**UTICA COLLEGE**

**GRADUATE STUDIES**

**MASTER’S IN DATA SCIENCE**

MSDS THESIS

**Pan-Collagen Gene Copy Number Variation Survival Analysis of Ovarian Neoplasms from the TCGA Database**

by

Dr. Robert I. Hodges

PharmD, Medical University of South Carolina, 2010

MBA, The Citadel Graduate College, 2012

Submitted for fulfillment of requirements for degree of

Master of Science in Data Science

August 2021

**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 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 Dr. Brandee Rockefeller for time given in mentoring me and providing guidance on this thesis. Her guidance has made this research much more fluid, better outlined, and more professional.

I would also like to thank Dr. Michael McCarthy for helping guide me from the beginning of this journey in the master’s in data science program and offering the thesis research project as an option. I also thank him for his time and commitment in helping sharpen this research and other data science skills.

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 helps keep me on the progressive path and 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 the master’s in data science program. Her tolerance and understanding have been a key part of my growth in data science and her support has been overwhelming. I love you and thank you.

**Table of Contents** Page Number

Personal Acknowledgements………………………………………………………………. ii

Table of Contents………………………………………………………………………….. iv

Abstract ……………………………………………………………………………………. v

Introduction………………………………………………………………………………… 1

Literature Review…………………………………………………………………………... 3

Methods……………………………………………………………………………………... 4

Data…………………………………………………………………………………………. 10

Results………………………………………………………………………………………. 12

Discussion…………………………………………………………………………………... 15

References…………………………………………………………………………………... 19

Appendix A…………………………………………………………………………………. 24

Appendix B…………………………………………………………………………………. 24

Appendix C…………………………………………………………………………………. 26

Appendix D…………………………………………………………………………………. 29

Abbreviations……………………………………………………………………………….. 31

**Abstract**

The Cancer Genome Atlas (TCGA) is a large longitudinal database of cancer patients with varying cancer types which has led to many studies being published (*The Cancer Genome Atlas Program*, 2019). This study, also based on TCGA, performed survival analysis, log-rank scoring, 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 and analysis was performed in RStudio. Stratification was performed on copy number variation showing stratification alignments of deletion, normal, and duplication. A Kaplan-Meier survival curve was performed along with log-rank tests with a Cox proportional hazard model which found that 3 genes: COL12A1, COL4A3BP, and COL5A3 to have statistically significance 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 neoplasm patients.

**Introduction**

The Cancer Genome Atlas (TCGA) has analyzed over 20,000 cancer observations across 33 different cancer types (Liu et al., 2018). 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% 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 (Xu et al., 2019). Collagen has been studied to some extent as various mechanisms have been discovered with regards to function in ovarian tumors; however, further research is needed (Ricciardelli & Rodgers, 2006).

Collagen is an extracellular matrix (ECM) that forms a barrier around organs and blood vessels (Dipiro et al., 2017, p. 917). Collagen is the most abundant type of ECM located in the ovaries (Cho et al., 2015). Multiple factors contribute to the growth of tumors including ECM remodeling, various growth factors, and other tissue inhibitors. Collagen is an essential component in homeostasis and ECM remodeling, such as ECM stiffness and elasticity, are implicated in ovarian tumorigenesis progression (Cho et al., 2015).

Collagen ECMs are also essential in regulating ovarian cell morphology through various mechanisms such as cell communication and cell shape (Woodruff & Shea, 2007). For example, in tumorigenesis, collagen is remodeled into thick fibrils (Cho et al., 2015). This structure of fibrils is regulated by the COL12A1 gene, which is the collagen type XII alpha 1 chain in homo sapiens (NCBI – COL12A1, 2020). However, research also focuses on drug resistance mechanisms which can be due to functional abnormalities rather than structural like ECM (Holohan et al., 2013). For example, the COL4A3BP gene, also known as the ceramide transfer protein (CERT) gene produces a protein which is a regulator of ceramide transport and has been implicated in multidrug resistance in certain cancer treatments such as colorectal and breast cancer (Lee et al., 2011).

