TODO

Label all figures and have explanation of each one.

Redo figures using basic R survival package and not Survminer as that is expectation from publishers. Researchers aren’t as familiar with Survminer or trust it’s results.

**Abstract**

The Cancer Genome Atlas (TCGA) is a large longitudinal database of cancer patients with various cancer types (*The Cancer Genome Atlas Program*, 2019). This study performed survival analysis, log-rank scoring, and Cox proportional hazard models comparing copy number variation in ovarian cancer patients from the TCGA repository. Data mining? (maybe be better word) was performed initially in Alteryx Designer and secondary data cleansing and analysis was performed in RStudio. Stratification was performed on copy number variation showing alignments of deletion, normal, and duplication. Results found 3 genes; COL12A1, COL4A3BP, and COL5A3, within statistical significance regarding survival and copy number variation abnormalities. This study provides further evidence that collagen genetics have multiple effects on tumors in ovarian cancer regarding survival.

**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 cancer types with multiple publications emanating from this database alone, including novel discoveries regarding ovarian cancer.

Ovarian cancer has a relative survival risk of 80% and is considered the most lethal gynecological cancer type (Cancer.org, 2014). One factor gaining mention in the scientific literature demonstrated to impact 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, more research is needed (Ricciardelli & Rodgers, 2006).

In this study, Kaplan-Meier survival analysis (KM), log-rank tests, and Cox-proportional hazard modeling was performed on patient copy number variation (CNV) mutations in 55 different collagen genes in ovarian cancer from the TCGA database. There are 55 null hypotheses which state there are no statistical differences in survival of ovarian cancer patients with any of the 55 collagen genes based on CNV stratification. The alternative 55 hypotheses states that 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 copy numbers.

Survival analysis is essential in gene exploration as we can potentially refine medical treatment for a more individualized approach with the objective to decrease morbidity and mortality.

**Literature Review**

TCGA has produced multiple studies that contribute to human understanding of genetics (The Cancer Genome Atlas – Publications, 2019). However, TCGA data needs to be handled with respect and caution. There are methods available that can re-identify patients based on medical data (SOURCE HERE).

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).

Ovarian cancer has been studied with regards to collagen in the past. A PubMed literature search was performed using MeSH terms for “Ovarian Neoplasms” and “Collagen”. Literature review was performed for 293 results regarding genetic studies involving collagen, and ovarian cancer.

There is an abundance of data investigating the relationship between collagen and ovarian cancer; however, little was noted regarding CNV. CNV has previously been researched in other types of cancer. Gene expression is generally the most commonly researched topic with regards to neoplasms and genetics as normalization techniques in high-throughput RNA sequencing are more available and widely used (Dillies et al., 2012).

One study linked collagen gene expression to metastasis promotion through the TGF-β1 signaling pathway (Cheon et al., 2013). Another linked enhanced COL2A1 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 cancer (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).

Gene-drug interactions in ovarian cancer has been researched previously but with respects to gene expression instead of CNV (Teng et al., 2013).

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

**Methods**

Patient clinical and survival data was pulled from TCGA database under the TCGA-OV project which is ovarian cancer. CNV data was pulled 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 can be referenced in Appendix A. Data cleaning included transposition, removing null value rows, and removing independent variable columns not being studied. Data cleansing 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 in with SurvMiner and Survival packages in RStudio. Charts were made in Microsoft Excel.

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 (SOURCE HERE). Censoring is either due to loss of follow-up data or non-occurrence of event in this study.

Kaplan-Meier survival curves (KM) 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. Primary Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene | CNV Stratification | P-value | Concordance | Coefficient | Exp(Coef) – Hazard Ratio (95% CI) |
| COL12A1 | | 0.046\* | 0.52 |  |  |
|  | COL12A1DUPLICATION | 0.0245\* |  | -0.36 | 0.6977 (0.5098 - 0.9547) |
|  | COL12ANORMAL | 0.0629 |  | -0.2451 | 0.7826 (0.6044 - 1.0133) |
| COL4A3BP | | 0.021\* | 0.516 |  |  |
|  | COL4A3BPDUPLICATION | 0.00696\*\* |  | 0.80615 | 2.239 (1.2470 – 4.021) |
|  | COL4A3BPNORMAL | 0.58518 |  | 0.07492 | 1.078 (0.8236 – 1.410) |
| COL5A3 | | 0.014\* | 0.549 |  |  |
|  | COL5A3DUPLICATION | 0.00548\*\* |  | -0.3818 | 0.6827 (0.5214 – 0.8937) |
|  | COL5A3NORMAL | 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.

