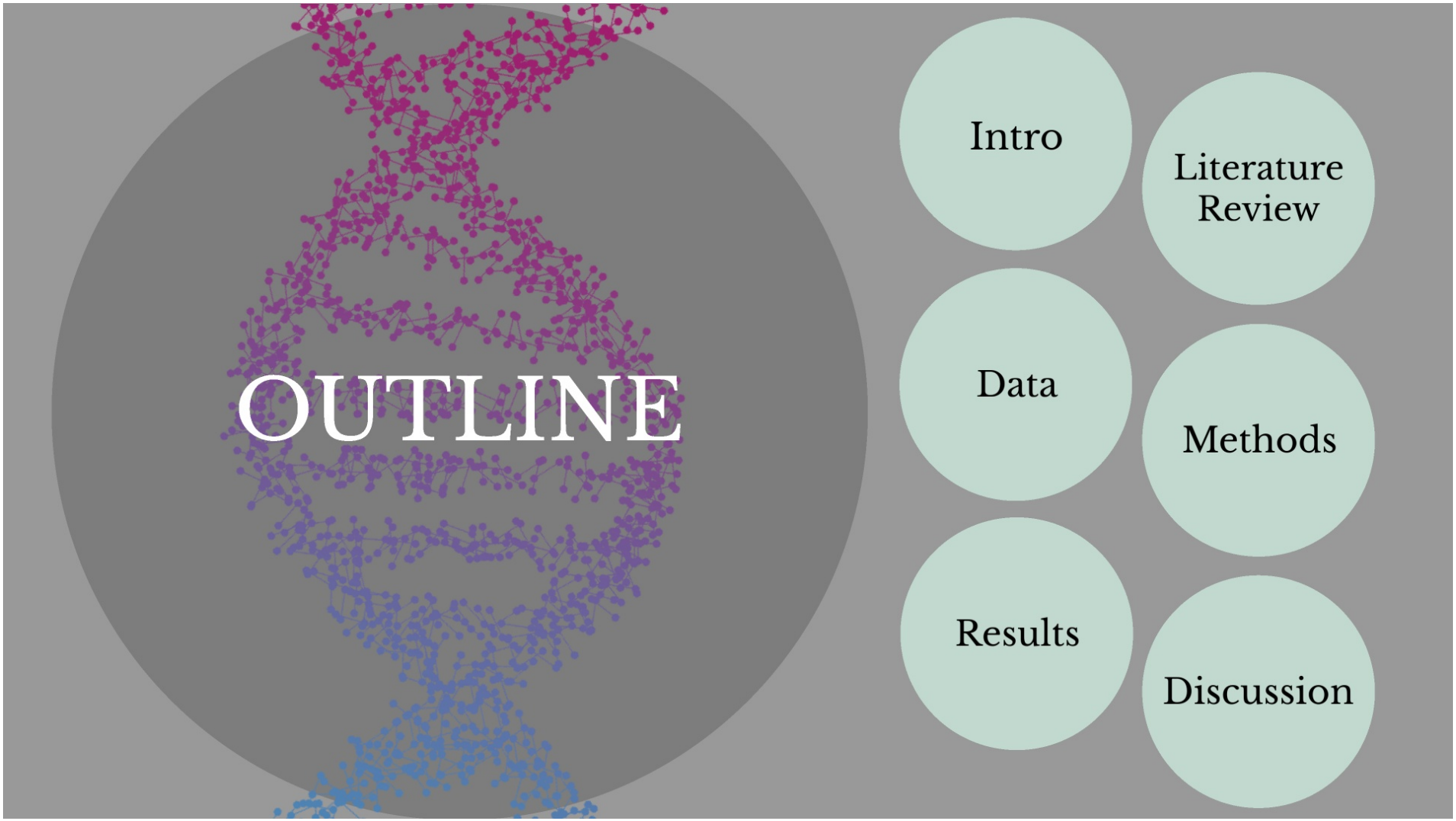
Pan-Collagen Copy Number Variation Survival Analysis in Ovarian Cancer

