Christopher Kim
Pranav Pammidimukkala
Leon Zha

Investigating Correlations in Age, Metastasis, and *TP53* Expression in Colorectal Cancer Patients

Introduction

Colorectal cancer (CRC) occurs in the colon and rectum. It is the third most common cancer type and is associated with the fourth greatest number of cancer-related deaths (Mármol et al., 2017). Given the prevalence and potency of CRC, in this paper, we sought to investigate the correlations between age, metastasis, and *TP53* mutations in patients with CRC to better understand how each factor affects each other as well as overall survival.

Two databases were employed in this project: The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Created by the joint efforts of the National Cancer Institute (NCI) and National Human Genome Research Institute, TCGA is a public database containing 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data on 33 cancer types (https://www.cancer.gov/tcga). The second source of data is the Clinical Proteomic Tumor Analysis Consortium (CPTAC), another public source of data under the NCI. Unlike the multi-omic nature of TCGA, CPTAC specializes in proteomic data (https://proteomics.cancer.gov/programs/cptac).

We analyzed data from TCGA and CPTAC in order to examine correlations between age, metastatic status, and *TP53*. Age is a relevant topic of investigation as many studies have shown that age is a primary cancer risk factor (Balducci 2006). Not only is the biological likelihood of carcinogenesis higher, but exposure and accumulation of environmental carcinogens also increase with time. Furthermore, aging has also been shown to have correlation with metastasis. Metastasis is the spread of cancerous cells from their original location to other parts of the body, where they form new cancerous tumors of the same type as the original. Studies have found that metastasis status is becoming increasingly common among younger patients while incidences remain relatively constant among elderly patients (Zahir et al., 2022). *TP53,* commonly known as the “guardian of the genome,” is also relevant for cancer research. *TP53* codes for the protein p53, the main function of which is to suppress tumor growth. *TP53* mutations are observed in approximately 50% of CRC patients, in addition to being associated with lower survival rates in patients that receive chemotherapy.

Methods

The clinical, genomic, and transcriptomic data used in this project was obtained from TCGA, whereas proteomic data was drawn from the CPTAC. TCGA data were processed using R and comprised most of the analysis. In an R script, the TCGA data, as well as the package “TCGAbiolinks” were loaded. The package “TCGAbiolinks” and the TCGA data were loaded onto R, after which colorectal cancer data were accessed using the accession code “COAD.” Patients were categorized into two groups: "Young" if they were younger than 50 and "Old" if they were 50 or older. Metastatic status was also disaggregated into two groups: Stage I and Stage II patients were considered to have no metastasis, while Stage III and Stage IV patients were categorized as metastasized. After data was categorized based on age and metastatic status, five types of figures were generated:

i) A bar plot comparing age groups and metastatic status, with the data sourced from the clinical data from TCGA

ii) A box plot comparing age groups and *TP53* counts, with the gene counts data coming from the SummarizedExperiments object in TCGA

iii) A Kaplan-Meier Curve comparing the survival rates of "Young" and "Old" patients, sourced from the clinical object in TCGA

iv) Two oncoplots (one for each age category) comparing the top 5 most commonly mutated genes, sourced from the MAF object in TCGA

v) A lollipop plot comparing the number and physical location of mutations in *TP53* between “Young” and “Old” patients, with the data being sourced from the MAF object in TCGA

The final three figures included in this project were generated using Python. Using CPTAC data, spearman correlation heatmaps were created to compare the RNA expression and protein levels of the 4 most commonly mutated genes shared by both age groups for which protein data was available.

