Title
Application of Statistical Models to Study Protein Stability in Subcellular Compartments

Author
Xiao, Shaohua

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Application of Statistical Models to Study Protein Stability in Subcellular Compartments

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Applied Statistics

by

Shaohua Xiao

2020

ABSTRACT OF THE THESIS

Application of Statistical Models to Study Protein Stability

in Subcellular Compartments

by

Shaohua Xiao

Master of Science in Applied Statistics

University of California, Los Angeles, 2020

Professor Hongquan Xu, Chair

Gap junction protein connexin-43 (Cx43) is essential for intercellular communication and synchronous muscle contraction of the mammalian heart. Cx43 is enriched in the intercalated discs in normal hearts, but redistributed in arrhythmic failing hearts. To understand the pathophysiological mechanisms for arrhythmias and guide therapeutic interventions, it will be important to assess whether the mis-localized Cx43 remains stable or undergoes rapid turnover. The M213L mutation in Cx43 has been shown to impair delivery of the protein to the intercalated discs. Stabilities of the wild type and M213L mutated Cx43 protein in subcellular fractions have been reported. Here we applied multiple statistical models to the protein stability data to further analyze the effects of subcellular compartments and mutation on protein

degradation. Our results indicated that fraction, but not the M213L mutation, affected Cx43 degradation, with the protein in the non-junctional fraction being degraded rapidly.

The thesis of Shaohua Xiao is approved.

Nicolas Christou

Ying Nian Wu

Hongquan Xu, Committee Chair

University of California, Los Angeles

2020

DEDICATION

This thesis is dedicated to my parents for their unwavering support and encouragement.

TABLE OF CONTENTS

1 Introduction..... 1

2 Experimental procedure and data collection..... 3

3 Data analyses and results..... 5

3.1 Exponential time constant as the response variable.....6

3.2 Percentage of protein to the initial amount as the response variable.....10

3.3 Ratio of labeled to total protein as the response variable.....15

3.4 MANOVA analyses.....19

4 Summary.....22

Appendix: Supplemental figures.....24

References.....30

LIST OF FIGURES

1. Schematic diagrams of experimental workflow and data collection..... 4

2. Treatment curves..... 5

3. Interaction plot of fraction and time (model 3).....13

4. Interaction plot of fraction and time (model 5).....17

LIST OF TABLES

1. The exponential time constant, τ_1, of Cx43 protein……..……..…… 7

2. Estimated parameters of models 1 and 2…….. 8

3. ANOVA analysis of full models 1 and 2…… 8

4. The exponential time constant, τ_2, of Cx43 protein…… 9

5. Estimated parameters of models 3 and 4 ……..…11

6. ANOVA analysis of full models 3 and 4……...…12

7. Estimated parameters of models 5 and 6……...16

8. ANOVA analysis of full models 5 and 6……...…17

9. Shapiro-Wilk normality test of the response variables in MANOVA analyses……19

10. MANOVA and multiple univariate ANOVA analyses with “percent_of_0hr” as the response variables…… 20

11. MANOVA and multiple univariate ANOVA analyses with “ratio” as the response variable...21

ACKNOWLEDGEMENTS

The dataset used in this thesis is the raw data from which published data in the following reference is derived.

Reference:

Xiao S, Shimura D, Baum R, Hernandez DM, Agvanian S, Nagaoka Y, Katsumata M, Lampe PD, Kleber AG, Hong TT, Shaw RM. Auxiliary trafficking subunit GJA1-20k protects connexin-43 from degradation and limits ventricular arrhythmias. *J Clin Invest*. 2020;134682. doi:10.1172/JCI134682

CHAPTER 1

Introduction

Connexin-43 (Cx43) protein forms gap junction channels between cardiomyocytes, one of the major cell types in the mammalian heart [1], enabling rapid propagation of electrical impulses and synchronous heart muscle contraction [2]. Cx43 channels are mainly localized to specialized membrane compartments, the intercalated discs, on the longitudinal ends of cardiomyocytes. In the diseased heart, Cx43 expression is down regulated and redistributed laterally [3]. Rescuing the localization of Cx43 channels to the intercalated discs could be potential therapies for arrhythmias [3]. Understanding the fate of the mis-localized Cx43 protein, whether it is rapidly degraded or remains stable, will help decide what biological processes therapeutic interventions should target.

The M213L mutation in Cx43 has been shown to impair the targeting of the protein to the intercalated discs in the heart of a mouse model [2] and to the junctional fraction in cultured cells [4]. The M213L mutant mice display sudden cardiac death phenotype and reduced expression of the Cx43 protein [2], similar to phenotypes observed in failing hearts. To investigate how the mutation and different subcellular locations affect the stability of Cx43, the authors performed pulse-chase experiments in cultured cells, a common method to study protein degradation. The authors calculated the exponential time constant, τ, of the wild type and M213L mutated Cx43 in different cellular fractions. They found that fraction, but not mutation, was significantly associated with Cx43 protein degradation [2].

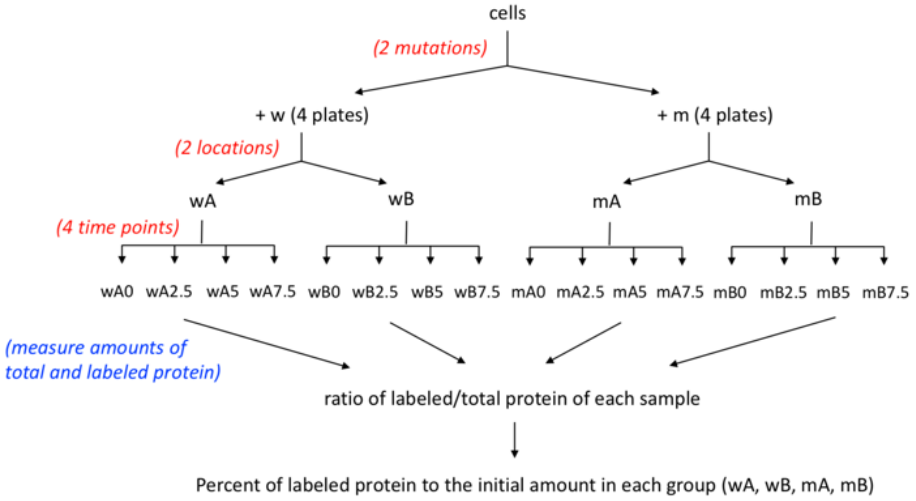
In this thesis project we applied multiple statistical models with R to further analyze the pulse-chase experimental data, which were obtained from the authors of reference [2]. The statistical approaches included ANOVA test [5-6], linear regression [7-8], randomized block design, and multivariate analysis [9]. The results indicated that subcellular fraction affected Cx43 protein degradation, whereas mutation did not. The protein was more stable in the junctional membrane fraction than in the non-junctional fraction. Our findings are consistent with what is reported in reference [2]. It is likely that the mutated Cx43 protein unable to traffic to the junctional membrane is effectively degraded. Therefore, when targeting Cx43 as therapeutic interventions for arrhythmias, we should consider increasing the production of Cx43 and facilitating its delivery to the junctional fraction at the same time.

CHAPTER 2

Experimental procedure and data collection

Figure 1A outlines the workflow of the pulse-chase experiments to measure stability the Cx43 protein (based on reference [2]). Cells were divided into two groups, which were randomly chosen to express either the wild type (w) or the M213L mutated protein (m). The Cx43 protein produced in cells within an hour was labeled, which was referred to the "pulse" phase of the experiment. The level of the labeled protein was monitored over 7.5 hours with 2.5-hour intervals, which was the "chase" phase of the experiment. At four time points (0, 2.5, 5, 7.5) during the chase period, aliquots of cells were fractionated into the non-junctional ("A") and junctional membrane ("B") fractions. Labeled and total Cx43 proteins in each subcellular fraction were purified and transferred onto a membrane for detection (Figure 1B). Intensities of the protein bands were quantified. The amount of labeled protein was normalized to that of the total protein in the same fraction at each time point, which resulted in an output variable, "ratio". The ratios of protein with the same mutation and in the same subcellular fraction were divided by the ratio at the 0-hour time point, generating another output variable, "percent_of_0hr" (Figure 1B). The same procedure was performed four times, each of which was treated as a batch. We coded the wild type protein and M213L mutation as "-1" and "1", respectively, for the variable "mutation". The non-junctional fraction and junctional membrane fraction, the two levels of the "fraction" variable, were coded as "-1" and "1", respectively. The original design of the experiment had 64 observations. Because of problems with some samples, 4 observations corresponding to fraction "1" and mutation "1" in batch 1 were excluded from the analysis. The dataset used in this thesis has 60 observations.