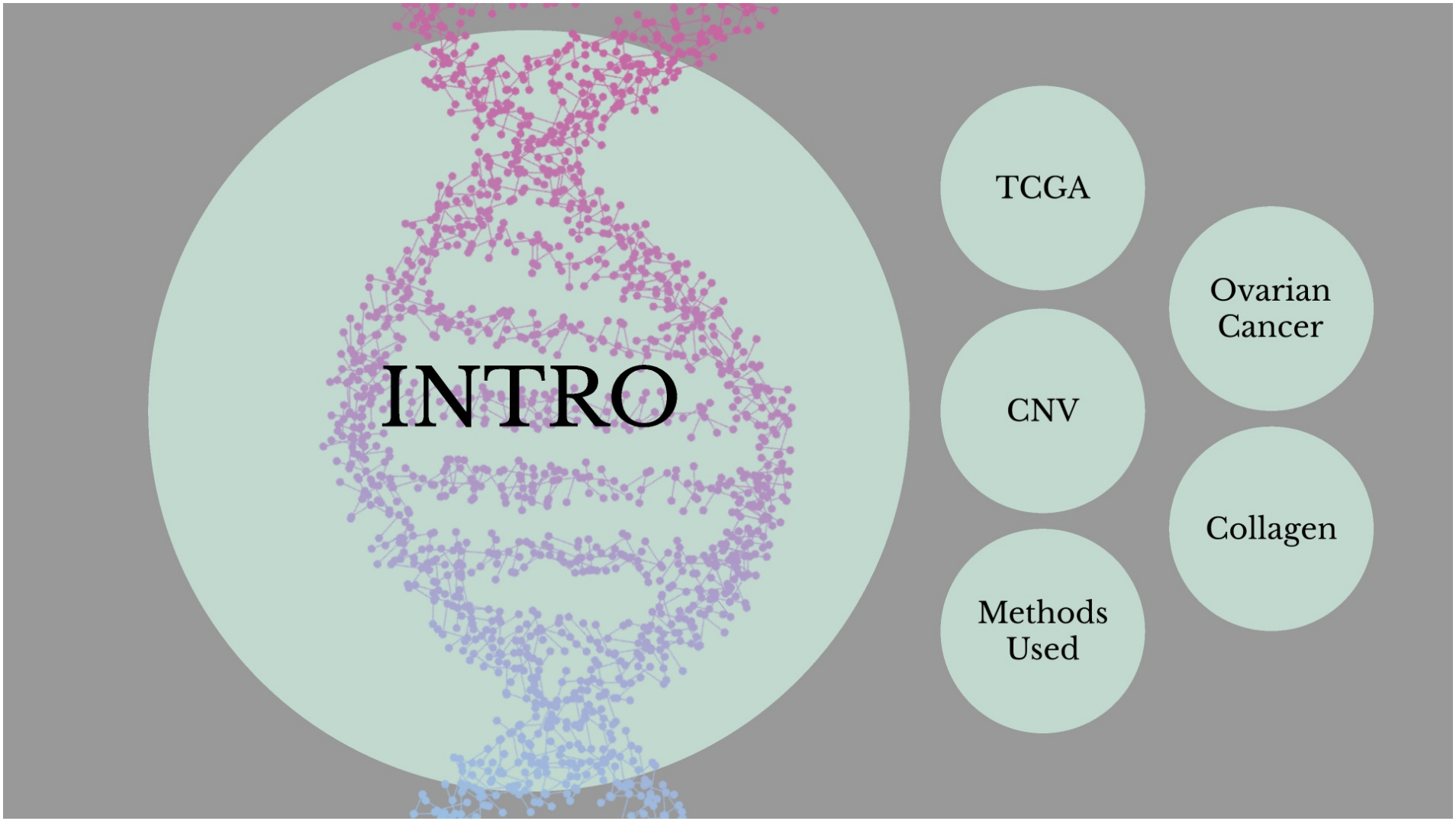
Robert Hodges, PharmD, MBA, RPh

Thesis

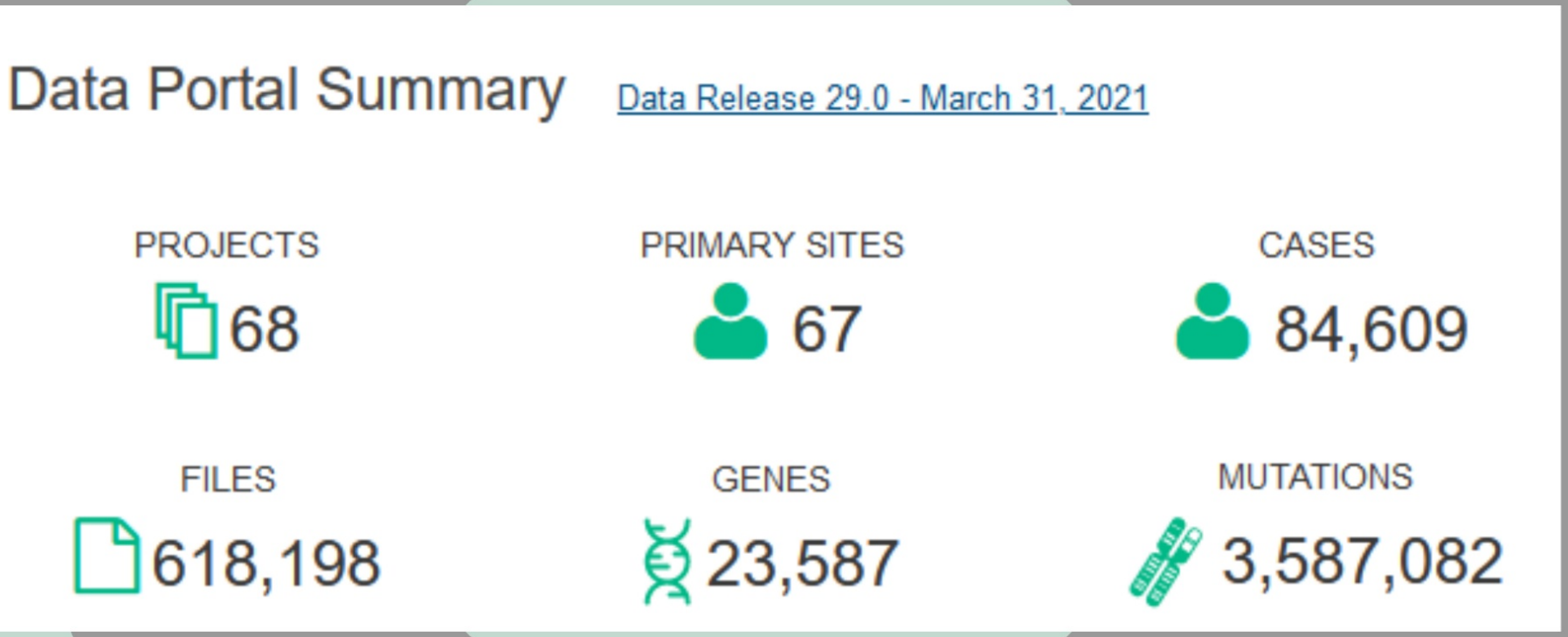
Master of Science in Data Science

July 2021


OUTLINE
Intro
Literature Review
Data
Methods
Results
Discussion

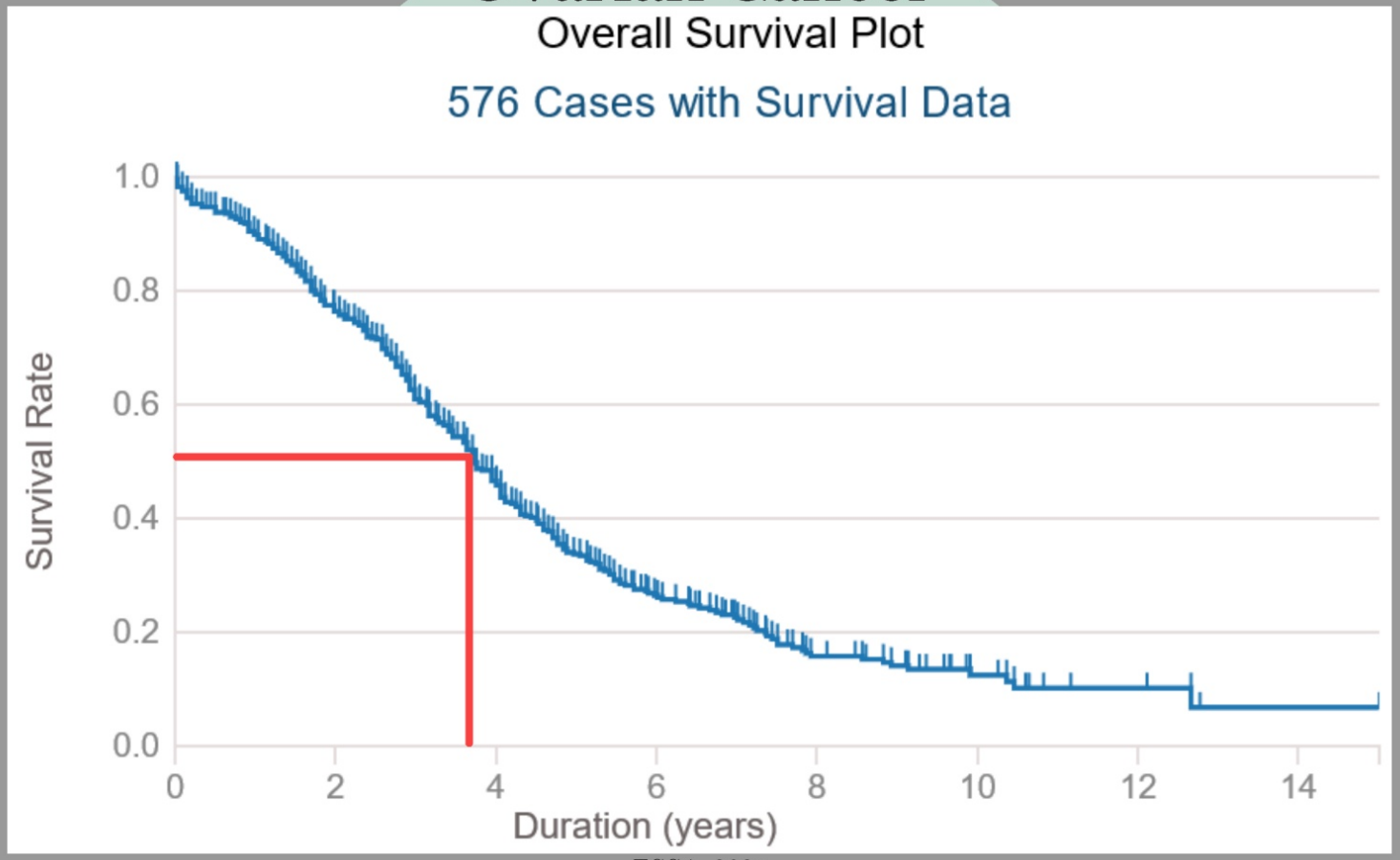

INTRO
TCGA
Ovarian Cancer
CNV
Collagen
Methods Used

TCGA

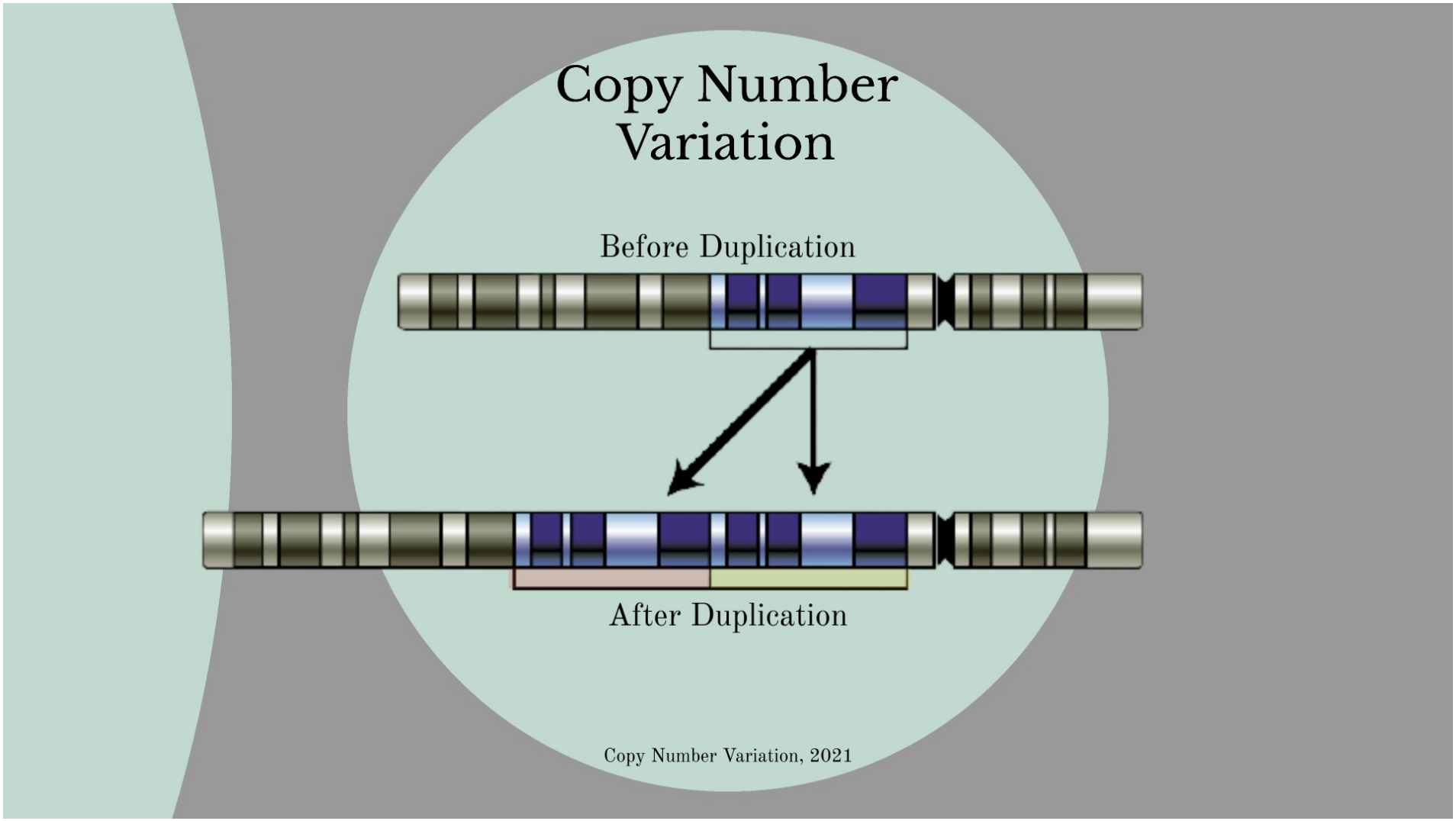


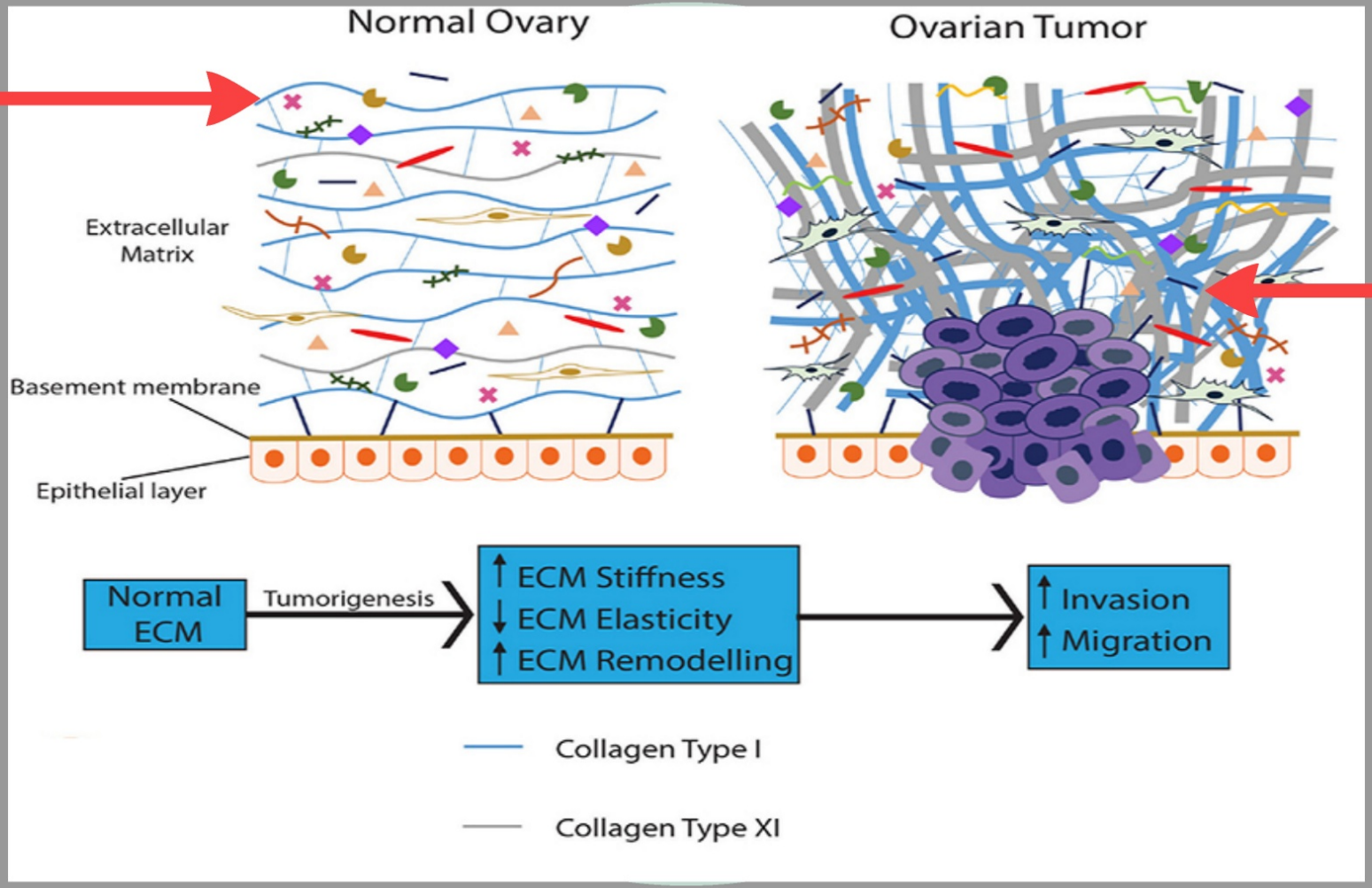
TCGA, 2021

Ovarian Cancer



TCGA, 2021

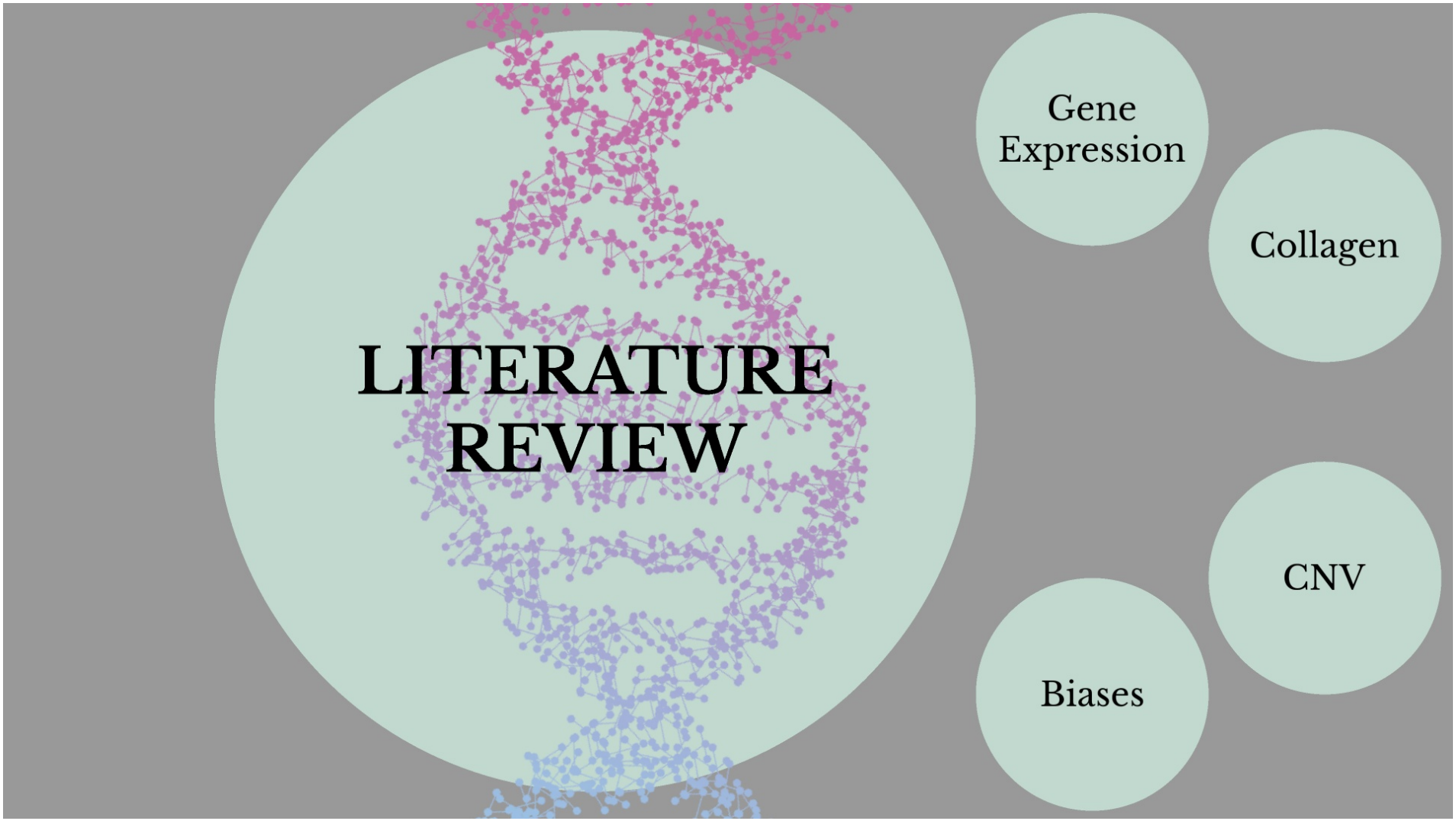

