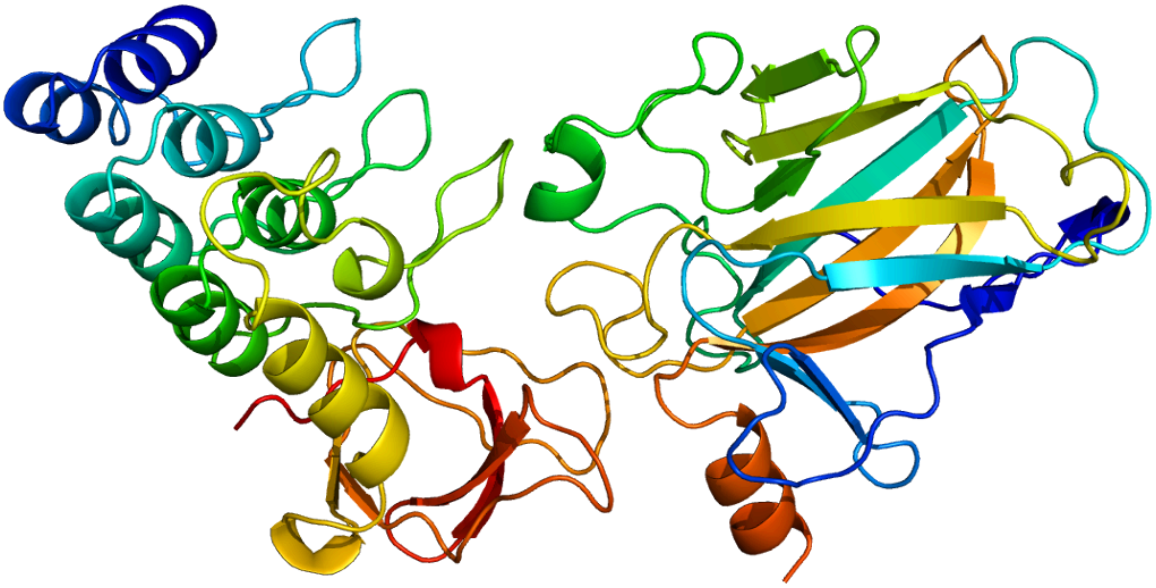
Variability in Expression of TP53BP2 in Common Cancer Tissues

Introduction

The modern research of cancer is highly interdisciplinary in nature, incorporating a variety of methods and techniques to treat the disease. A relatively well-researched aspect of the ongoing hunt for cancer treatment comes in the form of studying the almighty cancer suppressor protein, P53. Simply, P53 is the product of the TP53 gene. The protein is responsible for several important functions that aid in combating cancerous cells, including preventing the proliferation of cells with damaged DNA, slowing or halting the cell cycle to allow for cellular repair, inducing apoptosis in damaged cells, inhibiting angiogenesis, and regulating the immune system. Each of these functions serve to ultimately stop the growth of, and eventually eliminate tumorigenic cells. Recent research has shifted focus from the protein itself towards the genes responsible for its efficacy, such as TP53BP2. Studies have shown that TP53BP2 codes for a family of binding proteins that regulate both P53 and the TP53 gene. While TP53BP2 is able to regulate P53 through direct binding, its regulation of TP53 is more interesting. It is able to regulate TP53 through downstream feedback mechanisms.



Source: https://en.wikipedia.org/wiki/TP53BP2

When it comes down to modern therapeutic developments, personalized medicine has been proven to be one of the most effective approaches. There isn't always a "one size fits all" solution. We need to look at the factors affecting TP53 to find out how best we can amplify cancer suppression within patients. With this in mind, we turn our attention towards the variable gene expression of TP53BP2. Our goal is to focus on the variability in the expression of TP53BP2 throughout the most common cancer tissues–colon, breast and lung– in order to cross analyze expression with demographic information. We will be looking for possible correlation, patterns, or randomness in TP53BP2 expression among different age groups and sexes. We believe it is possible that patterns and correlations found in TP53BP2 expression could be used to create personalized treatments for patients in different demographic groups and provide insight into the types of research that needs to be conducted for new cancer treatments– i.e. potential epigenetic modifications to enhancer genes or binding proteins. To accomplish this, we will need to start by collecting and combining databases of cancer patients' age, sex, and level of expression.

Methods

Data Acquisition

The GTEX Portal Database houses a set of patients' expression data for thousands of different genes in dozens of different tissues throughout the human body, including tissues we are interested in. Each tissue has data that can be downloaded directly into xls format, which makes the data easily accessible through R. Our second

database, The Human Protein Atlas, houses data for patients that were tested for expression of TP53. The database is useful because it provides demographic data in the form of age and sex of the patients. Data on this database cannot be directly downloaded. However, it is accessible in table format, which can be copied over to an xls file, and thus be read by R. Because the 2 databases have the same primary key in patient IDs, our approach will be to clean up the data from both databases so that they are compatible with each other, and then integrate them using the same primary key (patient ID) using R.
Library packages used: ggplot2, dplyr, tidyr.