Understanding ECM organization and cell morphology is essential in understanding and gaining insights of malignant neoplasms and their processes (Cho et al., 2015). Copy number variation (CNV) has been linked to complicated traits and behaviors in diseases and drug resistance (Gamazon & Stranger, 2015). CNV is defined as structural variation which alters the number of copies of certain DNA regions (Thapar & Cooper, 2013). These structural variations affect base pairs with a duplication or deletion and have been identified as a facilitator for genomic disease states (Sharp et al., 2005). This is important as it has been estimated that 4.8% to 9.7% of the human genome is comprised of CNVs (Sharp et al., 2005). The research being conducted in this study evaluates the collagen genetic component behind ECM organization and cell morphology in ovarian cancer by exploring heterogenous pan-collagen gene CNV and patient survival relationships in ovarian cancer.

In this study, Kaplan-Meier survival analysis (KM), log-rank tests, and Cox-proportional hazard modeling was performed on ovarian cancer patient CNV mutations in 55 different collagen genes in ovarian cancer from the TCGA database. 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 in ovarian cancer patients regarding CNV. For survival analysis the dependent variable is the time to event, death, or loss of contact while independent variables are gene CNVs.

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

Hopefully, by increasing the genetic knowledge base, a considerable impact on patient outcomes may be achieved. This research has the potential to assist the need for 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. A PubMed literature search was performed using medical subject heading (MeSH) terms for “Ovarian Neoplasms” and “Collagen”. The literature review produced 293 results involving genetic studies involving collagen and ovarian neoplasms.

Gene expression is generally the most commonly 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 has been researched previously but with respects to gene expression instead of CNV (Teng et al., 2013). High-throughput RNA sequencing advances may be the reason why CNV has not been studied as often as gene expression.

Collagen has been shown to be involved in multiple aspects of tumorigenesis as previously mentioned (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 in turn decreasing tumor apoptosis has been studied as well in ovarian neoplasms and hypothesized with 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).

However, some investigation has been done regarding CNV and TCGA. One study investigated gene-drug interactions with regards to CNV in glioblastoma multiforme and lower grade brain glioma (Spainhour & Qiu, 2016).

As 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 similitude. This 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 cancer is caused by inherited germline variations and somatically acquired changes and these two types of neoplasm progenitors may work synergistically (Buckley et al., 2017).

**Methods**

This study uses a non-parametric Kaplan-Meier survival plots, log-rank tests, and semi-parametric Cox proportional hazard modeling. This study does not assume 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 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 multiple reasons including death or being lost to follow-up (Kaplan & Meier, 1958). The Kaplan-Meier survival curve makes a couple of assumptions. The first is that all survival patients are independent of each other, or rather one patient surviving does not affect another patient surviving or having an event. The second assumption is that censoring does not lead to an increased or decreased likelihood of events. The requirements to plot a Kaplan-Meier survival curve are 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). However, 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 descriptive only.

The hazard function and survival functions are both integral parts of survival analysis modeling. The survival function explains through probability of a subject surviving beyond a specific point in time (Kleinbaum & Klein, 2012, p. 54). Another definition for survival function which may be used is where a certain event has not occurred by a specific time point. 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 and will be discussion in more detail at the end of this section. The survival function and hazard function are related 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.

To obtain the survival function equation, one must first look at distribution function of survival time of an individual, also known as the cumulative distribution function of *T*, denoted in Figure 1 (Collet, 2015, p. 10).

*F(t) = P(T < t) = ∫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 the survival time is less than some value of the variable *t,* which is defined as a specific point in time. *∫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 (Collet, 2015, p. 12).

*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, p. 13). Now that 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 at death, or event time, *d*, showing the number of deaths, number of subjects alive, and total number of subjects. This is completed at every death event and the tables are treated as independent events. This procedure is known as the Mantel-Haenszel procedure (Collet, 2015, p. 47). The log-rank test will 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 (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 (Grace-Martin, 2018). The chi square equation is shown in figure 3.

*X2 = Σ(Ojt – Ejt)2 Ejt*

**Figure 3: Chi Square Equation.**

Ojt represents the observed number of events for the *jth* group over time while Ejt 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 Ejt. 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, it the hypothesis is considered to be statistically significant if the chi square value is greater than 3.84 which corresponds to being less than the alpha value of 0.05. Now that 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). This modeling technique allows researchers to investigate relationships between covariates and survival time (Cox, 1972). It 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).