Results

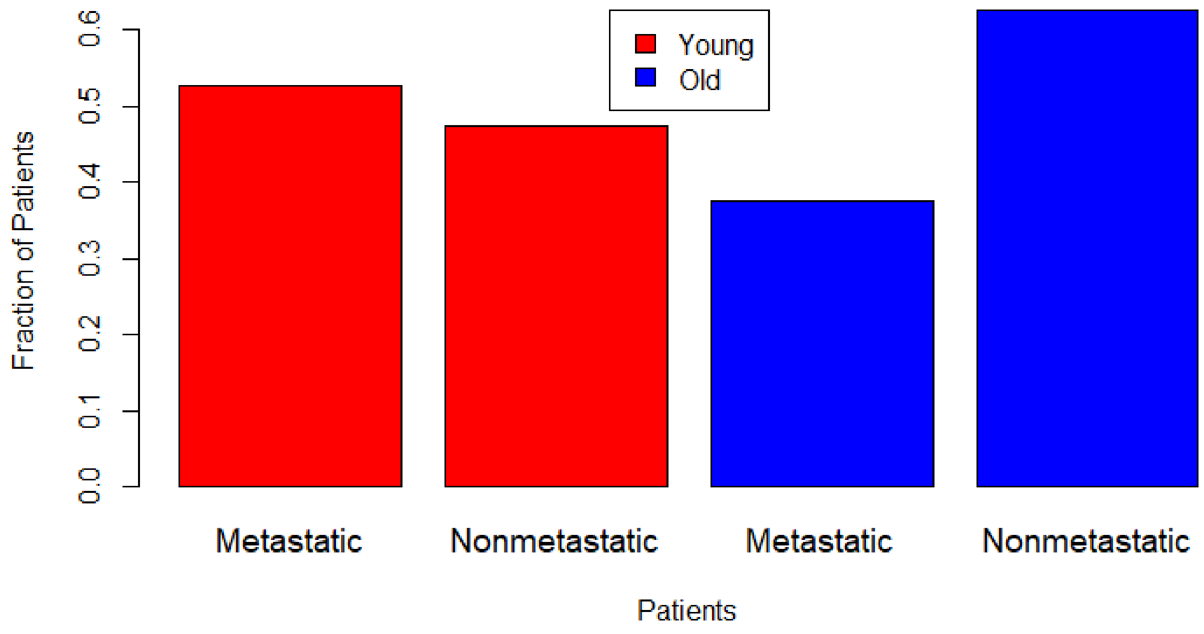

Figure 1. The proportion of patients with metastasis is higher for the "Young" category than the "Old" category.

Figure 1 is a bar plot depicting the metastatic status of "Young" and "Old" patients. Since there are many more "Old" patients than there are "Young" patients, the categories are presented in fractions of the total sample of "Young" and "Old" patients, respectively, as opposed to direct counts. It is clear that a greater percentage of "Young" patients have metastatic tumors than "Old" patients.

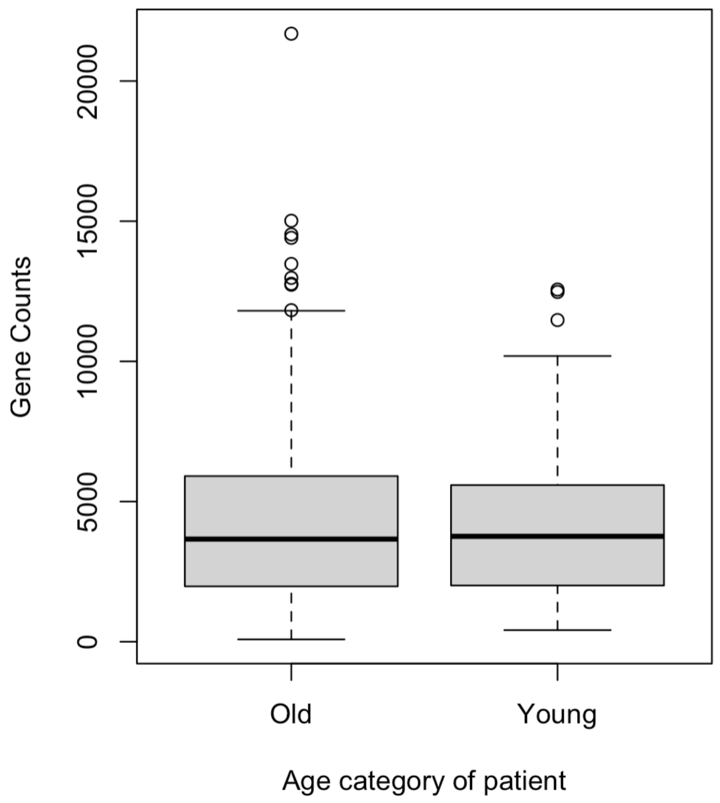

Figure 2. Boxplots comparing the distribution of TP53 gene counts between age categories. Similar means indicate no significant differences in mean TP53 *expression between patients in the Young and Old groups.*

