Pan-Collagen Gene Copy Number Variation Survival Analysis in Ovarian Cancer from TCGA

by

Robert Hodges, PharmD, MBA, RPh

A Thesis Submitted to the Faculty of
Utica College

August 2021

in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Data Science

Abstract

The Cancer Genome Atlas (TCGA) is a large longitudinal genetic database of cancer patients with varying cancer types which has led to many studies being published. Ovarian cancer is the most lethal gynecological cancer and genetics play a role in survival outcomes. This survival study performed Kaplan-Meier curves, log-rank tests, and Cox proportional hazard models comparing collagen gene copy number variation in ovarian cancer patients from the TCGA repository. Data cleaning was performed initially in Alteryx Designer and secondary data cleansing in RStudio version 1.4.1103. IBM SPSS was used for producing Kaplan-Meier curves as well as primary analysis. Secondary analysis and exploratory data analysis were performed with R-Studio. Stratification was performed on *copy number variation* showing stratification alignments of deletion, normal, and duplication. A Kaplan-Meier survival curve was performed along with a log-rank test and Cox proportional hazard model finding 3 genes: COL12A1, COL4A3BP, and COL5A3 to have statistically significant relationships between decreased survival and copy number variation abnormalities. This study supports current literature and provides evidence that collagen gene copy number variations can have varying survival outcomes in ovarian cancer patients. Keywords: Data Science, Dr. Michael McCarthy, survival analysis, proteomics, genome, neoplasm, ovary.

Personal Acknowledgements

I would like to thank Dr. John Christian Givhan Spainhour for mentorship, dedicated time to teaching over many phone calls, and helping guide me to this thesis subject. Christian has been a great friend over the years and has helped guide me in the world of data and statistics as I change careers. His teachings have pushed me to think more critically about my own thought processes and help hone my data science and research skills. His knowledge and expertise have been invaluable, and I am truly thankful and lucky to call him a mentor and a great friend.

I would also like to thank my committee members, Dr. McCarthy, Dr. Ho, and Dr. Rockefeller, for their guidance which has helped improve this project and helped me polish and sharpen all aspects of this research as well as learning more than I thought I would.

I would also like to thank a personal data scientist friend, Hansen Grider, for also guiding me on various data science issues and topics I have run across which has kept me on course. His friendship and knowledge have been truly significant.

I would also like to thank my primary data scientist mentor, Fred Frost, who has helped guide me in learning what I need to know to be successful in the data science world. His mentorship has been invaluable and keeps my career moving forward.

Most importantly I thank my fantastic wife, Vasilina Hodges, for putting up with me these past couple of years while working on this research and completing the data science program while also being pregnant with our first child, Victor Hodges. Her tolerance, understanding, and support has been the essential element of my growth in data science and her support has been overwhelming. I love you and thank you.

Table of Contents

List of Illustrative Materials

Introduction

The Cancer Genome Atlas (TCGA) has analyzed over 20,000 cancer observations across 33 different cancer types (Liu et al., 2018; *The Cancer Genome Atlas Program*, 2019). TCGA has been a colossal repository of genetic information pertaining to various neoplasm types with numerous publications emanating from this database alone, including novel discoveries concerning ovarian cancer (TCGA - Ovarian Serous Adenocarcinoma Study, 2018).

Ovarian cancer has a relative survival risk of 80%, which is defined as cancer survival in absence of other causes of death and is considered the most lethal gynecological cancer type (Cancer.org, 2014). One factor to be considered in ovarian cancer outcomes is the role of collagen, as collagen is a primary component in neoplasm fibrosis (Xu et al., 2019). Collagen can also affect neoplasm behavior through tyrosine kinase receptors, integrins, and various signaling pathways. Collagen has been studied to some extent as to these various mechanisms with regards to function in ovarian tumors; however, further research is needed (Ricciardelli & Rodgers, 2006).

Collagen is an extracellular matrix protein (ECM) that forms a barrier around and in support of organs and blood vessels (Dipiro et al., 2017). Collagen is the most abundant type of ECM protein located in ovarian tissue (Cho et al., 2015). Multiple factors contribute to the growth of tumors including ECM remodeling enzymes and growth factors. Collagen is essential in maintaining tissue homeostasis and ECM remodeling, such as ECM stiffness and elasticity, which are implicated in ovarian tumorigenesis (Cho et al., 2015).

Understanding ECM organization and cell morphology is essential in understanding and gaining insight into the mechanisms of malignancy (Cho et al., 2015). Collagen is also essential in regulating ovarian cell morphology through various mechanisms including cellular

communication and shape (Woodruff & Shea, 2007). For example, during tumorigenesis, collagen is remodeled into thick fibrils (Cho et al., 2015). This fibrillar structure is partially regulated by the COL12A1 gene, which is the collagen type XII alpha 1 chain in homo sapiens (National Center for Biotechnology Information – COL12A1, 2020).

Research has also focused on drug resistance mechanisms linked to functional abnormalities rather than structural (Holohan et al., 2013). For example, the COL4A3BP collagen gene, also known as the ceramide transfer protein (CERT) gene, codes for a protein that is a regulator of ceramide transport that interacts with ECM proteins and has been implicated in multidrug resistance in certain cancer treatments such as colorectal and breast cancer (Lee et al., 2011).

Copy number variation (CNV) has been linked to complex and abnormal behaviors in diseases and drug resistance (Gamazon & Stranger, 2015). CNV is defined as a structural variation that alters the number of copies of certain DNA regions (Thapar & Cooper, 2013). CNVs are important as it provides raw genetic material for gene divergence and expansion which has contributed to the evolution of humans (Perry, 2008). Copy number variation is important to human survival in that it increases genetic and phenotypic diversity of the species, but also linked to cancer (Li et al., 2017). This link to cancer is important as it has been estimated that 4.8% to 9.7% of the human genome is comprised of CNVs (Zarrei et al., 2015). While CNV contributes such a large portion to the human genetic profile, it has been shown that roughly 100 genes can be removed from the genome in a human without phenotypical consequences, thus making CNVs difficult to identify in importance (Zarrei et al., 2015). However, structural variations effect base pairs with a duplication or deletion and have been identified as a facilitator for genomic disease states (Sharp et al., 2005). CNV has also been found to affect gene

expression and a map of expression variation along with CNV information could help is understand certain disease states better, however, this relationship is not linear and is exceedingly complex (Gamazon & Stranger, 2015). This relationship seems to be found in cancer as some research suggests (Shao et al., 2019). Thus, it is important to understand CNVs and their effects on humans.

There have been issues in the past regarding CNV studies including sequencing techniques, pseudogenes, and technical artifacts, to name a few. Next generation sequencing (NGS) has replaced other methods as NGS has increased specificity in identifying CNVs and is also able to assess and identify pseudogene sequences (Kerkhof et al., 2017). NGS has had immense effects on genetic research by allowing researchers to expand our knowledge base and illustrating the importance of CNV effects (Zhao et al., 2020). Although, other methods are still valid such as Sanger Sequencing which is still the gold standard in validating multiple genetic errors (Arteche-López et al., 2021; Beck et al., 2016).

The research being conducted in this study evaluates the role of genetic abnormalities in collagen coding genes that may play a role in ECM organization and cell morphology by exploring heterogenous pan-collagen gene CNV and patient survival relationships in ovarian cancer by performing survival analysis.

A Kaplan-Meier survival analysis (KM), log-rank test, and Cox-proportional hazard modeling was performed on ovarian cancer patient CNV mutations in all 55 different collagen genes from the TCGA database. There are 46 genes that directly code for collagen in the human genome, three collagen subfamily genes that code for proteins that are collagen-like (COLEC genes), two collagen beta(1-0)galactosyltransferase genes (COLGALT genes), a collagen like subunit of acetylcholinesterase (COLQ), and two pro-collagen enhancer genes (PCOLCE genes)

(Gene Group, n.d.). The null hypothesis states no statistical differences in survival in ovarian cancer patients with any of the 55 collagen genes based on CNV stratification will be identified. The alternative hypothesis indicates there is a statistical difference in survival and death rates in ovarian cancer patients regarding CNV. For survival analysis, the dependent variable is the time to event, defined as death, while independent variables are gene CNVs. Censored data is included in the Kaplan-Meier model.

Survival analysis is essential in CNV exploration to potentially refine medical treatment for a more individualized approach in personalized medicine and target therapy. This importance has been emphasized in literature pertaining to genetics (Gamazon & Stranger, 2015).

Hopefully, by increasing the knowledge base of CNV effects on patient survivability, a considerable impact on patient outcomes with increased survivability may be achieved. The goal of this research is to increase the knowledge base of CNV effects in hopes for eventual targeted collagen or targeted gene therapies in ovarian neoplasm treatment.

Literature Review

TCGA has produced multiple studies that contribute to human understanding of genetics (The Cancer Genome Atlas – Publications, 2019). Ovarian cancer has been studied with regards to collagen in the past; however, little has been investigated regarding CNV and ovarian cancer. A PubMed literature search using medical subject heading (MeSH) terms for “Ovarian Neoplasms” and “Collagen yielded over 300 results involving genetic studies concerning collagen and ovarian neoplasms. A brief secondary search was also performed in Google Scholar regarding ovarian cancer, CNV, and collagen for the year 2021. The Google Scholar literature review allowed the most recent literature to also be included.

Gene expression is generally the most researched topic with regards to neoplasms as normalization techniques in high-throughput RNA sequencing are more available and widely used (Dillies et al., 2012). For example, gene-drug interactions in ovarian cancer have been investigated but with respects to gene expression instead of CNV (Teng et al., 2013). It is surprising that CNV has not be studied more considering that a systemic study of the human population has shown noticeable effects from CNV mutations (Shaikh, 2017). However, we are just now producing research indicating that CNV is a reliable and possibly better biomarker outcome in ovarian cancer patients (Paracchini et al., 2020).

Collagen has been shown to be involved in multiple aspects of tumorigenesis (Cho et al., 2015). One study linked collagen gene expression to metastasis promotion through the transforming growth factor (TGF-β1) signaling pathway (Cheon et al., 2013). Another linked the collagen gene COL2A1 and higher gene expression with delayed tumor relapse in high-grade ovarian cancer patients (Ganapathi et al., 2015). Drug resistance due to collagen gene expression by inhibiting molecular penetration and decreasing tumor apoptosis has been studied in ovarian neoplasms and was suggested to involve gene expression of COL5A3 (Januchowski et al., 2016). Another study found decreased gene expression levels of XI alpha 1 collagen gene COL11A1 with decreased ovarian tumor invasiveness and oncogenic potential (Wu et al., 2013).

Some investigation has been done regarding CNV and ovarian cancer. One study investigated gene-drug interactions with regards to CNV in glioblastoma multiforme and lower grade brain glioma (Spainhour & Qiu, 2016). Another study by the same author investigated CNV with drug exposure and survival data which allowed inference to drug-gene interactions which effect patient survival across all cancers in TCGA, which was then put into a portal called the gene-drug interaction for survival in cancer (GDISC) (Spainhour et al., 2017). One study

researched CNV from TCGA and tried to identify genes as strong candidates for therapeutic targeting (Li et al., 2017).

Another study performed a genome-wide analysis of CNV and ovarian cancer risk and found three CNV regions that are strongly associated with ovarian cancer risk, but may not contribute to susceptibility (Reid et al., 2019).

TCGA has been researched before with ovarian cancer and CNV. However, one study titled "Integrated Analysis of Germline and Somatic Variants in Ovarian Cancer" was limited to only three genes more commonly found in breast cancer and performed only a large-scale exome-wide analysis (Kanchi et al., 2014). Analysis of TCGA data in another study found that estrogen related receptor CNV is related to ovarian cancer histological grades of ovarian cancer (Huang et al., 2021).

Bias is a critical topic that should be covered in genetic research. An important bias to mention from TCGA literature, sequence homology can create technical artifacts, which in turn affects downstream analysis and mapping which potentially also cause concerns (Webster et al., 2019). These sequence homologies are created from shared evolutionary roots where DNA regions share high similarity or can even be pseudogenes. The technical artifact phenomenon has also been found in TCGA data (Buckley et al., 2017). This can possibly create bias in the genetic database where an estimated 5-10% of cancer patients may be more predisposed to certain cancer types (Garber & Offit, 2005). This is important as cancerR2d2rules can be caused in part by inherited genetics which may be passed to offspring in which these progenitors may work synergistically (Buckley et al., 2017).

There is typically a greater type I error bias in genetic research (Efron, 2005). Type I error bias is defined as having false positives. This is important as false positives could lead to

misleading outcomes for patients along with inefficient spending of time and money. Included in contributing to type I error bias are study methods which should also be carefully taken into consideration when selecting statistical methods for use in genetic research.

Methods

This study uses non-parametric Kaplan-Meier survival plots, log-rank tests, and semi-parametric Cox proportional hazard modeling for survival analysis. Non-parametric Kaplan-Meier curves are used as this study does not assume known distributions of events as it is likely not accurate to assume that hazard rates or probability of an event is constant; therefore, only non-parametric and semi-parametric methods are used.

The Kaplan-Meier survival curve (KM) is used in this study as it can account for censored data. Censoring is defined as the cutoff of survival time when the endpoint of interest has not been studied due to loss of follow-up of the particular patient (Kaplan & Meier, 1958). Censoring is either due to loss of patient follow-up data or non-occurrence of death in this study. The Kaplan-Meier survival curve makes a couple of assumptions. The first is that the survival of patients are independent of each other. The second assumption is that censoring occurs independently. The requirements to plot a Kaplan-Meier survival curve are patients' status of last observation and time to event. If one is to compare Kaplan-Meier curves, then data regarding what characteristic is being studied must be assigned to each group (Rich et al., 2010). The Kaplan-Meier curve is the "gold standard" in survival distribution estimation and is also used instead of the Nelson-Aalen estimator in this study because the Nelson-Aalen estimator is consistently larger in estimates when compared to the Kaplan-Meier estimation method (Bohoris, 1994). It is more appropriate to move into statistical details and other tests after the creation of Kaplan-Meier curves, involving hazard and survival functions, because the Kaplan-Meier curve

itself cannot account for differences in covariates, such as CNV, as the Kaplan-Meier curve is considered as a visual illustration only.

The hazard function and survival functions are both integral parts of survival analysis modeling. The survival function explains via probability of a subject surviving beyond a specific point in time (Kleinbaum & Klein, 2012). The point in time can be considered a device failure, the end of a study period, or in this case, death. The hazard function, otherwise known as a failure rate, is the rate of occurrence of a certain event during the given time interval. The hazard rate is also known as the Cox proportional hazard model. The survival function and hazard function are related and can be converted to each other (Schober & Vetter, 2018). For example, when the survival function is high, then the hazard rate is lower and there is increased survival, or less events take place which means they are inversely proportional.

To obtain the survival function equation, one must first look at the distribution function of survival time of an individual, also known as the cumulative distribution function of *T*, notated in Figure 1 (Collet, 2015).

$$F(t) = P(T < t) = \int_0^t f(u)\, du$$

Figure 1. Cumulative Distribution Function.

T will always be a positive number and is defined as a random variable associated with survival time. Figure 1 represents that survival time is less than some value of the variable *t*, which is defined as a specific point in time. The right part of the equation in Figure 1, $\int_0^t f(u)\, du$, is defined as the integral of the probability density function since any value of *t* can be a positive value. Therefore, this equation can be transformed into the survivor function shown in Figure 2 (Cox, 1972).

$$S(t) = P(T > t) = 1 - F(t)$$

Figure 2. Survival Function.

The survival function is then defined as the probability that survival time, *T*, is greater than a specific time, *t*; or that an individual survives beyond a specific time. Inversely, one can also define the survival function as the probability that one or more events take place after time *t* (Collet, 2015; Cox, 1972). Once probabilities have been obtained, statistical significance needs to be examined through log-rank testing.

A non-parametric log-rank test is constructed by separating each event time of the groups being studied and will help show the differences between the two groups in the Kaplan-Meier survival curve. A table is created showing the number of deaths (event), *d*, number of subjects alive, and total number of subjects. This is completed at every death event and the observations are treated as independent events. This procedure is known as the Mantel-Haenszel time-stratified procedure (Collet, 2015; Mantel & Haenszel, 1959). This procedure gives the (relative) risk ratio (RR). The RR is defined by the equation in Figure 3.

$$RR = \frac{a_i / n_a}{b_i / n_b}$$

Figure 3. Risk Ratio Equation.

The variable a_i is defined as the number of events for stratified group *a* while the variable b_i is defined as the number events for stratified group *b*. The variable *n* is the total population of each group as defined by the subscript. This equation gives the relative risk of one outcome group compared to another. For example, if group *a* was calculated with an $a_i / n_a = 0.1$ as the numerator and $a_b / n_b = 0.07$ then the RR would equal 1.43. This translates to group *a* having a 58% chance of the event when compared to someone without or with a different stratification.

The next step is to make sure this ratio is statistically significant. However, generally it is more informative to obtain the hazard ratios which are discussed in the Cox-proportional hazard model section.

The log-rank test will also give a probability value (p-value) for the difference between the two groups plotted on the Kaplan-Meier survival curve while assuming the null hypothesis of no difference between the two groups. The greater the gap in survival on the Kaplan-Meier survival curve between the two groups, the lower the p-value. The p-value is taken from a chi-square test where an alpha level has been set, generally 0.05, along with degrees of freedom (Pearson, 1900). Degrees of freedom is defined as the number of comparison groups minus one (Gosset, 1908; LaMorte, 2016). The log-rank chi square test is used instead of an analysis of variance (ANOVA) test in survival analysis because categorical data is being used. However, the log-rank chi square test can be considered a type of one-way ANOVA for survival analysis (Fisher, 1925; Grace-Martin, 2018). The chi square equation is shown in Figure 4.

$$X^2 = \Sigma \frac{(O_{jt} - E_{jt})^2}{E_{jt}}$$

Figure 4. Chi Square Equation.

O_{jt} represents the observed number of events for the observed *jth* group over time while E_{jt} represents the expected number of events in the *jth* group over time (LaMorte, 2016). The chi square number for each group are the sums for the observed and expected events computed in the chi square equation, at each event time. Expected events are calculated from the proportion of events occurring at each time with data from both groups totaled. This is better defined as total number of events divided by the total number at risk. The obtained value is then multiplied by number at risk in each group. The sum of this number is E_{jt}. The p-value is then determined from the chi square table of critical values where a statistically significant finding is where the

chi square value is greater than the corresponding critical value on the table (Fischer, 1925). Generally, the hypothesis is considered statistically significant if the chi square value is greater than 3.84 which corresponds to being less than the alpha value of 0.05. This is the same alpha value for this investigation. Once events have been plotted, and differences have been shown to be statistically significant, one should model the hazard or risk through the Cox proportional hazard model.

The Cox proportional hazard model is a regression and is considered the most utilized regression model in survival analysis (Chilamkurthy, 2020; Cox, 1972). This modeling technique allows researchers to investigate relationships between covariates and survival time. The model will allow a hypothesis about survival being equal or different to the data to be tested and is considered a natural extension of the log-rank test (Tibshirani, 1982). The Cox proportional hazard model allows the hazard to change over time but assumes that the hazard ratio is proportional or constant. For example, if the data presented a hazard ratio of male to female with males being twice as likely of an event than a female, it would assume this ratio is constant over time, or that the risk for the male is the same compared to a female at any point in time. Therefore, the Cox proportional hazard model is considered a semi-parametric model. However, assuming a constant ratio for the hazard model is considered unrealistic in the health sciences but allows for easier interpretation of the data as this research hasn't assumed any distributions (Kennedy, 2019; Zweiner et al., 2011). If one were to use a parametric model such as Weibull or exponential distribution models, without knowing the distribution, then potentially very inaccurate inferences are possible; and thus, a nonparametric Cox model is used.