The Cox proportional hazard model also makes other assumptions. 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. It also assumes that survival times are independent which means that a patient surviving is not dependent on another patient surviving or having an event. The Cox proportional hazard model also assumes that the baseline hazard is unspecified and that the treatment variables do not change over time (Cox, 1972). The most important assumption is that the survival curves do not cross each other (Kennedy, 2019). 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 4 (Collet, 2015, p. 13).

*H(t) = − log S(t)*

**Figure 4: 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*.

The hazard ratio is the exponential parameter estimate of proportional hazard models, or two groups with a hazard functions, 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 similar to the relative risk ratio, 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 5 (Bradburn et al., 2003).

*H(t) = H0(t) \* exp( b1X1 + b2 X2+…bi Xi )*

**Figure 5: Hazard Function Expanded**

H(t) is defined as the hazard function, which is determined through variables *X 1, X 2,… Xi.* H0(t) is defined as the baseline hazard which means that this is the hazard if all covariates of *X1, X2,… Xi* are equal to zero. The variable *t* is the specified survival time and *b1, b2,…bi* are denoted as the coefficients, or hazard ratios, which explain the impact of covariates *X1, X2,… Xi.* Thus, the Cox proportional hazards model may be written as a multiple linear regression of the log of hazard with the variables *X1, X2,… Xi*. H0(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 *b1, b2,…bi* are greater than zero, as *X1, X2,… Xi* 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. Thus, hazard ratios greater than one are negatively correlated with survival length (Bradburn et al., 2003).

**Data**

Patient clinical and survival data was obtained from TCGA database under the TCGA-OV project (ovarian cancer). 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. Additional data cleaning and exploratory data analysis (EDA) was performed on the new data frame with the DataExplorer package in RStudio (Version 1.2.5033) and analysis was completed with SurvMiner and Survival packages in RStudio.

55 collagen gene CNV columns were analyzed in EDA. Varying distributions of each gene were found and can be seen in Appendix B. CNV range for each gene spanned from negative two through positive two. Table 1 lists the reference to each value and stratified groupings into smaller groups for ease of analysis.

**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 |

The timeframe column titled “Days.to.Death” displayed a skewed right distribution with a high spike at the beginning of the timeframe. 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.

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.

**Results**

Primary results are displayed in Table 2.

**Table 2. Log-rank Results and Hazard Ratios**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | CNV Stratification | P-value | Coefficient | Exp(Coef) – Hazard Ratio (95% CI) |
| COL12A1 | DELETION | 0.046\* |  |  |
|  | DUPLICATION | 0.0245\* | -0.36 | 0.6977 (0.5098 - 0.9547) |
|  | NORMAL | 0.0629 | -0.2451 | 0.7826 (0.6044 - 1.0133) |
| COL4A3BP | DELETION | 0.021\* |  |  |
|  | DUPLICATION | 0.00696\*\* | 0.80615 | 2.239 (1.2470 – 4.021) |
|  | NORMAL | 0.58518 | 0.07492 | 1.078 (0.8236 – 1.410) |
| COL5A3 | DELETION | 0.014\* |  |  |
|  | DUPLICATION | 0.00548\*\* | -0.3818 | 0.6827 (0.5214 – 0.8937) |
|  | NORMAL | 0.04536\* | -0.2992 | 0.7414 (0.5531 – 0.9938) |

**\*Denotes statistical significance with p-value < 0.05.  
\*\*Denotes statistical significance with p-value < 0.01.**

All three gene survival models were statistically significant, however, not all CNV stratifications were significant. Only gene COL5A3 was within significance between deleted, normal, and duplicated CNV stratification.

All likelihood ratio tests and Wald tests were statistically significant for all three collagen gene models.

