Data Analysis of Salmonid Environmental DNA Measurements Obtained via Controlled Experiments and from several Pacific Streams

by

Robert Sneiderman
B.Sc., McGill University, 2018
A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Data Analysis of Salmonid Environmental DNA Measurements Obtained via Controlled Experiments and from several Pacific Streams

by

Robert Sneiderman
B.Sc., McGill University, 2018

Supervisory Committee

Dr. M Lesperance , Supervisor
(Department of Mathematics and Statistics)

Dr. M Hockings , Co-Supervisor
(EcoFish Research Ltd and the School of Enviromental Studies)

Dr. L Cowen , Member One
(Department of Mathematics and Statistics)

Dr. C Helbing , Outside Member
(Department of Biochemistry and Microbiology)

Supervisory Committee

Dr. M Lesperance , Supervisor
(Department of Mathematics and Statistics)

Dr. M Hockings , Co-Supervisor
(EcoFish Research Ltd and the School of Enviromental Studies)

Dr. L Cowen , Member One
(Department of Mathematics and Statistics)

Dr. C Helbing , Outside Member
(Department of Biochemistry and Microbiology)

ABSTRACT

Standard sampling and monitoring of fish populations are invasive and time-consuming techniques. The ongoing development of statistical techniques to analyze eDNA introduces a possible solution to these challenges. We analyzed and created statistical models for qPCR data obtained from two controlled experiments conducted on samples of Coho salmon.

The first experiment analyzed was a density experiment whereby varying numbers of Coho (1, 2, 4, 8, 16, 32 and 65 fish) were placed in separate tanks and eDNA measurements were taken. The second experiment dealt with dilution, whereby three Coho were placed into tanks, removed and eDNA was then sampled at dilution volumes of 20kL, 40kL, 80kL, 160kL and 1000kL.

Finally, we analyzed a set of field data from several streams in the Pacific North West for the presence of Coho salmon. In the field models, we considered the impact of environmental covariates as well as eDNA concentrations.

Our analysis suggests that eDNA concentration can be used as a reliable proxy to estimate Coho biomass.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank my advisor, Mary Lesperance for all her support and assistance. I would also like to thank the team at EcoFish Research, especially Morgan Hocking and Jeff MacAdams. As well, thank you to the team at Caren Helbing's lab, including Michael Allison and Lauren Bergman. Finally, I would like to acknowledge and thank my mother as I would not have had the opportunity to further my education without her support.

Chapter 1

Introduction to eDNA

The rapid and catastrophic decline in Earth's biodiversity is a clear and major obstacle facing humanity in the 21st century. Although this fact is widely accepted among the scientific community, the solution on how to combat this decline is still not known. One thing that is clear, however, is that researchers must expand and improve knowledge on the current state and distribution of biodiversity worldwide. In order to make informed and critical decisions regarding biodiversity, researchers must first monitor and model reliable distribution patterns, as well as estimate population sizes. Historically, this monitoring has involved invasive surveys that may even compound the issue of decline, especially when studying endangered or rare species. For this reason, it is critical that non-invasive, large scale biodiversity monitoring techniques are developed (Thomsen and Willerslev, 2015).

One such non-invasive technique for the monitoring of biodiversity involves the study and collection of environmental DNA. Environmental DNA, or eDNA, is mitochondrial or nuclear DNA that is released from an organism as it interacts with its environment. Common sources of eDNA include shed skin cells, hair, blood, urine and mucous (Coble et al., 2018). Shed DNA can be collected by researchers for analysis. There are numerous methods in which eDNA is collected and analyzed, such as from samples of lake water (Nevers et al., 2018) or other aquatic sources (Penarrubia et al., 2016). eDNA has also been obtained from non-aquatic environments such as from snow, ice, or from soil (Yoccoz et al., 2012).

New and emerging technology has advanced researchers' ability to collect and analyze this shed eDNA. In particular, studies of eDNA have assisted scientists in detecting the extant of a species in a certain area and have facilitated conservation efforts (Ficetola et al., 2008; Rees et al., 2015). Most commonly, environmental DNA is analyzed via DNA sequencing methods such as metagenomics and qPCR (Livak and Schmittgen, 2002).

The analysis of eDNA allows researchers to study species without the capture of the target organism. This is of benefit in particular for species that are endangered and allows researchers to conduct investigations with minimal environmental disruption. Before the common usage of eDNA, methods for studying aquatic diversity included fishing and trapping. These invasive methods are expensive, time consuming and directly impact the habitat in which the researcher is studying. Indeed, using classical sampling methods to study fish may negatively impact the species and environment (especially electrofishing) (Snyder, 2003; Strayer, 2010). Collection and study of eDNA however, is non-invasive and only requires minimal sampling.(Tsuji et al., 2017; Wilcox et al., 2016). Analysis of environmental DNA has already been used successfully as a surveillance method for rare fish species (Jerde et al., 2011).

The ability to sequence miniscule concentrations of eDNA directly from aquatic environments is revolutionizing the field of ecological monitoring and management (Collins et al., 2018). Although most studies regarding eDNA in aquatic environments have focused on freshwater, eDNA is increasingly being used to study ocean and marine systems. Corporations involved with aquatic conservation and fisheries management are progressively incorporating eDNA into their business and research (Cristescu and Hebert, 2018). Hence, it is critical that researchers understand the physical and chemical properties of eDNA, such as degradation or the impact of ecological covariates.

One species that researchers have studied using eDNA methodology is the invasive American signal crayfish, *Pacifastacus leniusculuscray*. This crayfish is problematic and is the current leading cause of decline among UK's native crayfish species. The practice of eDNA related techniques allows for non-invasive and potentially early detection of these crayfish *Pacifastacus leniusculuscray* and may allow for eradication before the levels were large enough to detect manually and before it is too late to act (Dunn et al., 2017).

Researchers have also used eDNA to study the Coastal tailed frog (*Ascaphus truei*). By using eDNA and testing water samples at a variety of lakes and watersheds, researchers were able to expand knowledge of known Coastal distributions in the region (Coble et al., 2018).

The association between eDNA concentration and species abundance is the subject of ongoing study. Research of native fish populations have shown that studying eDNA concentration in water samples may be able to predict fish abundance as well as, or better than more invasive and costly methods (Lacoursière-Roussel et al., 2016; Tillotson et al., 2018). Moreover, an ever-increasing number of non-invasive eDNA techniques produce reliable enough DNA to address nearly all questions that could be obtained via traditional methods (Beja-Pereira et al., 2009a).

In Florida, researchers collected eDNA to study the abundance and occupancy of invasive Burmese Pythons, *Python molurus bivitatus*, in the Everglades (Hunter et al., 2015). These pythons pose significant threats to native species. Burmese Pythons are considered to be largely mysterious and their population distribution in the Florida area is not well documented. These pythons mainly reside within inaccessible habitats, thus making standard sampling nearly impossible. Because they are so evasive, only a single study with the goal of estimating detection probability has been conducted in which standard sampling techniques were attempted (Reed et al., 2011). This standard study did not have very promising results and had very few sightings of the pythons. However, researchers using eDNA sampling were able to extensively expand knowledge of the python distribution via studies done on water samples taken throughout the everglades. The non-invasive study of eDNA allowed scientists to gain a better understanding of *Python molurus bivatatus* distribution in the area.

Overall, the study and analysis of environmental DNA is an emerging method of science that has been used successfully for the detection and study of rare species. The less invasive methods utilized while studying eDNA are becoming increasingly important as biodiversity continues to decline worldwide (Bergman et al., 2016).

In Chapter 3 of this thesis, we use water samples from tanks in which the biomass of Coho is known to study the relationship between eDNA concentration and biomass. In Chapter 4, we extend this result by considering the impact of flow or dilution

on eDNA. In Chapter 5, we bridge the gap to the field, where we build models for predicting Coho biomass using eDNA concentrations and several environmental covariates.

Chapter 2

Statistical Methods

2.0.1 Summary of qPCR

Real time PCR or Quantitative polymerase chain reaction ('qPCR') is a well-established laboratory technique that allows for the real time detection and quantification of microbial DNA (Kralik and Ricchi, 2017).

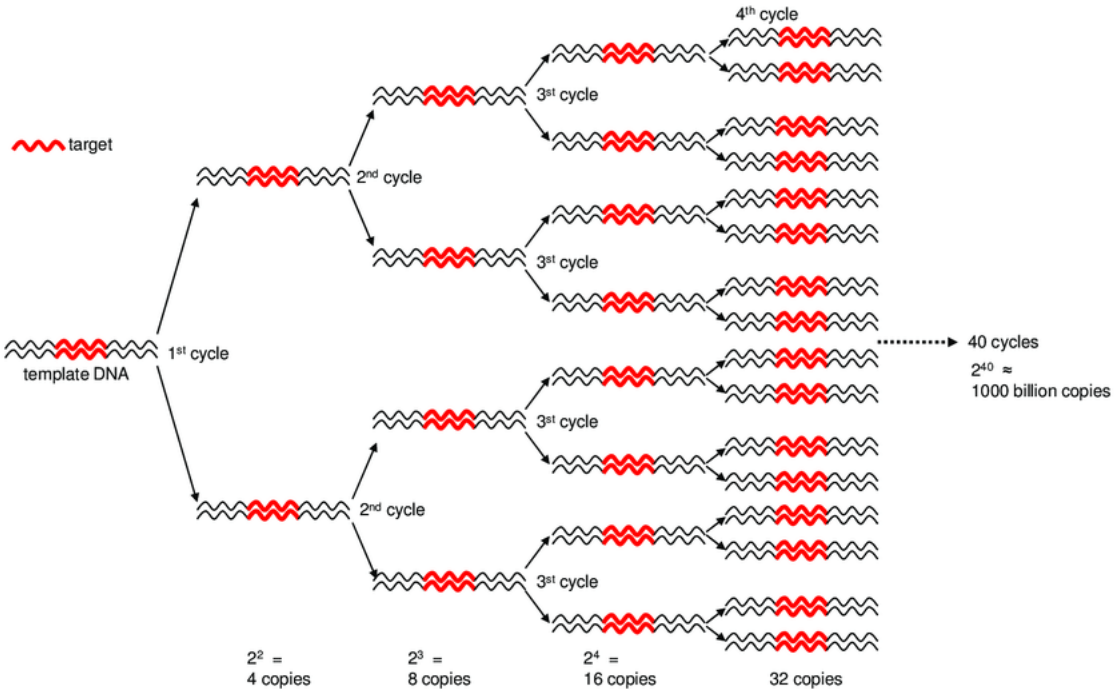

Figure 2.1: Assuming that the qPCR runs with perfect efficiency, we expect to see a doubling of DNA after every cycle (Rodriguez-Lazaro and Hernández, 2013).

Figure 2.1 is a basic visualization of how qPCR works. At each cycle, assuming perfect efficiency, the amount of target DNA should double.

To measure concentration of DNA, a fluorometer detects levels of fluorescence as the thermal cycler runs, and the level of fluorescent signal reflects the amount of target DNA in the sample. During the first cycles, the amount of fluorescence produced is not significant enough to separate from background light. As the experiment progresses, the fluorescent signal may increase above a certain detectable level (corresponding to the initial number of template DNA). This point is known as the 'Quantification Cycle' or the Cycle Threshold (CT). This value allows for the detection and quantification of target DNA (Schmittgen and Livak, 2008). Typically, a specified number of cycles is decided on as a 'cutoff' value. That is, if we have gone through a specified number of cycles, and still have not detected a signal, we would consider this to be indicative that the target DNA was not present in the sample.

In other words, as the experiment progresses, fluorescence is observed when the specified target DNA is detected. If it only takes a few cycles, this results in a low Cycle Threshold ('CT') score and indicates high levels of target DNA were present in the sample. The more cycles it takes to detect a visible fluorescence signal, the higher the CT score, indicating that less of the target DNA is present in the sample (Hindson and Ness, 2011; Willems et al., 2008).

Another concept that is often used when discussing eDNA related studies is the 'Limit of Detection' or 'LoD'. With respect to eDNA studies, the 'LoD' is defined as the minimum amount of target DNA required to conclude with high probability that specified target DNA exists in the sample. Researchers often agree upon a CT value to determine a LoD. For example, in our later research we define the LoD to be CT < 50. Hence, if there is still no visible fluorescence after 50 cycles (i.e. CT >50), we conclude that the sample does not contain the target DNA. In other cases, LoD may be defined in terms of the number of copies of target DNA.

2.0.2 Biomass and eDNA concentration

The study by Takahara et al. (2012) deals directly with estimation of fish biomass from eDNA concentration and helps to illustrate how statistical analysis is performed on eDNA data.

In this study, researchers hypothesized that eDNA release from the common carp, *Cyprinus carpio* is positively correlated with the carp's biomass. Furthermore, researchers extended their knowledge of eDNA to study carp in the field, where they incorporated several environmental covariates into their statistical models.

To study collected samples of carp eDNA, researchers used a carp specific mitochondrial gene fragment. To construct the standard curve, qPCR was conducted on the target species using a pGEM-T Easy Vector, and a dilution series of the carp DNA containing 30 to 30,000 copies of DNA was created. This involved plotting the absorbance obtained using known concentrations of carp DNA. Using the standard curve for carp, researchers were able to obtain estimates for the number of copies of carp DNA in each sample they ran. Water that was collected as samples was immediately filtered using a centrifuge. Collected samples were amplified as standards in triplicate in all qPCR assays and researchers used the mean estimate from the triplicates as the response variable in a linear regression to evaluate the relationship between carp biomass and eDNA concentration.

Since one copy of carp DNA was detected in each triplicate, the Limit of Detection (LoD) for carp DNA was chosen to be a detection of at least one copy of carp specific DNA. Three technical replicates were run on each sample and if a technical replicate showed a negative result, it was assigned a value of zero. The mean value obtained in the three replicates was the quantity chosen to be used in regression. In addition, three wells containing no carp DNA were used as negative controls. In each of the three wells chosen to be negative controls, no eDNA of carp was detected.

The first experiment was done using tanks. Juvenile carp in the Mie Prefecture of Japan were captured and transferred to a lab in Kyoto. To study the impact of biomass, twelve tanks were used. Four tanks had only one carp, four tanks had five carps and the other four tanks contained ten carps. The water temperature was kept constant among the separate tanks. On the sixth day, triplicate samples of 50mL

were collected from each tank.

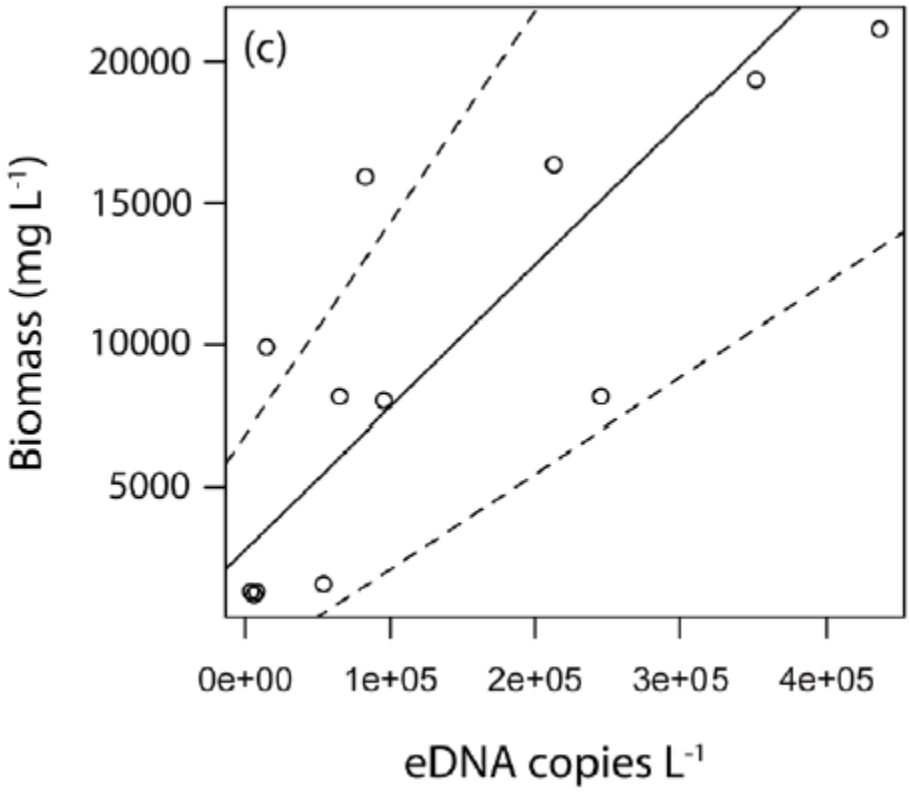

Figure 2.2: There was a significant positive correlation between the number of eDNA copies and carp biomass per 1 L (y = 0.050x+2789, R^2 =0.66, p= 0.001). There are twelve points, one for each tank used in the tank experiment. There is an overlapping of three points in the bottom left hand side of the figure. (Takahara et al., 2012).

Figure 2.2 shows the result of linear regression applied to the qPCR data. Using linear regression, researchers were able to obtain estimates of coefficients and other statistical quantities. In this experiment, linear regression was able to provide an adequate analysis with moderate explanatory power. The $R^2 = 0.66$ indicates that this linear model does a moderate job in explaining the variance in the data. The dashed lines show the associated confidence bands for the regression estimate. We also show later in our analysis that linear regression can be used as a reliable method for obtaining estimates of Coho biomass.

To bridge the research into the field, researchers collected water samples from Lake Biwa in Japan. The area of the lake in which they sampled was known to be inhabited by the common carp. From 21 distinct sites on the lake, 2-L samples of water were collected. Moreover, several environmental covariates were recorded such as water temperature, conductivity, etc. The samples were immediately transferred back to the lab where they were extracted and analyzed using the same standard curve for carp. All sequences from which the eDNA analysis indicated presence of carp, were confirmed to be regions of the lake that did contain carp.

In this experiment, researchers made use of General Linear Models and Stepwise selection. By comparing R^2 values, researchers were able to make conclusions regarding which covariates should be included in models. The model that included water temperature as a covariate turned out to work best in the field experiment and several environmental covariates such as pH or chlorophyll were seen to not be significant.

Table 1. Parameter coefficients and adjusted R^2 values of the full and best GLMs for the relationships between $\log_{10}$ (eDNA concentration+1) and six environmental factors.

	Full model	Best model
Habitat type (shore or offshore)	0.151	
Water temperature	**0.261**	**0.249**
Conductivity	−0.037	
Dissolved oxygen	−0.057	
pH of water	−0.052	
Chlorophyll *a* in water column	−0.100	
(Intercept)	**2.489**	**2.600**
Adjusted R^2	0.449	**0.539**

The factors were standardized and centered. The best model was selected by a stepwise procedure using the Akaike Information Criterion. Bold values indicate a significant contribution ($p<0.05$).
doi:10.1371/journal.pone.0035868.t001

Figure 2.3: Table summarizing the various estimates and adjusted R^2 values for a full model and the best model. (Takahara et al., 2012).

Figure 2.3 summarizes the result of the general linear model that researchers created to evaluate the relationship between eDNA concentration and several environmental covariates. The full model contains all the above environmental covariates and has an adjusted R^2=0.449. On the other hand, the best model only contains an intercept and water temperature, and the adjusted R^2 is improved to 0.539.

Prior to fitting a general linear model, researchers used a variance inflation factor (VIF) to assess whether there was collinearity among the covariates. This was done by dividing the variance estimates of the full model by the variance estimates of a model that only included one term at a time. Calculating the VIF for each covariate indicated that there was not sufficient evidence to conclude that collinearity existed among any of the covariates (this was done by ensuring that the VIF was not too large, in particular, less than a common cutoff value of 5 for all covariates). The eDNA estimates were obtained as above using qPCR and comparison to the standard curve, and a log10 transformation was applied to the estimates in an effort to normalize the results. A Shapiro Wilks test applied to the residuals confirmed normality of the estimates at a 95 percent confidence level. To select the best GLM, researchers compared models using stepwise selection based on AIC criteria. In this case, the best general linear model included the temperature of the water as an environmental covariate. Moreover, the estimate for water temperature was significantly positive.

In Chapter 5, we also use a stepwise selection approach to determine which environmental covariates to include in our field models.

2.0.3 eDNA and Species Occupancy

The study by Berger and Aubin-Horth (2018) further helps to illustrate the methods in which eDNA can be used to make inference regarding target organism occupancy. The *Schistocephalus solidus*-threespine stickleback pair is a commonly researched parasite/host pair. *Schistocephalus solidus* is a parasitic flatworm, while the abdominal cavity of the *Gasterosteus aculeatus* (threespine stickleback) is one possible host location for this parasite. Conventional methods for detecting if a stickleback is infected with the parasite is by simple visual detection. However, this results in not identifying parasites that are too small or have not yet grown to full size (past a certain mass threshold).

An alternative, non-invasive method to detect the parasitic tapeworm involves using qPCR to test for the presence of *Schistocephalus solidus* eDNA. Researchers took samples from the abdominal cavity of sticklebacks and performed qPCR to test for *Schistocephalus solidus* DNA. Using this method, scientists were able to correctly assign the status to 98 percent of n=151 fish. Moreover, not only did qPCR allow for the detection, but it also allowed for a comparative quantification of eDNA. That is, researchers were able to get an idea of how large the parasite was simply based on the results of the qPCR. Lower CT scores indicated higher presence of eDNA and was correlated with larger parasite mass.

Researchers used minnow traps in Lac Temiscouata and Isle-Vertes (Quebec, Canada) to capture the sticklebacks. The fish were transferred to Laval University where they were maintained according to their natural requirements. 96 Fish were taken from Lac Temiscoutata and 55 from Isle-Vertes. By the time the experiment began, all fish were considered to be adult fish. The fish were each subsequently isolated into 2 L tanks in the lab.

The fish were removed individually and placed on sponges. A small 1ml volume syringe filled with 100 μL of a phosphate buffer was injected into the abdominal cavity of the fish. Without removing the needle, the syringe was pulled back until it had 100 μL of the recently injected buffer back in the needle. The needle was then removed, and the fish was returned to its tank. This removed liquid was added directly to a tube containing 700 μL of Longmire Lysis preservation buffer. This was done twice for each of the fish on two consecutive days. The samples were then stored in an appropriate environment to maintain the delicate structure of the DNA. Note that on the second sample, the phosphate buffer was added to the same tube containing the 700μL of lysis buffer from the previous day. Hence, each tube had up to 900 μL of liquid. Once the fish had been sampled twice, they were killed with an injection of MS-222 and dissected. Fish size, sex and mass were recorded as well as the mass and number of any *Schistocephalus solidus* parasites in the fish. A 'PI' value was calculated where PI= [total mass of parasite (mg) / total mass of host plus parasite (mg)]x100.

30μL of proteinase K was added to the Lysis buffer tubes and incubated at 55° C overnight. The sample was then transferred to a new 2 mL tube and 950 μL

of phenol:choloroform:isoamyl alcohol was added. This new tube was then shaken and went through the centrifuge. Several other chemicals were added and the tube containing the DNA and buffers was stored overnight.

To detect DNA from the parasite, researchers used a pair of primers that amplified a specific sequence from a nuclear gene in the *Schistocephalus solidus* genome. The specific gene sequence has no known analogs in other species. PCR assays were then tested on genomic DNA to confirm that the primers indeed amplified the parasite DNA and did not falsely amplify other DNA from the stickleback.

Prior to the experiment, researchers obtained what is known as a 'Standard Curve' for *S.solidus* genomic DNA. A 'Standard Curve', also known as a 'Calibration Curve' allows researchers to make inference about DNA concentrations using previous knowledge of that species DNA (Larionov et al., 2005). In this case, researchers used known concentrations of *S. solidus* DNA obtained from three individuals and prepared a wide range of dilutions. At the low end, researchers prepared a 0.01 ng μL^{-1} sample and at the high end a 0.9 ngμL^{-1} sample. Researchers also created multiple samples with known concentrations in between the minimum and maximum values. These solutions are referred to as the 'standard solutions'. The fluorescence produced by cycling these standard solutions was then measured using a spectrophotometer and the results were recorded. A graph was produced showing the relation between concentration of eDNA and the CT value. This is known as a 'Standard Curve'. Hence, the Standard curve allowed researchers to make conclusions about new DNA concentrations using prior knowledge of the properties of that specific target DNA.

Finally, researchers were able to begin to analyze the eDNA. The removed eDNA was amplified using qPCR and the primers described. CT values were recorded when the fluorescence reached a certain threshold. Researchers included negative controls (no DNA and no primers) in each plate. Each reaction was done in triplicate.

Once the qPCR amplification had taken place, the presence or absence of eDNA in the fish was determined by comparing the obtained CT values with the standard curve. If more than 70 cycles of qPCR were needed, this was considered to be 'undetermined' for the presence of parasite eDNA.

For each fish, if at least one of the three CT values was less than 70, this was

considered a positive identification of the parasite. If the CT was undetermined for all three replicates (more than 70 cycles) it was also considered a negative. The CT value used for analysis was the lowest CT value obtained among the three replicates. Hence, if the CT was only undetermined for one sample, then the lowest CT score among the other two replicates was used. If it was undetermined for two replicates, then the CT of the remaining replicate was used.

Statistical analysis was done using the R programming language. Researchers used a non-parametric Spearman correlation to test whether the eDNA concentration (estimates obtained by comparing CT values collected during qPCR to the previously defined standard curve) negatively correlated with the parasite mass. The results of the experiment were promising. In total, of the 151 fish collected, 35 were infected with parasites. Using eDNA, researchers correctly predicted 32 of the 35 infected fish. Hence only 3 infected fish were incorrectly classified. Of the 116 fish that had no parasites, researchers never falsely predicted infection. A CT value of less than 70 was used as the detection cutoff.