Throughout each tissue we investigate, the process of data acquisition and data integration remains nearly identical, with the only difference being the names of the tissues. To simplify, we will only go through how we acquired and obtained data for colon tissue, but the process remains identical for Lung and Breast tissue as well. Additionally, we did our best to clean up the data and make sure that the data is of quality and consistent among all of the files.

GTEX Portal Data

The first step is to give R access to the gene expression for Colon tissue from GTEX Portal (The direct link can be found here*)* The dataset itself contains odd formatting, so a separate variable must be created for the column heads which contain patient IDs.

```
master<-read.csv("bigtable.csv",header=FALSE)
book1<-read.csv("Book1.csv", header=FALSE)
```

Our focus is on the TP53BP2 gene (ENSG00000143514.17), so we filtered the data to only display the gene of interest. In addition to this, we also proceeded to combine the column heads containing patient ID's with the raw expression data for TP53BP2 itself.

```
master<-master %>%
  filter(V1 == "ENSG00000143514.17")
joined_master<-rbind(book1, master)
```

This leaves the data to be formatted in a horizontal manner, similar to wide formatting. This makes the data incredibly difficult to view and digest, so we opted to shift the data into a vertical format. At the same time, we also removed the gap rows that didn't provide information about expression or patients.

```
long_master <- tibble(
  tissue = as.character(joined_master[1, ]),
  Expression_Level = as.numeric(as.character(joined_master[2, ]))
)
#removing gap rows
N<- 2
long_master<- long_master[-(1:N),]
```

Now, this data is ready to be integrated with data from The Human Protein Atlas.

GTEx sample id	Sample description	nTPM
GTEX-15FZZ-1826-SM-6LLJP	30-39 years, female	32.7
GTEX-144GM-1126-SM-79OJK	20-29 years, male	29.2
GTEX-13W3W-2226-SM-5LU4M	60-69 years, female	26.7
GTEX-VJYA-2126-SM-4KL1O	60-69 years, male	26.7
GTEX-13O1R-2426-SM-5KLZZ	60-69 years, male	24.6
GTEX-ZYFD-2226-SM-5E43P	50-59 years, male	24.4
GTEX-Y9LG-1326-SM-4VBQB	30-39 years, male	24.1
GTEX-1AMEY-1826-SM-72D5L	30-39 years, female	23.5
GTEX-1F5PL-1926-SM-7MXTP	40-49 years, female	23.1
GTEX-1339X-1726-SM-5P9J9	40-49 years, male	22.4

Figure 1: The Human Protein Atlas patient expression levels of TP53.

The Human Protein Atlas

The data from The Human Protein Atlas is more refined, although it comes only in the form of a visible table. However, it can be copied into a .xls file and read by R. Although the data was simply copy and pasted, the formatting needs to be adjusted. Here, we took data on colon tissue patient data (figure 1). After pasting it into a .xls file, we had R read it so that we could separate the 'Sample description' into columns of age and sex.

```
colon <- read_csv("Book2.csv")
colon <- colon %>% separate(`data`, c("Age", "Sex"), sep = ",")
colon <- colon %>% select(-expression)
```

We also removed the TP53 expression data (nTPM), as we are solely focused on TP53BP2.

Data Integration

After preparing both datasets to be compatible with each other, they are ready to be merged using the same *primary key:* patient IDs.

```
merged_data_colon <- inner_join(long_master, colon, by = c("tissue" = "Patient"))
```

In addition to this, we can also export it into our own csv file.

```
write.csv(merged_data_colon, "merged_data_colon.csv")
```

Results

We were able to successfully integrate the TP53BP2 expression data and the demography data for the patients into 3 separate files. We can visualize the different levels of expression by demographic markers.

```
ggplot(merged_data_colon, aes(x = Age, y =
Expression_Level, fill = Sex)) +
  stat_summary(fun = mean, geom = "bar",
position = "dodge") +
  stat_summary(fun.data = mean_se, geom =
"errorbar", width = 0.2, position =
position_dodge(width = 0.9)) +
  labs(title = "TP53BP2 Expression by Age and
Sex in colon tissue", x = "Age", y =
"Expression Level", fill = "Sex")
```

In the plots below, the x-axis represents the age of patients, grouped into 10-year intervals from 20 to 79 years. The y-axis displays the expression level of TP53BP2, measured in transcripts per million (TPM). Within each age group, the data is further divided by sex, distinguishing between male and female patients.