A



B

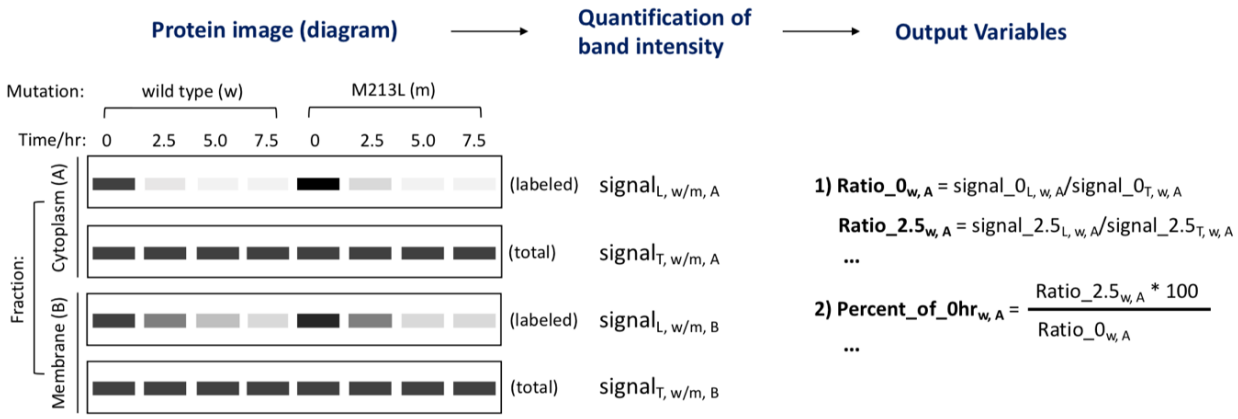


Figure 1. Schematic diagrams of experimental workflow (panel A) and data collection (panel B) (based on [2]). The subscript "L" and "T" indicate the labeled and total protein, respectively.

CHAPTER 3

Data analysis and results

We first visualized the data by plotting the amount of labeled protein, expressed as the percentage to the level at the 0-hour time point, against time, coloring by fraction and mutation (Figure 2). The treatment curves colored by fraction showed clear separation, especially after the 5-hour time point (Figure 2A). The curves colored by mutation were largely overlapped across time points (Figure 2B). The observations suggest that subcellular fraction might have an effect on protein degradation.

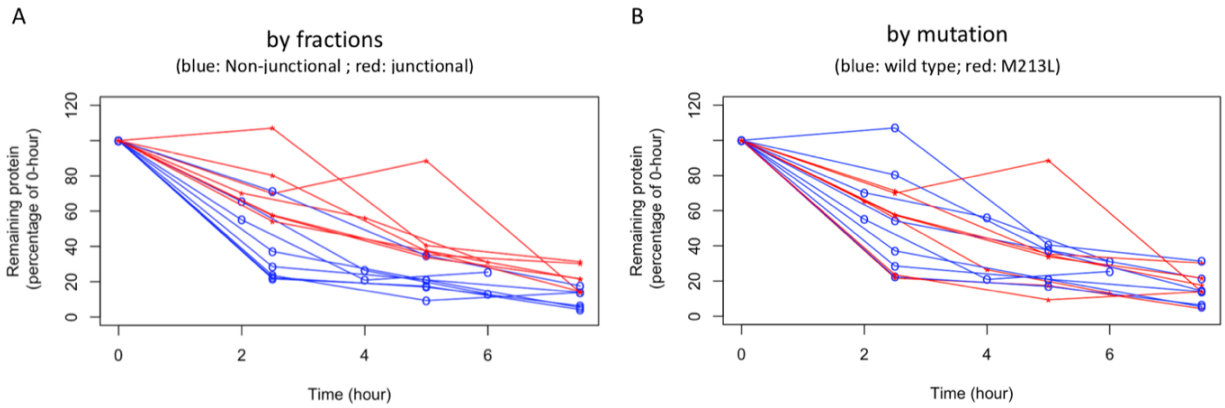


Figure 2. Treatment curves. The amounts of labeled wild type and M213L mutated Cx43 protein in junctional membrane and non-junction fractions are plotted against time, colored by subcellular fraction (A) and mutation (B).

To quantitatively analyze the contribution of mutation and subcellular fraction on Cx43 protein degradation we used multiple statistical models to fit the data. The methods included 1) using exponential regression to deduce the time constant of protein, τ, follow by a linear regression model with τ as the response variable; 2) a linear regression model and the response variable was the percentage of the labeled protein to the level at the 0-hour time point (the "percent_of_0hr"

variable in Figure 1B); 3) a linear regression model and the response variable was the ratio of the level of the labeled protein to that of the total protein (the "ratio" variable in Figure 1B); 4) multivariate regression models with the "percent_of_0hr" and "ratio" at various time points as the dependent variables.

3.1 Exponential time constant as the response variable

A common way to quantify and analyze protein degradation is to fit the amount of the remaining protein, the variable "percent_of_0hr" in this case, at various time points to an exponential decay model. The time constant, τ, is deduced from the exponential equation $y = ae^{-t/\tau}$, where y and t denote the percentage of the remaining labeled protein and time, respectively. The "a" is the initial amount of labeled protein, which in this case should be around 100%. Significance of the effect of mutation and fraction is evaluated using two-way ANOVA [2].

In this project, we first reproduced the results reported by Xiao et al [2] by fitting a linear model of log-transformed "percent_of_0hr", which is theoretically equivalent to fitting an exponential equation, to calculate τ ("τ_1" in Table 1). A linear model (model 1) was used to fit the data and identify significant factors associated with τ (Table 1). The four replicates of runs, specified in the batch variable, were treated as blocks. The residual plot, QQ plot, and Boxcox plot of the model showed that no transformation of the response variable was necessary (Supplemental Figure S1). The estimated coefficients of the full model are shown in Table 2. Backward and forward stepwise regression were used to choose the model with the least AIC and all significant predictors. The final fitted model was:

$$\tau = 3.35 + 1.67 * \text{fraction1} \quad (1)$$

The adjusted R^2 of the final model was 0.5784 (Table 2). The regression model and ANOVA revealed that subcellular fraction, but not mutation or batch effect, was significant at $\alpha = 0.05$ (Table 2 and 3). The coefficient for "fraction1" in the final model was 1.67, whose sign indicated that the exponential time constant in the junctional membrane fraction (fraction "1") was greater than that in the non-junctional fraction (fraction "-1").

Table 1. The exponential time constant, τ_1, of Cx43 protein.

run	mutation*	fraction*	batch*	τ_1^
1	-1	-1	1	3.9378
2	1	-1	1	2.8540
3	-1	1	1	5.3603
4	-1	-1	2	4.0211
5	1	-1	2	2.5758
6	-1	1	2	5.0109
7	1	1	2	6.1254
8	-1	-1	3	2.9145
9	1	-1	3	3.6467
10	-1	1	3	5.6108
11	1	1	3	4.3935
12	-1	-1	4	2.6622
13	1	-1	4	4.2159
14	-1	1	4	3.7994
15	1	1	4	4.8904

*mutation: "-1" – wild type, "1" – M213L mutation
*fraction: "-1" – non-junctional fraction, "1" – junctional membrane fraction
*batch: replicates of runs
^Exponential decay was fit with linear regression of logarithmically transformed response variable, "percent_of_0hr", over time.

Table 2. Estimated parameters of models 1 and 2. The response variable is the exponential time constant, τ.