Copy Number Variation
Before Duplication
After Duplication
Copy Number Variation, 2021



Cho et al., 2015

Survival Analysis

- Kaplan-Meier Curve
- Log-rank tests
- Cox-proportional hazards model


LITERATURE REVIEW
Gene Expression
Collagen
CNV
Biases

Gene Expression

- Most studies focus on expression and SNPs
  - Clancy, 2014
- More CNV research needed

Collagen & Cancer

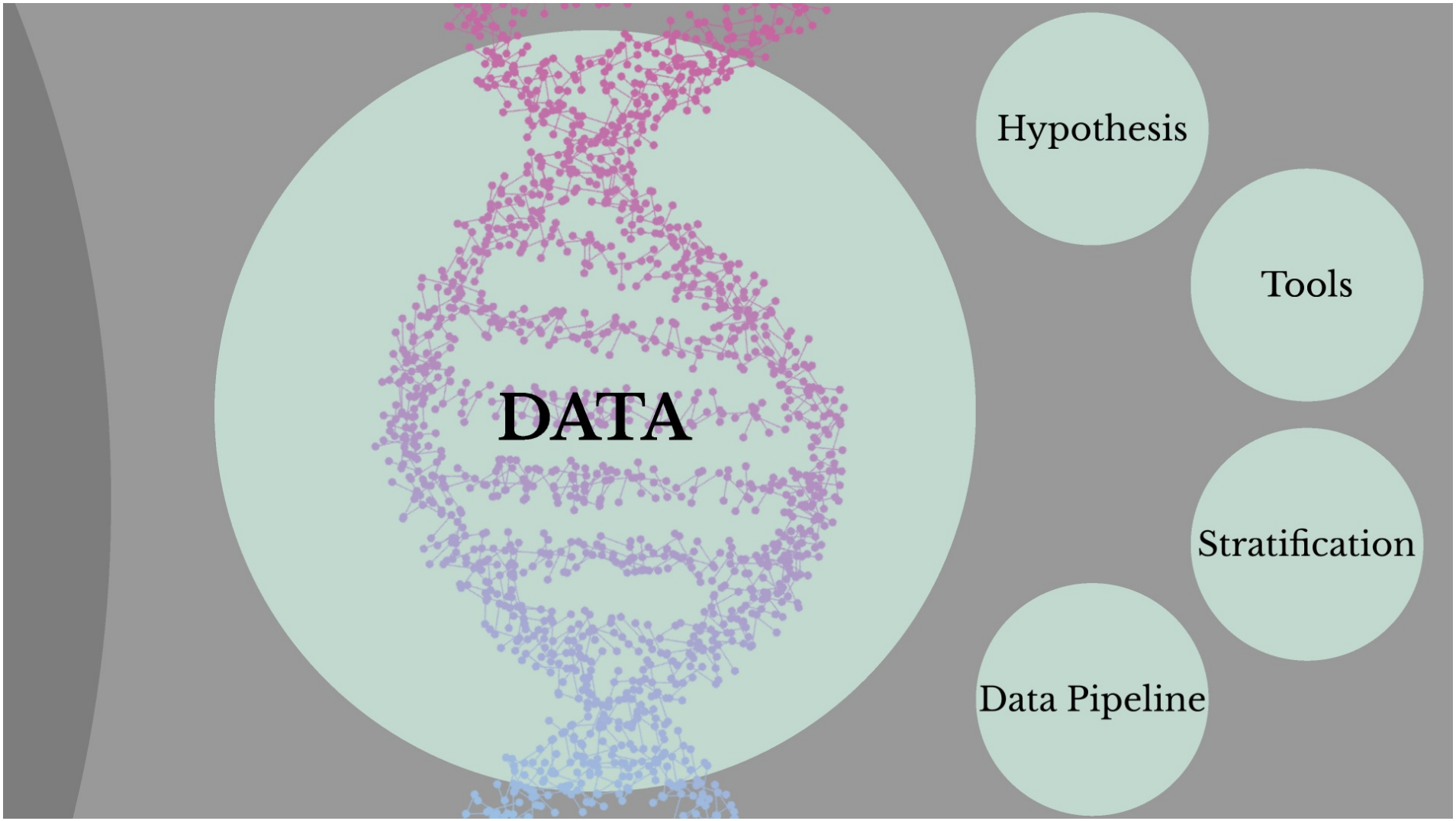
- Involved in tumorigenesis
  - Cho et al., 2015
- Gene expression studies most common it seems

