) patients. From the trajectories plotted above, it can be concluded that metastatic TCGA-SKCM patients have a higher probability of survival than non-metastatic TCGA-SKCM patients for the first ~1500 days. However, differences between survival probabilities between metastatic and non-metastatic patients after this time frame cannot be drawn since there is not enough data on survival probability after this time frame for the non-metastatic group. This could possibly be due to very high death rates after this time frame in non-metastatic patients. This conclusion is in opposition to the general consensus that metastatic cancers are more lethal than non-metastatic cancers (Chen et al, 2017). One possible reason that previously published work does not support this analysis could be because non-metastatic patients sought treatment only when their condition was extremely

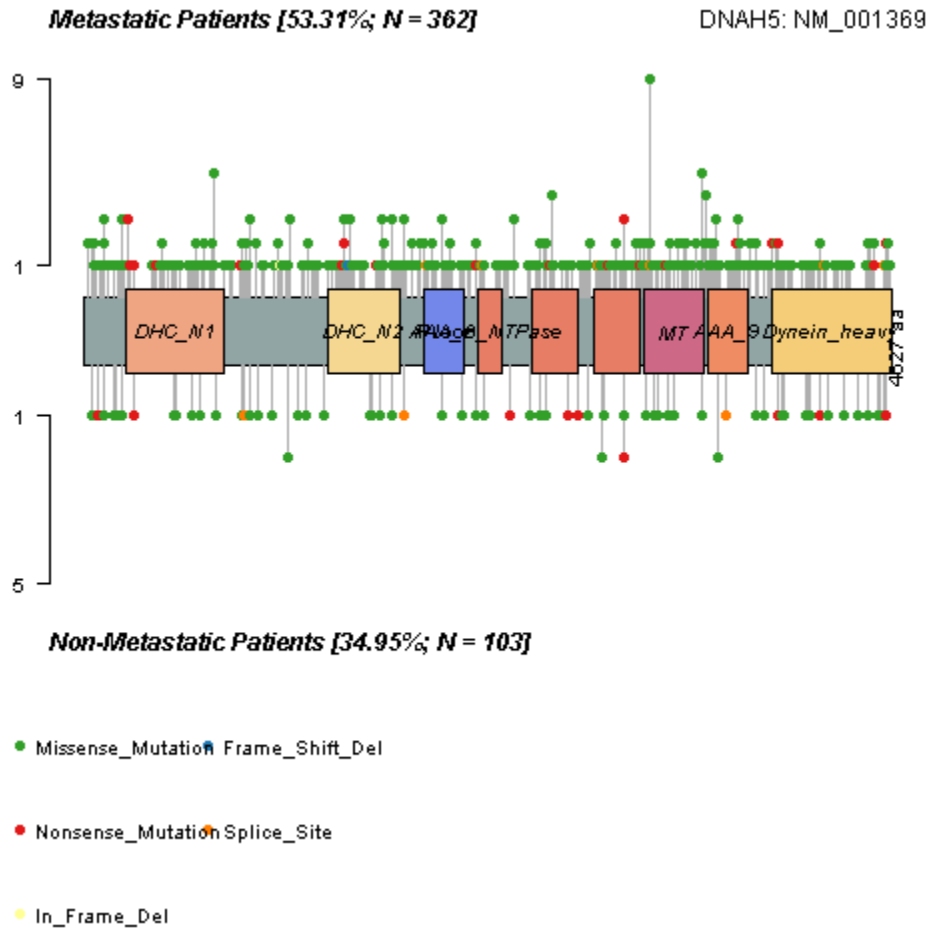
severe, and thus our data collected on non-metastatic patients could be skewed, as outlined in Dos-Anjos, et al (2017).