	Model 1		Model 2	
	full	final	full	final
(Intercept)	3.5505 ***	3.3535 ***	1.1073 **	0.9983 ***
mutation1	-0.0608		0.0709	
fraction1	1.5614 *	1.6737 ***	0.7639 **	0.7459 ***
batch2	0.0556		-0.3140	
batch3	-0.2363		-0.1850	
batch4	-0.4857		-0.0789	
mutation1:fraction1	0.3074		0.0179	
adjusted R^2	0.3932	0.5784	0.5478	0.6492
AIC	0.15	-8.03	-30.04	-36.57
Significance codes	. p<0.1; * p<0.05; ** p<0.01; *** p<0.001			

Table 3. ANOVA analysis of full models 1 and 2.

Model 1					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mutation	1	0.0155	0.0155	0.0208	0.8890
fraction	1	10.4544	10.4544	14.0354	0.0057 **
batch	3	0.6718	0.2239	0.3006	0.8242
mutation:fraction	1	0.0850	0.0850	0.1142	0.7441
Residuals	8	5.9589	0.7449		
Model 2					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mutation	1	4.33E-05	4.33E-05	0.0004	0.9839
fraction	1	2.0890	2.0890	20.9961	0.0018 **
batch	3	0.1950	0.0650	0.6532	0.6030
mutation:fraction	1	0.0003	0.0003	0.0029	0.9583
Residuals	8	0.7960	0.0995		

Significance Codes: . p<0.1; * p<0.05; ** p<0.01; *** p<0.001

We also calculated τ using nonlinear least square (nls) models, where we could keep the "a" in the exponential decay equation $y = ae^{-t/\tau}$ closer to the starting percentage, 100%. The list of

estimated τ, "τ_2", is shown in Table 4. Based on the Boxcox plot (Figure S2), we performed logarithmic transformation of the response variable, "percent_of_0hr", to stabilize the variance of residuals. We fit the data with a linear model ("Model 2" in Table 2). The residual plot and QQ plot of model 2 are shown in Figure S2. Based on the result of stepwise regression selection, the final regression model was:

$$log(\tau) = 1.00 + 0.75 * \text{fraction1} \quad (2)$$

The adjusted R^2 of the model was 0.6492 (Table2), which indicated a slightly better fit of data than model 1.

Table 4. The exponential time constant, τ_2, of Cx43 protein.

run	mutation	fraction	batch	τ_2^
1	-1	-1	1	3.2844
2	1	-1	1	3.3478
3	-1	1	1	5.8083
4	-1	-1	2	2.5925
5	1	-1	2	1.9538
6	-1	1	2	4.7866
7	1	1	2	5.3252
8	-1	-1	3	1.9988
9	1	-1	3	1.9307
10	-1	1	3	6.9774
11	1	1	3	8.0344
12	-1	-1	4	2.7646
13	1	-1	4	4.9483
14	-1	1	4	5.1507
15	1	1	4	4.6930

*mutation: "-1" – wild type, "1" – M213L mutation
*fraction: "-1" – non-junctional", "1" – junctional membrane fraction
*batch: replicates of runs
^Exponential decay was fit using nonlinear least square estimation.

Both the regression model and ANOVA analysis revealed that fraction had a significant effect on protein degradation at $\alpha = 0.05$, whereas mutation, batch, or the interaction term between mutation and fraction did not (Tables 2 and 3). The sign of the estimated coefficient for "fraction1" was positive, indicating that the exponential time constant of Cx43 was greater, hence higher protein stability, in the junctional membrane fraction than in the non-junctional fraction.

Since the question of interest is to examine the significant factors contributing to/associated with protein decay, one of the advantages of using τ as the response variable is that it is straightforward to interpret the result. However, the residual degrees of freedom of the full models 1 and 2 is 8, which is compromised compared to the total number of runs of the experimental design, 60. This approach also assumes that the data of protein degradation follows exponential decay. How well the exponential models fit the data will likely affect the estimated τ and the identification of significant factors.

3.2 Percentage of protein to the initial amount as the response variable

The data were also fit with a randomized block design model, with the percentage of the labeled protein relative to the starting amount, "percent_of_0hr", as the dependent variable. We wanted to test the effect of mutation, fraction, time, and batch (blocking variable) on the response variable. The "time" variable was treated as categorical with four levels, since data were collected at four time points during the "chase" period of the experiments. The levels were designated as "time0", "time2.5", "time5", and "time7.5". We used a linear model with two-factor interactions (model 3), to fit the data and performed Boxcox transformation (Figure S3) of

the response variable. The estimated parameters of the full regression model are shown in Table 5. The residual plot and QQ plot of the model are shown in Figure S3.

Table 5. Estimated parameters of models 3 and 4. The dependent variable is "percent_of_0hr".

	Model 3		Model 4	
	full	final	full	final
Intercept	10.497 ***	10.5723 ***	9.7056 ***	9.5667 ***
mutation1	0.1584		0.1172	
fraction1	0.1584	0.0818	0.9322 .	0.8119 .
time_fct2.5	-3.7513 ***	-3.7973 ***		
time_fct5	-5.6379 ***	-5.4870 ***		
time_fct7.5	-6.5414 ***	-6.6031 ***		
time			-1.3018 ***	-1.2793 ***
I(time^2)			0.0538 *	0.0512 *
batch2	-1.0133 *	-1.0081 **	-0.5124	
batch3	-0.6967 .	-0.6915 .	-0.1958	
batch4	-0.5948	-0.5895	-0.0938	
mutation1:fraction1	-0.1511		-0.1770	
mutation1:time_fct2.5	-0.0921			
mutation1:time_fct5	0.3018			
mutation1:time_fct7.5	-0.1233			
mutation1:time			0.0176	
fraction1:time_fct2.5	2.1618 **	2.1684 **		
fraction1:time_fct5	2.2604 **	2.2388 **		
fraction1:time_fct7.5	1.3764 .	1.3852 **		
fraction1:time			0.1897 .	0.1910 *
adjusted R^2	0.8535	0.8664	0.8408	0.8513
AIC	10.83	1.8	12.31	3.15
Significance codes	. p<0.1; * p<0.05; ** p<0.01; *** p<0.001			

We performed stepwise regression, and chose the following final regression model:

$$\sqrt{percent\_of\_0hr}$$

$$= 10.57 - 0.08 * fraction1 - 3.80 * \text{time}2.5 - 5.49 * \text{time}5 - 6.60$$

$$* time7.5 - 1.01 * batch2 - 0.69 * \text{batch}3 - 0.59 * \text{batch}4 + 2.17$$

$$* fration1{:}\,\text{time}2.5 + 2.24 * fraction1{:}\,\text{time}5 + 1.39$$

$$* \text{fraction}1{:}\,\text{time}7.5 \qquad (3)$$

The adjusted R^2 of the final model was 0.8664 (Table 5). ANOVA analysis indicated that the effect of fraction, time, and the interaction between fraction and time, but not mutation or batch, was significant at $\alpha = 0.05$ (Table 6).

Table 6. ANOVA analysis of full models 3 and 4. The dependent variable is "percent_of_0hr".

Model 3

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mutation	1	0.1003	0.1003	0.1047	0.7478	
fraction	1	31.2214	31.2214	32.5806	9.05E-07	***
time_fct	3	292.7586	97.5862	101.8344	8.00E-20	***
batch	3	7.1889	2.3963	2.5006	0.0718	.
mutation:fraction	1	0.0821	0.0821	0.0857	0.7711	
mutation:time_fct	3	0.3835	0.1278	0.1334	0.9397	
fraction:time_fct	3	12.1505	4.0502	4.2265	0.0104	*
Residuals	44	42.1645	0.9583			

Model 4

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mutation	1	0.1003	0.1003	0.0963	0.7577	
fraction	1	31.2214	31.2214	29.9651	1.50E-06	***
time	1	291.2411	291.2411	279.5214	6.95E-22	***
I(time^2)	1	6.0454	6.0454	5.8021	0.0198	*
batch	3	2.3676	0.7892	0.7574	0.5234	
mutation:fraction	1	0.0821	0.0821	0.0788	0.7801	
mutation:time	1	0.0061	0.0061	0.0058	0.9395	
fraction:time	1	3.9314	3.9314	3.7732	0.0578	.
Residuals	49	51.0545	1.0419			

Significance Codes: . p<0.1; * p<0.05; ** p<0.01; *** p<0.001

The interaction plot of fraction and time demonstrated that in fraction "1", the amounts of remaining protein at the 2.5 and 5 hour time points are higher than those in fraction "-1", although the protein level at the 0 hour time point in both fractions were the same (Figure 3). The positive sign of the interaction terms of fraction and time, together with the slopes of the interaction plots at the 2.5- and 5-hour time points implied that Cx43 protein degraded faster in the non-junctional fraction than the junctional membrane fraction.