We define True Positive (TP) to be the number of infected fish that were classified as infected, True Negative (TN) as the number of non-infected fish that were classified as non-infected, False Positive (FP) to be the number of fish incorrectly classified as infected and False Negative (FN) as the number of fish incorrectly classified as having no parasite. In this example, TP=32, TN=116, FP=0 and FN=3. Using these definitions, we calculate some key statistical values such as accuracy, sensitivity and specificity (Hastie et al., 2001).

Firstly, we calculate accuracy= $\frac{TP+TN}{TP+TN+FP+FN} = \frac{32+116}{32+116+3} = 0.9801$
Secondly, we calculate sensitivity=$\frac{TP}{TP+FN} = \frac{35}{35+3} = 0.914$
and we calculate the specificity=$\frac{TN}{TN+FP} = \frac{116}{116+0} = 1$

Hence, we see that the researchers were very accurate. The testing also had perfect specificity, which means that researchers never predicted that a healthy fish had a parasite.

A scatterplot of parasite mass versus CT confirms what we would expect. High parasitic mass is associated with lower CT values! This is confirmed by the negative Spearman correlation of -0.41. Although the variance appears to be somewhat high despite a clear trend. In this experiment, the study of eDNA alone produced highly accurate identification of infected versus non-infected species The plot shows the lowest CT value obtained from the three replicates for each of the organisms that were confirmed visually to have a parasite (Berger and Aubin-Horth, 2018).

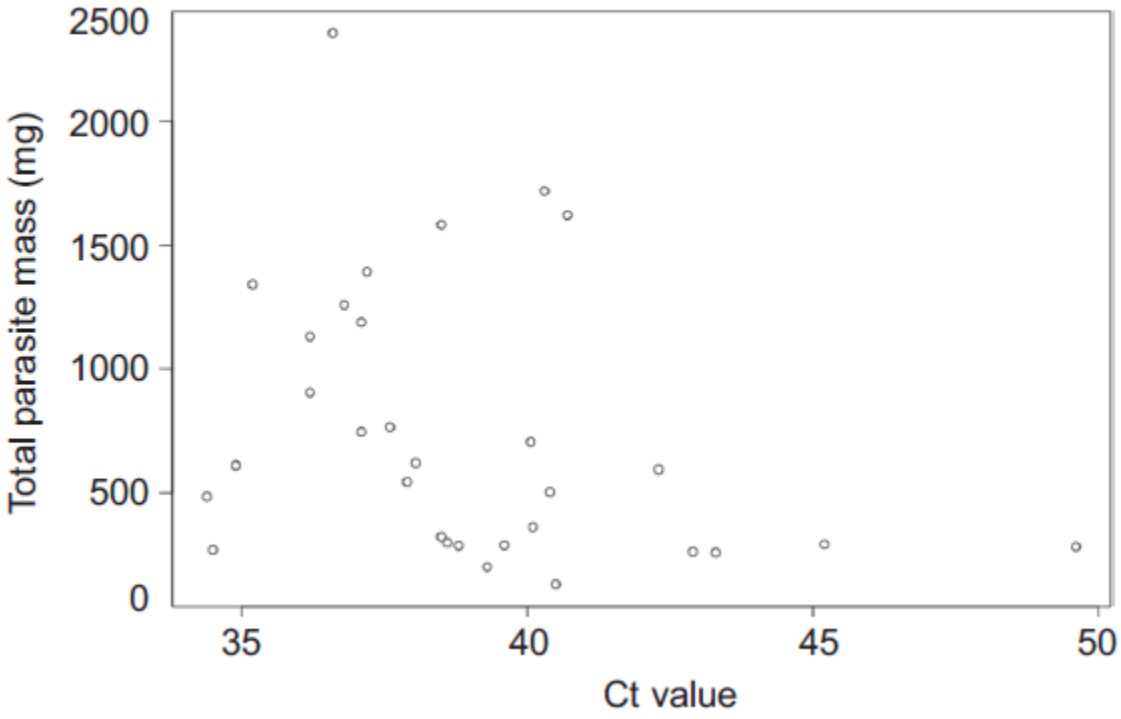

Figure 2.4: eDNA concentration co-varies with total parasite mass. 'eDNA concentration co-varies with total parasite mass. Low cycle threshold (CT) values, which represent high levels of eDNA, are associated with a high parasite mass (Spearman correlation=-0.41, P=0.01, n=31) '. The CT value plotted is the lowest value obtained from among the three replicates. (Berger and Aubin-Horth, 2018).

Figure 2.4 is a visualization of the CT scores obtained for the diseased fish. In general, parasite mass was negatively correlated with CT score. One fish was removed from the original 32 as it died before dissection could be performed. Hence there are 31 points plotted.

2.1 Environmental Factors

As interest in the study of eDNA grows, it is critical that researchers understand how environmental factors could influence detection protocols. The impact of environmental 'covariates' such as weather, proximity to target organism or the time of day and how these impact eDNA concentrations and collection is still a subject of ongoing study (Lodge et al., 2012). In Chapter 5 we consider how these environmental covariates may impact eDNA concentrations in the field.

To further understand how environmental covariates may impact eDNA measurements, we consider the following case study. The Rocky Mountain tail frog *Ascaphus montanus* and the Idaho giant salamander *Dicamptodon aterrimus* are two species that are endemic to mountainous regions of Western North America (Lannoo, 2005). Both species are considered to be 'secretive' and are especially difficult to study using standard sampling techniques. It is crucial that scientists and investigators understand how the results obtained via eDNA methodology may differ (or not) from standard techniques such as transect sampling or electrofishing.

In one study (Pilliod et al., 2013), researchers collected eDNA samples from 13 streams in the South Fork Salmon River Sub-Basin, located in Idaho, USA, during the summer of 2011. In addition, multiple negative controls were taken more than 100km south of where either species is known to live.

In this study, researchers described four main goals. The first goal was to "Develop and test alternative field protocols for sampling stream water for eDNA". The second goal was to "Compare estimates of detection probability, density, biomass and occupancy obtained via different methods". The third goal was to examine how covariates (time of day, location in stream, etc) influenced eDNA detection and the fourth and final goal was to examine factors influencing precision of eDNA concentration estimates.

Four of the original 13 streams were chosen to compare three distinct methods of sampling eDNA (In-Stream, Grab-and-filter and Grab-and-Hold). Researchers compared three field methods in four streams, direct filtration of stream water in the field (in-steam), water collected in a 1L Nalgene followed by immediate filtration (grab-and-filter) and water collected in a 1L Nalgene, stored over the night and filtered

in the lab the next day (grab-and-hold). Each of these three sampling methods has advantages and disadvantages with respect to time, money and effort (Goal 2).

Table 1. The average probability of detecting (*p*) Rocky Mountain tailed frog or Idaho giant salamander eDNA from water samples collected in four streams using different field methods.

Species	In-stream	Grab-and-filter	Grab-and-hold	Species *p*
Rocky Mountain tailed frog	1	0.83	0.92	0.92
Idaho giant salamander	1	0.92	1	0.97
Method *p*	1	0.88	0.96	

Note: Detection probability *p* was estimated as the number of replicates where a species' DNA was detected divided by the number of replicates collected in each stream ($n = 3$). Detection probability for each species (Species *p*) uses the same approach, but ignores method ($n = 9$). Likewise, Method *p* ignores species ($n = 6$). See Materials and methods for description of each water collection protocol.

Figure 2.5: Comparison of three eDNA sampling methods. (Pilliod et al., 2013)

Figure 2.5 illustrates a comparison of three eDNA sampling methods (Goal 2). In stream methodology had perfect detection, but all three methods produced similar estimates. Note that 'method p' ignores species while 'Species p' ignores method and averages over all methods. All three eDNA collection methods produced high detection rates. However, only 'In-Stream' filtering led to perfect detection for both species. Researchers found not enough evidence to conclude that any of the three eDNA collection methods were superior to any other.

To research Goal 1, the researchers a set up standard sampling techniques (using thirty 1m transects randomly placed within 1km upstream of each of the 13 streams where they had taken eDNA samples). Researchers then counted and weighed all larvae from either species on all the transects. The reason they chose larvae is because this is the most common life stage of the amphibians and thus can be used to infer information regarding adult populations. Larval density and biomass were calculated for each of the 30 transects and the average values were calculated.

Researchers then performed Backpack electrofishing on the four specified streams 500m upstream from where they had collected eDNA. Researchers then estimated giant salamander density by dividing the number of larvae caught by the total area they searched. Electrofishing allowed the researchers to assure that their methods of kick-net sampling were reliable. Kick-net sampling involves holding a net under the water and kicking the bottom of the substrate in an effort to direct organisms and materials towards and into the net. Indeed, results obtained via electrofishing were

excellent predictors of how many larvae had been captured using kick-net sampling (which was done immediately prior to the start of the experiment in each of the four specified streams). (Linear regression obtained a R^2 of 0.96 and p-value of p=0.019).

To investigate the impact of covariates (Goal 3), researchers chose three of the thirteen streams for additional analysis. These were the Deadwood River stream, the East Fork Deadwood stream and the Weir Creek. In the Deadwood river and East Fork, eDNA concentration was measured using the in-stream method for tailed frogs in two reaches 50m apart over a 48-hour period. In Weir Creek, multiple eDNA samples were taken (every 50m for 2km). Following this collection, a three-day multiple removal process (electrofishing) of salamanders took place. The salamanders were removed, weighed and the exact location in which they were removed was recorded. This capture data was used to create a population distribution model for salamanders along the 2km Creek. This data was used to study whether eDNA concentration was impacted by the abundance of salamanders upstream of where the sample was taken.

For each of the thirteen streams, three replicate samples of surface water were collected. Each sample was taken by pumping 1L of water through a disposable filter funnel with a 47mm diameter cellulose nitrate filter paper. Negative controls were simply 1L of store-bought distilled water. Genetic analysis (qPCR) was then performed on the filter papers. DNA extracted from both species was used to create serial dilutions (standard curves) for comparison. In particular, samples of tissue taken from the tails of each species was used. To quantify the DNA concentration in the original samples, researchers used a NanoDrop spectrophotometer to estimate the amount of DNA in each sample.

To compare eDNA concentration estimates obtained from the standard curve among the three different field methods (Goal 4), researchers used a mixed effect model with 'stream' as the random effect. Analysis was performed in SAS. Mixed models are extensions of linear models that allow for the inclusion of both fixed and random effects.

Researchers also compared stream-level occupancy and detection probability estimates obtained via standard sampling versus those obtained using eDNA methodology. To test whether eDNA concentration was correlated with abundance, researchers used a General Linear Model. Each of the thirteen streams was a sampling unit and

the predictor variables were mean density, mean biomass or proportion of transects occupied (occupancy). The response variable was mean eDNA concentration and was calculated using the three replicate subsamples from each stream (obtained by referring to the standard curve).

To test for the effect of covariates, further analysis was carried out on the samples obtained from the three specified streams above (Deadwood, East Fork and Weir). Sample location was included as a predictor, along with time of day.

One main finding of this study is that eDNA detection probabilities do not appear to be influenced by water collection methods. In this study, eDNA methodology resulted in better detection rates then simple transect methods. Finally, in this study, covariates such as time of collection did not appear to be of major impact on calculated eDNA concentration. In Chapter 5, we consider the impact of environmental covariates on eDNA concentrations in the field.

2.1.1 eDNA movement

eDNA of aquatic species is shed into streams or rivers. Hence, it is crucial that researchers understand how measurement of eDNA is impacted by movement in systems such as complicated, flowing water (Shogren et al., 2017). In standing or still waters, measurement of eDNA has been widely used to estimate target species abundance (Takahara et al., 2012). However, studies of eDNA within moving water systems are much less common. In order to study eDNA collected from moving water systems, it is important that researchers understand the biological and physical properties that influence retention and transport of eDNA travelling within these waters. In Chapter 4, we fit a variety of non-linear models to account for flow and dilution.

Researchers have suggested three major mechanisms that impact eDNA moving systems. Firstly, there is *Transport* (downstream movement driven by bulk water flow), secondly there is *retention* (deposition or capture by the streambed) and lastly, there is *resuspension* (eDNA embedded in the streambed may become loose or free again) (Shogren et al., 2017).

It has been demonstrated that eDNA in streams is significantly retained by the streambed as it moves down the stream and that concentration of eDNA decreases with downstream distance. Also, although eDNA is deposited on the streambeds, some of this eDNA will end up resuspended in the moving stream (Jerde et al., 2016). More specifically, it has been shown that 'fine' substrates such as sand retain more eDNA than 'course' substrates such as pea gravel (Shogren et al., 2016).

In a study (Shogren et al., 2017), researchers asked three main questions corresponding to each of the three major mechanisms. 1), "How far does an average eDNA particle travel in streams with different hydrologic signatures and how might eDNA detection be used to infer where a species is located?" 2) "Does surface-subsurface exchange trap eDNA in porous, benthic substrate interstices and does the presence of biofilms of organic matter play a role?" 3) "Can resuspension from the streambed result in eDNA detection after the source of eDNA has been removed, and how does the characteristic of the streambed impact this?"

To study these mechanisms, researchers proposed three hypotheses. The first hypothesis was that downstream transport of eDNA will be limited by hyporheic

exchange rates (how quickly water is moving in and out of the stream). In Chapter 4 we also consider the impact of flow on eDNA movement in a controlled experiment. Figure 2.6 illustrates the three proposed governing mechanisms of eDNA movement in streams.

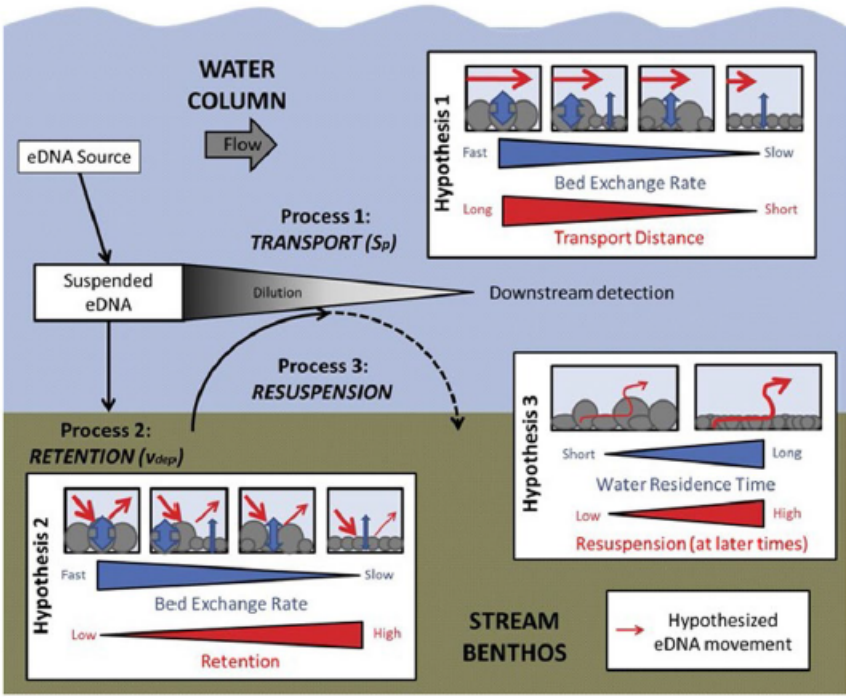

Figure 2.6: Conceptual diagram depicting the three governing processes of eDNA movement in streams: 1) Transport, 2) Retention, and 3) Resuspension, and associated hypotheses for eDNA movement (Shogren et al., 2017)

Researchers conducted experiments in four streams at the University of Notre Dame Linked Experimental Ecosystem Facility (ND-LEEF) in the summer of 2014. Each of the four streams are 0.4m wide, 7-10cm deep, and 60m long. The water is sourced from a head reservoir fed by groundwater. The four streams are nearly identical in all aspects except that each stream has differing benthic substrate lining the bottom.

Researchers collected water from a local pond that was known to have a high density of carp. This collected water was then pumped into the head of each of the four streams at a rate of 100mL/min. This was meant to simulate a carp at the

head of the stream, whereby it was hypothesized the eDNA would flow or 'transport' downstream. The injection of pond water was done for four hours. At the four-hour mark, 15 samples of 250mL were collected at predefined distances from the head of each stream. Once the pump was turned off (simulating fish removal), researchers continued to collect samples at regular timed intervals. The samples were stored and transferred to the lab, where qPCR was performed.

To quantify the number of DNA copies in each extract, researchers created a synthetic standard curve and included it on each qPCR plate along with the DNA extracts. Since researchers knew the amount of DNA in their dilution series, they were able to create a standard curve for carp concentration. By comparing the known molecular weights with the obtained CT scores, the researchers were able to apply linear regression to create the standard curve. Moreover, the researchers were able to estimate the number of DNA copies in samples by dividing the estimated molecular weight of the DNA by Avogadro's number. The LoD used in this study was 30 copies of DNA per reaction. The LoD was chosen based on the creation of the standard curve, as 30 copies of DNA per reaction resulted in a detection in 95% of reactions.

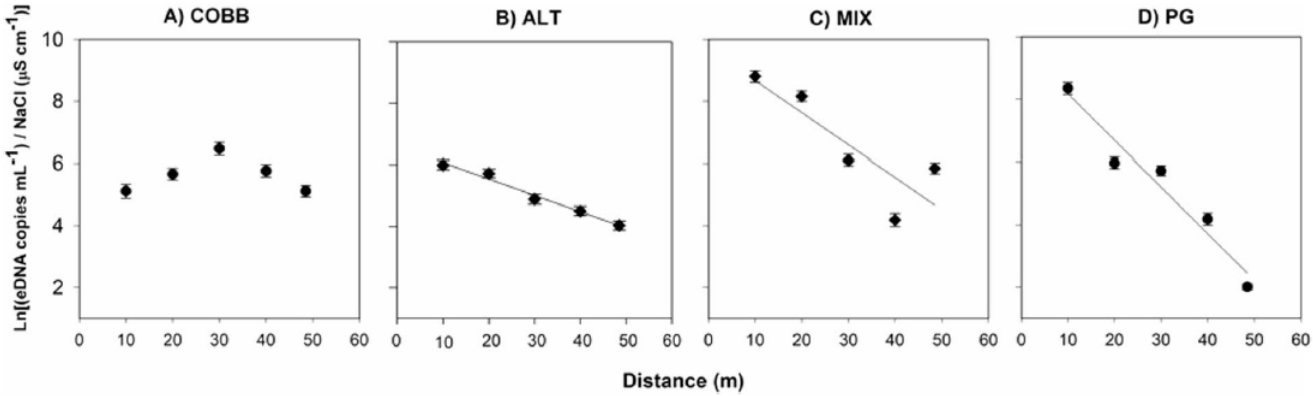

Figure 5. Regressions for COBB (**A**), ALT (**B**), MIX (**C**), and PG (**D**) natural-log transformed concentrations over distance used to estimate parameters reflected in Table 1. Each dot represents the mean of three replicate field samples ± SE bars.

Figure 2.7: Results of the flow experiment (Shogren et al., 2017). Stream PG has a smooth substrate, while COBB has course substrate. ALT and MIX have substrates in between course and smooth.

Figure 2.7 illustrates the results of the experiment. In general, eDNA concentration decreased as the distance from the head of the streams increased. However,

differences in substrate also resulted in different concentrations down the streams. COBB, ALT, MIX and PG are codes for each of the four streams. NaCl was added to the DNA samples as a solute as NaCl helps to prevent unwanted dilution of DNA. The points plotted are the log transformed mean number of estimated copy numbers from each set of replicates, with the associated line of best fit created using linear regression.

2.1.2 Temperature

Water temperature is known to be highly influential to the metabolism of fish (Selong et al., 2001). The exact impact that temperature has on eDNA release is still not completely understood. It has been shown that in general, higher water temperature does increase the metabolism of fish.

One study (Lacoursière-Roussel et al., 2016) researched the question of how water temperature and the eDNA capture method impact the relationship between eDNA concentration and fish abundance. Researchers collected Brook Charr fingerlings from a fish hatchery in Cap-Sante, Quebec. The fish were stored and returned alive to the lab. Fifteen tanks were prepared with water at 7° C, and another fifteen tanks were prepared with water 14° C.

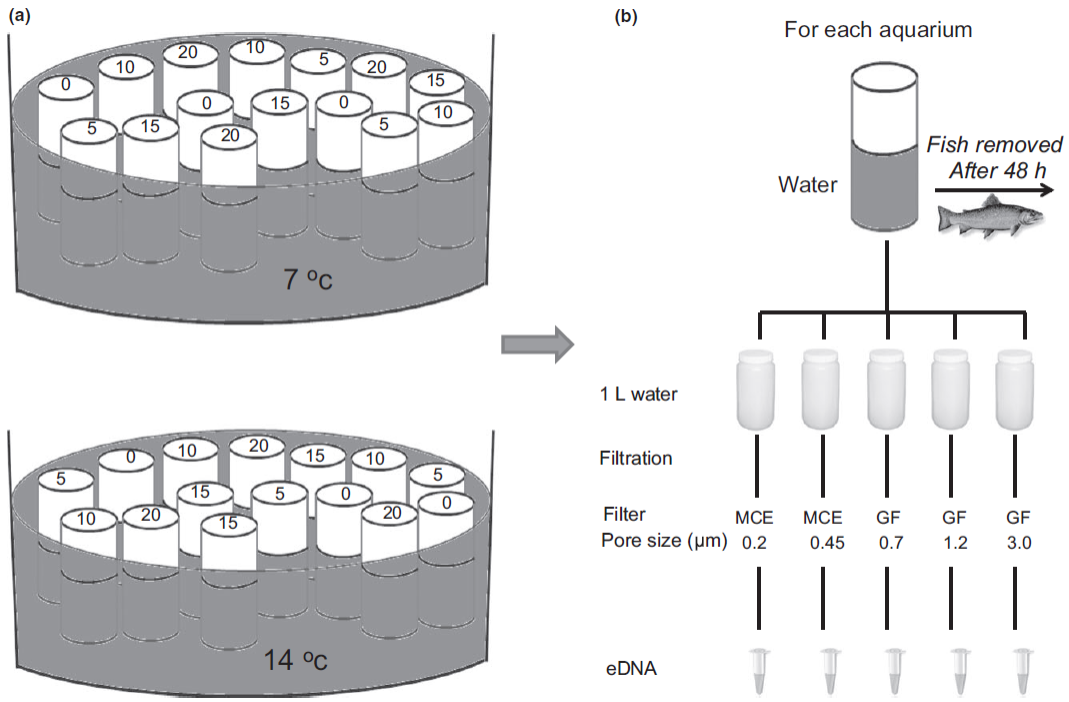

Figure 2.8: Experimental design of Brook Charr exposed to 7° C and 14°C and (b) eDNA collection procedures. Fish abundance for each water temperature was randomly assigned. eDNA was captured via filtration (MCE, mixed cellulose ester filters; GF, glass fiber). (Lacoursière-Roussel et al., 2016)

Figure 2.8 illustrates the experimental design. For each of the two distinct temperatures, 0, 5, 10, 15 and 20 Brook Charr were placed into the tanks. In total there were 30 tanks. For each temperature, each abundance of fish was repeated three times. That is, three tanks at 7° C contained 5 fish and three tanks at 14° C also contained 5 fish. The same was done for each number of fish. The zero fish triplicates represent negative controls, i.e. an abundance of 0.

The total biomass of fish in each tank was measured by comparing the weight of the tanks prior to the addition of the fish to the weight after they were added. The fish were then given five days to live in each tank. Each tank was then removed and let to sit in the dark for 48 hours, at which point the fish were removed. Once removed, the water in each 1L triplicate was mixed and then filtered using various filters and pore sizes to collect eDNA. DNA was extracted using the salt extraction method. A 139-basepair unique to Brook Charr was used as the target DNA. A standard curve was created using known 7 known dilutions of Brook Charr eDNA. eDNA concentration of samples was then quantified using real-time Taq-Man PCR and comparison to the 7-point standard curve previously created.

A Mann-Whitney test was used to test for significance in the mean eDNA concentration between the two separate temperatures (there were 150 fish for both temperatures for a total of 300 fish). A Cook's distance test was used to detect potential outliers and influential points. At all levels of fish abundance, higher eDNA levels (more concentrated) were found in the 14° C samples compared to the 7° C samples when using glass fiber (GF) filters. For a few filters that are known to have a low level of eDNA collection, such as MCE(2 μL), differences in eDNA at the two temperatures were not detected.

The results of this experiment showed that the concentration of eDNA did increase significantly at higher temperatures. Researchers hypothesized this may be due to the fact that fish mobility is increased at higher temperatures (Petty et al., 2014), moreover at higher temperature the rate at which skin cells are shed and other secretions are released is increased. Hence, temperature may need to be considered when a data analysis on eDNA on aquatic species is performed. This is particularly important when using fine filters that are known to capture high levels of eDNA. In Chapter 5, we consider the possible impact of temperature on eDNA concentrations

as well.

Chapter 3

Density Experiment

3.0.1 Introduction

The *Density Experiment* was conducted in the summer of 2015 (MacAdams, 2018; Hocking M.D et al., 2020). The goal was to investigate the relationship between Transformed CT values ('TCT') and Coho Salmon 'density' (or the similar measurement of biomass). CT is the 'Cycle Threshold' and we define TCT=50-CT. The experiment consisted of manipulating juvenile Coho salmon densities in treatments of 1, 2, 4, 8, 16, 32, and 65 fish in replicated 10,000L tanks.