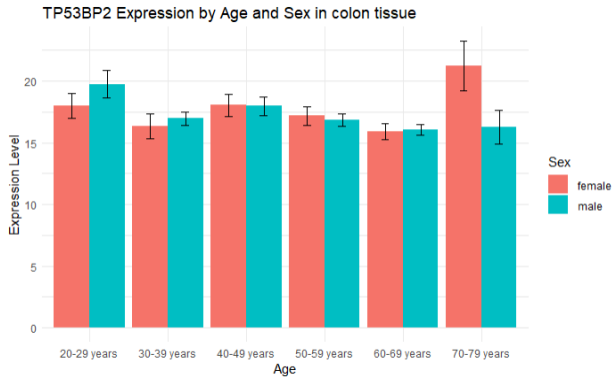


Figure 2: Tp53BP2 Expression by demographic markers in colon

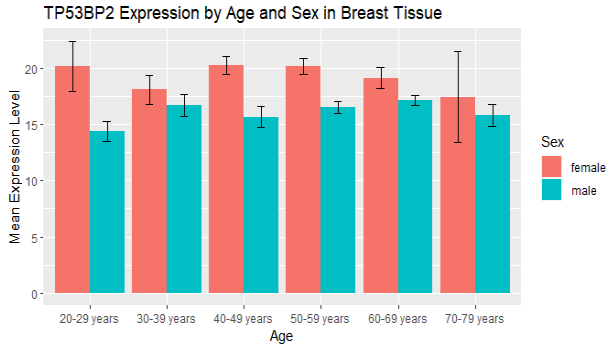


Figure 3: Tp53BP2 Expression by demographic markers in breast tissue

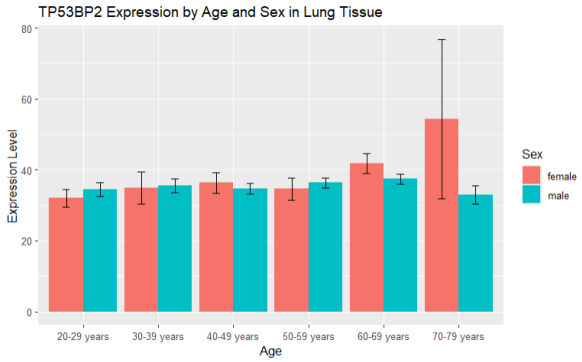


Figure 4: Tp53BP2 Expression by demographic markers in lung tissue

Based on the plots, the 20–29-year age group shows a slightly higher level of TP53BP2 expression in colon tissue, with males exhibiting higher expression levels than females in this group. There is also a notable peak in expression levels among females aged 70–79 years. However, there is no clear trend in TP53BP2 expression across age groups and sexes in colon tissue. On the other hand, in breast tissue, females exhibit consistently higher TP53BP2 expression levels compared to males across all age groups. In the group with the lung tissue, there is very evidently a huge amount of expression of TP53BP2 in females aged 70-79 years along with slightly higher expression in females aged 60-69. From these results it can be interpreted that at age 70-79 and maybe older ages as a whole corresponds to higher TP53BP2 expression.

Discussion

The expression level of TP53BP2 across different demographic groups in three cancer-prone tissues exhibits different patterns. In colon tissue, age- or sex-related mechanisms may influence TP53BP2 expression, as the eldest female group indicated an unusually high expression level. However, since no consistent trends indicative of age- or sex-related mechanisms were observed across other groups, we cannot make a definitive correlation between those demographic factors with TP53BP2 expression level. The clear pattern shown in the expression result of breast tissue indicates the possible protective role against breast cancer development, as females are more susceptible to breast cancer. This high expression level in females may correlate to higher cancer risk in females, indicating some potential protective functions of TP53BP2 in preventing breast cancer risk, possibly through its regulatory role in tp53 activity. This sex-based difference in expression level may be influenced by the hormone differences, as female sex hormones such as estrogen and progesterone may promote the expression of TP53BP2 and enhance its effect in protecting against cancer. In the lung tissue, it very clearly correlates that in the oldest age group for females the levels of TP53BP2 levels were quite high. This may be due to a large variety of factors one being age or sex related factors, for example aging can be associated with increased activity in the tp53 pathway or even hormonal related factors due to postmenopausal changes in women of that age category. Additionally environmental factors may play a role as well, it's unclear the history of the women chosen for the study but maybe only a group of female smokers for example, was picked for that age category. Overall, our project successfully integrated data and generated graphs that provide some initial understanding of the variability in TP53BP2 expression across colon, breast, and lung tissues. However, there are areas for improvement in further analysis. First, while we observed patterns and trends in the data, conducting statistical analyses, such as t-tests, would be necessary to confirm the significance of these differences. Second, it would be better if we could expand our analysis to more tissues that are prone to cancer. Currently, we are only working with three tissues, and our understanding of the expression result of TP53BP2 in different tissues is limited. For example, there may be a tissue presenting a more significant pattern with TP53BP2 expression level and those demographic factors. In addition, our current analysis focuses only on normal tissues with a high risk of cancer. It could be a better analysis if we could compare the expression level in those normal tissues with that in corresponding cancerous tissues. Additionally, our analysis was limited by the sample size of each of the patients in the study, we were not able to find additional information on the patients that were sampled, therefore it is hard to know what different biases were at play that could affect our data. Our analysis was also cross-sectional, meaning that we only performed analysis at a certain point of time which may not provide the full picture for how TP53BP2 plays a role in the TP53 pathway. Overall, our data

analysis has been a great starting point for understanding. For future studies perhaps if metadata that is more thorough is released it can be helpful to create different analyses across many different tissues again. This comparison can give us a better insight into how TP53BP2 expression is impacted by cancer in different tissues and could guide future research on the gene's role in cancer development and treatment strategies.