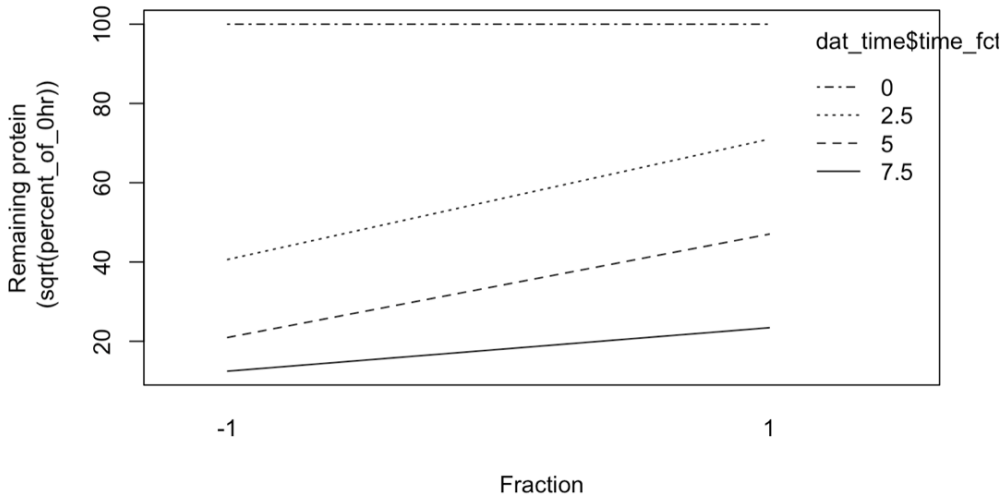


Figure 3. Interaction plot of fraction and time (model 3).

Because the values of the factor "time" is continuous and protein decay is expected to follow an exponential equation, we were interested in assessing the linear and quadratic effect of "time" on the response variable. We fit a second order, randomized block model to the data, treating "time" as numeric and using "time" and "time^2" to reflect the respective linear and quadratic components of the variable. The Boxcox plot (Figure S4) suggested a square-root transformation of the dependent variable. Estimated coefficients of the full model (model 4) are listed in Table

5. Residual plots and QQ plots are shown in Figure S4. The final fitted model after stepwise selection was:

$$\sqrt{percent\_of\_0hr} = 9.57 + 0.81 * \text{fraction1} - 1.28 * \text{time} + 0.05 * (time^2) + 0.19 * \text{fraction1:time} \quad (4)$$

The adjusted R^2 of the reduced model was 0.8513 (Table 5). Both the regression model and ANOVA analysis revealed that the linear and quadratic effects of time were significant at α = 0.05 (Tables 5 and 6). This is consistent with the exponential regression of the "percent_of_0hr" variable over time in models 1 and 2. The ANOVA result also indicated the significance of the effect of fraction on the dependent variable (Table 6), while the regression model identified the significant interaction term between fraction1 and time at α = 0.05 (final model 4, Table 5).

The response variable in model 3 and 4 is "percent_of_0hr", whereas the response variable in models 1 and 2 is the exponential time constant, which is calculated by fitting the values of the "percent_of_0hr" variable at all the time points. Therefore, models 3 and 4 relax the grouping of the "percent_of_0hr" data of a certain mutation in a certain fraction, which might increase the variance of the dependent variable. However, these models gain more degrees of freedom. The residual degrees of freedom of the full models 2 and 3 is 8 and 44, respectively. The overall fit of model 3 to the data is improved compared to model 2, with the respective adjusted R^2 of the final model 2 and model 3 at 0.6492 and 0.8664 (Tables 2 and 5).

3.3 Ratio of labeled to total protein as the response variable

The variable "ratio" is directly proportional to the amount of labeled protein in a specific subcellular fraction, rather than the percentage of the remaining labeled protein to the initial amount. We used "ratio" as the response variable and fit a linear model to examine the effect of mutation, fraction, time, and batch on protein stability. We treated "time" as a categorial variable with four levels as described for model 3. Estimated coefficients of the full model, with logarithmic transformation of the response variable, are shown in Table 7 (Model 5). The Boxcox plot, residual plots and QQ plots of the model are shown in Figure S5. We performed stepwise regression and selected the final model as:

$$log(ratio) = -0.85 + 0.21 * mutation1 + 0.15 * fraction1 - 1.01 * time2.5 - 1.62 * time5 - 2.24 * time7.5 - 2.20 * batch2 - 1.36 * batch3 + 0.07 * batch4 - 0.30 * mutation1:fraction1 + 0.64 * fraction1:time2.5 + 0.81 * fraction1:time5 + 0.74 * fraction1:time7.5 \quad (5)$$

The adjusted R^2 of the final model is 0.9307 (Table 7), which was higher than the models using other response variables. The ANOVA analysis indicated that fraction, time, batch, and the interaction term between fraction and time were significantly associated with the response variable at $\alpha = 0.05$ (Table 8). The interaction plot of fraction and time also revealed that the amounts of labeled protein at the 2.5, 5, and 7.5 hour time points were higher in fraction "1" than fraction "-1", whereas the amount of labeled protein at the 0 hour time point in the two subcellular fractions were similar (Figure 4). The interaction plots, as well as the positive sign of the coefficients of the interaction terms between fraction and time in the regression model,

suggested that the turnover rate of the Cx43 protein was faster in the non-junctional fraction than in the junctional fraction.

Table 7. Estimated parameters of models 5 and 6. The dependent variable is "ratio".

	Model 5		Model 6	
	full	final	full	final
Intercept	-0.8542 ***	-0.8539 ***	-1.0982 ***	-1.1275 ***
mutation1	0.2120	0.2112 .	0.1773	0.2112 .
fraction1	0.1481	0.1480	0.3266 .	0.3277 *
time_fct2.5	-1.0023 ***	-1.0098 ***		
time_fct5	-1.6448 ***	-1.6218 ***		
time_fct7.5	-2.2273 ***	-2.2443 ***		
time			-0.3257 ***	-0.3087 ***
I(time^2)			0.0017	
batch2	-2.2042 ***	-2.2042 ***	-2.0016 ***	-1.9986 ***
batch3	-1.3607 ***	-1.3607 ***	-1.158 ***	-1.1550 ***
batch4	0.0688	0.0688	0.2714 *	0.2744 *
mutation1:fraction1	-0.3047 .	-0.3047 .	-0.3189 .	-0.3176 .
mutation1:time_fct2.5	-0.0150			
mutation1:time_fct5	0.0460			
mutation1:time_fct7.5	-0.0340			
mutation1:time			0.0095	
fraction1:time_fct2.5	0.6407 *	0.6417 *		
fraction1:time_fct5	0.8117 **	0.8085 **		
fraction1:time_fct7.5	0.7409 **	0.7433 **		
fraction1:time			0.1038 **	0.1035 **
adjusted R^2	0.9262	0.9307	0.9302	0.9327
AIC	-114.76	-120.61	-121.61	-125.43
Significance codes	. p<0.1; * p<0.05; ** p<0.01; *** p<0.001			

Table 8. ANOVA analysis of full models 5 and 6. The response variable is "ratio".