Figure 2 compares the gene counts (i.e. expression) of *TP53* in the "Young" category (< 50yrs) and "Old" category (≥ 50yrs) patients. The minimum and lower quartile of the "Young" patients' are greater and their upper quartile and maximum are smaller than that of "Old" patients. In other words, the gene counts of "Young" patients have a smaller range or spread. However, the median gene counts of both age groups are almost identical – approximately 3000. This indicates that there is likely not a statistically significant difference in *TP53* expression levels between age categories.

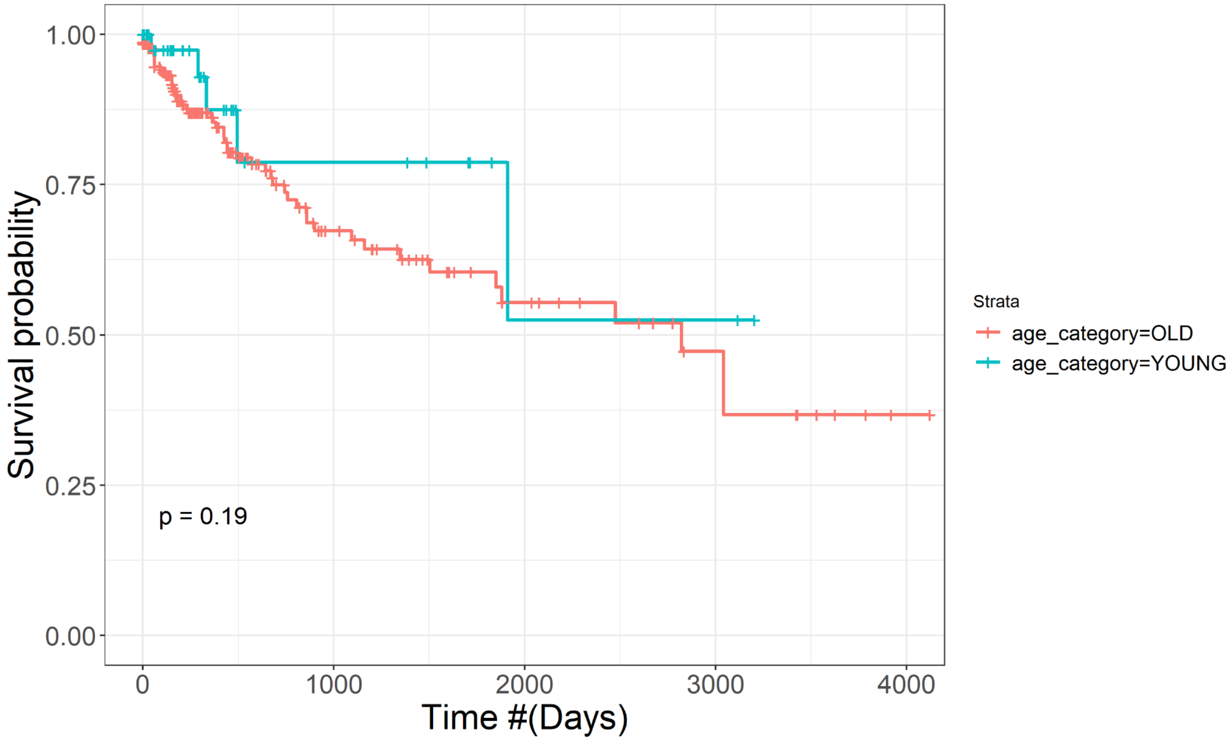

Figure 3. Kaplan-Meier survival curve showing that there are no statistically significant differences in survival probability over time found between the "Young" and "Old" groups.