Figure 3.1: Sampling schedule for the Density Experiment (Bergman, 2020).

Figure 3.1 (MacAdams, 2018) provides an overview of the experimental schedule. For each main tank, Coho were added throughout the week of the experiment (indicated by the green arrows). On day one, one fish was added to the tank. On each of the following days of the week, fish were added until on day 7 there were 65 fish.

There were four 10,000L Tanks, numbered 19, 20, 21 and 24 for the main density experiment. From each tank, sample replicates of water were taken, usually five per tank. For each fish, the biomass in grams was also recorded. The average mass of

each fish used in the experiment was 5.69 grams. As the amount of water in each tank in this experiment is fixed, we expect that as biomass increases, that the CT score should decrease (as we would expect more eDNA), hence the TCT should increase.

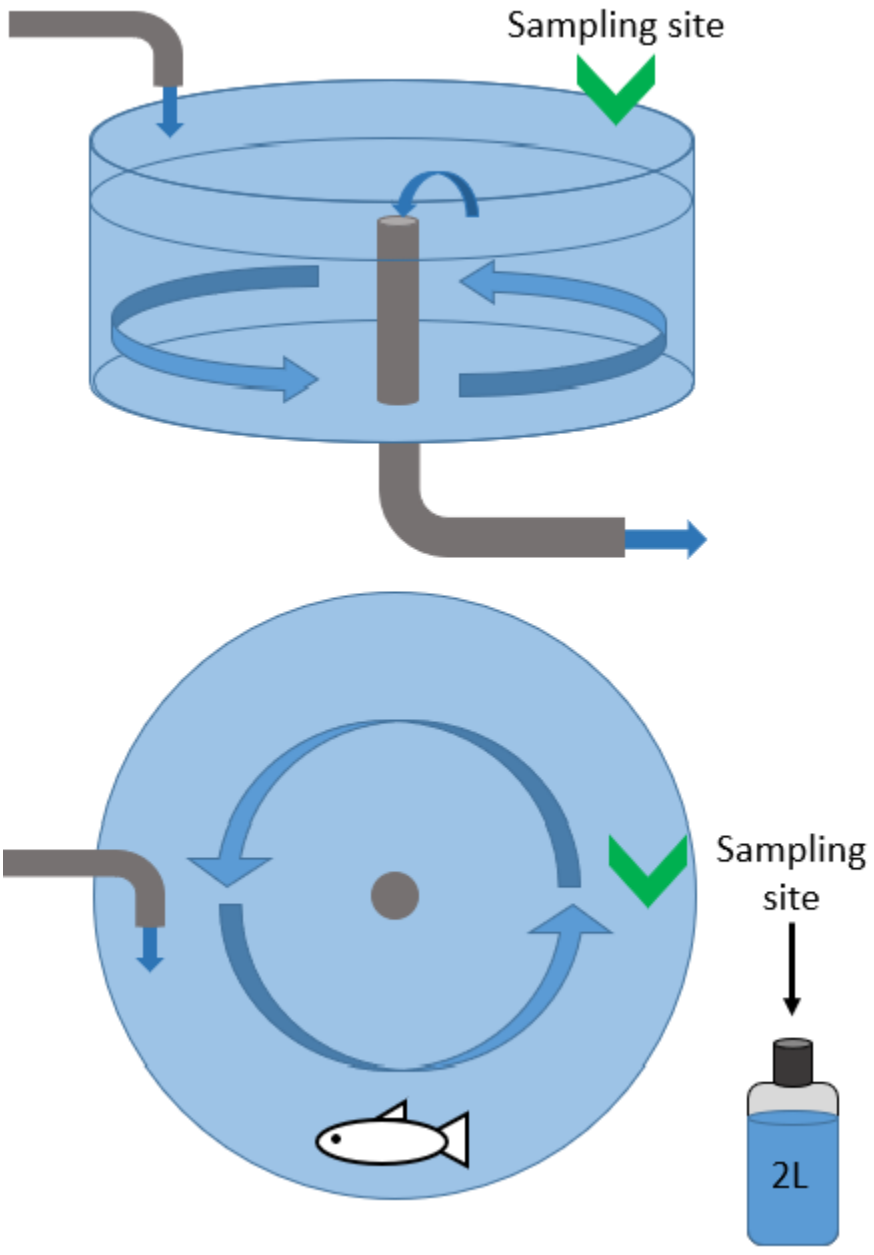

Figure 3.2: The general sampling methodology for the density experiment (Bergman, 2020).

Figure 3.2 illustrates the method in which samples were taken from the tanks. To sample from the tanks, 2L bottles were filled with water by submersion approximately 0.5m from the edge of the tank. The water was then filtered through 47mm diameter 0.45 μm mixed-cellulose ester (MCE) filter membranes. From each sample replicate, eight technical replicates were obtained. These technical replicate samples were then evaluated using qPCR for the presence of Coho eDNA. Each unique set of eight technical replicates were assigned a unique "sort.code". That is, every member coming from the same group of eight technical replicates has the same sort code. Note that one set of technical replicates, corresponding to sort.code=128 was discarded before analysis due to lack of integrity in the lab (Sample.replicate 3 of Tank 20, 65 Fish).

In order to detect Coho salmon specifically, a DNA assay, eONKI4 was used in combination with qPCR. To ensure quality of the eDNA, samples first underwent integritE-DNA tests. These tests combine a probe and a primer to amplify chloroplast DNA that appear pervasively in freshwater systems. Samples that failed the integritE-DNA test were cleaned using the Zymo OneStepTM PCR Inhibitor Removal Kit and tested again. If the sample failed a second time, it underwent an inhibitor removal. If the sample still failed to pass the integritE-DNA test after inhibitor removal, it was excluded from the data to minimize the presence of false negatives.

For 1, 2, 4, 8, 16, and 32 fish, 6*4*5*8= 960 observations were recorded. For each set of these six numbers of Coho, there were four tanks, for each tank there were five sample replicates, and for each sample replicate there were eight technical replicates. For 65 Fish, there were 3*5*8+4*8=152 observations. For 65 fish there were three tanks with five sample replicates and one tank with four, for each sample replicate there were eight technical replicates.

A small pilot experiment was also conducted during the summer. The pilot experiment used smaller tanks, tanks 1, 2, 3 and 4. Samples were taken from smaller tanks that only contained hatchery water and no Coho. The pilot experiment was used to gain insight regarding possible background signals in the hatchery waters.

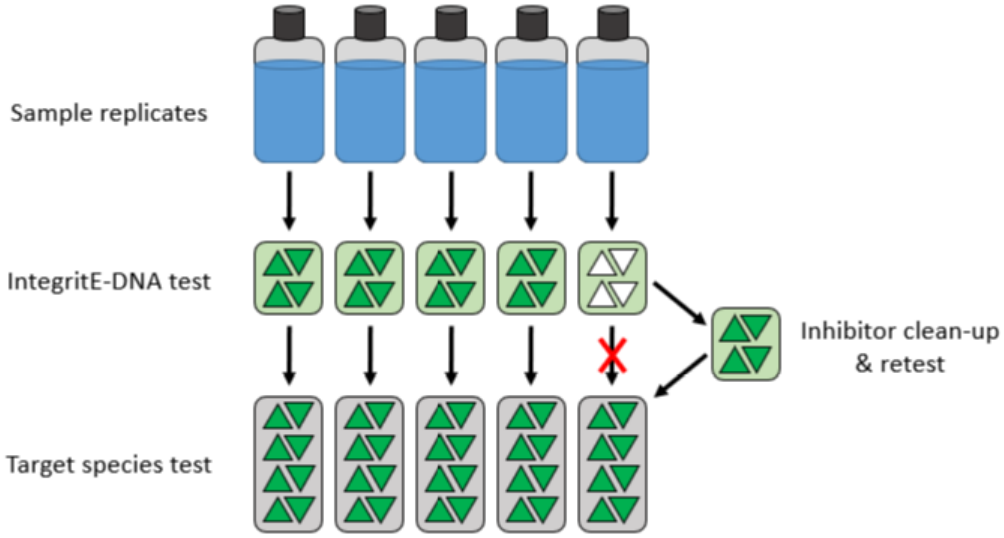

Figure 3.3: Basic procedure for analyzing eDNA. Sample replicates were first certified using an integritE-DNA test before testing for Coho eDNA (Bergman, 2020).

Figure 3.3 is a visualization of the process of DNA integrity validation. integritE-DNA tests were performed on samples prior to using qPCR to search for Coho eDNA. Before the experiment began, the tanks were bleached to clean any residual eDNA or materials that may have been present in the tanks. integritE-DNA test combines a probe and a primer to amplify chloroplast DNA that is pervasive in freshwater systems.

3.1 Summary Statistics

We now provide several tables that summarize features of the experiment. For example, we include tables listing the numbers of fish and which tanks they belonged to. Tables were created in R using the 'kableExtra' package (Zhu, 2019).

Summary of Biomass (gram)

Day	Tank	Biomass	Fish
1	19	5.69	1
1	20	5.69	1
1	21	5.69	1
1	24	5.69	1
2	19	11.38	2
2	20	11.38	2
2	21	11.38	2
2	24	11.38	2
3	19	21.78	4
3	20	21.98	4
3	21	21.98	4
3	24	21.98	4
4	19	43.88	8
4	20	42.68	8
4	21	45.38	8
4	24	43.08	8
5	19	88.18	16
5	20	82.98	16
5	21	88.08	16
5	24	84.68	16
6	19	188.58	32
6	20	173.78	32
6	21	182.58	32
6	24	175.48	32
7	19	380.38	65
7	20	358.68	65
7	21	363.28	65
7	24	377.98	65

Table 3.1: Table summarizing biomass per tank and day.

Table 3.1 provides the weight and number of the fish in each tank. For every day for a week, the amount of fish in the tanks (and hence the biomass) was doubled (except on the seventh day, it was doubled plus one to make 65 fish)

Number of Sample.replicates by number of Fish and Tank number.

Fish	Tank							
	1	2	3	4	19	20	21	24
0	24	3	3	3	5	5	5	5
1					5	5	5	5
2					5	5	5	5
4					5	5	5	5
8					5	5	5	5
16					5	5	5	5
32					5	5	5	5
65					5	4	5	5

Table 3.2: Summary of the number of sample replicates for each corresponding number of fish and tank number.

Table 3.2 provides the number of sample replicates that were taken from each tank number, and how many fish were present in that tank. Tanks 1, 2, 3, 4, 19, 20, 21 and 24 all had samples taken in which there were zero fish. However, for most levels of fish density, such as 16 fish or 32 fish, only tanks 19, 20, 21 and 24 were used. In general, five sample replicates were taken from each of the four main tanks for each level of fish density (except for tank 20, 65 fish).

Summary of TCTs by Tank and Fish over samples and techincal replicates.

Fish	Tank	Min TCT	Max TCT	Median TCT	Mean TCT
1	19	0.00	17.42	15.220	14.38
1	20	14.52	19.00	16.355	16.47
1	21	0.00	17.77	13.455	12.59
1	24	15.31	18.37	16.655	16.79
2	19	12.94	16.65	15.370	15.31
2	20	13.18	17.68	16.050	16.03
2	21	0.00	15.33	13.620	13.15
2	24	12.84	17.20	15.395	15.37
4	19	13.44	18.07	15.990	16.00
4	20	13.74	17.89	16.730	16.58
4	21	11.47	16.30	14.430	14.36
4	24	14.44	18.78	17.230	17.05
8	19	15.68	20.74	16.905	17.81
8	20	14.17	18.97	17.800	17.41
8	21	14.79	17.31	15.935	16.13
8	24	17.15	20.18	19.125	18.95
16	19	16.81	19.87	18.390	18.46
16	20	15.42	18.84	17.720	17.51
16	21	15.60	17.96	16.635	16.68
16	24	16.70	19.36	18.665	18.36
32	19	17.87	20.78	19.805	19.32
32	20	19.44	21.90	20.855	20.74
32	21	17.88	19.90	18.720	18.78
32	24	18.45	21.01	19.700	19.76
65	19	19.98	22.47	21.270	21.33
65	20	19.84	23.39	21.950	21.75
65	21	17.72	20.65	18.925	18.99
65	24	19.43	21.66	20.810	20.68

Table 3.3: Summary of the minimum, maximum and median TCT for each number of fish and tank.

Table 3.3 provides several summary statistics for TCT values obtained. For each number of fish and tank, there were five sample replicates, and for each sample replicate there were eight technical replicates. Hence the maximum, minimum and median calculations use 5 x 8= 40 measurements each. Because one set of technical replicates was discarded from 65 fish, tank 20, only 32 measurements were used for that set of summary statistics.

Zero Fish Summary-Pilot study.

Date (2015)	Fish	Tank	Sort Code	Min TCT	Max TCT	Median TCT	Mean TCT
08-19	0	1	141	0	0	0	0
08-19	0	1	142	0	0	0	0
08-19	0	1	143	0	11.2	0	1.4
08-20	0	1	144	0	14.03	0	3.47
08-20	0	1	145	0	12.25	0	3.02
08-20	0	1	146	0	13.2	0	4.86
08-21	0	1	147	0	12.14	0	1.52
08-21	0	1	148	0	0	0	0
08-21	0	1	149	0	0	0	0
08-22	0	1	150	0	0	0	0
08-22	0	1	151	0	12.24	0	4.21
08-22	0	1	152	0	0	0	0
08-23	0	1	153	0	0	0	0
08-23	0	1	154	0	12.76	0	1.6
08-23	0	1	155	0	0	0	0
08-24	0	1	156	0	0	0	0
08-24	0	1	157	0	0	0	0
08-24	0	1	158	0	10.2	0	1.27
08-25	0	1	159	0	12.77	0	1.6
08-25	0	1	160	0	0	0	0
08-25	0	1	161	0	13.08	0	3.13
08-12	0	1	162	0	0	0	0
08-12	0	1	163	0	12.8	0	1.6
08-12	0	1	164	0	0	0	0
08-12	0	2	165	0	0	0	0
08-12	0	2	166	0	0	0	0
08-12	0	2	167	0	0	0	0
08-12	0	3	168	0	0	0	0
08-12	0	3	169	0	0	0	0
08-12	0	3	170	0	11.64	0	4.31
08-12	0	4	171	0	0	0	0
08-12	0	4	172	0	0	0	0
08-12	0	4	173	0	0	0	0

Table 3.4: These samples correspond to the Pilot experiment

Table 3.4 summarizes the samples that were taken concurrently over the course of the seven days in which the density experiment took place. The sampling began on August 19, 2015 and continued until August 25, 2015. These samples correspond mostly to a smaller tank (tank 1) and were taken during the pilot experiments. Samples corresponding to sort codes 162-173 are also part of the pilot experiment but

were taken earlier on a single day (August 12).

Finally, we include a table that summarizes the negative controls for the four main tanks.

Zero Fish Summary.

Fish	Tank	Sort Code	Min TCT	Max TCT	Median TCT	Mean TCT
0	19	174	0	0.00	0	0.00
0	19	175	0	0.00	0	0.00
0	19	176	0	0.00	0	0.00
0	19	177	0	0.00	0	0.00
0	19	178	0	0.00	0	0.00
0	20	179	0	0.00	0	0.00
0	20	180	0	0.00	0	0.00
0	20	181	0	0.00	0	0.00
0	20	182	0	0.00	0	0.00
0	20	183	0	0.00	0	0.00
0	21	184	0	0.00	0	0.00
0	21	185	0	0.00	0	0.00
0	21	186	0	0.00	0	0.00
0	21	187	0	0.00	0	0.00
0	21	188	0	0.00	0	0.00
0	24	189	0	0.00	0	0.00
0	24	190	0	0.00	0	0.00
0	24	191	0	10.73	0	1.34
0	24	192	0	0.00	0	0.00
0	24	193	0	0.00	0	0.00

Table 3.5: Pre-fish, negative controls, taken over all of the tanks. The calculations are taken over the eight technical replicates.

Table 3.5 summarizes the pre-fish negative controls. Samples were taken prior to the start of the experiment from the four main experiment tanks (tanks 19, 20, 21 and 24).

3.2 Density Plots

We now visualize relationships between TCT and number of fish in each tank. Recall that for each sample replicate, there are eight technical replicates.

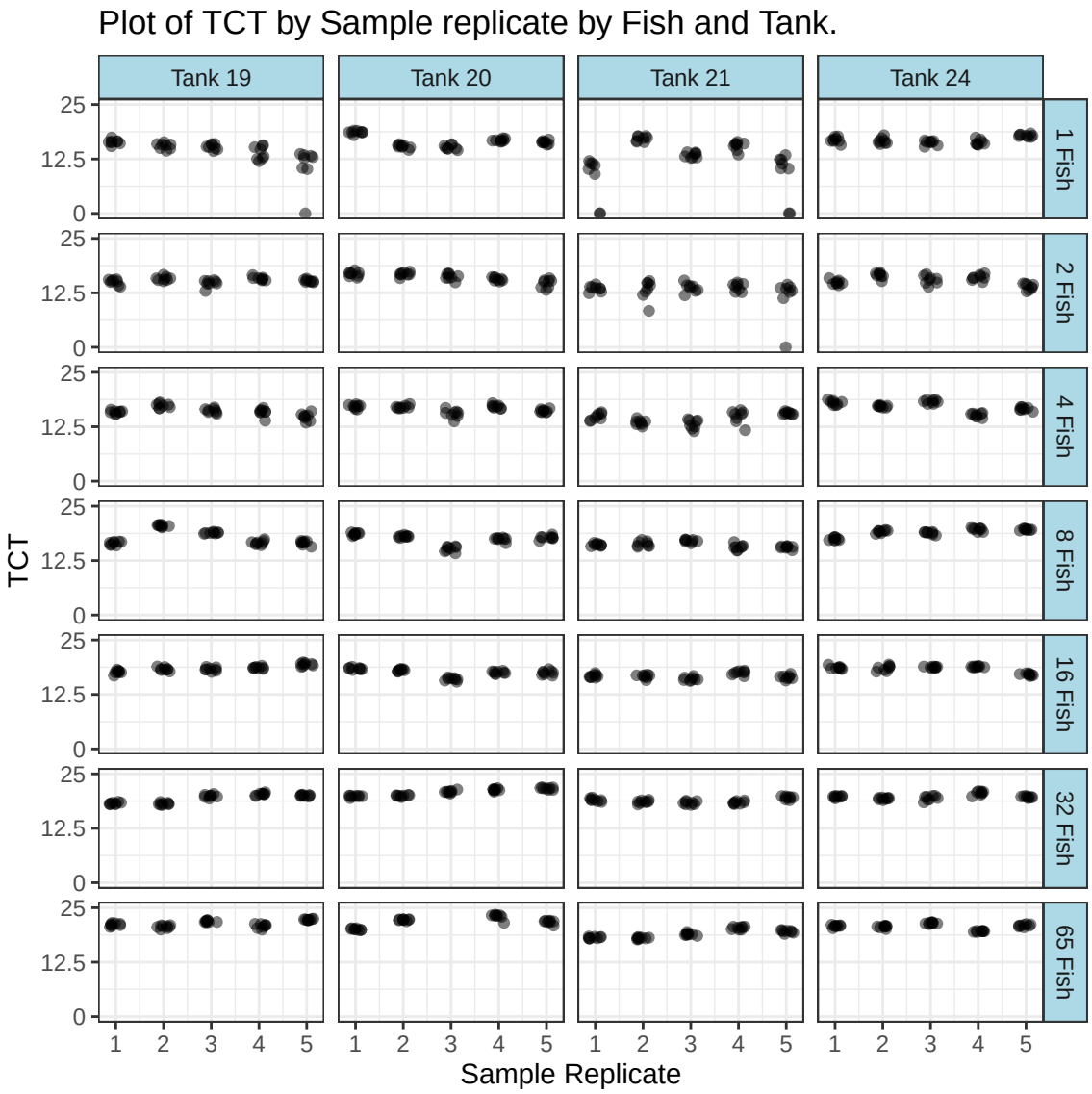

Figure 3.4: Exploratory plots that plot the TCT for each number of fish in each tank. Each sample replicate has 8 technical replicates, each technical replicate has a TCT.

Figure 3.4 shows the associated TCT values for each technical replicate. In general,

we notice TCT tends to increase as the number of fish increases. Although we have a relatively consistent increase in TCT as number of fish increases, there are several outliers. In particular, there are outliers for 1 Fish, Tank 19 and for 1 Fish, Tank 21. This may indicate that the presence of eDNA may not always be picked up in the lab. Although the sample had several technical replicates which did contain eDNA, some from the same sample replicate did not. As the density of fish increased, we see that we no longer have any major outliers, and the test for eDNA did not fail to pick up the presence of Coho eDNA.

We also plot TCT values for the sample replicates taken for zero fish. These are negative controls corresponding to sort codes 162-193 in Table 5. They were taken over a variety of tanks and days. The plots for tanks 1, 2, 3 and 4 correspond to the pilot experiment and give an idea of hatchery signal. The plots for tanks 19, 20, 21 and 24 are from 'true' pre-fish negative controls for the density experiment.

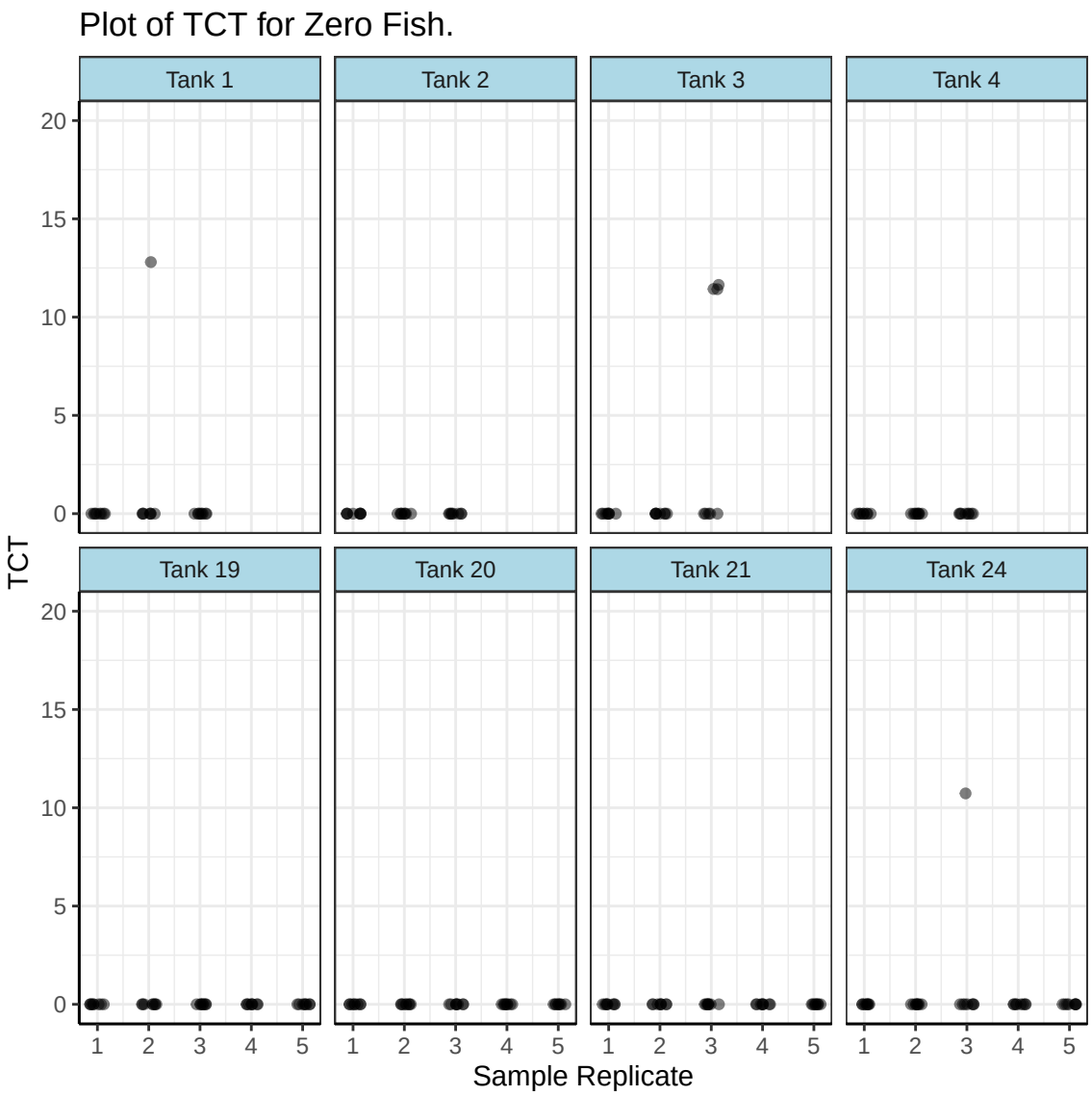

Figure 3.5: Plots of TCT for zero fish (Negative Controls).