In the Cox-proportional hazard model, the coefficients were negative for COL5A3NORMAL and COL5A3DUPLICATION when compared to the deletion. This means that normal and duplicated CNV levels 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 was significant.

MAKE TABLE FOR HAZARD RATIOS

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 in this stratification. In gene COL4A3BP for duplicated CNV is a hazard ratio of 2.239 which translates to a 2.24:1 ratio for increased chance of death which is a 224% probability increase in death for this gene stratification. 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.

Concordance also was higher with COL5A3 compared to the other two genes, showing that we could explain more of the data with the KM curve for this model. Shown in Appendix C are Kaplan-Meier curves (KM) for all 3 genes. Gene COL5A3 has a very distinct and visible survival difference between 1100 days and 1700 days roughly, but then comes back to where all 3 stratifications do not show much difference around day 2000, but then splits off again. Duplication of CNV seems to show greater probability of survival through most of the KM curve for COL5A3.

Gene COL12A1 also shows deletion of CNV with a lower survival probability starting around 1300 days, however, COL12ANORMAL was not statistically significant in this model. When comparing deletion against duplication of CNV in the KM curves, 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. Normal CNV was not statistically significant and can be seen in the KM curves to closely mimic the deletion CNV curve.

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

**Table 3. Log-rank 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 |

Log-rank scores show the numerical differences in events that are expected vs observed. Highlighted in red are the observed events which are higher than the expected events, showing a decreased probability of survival in the models.

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 good fit to the data.

Table 4 lists median days of survival with upper and lower limits of confidence intervals with asterisk marking statistical significance.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | N | Events | Median | 0.95LCL | 0.95UCL |
| COL12A1=Deletion\* | 196 | 118 | 1259 | 1091 | 1380 |
| COL12A1=Duplication\* | 134 | 60 | 1451 | 1204 | 2148 |
| COL12A1=Normal | 234 | 113 | 1249 | 1249 | 1686 |
| COL4A3BP=Deletion\* | 410 | 207 | 1336 | 1204 | 1492 |
| COL4A3BP=Duplication\* | 21 | 12 | 457 | 256 | NA |
| COL4A3BP=Normal | 133 | 72 | 1384 | 1213 | 1686 |
| COL5A3=Deletion\* | 193 | 110 | 1102 | 1046 | 1264 |
| COL5A3=Duplication\* | 217 | 105 | 1446 | 1278 | 1595 |
| COL5A3=Normal\* | 154 | 76 | 1516 | 1364 | 1757 |

COL12A1 does have overlap in the limits of confidence intervals.

COL4A3BP shows an 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 and normal CNV.

**Discussion**

These results are novel have not been previously reported. However, these findings should be further investigated based on possibility of false discovery, previously mentioned as a risk with this type of genetic research (Efron, 2005). The reasoning for this is having an alpha level set at 0.05. There are an estimated 25,000 genes in the human body. There are 1,300 genes that could potentially have a finding that is random or not truly exist with an alpha set at 0.05 (Genetics Home Reference, 2019). However, very few genetic studies seem to use any controls for false discovery and may be due to lack of quality and quantity of data (Dahiru, 2011).

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 (SOURCE HERE). Gene sequencing is becoming cheaper and more accurate (REFERENCE HERE). Thus, a strong argument can be made for keeping alpha at 0.05 with confidence intervals and not controlling for false discovery. While objectivity is a strong goal to maintain in any study and can be difficult to achieve, finding appropriate data in genetics is just as difficult.

Another argument for not controlling for false discovery are the underlying mechanisms showing collagen has 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.

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.

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). The methods for CNV detection with high-throughput sequencing contributes to false discovery rates (SOURCE HERE). An increase in observations would be appropriate as well. It is also possible that a combination of TCGA repository and future cancer databases could be utilized and combined to increase observation numbers, but caution should be used as previously stated. It is possible that this future research structure may not be feasible with technology advancing in genomics at a fast rate. However, TCGA is such a huge longitudinal archive of genetic information, more CNV survival analysis studies should be performed on other cancers to help create a baseline knowledge of CNV and collagen effects on neoplasms.