Figure 3 is a Kaplan-Meier survival plot that compares the change in survival possibility with progression in time between "Young" and "Old" patients. "Young" patients have a greater probability of survival in most time points except for the period of time between approximately 1900 and 2500 days. The survival probability of "Young" and "Old" patients plateaus around 1900 days with an approximate survival probability of 0.52, and around 3000 days with an approximate survival probability of 0.37, respectively. However, as the p-value > 0.05, it's likely that there's no statistical significance between the survival probabilities of the two ages.

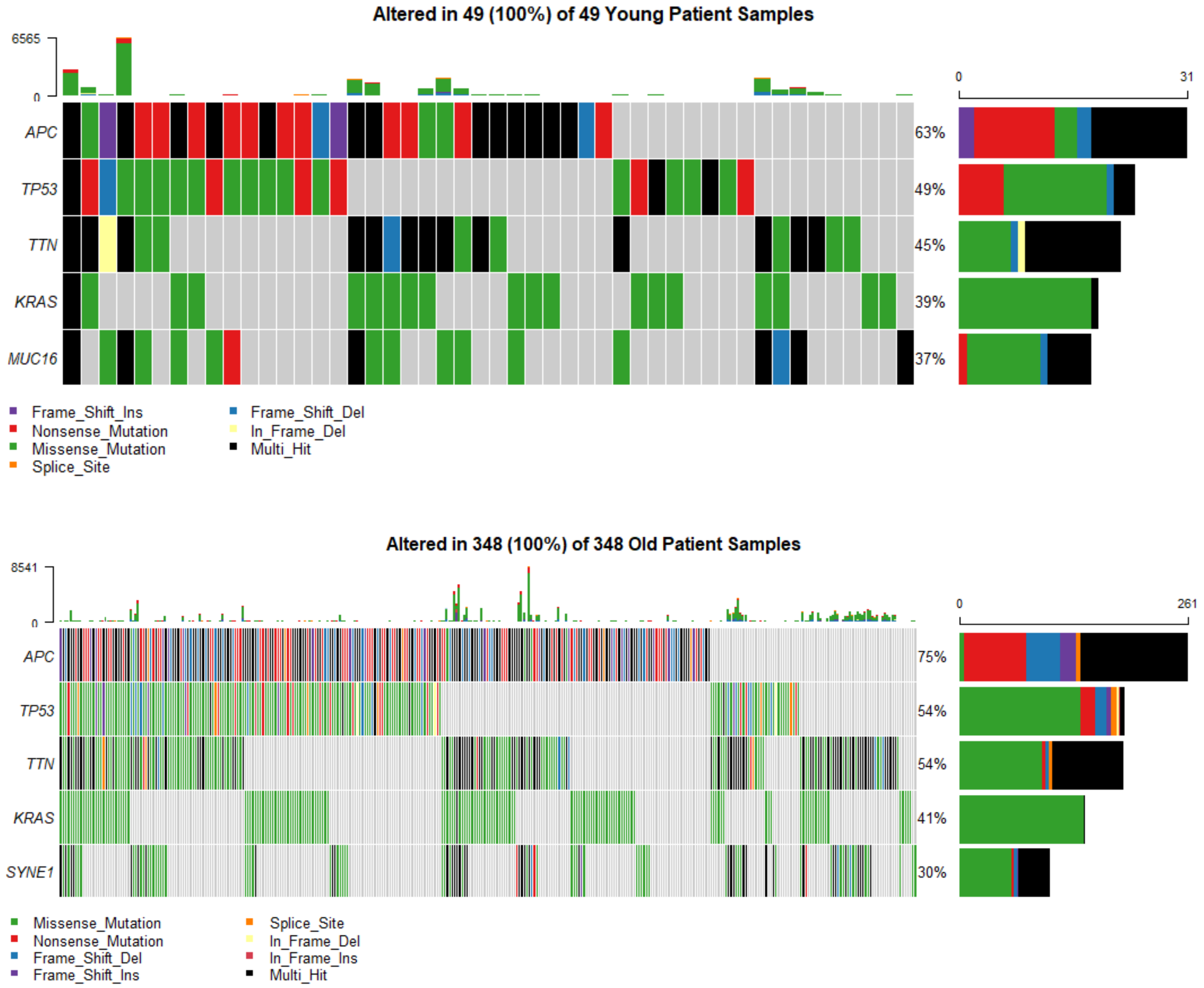

Figure 4. Oncoplots indicating that "Young" and "Old" patients share the top 4 most mutated genes, of which TP53 *is one. Along the x-axis are individual patients, while the y-axis indicates the genes.*

Figure *4* consists of two side-by-side oncoplots of the top 5 most frequently mutated genes in the "Young" and "Old" cohort, respectively. The top 4 most frequently mutated genes *APC, TP53,* *TTN,* and *KRAS*, are shared between the two age categories, while the fifth differed. *APC* mutations were by far the most varied, whereas the mutations for the other 5 genes were generally dominated by missense mutations. *TP53* was the second most commonly mutated genes in both groups, and is split into two groups: patients that exhibit *TP53* mutations and *APC* mutations, and those that only exhibit *TP53* mutations.

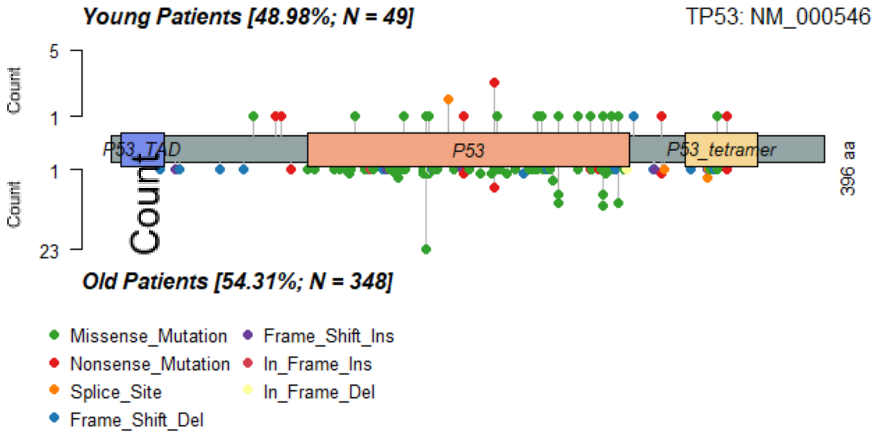

Figure 5. Lollipop plot of mutations in TP53 *for both age categories*

Figure *5* is a lollipop plot of *TP53* for both age categories. "Old" patients had a much greater number of unique mutations, specifically in the P53 region. The greatest frequencies of mutations were also greater in "Old" patients than in "Young" patients: the frequency of the highest occurring mutation in "Old" patients was 23 times, while that in "Young" patients was 5. However, the above results may be attributed to the different sizes of the two age categories, as one would expect the “Old” patients to have more mutations on account of having more data.