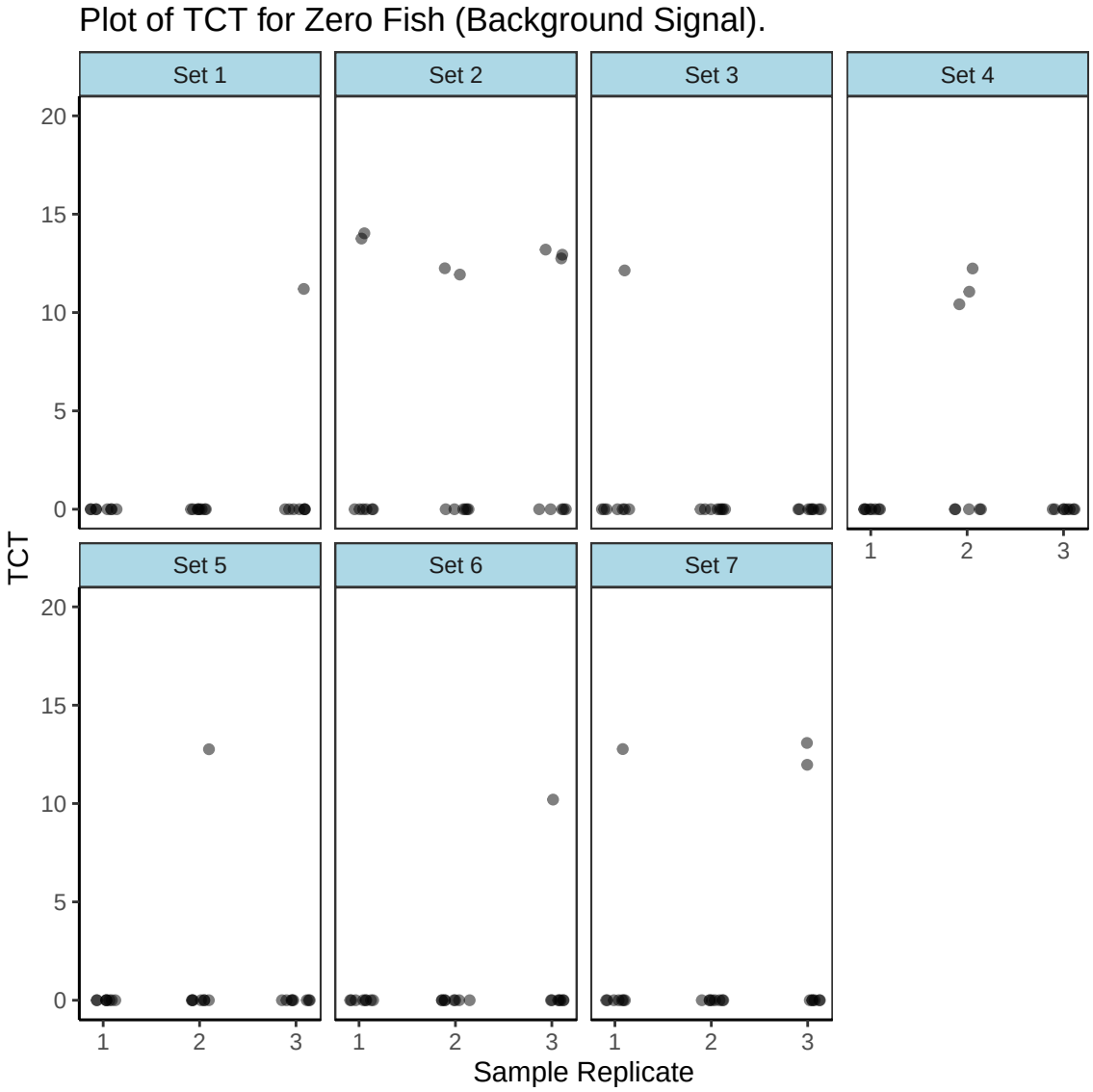

Figure 3.6: Additional plots of TCT for zero fish taken from tank 1.

Figure 3.5 and Figure 3.6 are both plots of TCT values obtained from tanks in which no fish were present. Figure 3.5 contains the negative controls for the main experiment, as well as TCT values obtained from the pilot experiment. Figure 3.6 contains additional plots of TCT measurements taken from the pilot experiment. In general, the hatchery water appears to contain a small background signal of Coho.

3.3 Models for Density (Median)

We fit a variety of competing linear models using the 'lm' function in R (R Core Team, 2013). The 'lm' function fits a 'linear model'. We use the median transformed CT of the technical replicates as our response variable. Initially, we fit a simple linear model, l.one.line, where we fit a least squares line of best fit for our response variable by considering only a single input predictor, log2(biomass). To plot models, we used the base R plot function and the ggplot2 package (Wickham, 2016), as well as the 'ggthemes' package (Arnold, 2019) and the 'latex2exp' package(Meschiari, 2015). Next, we fit two 'multiple linear models'. lm.tfac is a multiple linear model with the two predictors, log2(biomass) and tank. We also fit lfull.tfac, which incorporates both log2(biomass) and tank and also the interaction between log2(biomass) and tank.

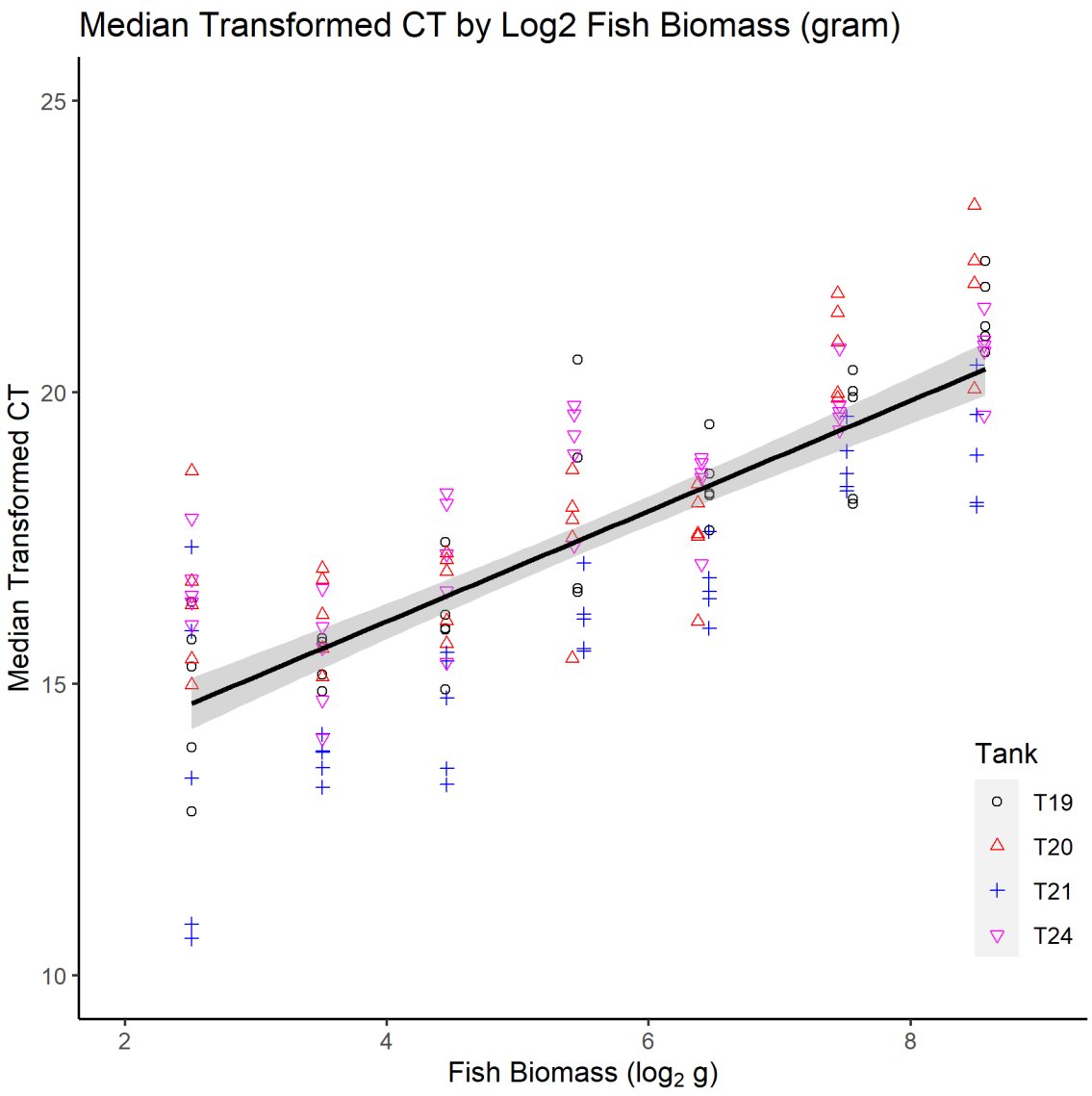

Figure 3.7: Median Transformed CT versus Log2(biomass). The fitted regression line, l.one.line and the 95 % confidence intervals are included. Each of the four tanks has an associated color and shape.

```
Call:
lm(formula = TCTmed ~ l2biom,
data = eco.sum.out.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0316 -0.9273  0.1096  0.9814  3.9834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.28782    0.36288   33.86   <2e-16 ***
l2biom       0.94633    0.06244   15.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.466 on 137 degrees of freedom
Multiple R-squared:  0.6264,Adjusted R-squared:  0.6237
F-statistic: 229.7 on 1 and 137 DF,  p-value: < 2.2e-16
```

Table 3.6: Summary of our first simple model, l.one.line. This model only includes an intercept and a biomass term.

Table 3.6 is a summary of a simple linear model (l.one.line) fit using the lm function. The estimated intercept of the model is 12.288 with a standard error of 0.362. The model also produces an estimate for the slope of 0.946 with a standard error of 0.0624. The p-value for both estimates is very small. The p-values are in the $Pr(> |t|)$ column. The p-value refers to the probability that we observe what we did, given that the null hypothesis is true. The null hypothesis in this case is that the associated parameter is zero. Since the p-value is so small, we can reject the null hypothesis and conclude that those terms are significantly different from zero.

Figure 3.7 confirms visually that as fish biomass increases, log2(biomass) in particular, median TCT also increases. This makes sense, as with more fish we would expect more residual eDNA, and thus a lower CT score (and hence a higher TCT score). The R^2 is 0.6264 which means that this simple model already does a relatively good job at explaining variation in the data.

We also plot regression lines obtained by considering each tank on its own and allowing the intercept to vary.

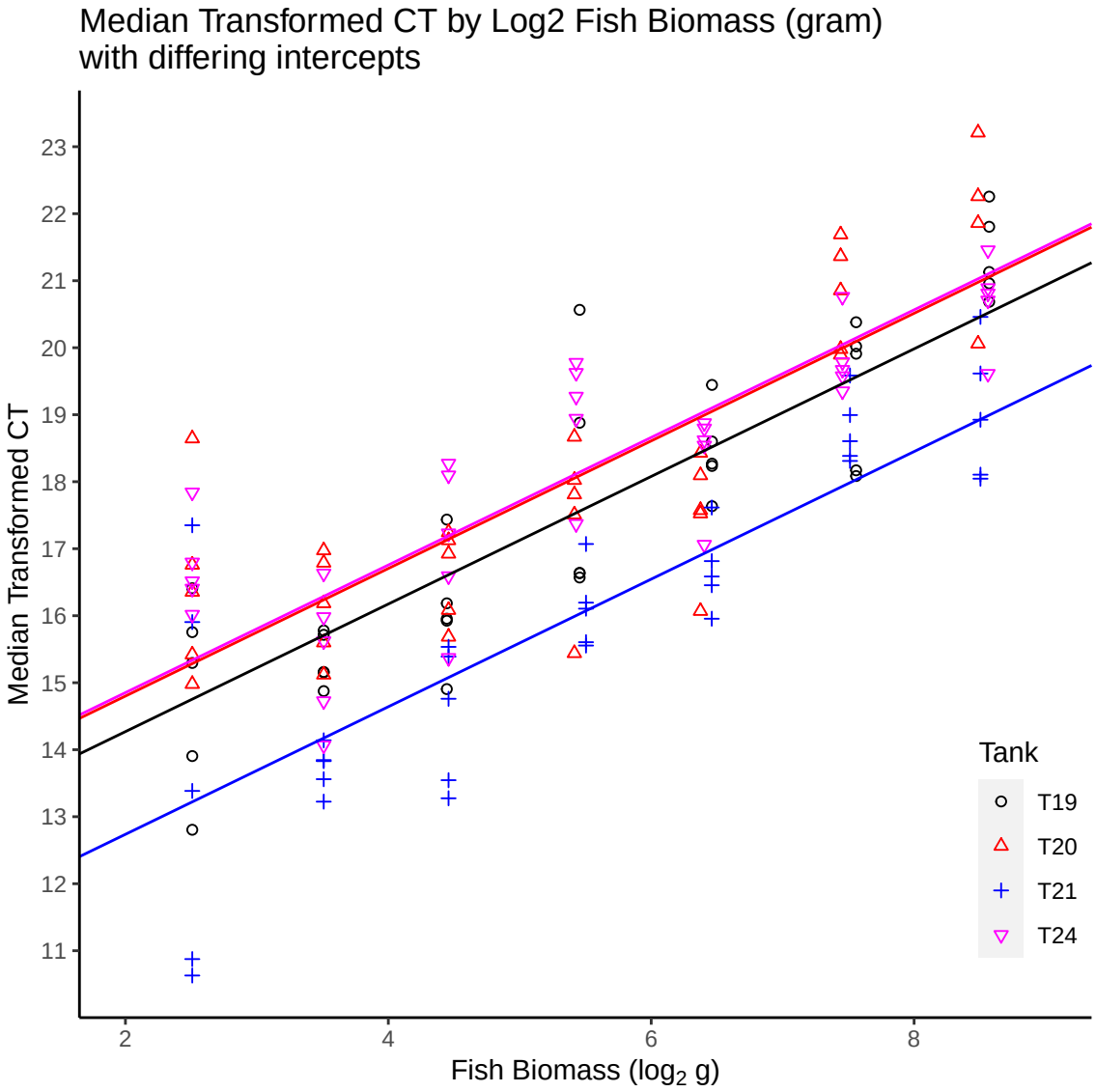

Figure 3.8: Lines of best fit for Median Transformed CT versus Log2(biomass) for each specific tank. This is equivelent to the model lm.tfac.

Figure 3.8 is the plot of the model lm.tfac. In this model, only the intercept is allowed to vary for each tank, but the slope remains the same. We see again tank 21 has a siginficantly lower intercept than the other tanks.

```
Call:
lm(formula = TCTmed ~ l2biom + tankF,
data = eco.sum.out.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8959 -0.6706 -0.0863  0.7260  4.1286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.36587    0.34637  35.701  < 2e-16 ***
l2biom       0.95211    0.05106  18.645  < 2e-16 ***
tankF20      0.53062    0.28873   1.838   0.0683 .
tankF21     -1.53280    0.28655  -5.349 3.71e-07 ***
tankF24      0.58199    0.28655   2.031   0.0442 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.199 on 134 degrees of freedom
Multiple R-squared:  0.7558,Adjusted R-squared:  0.7485
F-statistic: 103.7 on 4 and 134 DF,  p-value: < 2.2e-16
```

Table 3.7: This model, lm.tfac considers each tank as a predictor.

Table 3.7 provides a summary for the model lm.tfac. This model has a common slope parameter for log2(biomass), but allows the intercept to change for each tank. The intercept changes depending on which tank the point belongs to according to the estimates in Table 3.7.The baseline for tank comparisons is tank 19. Compared to tank 19, tank 20 and tank 24 appear to result in slightly higher median TCT values. This can be seen by the coefficients that are greater than 0, which means for those tanks we have in increase in median TCT compared to our baseline tank. For tank 24, the p-value is 0.0442, which indicates significance. However, we see a clear effect in Tank 21, which results in a lower median TCT compared to the other tanks, with a very small p-value of 3.71e-07 indicating strong significance. One possible explanation for this is that tank 21 obtained a more complete bleaching than the other tanks, hence resulting in less eDNA when sampled. This would cause a higher CT and hence a lower TCT. The adjusted R^2 for this model is 0.746, which is much larger than the adjusted R^2 of l.one.line (0.623). This means that our model does a good job at explaining the variation in the data. The adjusted R^2 contains a penalty

term for the number of predictors in the model. Note again that points belonging to tank 21 appear to have lower median transformed CT than the other tanks. This is possibly due to bleach in the tank.

Finally, we also consider a model that includes the possible interactions between tank number and biomass. This model allows for a different slope and intercept for each tank and also allows for interactions between tank and biomass. This is the model plotted in Figure 3.8, and we will see that results obtained from this model are equivelent to considering each tank in isolation.

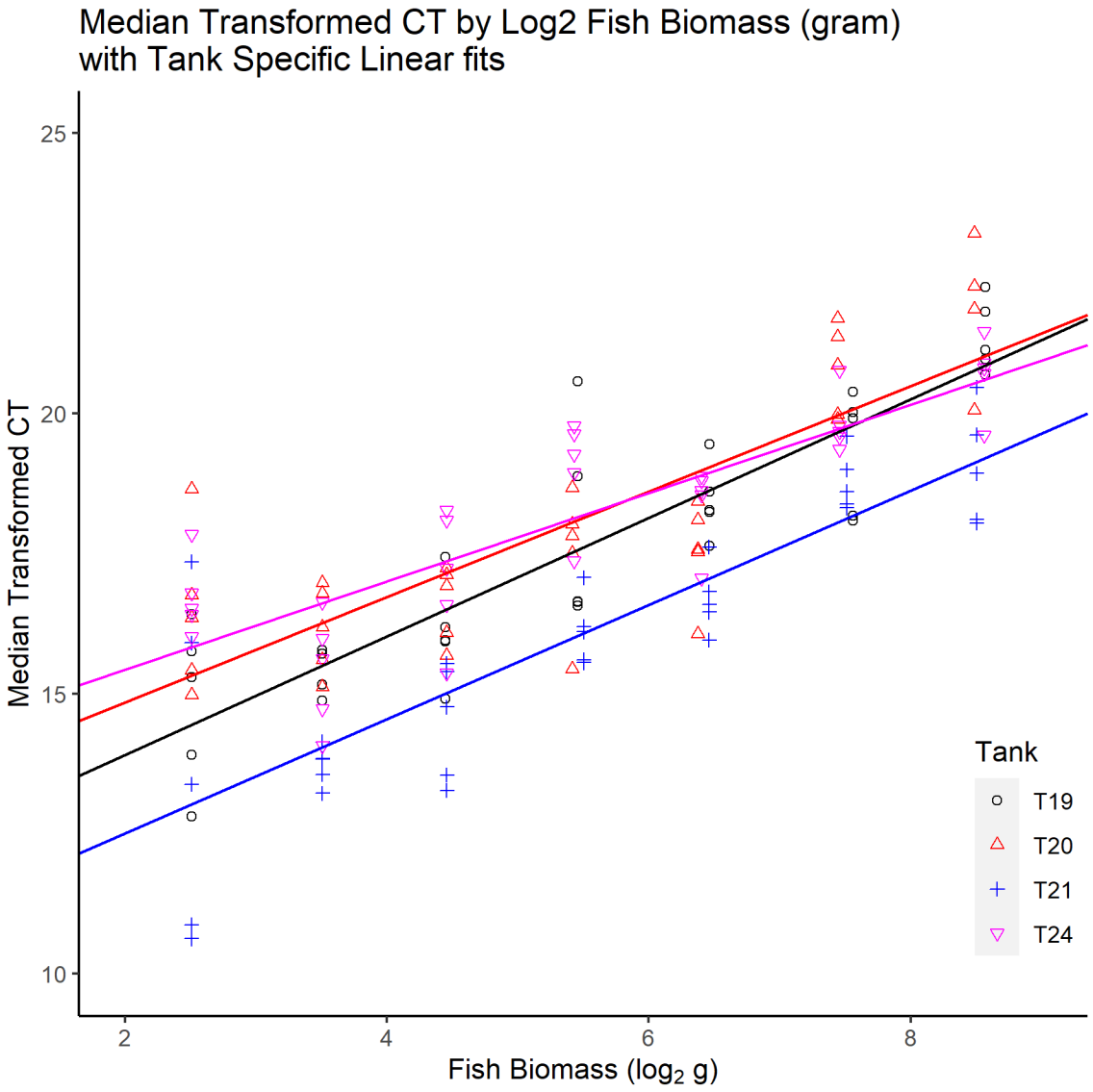

Figure 3.9: Lines of best fit for Median Transformed CT versus Log2(biomass) for each specific tank. This is equivelent to the model lfull.tfac.

Figure 3.9 is a plot of regression lines obtained by applying regression over each individual tank. In this regression, each tank has its own intercept and own slope.

Figure 3.9 seems to indicate that tank likely impacts the result of median TCT. Thus we now fit a model that includes tank as a predictor. We summarize these in the following table:

Summary of Tank Specific Simple Linear Regressions

Tank	Intercept	Slope	R^2
19	11.786	1.057	0.806
20	12.958	0.941	0.671
21	10.463	1.019	0.725
24	13.849	0.788	0.699

Table 3.8: Table summarizing simple linear regression on log2(biomass) when each tank is considered in isolation for median TCT.

Table 3.8 summarizes the results of applying simple linear regression to each tank in isolation. We see we obtain separate estimates for each intercept and slope. Tank 21 has the smallest intercept, while tank 24 has the largest intercept. The slopes are similar, however tank 24 has a noticeably smaller slope. The R^2 are all quite high, but tank 19 alone has the highest R^2.

```
Call:
lm(formula = TCTmed ~ l2biom + tankF + l2biom * tankF,
data = eco.sum.out.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8844 -0.6266 -0.0271  0.7038  4.3296

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.78627    0.58455  20.163   <2e-16 ***
l2biom          1.05747    0.09973  10.603   <2e-16 ***
tankF20         1.17132    0.83935   1.396   0.1652
tankF21        -1.32314    0.83023  -1.594   0.1134
tankF24         2.06260    0.82886   2.488   0.0141 *
l2biom:tankF20 -0.11674    0.14522  -0.804   0.4229
l2biom:tankF21 -0.03802    0.14182  -0.268   0.7891
l2biom:tankF24 -0.26992    0.14181  -1.903   0.0592 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.193 on 131 degrees of freedom
Multiple R-squared:  0.7634,Adjusted R-squared:  0.7508
F-statistic: 60.39 on 7 and 131 DF,  p-value: < 2.2e-16
```

Table 3.9: A model, lfull.tfac, that allows for interactions between all terms.

Table 3.9 provides a summary for the full model (lfull.tfac) that includes interactions. We obtain similar estimates to Table 3.7. However, we now have included interactions. The interaction terms are not statistically significant. As well, the adjusted R^2 is only slightly increased to 0.7508. Thus, the interaction between tank and biomass in our modelling is not warranted. This is further evidence that accounting for biomass as a predictor in insolation may be warranted. Notice we obtain identical results to those given in Table 3.8.

We apply ANOVA to compare the current models:

```
Analysis of Variance Table

Model 1: l.one.line
Model 2: lm.tfac
Model 3: lfull.tfac

Model 1: TCTmed ~ l2biom
Model 2: TCTmed ~ l2biom + tankF
Model 3: TCTmed ~ l2biom + tankF + l2biom * tankF
  Res.Df    RSS Df Sum of Sq       F   Pr(>F)
1    137 294.51
2    134 192.54  3   101.970 23.8738 2.17e-12 ***
3    131 186.51  3     6.034  1.4128    0.242
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3.10: ANOVA to compare l.one.line, lm.tfac and lfull.tfac.

Table 3.10 provides the results of applying ANOVA on the three previously models. We are testing the significance of the three models sequentially. Model 3, lfull.tfac allows for both the slopes and intercepts to change and it also allows for interactions between the terms. Model 2, lfull.tfac, has constant slopes but differing intercepts. Hence the test of model 3 versus model 2 is a test of the hypothesis of a common slope versus different slopes. Since we have a large p-value (0.242), we can conclude that the additional slope parameters are not significantly different from zero. The comparison between model 2 and model 1 is the test of non-differing intercepts for each tank. Because the p-value for this test is so low, we reject this hypothesis. That is, since the p-value (2.17e-12) is so small, we safely reject the null hypothesis. Hence, the effect of tank (and hence differing intercepts) is significant and should be included in any modeling.

We now fit a model similar to l.one.line. However, we now collapse over each tank by taking the median value of TCT for each tank. This model is called l.tankregression.med.

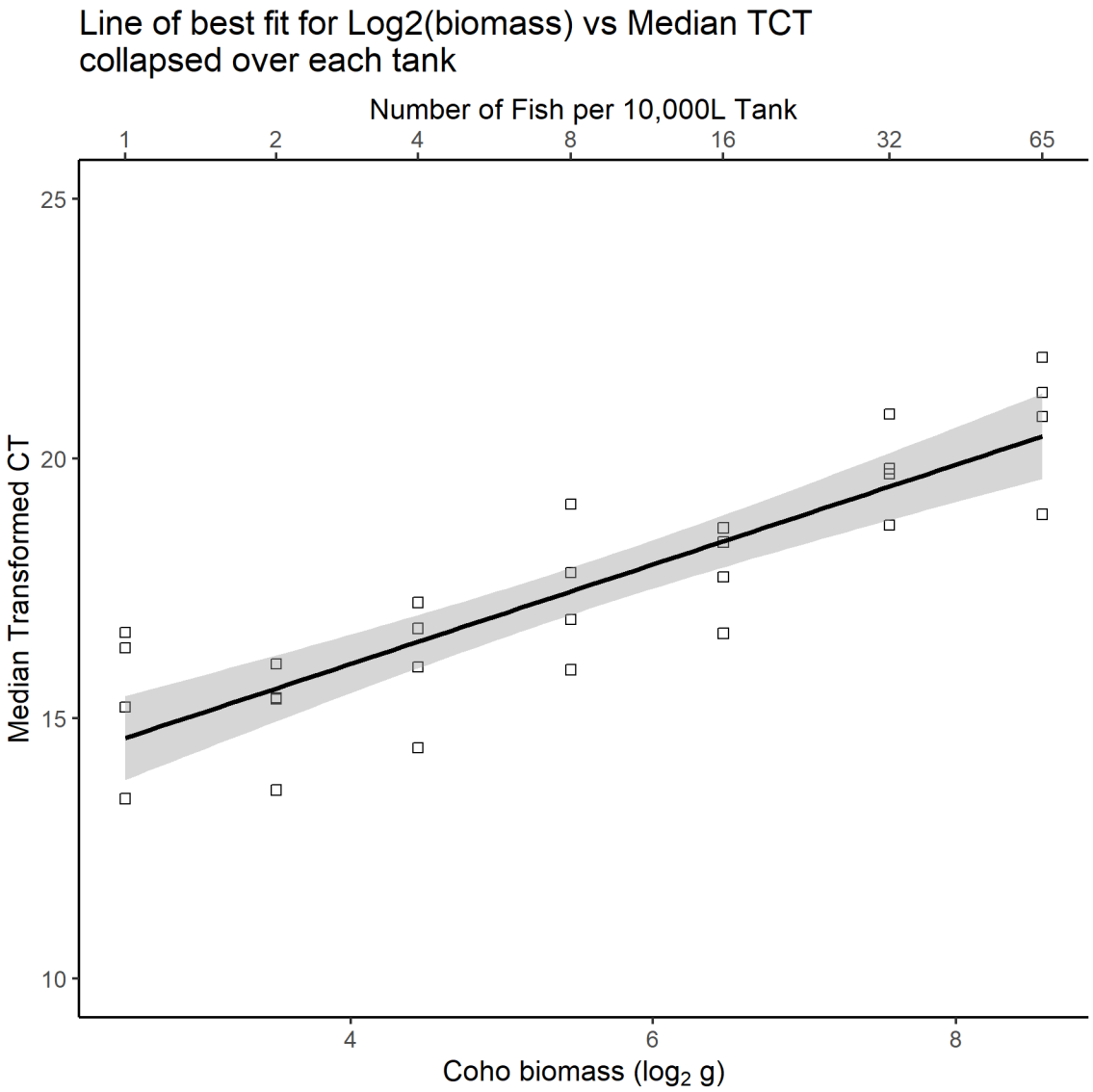

Figure 3.10: Regression line showing the relationship between biomass and Median TCT. Included are the confidence bands about the regression line. The line is the model l.tankregression.med. The R^2 is 0.748. Points shown represent the Median TCT for each of the four tanks for each of the seven unique numbers of fish.