In the Cox-proportional hazard model, the COL5A3 coefficients were negative for normal and duplication CNV when compared to deletion CNV. This means that normal and duplicated CNV stratifications in this gene have increased survival probability as it is negatively correlated to the event. Coefficients for COL12A1 were also negative showing the same survival probability increase. 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 COL12A1 duplication at 0.6977 which approximately translates to a 0.7:1 ratio for chance of death. This is approximately a 30% decreased probability of death with this gene duplication. In gene COL4A3BP, for duplicated CNV, there is a hazard ratio of 2.239. This translates to a 2.24:1 ratio for increased chance of death which is a 224% probability increase in death for with this gene duplication. Hazard ratios for COL5A3 were both statistically significant for duplication and normal CNV at 0.6827 and 0.7414, respectively. These translate to a ratio of 0.68:1 ratio and 0.74:1 ratio for chances of death; or a 32% and 26% decreased chance of death, respectively, for duplication and normal CNV.

Shown in Appendix C are KM plots for all three genes. Gene COL5A3 has a very distinct and visible survival difference between 1,100 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. Duplication of CNV seems to show greater probability of survival through most of the KM plot for COL5A3.

Gene COL12A1 illustrates deletion of CNV with a lower survival probability starting around 1,000 days, however, normal CNV was not statistically significant in this model. When comparing deletion against duplication CNV in the KM plots, there is a clear distinction between the two since normal is not statistically significant.

COL4A3BP is only significant with regards to duplication when compared to deletion CNV. Normal CNV was not statistically significant and can be seen in the KM plots to closely mimic the deletion CNV curve.

Table 3 displays log-rank scores of each gene with asterisks marking statistical significance.

**Table 3. 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 |

**\*Denotes statistical significance with p-value < 0.05**. **-Highlighted in red are observed events which are greater than expected events, showing a decreased probability of survival in the models.**

Log-rank scores show the numerical differences in events that are expected versus observed.

Hazard models plotted over time are displayed in Appendix D. The beta line is not within the 95% confidence interval one hundred percent of the time for any of the three models. However, these models still have a satisfactory fit for the data.

Table 4 lists median days of survival with upper and lower limits of confidence intervals.

**Table 4. Median Days of Survival with Confidence Interval Limits**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | N | Events | Median | 0.95LCL | 0.95UCL |
| COL12A1=Deletion\* | 196 | 118 | 1,259 | 1,091 | 1,380 |
| COL12A1=Duplication\* | 134 | 60 | 1,451 | 1,204 | 2,148 |
| COL12A1=Normal | 234 | 113 | 1,249 | 1,249 | 1,686 |
| COL4A3BP=Deletion\* | 410 | 207 | 1,336 | 1,204 | 1,492 |
| COL4A3BP=Duplication\* | 21 | 12 | 457 | 256 | NA |
| COL4A3BP=Normal | 133 | 72 | 1,384 | 1,213 | 1,686 |
| COL5A3=Deletion\* | 193 | 110 | 1,102 | 1,046 | 1,264 |
| COL5A3=Duplication\* | 217 | 105 | 1,446 | 1,278 | 1,595 |
| COL5A3=Normal\* | 154 | 76 | 1,516 | 1,364 | 1,757 |

**\*Denotes statistical significance with p-value < 0.05.**

COL12A1 has overlap in the limits of confidence intervals. However, comparison of normal to duplication CNV in this gene displays no overlap of the confidence interval.

COL4A3BP shows a “NA” value meaning the confidence interval could not be calculated for this as the data for this stratification is sparse. More data is needed for this and should be investigated more.

COL5A3 shows clear non-overlap in the limits of confidence intervals when comparing deletion to normal CNV and deletion to duplication CNV.

**Discussion**

These are novel results and have not been previously published. 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. There are an estimated 30,000 genes in the human body (Human Genome Project FAQ, 2013). There are 1,500 genes that could potentially have a finding that is random with an alpha set at 0.05 (Genetics Home Reference, 2019). However, very few genetic studies seem to use any controls for false discovery, such as a Bonferroni p-value correction, and are generally due to lack of quality and quantity of data (Dahiru, 2011). This bias should be kept in mind as the primary limitation to this study is quality and quantity of data. 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). 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 controlling for false discovery (Ulrich, 2016). 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.