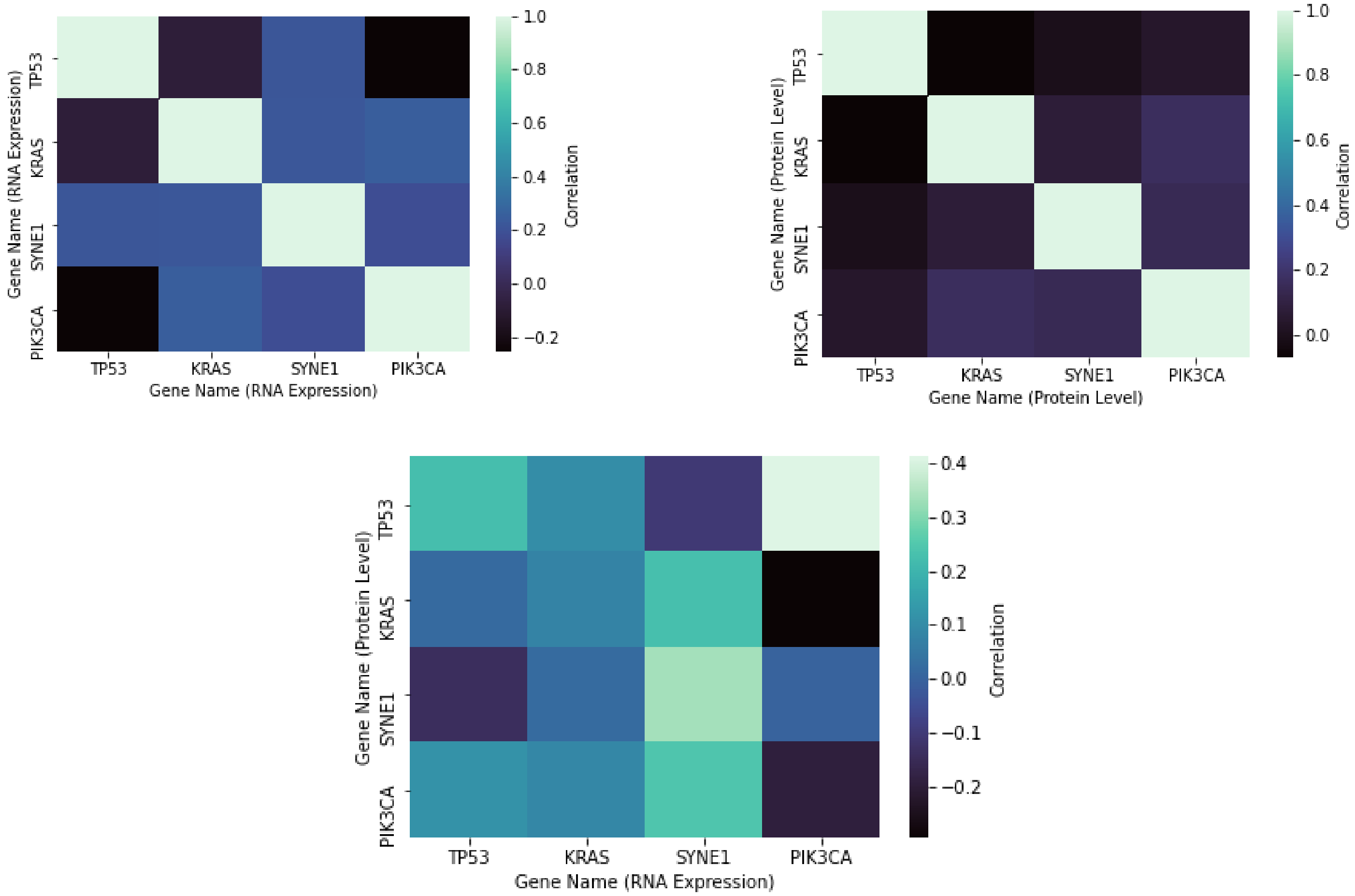

Figure 6. Top left: Spearman heatmap comparing correlation between RNA expression for TP53, KRAS, SYNE1, *and* PIK3CA. *Top right: Spearman heatmap comparing correlation between protein level for the same four genes as before. Bottom center: Spearman heatmap comparing correlation between RNA expression and RNA expression for the same for genes as before.* TP53 *tends to be weakly or negatively correlated when comparing between RNA expression only or protein levels only, but has positive correlations when comparing RNA expression and protein levels.*

Figure 6 is a Spearman correlation plot of four genes – *TP53, KRAS, SYNE1, PIK3CA* – that were commonly expressed in CRC. These four genes were chosen from the available data as good candidates to investigate due to their mutation frequency in CRC patients. These four genes were the result of the intersection of the genes which had protein level data from CPTAC with the union of the top 20 most frequently mutated genes from the “Young” and “Old” cohorts. Lighter colors represent a greater $\log_2$ fold change of RNA expression over protein levels. When looking at RNA expression only comparisons, TP53 is negatively correlated with PIK3CA and KRAS. When looking at the protein level only comparisons, TP53 is weakly positively correlated with PIK3CA, KRAS, and SYNE1. However, when looking at the RNA expression and protein level comparisons, there is a relatively strong positive correlation between PIK3CA RNA expression and TP53 protein levels.