```
Call:
lm(formula = medV ~ l2biomTank, data = new.tankmed)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0475 -0.7029  0.2458  0.6363  2.0326

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2194     0.6388  19.130  < 2e-16 ***
l2biomTank    0.9580     0.1090   8.791 2.89e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.166 on 26 degrees of freedom
Multiple R-squared:  0.7482,Adjusted R-squared:  0.7386
F-statistic: 77.27 on 1 and 26 DF,  p-value: 2.89e-09
```

Table 3.11: Model: ltankregression.med

Figure 3.10 is a plot made of our final model, ltankregression.med. Each point represents the Median TCT for each of the four main tanks. Each distinct biomass value corresponds to a unique number of fish. The results of this final model are good, as we have quite a large R^2 value. Table 3.11 summarizes the result of taking the median of TCT over each tank and fitting a simple regression model. The only independent variable considered is log2 of Coho biomass. By collapsing over each tank, we are able to capture a large portion of the variation in our data.

3.4 Models for Density (Mean)

We now consider modeling of the response variable mean TCT. Previously, we considered median TCT. We take an identical approach to the previous models. That is, we first create a simple linear model for predicting the response mean TCT. That is, we create a model l.one.line.mean that only considers log2(biomass) as a predictor.

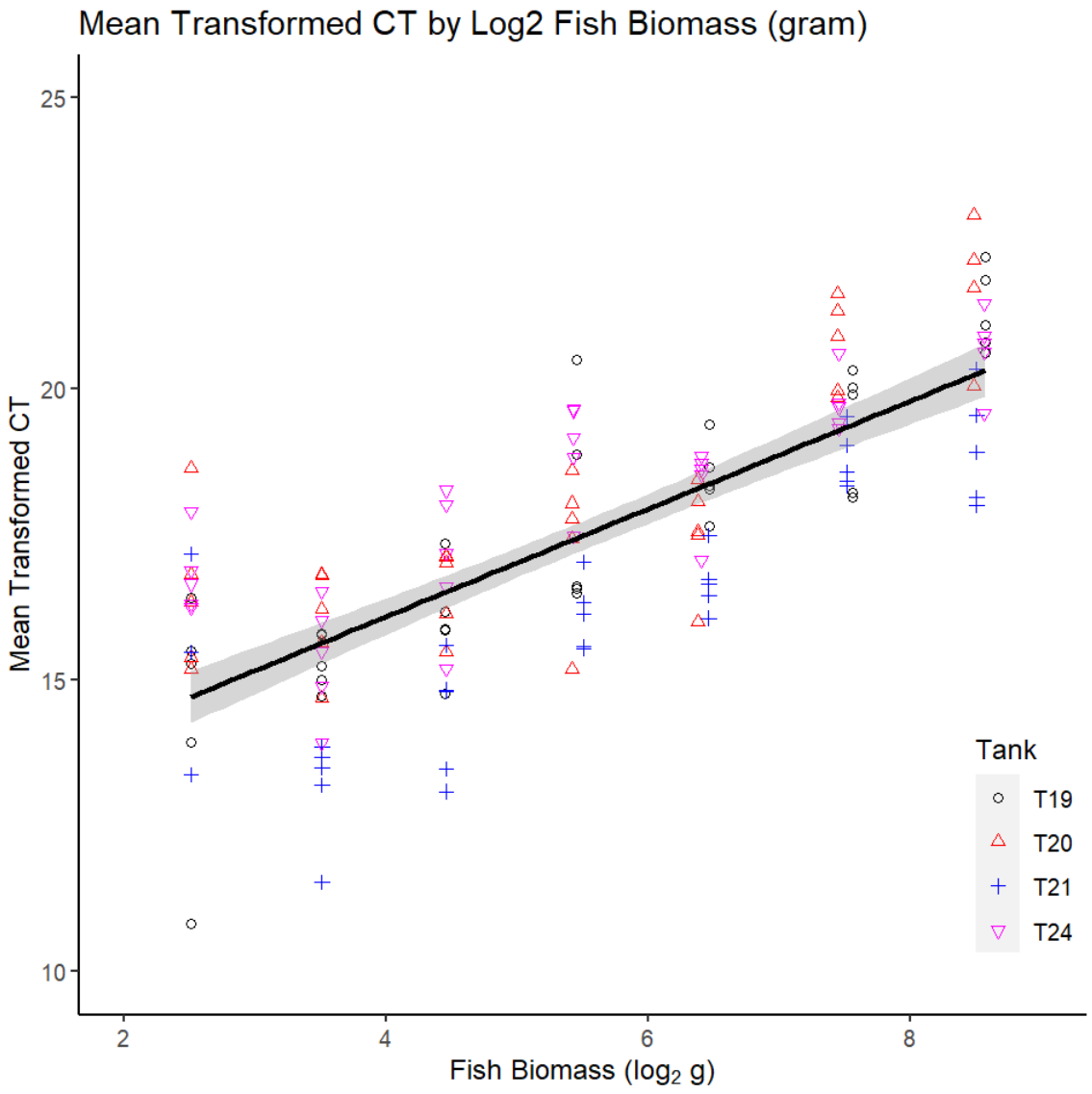

Figure 3.11: Mean TCT versus Log2(biomass). The fitted regression line, l.one.line.mean and the 95 % confidence intervals are included. Each of the four tanks has an associated color and shape

Figure 3.11 plots l.one.line.mean and the associated 95 % confidence bands. Again, we see that the mean TCT from tank 21 tends to be less than the mean TCT from the other three tanks. In general, mean TCT increases as log2(biomass) increases.

```
Call:
lm(formula = TCTmean ~ l2biom,
data = eco.sum.out.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2567 -0.8728  0.1739  1.0170  4.2183

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.93296    0.40431   29.51   <2e-16 ***
l2biom       0.99264    0.06957   14.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.634 on 137 degrees of freedom
Multiple R-squared:  0.5978,Adjusted R-squared:  0.5948
F-statistic: 203.6 on 1 and 137 DF,  p-value: < 2.2e-16
```

Table 3.12: Model: loneline.mean. Simple linear regression for mean TCT which only considers log2(biomass).

Table 3.12 provides a summary of the simple linear model, l.one.line.mean. The estimates obtained are very similar to the estimates in Table 3.6.

Next, we consider a model for mean TCT with biomass as a predictor, and also consider the impact of tank. We first plot regression lines obtained by considering a model with a common slope but allow for differing intercepts.

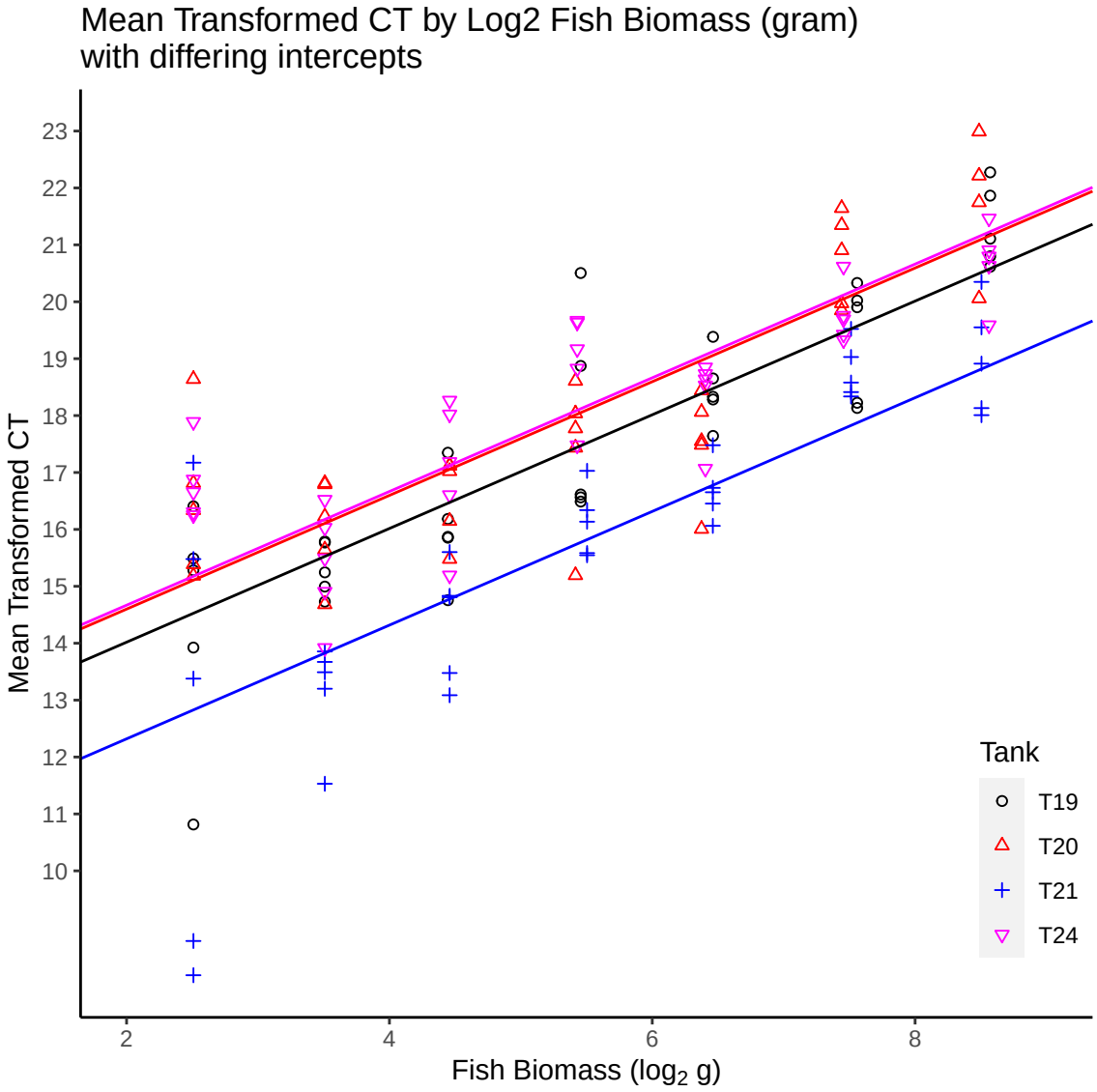

Figure 3.12: Lines of best fit by allowing intercept to differ over each tank. This is a plot of the model lm.tfac.mean.

```
[1] "Model: lmtfac.mean"

Call:
lm(formula = TCTmean ~ l2biom + tankF,
data = eco.sum.out.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6614 -0.6205  0.0259  0.7544  4.3449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.02003    0.38709  31.053  < 2e-16 ***
l2biom       0.99900    0.05707  17.506  < 2e-16 ***
tankF20      0.58147    0.32266   1.802   0.0738 .
tankF21     -1.69833    0.32023  -5.304 4.57e-07 ***
tankF24      0.64966    0.32023   2.029   0.0445 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.34 on 134 degrees of freedom
Multiple R-squared:  0.7354,Adjusted R-squared:  0.7275
F-statistic: 93.13 on 4 and 134 DF,  p-value: < 2.2e-16
```

Table 3.13: Model:lm.tfac.mean. A model that allows for differing intercepts depending on which tank a sample came from.

Table 3.13 summarizes the result of lm.tfac.mean, which is a model including the tank as a predictor. The estimated intercept is 12.02 for lm.tfac.mean, while the estimated intercept for the median TCT counterpart, lm.tfac was 12.365, which can be seen in Table 3.7. R again indicates that effect of tank 21 is highly significant with an estimate of -1.698. That is, tank 21 produces much smaller mean TCT values compared to the other tanks.

Finally, we build a full model, lfull.tfac.mean, that includes biomass, tank and the interactions between biomass and tank.

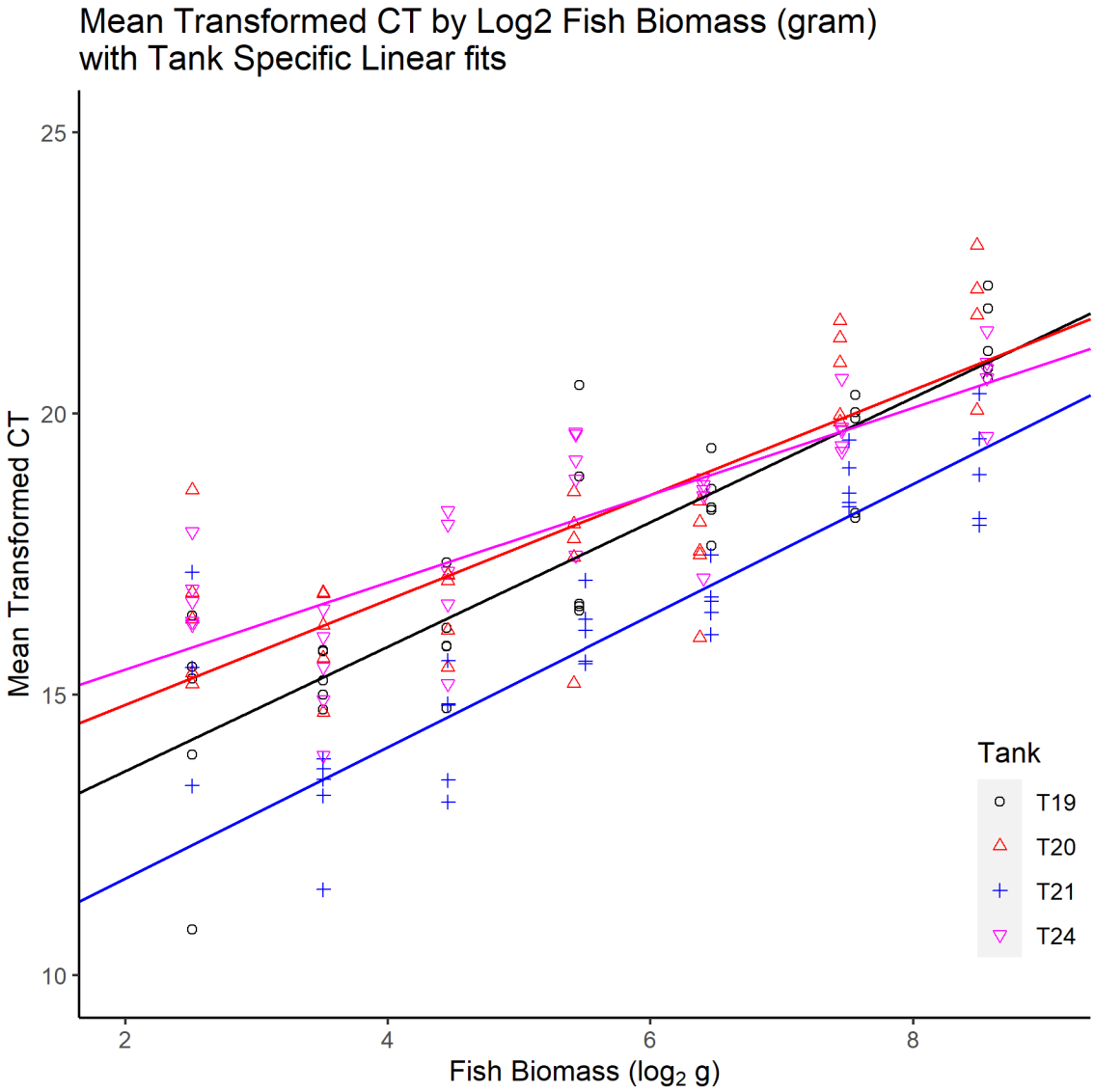

Figure 3.13: Lines of best fit by allowing intercept and slope to differ over each tank. This is a plot of the model lfull.mean.

Summary of Tank Specific Simple Linear Regressions

Tank	Intercept	Slope	R^2
19	11.423	1.107	0.785
20	12.955	0.933	0.660
21	9.377	1.171	0.696
24	13.888	0.776	0.690

Table 3.14: Table summarizing simple linear regression on log2(biomass) when each tank is considered in isolation for mean TCT.

Table 3.14 summarizes the results of applying simple linear regression to each tank in isolation. We see we obtain separate estimates for each intercept and slope. Tank 21 has the smallest intercept, while tank 24 has the largest intercept. The slopes are similar, however tank 24 has a noticeably smaller slope. The R^2 are all quite high, but tank 19 alone has the highest R^2.

```
[1] "Model: lfull.mean"

Call:
lm(formula = TCTmean ~ l2biom + tankF + l2biom * tankF,
data = eco.sum.out.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1479 -0.5807  0.0048  0.7632  4.8583

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     11.4232     0.6451  17.708  < 2e-16 ***
l2biom           1.1075     0.1101  10.063  < 2e-16 ***
tankF20          1.5322     0.9263   1.654  0.10049
tankF21         -2.0464     0.9162  -2.234  0.02721 *
tankF24          2.4651     0.9147   2.695  0.00796 **
l2biom:tankF20  -0.1744     0.1603  -1.088  0.27840
l2biom:tankF21   0.0635     0.1565   0.406  0.68561
l2biom:tankF24  -0.3311     0.1565  -2.116  0.03628 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.317 on 131 degrees of freedom
Multiple R-squared:  0.7501,Adjusted R-squared:  0.7367
F-statistic: 56.17 on 7 and 131 DF,  p-value: < 2.2e-16
```

Table 3.15: Model: lfull.mean.

Figure 3.13 is the plot of the model lfull.mean. The intercept and slope is allowed to vary over each tank. Table 3.15 provides the summary of lfull.mean. For the full model with mean TCT as the response and both tank and the interaction between tank and biomass as predictors, the least squares estimate of the intercept term is 11.423. For the full model on median TCT, the least squares estimate of the intercept is 11.786. This model provides identical results to considering each tank in isolation, as seen by comparing the estimates to those given in Table 3.14.

We again create an ANOVA table:

```
[1] "Model 1: l.one.line.mean"
[1] "Model 2: lm.tfac.mean"
[1] "Model 3: lfull.mean "
Analysis of Variance Table

Model 1: TCTmean ~ l2biom
Model 2: TCTmean ~ l2biom + tankF
Model 3: TCTmean ~ l2biom + tankF + l2biom * tankF
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1    137 365.60
2    134 240.47  3   125.134 24.0559 1.824e-12 ***
3    131 227.15  3    13.322  2.5611    0.0577 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3.16: ANOVA table for Mean TCT models.

Table 3.16 shows the results of applying an ANOVA test on the three models for Mean TCT. The interpretation is the same as it was with the anova table for median TCT. Since lm.tfac.mean has a very small p-value, we conclude that we need to include the tank effect in our model (that is, we reject the hypothesis that differing intercepts is zero). Since the p-value associated with lfull.mean is 0.0577. This indicates that including interaction for modelling mean TCT is marginally significant. The R^2 for lm.tfac.mean is 0.735 while the R^2 for lfull.mean is 0.750. Although the R^2 increases, it is only by a small amount. Since the p-value is on the boundary of significance, we may still choose to ignore the interaction term as it adds complexity and only slightly increases the R^2.

We now fit a model similar to l.one.line.mean. However, we now collapse over each tank by taking the mean value of TCT. This model is called l.tankregression.

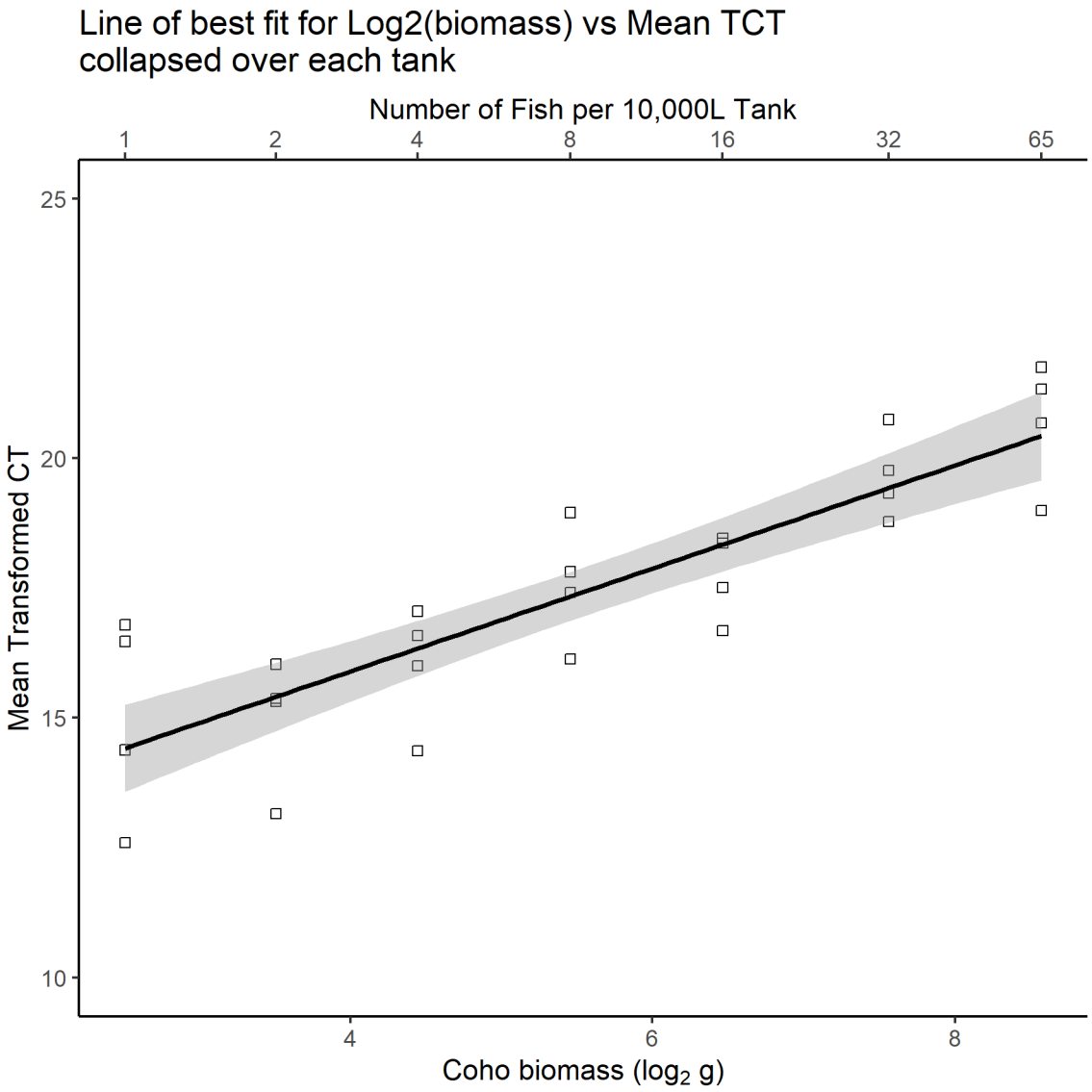

Figure 3.14: Regression line showing the relationship between Mean TCT and biomass. Included are the confidence bands about the regression line. The line is the model l.tankregression. The R^2 is 0.748. Points shown represent the Mean TCT for each of the four tanks for each of the seven unique numbers of fish.

```
Call:
lm(formula = medV ~ l2biomTank, data = new.tankmed)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0475 -0.7029  0.2458  0.6363  2.0326

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2194     0.6388  19.130  < 2e-16 ***
l2biomTank    0.9580     0.1090   8.791 2.89e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.166 on 26 degrees of freedom
Multiple R-squared:  0.7482,Adjusted R-squared:  0.7386
F-statistic: 77.27 on 1 and 26 DF,  p-value: 2.89e-09
```

Table 3.17: Model: l.tankregression

Table 3.17 summarizes the result of taking the mean of TCT over each tank and fitting a simple regression model. The only variable considered is the Coho biomass. Figure 3.14 is the regression line obtained by applying simple linear regression over the points collapsed by taking the mean over each tank. The R^2 is high, indicating our model does a good job at explaining variation in the data. Moreover, it has a simple interpretation (Mean TCT increases linearly with log2(biomass) and thus reduces complexity.