CNV & Cancer

- Gene-drug interactions
  - Spainhour & Qui, 2016
- CNV and drug exposure/survival data
  - Spainhour et al., 2017
- CNV survival predictor across cancers
  - Van Dijk et al., 2021

Biases

- Technical artifacts & pseudogenes
  - Buckley et al., 2017
- Accuracy has improved over time
  - Gao et al., 2019


DATA
Hypothesis
Tools
Stratification
Data Pipeline

Hypothesis

Null - no statistical difference in survival

Alternate - statistical difference in survival

DV - time to event (death)

IV - CNVs for 55 collagen genes

Alpha = 0.05

Tools

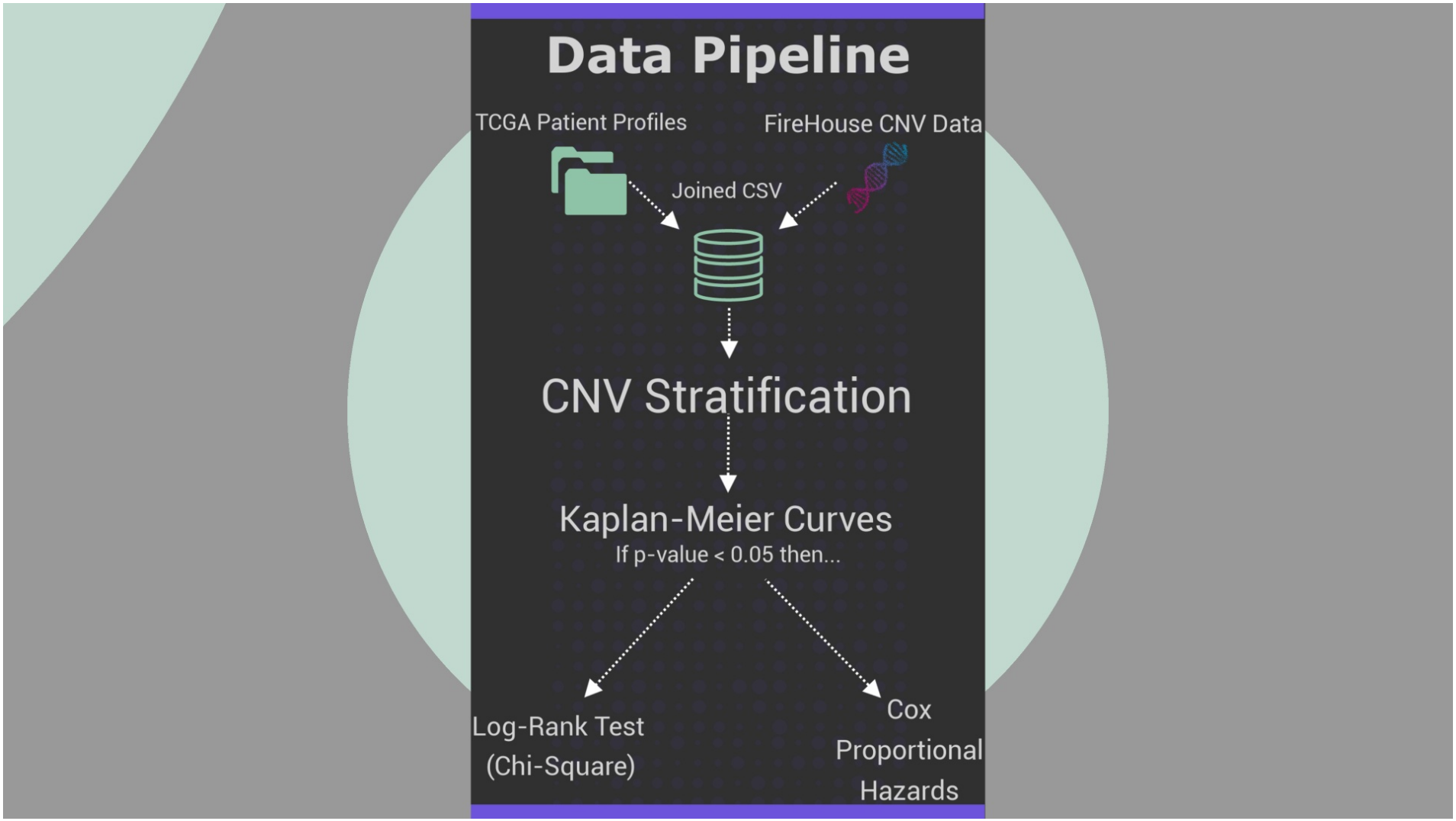
- Alteryx Designer
  - Data wrangling
- R-Studio
  - EDA + data wrangling
- SPSS
  - Survival analysis
  - Breslow method

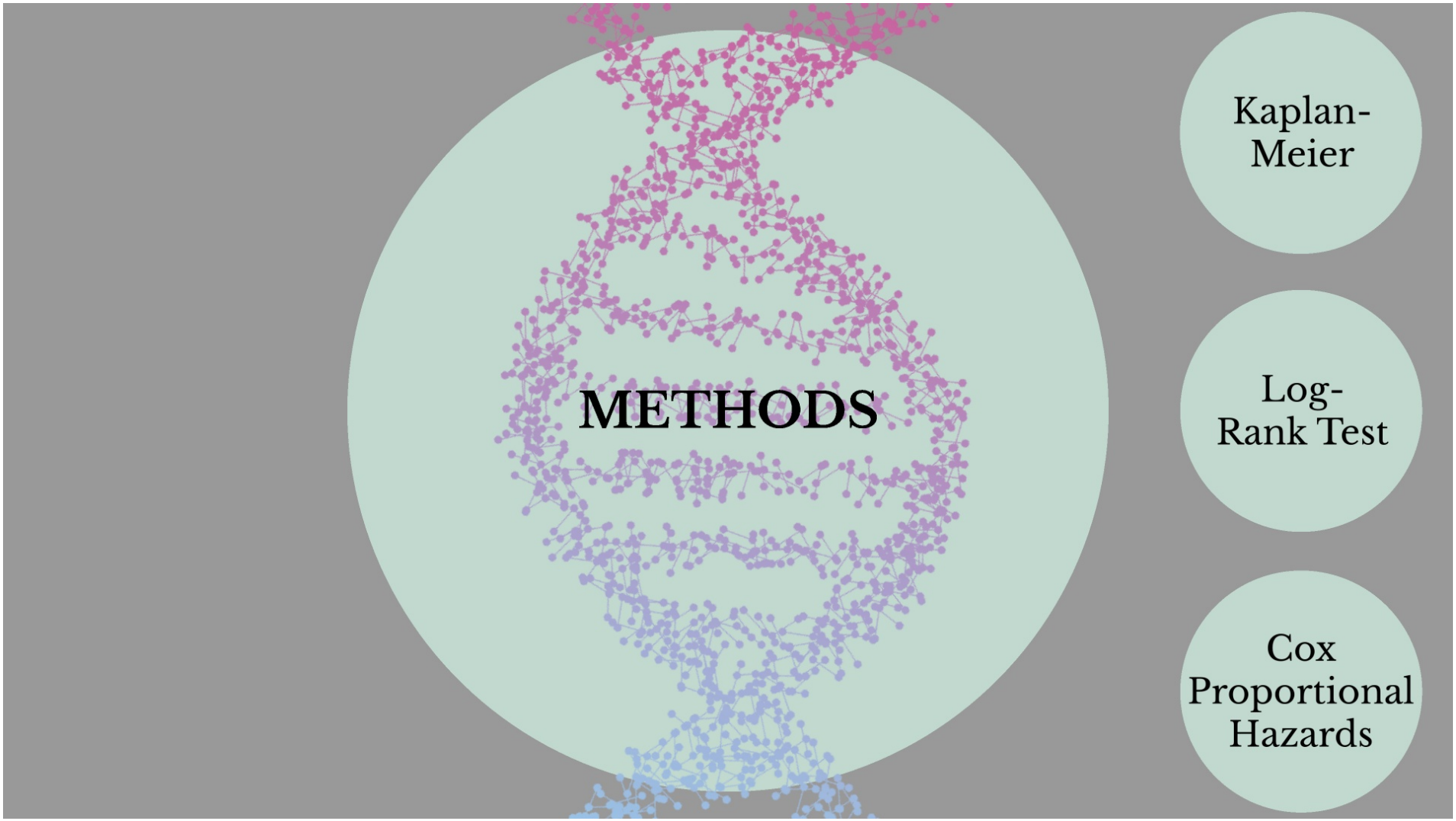
CNV Groupings

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

Examples of studies with same groupings

- Kim et al., 2016
- Spainhour & Qui, 2016


Data Pipeline
TCGA Patient Profiles
FireHouse CNV Data
Joined CSV
CNV Stratification
Kaplan-Meier Curves
If p-value < 0.05 then...
Log-Rank Test
(Chi-Square)
Cox
Proportional
Hazards