The Cox proportional hazard model also makes other assumptions (Cox, 1972). Much like the Kaplan-Meier survival curve, the first assumption is that censoring does not lead to an

increased or decreased likelihood of events, or rather that the censoring is *non-informative*. The Cox model also assumes that survival times are *independent*. This *independence* is defined as the patient's survival or event is not dependent on another patient's survival or event. Censoring generally will affect the survival curve shape primarily when many censored observations occur at a specific point in time leading to large flat lines in the survival model (Prinja et al., 2010). The Cox proportional hazard model also assumes that the baseline hazard is unspecified and that the treatment variables do not change over time. The most important assumption is that the survival curves do not cross each other (Kennedy, 2019; Zweiner et al., 2011). This assumption can be met by using the Kaplan-Meier survival curve and investigating to see if survival curves cross (Zweiner et al., 2011). If the survival curves do not cross, then this assumption has been met. The last assumption is that the log hazard rate is a linear function of the variables, much like logistic regression, where the log odds are the linear function of the variables.

The Cox proportional hazard model is expressed via the hazard function which is defined as the cumulative risk of an event occurring by time, *t*. The simple form of the hazard function equation is shown in Figure 5 (Collet, 2015).

$$H(t) = -\log S(t)$$

Figure 5. Hazard Function.

The hazard function also illustrates the cumulative number of expected events that occur from time zero until a specified time, *t*, and encapsulates the risk of death up until time *t* and is communicated through the hazard ratio.

The hazard ratio is the exponential parameter estimate of proportional hazard models, or two groups with a hazard function, which may then be used to approximate the ratio of hazard rates between a comparison or control group and a treatment group (Bradburn et al., 2003). The

hazard ratio is akin to the relative risk ratio (RR) as discussed earlier, although not the same (Sutradhar & Austin, 2008). Relative risk ratio does not factor the timing of the event like hazard ratio. The hazard ratio is evaluated by looking at the values which signifies if the hazard ratio is higher or lower than the comparison or control group.

Another way of defining the hazard ratio is through the equation in Figure 6 (Bradburn et al., 2003).

$$H(t) = H_0(t) * exp^{(b_1X_1 + b_2X_2 + \ldots b_iX_i)}$$

Figure 6: Hazard Function Expanded.

H(t) is defined as the hazard function, which is determined through variables $X_1, X_2, \ldots X_i$. $H_0(t)$ is defined as the baseline hazard which means that this is the hazard if all covariates of $X_1, X_2, \ldots X_i$ are equal to zero. The variable t is the specified survival time and $b_1, b_2, \ldots b_i$ are notated as the coefficients, or hazard ratios, which explain the impact of covariates $X_1, X_2, \ldots X_i$. Thus, the Cox proportional hazards model may be written as a multiple linear regression of the log of hazard with the variables $X_1, X_2, \ldots X_i$. $H_0(t)$, which is the baseline hazard, can be roughly interpreted as an "intercept". However, this "intercept" will vary over time, thus, it is not considered to be the same type of intercept seen in standard linear regressions. When coefficients $b_1, b_2, \ldots b_i$ are greater than zero, as $X_1, X_2, \ldots X_i$ increase in value, the survival likelihood decreases, or the likelihood of an event increases. If a hazard ratio is greater than one, then covariates are positively correlated with an event happening and thus, hazard ratios greater than one are negatively correlated with survival time (Bradburn et al., 2003).

Data

Patient clinical and survival data was obtained from The Cancer Genome Atlas (TCGA) database under the TCGA-OV project (ovarian cancer). Copy number variation (CNV) data was

obtained from The Broad Institute of MIT and Harvard, which is part of TCGA, from the ovarian cancer archives in the form of comma separated value (CSV) files (*Broad GDAC Firehose*, 2016). The original files were not modified for data integrity purposes. Data was initially cleaned and joined using Alteryx Designer (Version 2019.4.8.22007) and data cleaning workflow can be referenced in Appendix A (Alteryx, 2006). Additional data cleaning and exploratory data analysis (EDA) was performed on the new data frame with the DataExplorer and SurvMiner packages in RStudio (Version 1.2.5033) (RStudioTeam, 2021). Survival analysis methods were completed in IBM's SPSS Statistics (Version 25) (IBM Corp, 2017).

Figure 7 illustrates the data pipeline in this study.

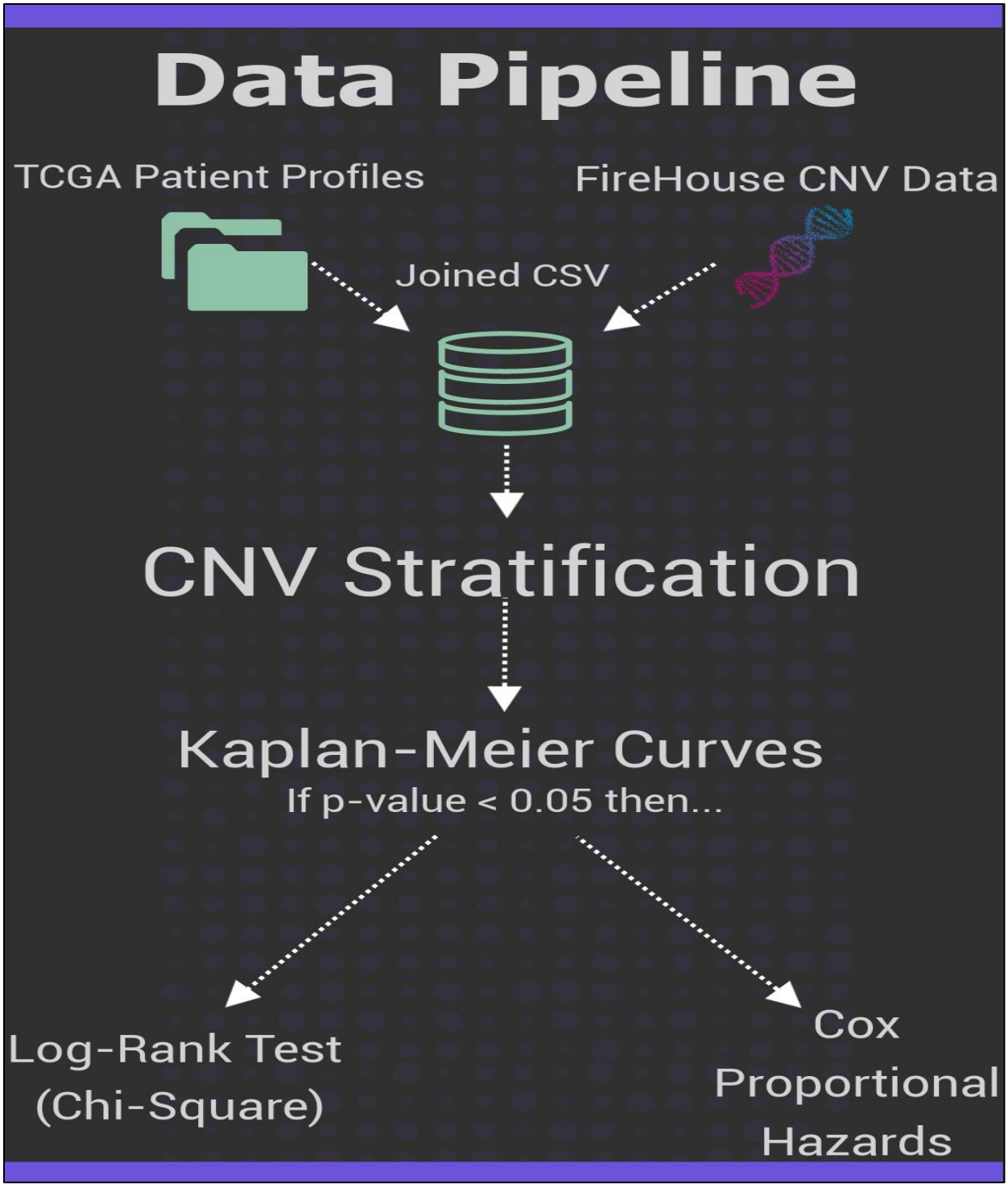


Figure 7. Pipeline of TCGA Data

55 collagen gene CNV columns were analyzed in EDA. Patient event profiles were joined with patient genetic data to create a joined data file. Table 1 explains the CNV

stratification from Figure 7 and lists the reference to each value and stratified groupings into smaller groups for ease of analysis as well as increasing power of the study. CNV range for each gene spanned from negative two through positive two initially.

Table 1. Copy Number Variation (CNV) Description and Stratified Groupings

Value	Description	Stratified Groupings
-2	Complete Deletion	Deletion
-1	Partial Deletion	
0	Normal	Normal
1	Partial Duplication	Duplication
2	Complete Duplication	

*** Varying distributions of each gene were processed in EDA and can be seen in Appendix B.**

Once Kaplan-Meier curves were performed, any model with a p-value less than 0.05 was used for further modeling of log-rank tests and Cox-proportional hazard models.

The timeframe column titled "Days.to.Death" displayed a skewed right distribution with a high spike at the beginning of the timeframe and can be seen in Appendix B. Maximum days to death was 5,481 days and minimum was eight days with a median of 864 days and a mean of 989 days. There were 564 observations with the status column displaying 291 events of death and 273 censored non-events. Censoring is defined as where the event, death, did not occur during the observation (Prinja et al., 2010). Censoring is either due to loss of patient follow-up data or non-occurrence of death in this study. Table 2 shows the summary of the censored data for each stratification where N is defined as the population or observations. 48.4% of the data had censored events.

Table 2. Summary of Censored Data

Censored Observations Summary				
			Censored	
COL12A1	Total N	N of Events	N	Percent
A-Norm	234	113	121	51.7%
B-Del	196	118	78	39.8%
C-Dup	134	60	74	55.2%
Overall	564	291	273	48.4%
COL5A3				
A-Norm	154	76	78	50.6%
B-Del	193	110	83	43.0%
C-Dup	217	105	112	51.6%
Overall	564	291	273	48.4%
COL4A3BP				
A-Norm	133	72	61	45.9%
B-Del	410	207	203	49.5%
C-Dup	21	12	9	42.9%
Overall	564	291	273	48.4%

Kaplan-Meier (KM) survival curves were applied to all independent variables. Genes COL12A1, COL4A3BP, COL5A3 were found statistically significant with a p-value threshold of < 0.05. Log-rank tests and Cox-proportional hazard models were applied to the three significant findings.

It is important to mention that the log-rank test is used instead of the Prentice modified Wilcoxon test, which is more sensitive and may make this study more susceptible to type I error bias (Peto & Peto, 1972; Kalbfleisch & Prentice, 2002). The Wilcoxon test is more sensitive as it compares weights of each event by the population at risk at event time, rather than weighing all events equally like the log-rank test (Visintainer, 2016). The use of the log-rank test helps control for this type I error bias.

There are roughly five primary equations that can be used within the Cox-proportional hazard model, to which the Efron and Breslow methods are generally considered to be the most

common (Xu, 2019). The Efron method is generally seen as more accurate than the Breslow method but more difficult to use (Breheny, 2019; Li, 2019). This issue is generally brought up when two groups have a tie in events (Breslow, 1972). The Breslow method is used in this study within the Cox-proportional hazards model as it is the baseline estimation method used in SPSS software and controls for type I error bias (Singer & Willett, 2003). The use of this estimation method may potentially control for type I error bias, although may decrease accuracy when there are ties in events, to which this data has a relatively small amount of cases as time is measured in days in this study.

Results

Log-rank tests and Cox-proportional hazard model primary results are displayed in Table 3, which include chi square values, p-values, coefficients, and hazard ratios with confidence intervals. All three gene survival models were statistically significant overall as defined by the p-value in Table 3 under the normal stratification, however, not all CNV stratifications were significant.

Table 3. Log-rank Results and Hazard Ratios

Gene	Coefficient	P-value	Chi-Square	Hazard Ratio	Lower 95% CI	Upper 95% CI
COL5A3 Normal		0.015*	8.56			
COL5A3 Deletion	0.299	0.045*		1.349	1.006	1.808
COL5A3 Duplication	-0.083	0.586		0.921	0.684	1.240
COL4A3BP Normal		0.026*	7.698			
COL4A3BP Deletion	-0.075	0.586		0.928	0.709	1.214
COL4A3BP Duplication	0.731	0.019*		2.077	1.125	3.835
COL12A1 Normal		0.047*	6.179			
COL12A1 Deletion	0.245	0.063		1.277	0.987	1.654
COL12A1 Duplication	-0.115	0.474		0.891	0.650	1.221

***Statistical significance with p-value < 0.05.**

In Table 3, the COL5A3 coefficients were negative for duplication CNV and positive for deletion CNV when compared to normal CNV. This translates to an increased survival for duplication and a significant decrease in survival for deletion mutations. However, the duplication stratification was not statistically significant.

The coefficients can be translated as percentages of the event per unit time. For example, COL4A3BP duplication translate to roughly a 73% increase in probability of event per unit time while the COL4A3BP deletion would translate to a 7.5% decrease in probability of event per unit time.

Hazard ratios for COL4A3BP were only significant between the normal CNV and duplication mutations, showing that the duplication mutation has significant decreased survival. The large coefficient value of 0.731 corroborates the decrease in survival for duplication mutations.

The coefficients for the COL12A1 gene with deletion and duplication mutations were not statistically significant while the overall model was statistically significant. COL4A3BP had positive coefficients which imply lower probability of survival with duplicated and normal CNV, however, only duplication and deletion CNV was statistically significant.

Hazard ratios for statistically significant findings include gene COL4A3BP, for duplicated CNV, with a hazard ratio of 2.077. This translates to just over a 2:1 ratio for increased chance of death which indicates that there is a 67.5% probability that a patient with this gene duplication will have an event before a patient with the normal stratification. Hazard ratios for COL5A3 were statistically significant for deletion mutations at 1.349. This translates roughly to a ratio of 1.35:1 ratio for chances of death; or a 57.4% probability of death for a patient with the COL5A3 deletion mutation before a patient with the normal stratification. A hazard ratio of one for each stratification would be equivalent to each stratification having a 50% probability that a patient with this mutation will have an event before someone who doesn't have a mutation.

A forest plot, shown in Figure 8, displays a graphical representation of the hazard ratios. Note that the normal stratification is not displayed as it is located at the vertical line on the x-axis at the one marker for baseline comparison to the other stratified CNV mutations.

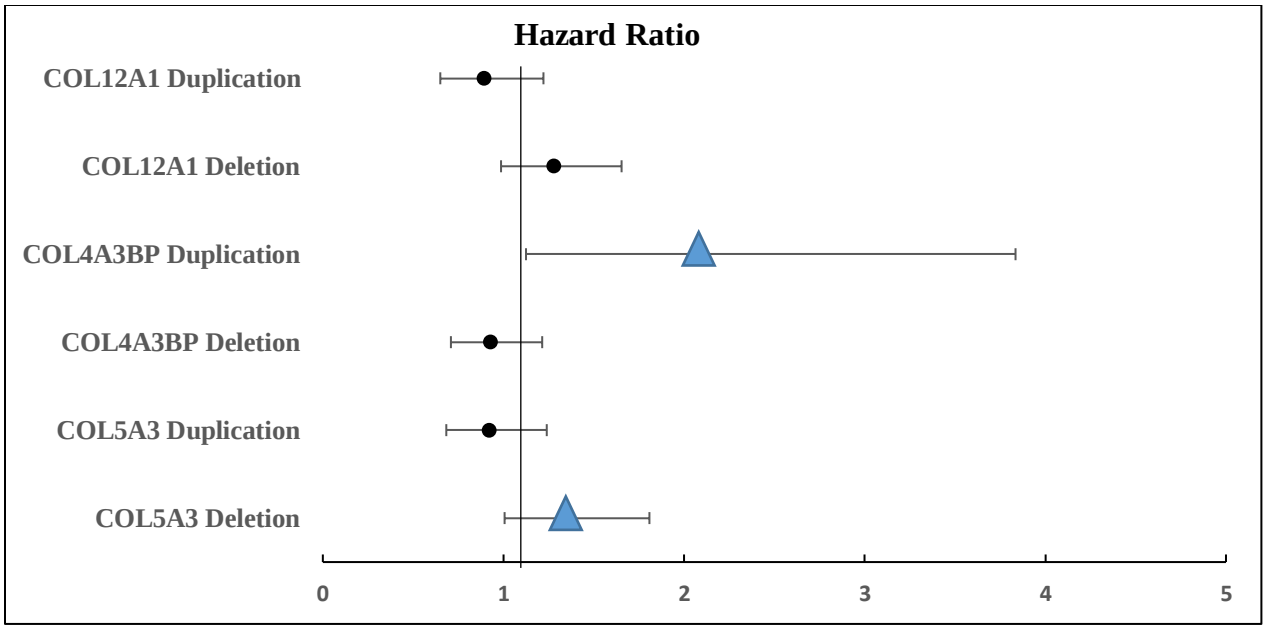


Figure 8. Forest Plot of Hazard Ratio.

Figure 8 illustrates that the hazard ratio for COL4A3BP duplication stratification confidence interval is quite wide due primarily to lack of data, yet it was statistically significant. Statistically significant stratifications have markers highlighted in blue and shaped as a triangle. All markers right of the line marked at hazard ratio of one are considered to have an increased chance of death, and markers less than one (1.0) are considered to have increased survival compared to the baseline of no mutations of respective collagen genes. The x-axis value of one is the baseline and is the normal stratification.

Figure 9 displays a bar chart of coefficients which are correlated directly to the hazard ratios. Highlighted in blue are the statistically significant findings for each stratified gene. Values to the right of baseline, at value zero, translate to an increased chance of event, and negative values to the left of the baseline represent stratifications with a decreased chance of event. The zero marker is baseline and is the normal stratification. The coefficients can be translated to a percent chance of events per unit time. For example, the COL4A3BP duplication can be translates as roughly a 70% increase in probability of an event per unit time.