Figure 3.14 is a plot made of our final model. Each point represents the Mean TCT for each of the four main tanks. Each distinct biomass value corresponds to a unique number of fish.

3.5 Robust Models

We model using a robust fit from the 'robust' package, and the lmRob function (Wang et al., 2020). This fits a 'robust model' which is less sensitive to outliers than a regular least squares, linear fit. The function lmRob automatically chooses an appropriate algorithm to efficiently fit a model.

Again, we use median TCT as our response variable. Initially, we fit a 'simple robust model', lr, to our response variable, Median TCT, with a single input predictor, log2(biomass). Next, we fit two 'multiple linear models'. lrparallel.tfac is a multiple linear model with the two predictors log2(biomass) and tank. Finally, we fit lrfull.tfac, that incorporates both log2(biomass) and tank as well as the interaction between log2(biomass) and tank.

```
Call:
robust::lmRob(formula = eco.md ~ l2biom,
data = eco.sum.out.dat)

Residuals:
     Min       1Q   Median       3Q      Max
-4.12508 -0.97680  0.09877  0.95163  3.88992

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.4349     0.3783   32.87   <2e-16 ***
l2biom        0.9250     0.0646   14.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.508 on 137 degrees of freedom
Multiple R-Squared: 0.5464

Test for Bias:
            statistic p-value
M-estimate     0.5861   0.746
LS-estimate   -3.2261   1.000
```

Table 3.18: Model: lr

Table 3.18 provides the results of fitting a simple linear model for median TCT using lmRob. For a unit increase in log2(biomass) this model predicts an increase in

0.9250 for median TCT. This also makes sense, as log2(biomass) increases, we expect that the volume of eDNA in the tank will increase, and hence CT will decrease (and TCT will increase).The p-value for both intercept and log2(biomass) are extremely low, indicating that both play an important role in the model. The estimates for the intercept in both the robust and linear model are quite similar. The estimate for log2(biomass) is slightly less for the robust fit. We now fit a robust model for median TCT that includes biomass as a predictor and also considers tank as a possible predictor.

```
Call:
robust::lmRob(formula = eco.md ~ l2biom + tankF,
data = eco.sum.out.dat)

Residuals:
     Min       1Q   Median       3Q      Max
-3.01958 -0.56948 -0.02898  0.76353  4.58354

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.83864    0.33155  35.707  < 2e-16 ***
l2biom       1.03185    0.04907  21.027  < 2e-16 ***
tankF20      0.67319    0.27469   2.451   0.0155 *
tankF21     -1.66050    0.27129  -6.121 9.63e-09 ***
tankF24      0.59710    0.26948   2.216   0.0284 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.038 on 134 degrees of freedom
Multiple R-Squared: 0.6188

Test for Bias:
            statistic p-value
M-estimate      7.002 0.22047
LS-estimate    11.162 0.04827
```

Table 3.19: Model: lrparallel.tfac

Table 3.19 summarizes the model lrparallel.tfac. For this model, lmRob produces similar conclusions for both the intercept and for the log2(biomass). However, we now also have estimates for the tank. The data indicates that compared to the baseline of tank 19, tank 20 and 24 produce slightly higher median TCT estimates, while on

the other hand Tank 21 produces lower TCT. These differences are possibly due to differences in bleaching and cleaning of the tanks.

Finally, we fit a robust model (lrfull.tfac) for the response median TCT that includes biomass, tank, and the possible interactions between biomass and tank.

```
Call:
robust::lmRob(formula = eco.md ~ l2biom + tankF + l2biom * tankF,
    data = eco.sum.out.dat)

Residuals:
     Min       1Q   Median       3Q      Max
-3.02633 -0.55631  0.06038  0.73640  5.14548

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.68910    0.85192  13.721  < 2e-16 ***
l2biom          1.05892    0.14248   7.432 1.23e-11 ***
tankF20         0.74177    1.19365   0.621   0.5354
tankF21        -2.51491    1.21406  -2.071   0.0403 *
tankF24         2.34823    1.14755   2.046   0.0427 *
l2biom:tankF20 -0.01331    0.20598  -0.065   0.9486
l2biom:tankF21  0.14914    0.20416   0.731   0.4664
l2biom:tankF24 -0.29643    0.19354  -1.532   0.1280
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.054 on 131 degrees of freedom
Multiple R-Squared: 0.6495

Test for Bias:
            statistic p-value
M-estimate      3.963  0.8605
LS-estimate    11.736  0.1634
```

Table 3.20: Model: lrfull.tfac

Table 3.20 summarizes the estimates of lrfull.tfac. The robust model, lrfull.tfac containing log2(biomass), tank and their interaction terms produces similar results to lrparallel.tfac. The multiple R^2 for this model is only slightly larger than lrparallel.tfac at 0.6495. The interaction terms between tank and log2(biomass) are not significant.

We confirm this with an 'anova' test.

Recall that anova compares a variety of nested models. Our first model is lr, which included log2(biomass). Our second model, lrparallel.tfac includes for log2(biomass) and tank.

```
[1] "Model 1: lr"
[1] "Model 2: lrparallel"
[1] "Model 3: lrfull.tfac "

Response: eco.md
     Terms     Df RobustF     Pr(F)
[1,] Model 1:   1
[2,] Model 2:   1  3 25.8678 6.294e-06 ***
[3,] Model 3:   1  3  3.4587    0.3091
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3.21: Robust ANOVA

Table 3.21 summarizes the results of applying the ANOVA function to the three above models. We see that compared to model 1 (lr), model 2 (lrparallel) has a p-value of 6.294e-6, which is very small. We thus reject the null hypothesis that the tank effect is zero. On the other hand, for our third model lrfull.tfac, the p-value comparing lrfull.tfac versus lrparallel is 0.3091. We do not reject the hypothesis that the coefficient for the interaction term is zero.

3.6 Residual Analysis

We now consider the residuals for some of the models we previously fit. To create the residual plot for l.tankregression we used the 'broom' package (Robinson et al., 2020).

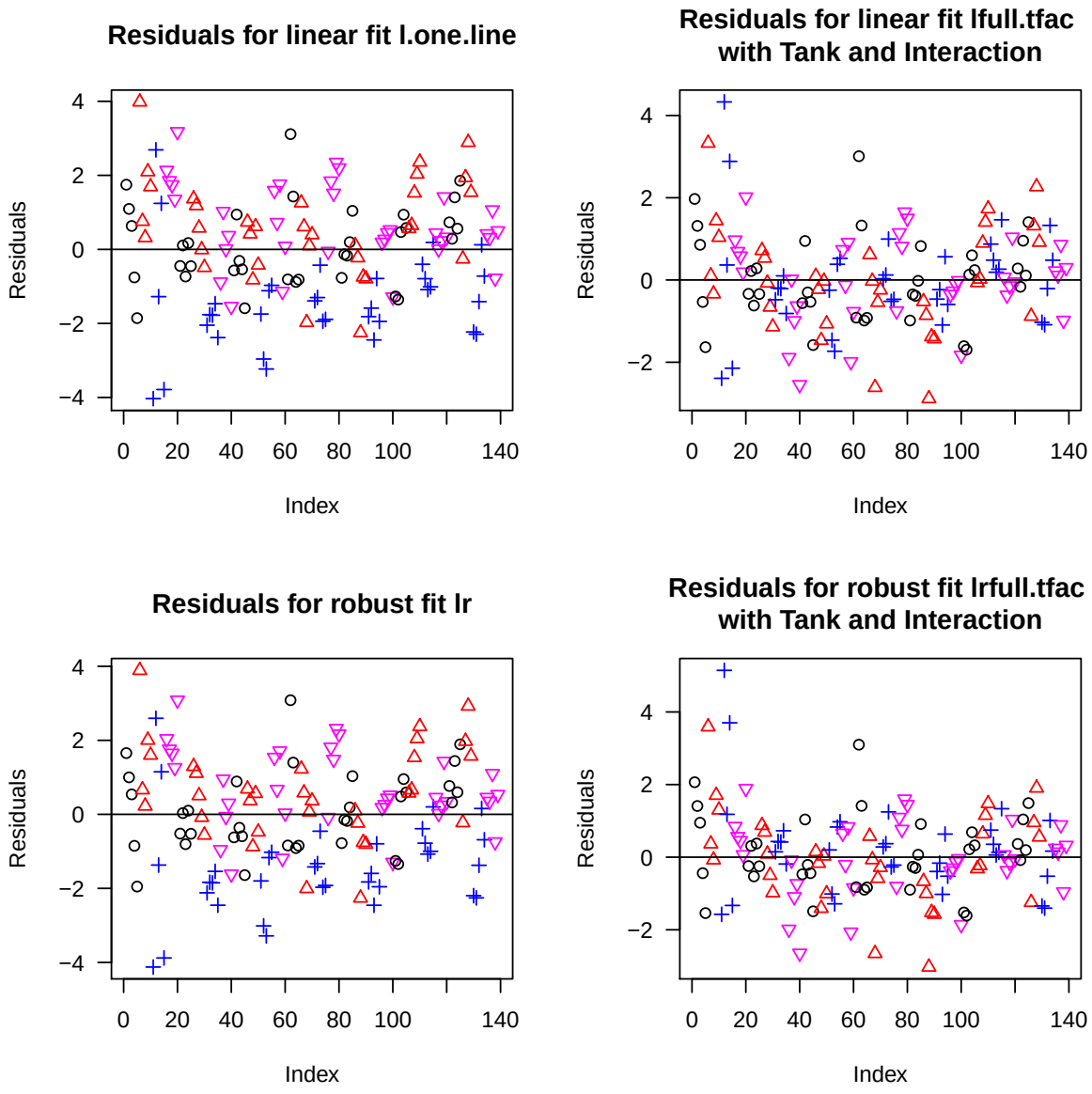

Figure 3.15: Residual plots for four models, l.one.line and lfull.tfac are standard linear models, while lr and lrfull.tfac are robust models. Tanks 19, 20, 21 and 24 are represented as black circles, red triangles, blue crosses and pink inverted triangles respectively.

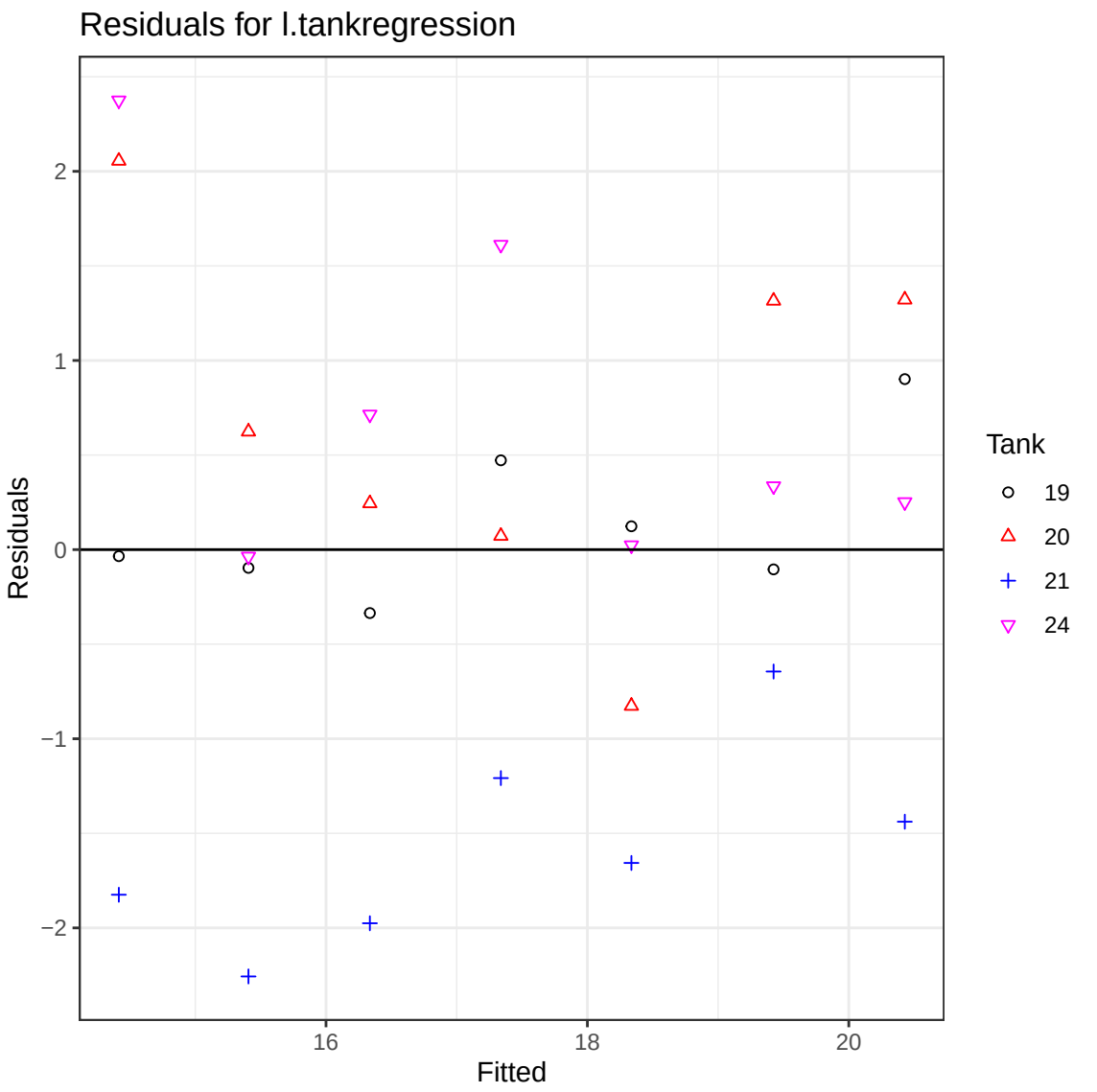

Figure 3.16: Residual plots for our final model for mean TCT, l.tankregression.

Figure 3.15 summarizes the residuals from four of the models we fit. The residuals appear to be randomly distributed about the x-axis which is a good indication that our model is making valid assumptions. Figure 3.16 is a plot of the residuals for l.tankregression. The residuals again appear to be randomly distributed about the x-axis. Although Tank 21 has consistently negative residuals. The legend indicates what tank each point belonged to.

3.7 Density Conclusions

The density experiment provided an excellent opportunity to validate the relationship between Coho density/biomass and eDNA accumulation. Using a Coho specific DNA primer, we were able to detect Coho at all levels of density, ranging from 1 fish to 65 fish. To model mean TCT of Coho eDNA, we chose models that contained biomass as a predictor. The impact of tank appeared to be due to human error in the cleaning of the tanks. Overall, we confirmed that Coho biomass is highly correlated with an increase in mean TCT. We fit both median and mean TCT, and it would be up to a researcher to decide which they prefer to work with. Robust models provided very similar estimates compared to their linear model counterparts and we thus chose to continue only with linear models.

One highlight of our analysis was the creation of Figure 3.14, which summarizes much of the results of the experiment in one concise image. Collapsing by taking the mean TCT value over each tank proved to be useful and provided a simple model, l.tankregression. This model performed well, evidenced by the high R^2 and the distribution of the model residuals (Figure 3.16).

In the density experiment, Coho eDNA was detected with 100% certainty at all densities ranging from 1 to 65 fish per tank. In regard to detection of technical replicates, perfect detection was achieved at all densities from 4 to 65 fish. For 2 fish, one technical replicate failed to detect Coho. For 1 fish, only three sample replicates indicated non-detection. We thus were also able to validate the performance and high levels of sensitivity of the eONKI4 DNA assay in the detection of Coho Salmon. A small pilot experiment also was used to gain insight regarding the background signal of Coho in the hatchery water. As seen in Figure 3.5 and Figure 3.6, the hatchery water contained a small but non-consistent amount of evidence of Coho eDNA.

Chapter 4

Dilution Experiment

4.0.1 Introduction

As discussed in Chapter 2, environmental covariates may directly influence eDNA persistence and collection. One such variable is the rate of flow, or 'current' of a water system. The goal of the 2016 'dilution' experiment was to investigate the relationship between flow rate and Coho eDNA collection in a controlled study. In the 'density' experiment of Chapter 3, Coho were added daily to tanks and measurements were taken throughout the week. In this experiment, three juvenile Coho were allowed to acclimate to four tanks (tanks 19, 20, 21 and 24) and were subsequently removed. After the Coho were removed, water was allowed to flow out of each tank at a known rate as it was replaced, or diluted, by hatchery water. Measurements were obtained at several intervals of varying levels of flow. Moreover, several control samples were taken from the hatchery kitchen sink and from the hatchery pond.

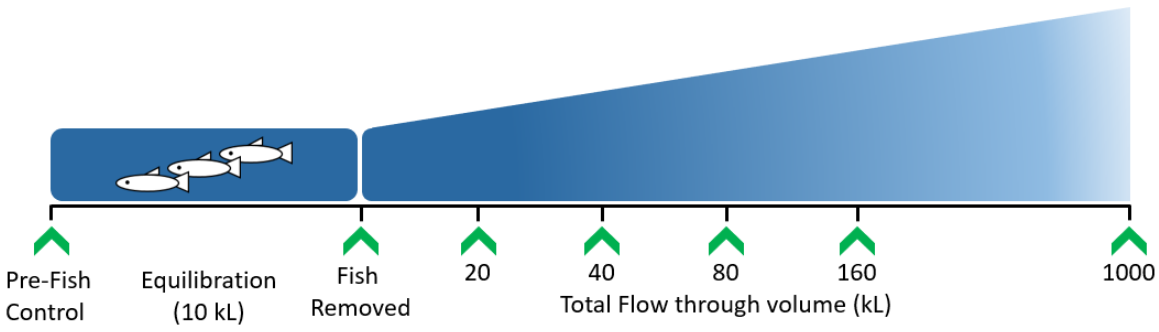

Figure 4.1: The Dilution Experiment (MacAdams, 2018; Hocking M.D et al., 2020)

Figure 4.1 depicts the dilution experiment. Three juvenile Coho salmon were placed in a minnow trap in each of four experimental tanks. The fish were left to equilibrate for nine days with flow through (140-155 L/minute). Traps containing the fish were then removed and accumulated eDNA was diluted via hatchery water flow-through. Pre-fish negative controls were collected for each tank, and fish-positive controls were collected after the nine-day equilibration period. Triplicate samples, indicated by the green arrows, at specified times representing aggregate flow-through volumes of 20, 40, 80, 160 and 1000 kL. Five negative control samples were taken concurrently from the hatchery pond (n = 3) and the hatchery kitchen sink (n= 2).

Each sample consisted of eight technical replicates, and each set of eight technical replicates was assigned a unique sort code. Technical replicates with sort codes equal to 31 and 84 did not pass integritE tests and were thus removed from the data set. These corresponded to samples from 20kL Flow, Tank 19 and from 1000kL Flow, Tank 20. There were an additional extra set of 12 sample replicates corresponding to 10kL Flow. These were stored in the excel file with special characters to indicate they were not to be included in analysis (because the samples had been taken after only three days, while the fish were still in the tanks opposed to after nine days). We include several tables summarizing how many samples were taken and the general method of the experiment.

Number of Sample.replicates by Flow/Tank

Flow (kL)	Log2(Flow)	Tank					Total
		1	19	20	21	24	
0			3	3	3	3	12
10	3.32		4	4	3	3	14
20	4.32		2	3	3	3	11
40	5.32		3	3	3	3	12
80	6.32		3	3	3	3	12
160	7.32		3	3	3	3	12
1000	9.97		4	3	4	4	15
pond		3					3
sink		2					2

Table 4.1: Summary of the number of sample replicates for each corresponding number of fish and Flow.

Table 4.1 provides an overview summary of the samples taken for the dilution

experiment. Also included is the number of sample replicates for sink and pond. Most levels of flow had three or four replicates taken from each tank. Two sample replicates were discarded since they did not appear to be valid measurements due to failure to pass DNA integritE tests. One was removed from tank 19 and 20kL flow and another was removed from tank 20 and 1000kL flow. The pond water had 3 sample replicates taken and the sink water had 2. Note that for each sample replicate, there are eight associated technical replicates.

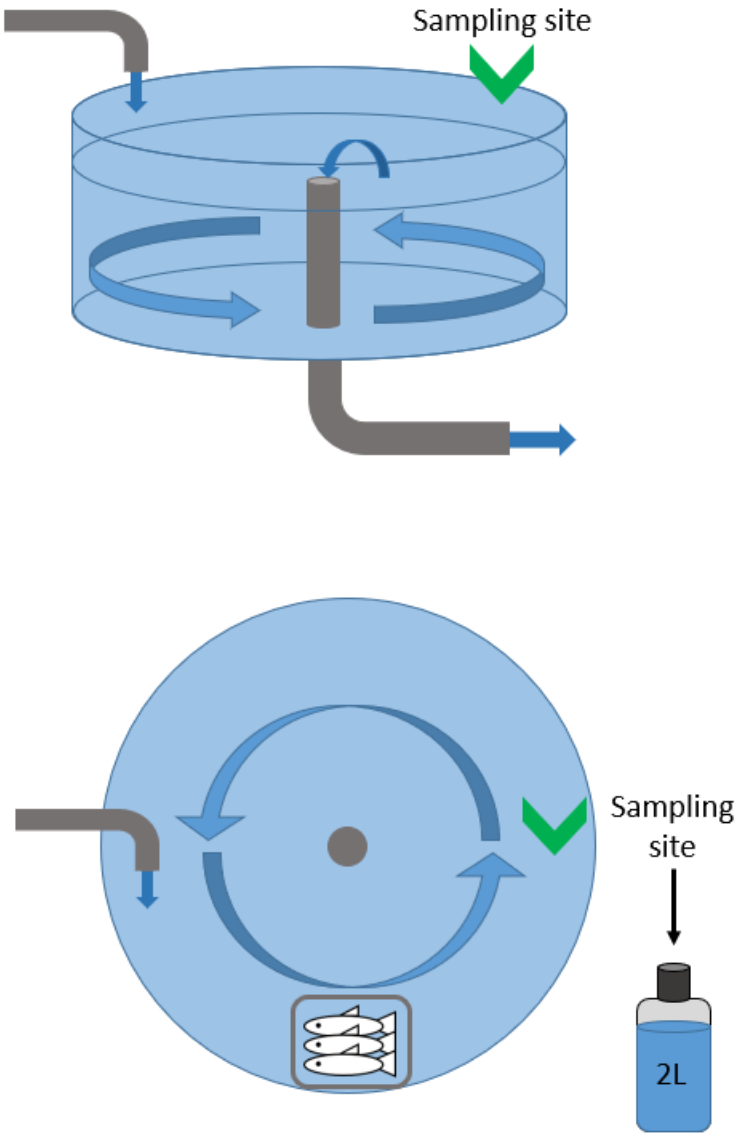

Figure 4.2: Sampling routine for the flow experiment (Bergman, 2020).

Figure 4.2 represents the general method in which samples were taken. eDNA samples were taken from the opposite side of the inflow pipe. The tanks were all of size 1000kL. Three fish in a minnow trap were in each tank except for the pre-fish negative controls.

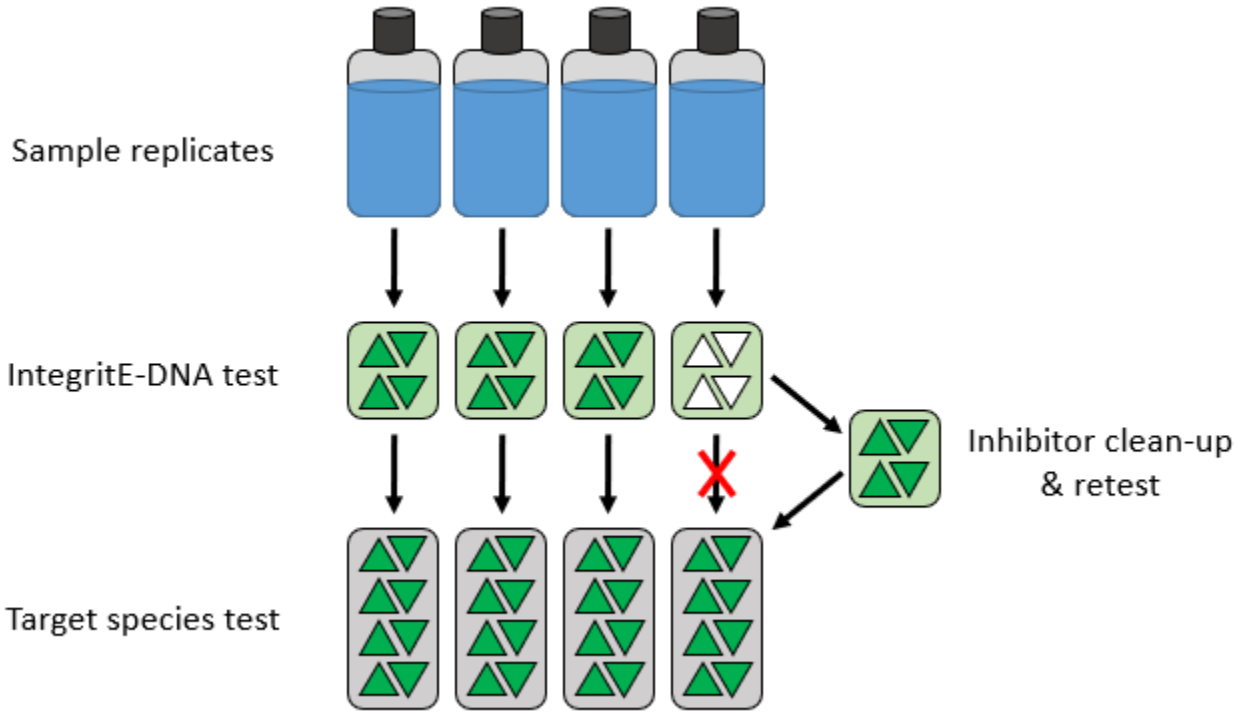

Figure 4.3: Analysis procedure for the dilution experiment (Bergman, 2020).

Figure 4.3 represents the method in which quality of eDNA was assessed before analysis. Sample replicates were first tested with integritE-DNA test. Samples which failed to pass integritE test were cleaned and retested before analysis.

4.1 Flow Plots

First, we examine the 'Pre-Fish Negative controls'. We plot the Sample.replicate number on the x-axis, and the transformed CT (TCT) of the associated eight technical replicates on the y-axis. These plots correspond to tanks which had just been filled by the infill pipe, and no additional water had been allowed to flow through the tank.

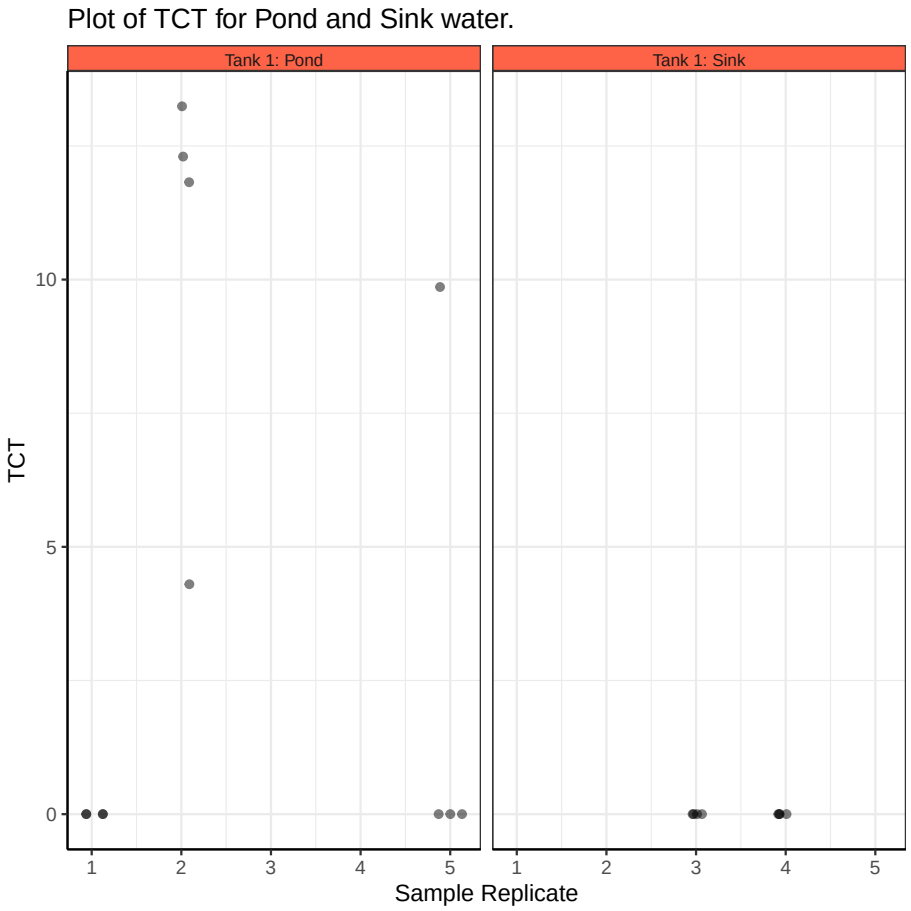

Figure 4.4: Transformed CT values obtained from samples taken from pond and sink water. Samples were taken from a small tank, tank 1.

Figure 4.4 are the plots of TCT values obtained from the hatchery sink and the hatchery pond. The water was transferred to tank 1 where sample replicates were then obtained. The goal of these measurements was to observe possible background hatchery signal of Coho eDNA. While the sink water showed no sign of Coho eDNA, the samples from the pond did indicate presence of Coho.

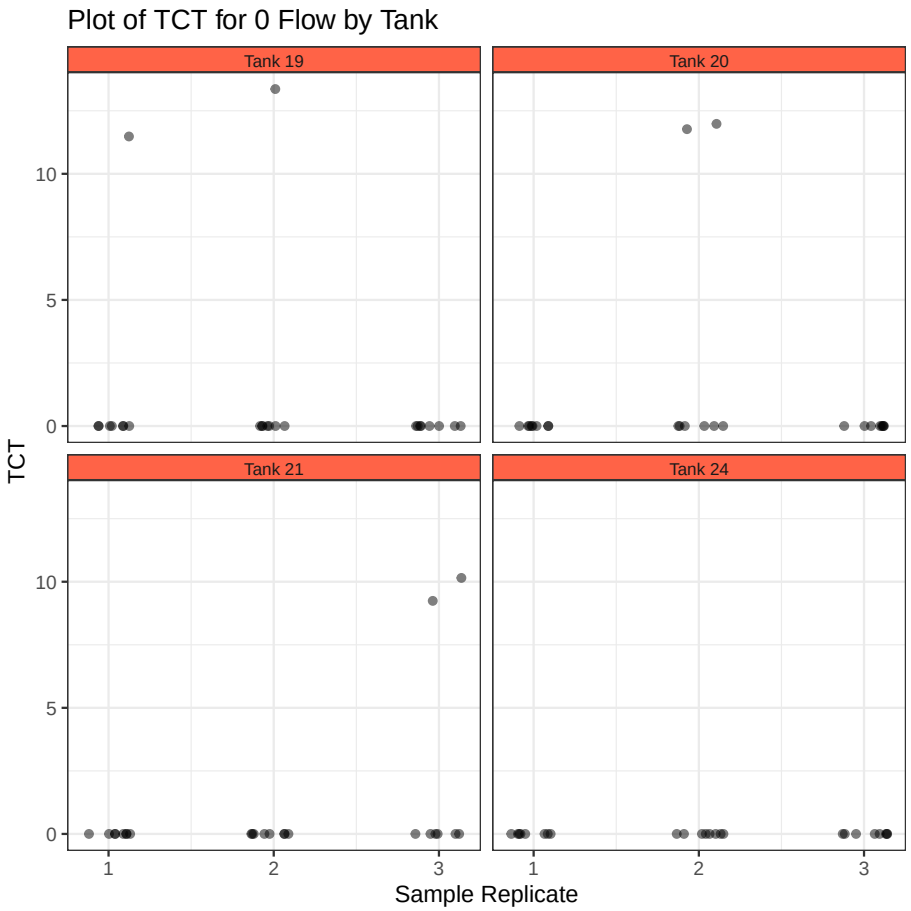

Table 4.2: Transformed CT values obtained from pre-fish negative controls.

Table 4.2 are the TCT plots obtained from the tanks before the fish were added. The samples help to measure a hatchery background signal of Coho eDNA. Most samples indicate no presence of Coho eDNA, although there exist several outliers. This may indicate that the hatchery water itself contains trace amounts of Coho eDNA.

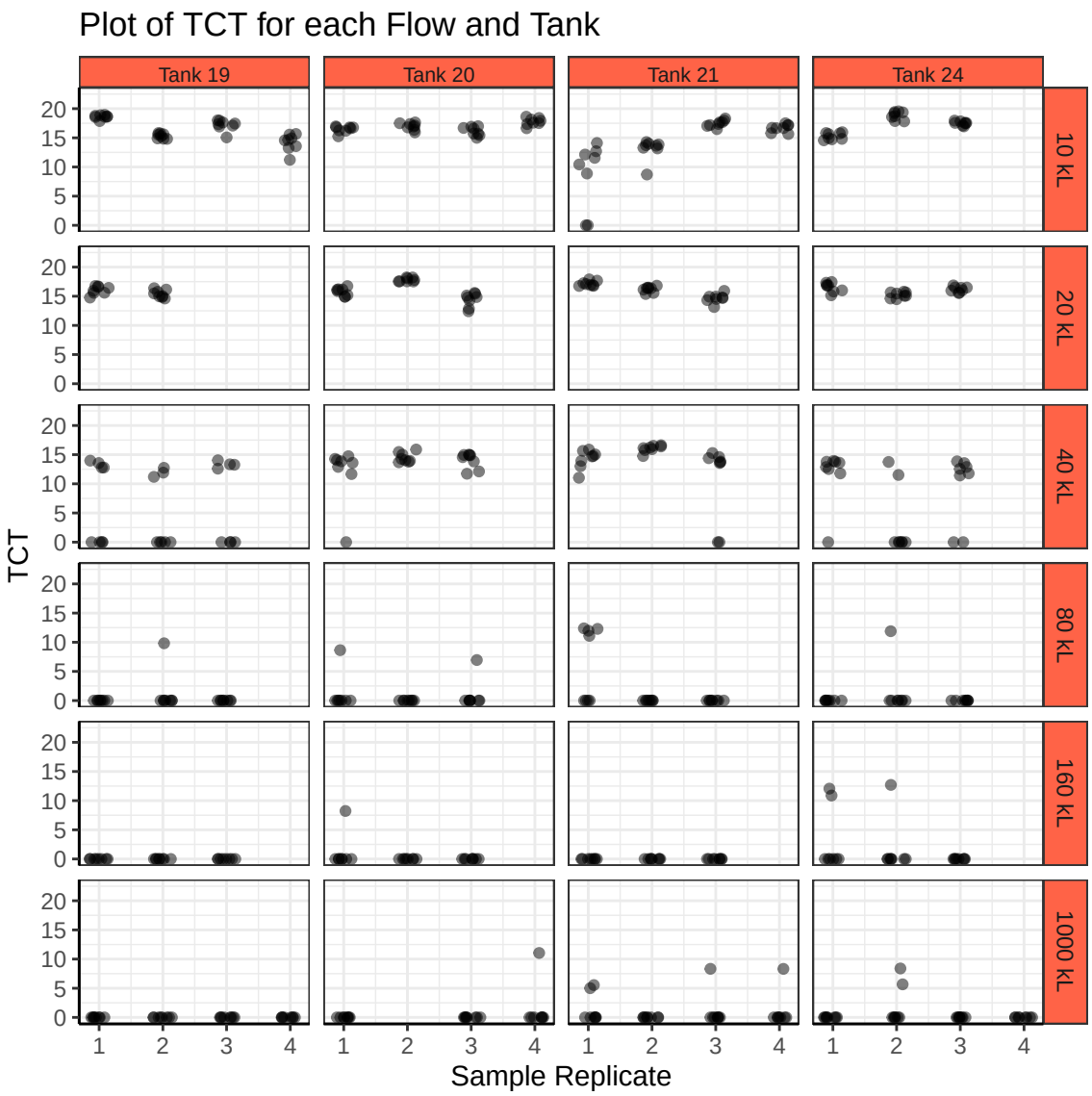

Figure 4.5: Transformed CT values obtained from samples taken from each tank at differing levels of dilution.

Table 4.5 are the TCT plots obtained from the tanks after the indicated levels of dilution. Note that the 10kL corresponds to samples that were taken right after the fish were removed (prior to dilution). When only 20kL of water had been allowed to flow through the tank, we consistently detected eDNA. At 40kL of flow, we still had consistent detection, but the number of non-detects began to increase more significantly. By the time we reach 80kL, detection of Coho eDNA was no longer consistent. Presence of Coho eDNA was not being picked up, except in a small number of cases.

For flows greater than 80kL, we are no longer consistently detecting Coho eDNA. However, there does exist the rare detection of Coho eDNA at even the higher levels of flow. This may be due to the background noise in the hatchery water that we previously discussed.

4.2 Flow Models

4.2.1 Median Transformed CT

We now fit statistical models for the response variable median TCT. Only data that occurred once the minnow trap was removed is included in this analysis. This corresponds to the six extraction points in Figure 4.5. In general, as log2(Flow) increases, we would expect 'CT' to rise and hence 'TCT' to decrease as the shed eDNA was washed out of the tank. We first fit a simple linear model, flow.l.one.line. This model only considers log2(Flow) as a predictor.

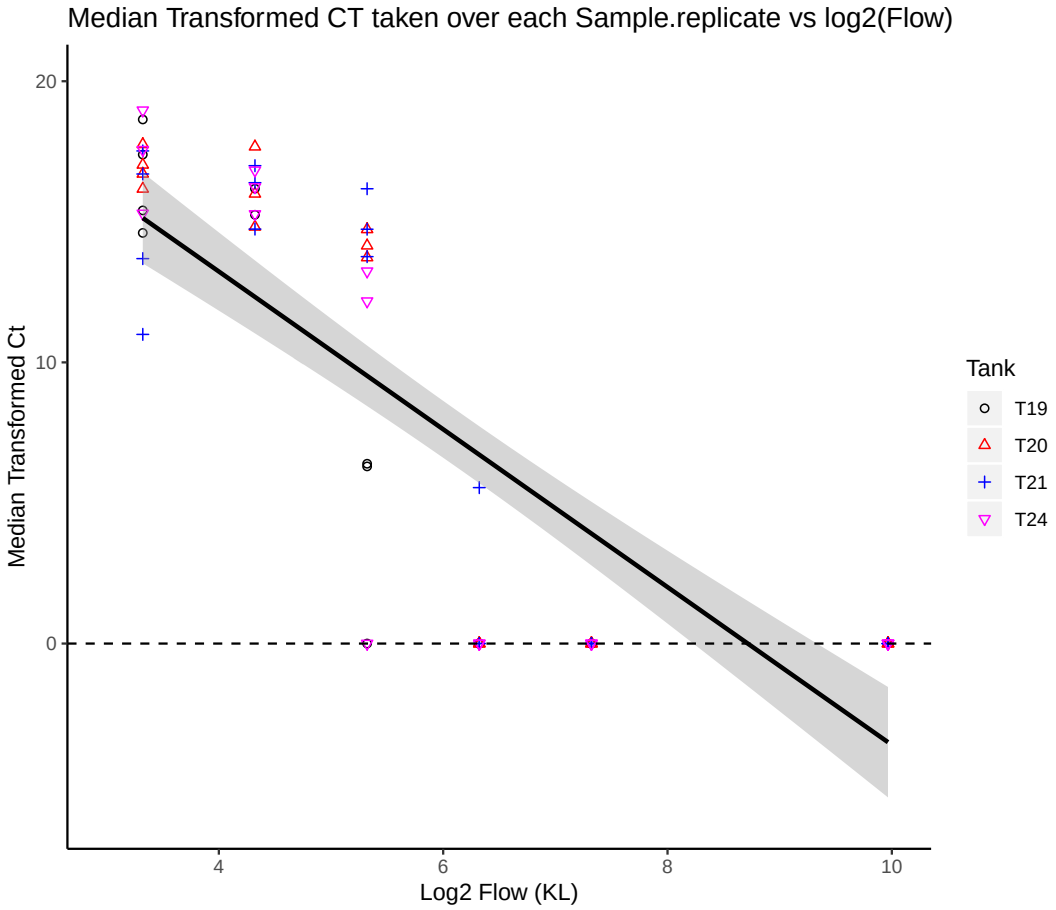

Figure 4.6: Plot of Median TCT vs Log2(Flow) in kL, included is the simple linear model, flow.l.one.line and its associated 95% confidence bands.