2) Mutation differences between metastatic and non-metastatic patients for multiple genes



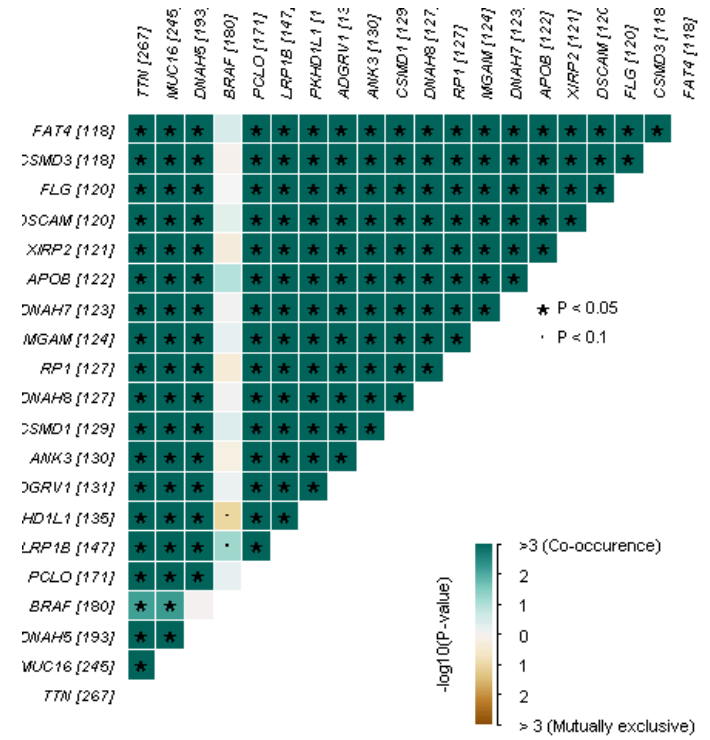
The co-oncoplot above shows the top 5 most commonly mutated genes in the metastatic group and in the non-metastatic group. From the mutation distribution above, it can be concluded that there are differences in mutation rates for multiple genes, namely *TTN*, *MUC16*, *DNAH5*, and *PCLO*. A significant difference in the proportion of mutation in the *BRAF* gene cannot be concluded since there is only a 1% difference in rate between the two groups. This finding is supported by scientific literature. The BRAF gene is an oncogene that, once mutated, can lead to the development of a tumor. Its relatively equal prominence between the two groups makes sense, since both groups have a cancer diagnosis and only differ in terms of metastasis, a process in which the *BRAF* gene is not directly related (Dhomen & Marais, 2007).

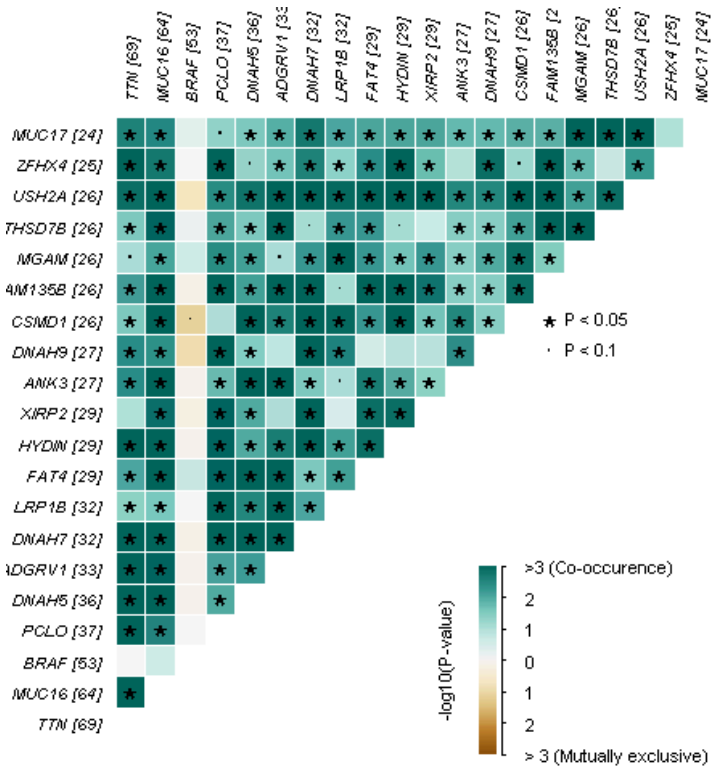
3) Mutation differences for specific genes of interest