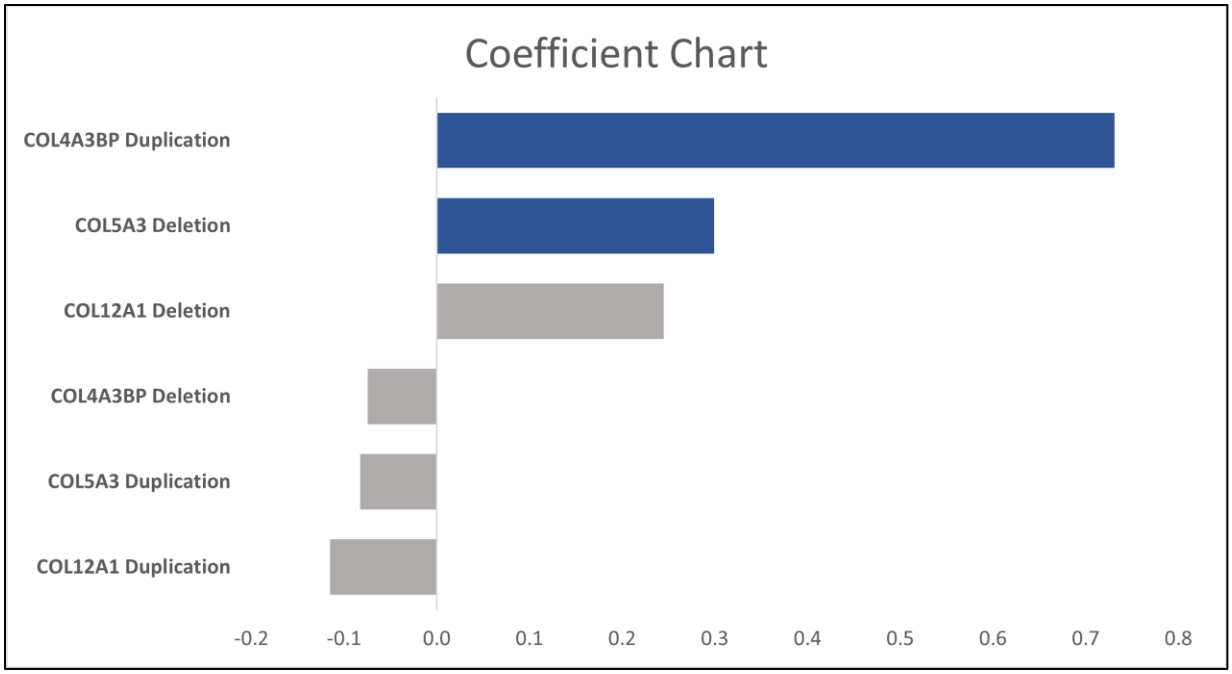


Figure 9. Coefficient Bar Chart

Figures 10a, b, and c show Kaplan Meier survival plots for each collagen gene. Each vertical drop in the line represents a patient event (death) and each tick mark represents a censored event. As stated previously, censoring is either due to loss of patient follow-up data or non-occurrence of death in this study. The blue lines denote a normal CNV, while red and green lines denote deletion and duplication stratifications, respectively.

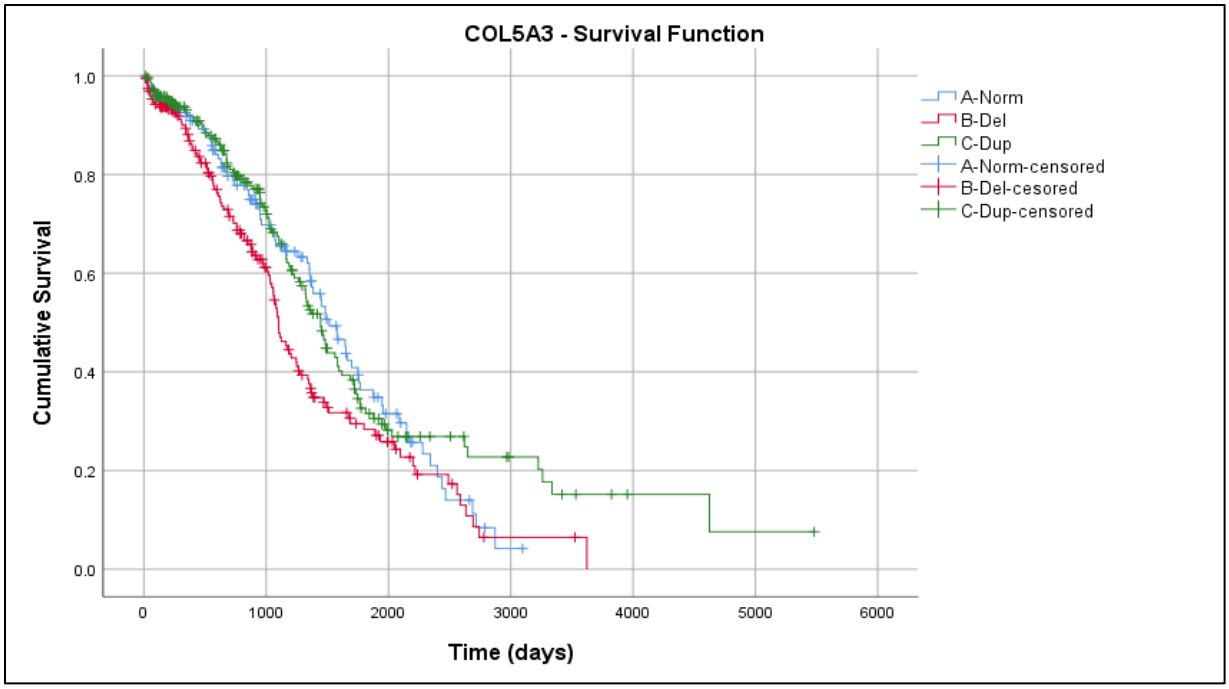


Figure 10a. COL5A3 Kaplan-Meier Curve.

Gene COL5A3 in Figure 10a has a very distinct and visible deletion mutation survival difference between roughly 300 days and 1,700 days then converges again to where all three stratifications do not show considerable differentiation at day 2,000, then splits off again. It is important to note that the deletion curve only just crosses the normal CNV curve at around 2,400 days.

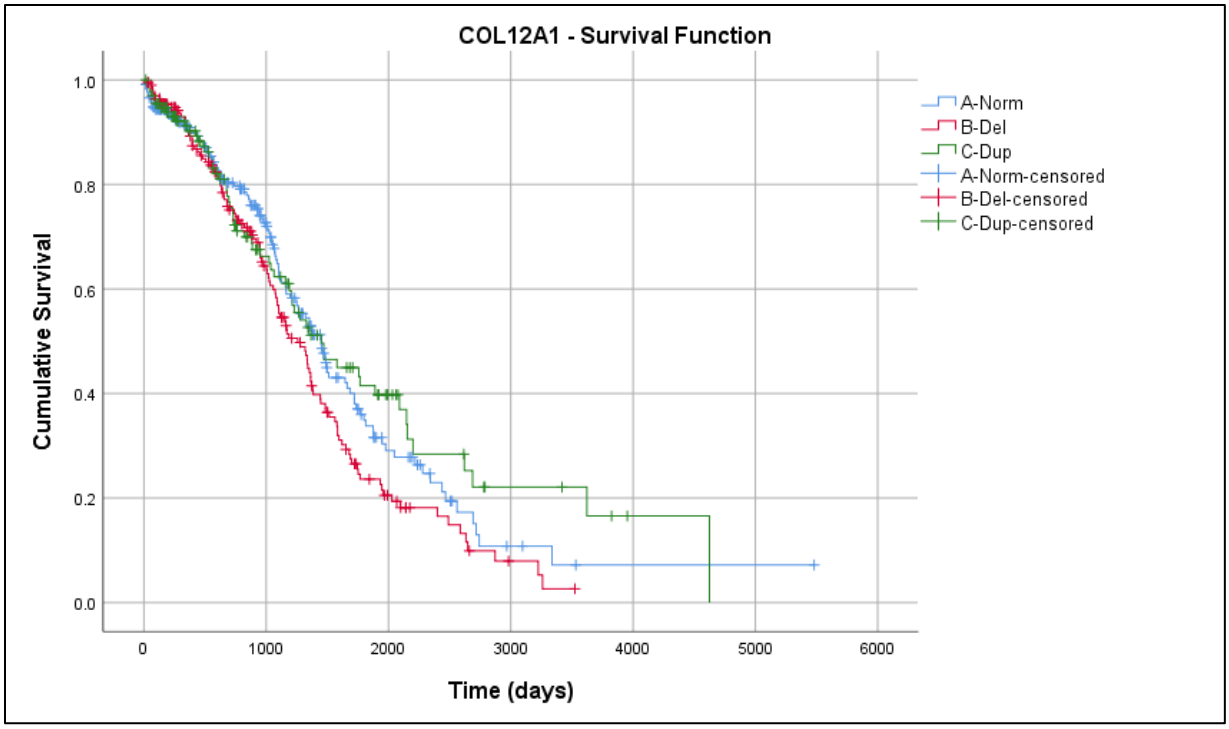


Figure 10b. COL12A1 Kaplan-Meier Curve.

Gene COL12A1 shown in Figure 10b illustrates deletion of CNV with a lower survival probability starting just about 1,000 days, however, these mutation stratifications were not statistically significant. This figure also has survival curves crossing with the duplication stratification crossing both deletion and normal CNV lines.

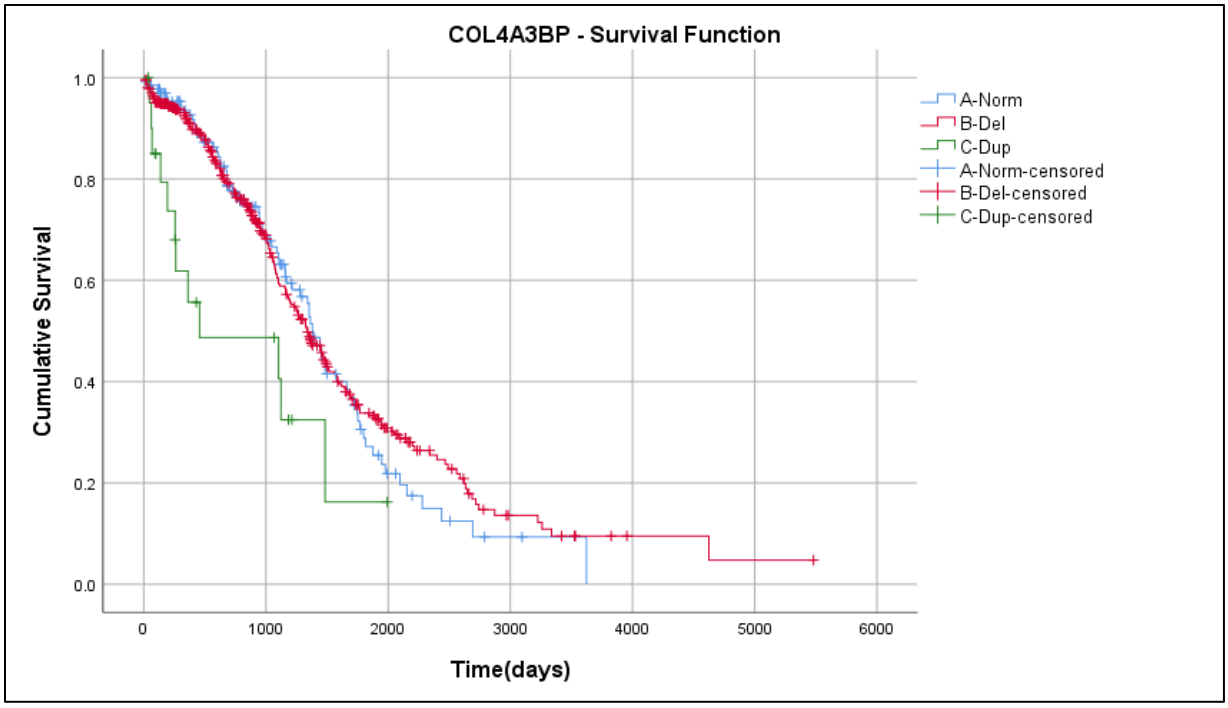


Figure 10c. COL4A3BP Kaplan-Meier Curve.

*** Shown in Appendix C are additional Kaplan-Meier curves with confidence intervals presented (created in R-Studio)**
****Shown in Appendix D are hazard function models.**
***** Shown in Appendix E is the Kaplan-Meier curve and the hazard function at the mean of covariates.**

Collagen gene COL4A3BP in Figure 10c is only significant with regards to duplication when compared to normal CNV. However, the data for this stratification was sparse as can be seen by the number of events as compared to the other stratifications. This model does not display curve overlap with the statistically significant duplication stratification but does show curve overlap between normal and deletion CNV stratifications.

Table 4 displays observed against expected outcomes from the log-rank test of each gene with asterisks marking statistical significance.

Table 4. Observed and Expected Scores of Each Stratified Gene.

Gene	N	Observed	Expected
COL12A1 = Deletion	196	**118**	98.8
COL12A1 = Duplication	134	60	71.6
COL12A1 = Normal*	234	113	120.7
COL4A3BP = Deletion	410	207	215.64
COL4A3BP = Duplication*	21	**12**	5.63
COL4A3BP = Normal*	133	**72**	69.73
COL5A3 = Deletion*	193	**110**	87.5
COL5A3 = Duplication	217	105	121.9
COL5A3 = Normal*	154	76	81.5

***Statistical significance with p-value < 0.05.**
****Red bolded text indicates stratifications with lower survival probability.**

Log-rank scores show the numerical differences in events that are expected versus observed. Highlighted in red are observed events which are greater than expected events, indicating a decreased probability of survival.

Table 5 lists mean and median days of survival with upper and lower limits of confidence intervals.

Table 5. Mean and Median Days of Survival with Confidence Interval Limits

	Mean*				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
COL4A3BP-Norm**	1514.13	105.79	1307	1721	1384	53.39	1279	1489
COL4A3BP-Del	1733.72	102.34	1533	1934	1336	74.89	1189	1483
COL4A3BP-Dup**	857.09	180.28	504	1210	457	604.32	0	1641
COL4A3BP Overall	1657.58	82.77	1495	1820	1354	60.36	1236	1472
COL12A1-Norm**	1724.72	141.70	1447	2002	1448	90.53	1271	1625
COL12A1-Del	1392.60	77.25	1241	1544	1259	95.19	1072	1446
COL12A1-Dup	1921.35	189.72	1550	2293	1451	184.25	1090	1812
COL12A1 Overall	1657.58	82.77	1495	1820	1354	60.357	1236	1472
COL5A3-Norm**	1550.94	85.39	1384	1718	1516	115.28	1290	1742
COL5A3-Del**	1363.36	88.69	1190	1537	1102	43.33	1017	1187
COL5A3-Dup	1938.09	157.53	1629	2247	1446	68.07	1313	1579
COL5A3 Overall	1657.58	82.77	1495	1820	1354	60.36	1236	1472

***Estimation limited to largest survival time if censored.**
****Statistically significant**

COL5A3 has overlap in the mean limits of confidence intervals between normal and deletion. There are other overlaps, however, only statistically significant stratifications are mentioned.

COL4A3BP does not have a mean overlap in confidence intervals between normal and duplication, which are also statistically significant.

Figure 11 illustrates the mean survival time, measured in days, for each stratification. Triangle markers colored in blue indicate the statistically significant results while the vertical line at 1,657 indicates the overall mean survival for each stratified group as a baseline.

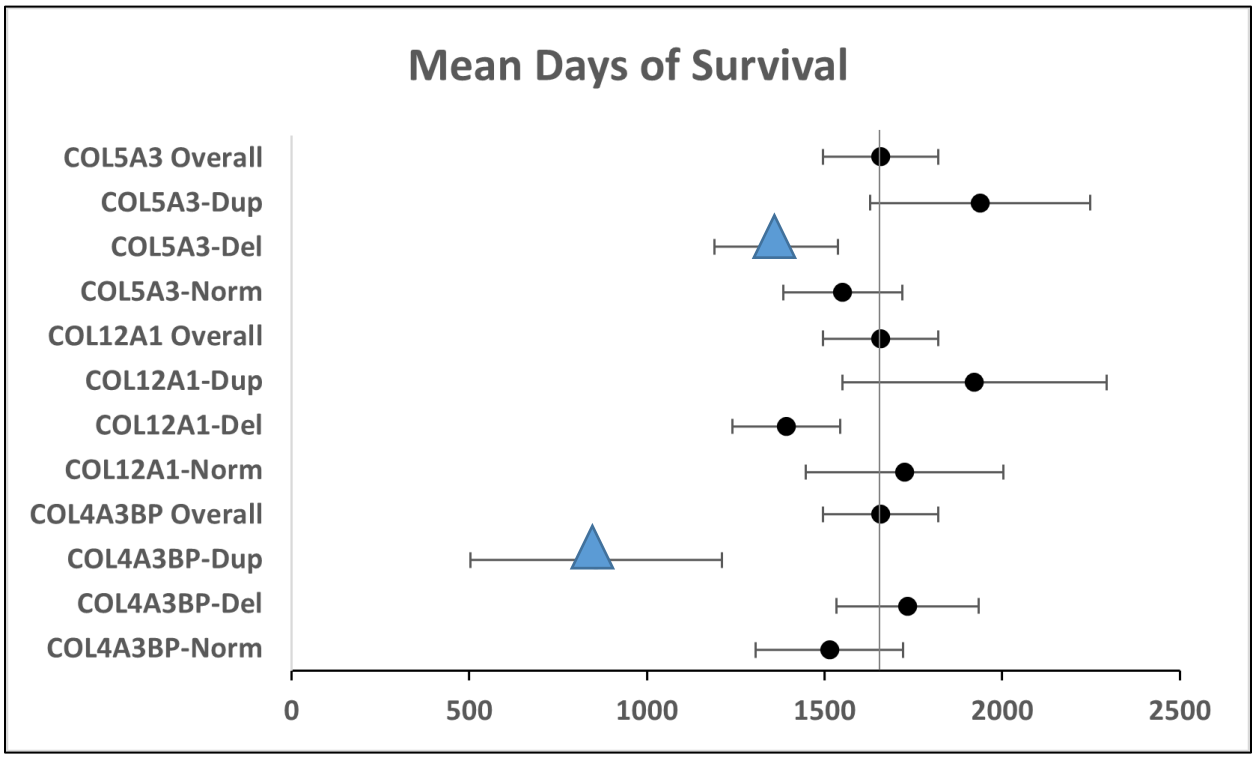


Figure 11. Mean Days of Survival Forest Plot

In testing for the assumptions of the Cox-proportional hazard model, figures in Appendix F are introduced which display the log minus log (LML) plots for each gene.

In each of these genes, the curves never meet and are considered curved parallel, which indicates that the assumptions of the proportional hazards model may have been met and not violated (In & Lee, 2019). These plots primarily illustrate that one stratification is easily distinguishable from the other.

To illustrate the hazard ratios over time, it is appropriate to use a time varying coefficient model for each gene (Zhang et al., 2015). Figures 12, 13, and 14 display how hazard ratios vary over time. The green line is the beta line and is the reference for no effect. The hazard ratio for

each statistically significant finding is represented by the blue line and represents the average hazard over time, except for COL12A1, which had no stratified significant finding. The x-axis is not to linear scale as this visualization has transformed the hazard over time to illustrate time variation. The dashed lines represent the upper and lower limits of the 95% confidence intervals. None of these hazard ratio models are time constant, although Figure 16 is closer to being time constant than the others.