Discussion

Figure 1 suggests that a higher number of patients expressed metastasis in younger patients than in older patients, with approximately 52% in “Young” patients and 38% in “Old” patients. These results are consistent with current literature, with metastasis displaying a negative correlation with patient age (Pretzsch et al.). Though the exact reasons are yet to be elucidated, one possible explanation behind this is the increase of anti-inflammaging factors, or factors that reduce endogenous inflammations that occur due to aging, in older individuals. Inflammation is a major factor in tumor growth, and hindering this could potentially slow down metastatic spread (Pretzsch et al.).

It was expected that younger patients would have a lower percentage of patients with metastasis due to a lower hypothesized level of mutations in the *TP53* gene. A potential explanation for the observed trends is that older patients with metastasis are more likely to be diagnosed with Mucinous adenocarcinoma, while younger patients with metastasis are more likely to be diagnosed with right-sided location cancer (Yang et al., 2018). This could result in different metastasis data based on diagnosis, which could account for the higher percentage of younger patients with metastasis. The results of this study suggest that the *TP53* gene is not correlated with metastasis in CRC patients.

Older age is correlated with increased gene mutations (Risques & Kennedy, 2018). *TP53* is a tumor suppressor gene, and it was initially expected that mutations in this gene would lead to decreased function in older patients. However, Figure 2 indicates that there is no significant difference in *TP53* gene expression levels between the two age categories. The results from Figure 3 seem to support this finding, as there is no statistically significant difference in the survival of patients in the “Young” and “Old” categories. One explanation for this is that there are confounding variables such as unknown gene interactions, resulting in the *TP53* gene having less of an effect on the disease alone. Another explanation could be due to the smaller sample size of younger patients, which could have resulted in skewed results. However, literature suggests that older individuals express higher levels of p53 protein, which challenges both the expected and experimentally obtained results (Wu and Prives, 2018).

Figure 4 indicates that *TP53* is the second most frequently mutated gene. *TP53* was expected to show up among the top most mutated genes due to its well established role as tumor suppressor, so this result was not surprising. Furthermore, patients that exhibited *TP53* mutations can also be broken up into two groups: those that have *TP53* mutations with *APC* mutations, and those that have solely *TP53* mutations. Finally, the TP53 mutations were, like the other genes aside from *APC*, most frequently missense mutations. This finding confirms the importance of *TP53* in colorectal cancer and emphasizes the need for more in depth research.

At the same time, attention should also be paid to *APC,* the gene that is the most mutated gene in both young and old patient categories. Like *TP53*, *APC* is also a tumor suppressor gene, and mutations in this gene could potentially lead to a lower function in patients. Furthermore, the gene had many different types of mutations present, as *APC* wasn't dominated by missense mutations like the other genes. Due to the prominence of *APC* in this study, further research into *APC* should be taken to understand its relationship with overall patient survival.

Figure 5 also reveals that older patients had many more mutations in the *TP53* gene. Many of the extra mutations present in older patients are missense mutations, which could have an effect on *TP53* function. This could potentially explain the results of Figure 1, as the mutations could have an unknown effect on the progression of metastasis. A more in-depth look at the role of specific mutations could make any mechanism more apparent.