The co-lollipop plot above displays the mutation distribution along the region of the gene of interest *DNAH5*. The axis is the location of amino acids, with the domains of the protein labeled, and mutations are marked based on type. The mutation markers above the axis represent the metastatic group of patients whereas the markers below the axis represent the non-metastatic group of patients. From this co-lollipop plot, it can be concluded that missense mutations are the most common type of mutation within both groups, and that mutations are relatively evenly spaced across the gene. The exact missense mutations (i.e. point nucleotide substitutions) cannot be concluded from the plot above, as it does not depict specifics (i.e. amino acid substitutions) about the mutations. One main conclusion that can be drawn from the plot is that there is a higher likelihood of mutations within the *DNAH5* gene in the metastatic group than the non-metastatic group. This is in accordance with findings from Li et al (2016) that mutations within the *DNAH5* gene are associated with oncogenic metastasis.

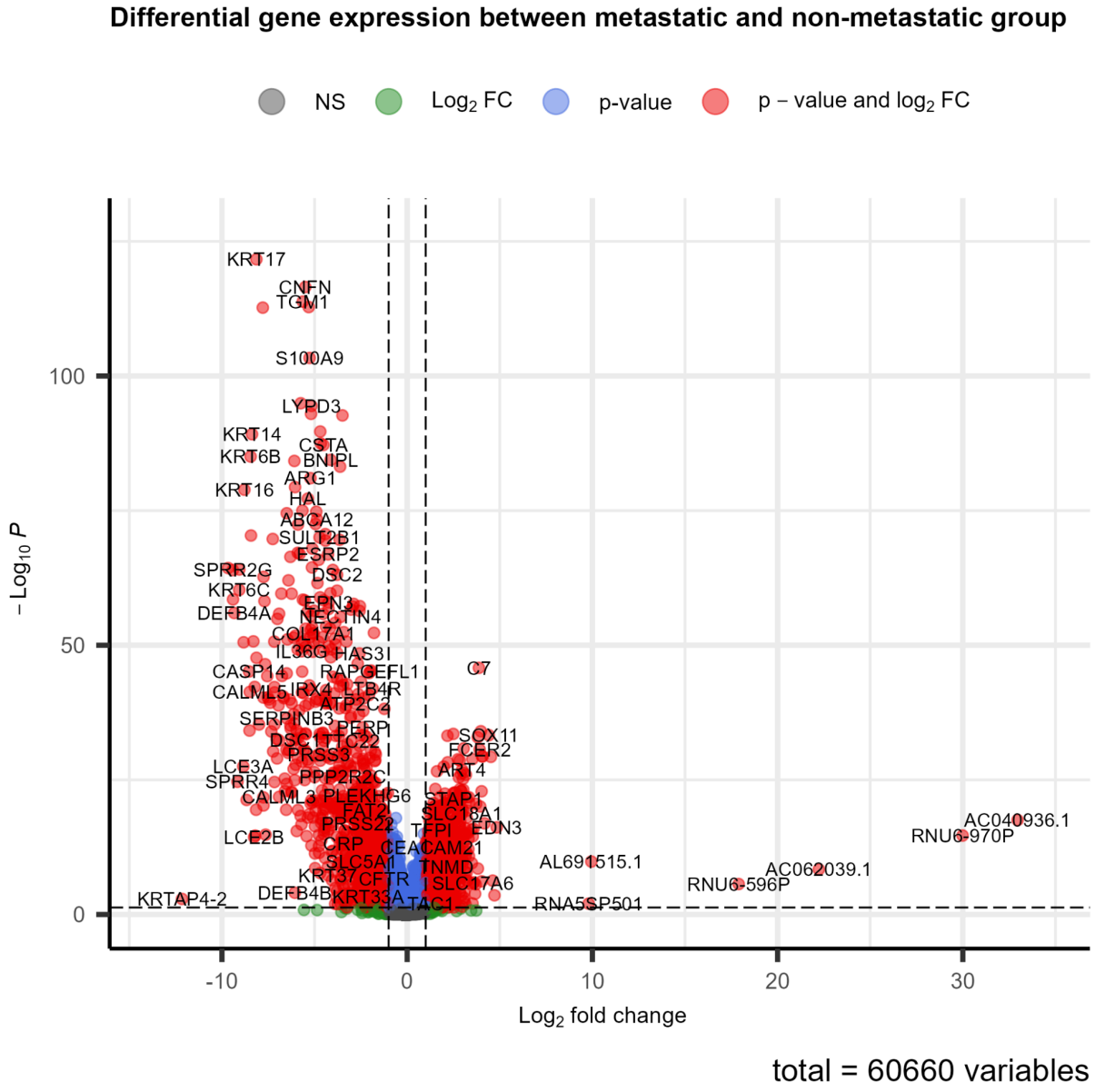
4) Co-occurrence or mutual exclusion of common gene mutations: one for metastatic patients, one for non-metastatic patients