Another 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. One would not randomly pick a set of genes to investigate where the gene is not expressed in the tissue being studied. Thus, this strengthens the argument for using alpha at 0.05 and not using a Bonferroni p-value correction.

The methods for CNV detection with high-throughput sequencing also contribute to false discovery rates (Jiang et al., 2018). An increase in observations would be appropriate to help control for this. It is also possible that a combination of TCGA repository and future cancer databases could be utilized and combined to increase observation numbers, such as the Genotype Tissue Expression Project (GTEx) which has been combined with the TCGA recently (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 collagen effects on neoplasms.

Another bias worth mentioning is that data collection was done by multiple people at different locations, as this is secondary data. 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).

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). It is most likely that the evolutionary artifacts have not introduced much bias into this study. However, one should be aware that they exist and that they are not able to be controlled for in this research, as quantity and quality of data is are limiting factors. Evolutionary artifacts would most likely not make a difference in treatment options that may arise from research such as this.

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 and small dose alterations may lead to toxicity, thus, chemotherapy dosing changes based on genetics may be unlikely from CNV studies such as this (Eaton & Lyman, 2019). 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.

**References**

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

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

Buckley, A. R., Standish, K. A., Bhutani, K., Ideker, T., Lasken, R. S., 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>

Collett, D. (2015). *Modelling survival data in medical research* (3rd ed.). Crc Press Taylor & Francis Group.‌

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#  
‌

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>

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/

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

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

‌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>

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>

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

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>  
‌

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>

*NCBI - COL12A1*. (2020, October 25). Www.Ncbi.Nlm.Nih.Gov; NCBI. <https://www.ncbi.nlm.nih.gov/gene/1303>

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

‌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

‌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

‌Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re identifications in incomplete datasets using generative models. *Nature Communications*, *10*(1). <https://doi.org/10.1038/s41467-019-10933-3>

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>

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

‌  
‌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](https://doi.org/10.1186/s12859-016%091255-7)

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

‌TCGA - Ovarian Serous Adenocarcinoma Study. (2018, September 5). [Www.Cancer.Gov](http://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

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

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

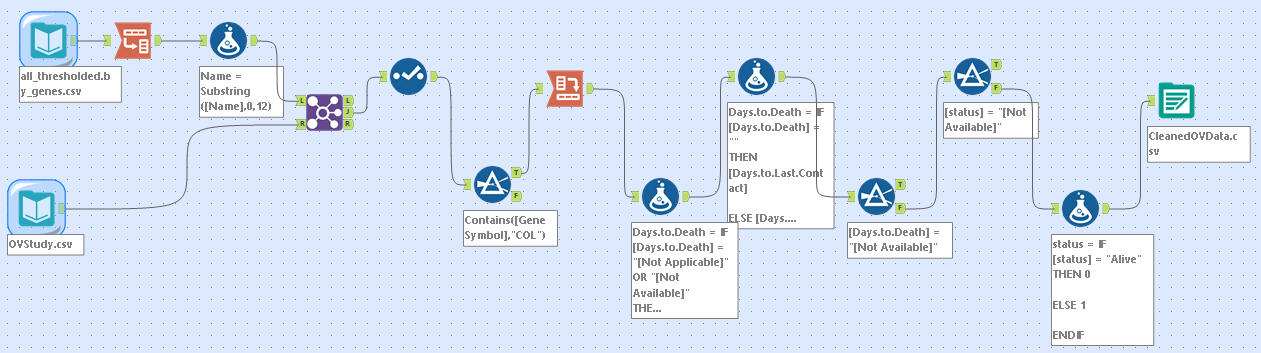
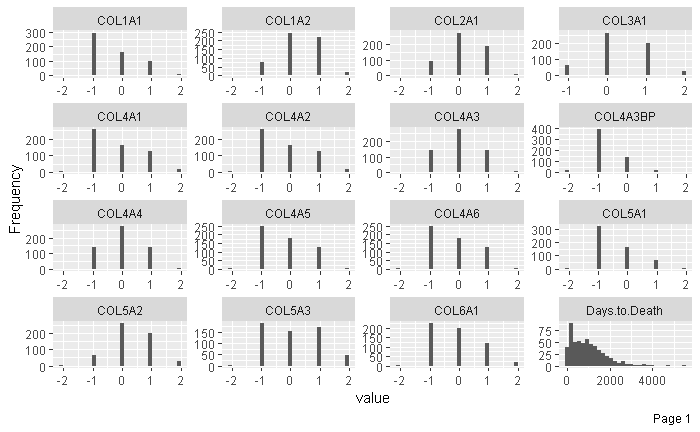
Webster, T. H., Couse, M., Grande, B. M., Karlins, E., Phung, T. N., Richmond, P. A., Whitford, W., & Wilson, M. A. (2019). Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience*, *8*(7). <https://doi.org/10.1093/gigascience/giz074>