Studies such as this can be considered foundational to persue further by scientist with the hopes of providing personalized oncological treatments. One could certainly envision a genetic panel being performed on a patient before treatment is started to help calculate a dose of chemotherapy, should be that be most appropriate option. However, most antineoplastic molecules are of a narrow therapeutic index and small dose alterations may lead to toxicity (Eaton & Lyman, 2019). A two to three gene CNV signature test could possibly be implemented from studies such as this one to present information about probability for survival to the patient and physicians. This potential CNV signature could also be assimilated into a gene expression panel for a more complete study in predictive power of genetics and proteomics in patient survival.

**Work Cited/ References**

*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; 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>

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>  
‌

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>

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

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>

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

‌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>

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>

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>

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>

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>

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

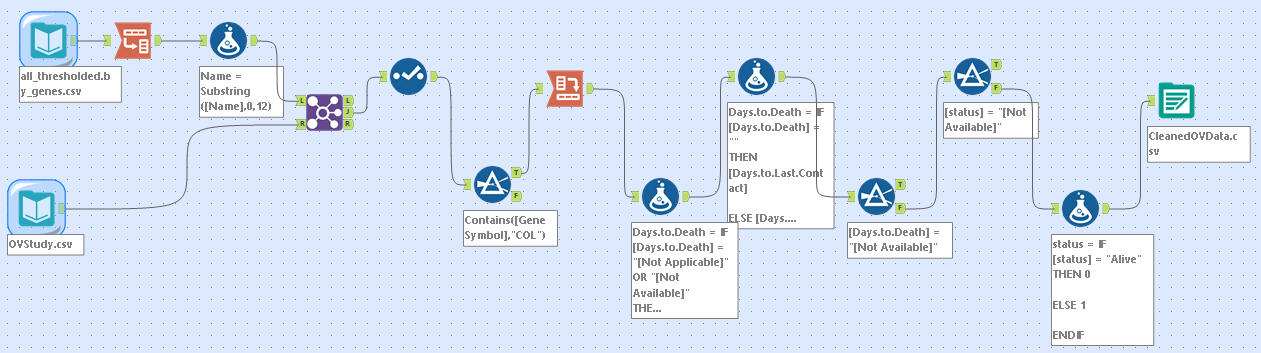
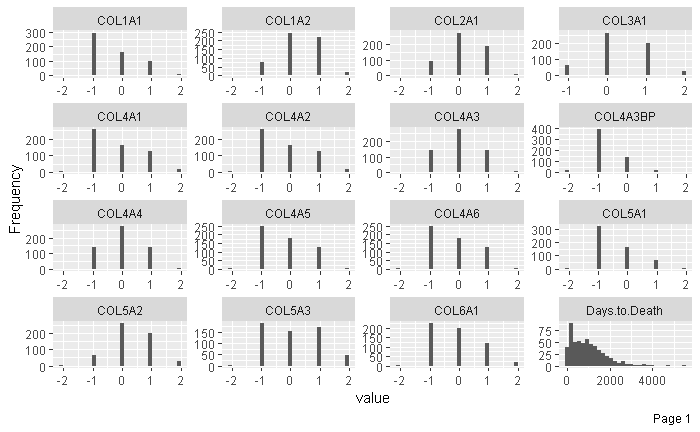
*The Cancer Genome Atlas - Publications*. (2019, July 24). National Cancer Institute; Cancer.gov. <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/publications>  
‌