Model 5						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mutation	1	0.1318	0.1318	1.1154	0.2967	
fraction	1	2.6370	2.6370	22.3189	2.38E-05	***
time_fct	3	29.1537	9.7179	82.2506	4.52E-18	***
batch	3	55.4464	18.4821	156.4296	1.74E-23	***
mutation:fraction	1	0.3342	0.3342	2.8287	0.0997	.
mutation:time_fct	3	0.0187	0.0062	0.0529	0.9838	
fraction:time_fct	3	1.5439	0.5146	4.3557	0.0090	**
Residuals	44	5.1986	0.1181			
Model 6						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mutation	1	0.1318	0.1318	1.1786	0.2829	
fraction	1	2.6370	2.6370	23.5853	1.26E-05	***
time	1	33.2987	33.2987	297.8260	1.83E-22	***
I(time^2)	1	0.1264	0.1264	1.1308	0.2928	
batch	3	51.2788	17.0929	152.8803	7.18E-25	***
mutation:fraction	1	0.3342	0.3342	2.9892	0.0901	.
mutation:time	1	0.0017	0.0017	0.0155	0.9014	
fraction:time	1	1.1771	1.1771	10.5281	0.0021	**
Residuals	49	5.4785	0.1118			

Significance Codes: . p<0.1; * p<0.05; ** p<0.01; *** p<0.001

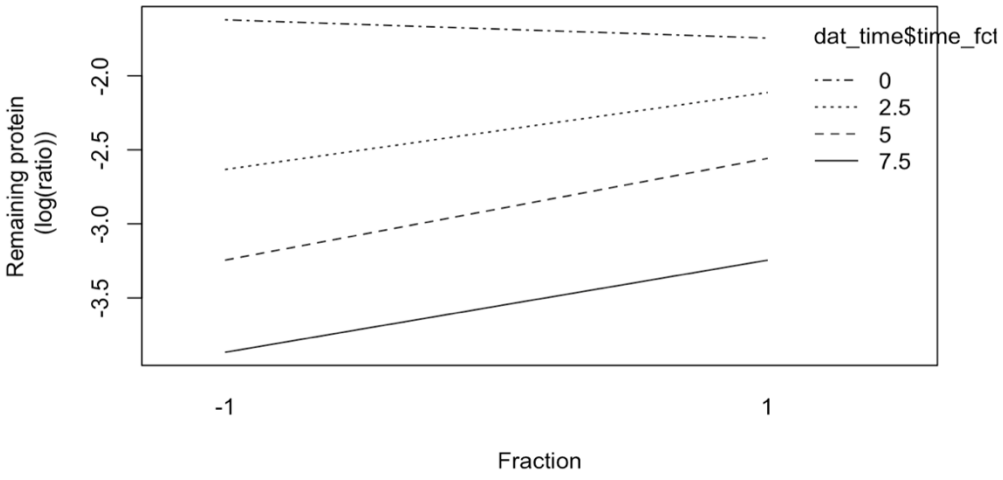


Figure 4. Interaction plot of fraction and time (model 5).

We also fit a second-order model, with "time" as numeric and "ratio" as the response variable, to assess the linear and quadratic effect of time. The linear and quadratic components of "time" was assessed with terms "time" and "time^2", respectively. The estimated parameters of the full model, with the response variable logarithmically transformed, are listed in Table 7 (Model 6). The Boxcox plot, residual plots and QQ plots of the model are shown in Figure S6. The final fitted model based on stepwise regression was:

$$log(ratio) = -1.13 + 0.21 * Mutation1 + 0.33 * fraction1 - 0.31 * time - 2.00 * batch2 - 1.16 * batch3 + 0.27 * batch4 - 0.32 * mutation1{:}fraction1 + 0.10 * fraction1{:}time \quad (6)$$

The adjusted R^2 of the reduced model was 0.9327 (Table 7). The regression and ANOVA results revealed that fraction, time, batch, and the interaction between fraction and time were significant factors associated with the dependent variable at $\alpha = 0.05$ (Tables 7 and 8). Interestingly, the quadratic effect of time was not significant even at $\alpha = 0.1$.

The blocking effect of "batch" is significant in models 5 and 6 at $\alpha = 0.05$, unlike in models 1, 2, and 4. This can be explained by the high variance in the response variable, "ratio", among batches in models 5 and 6. On the other hand, the response variable in models 1 to 4 is either the exponential time constant or the variable "percent_of_0hr". The calculation of these response variables involved dividing of the amount of protein at later time points by that at the 0-hour time point (Figure 2). Thus, the transformation of "ratio" to time constant or "percent_of_0hr" might have reduced the variance in the response variable among batches, resulting in similar effect to including "batch" as a blocking variable in a regression model.

3.4 MANOVA analyses

To investigate how the amount of protein at a specific time point was affected by or associated with mutation and fraction, we grouped variables "percent_of_0hr" and "ratio" by time points and used them as the dependent variables in MANOVA tests. In a model assessing the effect of factors on "percent_of_0hr", there were three dependent variables, corresponding to the percentage of protein at the 2.5, 5, and 7.5-hour time points. The 0-hour time point was excluded because the value of "percent_of_0hr" was always 100%. The response variables were logarithmically transformed so that a normal distribution was satisfied (Table 9). A linear model of "mutation" and "fraction", with "batch" as a blocking variable, was fit to the data. The MANOVA analysis indicated that "fraction", but not "mutation" or "batch", was a significant factor at $\alpha = 0.05$ (Table 10). Multiple univariate ANOVA further revealed that "fraction" was significant to all three response variables, suggesting that any of the 3 timepoints could be used to study the effect of fraction on the remaining amount of protein expressed as "percent_of_0hr" (Table 10).

Table 9. Shapiro-Wilk normality test of the response variables in MANOVA analyses.

	original		log-transformed	
	W	p-value	W	p_value
percent_of_0hr_2.5	0.9332	0.3044	0.8954	0.0810
percent_of_0hr_5	0.8395	0.0124	0.9680	0.8272
percent_of_0hr_7.5	0.9252	0.2309	0.8952	0.0804
ratio_0	0.8657	0.0292	0.8888	0.0644
ratio_2.5	0.8772	0.0430	0.9220	0.2063
ratio_5	0.8597	0.0239	0.9241	0.2221
ratio_7.5	0.8461	0.0153	0.9528	0.5689

Table 10. MANOVA and multiple univariate ANOVA analyses with "percent_of_0hr" as the response variables. Responses 1, 2, and 3 represents "percent_of_0hr" at 2.5-, 5-, and 7.5-hour time points, respectively.

```
                 Df  Pillai approx F num Df den Df  Pr(>F)
mutation          1 0.01570   0.0319      3      6 0.99155
fraction          1 0.73310   5.4934      3      6 0.03718 *
batch             3 0.76857   0.9185      9     24 0.52620
mutation:fraction 1 0.23651   0.6196      3      6 0.62754
Residuals         8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
 Response 1 :
                 Df  Sum Sq Mean Sq F value  Pr(>F)
mutation          1 0.01374 0.01374  0.0970 0.76344
fraction          1 1.52463 1.52463 10.7595 0.01118 *
batch             3 0.87110 0.29037  2.0492 0.18558
mutation:fraction 1 0.05860 0.05860  0.4136 0.53816
Residuals         8 1.13361 0.14170
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 2 :
                 Df  Sum Sq Mean Sq F value  Pr(>F)
mutation          1 0.00054 0.00054  0.0029 0.95858
fraction          1 2.44888 2.44888 13.1281 0.00675 **
batch             3 0.21101 0.07034  0.3771 0.77220
mutation:fraction 1 0.03399 0.03399  0.1822 0.68073
Residuals         8 1.49229 0.18654
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 3 :
                 Df  Sum Sq Mean Sq F value  Pr(>F)
mutation          1 0.02836 0.02836  0.0847 0.77845
fraction          1 2.03876 2.03876  6.0886 0.03886 *
batch             3 0.86421 0.28807  0.8603 0.49988
mutation:fraction 1 0.00001 0.00001  0.0000 0.99563
Residuals         8 2.67877 0.33485
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In another MANOVA test, the response variables were "ratio" grouped by 0, 2.5, 5, and 7.5-hour time points. We also transformed the response variables logarithmically to achieve normal distribution of the data (Table 9). A linear model of "mutation", "fraction", and the blocking variable "batch" was fit to the data. MANOVA result showed that none of the factors or their interaction was significant at $\alpha = 0.05$ (Table 11). However, multiple univariate ANOVA revealed that "fraction" was significantly associated with protein level at 2.5- and 5-hour time

points at $\alpha = 0.05$ (Table 11). Therefore, we could use either of the two time points to assess the effect of "fraction" on the amount of remaining protein expressed as "ratio". The effect of "batch" was significant to all the dependent variable at $\alpha = 0.05$, implying high variance among batches at all the time points. Taken together, 2.5 and 5 hours would be reasonable time points to study the decay of Cx43 protein in different subcellular fractions.

Table 11. MANOVA and multiple univariate ANOVA analyses with "ratio" as the response variables. Responses 1, 2, 3, and 4 represents "ratio" at 0-, 2.5-, 5-, and 7.5-hour time points, respectively.