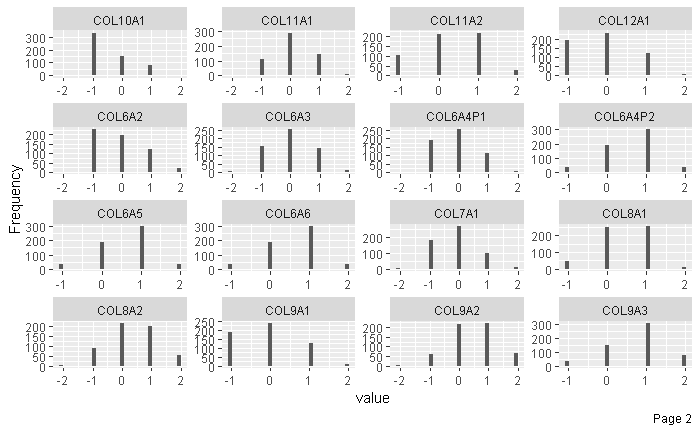
Woodruff, T. K., & Shea, L. D. (2007). The Role of the Extracellular Matrix in Ovarian Follicle Development. *Reproductive Sciences*, *14*(8\_suppl), 6–10. https://doi.org/10.1177/1933719107309818

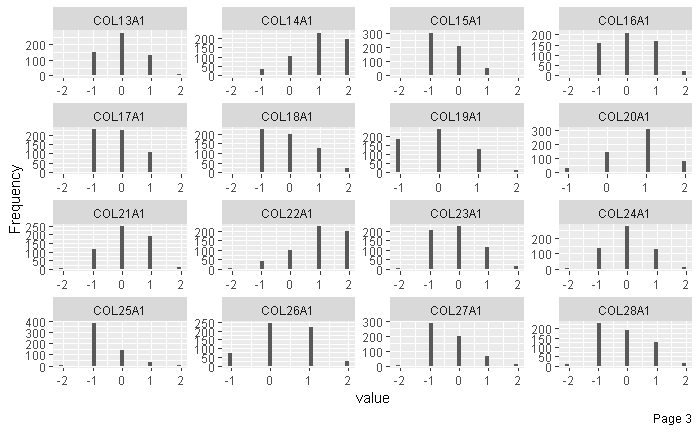
‌‌Wu, Y.-H., Chang, T.-H., Huang, Y.-F., Huang, H.-D., & Chou, C.-Y. (2013). COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene*, *33*(26), 3432–3440. <https://doi.org/10.1038/onc.2013.307>

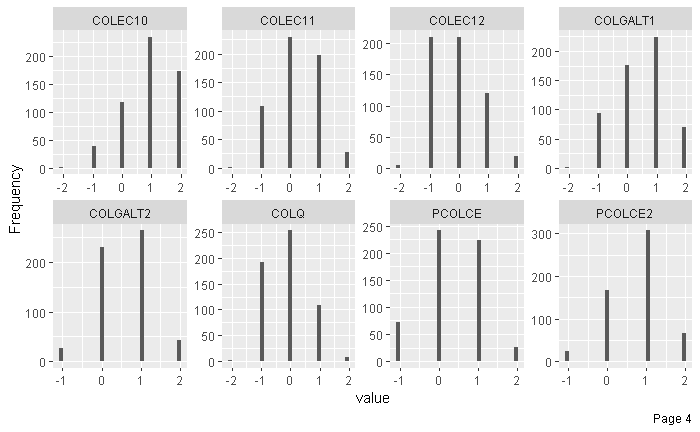
Xu, S., Xu, H., Wang, W., Li, S., Li, H., Li, T., Zhang, W., Yu, X., & Liu, L. (2019). The role of collagen in cancer: from bench to bedside. *Journal of Translational Medicine*, *17*(1). https://doi.org/10.1186/s12967-019-2058-1