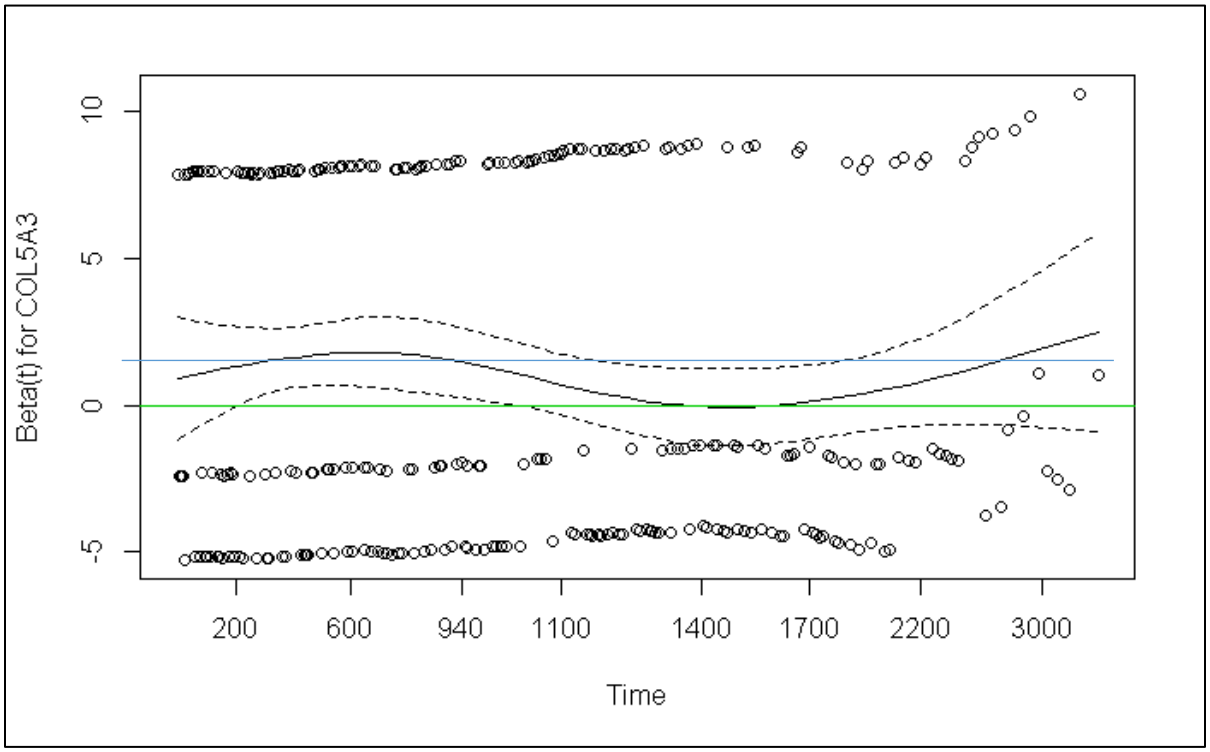


Figure 12. Time-Varying Hazard for COL5A3.
***Blue line represents duplication with HR = 1.349.**

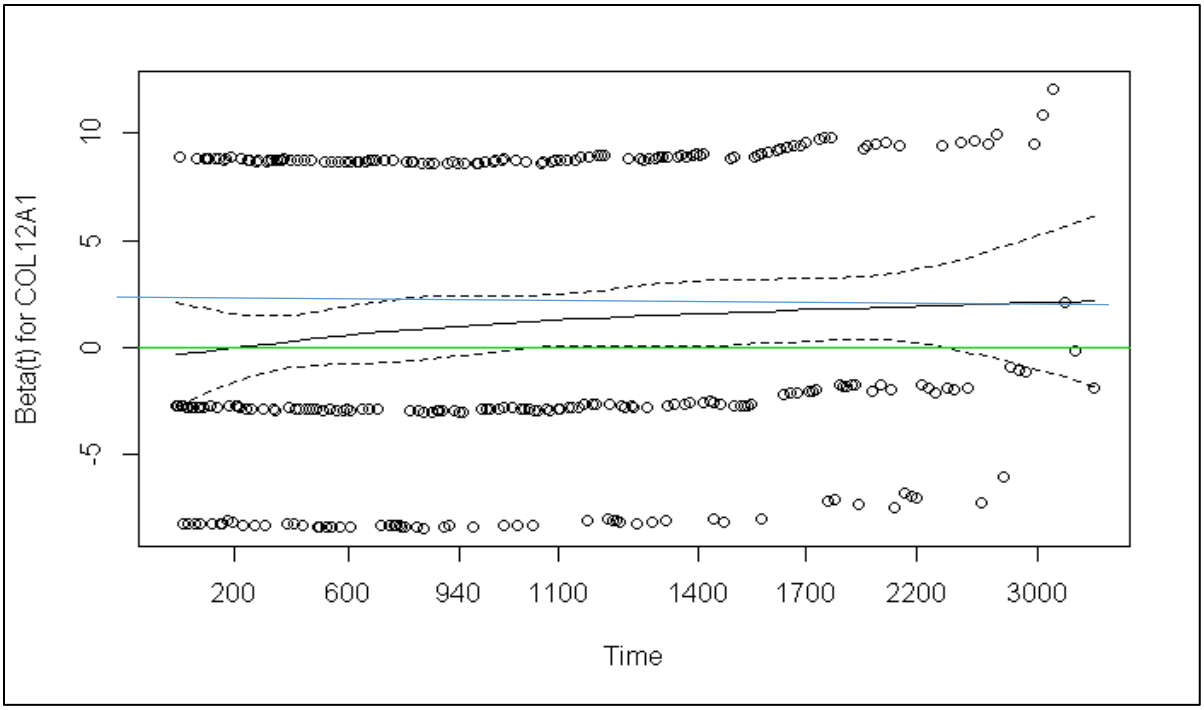


Figure 13. Time-Varying Hazard for COL12A1.
***Blue line represents deletion with HR = 2.077**

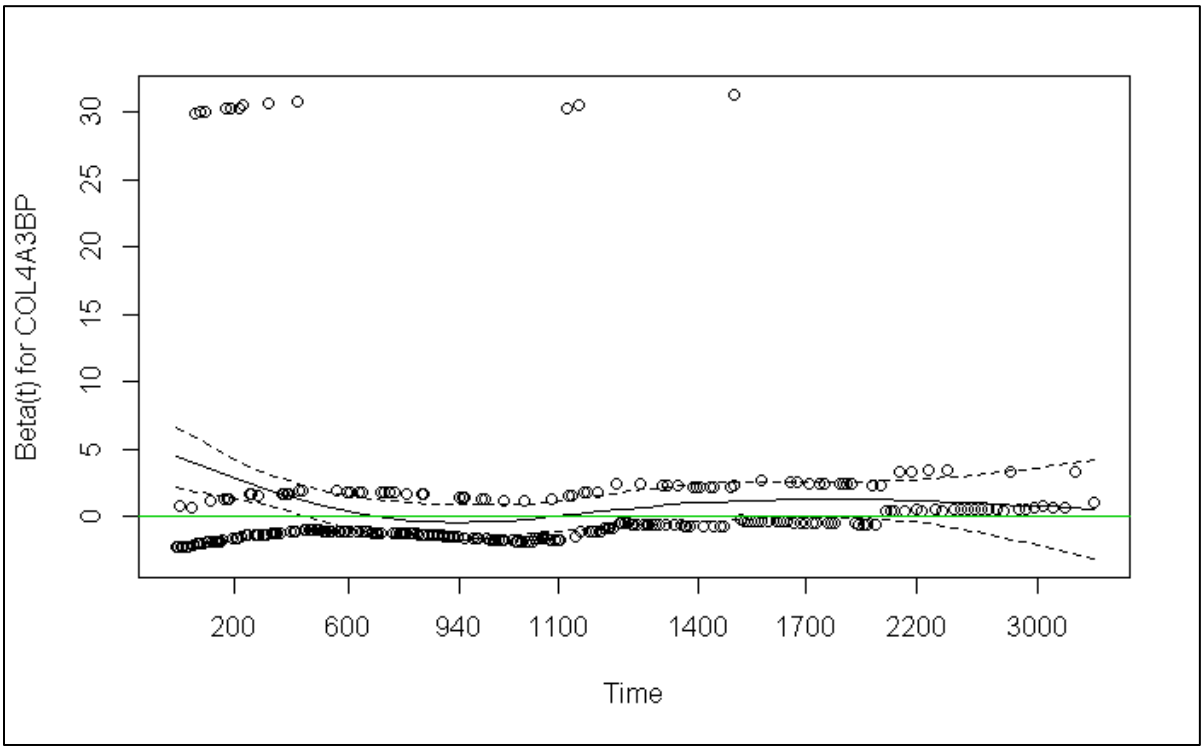


Figure 14. Time-Varying Hazard for COL4A3BP.
***No blue line as no stratified finding was statistically significant.**

The time-varying coefficient models, according to these plots, illustrate that these survival models violate one of the assumptions of the Cox-proportional hazards model, which is that the hazard is constant over time.

Discussion

The micro-ecosystem of neoplasms includes numerous types of cells, growth factors, ECMs, and enzymes which contribute to remodeling (Cho et al., 2015). These specific parts of the tumor micro-environment can all potentially propagate tumor invasion and growth.

Collagen, an ECM protein, is found in all tissues in the human body and is important in tissue homeostasis (Bonnans et al., 2014). Collagen also undergoes continual but controlled remodeling modulated by interactions between cells and its micro-habitat (Cho et al., 2015). Multiple and various interactions which inhibit, or secret, ECM parts may create a pro-tumor environment, which may be defined as host-tumor interactions (Zigrino et al., 2005).

Remodeling changes, such as increased ECM stiffness due to activated fibroblasts, along with abnormal ECM composition, are correlated and documented in multiple cancers (Cho et al., 2015). These remodeling changes lead to increased invasion and migration of tumors but were originally thought of as protection to tumor cell migration.

An important component of collagen which provides tensile strength and elasticity are fibrous proteins (Cho et al., 2015). However, these can become atypical with some tumors which propagate tumor cell growth.

One of the CNV genes in this study has function regarding fibrous proteins. The COL12A1 gene is a protein coding gene which is associated with fibrils and type I collagen which has been found to interact between the fibrils and the protein matrix (Weizmann Institute of Science - COL12A1, 2021). This translates to having effects on tensile strength of the ECM.

This could potentially have effects on tumorigenesis or drug resistance mechanisms such as previously found cisplatin resistance (Helleman et al., 2005).

Numerous fibrous proteins may also affect collagen function in other ways. Collagen I and fibronectin have mechanisms involved in cell adhesion to various surfaces (Moser et al., 1996; Kenny et al., 2008). The same mechanism has also been found with hyaluronan which also plays a role in ovarian cancer metastasis (Ween et al., 2011). This can potentially affect clearance of apoptotic cells as fibronectin has been to inhibit apoptosis (Sakai et al., 2001).

The COL4A3BP gene has a different coding mechanism than the other two statistically significant gene findings but also codes for other proteins. This gene codes for phosphorylation of a terminal on a chain of type IV collagen and affects the ceramide transporter (CERT-I) protein I (Weizmann Institute of Science – COL4A3BP, 2021). CERT proteins are known for interacting with ECM components such as collagen type IV, as well as having effect on the clearance of apoptotic cells (Bode et al., 2014). This may affect the signaling in cell death which might have a direct effect on tumorigenesis (Scheffer et al., 2011). Clearing apoptotic cells is considered to have an anti-inflammatory affect, as well as an anti-tumor affect, and defects in the clearance of apoptotic cells is linked to multiple disease states (Poon et al., 2014). It is possible this is a mechanism behind why mutations in this CNV affect survival of ovarian cancer patients. However, this mechanism with CNV mutations should be researched more.

The third statistical finding also relates to adhesion mechanisms which may contribute to cancer tumorigenesis. The COL5A3 gene is a protein encoding gene which codes for a type of less abundant fibrillar collagen (Weizmann Institute of Science – COL5A3, 2021). This gene also codes for heparin binding of alpha3(V) chains which may affect adhesion which can affect cell apoptotic clearance (Imamura et al., 2000). Mutations in this gene also cause connective

tissue abnormalities, which could potentially affect the ovaries, but are more commonly seen in the disease grouping called Ehlers-Danlos Syndrome type III (Weizmann Institute of Science - MalaCards, 2021). This could potentially be a causal mechanism in ovarian cancer as collagen is a central component in neoplasm fibrosis (Xu et al., 2019).

While the models, according to the time-varying hazard models, violate assumptions of the Cox-proportional hazard model, they are still useful in mapping the changing hazard ratio (Zhang et al., 2015). However, generally these models are still considered valid provided there is not an extreme crossing of survival curves in the Kaplan-Meier survival plots, which is the case in this study. Other methods have been developed that may handle time-varying covariates more appropriately, such as using the standard deviation of longitudinal measurements for example (Muntner et al., 2011). The independence of censoring is the most important assumption in survival analysis and if this assumption is met, it is assumed the models are useful (Prinja et al., 2010).

The survival analysis Cox models mostly meet the Cox-proportional hazard model assumptions. However, these findings should be further investigated based on the possibility of false discovery previously mentioned in other literature as this is a risk with genetic research (Efron, 2005). The reasoning for this is having an alpha level set at 0.05, which is also what most genetic research uses for alpha (Columbia Public Health, 2019). There are an estimated 30,000 genes in the human body (Human Genome Project FAQ, 2013). If one multiplies 30,000 genes by the alpha (0.05) then there are 1,500 genes that could potentially have a type I error finding that is random (Genetics Home Reference, 2019). Type I error bias should be kept in mind as the primary limitation to this study. However, very few genetic studies seem to use any controls for false discovery, such as a Bonferroni p-value correction for example and this is

generally due to lack of quality and quantity of data (Dahiru, 2011). The Bonferroni p-value correction is defined as dividing alpha by some factor in the study (Bonferroni, 1936). In this study, 30,000 genes could be used which would give a p-value threshold of 0.0000016 which can be quite difficult to meet considering the small amount of data found in genetic databases. Another factor that could be used are the 55 collagen genes which still leaves a difficult p-value threshold of 0.0009. This would greatly detract from the power of the study. The Bonferroni correction, while quite popular, may not be a valid method to use in complex or complicated situations such as rare interactions or low-frequency events which can be found in genetic research. To remedy this, there are researchers trying to develop more robust by decreasing type I error bias while not decreasing the study power (Yi et al., 2014).

There is a sample selection bias to take into account, which has been documented in TCGA literature (Solomon et al., 2008). TCGA has this bias because TCGA collects only quality samples of tumors or tissues for sequencing. It has been recommended that TCGA select multiple quality tumors or tissues for analysis. However, samples need to be readable during sequencing and having high quality samples means less errors during sequencing and less type I error bias. Because of this sample quality requirement, I disagree with the current literature on this issue.

There is also a sample bias in this study where most of the population were Caucasian women. There were very few African American, Pacific Islander, or Asian women in the study population. This may also be due to minority groups having a lack of access to healthcare or socioeconomic reasons (Benz et al., 2011; Riley, 2012). Therefore, it is difficult to extrapolate these results out to the general population.

Obtaining genetic data is quite difficult and expensive as the cost of TCGA as of 2015 is $375 million ("The Future of Cancer Genomics", 2015). There are also few genetic databases, and many are proprietary. Although gene sequencing is becoming cheaper and more accurate, a strong argument can be made for keeping alpha at 0.05 with confidence intervals and not strongly controlling for false discovery due to expense and availability of data (Ulrich, 2016).

There are other methods of survival analysis which may contribute to controlling for type I error. Bayesian survival analysis (BSA) has been mentioned as a potential model and has been found to control for type I error more than frequentist survival analysis but has increased type II error bias (Kelter, 2020). It should also be noted that while BSA may be less prone to type I error bias it generally requires subjective data analysis and expert knowledge of parameters and potential prior estimation to perform as well as frequentist survival analysis (Omurlu et al., 2009).

Other methods were applied as well to help control for type I errors such as the log-rank test being used over the Wilcoxon test and the Breslow method used in place of the Efron method in the Cox-proportional hazard model. The Wilcoxon test is generally used when the assumptions of the Cox-proportional hazard models are violated (Martinez, 2007). Otherwise, it is considered the standard to use the log-rank test in survival analysis. Another method for selecting the log-rank test over the Wilcoxon test is that the Wilcoxon test is generally more sensitive to differences early in the survival curve, or rather on the left side of the curve where the time value is lower or early. The log-rank test tends to be more sensitive later in time. This study didn't have differences early in time except for the COL4A3BP gene and thus, the log-rank is the better test to use.

An argument for not controlling for false discovery are the underlying mechanisms of collagen which have multiple and various effects on ovarian neoplasms (Xu et al., 2019). Selecting a subset of specific genes with known effects on a specific neoplastic tissue seems appropriate. Generally, one would not randomly pick a set of genes to investigate where the gene is not expressed in the tissue being studied. This strengthens the argument for using alpha at 0.05.

The methods for CNV detection with high-throughput sequencing also contribute to false discovery rates (Jiang et al., 2018). CNV detection is due to read coverage of the genome and relies on multiple variables which include mappability, local characteristics of gene sequence, and having CNV itself (Sims et al., 2014). An increase in observations would be appropriate to help control for this as well as appropriate normalization of data. It is also possible that a combination of TCGA repository and future cancer databases could be utilized in combination to increase observation numbers, such as the Genotype Tissue Expression Project (GTEx) which has been combined with the TCGA as of 2018 (Wang et al., 2018). It is possible that this future research structure may not be feasible with technology advancing in genomics at a fast rate as sequencing methods have been evolving rather quickly (Davis, 2015). However, TCGA is an enormous longitudinal archive of genetic information, and more CNV survival analysis studies should be performed on other cancer types to help create a baseline knowledge of CNV and effects on neoplasms.

While objectivity is a strong goal to maintain in any study and can be difficult to achieve, finding appropriate data in genetics is also difficult, therefore using a 95% confidence interval is acceptable; stipulating that said research is further explored or reproduced with other genetic data and statistical methods are applied appropriately.

Because there are evolutionary artifacts in the TCGA germline based on natural selection, future CNV research should possibly be performed on a new cancer repository (Webster et al., 2019). Evolutionary artifacts are characterized as when X and Y chromosomes share a common evolutionary source with high resemblance. Pseudogenes can be considered a type of evolutionary artifact and are defined as a DNA sequence that resembles a gene but is inactive as it has mutated over time, generally due to evolution (Tutar, 2012). It is most likely that the evolutionary artifacts and pseudogenes have not introduced much bias into this study as they are generally controlled for, identified, and researched during sequencing. However, one should be aware that they exist and that they are not able to be controlled for in this research by the author, as quantity and quality of data are limiting factors. Evolutionary artifacts would most likely not make a difference in treatment options that may arise from research such as this.

Another bias is that data collection was done by multiple people at multiple locations and sequencing done at three separate locations, as this is secondary data (The Cancer Genome Atlas Program, 2019). It is difficult to analyze the integrity of the data in TCGA. However, care has been taken in the data collection process for use in quantitative polymerase chain reaction (qPCR) and uses a genetic reference to allow qPCR normalization which has allowed greater consistency when performing research from the TCGA database (Krasnov et al., 2019). As previously stated, there has even been recent work where the GTEx was combined with the TCGA into a pipeline which unifies RNA sequencing data and therefore allows greater normalization (Wang et al., 2018).

Studies such as this can be considered foundational to future personalized cancer treatments and target therapy. However, most antineoplastic molecules are of a narrow therapeutic index which means that small dose alterations may lead to toxicity; thus,

chemotherapy dosing changes based on genetics may be unlikely from CNV studies at this time, but this may change in the future (Eaton & Lyman, 2019). There is potential for focused or targeted gene therapy which may be able to use information from studies such as this; especially as targeted gene therapy becomes advanced and viable (Zaimy et al., 2017). Clustered regularly interspaced short palindromic repeats (CRISPR) technology could potentially be used for correcting CNV mutations and is even currently being used to model genomic disorders based on CNV (Tai et al., 2016). However, CRISPR has its own issues in CNV screening regarding false positives as well (De Weck et al., 2018).