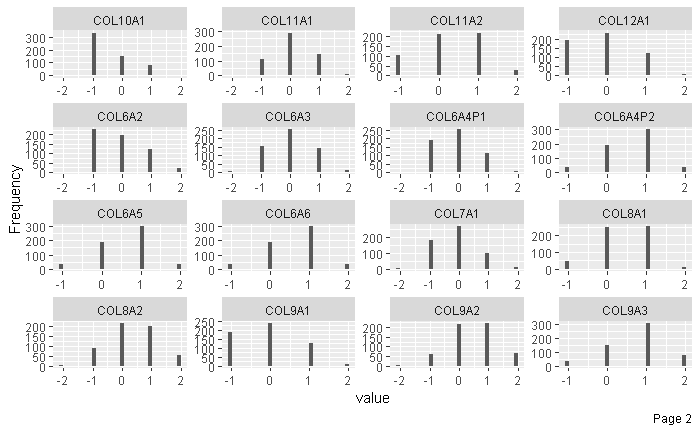
Webster, T. H., Couse, M., Grande, B. M., Karlins, E., Phung, T. N., Richmond, P. A., Whitford, W., & Wilson, M. A. (2019). Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience*, *8*(7). https://doi.org/10.1093/gigascience/giz074

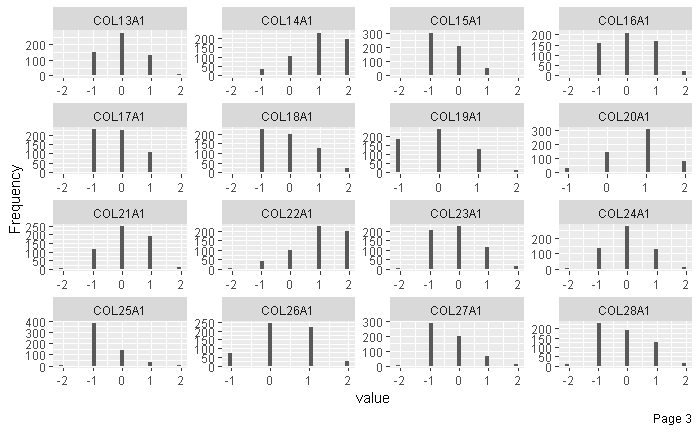
‌‌Wu, Y.-H., Chang, T.-H., Huang, Y.-F., Huang, H.-D., & Chou, C.-Y. (2013). COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene*, *33*(26), 3432–3440. <https://doi.org/10.1038/onc.2013.307>

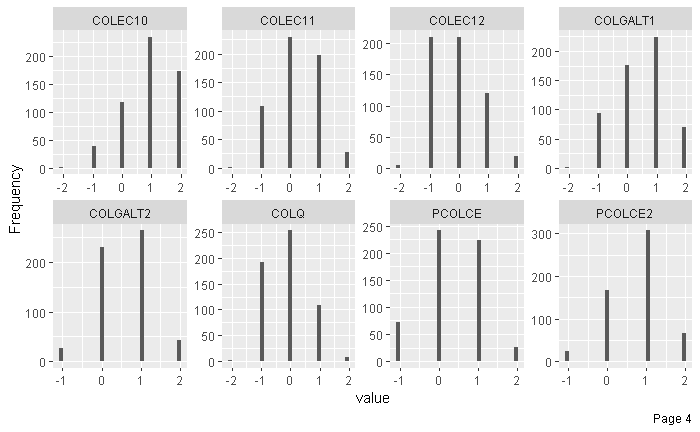
Xu, S., Xu, H., Wang, W., Li, S., Li, H., Li, T., Zhang, W., Yu, X., & Liu, L. (2019). The role of collagen in cancer: from bench to bedside. *Journal of Translational Medicine*, *17*(1). https://doi.org/10.1186/s12967-019-2058-1