```
[1] "Model:flow.l.one.line"

Call:
lm(formula = TCTmed ~ l2Flow,
data = flow.new.sum.dat)

Residuals:
   Min     1Q Median     3Q    Max
-9.522 -3.909  2.386  3.510  6.653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.4567     1.4605   16.75   <2e-16 ***
l2Flow       -2.8061     0.2223  -12.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.447 on 75 degrees of freedom
Multiple R-squared:   0.68,Adjusted R-squared:  0.6757
F-statistic: 159.3 on 1 and 75 DF,  p-value: < 2.2e-16
```

Table 4.3: A simple linear model for Median TCT that only considers log2(Flow) as a predictor. Model: flow.l.one.line

Figure 4.6 is the plot of the regression line flow.l.one.line. Median Transformed CT is taken over each Sample.replicate vs Log2(Flow). Notice that at a certain point, median TCT approaches zero. That is, we no longer are detecting eDNA past a certain value of log2(Flow). We later use alternative models that account for this fact as it does not make sense to predict negative median TCT values. Table 4.3 provides the estimates of the model flow.l.one.line. The estimated intercept is 24.457 with a standard error of 1.46 and the estimate for the flow term is -2.806. Both estimates have extremely small p-values, indicating high significance of both the intercept and slope. The R^2 is 0.68 which means our model does a reasonable job at explaining variation in the data. The $Pr(> |t|)$ is the p-value for the test of a zero parameter value. Since both the intercept and log2(flow) have small p-values, they should be included in the model.

We now proceed as in Chapter 3, whereby we consider the impact of tank and interactions on the TCT values. We fit a model flow.l.tank, that considers dilution as before and also considers the tank as a factor. This model thus allows the intercept to change depending on the tank, however, the slope for each tank remains constant.

```
[1] "Model:flow.l.tank"

Call:
lm(formula = TCTmed ~ l2Flow + TankF,
data = flow.new.sum.dat)

Residuals:
   Min     1Q Median     3Q    Max
-9.341 -4.522  1.768  3.634  5.997

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.2535     1.7365  13.391   <2e-16 ***
l2Flow       -2.7941     0.2242 -12.465   <2e-16 ***
TankF20       1.7271     1.4539   1.188    0.239
TankF21       1.7952     1.4343   1.252    0.215
TankF24       0.9581     1.4524   0.660    0.512
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.477 on 72 degrees of freedom
Multiple R-squared:  0.6887,Adjusted R-squared:  0.6714
F-statistic: 39.82 on 4 and 72 DF,  p-value: < 2.2e-16
```

Table 4.4: A model that allows for differing intercepts among tanks. Model: flow.l.tank.

Table 4.4 is a summary of our model that includes the tank number as a factor. As before, there is a high level of significance for the intercept and flow terms. Each tank has an associated coefficient term. The baseline tank is tank 19. For example, a sample corresponding to Tank 20 would have an intercept of 23.25+1.727 and a slope of -2.79. The individual p-values for each tank do not immediately appear significant. However, this does not guarantee that tank itself is not important. To conclude that, we would need to compare this model with a model that does not contain tanks, we can do this using ‘anova’. The R^2 for flow.l.tank 0.69, which is only slightly larger

than the R^2 for flow.l.one.line. Moreover, the adjusted R^2 which accounts for number of predictors is now actually less than the R^2 of flow.l.one.line.

We also fit a model flow.l.four.line that considers flow, tank as a factor and also the interaction between tank and biomass. This will allow not only for differing intercepts, but also differing slopes for each tank.

```
[1] "Model:flow.l.four.line"

Call:
lm(formula = TCTmed ~ l2Flow + TankF + l2Flow * TankF,
data = flow.new.sum.dat)

Residuals:
   Min     1Q Median     3Q    Max
-9.347 -4.091  1.887  3.784  6.051

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     21.8960     2.9888   7.326 3.40e-10 ***
l2Flow          -2.5767     0.4483  -5.748 2.25e-07 ***
TankF20          4.9543     4.2496   1.166    0.248
TankF21          2.7528     4.1618   0.661    0.511
TankF24          2.3573     4.2892   0.550    0.584
l2Flow:TankF20  -0.5315     0.6542  -0.812    0.419
l2Flow:TankF21  -0.1523     0.6288  -0.242    0.809
l2Flow:TankF24  -0.2239     0.6419  -0.349    0.728
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.55 on 69 degrees of freedom
Multiple R-squared:  0.6918,Adjusted R-squared:  0.6605
F-statistic: 22.12 on 7 and 69 DF,  p-value: 2.273e-15
```

Table 4.5: A model for which intercepts and slopes are allowed to differ for each tank. Model: flow.l.four.line.

Table 4.5 summarizes the model flow.l.four.line. This model includes tank and the interaction between tank and log2(Flow). The output indicates that neither tank, nor the interaction are significant. The interaction terms are multiplications of flow and the tank indicator terms. These terms would cause a different slope depending from which tank the sample was taken.

We now preform an 'anova' test to see which of our models is significant;

```
Analysis of Variance Table

Model 1: flow.l.one.line
Model 2: flow.l.tank
Model 3: flow.l.four.line

Model 1: TCTmed ~ l2Flow
Model 2: TCTmed ~ l2Flow + TankF
Model 3: TCTmed ~ l2Flow + TankF + l2Flow * TankF
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     75 1483.2
2     72 1442.8  3    40.388 0.6503 0.5855
3     69 1428.5  3    14.324 0.2306 0.8748
```

Table 4.6: ANOVA to compare the three flow models.

Table 4.6 provides the results of the anova test. The full model, flow.l.four.line allows both the intercept and slopes to differ depending on which tank the sample is taken from. We compare this model with the model flow.l.tank to test for the hypothesis of differing slopes between tanks. Because the p-value is large (p=0.8748) , we cannot reject the hypothesis that the coefficient for additional slopes is zero. Similarly, when comparing flow.l.tank with flow.l.one.line, the hypothesis is that the coefficient that determines differing intercepts among tanks is zero. Because the p-value is again large (p=0.5855), we cannot reject the hypothesis that the coefficient is zero. Hence, we conclude that both tank and the interaction between tank and Flow are not needed in our models.

4.2.2 Flow Models (Mean)

We now fit statistical models for the response variable mean TCT. First, we fit a simple linear model, flow.l.one.line.mean. This model only considers log2(Flow) as a predictor.

```
Call:
lm(formula = TCTmean ~ l2Flow,
data = flow.new.sum.dat)

Residuals:
   Min     1Q Median     3Q    Max
-6.862 -4.175  1.988  2.990  6.495

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.8486     1.3591   17.55   <2e-16 ***
l2Flow       -2.6868     0.2069  -12.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.138 on 75 degrees of freedom
Multiple R-squared:  0.6922,Adjusted R-squared:  0.6881
F-statistic: 168.7 on 1 and 75 DF,  p-value: < 2.2e-16
```

Table 4.7: A simple linear model for Mean TCT which only considers log2(Flow) as a predictor. Model: flow.l.one.line.mean

Table 4.7 summarizes flow.l.one.line.mean. The output provides estimates of an intercept of 23.848 and a slope term on log2(Flow) of -2.686. The R^2 is 0.692 and the adjusted R^2 is 0.688.

We now fit a model, flow.l.tank.mean that also includes the tank number as a factor. As before, R indicates high significance for the intercept and flow terms.

```
Call:
lm(formula = TCTmean ~ l2Flow + TankF,
data = flow.new.sum.dat)

Residuals:
   Min     1Q Median     3Q    Max
-7.428 -4.086  1.497  3.584  6.111

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.8150     1.6211  14.074   <2e-16 ***
l2Flow       -2.6769     0.2093 -12.793   <2e-16 ***
TankF20       1.5379     1.3572   1.133    0.261
TankF21       1.3651     1.3389   1.020    0.311
TankF24       0.9675     1.3559   0.714    0.478
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.179 on 72 degrees of freedom
Multiple R-squared:  0.6987,Adjusted R-squared:  0.682
F-statistic: 41.75 on 4 and 72 DF,  p-value: < 2.2e-16
```

Table 4.8: A model for Mean TCT that considers tank as a predictor in addition to log2(Flow). Model: flow.l.tank.mean

Table 4.8 summarizes the model flow.l.tank.mean. The estimated intercept for tank 19 is 22.82 and the slope coefficient is -2.677. The R^2 for the new mean model is 0.699 , however the adjusted R^2 has now decreased to 0.682. The individual p-values for each tank do not immediately appear significant. However, this does not guarantee that tank itself is not important. To conclude that, we need to compare this model with a model that does not contain tanks, we can do this using ‘anova’.

```
Call:
lm(formula = TCTmean ~ l2Flow + TankF + l2Flow * TankF,
data = flow.new.sum.dat)

Residuals:
   Min     1Q Median     3Q    Max
-7.312 -4.015  1.488  3.579  6.260

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    22.33649    2.78951   8.007 1.94e-11 ***
l2Flow         -2.60034    0.41842  -6.215 3.41e-08 ***
TankF20         3.88081    3.96616   0.978    0.331
TankF21         0.73721    3.88426   0.190    0.850
TankF24         1.34798    4.00320   0.337    0.737
l2Flow:TankF20 -0.38996    0.61058  -0.639    0.525
l2Flow:TankF21  0.10328    0.58683   0.176    0.861
l2Flow:TankF24 -0.06103    0.59908  -0.102    0.919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.247 on 69 degrees of freedom
Multiple R-squared:  0.7018,Adjusted R-squared:  0.6716
F-statistic:  23.2 on 7 and 69 DF,  p-value: 7.473e-16
```

Table 4.9: A model for Mean TCT that includes log2(Flow), Tank, and the interaction between log2(Flow) and Tank. Model: flow.l.four.line.mean.

This model includes tank and the interaction between tank and log2(Flow). Hence this allows for both intercepts and slopes to differ between each tank. The results indicate that neither tank, nor the interaction terms are significant.