```
                 Df  Pillai approx F num Df den Df  Pr(>F)
mutation          1 0.26449   0.4495      4      5 0.77072
fraction          1 0.75891   3.9349      4      5 0.08268 .
batch             3 1.29054   1.3212     12     21 0.27804
mutation:fraction 1 0.74055   3.5679      4      5 0.09777 .
Residuals         8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
 Response 1 :
                 Df  Sum Sq Mean Sq  F value    Pr(>F)
mutation          1  0.0109  0.0109   0.3075    0.5944
fraction          1  0.0599  0.0599   1.6915    0.2296
batch             3 11.0702  3.6901 104.1375 9.468e-07 ***
mutation:fraction 1  0.0758  0.0758   2.1388    0.1818
Residuals         8  0.2835  0.0354
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 2 :
                 Df  Sum Sq Mean Sq  F value    Pr(>F)
mutation          1  0.0491  0.0491   0.8609   0.38062
fraction          1  0.9799  0.9799  17.1644   0.00324 **
batch             3 17.7043  5.9014 103.3760 9.742e-07 ***
mutation:fraction 1  0.2678  0.2678   4.6907   0.06222 .
Residuals         8  0.4567  0.0571
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 3 :
                 Df  Sum Sq Mean Sq F value    Pr(>F)
mutation          1  0.0163  0.0163  0.0998 0.7601589
fraction          1  1.7413  1.7413 10.6801 0.0113895 *
batch             3 13.7279  4.5760 28.0654 0.0001346 ***
mutation:fraction 1  0.0083  0.0083  0.0509 0.8271291
Residuals         8  1.3044  0.1630
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 4 :
                 Df  Sum Sq Mean Sq F value    Pr(>F)
mutation          1  0.0742  0.0742  0.2635 0.6216230
fraction          1  1.3997  1.3997  4.9701 0.0563521 .
batch             3 13.7533  4.5844 16.2784 0.0009099 ***
mutation:fraction 1  0.0742  0.0742  0.2633 0.6217015
Residuals         8  2.2530  0.2816
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CHAPTER 4

Summary

We applied multiple statistical models to fit the data of Cx43 protein degradation to examine the effect of subcellular fraction and mutation on protein stability. The dependent variables in our models included the exponential time constant (models 1, 2), the percentage of remaining protein to the initial amount (models 3, 4), and the ratio of the amount of labeled protein to that of the total protein (models 5, 6). The analyses consistently indicated that fraction, but not mutation, had a significant effect on Cx43 protein degradation, with the protein being more stable in the junctional membrane fraction than in the non-junctional fraction (Tables 2-3, 5-8). The overall fit of the final models 5 and 6 was the best among the models tested, with the adjusted R^2 around 0.93 (Table 7). On the other hand, the adjusted R^2 of the final models 1 and 2 was 0.58 and 0.65, respectively (Table 2), which is probably due to the reduced degrees of freedom in the two models compared to the others. The difference between models 1 and 2 is how exponential regression was conducted to deduce the time constant of the Cx43 protein. In model 1 linear regression models were fit to logarithmically transformed response variable, whereas a nls (non-linear least square) method was used in model 2. MANOVA analyses with the response variables being the amount of proteins at various time points revealed that the effect of mutation and fraction on the degradation of Cx43 protein was evident at the 2.5 and 5-hour time points (Tables 10 and 11). For similar experiments in the future, extending the chase period to 7.5 hours might not be necessary, which could help shorten the duration of experiments.

The finding that the subcellular location of Cx43 affects protein stability, with the protein being more stable in the junctional fraction than the non-junctional fraction, implies that it is better to keep the protein on the membrane to extend its life. Therefore, to correct for the mis-localization of Cx43 and its overall reduced expression in arrhythmic failing hearts, we should consider increasing protein production, together with rapid delivery of the protein to the junctional fraction to minimize protein degradation.

APPENDIX

Supplemental figures

A B C

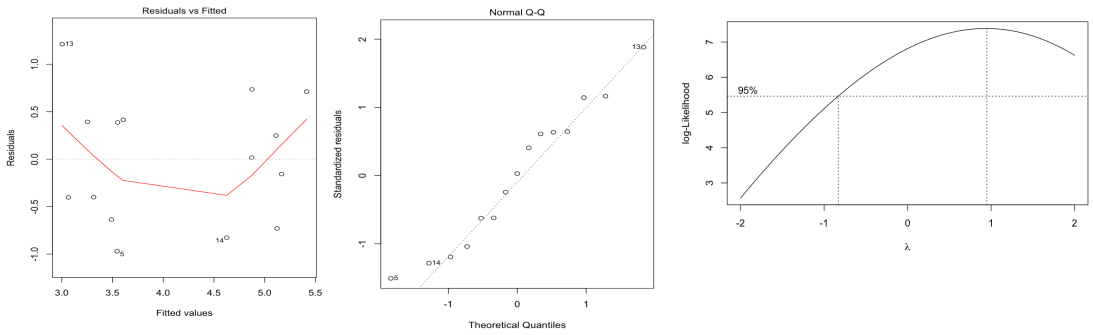


Figure S1. Residual plot (A), QQ plot(B), and the Boxcox plot (C) of model 1. $\lambda_{max} = 0.950$, suggesting that it is not necessary to transform the response variable, τ.