The somatic interactions between the top 20 genes with mutations in the metastatic group (*top*) and the non-metastatic group (*bottom*) are shown above. Somatic interactions encompasses gene expression and gene mutations across the genomes of all of the patients with metastatic cancer to detect mutual exclusion or co-occurrence with other genes. As depicted in color and by the significance symbols, almost all of the genes are in co-occurrence with one another at a statistically significant level in both groups, except for *BRAF*. It is to be noted that there are more genes co-occurring at a more statistically significant level in the metastatic group than the non-metastatic group. In fact, the *BRAF* gene only co-occurs with one other gene at a statistically significant level, and it even exists in somewhat mutual exclusion with *HD1L1*. The graphs do not contain any information in relation to the other group of interest, and they do not inform types of mutations. Scientific literature has been published that is in accordance with this conclusion of *BRAF*'s tendency towards mutual exclusion, especially in comparison to co-occurring mutations in other genes (Sahin et al, 2013).

5) Differential expression between non-metastatic and metastatic patients controlling for treatment effects, race, gender, and vital status



The volcano plot above visualizes differential gene expression between metastatic and non-metastatic cancer patients, controlling for treatment effects, race, gender, and vital status. Fold change is calculated with metastatic patients as the numerator and non-metastatic patients as the denominator. From this plot it can be concluded that there are a multitude of genes that have differential expression in metastatic patients. Red points on the right of the log2FC = +1 threshold line indicate genes that are upregulated in metastatic patients, and red points on the left of the log2FC = -1 threshold line indicate genes that are downregulated in metastatic patients.

There are more genes that are differentially downregulated in metastatic patients than genes that are upregulated in metastatic patients compared to non-metastatic patients. The volcano plot provides an accessible visualization of statistically significant differential gene expression, but it does not provide information on the type of mutation, number of patients analyzed, nor rates or prevalence of mutation. The findings of the volcano plot are in accordance with scientific findings of differential gene expression between metastatic and non-metastatic cell lines, especially in regards to cytokeratins and collagens (Günes & Carlsen, 2003). Indeed, one of the groups of genes most differentially expressed between the two groups is the *KRT* group of genes that are associated with keratins; as seen in the plot above, it is specifically downregulated in metastatic patients compared to non-metastatic patients.

References

Chen, Meng-Ting, et al. “Comparison of patterns and prognosis among distant metastatic breast cancer patients by age groups: a SEER population-based analysis.” *Scientific reports* 7.1 (2017): 9254.

Dhomen, Nathalie, and Richard Marais. “New insight into BRAF mutations in cancer.” *Current opinion in genetics & development* 17.1 (2007): 31-39.

Dos-Anjos, Caroline S., et al. “Assessment of the integration between oncology and palliative care in advanced stage cancer patients.” *Supportive Care in Cancer* 25 (2017): 1837-1843.

Günes, H., and S. A. Carlsen. “Identification of differentially expressed genes in isogenic highly metastatic and poorly metastatic cell lines of R3230AC rat mammary adenocarcinoma.” *Cell proliferation* 36.6 (2003): 333-346.

Li, Fei, et al. “Identification of TRA2B-DNAH5 fusion as a novel oncogenic driver in human lung squamous cell carcinoma.” *Cell research* 26.10 (2016): 1149-1164.

Sahin, Ibrahim Halil, et al. “Rare though not mutually exclusive: a report of three cases of concomitant KRAS and BRAF mutation and a review of the literature.” *Journal of Cancer* 4.4 (2013): 320.