Figure 6 indicates that *PIK3CA* RNA expression is greatly correlated with the presence of *TP53* protein relative to the other correlations. *PIK3CA* codes for the p110α protein, which is responsible for instructing cells when to grow and divide (Hamada et al., 2017). Intuitively, an overexpression of this gene is likely fundamental in the progression of certain cancers. The high correlation of the activity of both these genes in CRC patients could indicate that both of the genes play a big role in the onset of CRC in patients of all ages. These results match that of a study on lung cancer, which found that mutations in *TP53* and *PIK3CA* were found in many patients (VanderLaan et al., 2017). This could indicate that mutations in these two genes could increase the onset of tumors in multiple cancer types. The figure also shows a relatively low correlation between the expression of *TP53* RNA and *TP53's p53* protein presence. This could suggest errors in the post-translational processing that occurs to create the *p53* protein, which opens an avenue for future study. The low correlation could also be attributed to the difference in progress in the development of the technology used to detect RNA expression and the technology used to detect protein expression levels.

Overall, issues arise with the data when attempting to analyze metastasis and outcomes of younger and older patients. Younger patients are more likely to visit a health professional once their symptoms get worse, while older patients are more likely to have regular checkups. Thus, it is difficult to make concrete conclusions about the relationship between age and metastasis due to the fact that many studies do not collect data from young people that do not receive regular checkups. A future study can aim to resolve this, through a more in-depth study of age and metastasis in CRC. This study could utilize multiple datasets to reduce selection bias among patients.

References

Balducci L. Management of cancer in the elderly. Oncology (Williston Park, N.Y.). 2006 Feb;20(2):135-43; discussion 144, 146, 151-2. PMID: 16562648.

Clinical Proteomic Tumor Analysis Consortium. (NCI/NIH). (2006). https://proteomics.cancer.gov/data-portal

Hamada, T., Nowak, J. A., & Ogino, S. (2017). PIK3CA mutation and colorectal cancer precision medicine. *Oncotarget*, *8*(14), 22305–22306. https://doi.org/10.18632/oncotarget.15724

Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., Rodriguez Yoldi, MJ. (2017). Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. International Journal of Molecular Sciences, 18(1), 197, https://doi.org/10.3390/ijms18010197

Pretzsch, E., Nieß, H., Bösch, F., Westphalen C.B., Jacob, S., Neumann, J., Werner, J., Heinemann, V., Angele, M.K. (2022). Age and metastasis – How age influences metastatic spread in cancer. Colorectal cancer as a model. *Cancer Epidemiology*, *77*. https://doi.org/10.1016/j.canep.2022.102112

Risques, R. A., & Kennedy, S. R. (2018). Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS genetics*, *14*(1), e1007108. https://doi.org/10.1371/journal.pgen.1007108

The Cancer Genome Atlas Research Network. (NCI/NIH) (2016). https://www.cancer.gov/tcga

VanderLaan, P. A., Rangachari, D., Mockus, S. M., Spotlow, V., Reddi, H. V., Malcolm, J., Huberman, M. S., Joseph, L. J., Kobayashi, S. S., & Costa, D. B. (2017). Mutations in TP53, PIK3CA, PTEN and other genes in EGFR mutated lung cancers: Correlation with clinical outcomes. *Lung cancer (Amsterdam, Netherlands)*, *106*, 17–21. https://doi.org/10.1016/j.lungcan.2017.01.011

Wu, D., & Prives, C. (2018). Relevance of the p53–MDM2 axis to aging. Cell Death and Differentiation, 25(1), 169–179. https://doi.org/10.1038/cdd.2017.187

Yang, L., Yang, X., He, W., Liu, S., Jiang, C., Xie, K., Peng, K., You, Y., Zhang, B., & Xia, L. (2018). Comparisons of metastatic patterns of colorectal cancer among patients by age group: a population-based study. *Aging*, *10*(12), 4107–4119. https://doi.org/10.18632/aging.101700

Zahir Ahmed, Safia, et al. "Incidence of Age Migration of Colorectal Cancer in Younger Population: Retrospective Single Centred-Population Based Cohort Study." Annals of Medicine and Surgery, vol. 74, Elsevier Ltd, 2022, pp. 103214–103214, https://doi.org/10.1016/j.amsu.2021.103214.