A two to three gene CNV signature test could possibly be implemented from studies such as this to present information about probability for survival to the patient and physician. This potential CNV signature could also be assimilated into a gene expression and CNV panel for a more complete study in predictive power of genetics and proteomics in patient survival. This is important for future treatment as there has been research, also based on TCGA, exploring the importance of CNV-based risk scores, and found it to be potentially stronger than other ovarian cancer genomic markers and successfully used to predict cancer survival and highlights the importance of studies such as this (Graf et al., 2021; Paracchini et al., 2020; Van Dijk, et al., 2021).

It is also important to consider various social responsibilities when doing genetic research. Federal and state privacy laws should be accounted for when collecting data and performed appropriately so as not to release private health information without consent. Fortunately, this is not an issue within TCGA as patients were de-identified, although researchers should be aware of the social and ethic responsibilities and complications that arise with medical data.

Genetic laws are currently patchy, incomplete, and lacking thought (Molteni, 2019). Should a genetic data lake or warehouse be created, this could lead to unwarranted and unethical actions by researchers or others. Access to such databases could be used for nefarious purposes and the trafficking or transmission of such genetic information could lead to compromising people's genetic anonymity against their will.

Another social responsibility worth keeping in mind, is that studies such as survival analysis performed on genetic data will most likely have biases; and medical products, treatments, or tests designed from said research should have biases explicitly stated to the product, treatment, or test's respective creators. It would be highly unethical, and could cause harm to patients, if faulty and biased research are used to produce treatments or tests which may lead to incorrect conclusions or decisions about a patient's health. Thus, explicitly stating biases in research helps contribute to this process of preventing unethical and biased conclusions. It would be prudent to have findings replicated in another genetic database when bias may be a major complication. Because of these biases and the lack of research on CNV in general, more studies on collagen in patient genetic profiles for all types of cancer regarding CNV should be conducted. Further research should confirm the importance of using CNV as a metric for outcome prognosis, compare CNV metric to other biomarkers, and further explore the creation of signature panels to be used in practice.

References

Alteryx, Inc. Released 2006. Alteryx Designer, Version 2019.4.8.22007. Irvine, CA: Alteryx, Inc.

Arteche-López, A., Ávila-Fernández, A., Romero, R., Riveiro-Álvarez, R., López-Martínez, M. A., Giménez-Pardo, A., Vélez-Monsalve, C., Gallego-Merlo, J., García-Vara, I., Almoguera, B., Bustamante-Aragonés, A., Blanco-Kelly, F., Tahsin-Swafiri, S., Rodríguez-Pinilla, E., Minguez, P., Lorda, I., Trujillo-Tiebas, M. J., & Ayuso, C. (2021). Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-85182-w

Beck, T. F., Mullikin, J. C., & Biesecker, L. G. (2016). Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. *Clinical Chemistry*, *62*(4), 647–654. https://doi.org/10.1373/clinchem.2015.249623

Benz, J. K., Espinosa, O., Welsh, V., & Fontes, A. (2011). Awareness of racial and ethnic health disparities has improved only modestly over a decade. *Health Affairs (Project Hope)*, *30*(10), 1860–1867. https://doi.org/10.1377/hlthaff.2010.0702

Bode, G. H., Losen, M., Buurman, W. A., Veerhuis, R., Molenaar, P. C., Steinbusch, H. W. M., De Baets, M. H., Daha, M. R., & Martinez-Martinez, P. (2014). Complement Activation by Ceramide Transporter Proteins. *The Journal of Immunology*, *192*(3), 1154–1161. https://doi.org/10.4049/jimmunol.1301673

Bohoris, G. A. (1994). Comparison of the cumulative-hazard and Kaplan-Meier estimators of the survivor function. *IEEE Transactions on Reliability*, *43*(2), 230–232. https://doi.org/10.1109/24.294997

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche E Commerciali Di Firenze*.

Bonnans, C., Chou, J., & Werb, Z. (2014). Remodelling the extracellular matrix in development and disease. *Nature Reviews Molecular Cell Biology*, *15*(12), 786–801. https://doi.org/10.1038/nrm3904

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer*, *89*(3), 431–436. https://doi.org/10.1038/sj.bjc.6601119

Breheny, P. (2019). Tied survival times Estimating survival probabilities. In *https://myweb.uiowa.edu/pbreheny/7210/f15/notes/11-5.pdf*. University of Iowa. https://myweb.uiowa.edu/pbreheny/7210/f15/notes/11-5.pdf

Breslow, N. (1972). Discussion of the paper by D. R. Cox. *J R Statist Soc B*, *B*(34), 216–217. https://dlin.web.unc.edu/wp-content/uploads/sites/1568/2013/04/Lin07.pdf

Broad GDAC Firehose. (2016). Broadinstitute.Org. http://gdac.broadinstitute.org/

Buckley, A. R., Standish, K. A., Bhutani, K., Ideker, T., Lasken, R., Carter, H., Harismendy, O., & Schork, N. J. (2017). Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3770-y

Cancer.org, 2014; Survival Rates of Ovarian cancer. *American Cancer Society*. https://www.cancer.org/cancer/ovarian-cancer/detection-diagnosis-staging/survival-rates.html

Cheon, D.-J., Tong, Y., Sim, M.-S., Dering, J., Berel, D., Cui, X., Lester, J., Beach, J. A., Tighiouart, M., Walts, A. E., Karlan, B. Y., & Orsulic, S. (2013). A Collagen-Remodeling Gene Signature Regulated by TGF-β Signaling Is Associated with Metastasis and Poor Survival in Serous Ovarian Cancer. *Clinical Cancer Research*, *20*(3), 711–723. https://doi.org/10.1158/1078-0432.ccr-13-1256

Chilamkurthy, K. (2020, May 26). *The Cox Proportional-Hazards Model*. Medium. https://towardsdatascience.com/the-cox-proportional-hazards-model-da61616e2e50

Cho, A., Howell, V. M., & Colvin, E. K. (2015). The Extracellular Matrix in Epithelial Ovarian Cancer – A Piece of a Puzzle. Frontiers in Oncology, 5. https://doi.org/10.3389/fonc.2015.00245

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Dahiru, T. (2011). P-Value, a true test of statistical significance? a cautionary note. *Annals of Ibadan Postgraduate Medicine*, *6*(1). https://doi.org/10.4314/aipm.v6i1.64038

Davis, N. (2015, December 10). The evolution of high-throughput genome sequencing. *Search Magazine*. https://www.jax.org/news-and-insights/2015/december/the-evolution-of-high-throughput-genome-sequencing#

De Weck, A., Golji, J., Jones, M. D., Korn, J. M., Billy, E., McDonald, E. R., Schmelzle, T., Bitter, H., & Kauffmann, A. (2018). Correction of copy number induced false positives in CRISPR screens. *PLOS Computational Biology*, *14*(7), e1006279. https://doi.org/10.1371/journal.pcbi.1006279

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., & Jaffrezic, F. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, *14*(6), 671–683. https://doi.org/10.1093/bib/bbs046

Dipiro, J. T., Talbert, R., Yee, G., Matzke, G., Wells, B., & Posey, L. M. (2017). Pharmacotherapy-- a pathophysiologic approach (10th ed., p. 917). Mcgraw-Hill Education.

Eaton, K., & Lyman, G. (2019). Dosing of anticancer agents in adults. *UpToDate*. https://www.uptodate.com/contents/dosing-of-anticancer-agents-in-adults

Efron, B. (2005). Bayesians, Frequentists, and Scientists. *Journal of the American Statistical Association*, *100*(469), 1–5. https://doi.org/10.1198/016214505000000033

Fischer, R. (1925). *Statisical Methods for Research Workers*. Oliver and Boyd. http://psychclassics.yorku.ca/Fisher/Methods/

Gamazon, E. R., & Stranger, B. E. (2015). The impact of human copy number variation on gene expression: Figure 1. *Briefings in Functional Genomics*, *14*(5), 352–357. https://doi.org/10.1093/bfgp/elv017

Ganapathi, M. K., Jones, W. D., Sehouli, J., Michener, C. M., Braicu, I. E., Norris, E. J., Biscotti, C. V., Vaziri, S. A. J., & Ganapathi, R. N. (2015). Expression profile of COL2A1 and the pseudogene SLC6A10P predicts tumor recurrence in high-grade serous ovarian cancer. *International Journal of Cancer*, *138*(3), 679–688. https://doi.org/10.1002/ijc.29815

Garber, J. E., & Offit, K. (2005). Hereditary Cancer Predisposition Syndromes. *Journal of Clinical Oncology*, *23*(2), 276–292. https://doi.org/10.1200/jco.2005.10.042

Genetics Home Reference. (2019). *What is a gene?* Genetics Home Reference. https://ghr.nlm.nih.gov/primer/basics/gene

Gosset, W. S. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1),. https://doi.org/10.2307/2331554

Grace-Martin, K. (2018, August 6). *Six Types of Survival Analysis and Challenges in Learning Them*. The Analysis Factor. https://www.theanalysisfactor.com/the-six-types-of-survival-analysis-and-challenges-in-learning-them/

Graf, R. P., Eskander, R., Brueggeman, L., & Stupack, D. G. (2021). Association of Copy Number Variation Signature and Survival in Patients With Serous Ovarian Cancer. *JAMA Network Open*, *4*(6), e2114162. https://doi.org/10.1001/jamanetworkopen.2021.14162

Helleman, J., Jansen, M. P. H. M., Span, P. N., van Staveren, I. L., Massuger, L. F. A. G., Meijer-van Gelder, M. E., Sweep, F. C. G. J., Ewing, P. C., van der Burg, M. E. L., Stoter, G., Nooter, K., & Berns, E. M. J. J. (2005). Molecular profiling of platinum resistant ovarian cancer. *International Journal of Cancer*, *118*(8), 1963–1971. https://doi.org/10.1002/ijc.21599

Holohan, C., Van Schaeybroeck, S., Longley, D. B., & Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer*, *13*(10), 714–726. https://doi.org/10.1038/nrc3599

Huang, X., Ruan, G., & Sun, P. (2021). Estrogen-related receptor alpha copy number variation is associated with ovarian cancer histological grade. *Journal of Obstetrics and Gynaecology Research*, *47*(5), 1878–1883. https://doi.org/10.1111/jog.14741

Human Genome Project FAQ. (2013). Genome.Gov. https://www.genome.gov/human-genome project/Completion-FAQ

IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Imamura, Y., Scott, I. C., & Greenspan, D. S. (2000). The Pro-α3(V) Collagen Chain. *Journal of Biological Chemistry*, *275*(12), 8749–8759. https://doi.org/10.1074/jbc.275.12.8749

In, J., & Lee, D. K. (2019). Survival analysis: part II – applied clinical data analysis. *Korean Journal of Anesthesiology*, *72*(5), 441–457. https://doi.org/10.4097/kja.19183

Januchowski, R., Świerczewska, M., Sterzyńska, K., Wojtowicz, K., Nowicki, M., & Zabel, M. (2016). Increased Expression of Several Collagen Genes is Associated with Drug Resistance in Ovarian Cancer Cell Lines. *Journal of Cancer*, *7*(10), 1295–1310. https://doi.org/10.7150/jca.15371

Jiang, Y., Wang, R., Urrutia, E., Anastopoulos, I. N., Nathanson, K. L., & Zhang, N. R. (2018). CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-018-1578-y

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data Kalbfleisch/The Statistical*. Hoboken, Nj, Usa John Wiley & Sons, Inc. (Original work published 1980)

Kanchi, K. L., Johnson, K. J., Lu, C., McLellan, M. D., Leiserson, M. D. M., Wendl, M. C., Zhang, Q., Koboldt, D. C., Xie, M., Kandoth, C., McMichael, J. F., Wyczalkowski, M. A., Larson, D. E., Schmidt, H. K., Miller, C. A., Fulton, R. S., Spellman, P. T., Mardis, E. R., Druley, T. E., & Graubert, T. A. (2014). Integrated analysis of germline and somatic variants in ovarian cancer. *Nature Communications*, *5*(1). https://doi.org/10.1038/ncomms4156

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Kelter, R. (2020). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics, 36*, 1263–1288. https://doi.org/10.1007/s00180-020-01034-7

Kennedy, M. C. (2019). Survival Analysis | Statistics for Applied Epidemiology | Tutorial 11 [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=sJPti8Yh4k4

Kenny, H. A., Kaur, S., Coussens, L. M., & Lengyel, E. (2008). The initial steps of ovarian cancer cell metastasis are mediated by MMP-2 cleavage of vitronectin and fibronectin. *Journal of Clinical Investigation, 118*(4), 1367–1379. https://doi.org/10.1172/jci33775

Kerkhof, J., Schenkel, L. C., Reilly, J., McRobbie, S., Aref-Eshghi, E., Stuart, A., Rupar, C. A., Adams, P., Hegele, R. A., Lin, H., Rodenhiser, D., Knoll, J., Ainsworth, P. J., & Sadikovic, B. (2017). Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels. *The Journal of Molecular Diagnostics, 19*(6), 905–920. https://doi.org/10.1016/j.jmoldx.2017.07.004

Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: a self-learning text* (p. 54). Springer.

Krasnov, G. S., Kudryavtseva, A. V., Snezhkina, A. V., Lakunina, V. A., Beniaminov, A. D., Melnikova, N. V., & Dmitriev, A. A. (2019). Pan-Cancer Analysis of TCGA Data Revealed Promising Reference Genes for qPCR Normalization. *Frontiers in Genetics, 10*. https://doi.org/10.3389/fgene.2019.00097

LaMorte, W. (2016). *Comparing Survival Curves*. Sphweb.Bumc.Bu.Edu; Boston University School of Public Health. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_survival/BS704_Survival5.html

Lee, A. J. X., Roylance, R., Sander, J., Gorman, P., Endesfelder, D., Kschischo, M., Jones, N. P., East, P., Nicke, B., Spassieva, S., Obeid, L. M., Birkbak, N. J., Szallasi, Z., McKnight, N. C., Rowan, A. J., Speirs, V., Hanby, A. M., Downward, J., Tooze, S. A., & Swanton, C. (2011). CERT depletion predicts chemotherapy benefit and mediates cytotoxic and polyploid-specific cancer cell death through autophagy induction. *The Journal of Pathology*, *226*(3), 482–494. https://doi.org/10.1002/path.2998

Li, L., Bai, H., Yang, J., Cao, D., & Shen, K. (2017). Genome-wide DNA copy number analysis in clonally expanded human ovarian cancer cells with distinct invasive/migratory capacities. *Oncotarget, 8*(9), 15136–15148. https://doi.org/10.18632/oncotarget.14767

Li, Y. (2009). Modeling of Survival Data. In *Applied Survival Analysis* (pp. 1–42). University of Michigan. http://www-personal.umich.edu/~yili/lect4notes.pdf

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Hu, H., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., … Mariamidze, A. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, *173*(2), 400-416.e11. https://doi.org/10.1016/j.cell.2018.02.052

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *Journal of the National Cancer Institute*, *22*(4), 719–745. https://doi.org/10.1093/jnci/22.4.719

Martinez, R. L. M. C. (2007). Diagnostics for Choosing between Log-Rank and Wilcoxon Tests. *Dissertations*. https://scholarworks.wmich.edu/dissertations/895

Molteni, M. (2019, May 1). The US Urgently Needs New Genetic Privacy Laws. *Wired*. https://www.wired.com/story/the-us-urgently-needs-new-genetic-privacy-laws/

Moser, T. L., Pizzo, S. V., Bafetti, L. M., Fishman, D. A., & Stack, M. S. (1996). Evidence for preferential adhesion of ovarian epithelial carcinoma cells to type I collagen mediated by the αA2β1 integrin. *International Journal of Cancer*, *67*(5), 695–701. https://doi.org/3.0.co;2-4">10.1002/(sici)1097-0215(19960904)67:5<695::aid-ijc18>3.0.co;2-4

Muntner, P., Shimbo, D., Tonelli, M., Reynolds, K., Arnett, D. K., & Oparil, S. (2011). The Relationship Between Visit-to-Visit Variability in Systolic Blood Pressure and All-Cause Mortality in the General Population. *Hypertension*, *57*(2), 160–166. https://doi.org/10.1161/hypertensionaha.110.162255

National Center for Biotechnology Information - COL12A1. (2020, October 25). Www.Ncbi.Nlm.Nih.Gov; NCBI. https://www.ncbi.nlm.nih.gov/gene/1303

Omurlu, I. K., Ozdamar, K., & Ture, M. (2009). Comparison of Bayesian survival analysis and Cox regression analysis in simulated and breast cancer data sets. *Expert Systems with Applications*, *36*(8), 11341–11346. https://doi.org/10.1016/j.eswa.2009.03.058

Paracchini, L., Beltrame, L., Grassi, T., Inglesi, A., Fruscio, R., Landoni, F., Ippolito, D., Delle Marchette, M., Paderno, M., Adorni, M., Jaconi, M., Romualdi, C., D'Incalci, M., Siravegna, G., & Marchini, S. (2020). Genome-wide Copy-number Alterations in Circulating Tumor DNA as a Novel Biomarker for Patients with High-grade Serous Ovarian Cancer. *Clinical Cancer Research*, *27*(9), 2549–2559. https://doi.org/10.1158/1078-0432.ccr-20-3345

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series, 5*(50), 157–175. http://www.economics.soton.ac.uk/staff/aldrich/1900.pdf

Perry, G. H. (2008). The evolutionary significance of copy number variation in the human genome. *Cytogenetic and Genome Research*, *123*(1-4), 283–287. https://doi.org/10.1159/000184719

Peto, R., & Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series a (General)*, *135*(2), 185. https://doi.org/10.2307/2344317

Poon, I. K. H., Lucas, C. D., Rossi, A. G., & Ravichandran, K. S. (2014). Apoptotic cell clearance: basic biology and therapeutic potential. *Nature Reviews Immunology*, *14*(3), 166–180. https://doi.org/10.1038/nri3607

Prinja, S., Gupta, N., & Verma, R. (2010). Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine*, *35*(2), 217–221. https://doi.org/10.4103/0970-0218.66859

Reid, B. M., Permuth, J. B., Chen, Y. A., Fridley, B. L., Iversen, E. S., Chen, Z., Jim, H., Vierkant, R. A., Cunningham, J. M., Barnholtz-Sloan, J. S., Narod, S., Risch, H., Schildkraut, J. M., Goode, E. L., Monteiro, A. N., & Sellers, T. A. (2019). Genome-wide Analysis of Common Copy Number Variation and Epithelial Ovarian Cancer Risk. *Cancer Epidemiology Biomarkers & Prevention*, *28*(7), 1117–1126. https://doi.org/10.1158/1055-9965.epi-18-0833

Ricciardelli, C., & Rodgers, R. (2006). Extracellular Matrix of Ovarian Tumors. *Seminars in Reproductive Medicine*, *24*(4), 270–282. https://doi.org/10.1055/s-2006-948556

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C. J., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology–Head and Neck Surgery*, *143*(3), 331–336. https://doi.org/10.1016/j.otohns.2010.05.007

Riley, W. (2012). Health Disparities: Gaps in Access, Quality and Affordability of Medical Care. *Trans Am Clin Climatol Assoc.*, *123*, 167–174. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540621/

RStudio Team (2021). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Sakai, T., Johnson, K. J., Murozono, M., Sakai, K., Magnuson, M. A., Wieloch, T., Cronberg, T., Isshiki, A., Erickson, H. P., & Fässler, R. (2001). Plasma fibronectin supports neuronal survival and reduces brain injury following transient focal cerebral ischemia but is not essential for skin-wound healing and hemostasis. *Nature Medicine*, *7*(3), 324–330. https://doi.org/10.1038/85471

Scheffer, L., Rao Raghavendra, P., Ma, J., & K. Acharya, J. (2011). Ceramide Transfer Protein and Cancer. *Anti-Cancer Agents in Medicinal Chemistry*, *11*(9), 904–910. https://doi.org/10.2174/187152011797655087

Schober, P., & Vetter, T. R. (2018). Survival Analysis and Interpretation of Time-to-Event Data. *Anesthesia & Analgesia*, *127*(3), 792–798. https://doi.org/10.1213/ane.0000000000003653

Shaikh, T. H. (2017). Copy Number Variation Disorders. *Current Genetic Medicine Reports*, *5*(4), 183–190. https://doi.org/10.1007/s40142-017-0129-2\

Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., & Fan, X. (2019). Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Medical Genetics*, *20*(1). https://doi.org/10.1186/s12881-019-0909-5

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., & Eichler, E. E. (2005). Segmental Duplications and Copy-Number Variation in the Human Genome. *The American Journal of Human Genetics*, *77*(1), 78–88. https://doi.org/10.1086/431652

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence* (p. 525). Oxford University Press.