**Appendix A**

**  
  
Appendix B**







**Appendix C**

COL12A1

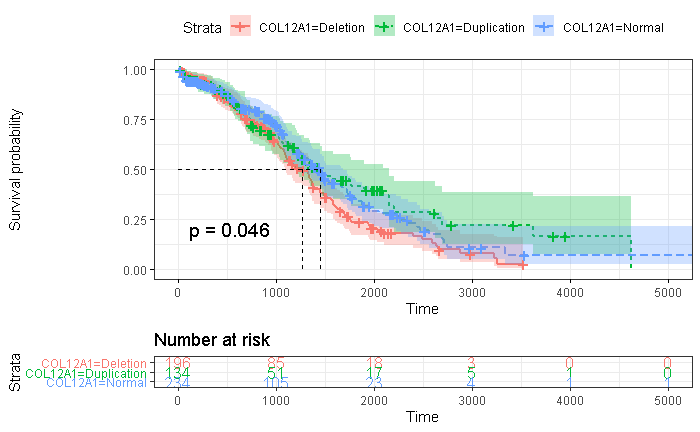


Figure COL12A1 survival has stratified curve of CNV changes (n = 564). Time is measured in days. Green illustrates duplicated copy number variation showing increased survival probability while red illustrates deletion of copy number variation displaying lower survival probability. Blue displays normal 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

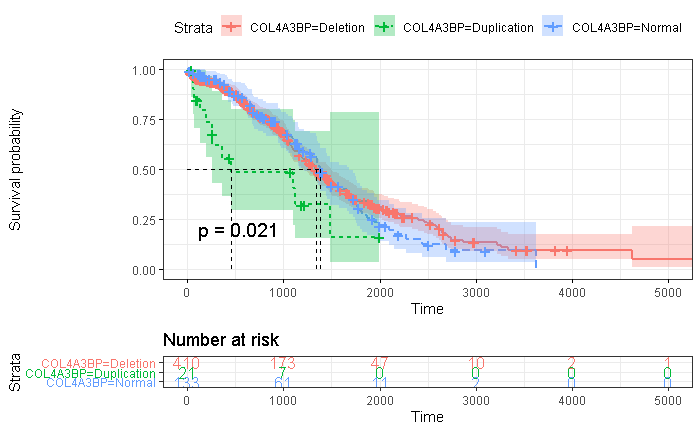


Figure COL4A3BP survival has stratified curve of CNV changes (n = 564). Time is measured in days. Green illustrates duplicated copy number variation displaying a decreased survival probability while red illustrates deletion of copy number variation. Blue displays normal copy number variation illustrating 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