METHODS
Kaplan-Meier
Log-Rank Test
Cox Proportional Hazards

Kaplan-Meier

- Most common for survival analysis
  - Prinja et al., 2010
- Censored data
  - No event or lost to followup
- Assumptions
  - Patient outcome = independent
  - Censoring ≠ ↑ or ↓ likelihood events

Kaplan & Meier, 1958

Log-Rank Test

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Square Formula

P-value from critical values table

Relative Risk

RR = (a/n) / (b/n)

Peto & Peto, 1972

Cox Proportional Hazards

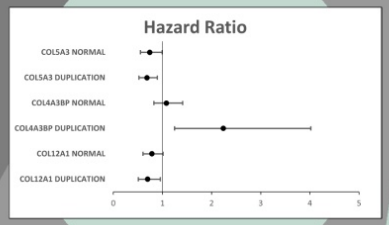
$$F(t) = P(T < t) = \int_o^t f(u)du$$

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

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

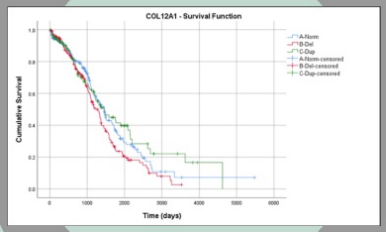
Cox, 1972

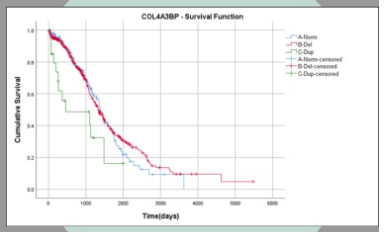
RESULTS

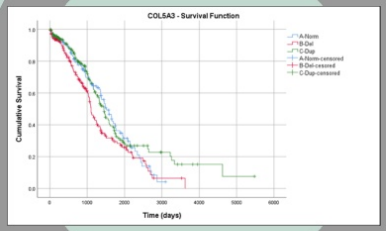


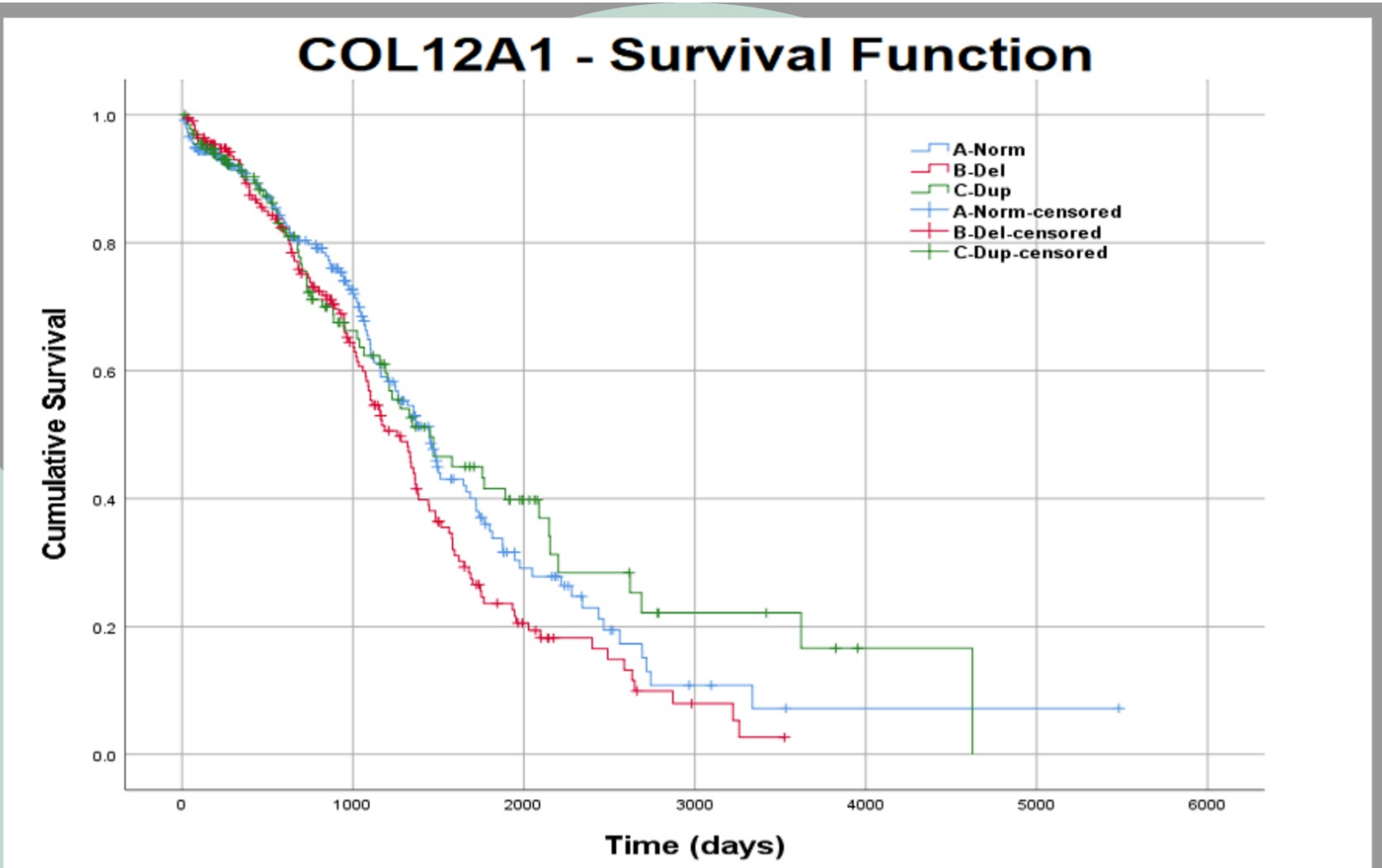
Gene	Coefficient	P-value	Chi-Square	Hazard Ratio	Lower 95% CI	Upper 95% CI
COL5A3 Normal		0.015*	8.56			
COL5A3 Deletion	0.299	0.045*		1.349	1.006	1.808
COL5A3 Duplication	-0.083	0.586		0.921	0.684	1.240
COL4A3BP Normal		0.026*	7.698			
COL4A3BP Deletion	-0.075	0.586		0.928	0.709	1.214
COL4A3BP Duplication	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

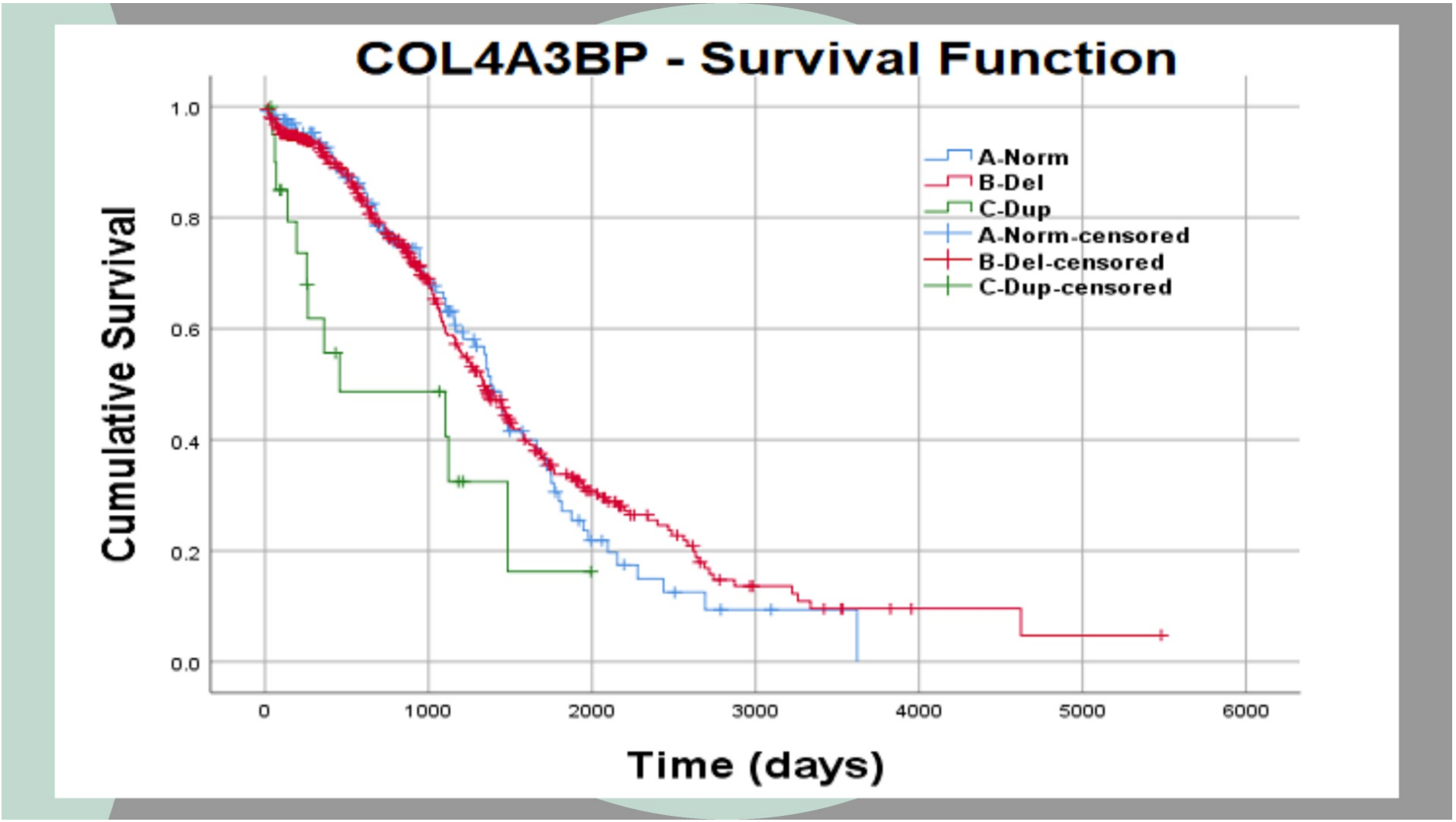
*Statistical significance with p-value < 0.05.