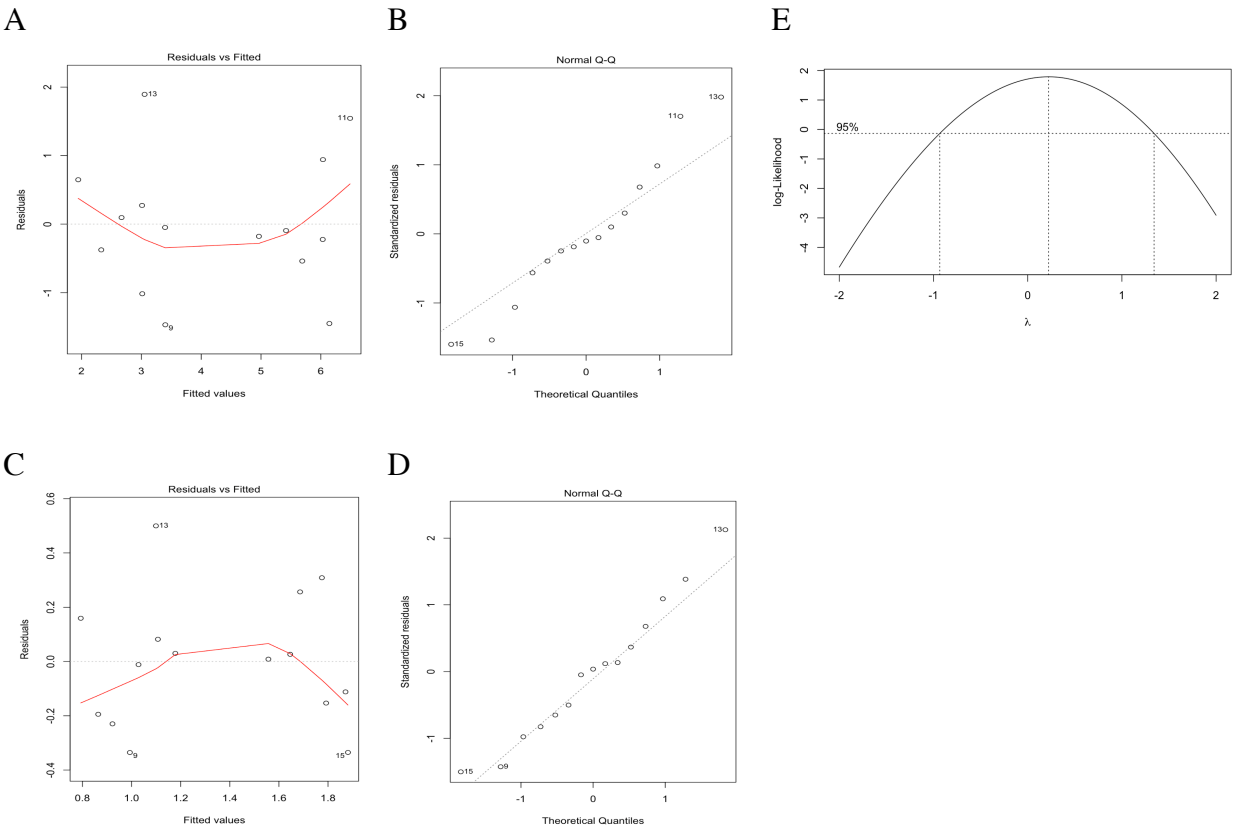


Figure S2. Residual plot of model 2 before (A) and after transforming τ (C), QQ plot before (B) and after transforming τ (D), and the Boxcox plot (E). $\lambda_{max} = 0.222$, suggesting a logarithmic transformation of τ.

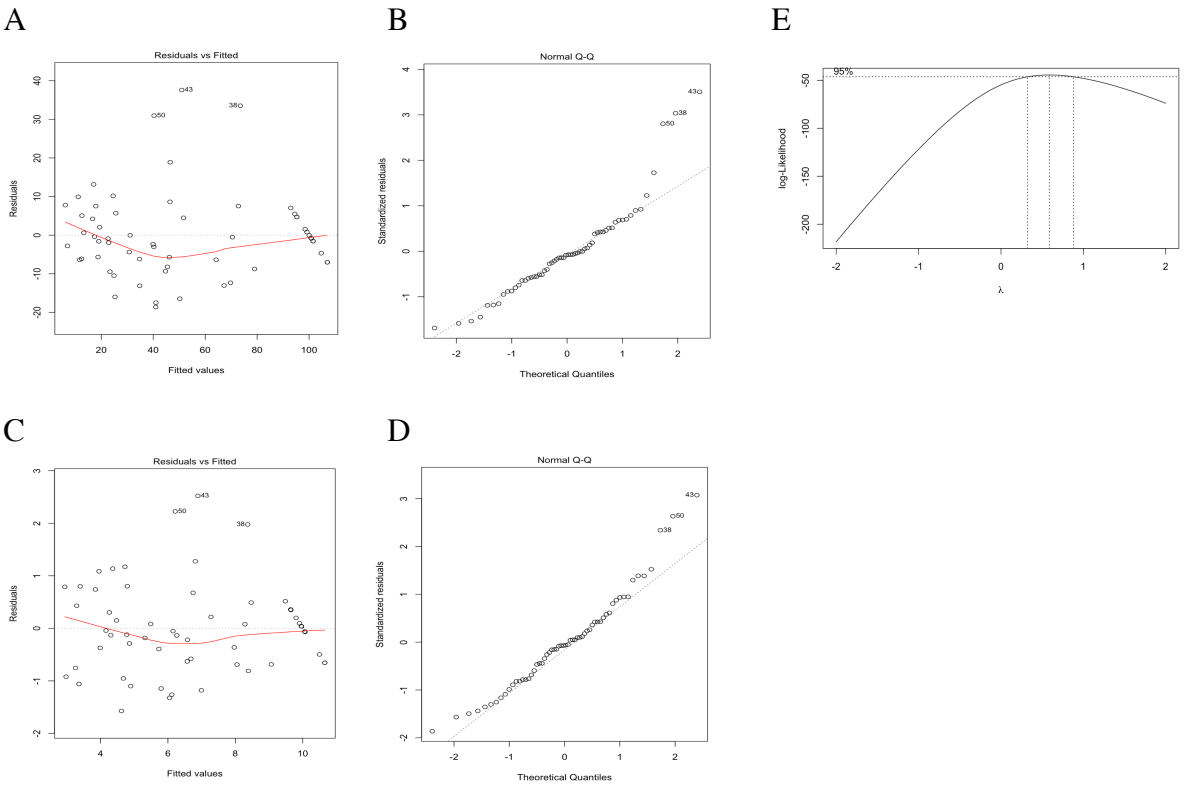


Figure S3. Residual plot of model 3 before (A) and after (C) transforming the response variable, QQ plot before (B) and after (D) transforming the response variable, and the Boxcox plot of model 3 (E). $\lambda_{max} = 0.586$, suggesting a square-root transformation of the response variable.

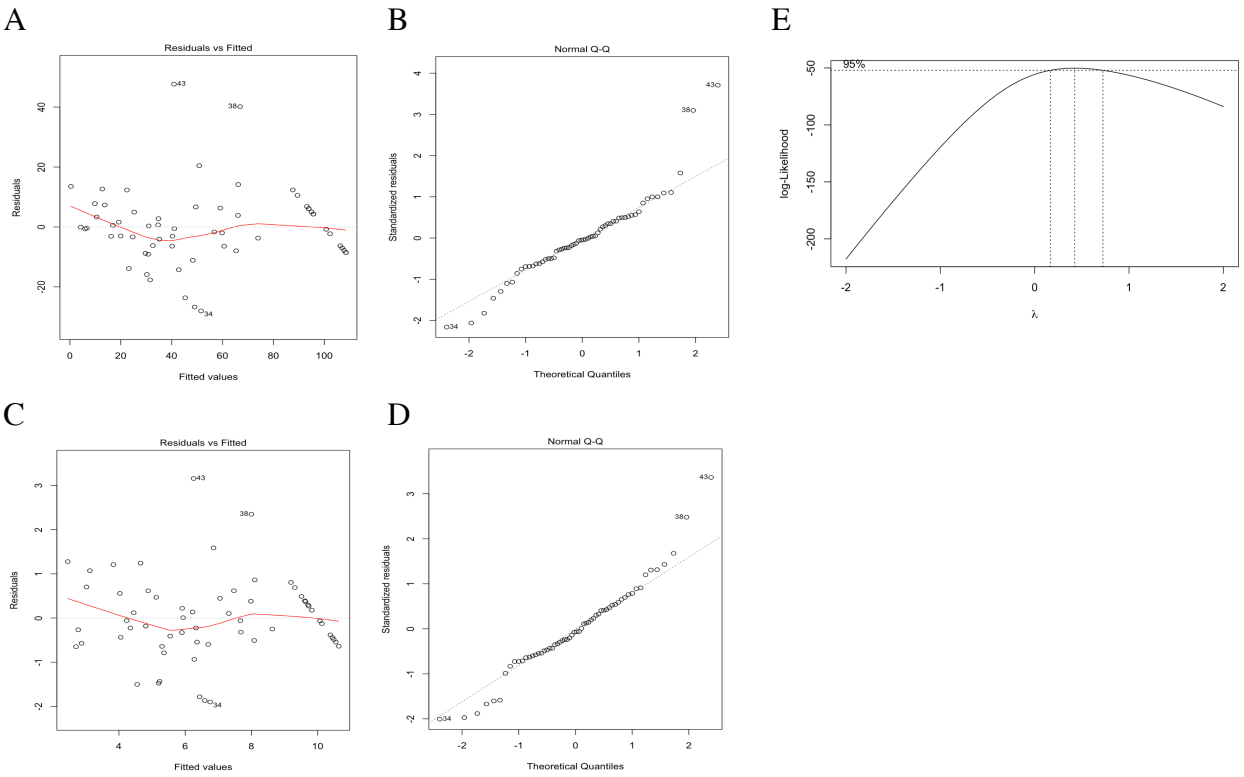


Figure S4. Residual plot of model 4 before (A) and after (C) transforming the response variable, QQ plot before (B) and after (D) transforming the response variable, and the Boxcox plot of model 4 (E). $\lambda_{max} = 0.424$, suggesting a square-root transformation the response variable.