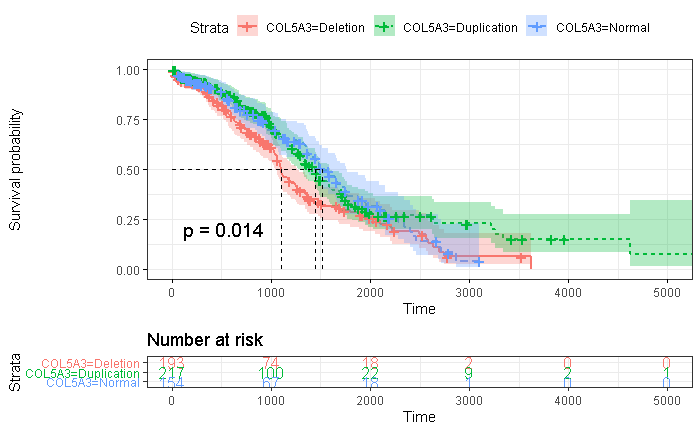
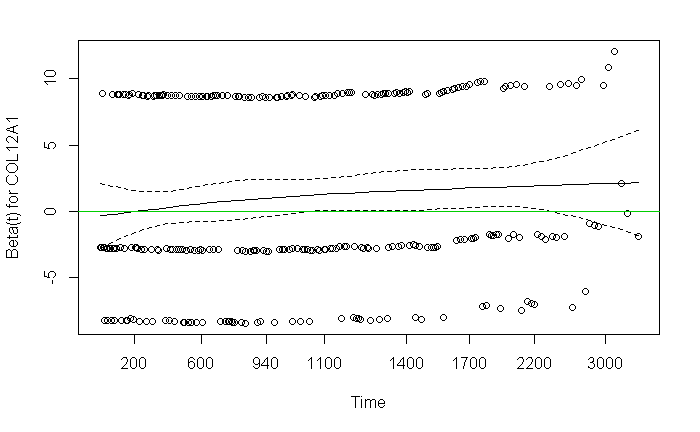
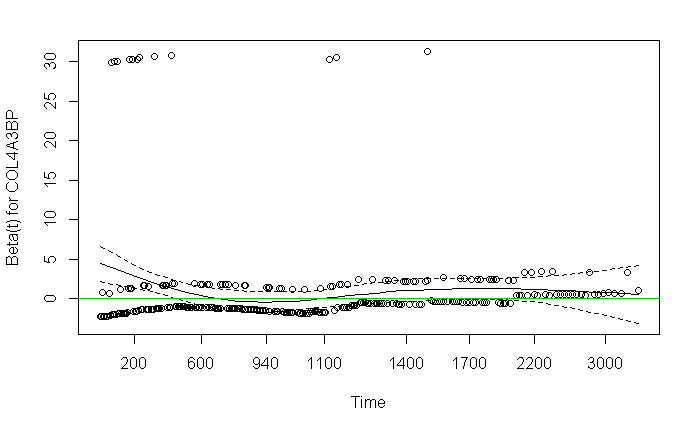


Figure COL5A3 survival has stratified curve of CNV changes (n = 564). Time is measured in days. Green illustrates duplicated copy number variation displaying an increased survival probability while red illustrates deletion of copy number variation illustrating a lower survival probability when compared 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.

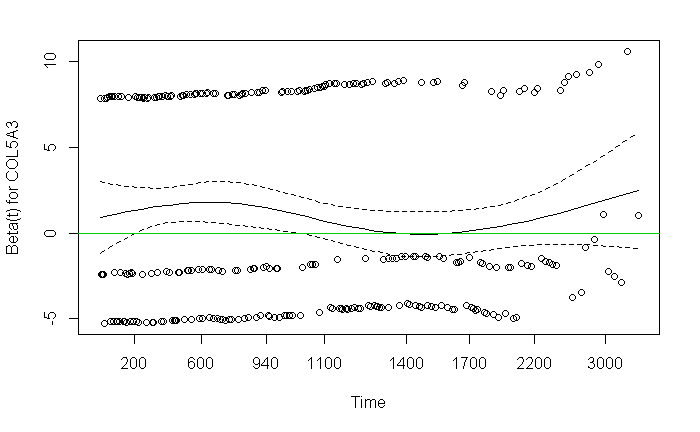
**Appendix D**

COL12A1

COL4A3BP



COL5A3



**Table of Abbreviations**

Analysis of Variance (ANOVA)

Ceramide Transfer (Protein) (CERT)

Copy Number Variation (CNV)

Collagen Gene 12A1 (COL12A1)

Collagen Gene 4A3BP (COL4A3BP)

Collagen Gene 5A3 (COL5A3)

Comma Separated Value (CSV)

Deoxyribonucleic Acid (DNA)

Exploratory Data Analysis (EDA)

Extra-cellular Matrix (ECM)

Genotype Tissue Expression Project (GTEx)

Kaplan-Meier (KM)

Medical Subject Heading (MeSH)

Not Applicable (NA)

Ovarian Cancer (OV)

Probability Value (p-value)

Quantitative Polymerase Chain Reaction (qPCR).

Ribonucleic Acid (RNA)

The Cancer Genome Atlas (TCGA)

Transforming Growth Factor (TGF)