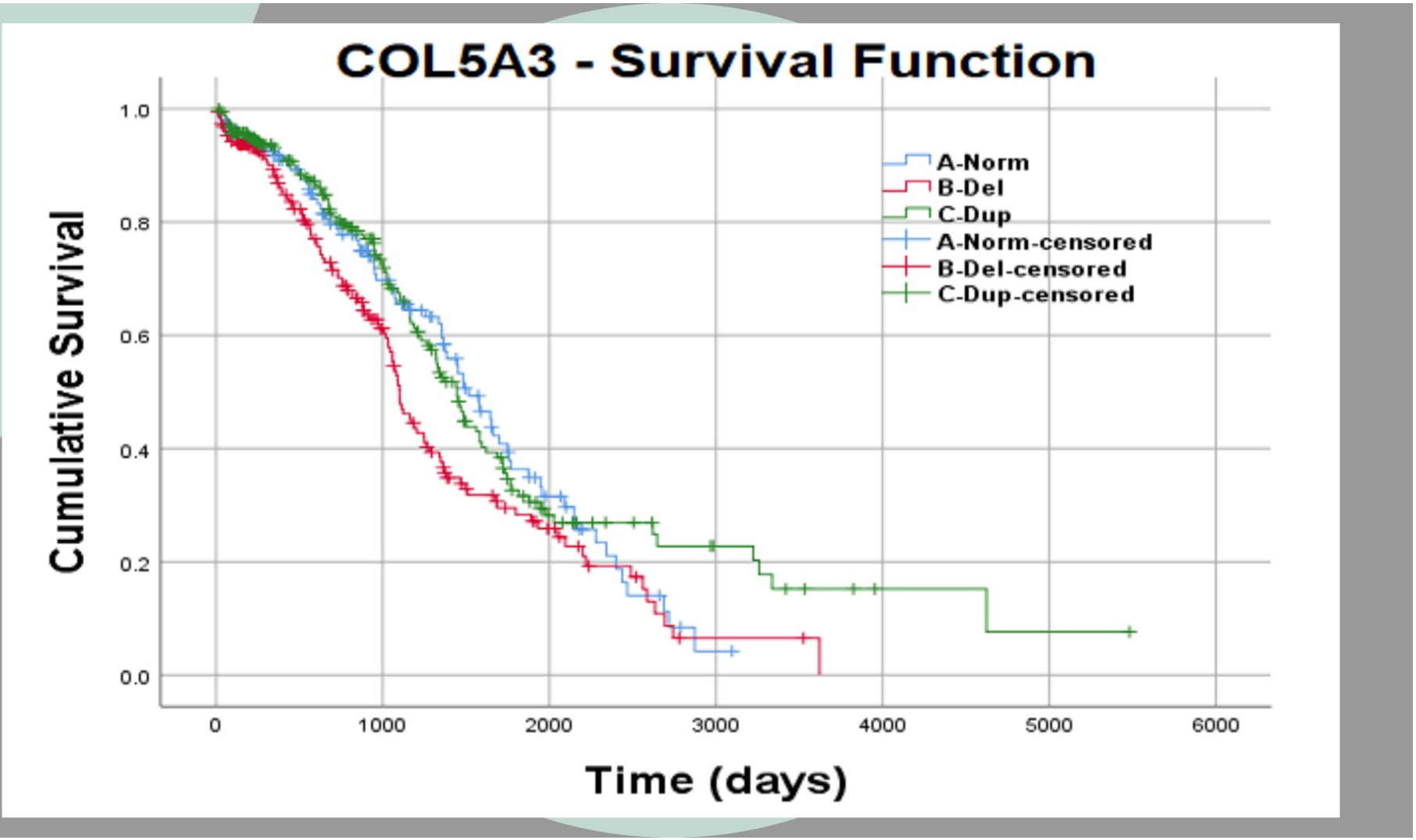







COL12A1 - Survival Function
Cumulative Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
Time (days)
A-Norm
B-Del
C-Dup
A-Norm-censored
B-Del-censored
C-Dup-censored


COL4A3BP - Survival Function
Cumulative Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
Time (days)
A-Norm
B-Del
C-Dup
A-Norm-censored
B-Del-censored
C-Dup-censored

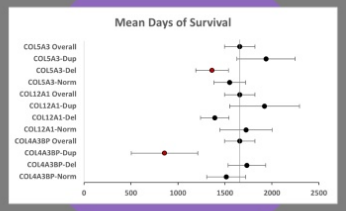

COL5A3 - Survival Function
Cumulative Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
Time (days)
A-Norm
B-Del
C-Dup
A-Norm-censored
B-Del-censored
C-Dup-censored

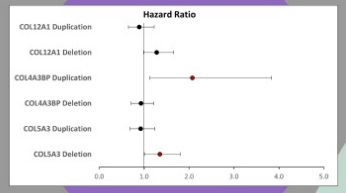
Results

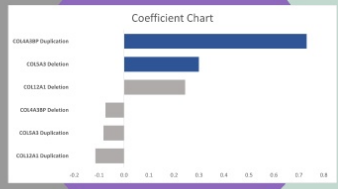
Gene	Coefficient	P-value	Chi-Square	Hazard Ratio	Lower 95% CI	Upper 95% CI
COL5A3 Normal		0.015*	8.56			
COL5A3 Deletion	0.299	0.045*		1.349	1.006	1.808
COL5A3 Duplication	-0.083	0.586		0.921	0.684	1.240
COL4A3BP Normal		0.026*	7.698			
COL4A3BP Deletion	-0.075	0.586		0.928	0.709	1.214
COL4A3BP Duplication	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

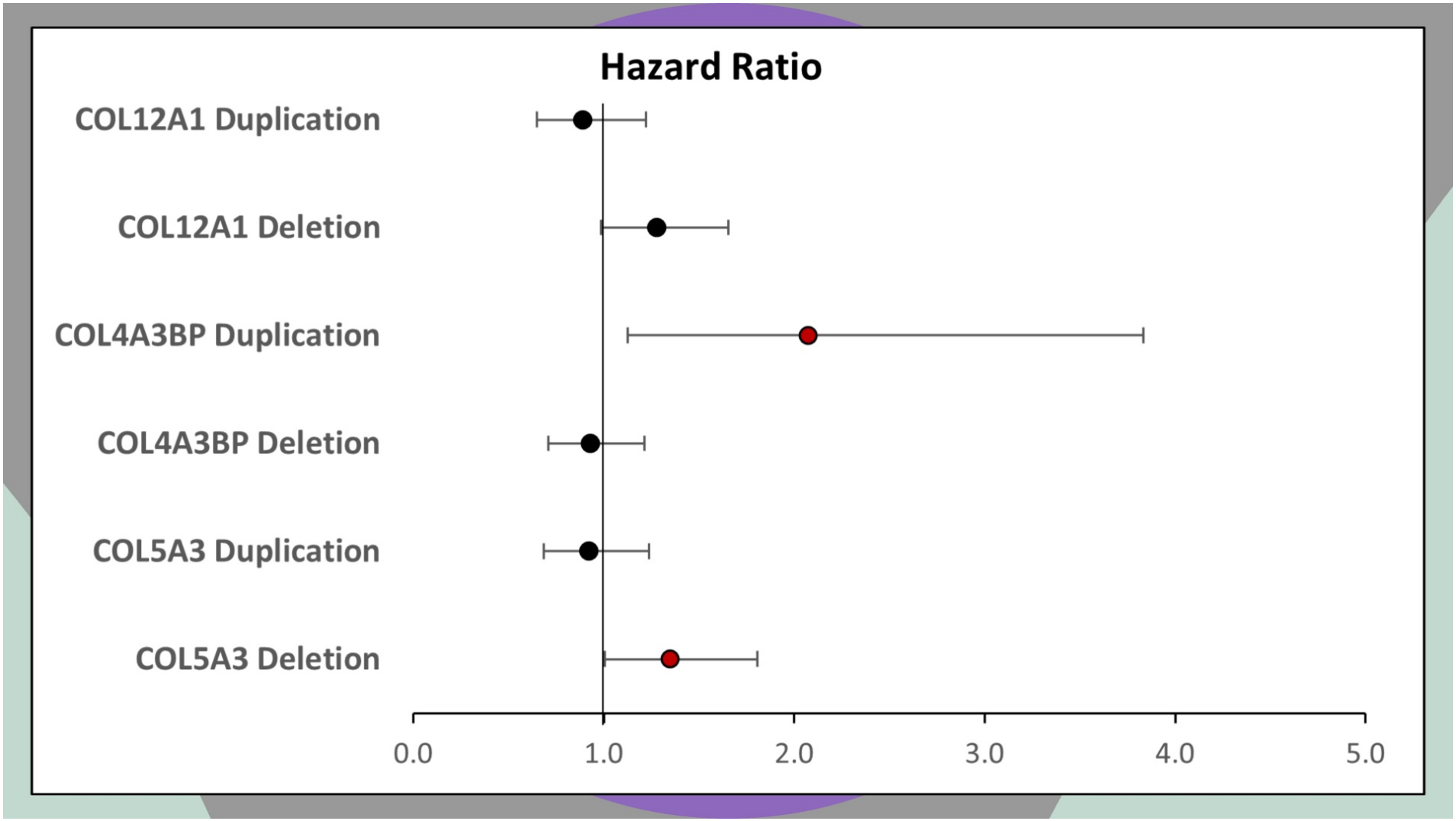
***Statistical significance with p-value < 0.05.**