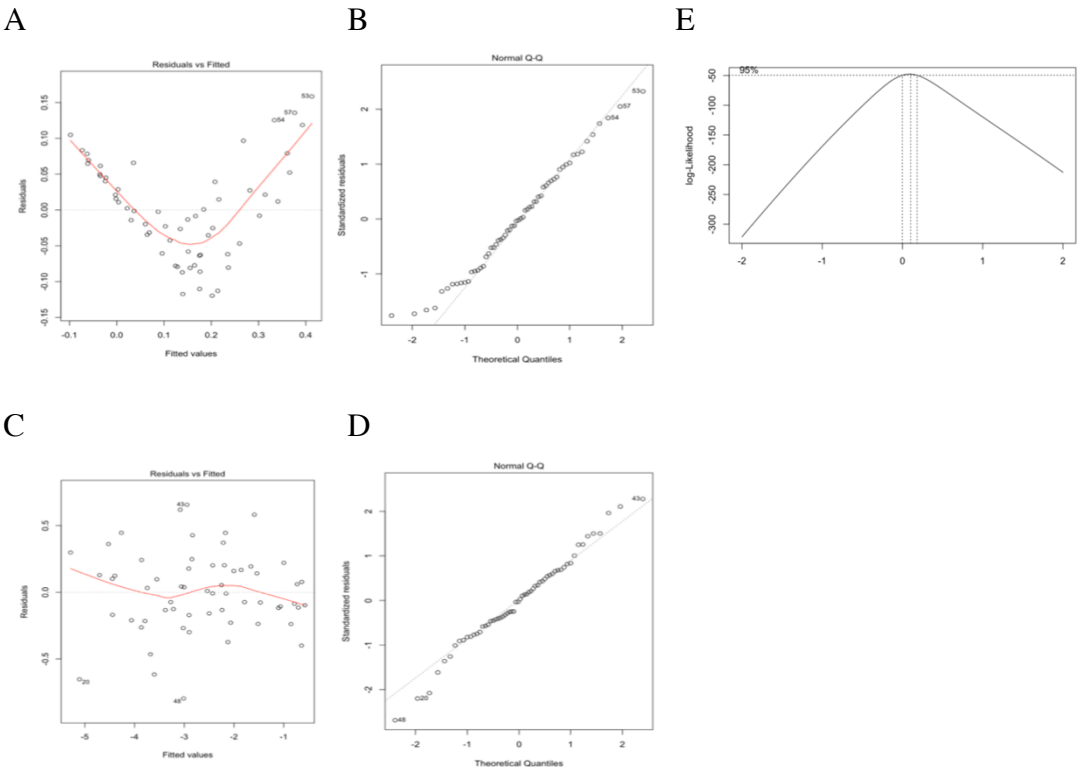


Figure S5. Residual plot of model 5 before (A) and after (C) transforming the response variable, QQ plot before (B) and after (D) transforming the response variable, and the Boxcox plot of the model (E). $\lambda_{max} = 0.101$, suggesting a logarithmic transformation the response variable.

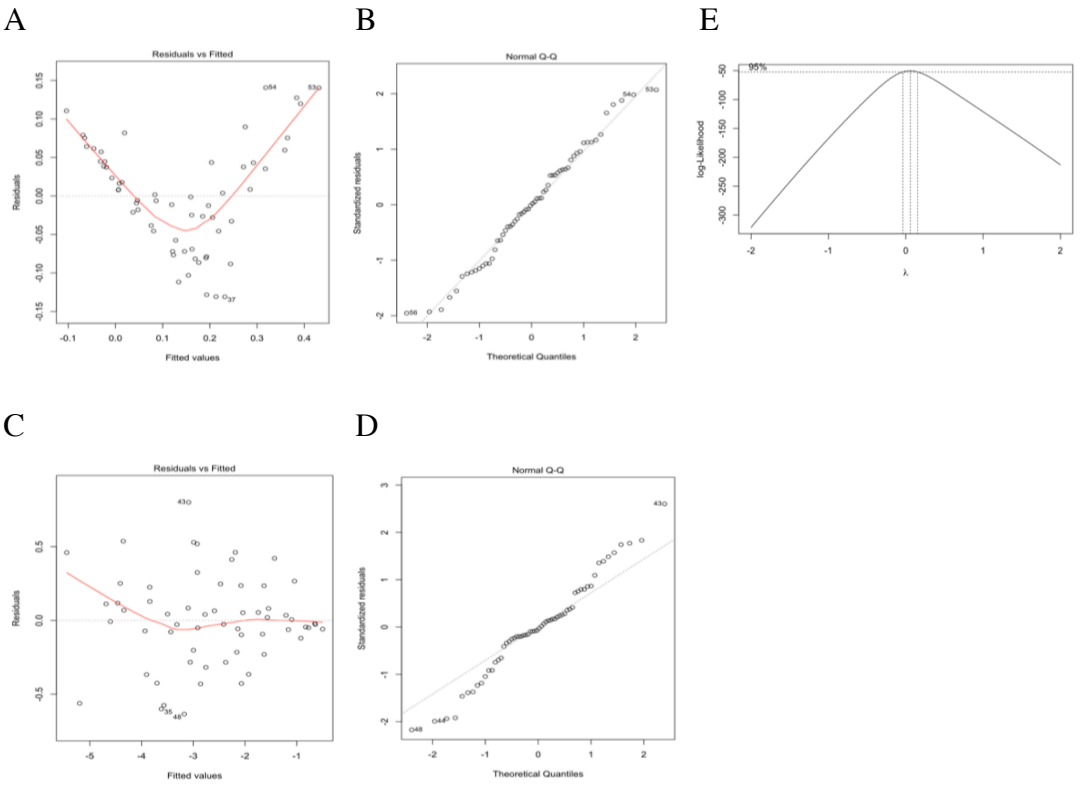


Figure S6. Residual plot of model 6 before (A) and after (C) transforming the response variable, QQ plot before (B) and after (D) transforming the response variable, and the Boxcox plot of the model (E). $\lambda_{max} = 0.061$, suggesting a logarithmic transformation the response variable.

REFERENCES

[1] Zhou P, Pu WT. Recounting Cardiac Cellular Composition. *Circ Res*. 2016;118(3):368-370. doi:10.1161/CIRCRESAHA.116.308139

[2] Xiao S, Shimura D, Baum R, et al. Auxiliary trafficking subunit GJA1-20k protects connexin-43 from degradation and limits ventricular arrhythmias. *J Clin Invest*. 2020;134682. doi:10.1172/JCI134682

[3] Fontes MS, van Veen TA, de Bakker JM, van Rijen HV. Functional consequences of abnormal Cx43 expression in the heart. *Biochim Biophys Acta*. 2012;1818(8):2020-2029. doi:10.1016/j.bbamem.2011.07.039

[4] Smyth JW, Shaw RM. Autoregulation of connexin43 gap junction formation by internally translated isoforms. *Cell Rep*. 2013;5(3):611-618. doi:10.1016/j.celrep.2013.10.009

[5] Montgomery DC. Design and Analysis of Experiments. 9th ed. Hoboken, NJ: John Wiley & Sons, Inc; 2017. 734 p.

[6] Wu CFJ, Hamada MS. Experiments: Planning, Analysis and Optimization. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2009. 760 p.

[7] Sheather SJ. A Modern Approach to Regression with R. New York: Springer Science+Business Media, LLC; 2009. 393 p.

[8] Faraway JJ. Linear Models with R. Boca Raton: Chapman & Hall/CRC; 2005. 240 p.

[9] Rencher AC. Methods of Multivariate Analysis. 2nd ed. New York; John Wiley & Sons, Inc; 2002. 738 p.