Solomon, D. A., Kim, J.-S., Ressom, H. W., Sibenaller, Z., Ryken, T., Jean, W., Bigner, D., Yan, H., & Waldman, T. (2009). Sample Type Bias in the Analysis of Cancer Genomes. *Cancer Research*, *69*(14), 5630–5633. https://doi.org/10.1158/0008-5472.can-09-1055

Spainhour, J. C. G., Lim, J., & Qiu, P. (2017). GDISC: a web portal for integrative analysis of gene–drug interaction for survival in cancer. *Bioinformatics*, *33*(9), btw830. https://doi.org/10.1093/bioinformatics/btw830

Spainhour, J. C. G., & Qiu, P. (2016). Identification of gene-drug interactions that impact patient survival in TCGA. *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/s12859-016 1255-7

Sutradhar, R., & Austin, P. C. (2018). Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Annals of Epidemiology*, *28*(1), 54–57. https://doi.org/10.1016/j.annepidem.2017.10.014

Tai, D. J. C., Ragavendran, A., Manavalan, P., Stortchevoi, A., Seabra, C. M., Erdin, S., Collins, R. L., Blumenthal, I., Chen, X., Shen, Y., Sahin, M., Zhang, C., Lee, C., Gusella, J. F., & Talkowski, M. E. (2016). Engineering microdeletions and microduplications by targeting segmental duplications with CRISPR. *Nature Neuroscience*, *19*(3), 517–522. https://doi.org/10.1038/nn.4235

TCGA - Ovarian Serous Adenocarcinoma Study. (2018, September 5). www.Cancer.Gov. https://www.cancer.gov/about-nci/organization/ccg/research/structural genomics/tcga/studied-cancers/ovarian

Teng, P.-N., Wang, G., Hood, B. L., Conrads, K. A., Hamilton, C. A., Maxwell, G. L., Darcy, K. M., & Conrads, T. P. (2013). Identification of candidate circulating cisplatin-resistant biomarkers from epithelial ovarian carcinoma cell secretomes. *British Journal of Cancer*, *110*(1), 123–132. https://doi.org/10.1038/bjc.2013.687

Thapar, A., & Cooper, M. (2013). Copy Number Variation: What Is It and What Has It Told Us About Child Psychiatric Disorders? *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(8), 772–774. https://doi.org/10.1016/j.jaac.2013.05.013

The Cancer Genome Atlas Program. (2019). National Cancer Institute; Cancer.gov. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

The Cancer Genome Atlas - Publications. (2019). National Cancer Institute; Cancer.gov. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/publications

The future of cancer genomics. (2015). *Nature Medicine*, *21*(2), 99–99. https://doi.org/10.1038/nm.3801

Tibshirani, R. (1982). A Plain Man's Guide to the Proportional Hazards Model. *Clinical & Investigative Medicine*, *5*(1), 63–68. http://statweb.stanford.edu/~tibs/ftp/plain.pdf

Tutar, Y. (2012). Pseudogenes. *Comparative and Functional Genomics*, *2012*, 1–4. https://doi.org/10.1155/2012/424526

Ulrich, T. (2016, September 13). Opinionome: Can DNA sequencing get any faster and cheaper? *BROADMINDED BLOG*. https://www.broadinstitute.org/blog/opinionome-can-dna sequencing-get-any-faster-and-cheaper

Van Dijk, E., van den Bosch, T., Lenos, K. J., El Makrini, K., Nijman, L. E., van Essen, H. F. B., Lansu, N., Boekhout, M., Hageman, J. H., Fitzgerald, R. C., Punt, C. J. A., Tuynman, J. B., Snippert, H. J. G., Kops, G. J. P. L., Medema, J. P., Ylstra, B., Vermeulen, L., & Miedema, D. M. (2021). Chromosomal copy number heterogeneity predicts survival rates across cancers. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-23384-6

Visintainer, P. (2016, May 19). *Which-statistical-test-is-best-for-comparing-two-survival-curves*. ResearchGate.net. https://www.researchgate.net/post/Which-statistical-test-is-best-for-comparing-two-survival-curves

Wang, Q., Armenia, J., Zhang, C., Penson, A. V., Reznik, E., Zhang, L., Minet, T., Ochoa, A., Gross, B. E., Iacobuzio-Donahue, C. A., Betel, D., Taylor, B. S., Gao, J., & Schultz, N. (2018). Unifying cancer and normal RNA sequencing data from different sources. *Scientific Data*, *5*(1). https://doi.org/10.1038/sdata.2018.61

Webster, T. H., Couse, M., Grande, B. M., Karlins, E., Phung, T. N., Richmond, P. A., Whitford, W., & Wilson, M. A. (2019). Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience*, *8*(7). https://doi.org/10.1093/gigascience/giz074

Ween, M., Oehler, M., & Ricciardelli, C. (2011). Role of Versican, Hyaluronan and CD44 in Ovarian Cancer Metastasis. *International Journal of Molecular Sciences*, *12*(2), 1009–1029. https://doi.org/10.3390/ijms12021009

Weizmann Institute of Science - COL12A1. (2021). Genecards.org. https://www.genecards.org/cgi-bin/carddisp.pl?gene=COL12A1

Weizmann Institute of Science - COL4A3BP. (2021). Genecards.org. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CERT1

Weizmann Institute of Science - COL5A3. (2021). Genecards.org. https://www.genecards.org/cgi-bin/carddisp.pl?gene=COL5A3

Weizmann Institute of Science - MalaCards. (2021). Malacards.org. https://www.malacards.org/card/ehlers_danlos_syndrome

Woodruff, T. K., & Shea, L. D. (2007). The Role of the Extracellular Matrix in Ovarian Follicle Development. *Reproductive Sciences*, *14*(8_suppl), 6–10. https://doi.org/10.1177/1933719107309818

Wu, Y.-H., Chang, T.-H., Huang, Y.-F., Huang, H.-D., & Chou, C.-Y. (2013). COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene*, *33*(26), 3432–3440. https://doi.org/10.1038/onc.2013.307

Xu, L. (2019, April 18). *Lecture 5THE PROPORTIONAL HAZARDSREGRESSION MODEL*. Math UCSD. https://www.math.ucsd.edu/~rxu/math284/slect5.pdf

Xu, S., Xu, H., Wang, W., Li, S., Li, H., Li, T., Zhang, W., Yu, X., & Liu, L. (2019). The role of collagen in cancer: from bench to bedside. *Journal of Translational Medicine*, *17*(1). https://doi.org/10.1186/s12967-019-2058-1

Yi, N., Xu, S., Lou, X.-Y., & Mallick, H. (2014). Multiple comparisons in genetic association studies: a hierarchical modeling approach. *Statistical Applications in Genetics and Molecular Biology*, *13*(1). https://doi.org/10.1515/sagmb-2012-0040

Zaimy, M. A., Saffarzadeh, N., Mohammadi, A., Pourghadamyari, H., Izadi, P., Sarli, A., Moghaddam, L. K., Paschepari, S. R., Azizi, H., Torkamandi, S., & Tavakkoly-Bazzaz, J. (2017). New methods in the diagnosis of cancer and gene therapy of cancer based on nanoparticles. *Cancer Gene Therapy*, *24*(6), 233–243. https://doi.org/10.1038/cgt.2017.16

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, *16*, 172–183. https://doi.org/A copy number variation map of the human genome

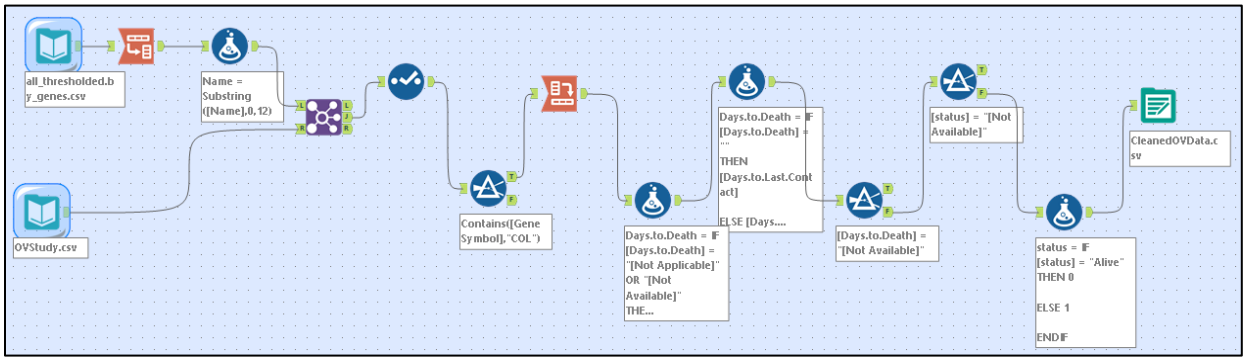
Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., & Groothuis-Oudshoorn, C. G. M. (2018). Time-varying covariates and coefficients in Cox regression models. *Annals of Translational Medicine*, *6*(7), 121–121. https://doi.org/10.21037/atm.2018.02.12

Zhao, F., Wang, Y., Zheng, J., Wen, Y., Qu, M., Kang, S., Wu, S., Deng, X., Hong, K., Li, S., Qin, X., Wu, Z., Wang, X., Ai, C., Li, A., Zeng, L., Hu, J., Zeng, D., Shang, L., & Wang, Q. (2020). A genome-wide survey of copy number variations reveals an asymmetric evolution of duplicated genes in rice. *BMC Biology*, *18*(1). https://doi.org/10.1186/s12915-020-00798-0

Zigrino, P., Löffek, S., & Mauch, C. (2005). Tumor–stroma interactions: their role in the control of tumor cell invasion. *Biochimie*, *87*(3-4), 321–328. https://doi.org/10.1016/j.biochi.2004.10.025

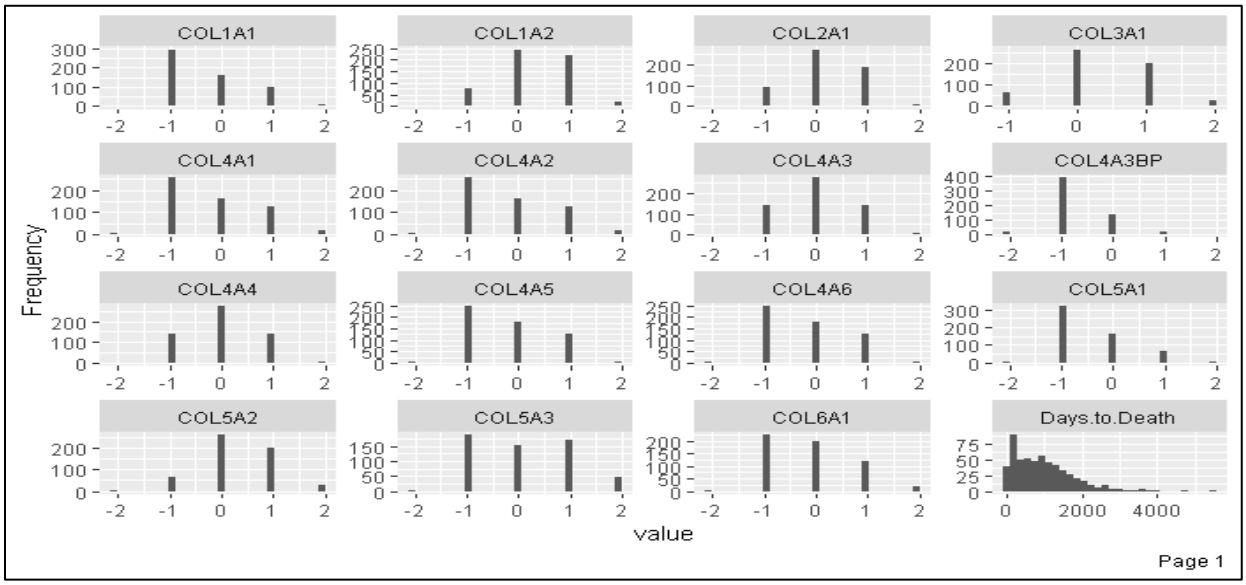
Zwiener, I., Blettner, M., & Hommel, G. (2011). Survival Analysis. *Deutsches Aerzteblatt Online*, *108*(10), 163–169. https://doi.org/10.3238/arztebl.2011.0163

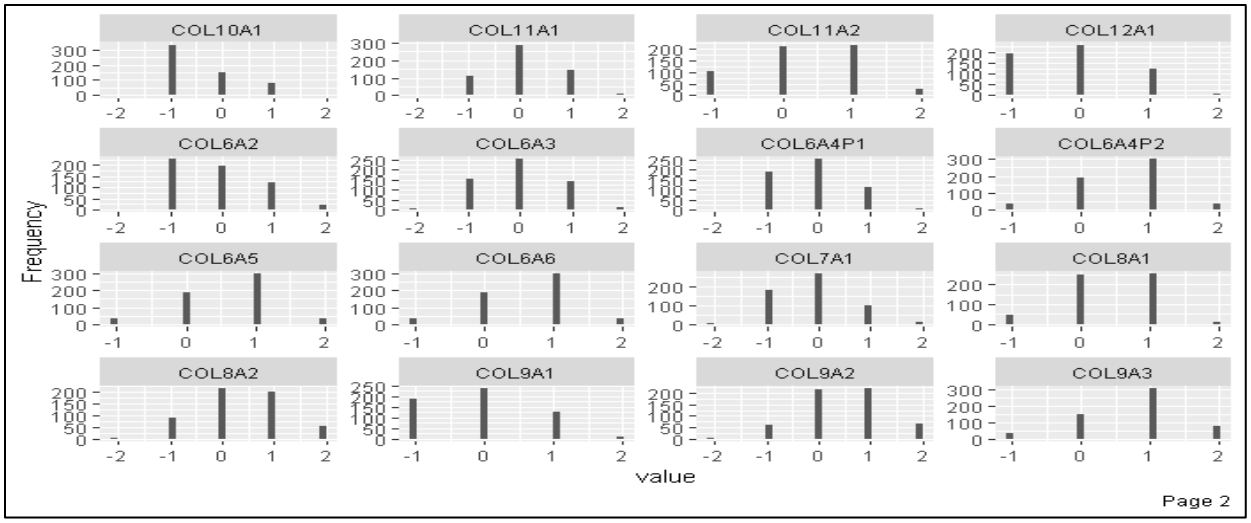
Appendix A – Alteryx Workflow

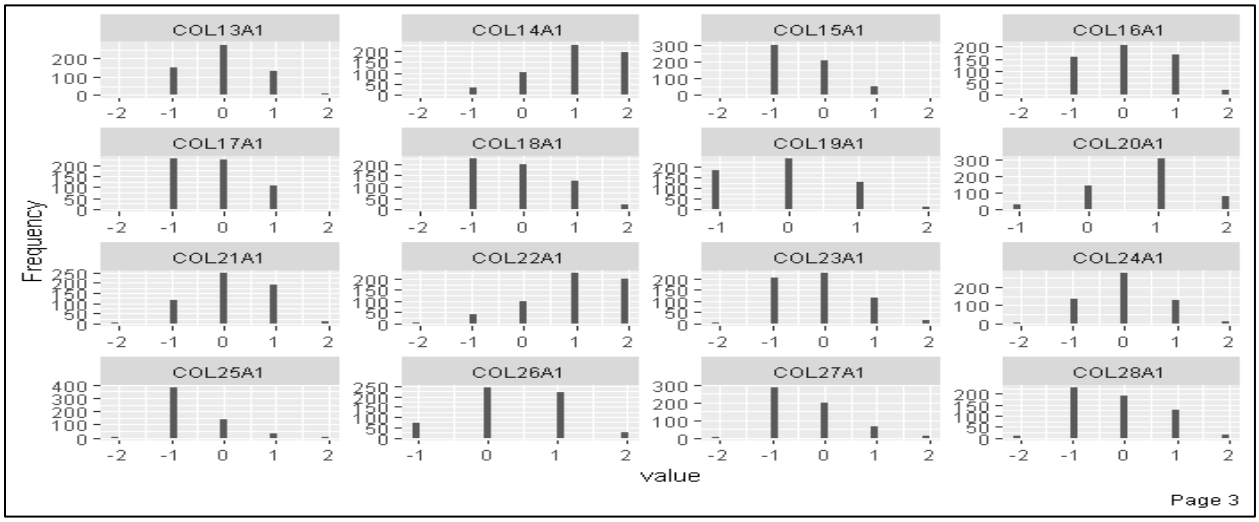


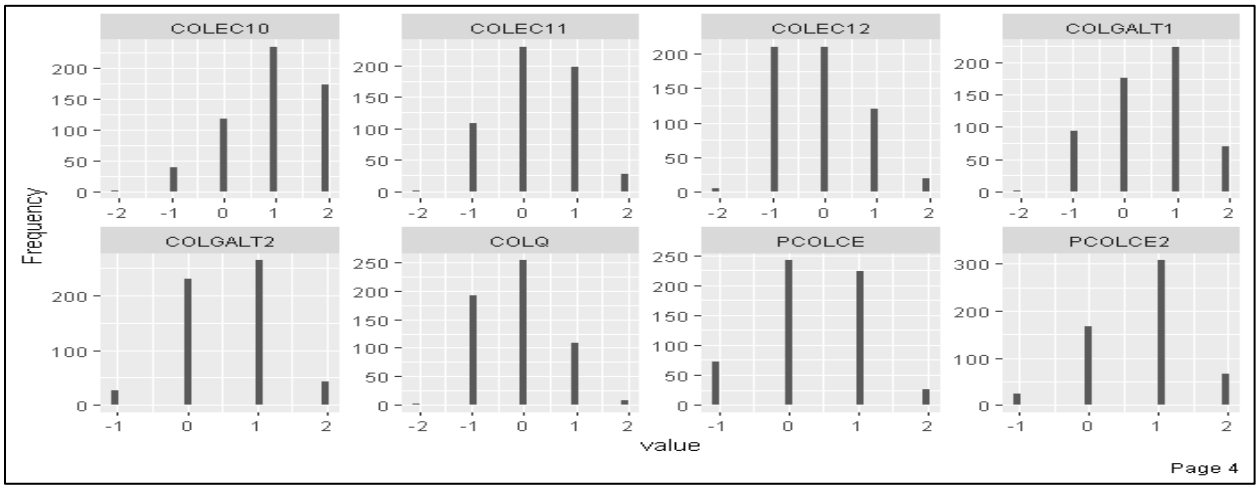
This figure shows the workflow in Alteryx for data wrangling. Two gene data was pivoted and then joined with the patient data on patient TCGA ID numbers. The data types were then changed, and genes were filtered down to only collagen genes. The collagen genes were checked to make sure they had expression in the ovaries. Any non-collagen genes or genes not related to collagen in any way were removed from analysis. Time to event labeled as the column "Days.to.Death" was put through various formulas and filtered. The event status was changed to a numerical variable. The final CSV file can be found in a GitHub Repository at https://github.com/hodgesr2/Pan-Collagan-Ovarian-Cancer-Study-from-TCGA.