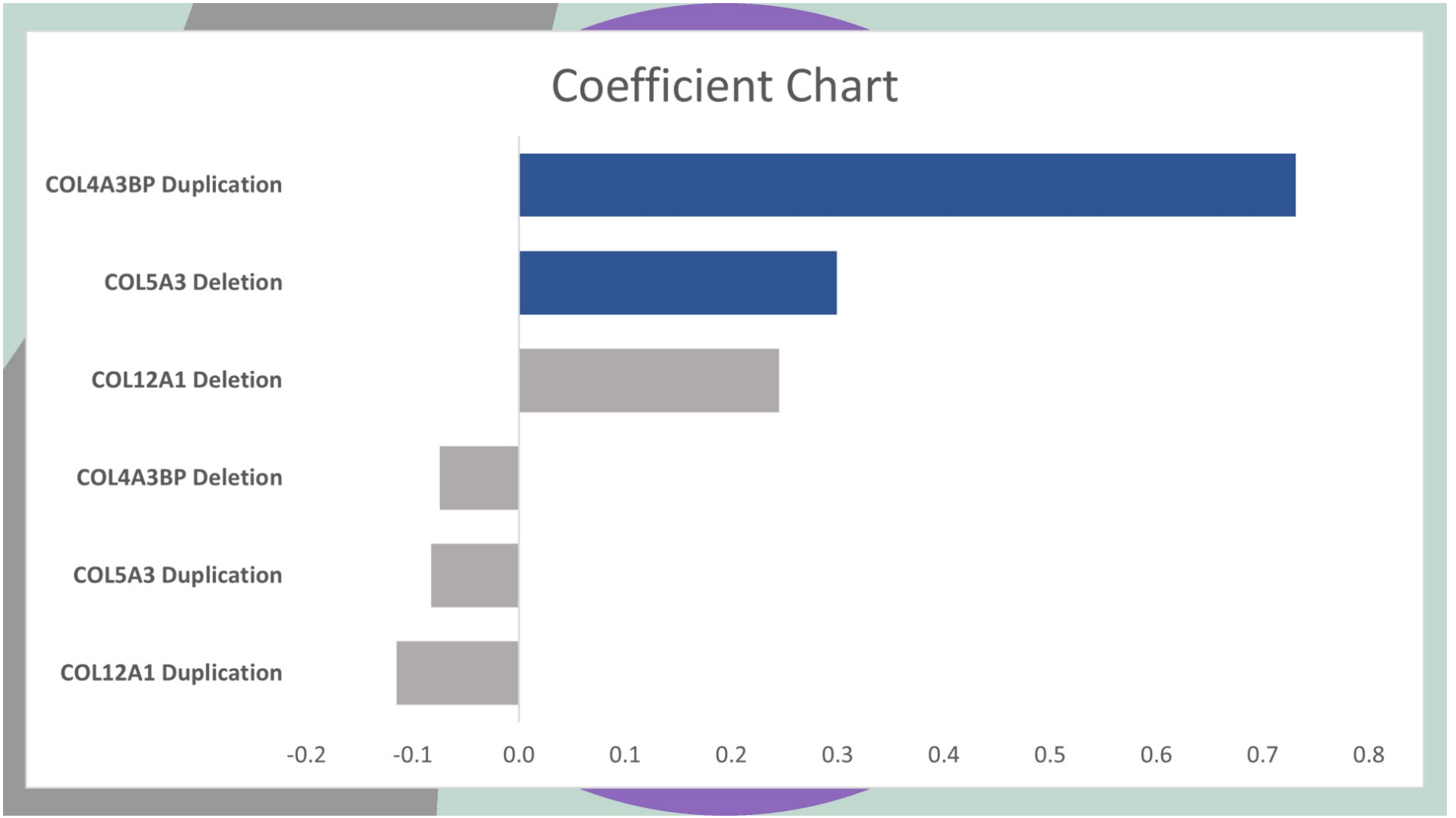
Hazard Ratio
Coefficient
Mean Days of Survival

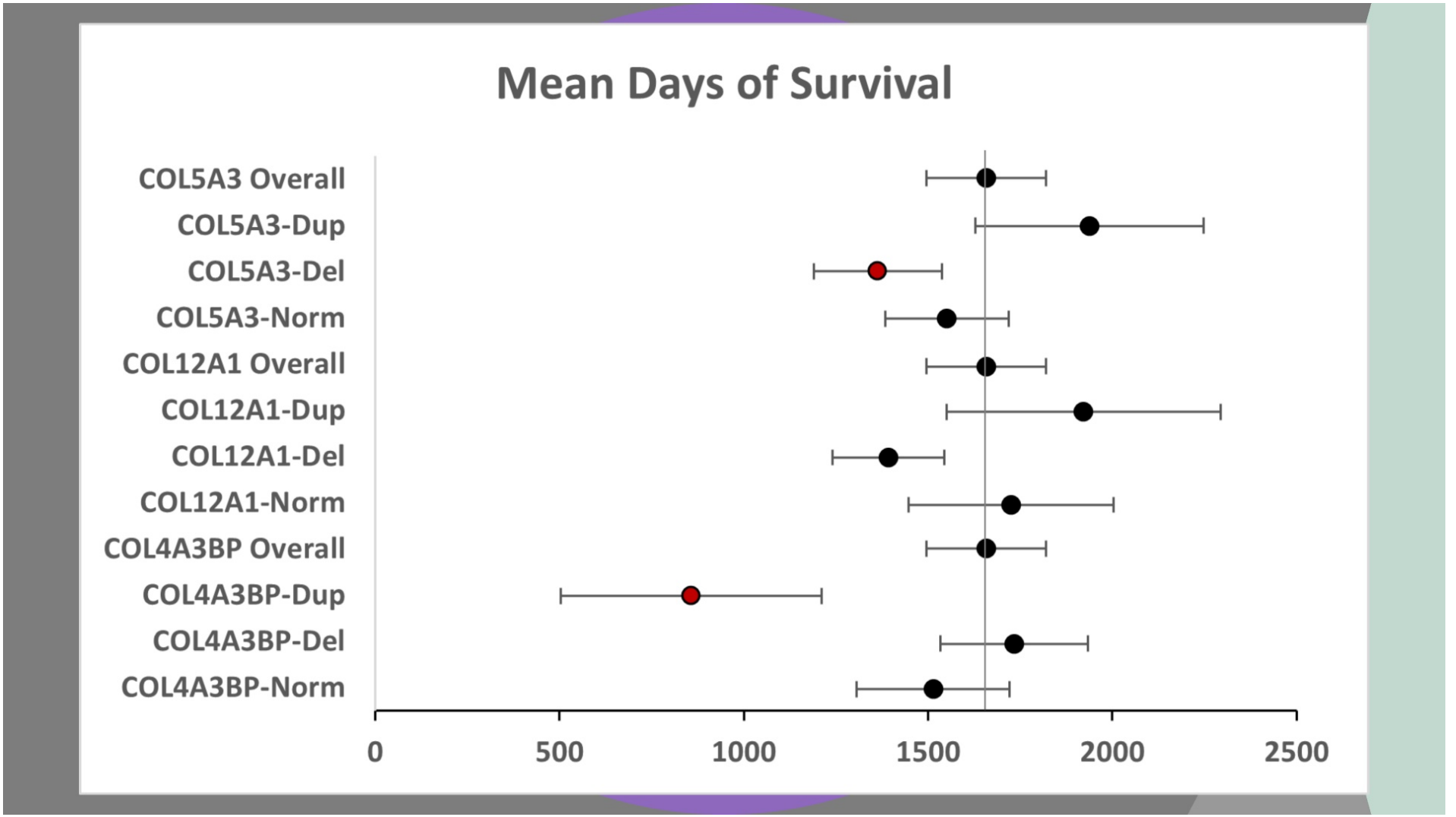


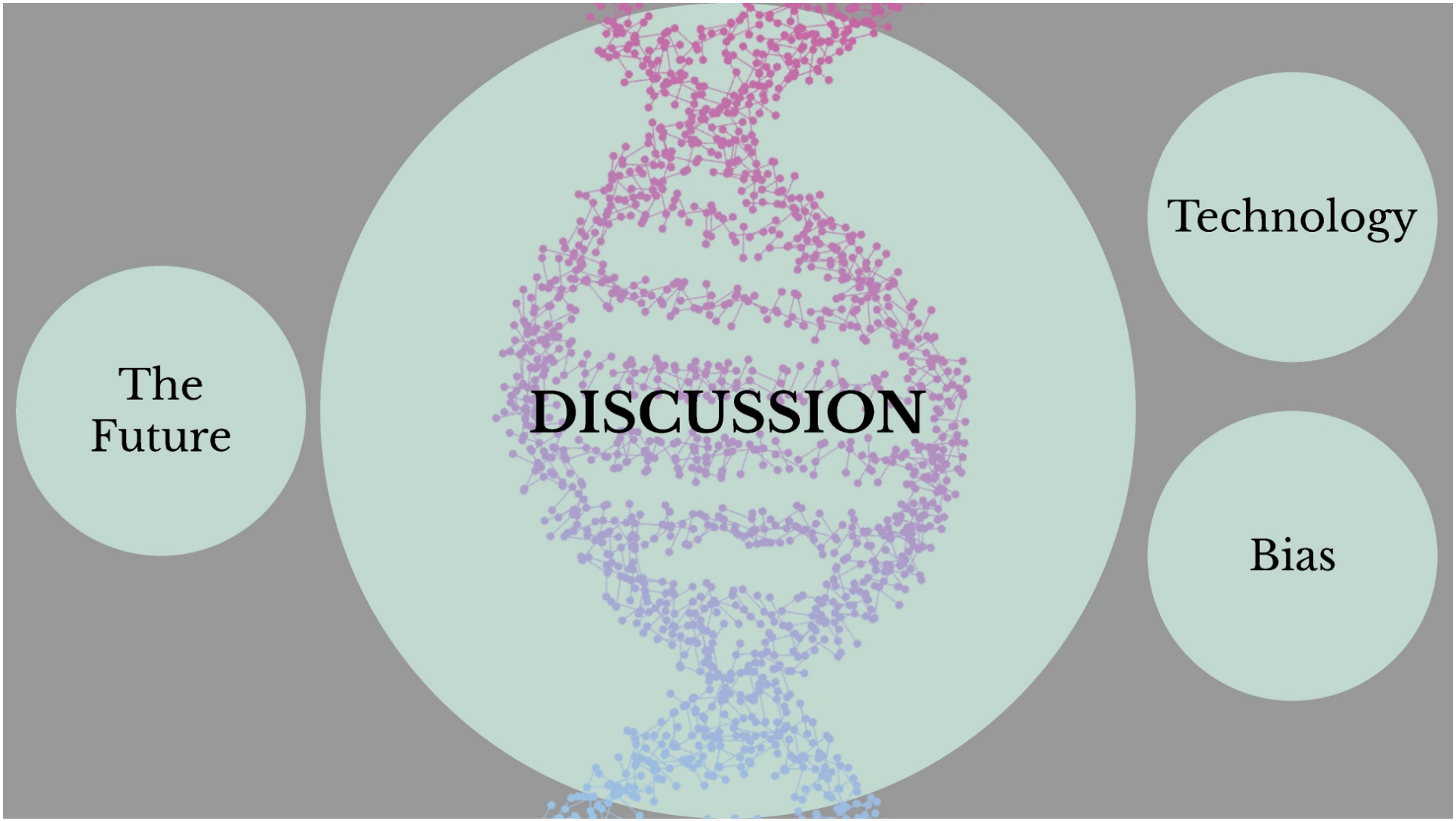





Hazard Ratio
COL12A1 Duplication
COL12A1 Deletion
COL4A3BP Duplication
COL4A3BP Deletion
COL5A3 Duplication
COL5A3 Deletion
0.0
1.0
2.0
3.0
4.0
5.0