**Appendix A**

**  
  
Appendix B**

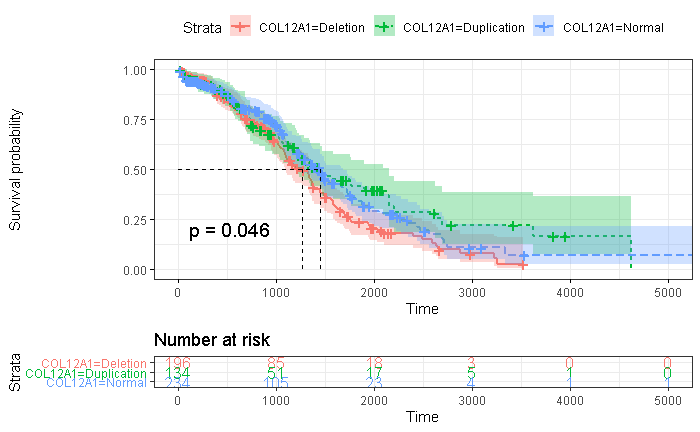


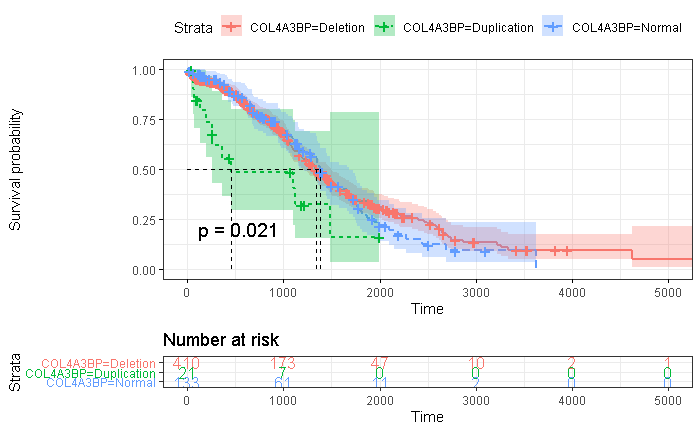




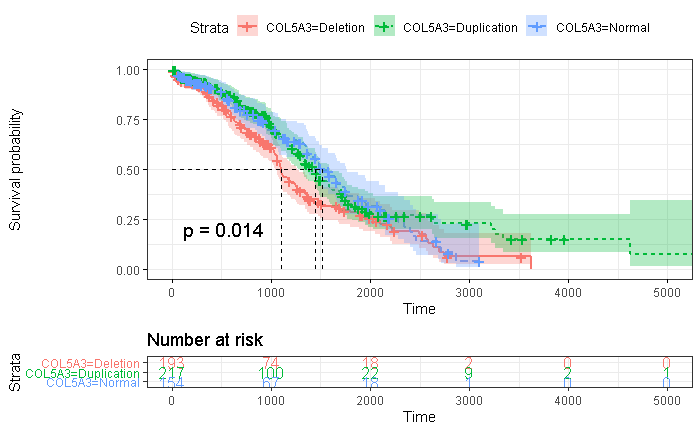
**Appendix C**

COL12A1

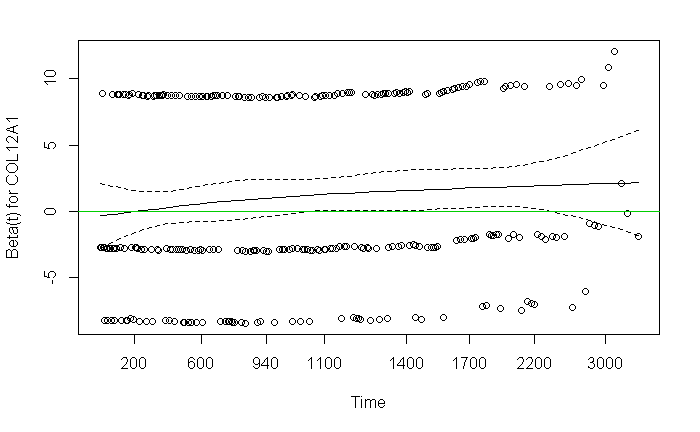
COL4A3BP



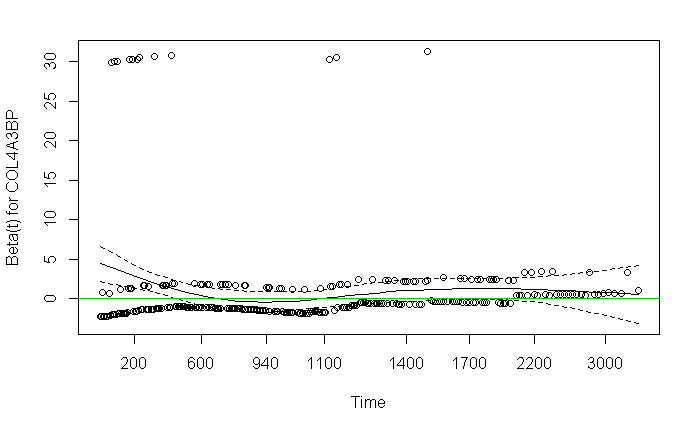
COL5A3



**Appendix D**

COL12A1

COL4A3BP



COL5A3