Appendix B – Distribution of Collagen Genes









Appendix B shows the various distributions of each stratified gene and the column “Days.to.Death”, which is the time to event, along with the population counts.

Appendix C

COL12A1

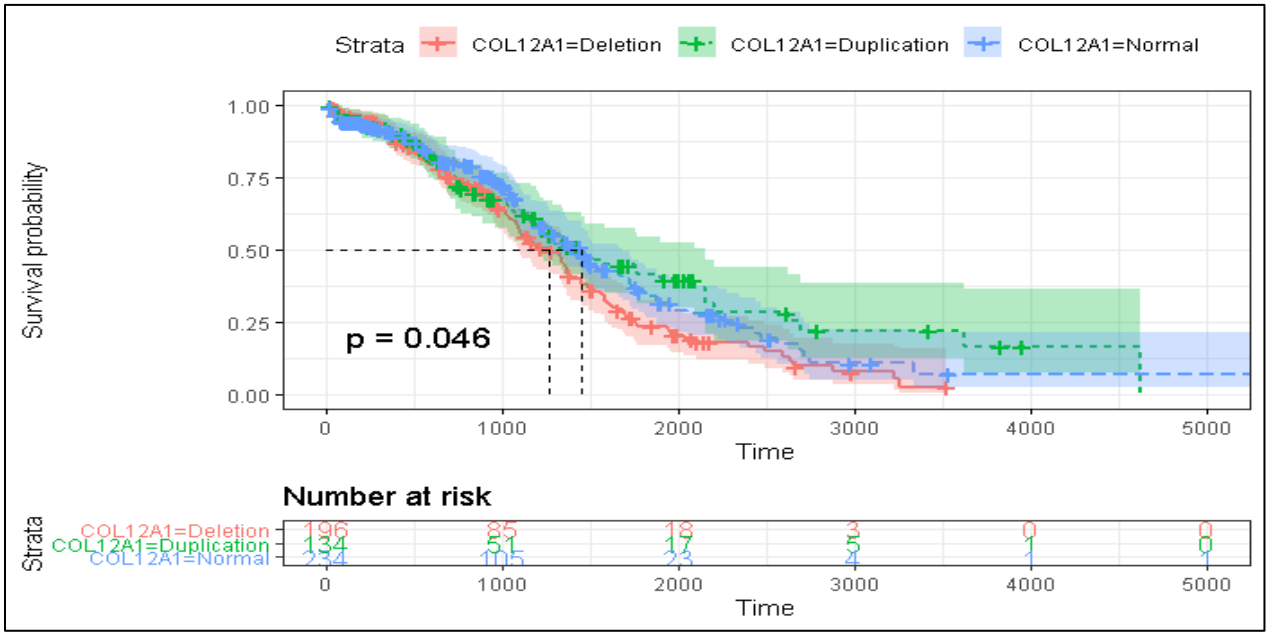


Figure COL12A1 survival has stratified curve of CNV changes (n = 564). Time is measured in days. Green illustrates duplicated copy number variation showing increased survival probability while red illustrates deletion of copy number variation displaying lower survival probability. Blue displays normal copy number variation. The model's p-value of 0.046 demonstrates a statistically significant model when stratifying between COL12A1 duplication and deletion with regards to survival time. The dotted lines display the 50% survival probability between duplication, deletion, and normal copy number variation.

COL4A3BP

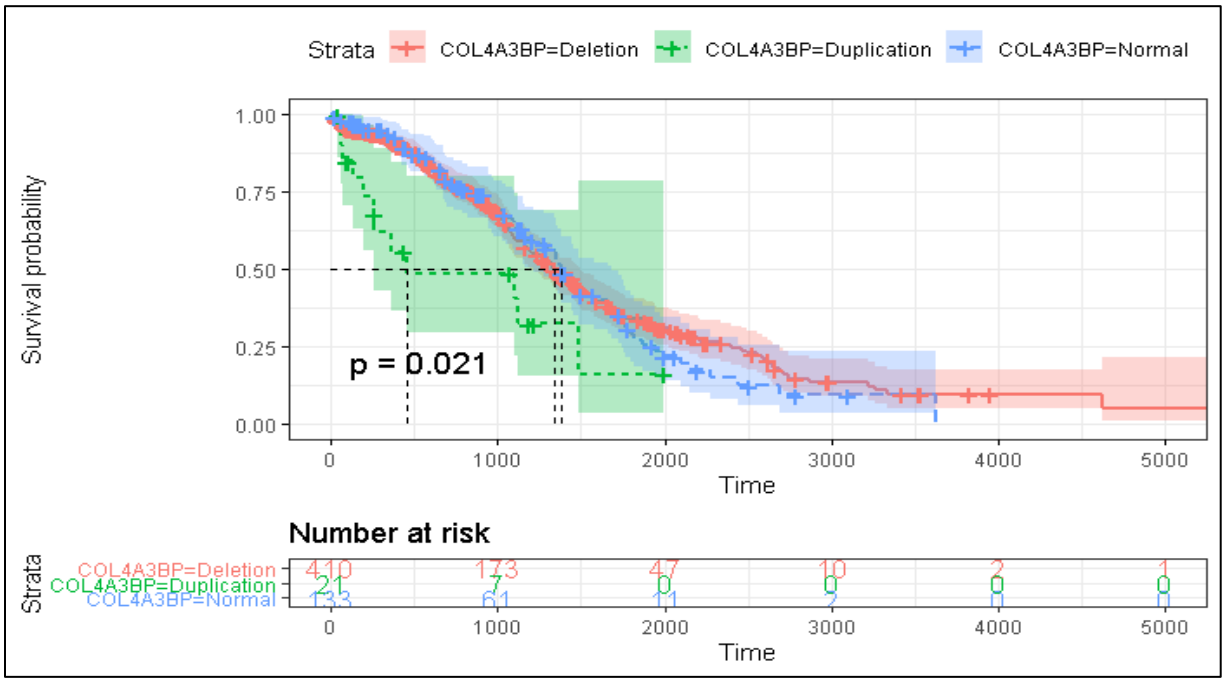


Figure COL4A3BP survival has stratified curve of CNV changes (n = 564). Time is measured in days. Green illustrates duplicated copy number variation displaying a decreased survival probability while red illustrates deletion of copy number variation. Blue displays normal copy number variation illustrating an increased survival probability compared to copy number variation duplication. The model's p-value of 0.021 demonstrates a statistically significant model when stratifying between COL12A1 duplication vs normal and deletion copy number variation with regards to survival time. The dotted lines display the 50% survival probability between duplication, deletion, and normal copy number variation.

COL5A3

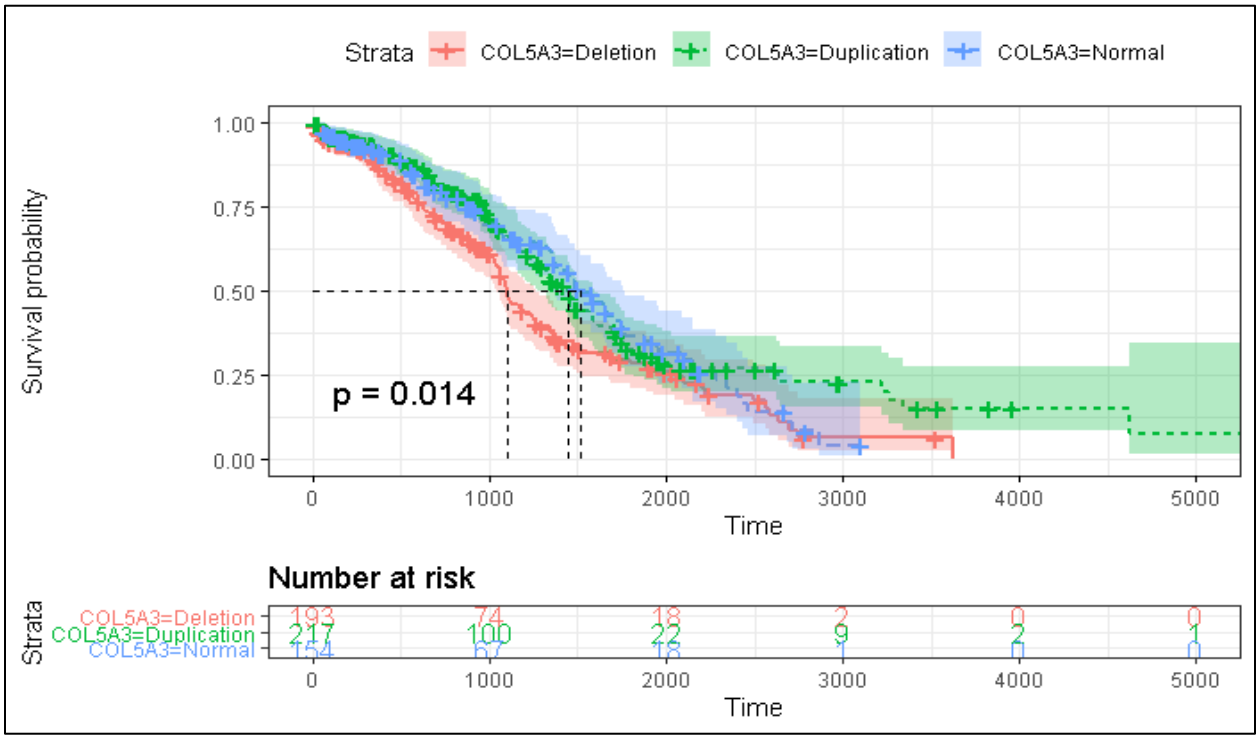
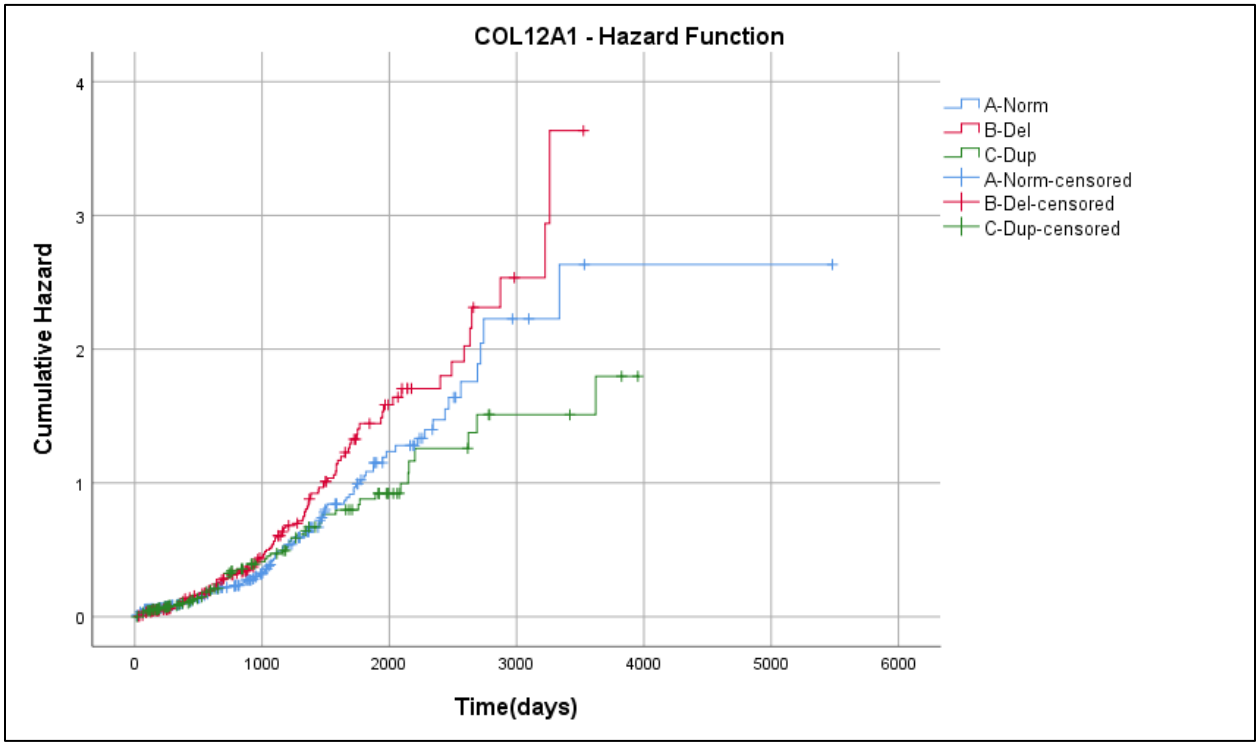


Figure COL5A3 survival has stratified curve of CNV changes (n = 564). Time is measured in days. Green illustrates duplicated copy number variation displaying an increased survival probability while red illustrates deletion of copy number variation illustrating a lower survival probability when compared to normal and duplicated copy number variation. Blue displays normal copy number variation. The model's p-value of 0.014 demonstrates a statistically significant model when stratifying between COL5A3 deletion vs normal and duplicated copy number variation with regards to survival time. The dotted lines display the 50% survival probability between duplication, deletion, and normal copy number variation.

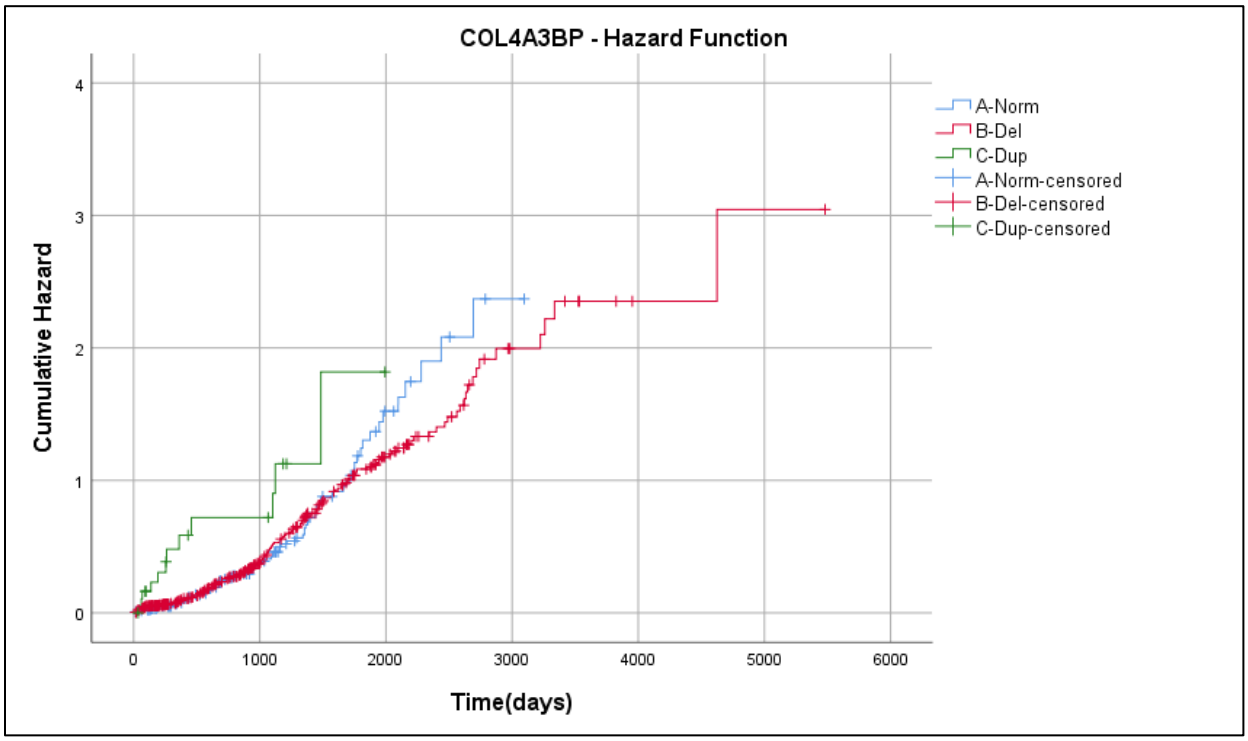
Appendix D

COL12A1



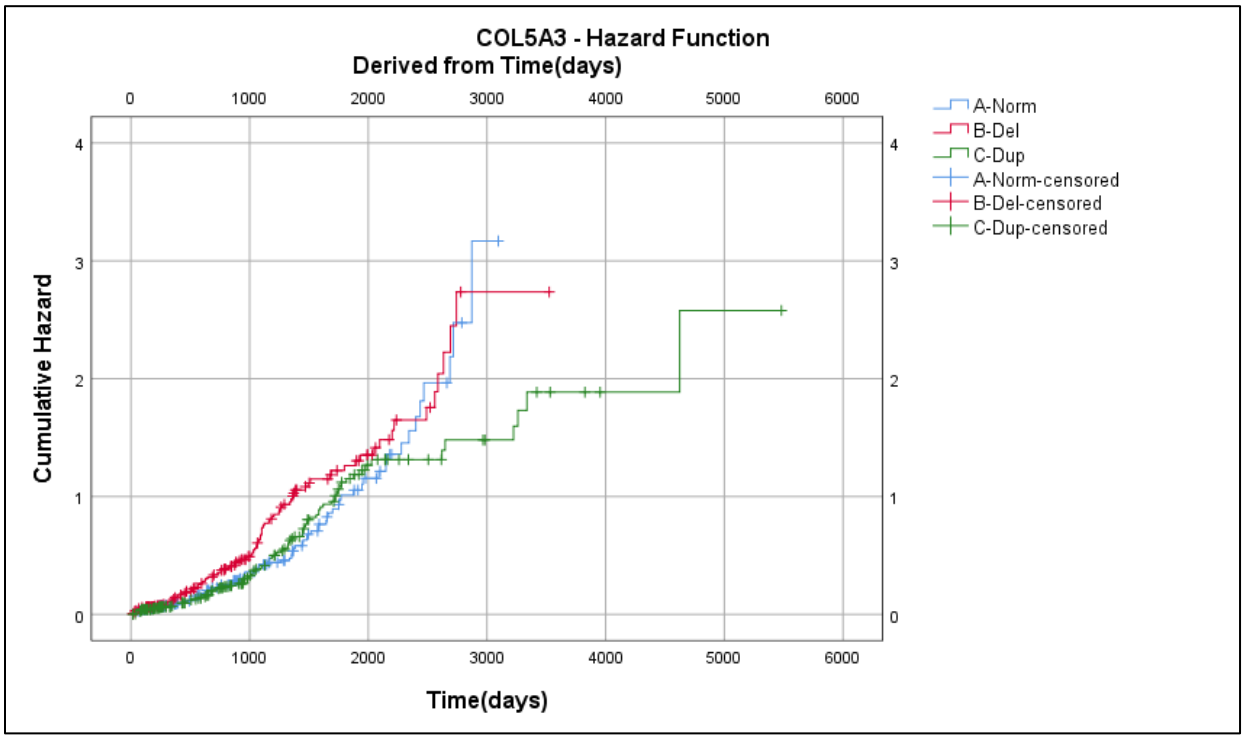
This figure shows the hazard function for COL12A1 which illustrates the cumulative hazard ratio over time for each stratification. The deletion stratification shows the worst survival probability.