Coefficient Chart
COL4A3BP Duplication
COL5A3 Deletion
COL12A1 Deletion
COL4A3BP Deletion
COL5A3 Duplication
COL12A1 Duplication
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Mean Days of Survival
COL5A3 Overall
COL5A3-Dup
COL5A3-Del
COL5A3-Norm
COL12A1 Overall
COL12A1-Dup
COL12A1-Del
COL12A1-Norm
COL4A3BP Overall
COL4A3BP-Dup
COL4A3BP-Del
COL4A3BP-Norm
0
500
1000
1500
2000
2500


The Future
DISCUSSION
Technology
Bias

Technology

- CNV Detection
  - Jiang et al., 2018
- Sequencing Technology
  - Sanger Sequencing - TCGA, 2021
- qPCR
  - Krasnov et al., 2019

False Discovery

- Type-I error in genetic research
  - George, 2012
- 30,000 genes at 5% error rate = 1,500 genes
- Most genetic studies set alpha at 0.05
  - Columbia Public Health, 2019

Controls

Other Controls

Controls

- Bayesian survival analysis
  - Kelter, 2020; Omurlu et al., 2019
- Log-rank vs Wilcoxon
  - Peti & Peto, 1972
- Breslow method - Cox Model
  - Breslow, 1972

Other Controls

- Known affects of collagen
  - Xu et al., 2019
- Control of evolutionary artifacts
  - Krasnov et al., 2019

The Future

- Personalized cancer treatments...dosing is difficult
  - Eaton & Lyman, 2019
- Targeted gene therapy
  - Liang et al., 2015
- CNV signature test
  - Van Dijk et al., 2012
- Social responsibility
- More studies needed

References

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

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

Cho, A., Howell, V. M., & Colvin, E. K. (2015). The Extracellular Matrix in Epithelial Ovarian Cancer – A Piece of a Puzzle. Frontiers in Oncology, 1(5). https://doi.org/10.3389/fonc.2015.00245

Clancy, S. (2014). Copy Number Variation | Learn Science at Scitable. Nature.com. https://www.nature.com/scitable/topicpage/copy-number-variation-445/

Columbia Public Health. (2019). Www.publichealth.columbia.edu. https://www.publichealth.columbia.edu/research/population-health-methods/false-discovery-rate

Copy number variation. (2021, June 6). Wikipedia. https://en.wikipedia.org/wiki/Copy_number_variation

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

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

Gao, G. F., Parker, J. S., Reynolds, S. M., Silva, T. C., Wang, L.-B., Zhou, W., Akbani, R., Bailey, M., Balu, S., Berman, B. P., Brooks, D., Chen, H., Cherniack, A. D., Demchok, J. A., Ding, L., Felau, I., Gaheen, S., Gerhard, D. S., Heiman, D. I., & Hernandez, K. M. (2019). Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. Cell Systems, 9(1), 24-34.e10. https://doi.org/10.1016/j.cels.2019.06.006

George, A. W. (2012). Controlling type 1 error rates in genome-wide association studies in plants. Heredity, 111(1), 86–87. https://doi.org/10.1038/hdy.2012.101

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

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

Kim, D. S., Kim, J. H., Burt, A. A., Crosslin, D. R., Burnham, N., Kim, C. E., McDonald-McGinn, D. M., Zackai, E. H., Nicolson, S. C., Spray, T. L., Stanaway, I. B., Nickerson, D. A., Heagerty, P. J., Hakonarson, H., Gaynor, J. W., & Jarvik, G. P. (2016). Burden of potentially pathologic copy number variants is higher in children with isolated congenital heart disease and significantly impairs covariate-adjusted transplant-free survival. The Journal of Thoracic and Cardiovascular Surgery, 151(4), 1147-1151.e4. https://doi.org/10.1016/j.jtcvs.2015.09.136

References, cont.

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

Liang, L., Fang, J-Y., & Xu, J. (2015). Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. Oncogene, 35(12), 1475–1482. https://doi.org/10.1038/onc.2015.209

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

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

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

Sanger Sequencing - TCGA. (2021). Tcga.org.uk. https://tcga.org.uk/dna-sequencing/

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

TCGA. (2021, March). Portal.gdc.cancer.gov. https://portal.gdc.cancer.gov

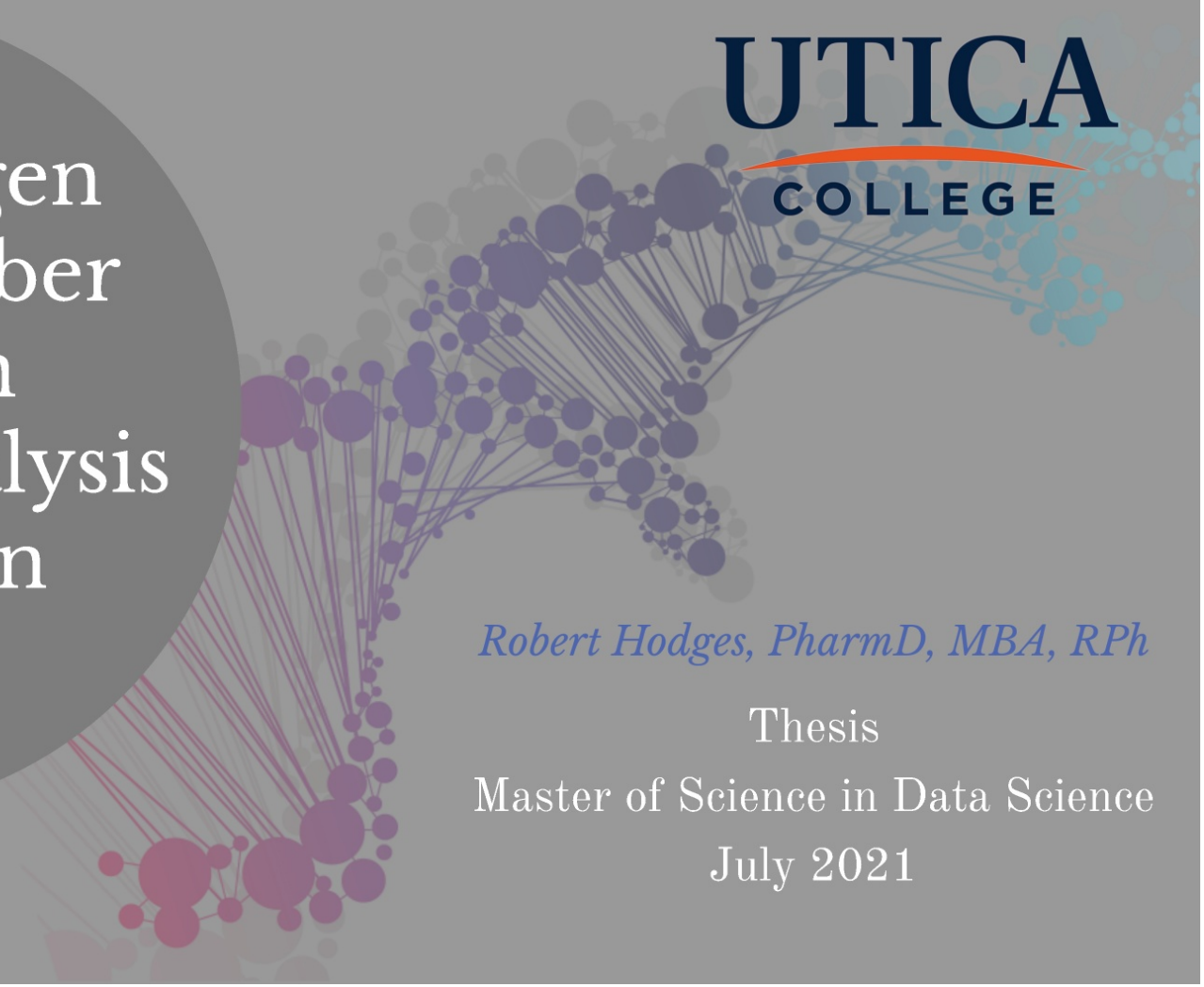
Van Dijk, E., van den Bosch, T., Lenos, K. J., El Makrini, K., Nijman, L. E., van Essen, H. F. B., Lansu, N., Boekhout, M., Hageman, J. H., Fitzgerald, R. C., Punt, C. J. A., Tuynman, J. B., Snippert, H. J. G., Kops, G. J. P. L., Medema, J. P., Ylstra, B., Vermeulen, L., & Miedema, D. M. (2021). Chromosomal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

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

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

GitHub: https://github.com/hodgesr2/Pan-Collagan-Ovarian-Cancer-Study-from-TCGA

Pan-Collagen Copy Number Variation Survival Analysis in Ovarian Cancer



UTICA COLLEGE

Robert Hodges, PharmD, MBA, RPh

Thesis

Master of Science in Data Science

July 2021