COL4A3BP



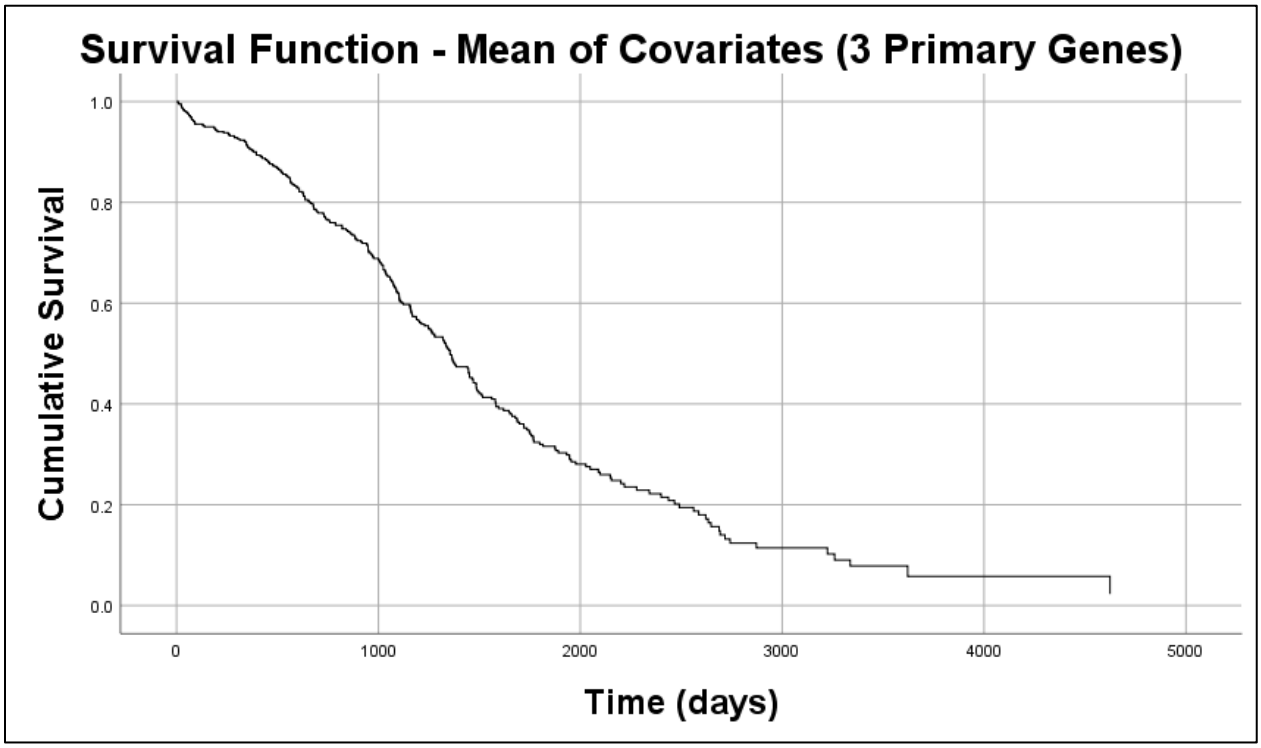
This figure shows the hazard function for COL4A3BP which illustrates the cumulative hazard ratio over time for each stratification. The duplication stratification shows the worst survival probability.

COL5A3

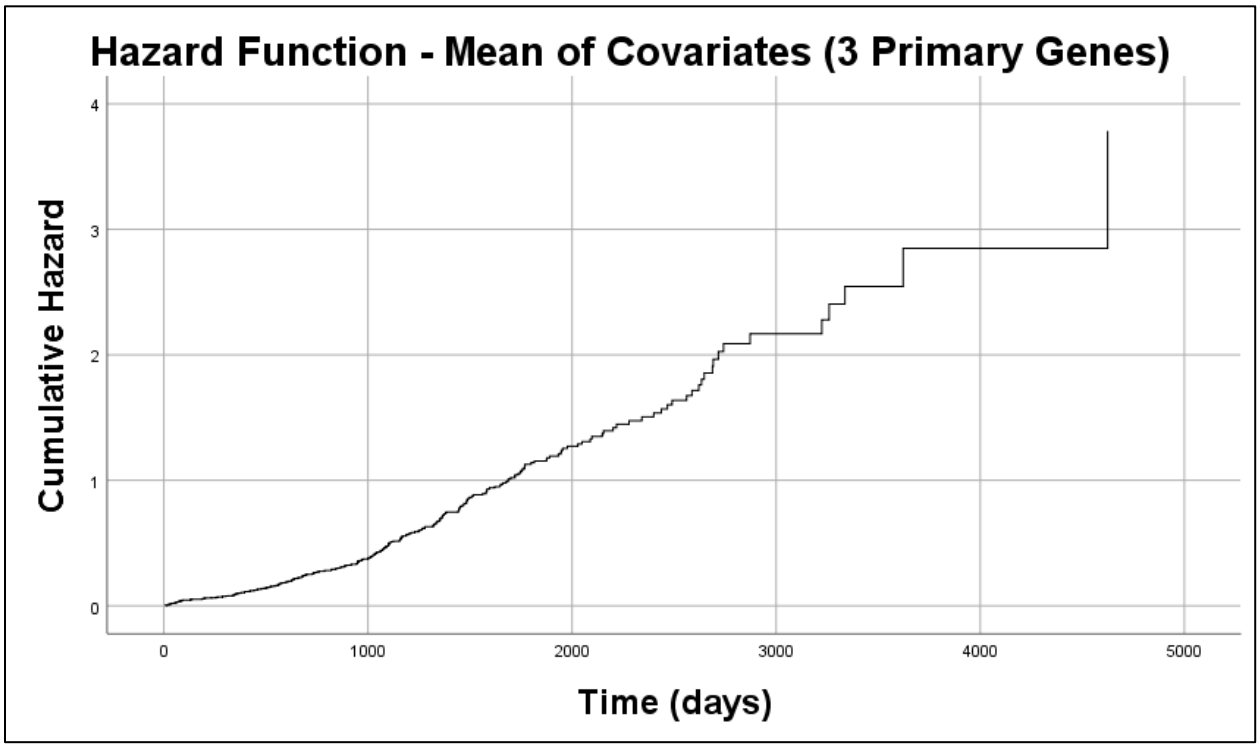


This figure shows the hazard function for COL5A3 which illustrates the cumulative hazard ratio over time for each stratification. The duplication stratification shows better survival probability.

Appendix E



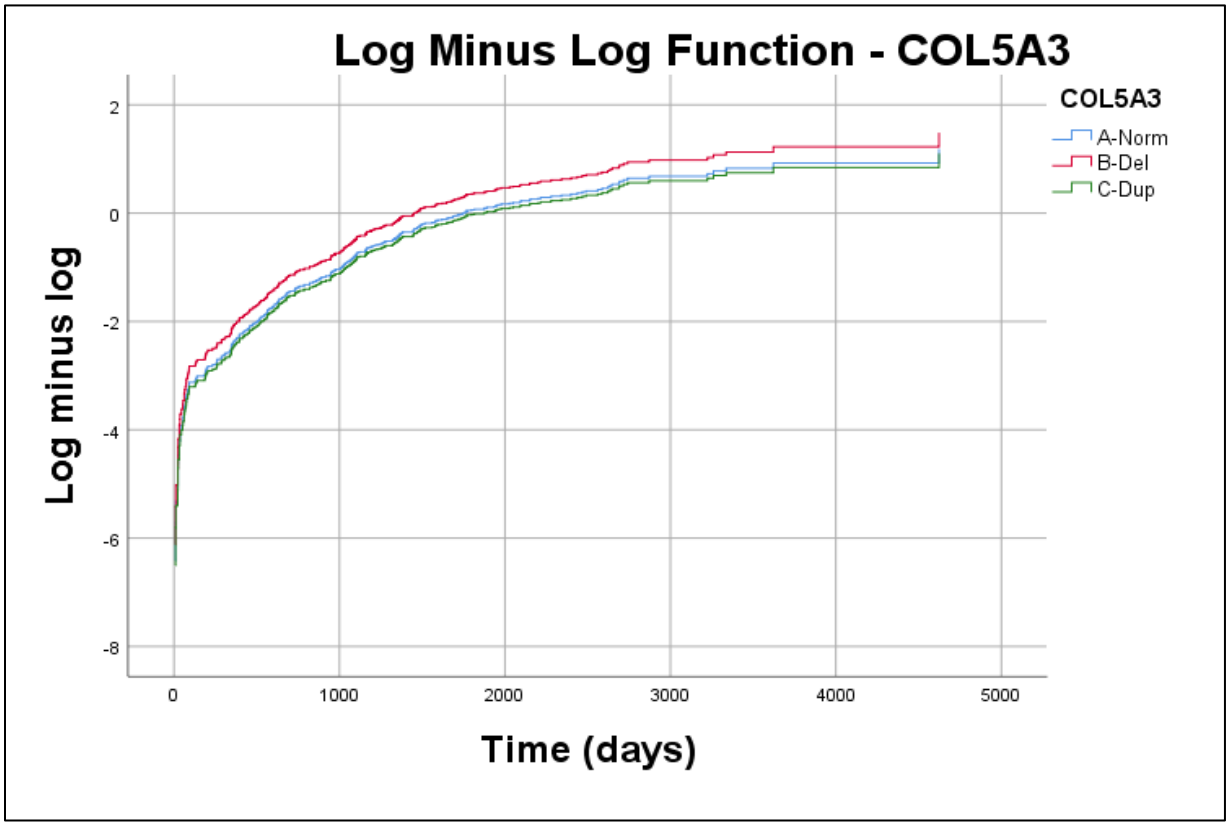
This figure illustrates the survival function over time for the mean of covariates which are the 3 primary genes with statistical findings (COL5A3, COL4A3BP, and COL12A1).



This figure illustrates the hazard function over time for the mean of covariates which are the 3 primary genes with statistical findings (COL5A3, COL4A3BP, and COL12A1).

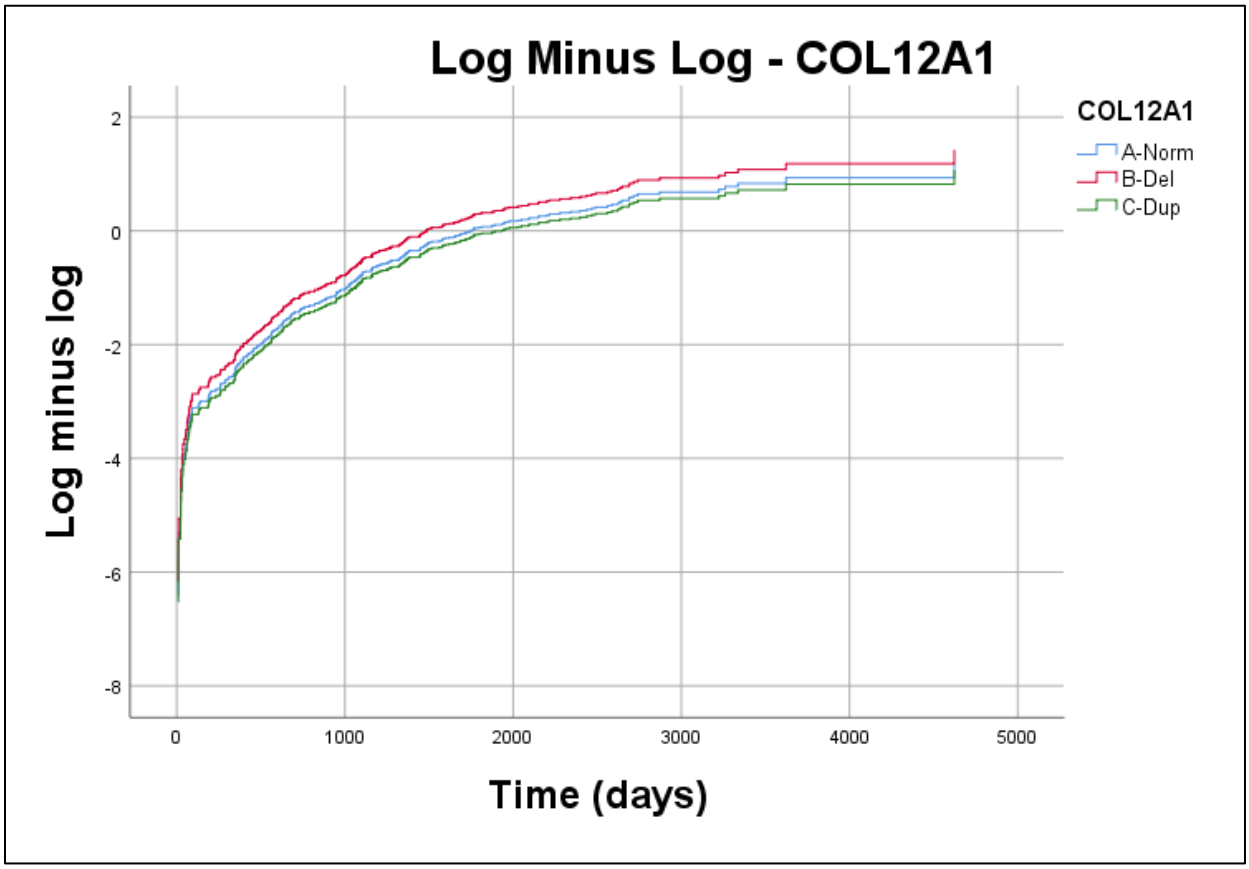
Appendix F

COL5A3



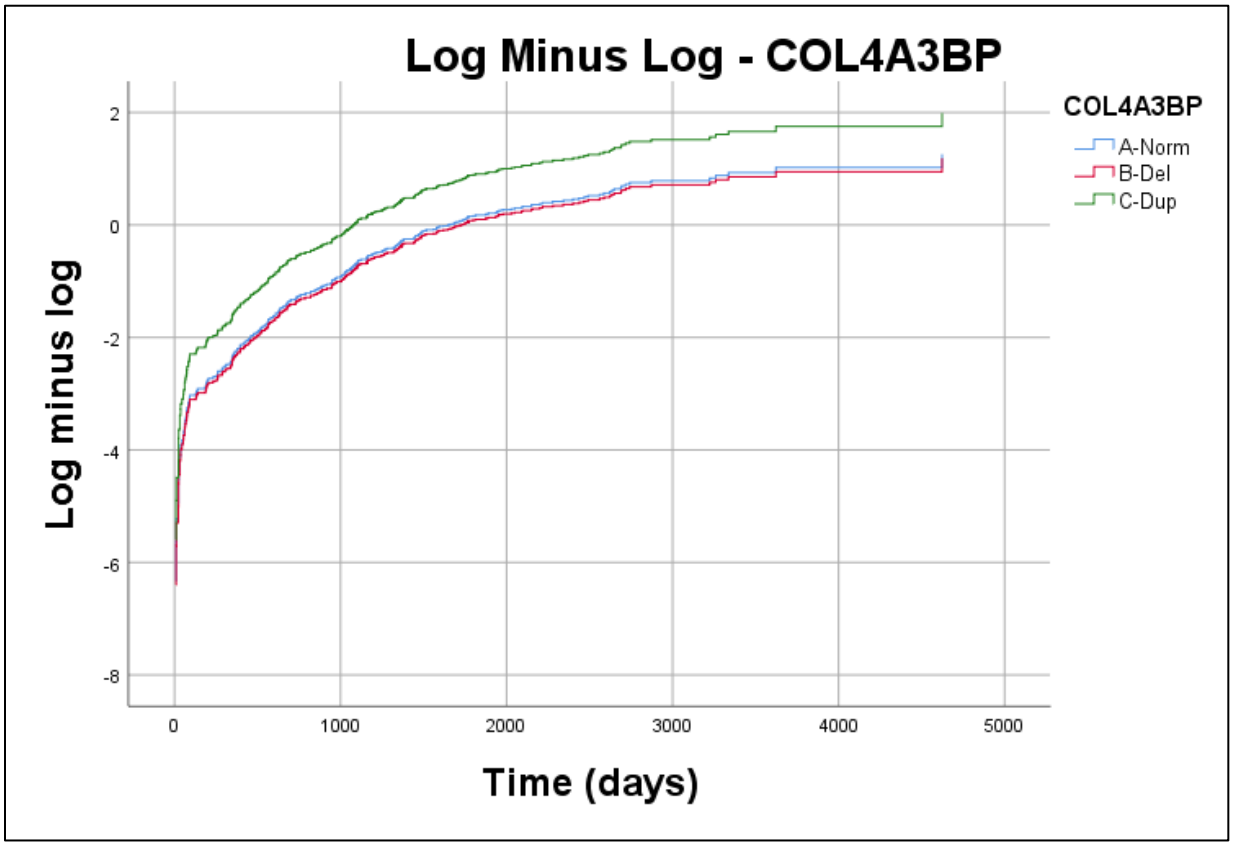
This figure for COL5A3 illustrates the model doesn't violate assumptions of the Cox proportional hazards model as all lines are parallel with each other and don't cross.

COL12A1



This figure for COL12A1 illustrates the model doesn't violate assumptions of the Cox proportional hazards model as all lines are parallel with each other and don't cross.

COL4A3BP



This figure for COL4A3BP illustrates the model doesn't violate assumptions of the Cox proportional hazards model as all lines are parallel with each other and don't cross.

Appendix G - Table of Abbreviations

Analysis of Variance (ANOVA)

Bayesian Survival Analysis (BSA)

Beta(1-0)galactosyltransferase Gene (COLGALT)

Ceramide Transfer (Protein) (CERT)

Clustered regularly interspaced short palindromic repeats (CRISPR)

Collagen-like Gene (COLEC)

Collagen-like subunit of acetylcholinesterase (COLQ)

Collagen Gene 12A1 (COL12A1)

Collagen Gene 4A3BP (COL4A3BP)

Collagen Gene 5A3 (COL5A3)

Comma Separated Value (CSV)

Copy Number Variation (CNV)

Deoxyribonucleic Acid (DNA)

Exploratory Data Analysis (EDA)

Extra-cellular Matrix (ECM)

Genome Reference Consortium Human Build 37 (GRCh37) (HG19)

Genotype Tissue Expression Project (GTEx)

Kaplan-Meier (KM)

Log Minutes Log (LML)

Medical Subject Heading (MeSH)

Next Generation Sequencing (NGS)

Not Applicable (NA)

Ovarian Cancer (OV)

Probability Value (p-value)

Pro-collagen enhancer gene (PCOLCE)

Quantitative Polymerase Chain Reaction (qPCR).

Ribonucleic Acid (RNA)

Risk Ratio (RR)

The Cancer Genome Atlas (TCGA)

The Office of The National Coordinator for Health Information Technology (ONC)

Transforming Growth Factor (TGF)