```
Analysis of Variance Table

Model 1: flow.l.one.line.mean
Model 2: flow.l.tank.mean
Model 3: flow.l.four.line.mean

Model 1: TCTmean ~ l2Flow
Model 2: TCTmean ~ l2Flow + TankF
Model 3: TCTmean ~ l2Flow + TankF + l2Flow * TankF
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     75 1284.4
2     72 1257.3  3    27.086 0.5007 0.6831
3     69 1244.3  3    13.030 0.2409 0.8675
```

Table 4.10: ANOVA table for flow models.

Table 4.10 shows the results of the 'anova' test to see which of our models for mean TCT is significant. We first consider the full model flow.l.four.line.mean that allows for different slopes and different intercepts. Comparing flow.l.four.l.mean with flow.l.tank.mean is a test of equality of slopes. Since the p-value is large, we cannot reject the null hypothesis that the terms are identical. That is, we cannot reject the hypothesis that the slopes do not differ among tanks. Next we compare flow.l.tank.mean with flow.l.one.line.mean, which is a test of differing intercepts. Similarly, we obtain a large p-value (p=0.6831) and hence cannot reject the hypothesis that the intercepts do not differ. Thus, as before, we conclude that working with a model that only includes log2(flow) as a predictor is sufficient.

4.3 Alternative Models

As seen above, once the dilution is large enough, we no longer detect the presence of Coho eDNA. Hence our statistical model should account for this asymptotic property. We fit and compare a variety of non-linear models. These include the 'broken stick' model for mean TCT, hyperbolic tangent models, Lowess models and finally a 'Bent Cable Model' (Toms and Lesperance, 2003). In this analysis we include models fit to the response variable Mean Transformed CT. The models for median Transformed CT are very similar.

4.3.1 Broken Stick Models

Firstly, we fit what is known as a broken stick model. The broken stick model is a non-linear regression model that allows for sharp changes in direction. The model includes a 'break point' , γ, that is used to fit a model. We use the R function 'nls' (Baty et al., 2015) which stands for 'Nonlinear Least Squares' to fit the model. This function allows us to obtain 'weighted' estimates of the parameters.

The Broken Stick model has the form:

$$\beta_0 + \beta_1 x + \begin{cases} 0 & \text{if } x \leq \gamma \\ \\ \beta_2(x-\gamma) & \text{if } x > \gamma \end{cases}$$

This can be written as:

$$\beta_0 + \beta_1 x \ \text{ for } \ x \leq \gamma \text{ , and}$$

$$\beta_0 + \beta_1 x + \beta_2(x-\gamma) = (\beta_0 - \beta_2\gamma) + (\beta_1 + \beta_2)x \ \text{ for } \ x > \gamma.$$

.

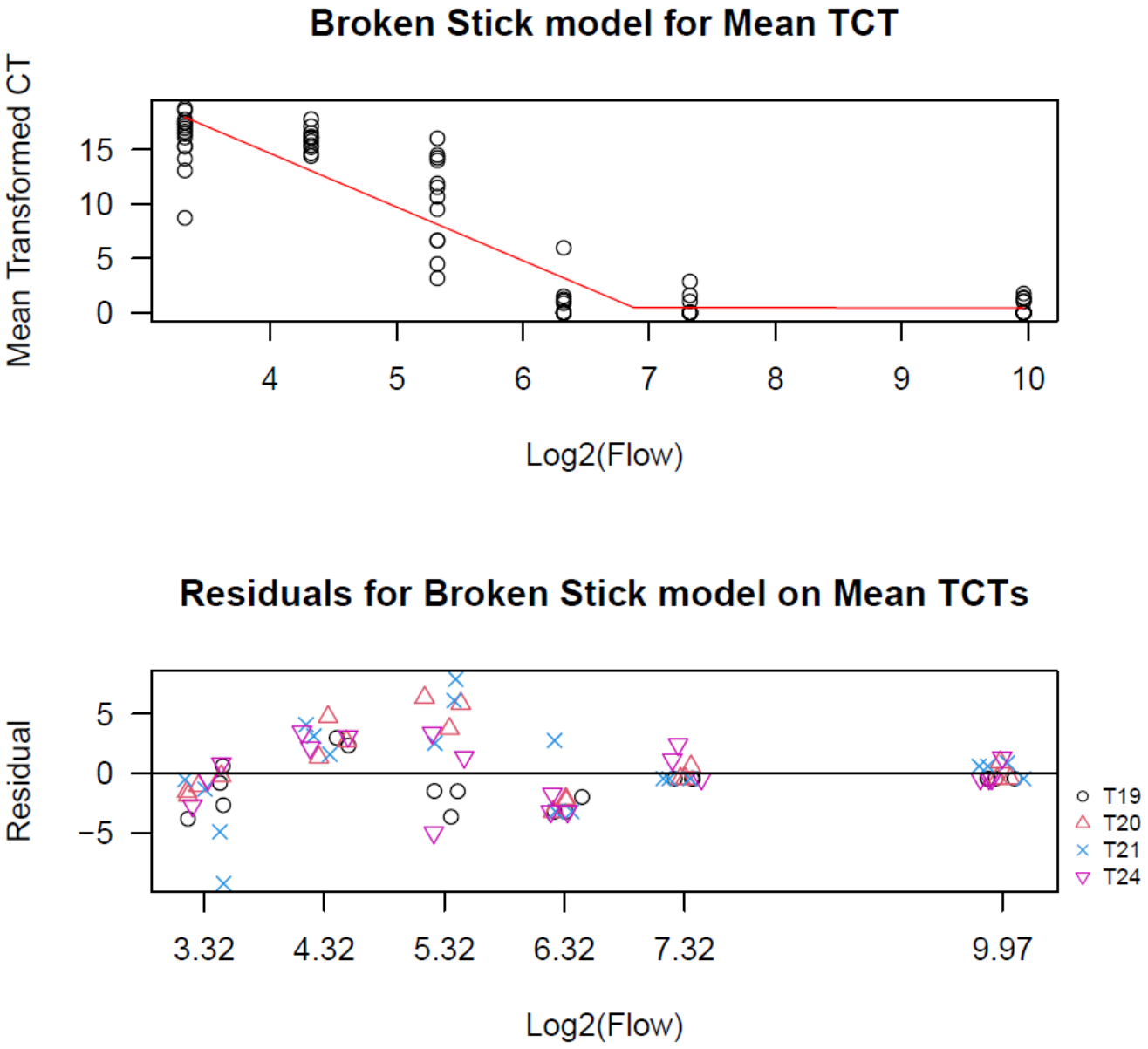

Figure 4.7: Broken Stick model for Mean TCT and the associated residuals for the model.

Figure 4.7 shows the broken stick regression line in red. The 'breakpoint', γ is estimated as 'Br' and is 6.877 (s.e 0.266). The estimate for β_0 is 33.374 (s.e 1.74), β_1 has an estimate of -4.931 (s.e 0.347) and β_2 is estimated to be 4.932 (s.e 0.555). Also included is the residual plot for the Broken-stick model. The fitted values are

the predicted mean TCT for a given log2(Flow) value. On the y-axis are the actual residuals, the difference between the true value and the fitted value. The 'RSS' (Residual Sum of Squares) for this model is 614.761.

4.3.2 Bent Cable Models

In the above analysis we fit 'broken stick models'. We can see that in the broken stick models, the break point causes a sharp turn. However, it may be more reasonable and biologically sound to fit a more flexible model. Hence, we make use of the "bent cable" model that allows for a more smooth model (Chiu et al., 2006). We use the bentcable function (Sonderegger, 2020).

The Bent Cable model takes the form:

$$\beta_0 + \beta_1 x + \begin{cases} 0 & \text{if } x < \alpha - \gamma \\ \beta_2 \frac{(x-\alpha+\gamma)^2}{4\gamma} & \text{if } \alpha - \gamma \leq x \leq \alpha + \gamma \\ \beta_2 (x - \alpha) & \text{if } x > \alpha + \gamma \end{cases}$$

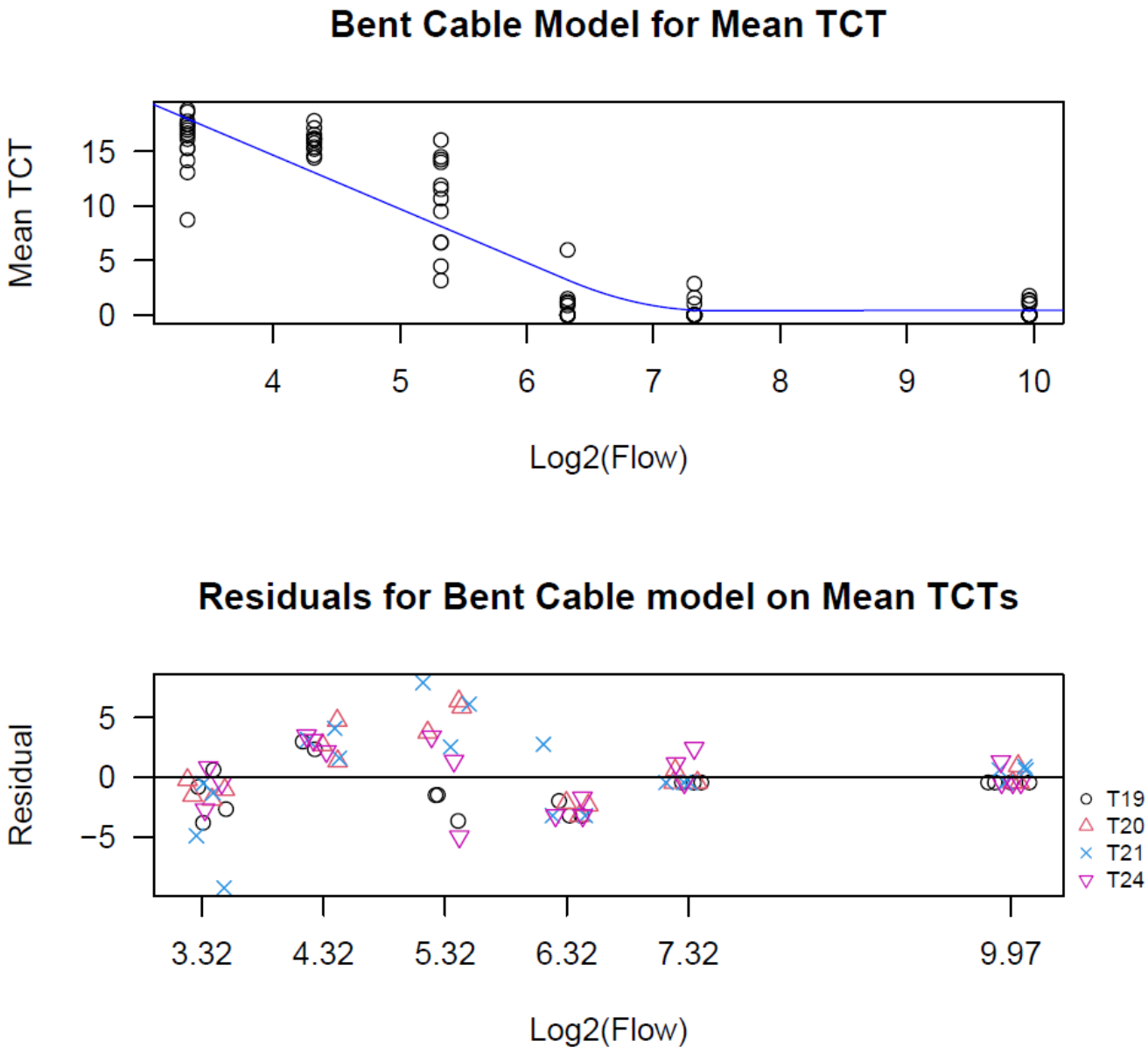

Figure 4.8: Bent Cable Model for Mean TCT and the associated residuals for the model.

Figure 4.8 is the plot of the bent cable model and the associated residual plot. The red line is the bent cable regression line. The α value is estimated as 6.89 (s.e 0.542), the γ value is estimated 0.565 (s.e 0.124) and the log-likelihood is -189.24. The estimate for β_0 is 34.350 (s.e 1.51), β_1 has an estimate of -4.925 (s.e 0.288) and β_2 is

estimated to be 4.940 (s.e 0.550). The parameter α is the breakpoint discussed above, while γ is a parameter that impacts the length of the quadratic portion connecting the two linear sections of the regression line. The RSS for this model is 614.767.

4.3.3 Hyperbolic Tangent Models

We can also fit models similar to 'bent cables' that utilize the hyperbolic tangent function to allow for smoothness.

The tanh model takes the form:

$$\beta_0 + \beta_1(x-\alpha) + \beta_2(x-\alpha)\tanh(\frac{x-\alpha}{\gamma})$$

where $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ is the hyperbolic tangent function.

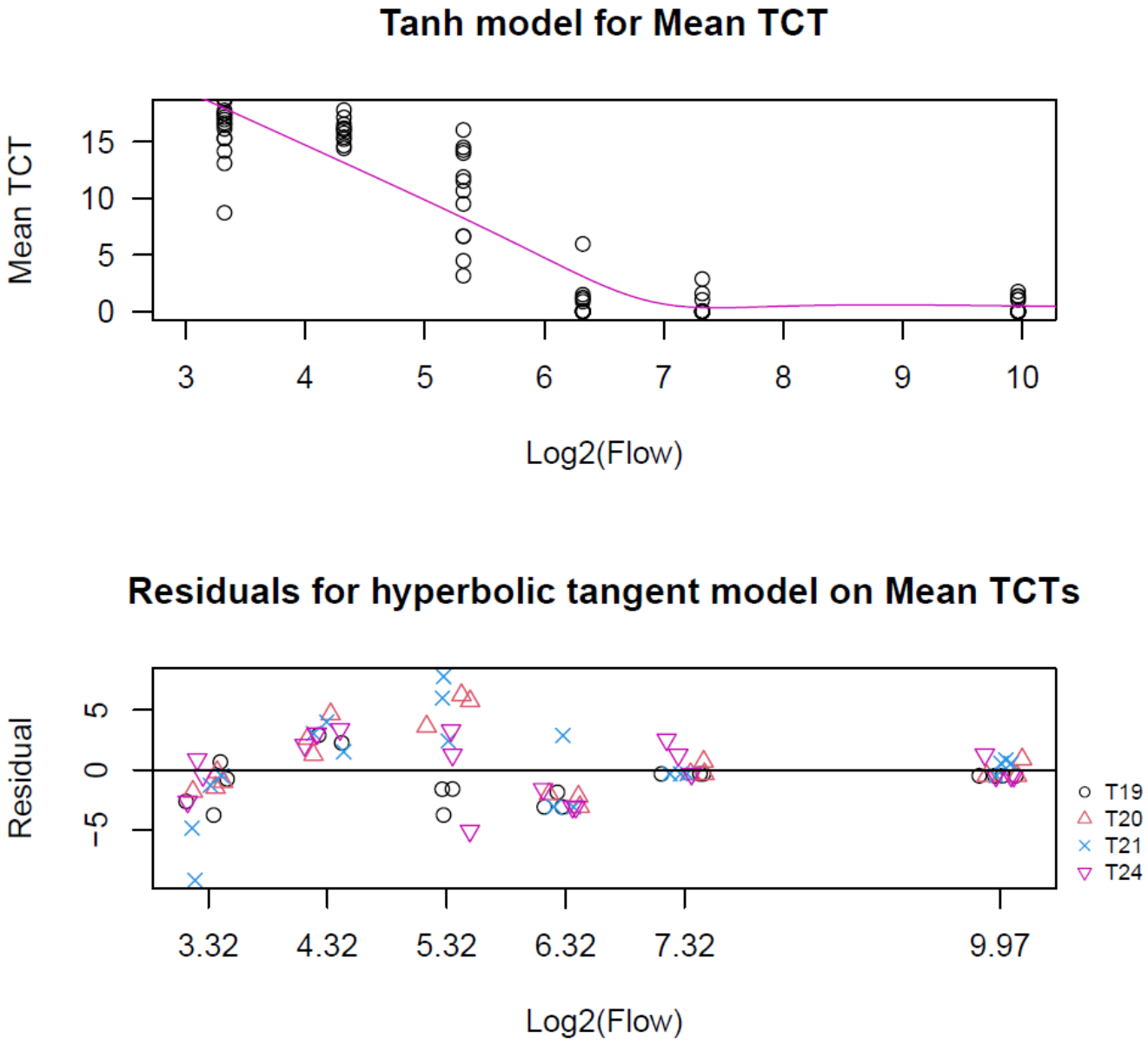

Figure 4.9: Hyperbolic Tangent (tanh) model for Mean TCT and the associated residuals for the model.

Figure 4.9 is the tanh model and residuals for mean TCT. The purple line is the tanh regression line. For mean TCT, our hyperbolic tangent model estimates the break point 'br' (α) to be $\alpha = 6.887$ (s.e 0.448). R estimates $\gamma = -0.776$ (s.e 4.82) , $\beta_0 = 0.8920$ (s.e 2.80), $\beta_1 = -2.46$ (s.e 0.350) and $\beta_2 = -2.33$ (s.e 1.03). The

residuals look good as they appear to be randomly distributed about the x-axis. The RSS for the tanh model is 595.239.

4.3.4 Lowess Models

We fit similar models using the 'lowess' function in R. 'lowess' stands for locally-weighted polynomial regression. The lowess models are quite 'sharp'. They both appear similar to our 'broken stick' models. We used a smoothing parameter value of 3/4 and allowed the algorithm to perform three iterations.

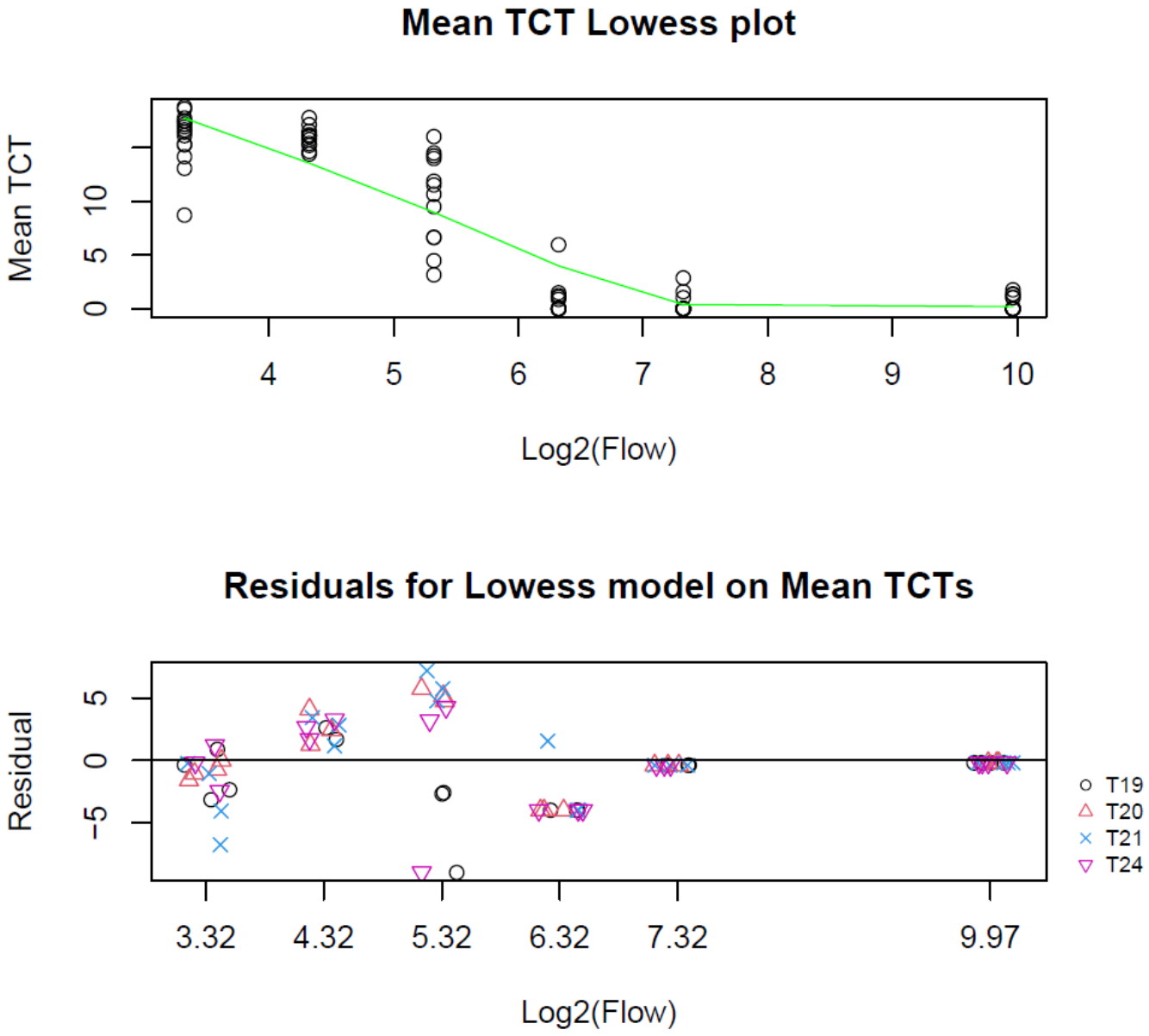

Figure 4.10: Lowess model for Mean TCT and the associated residuals for the model.

Figure 4.10 is the lowess model for mean TCT and the associated residuals. The purple line represents the line of best fit according to the lowess criteria. This model looks very similar to the broken-stick model. This model estimates an intercept of 24.42 (s.e 0.894) and a lowess parameter of -2.74 (s.e 0.136). The residuals look good as they appear to be randomly distributed about the x-axis. The RSS for the lowess

model is 589.698.

4.3.5 Model Comparison

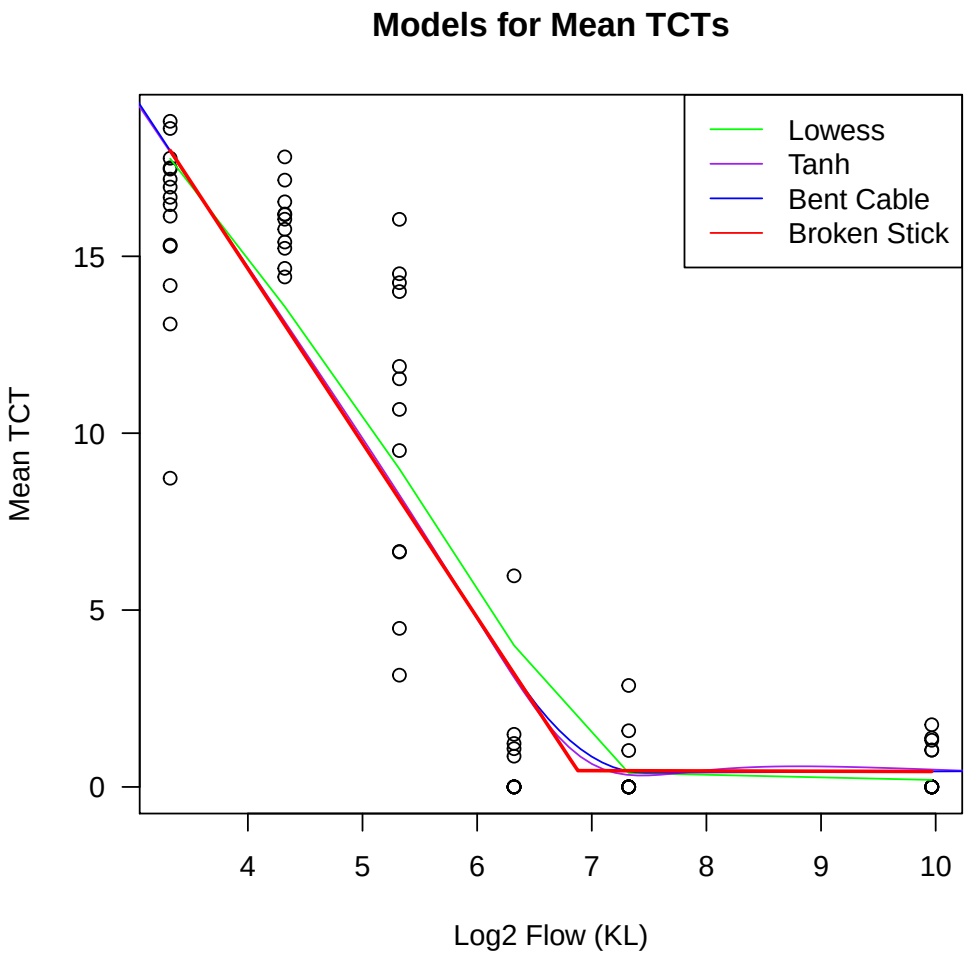

Figure 4.11: Model comparison for four distinct models for the response variable mean TCT. Here we plot a broken-stick model, a lowess model, a hyperbolic tangent model and a bent-cable model.

Figure 4.11 plots all of our non-linear models for mean TCT. They all appear to be quite similar to one another. The broken-stick model appears to have a very sharp turn at the breakpoint, while the bent cable makes a more smooth transition. The log-likelihood for the broken stick model is -189.24, which is the same as the log-likelihood for the bent-cable model. The log-likelihood for the Lowess and Tanh models is -185.34 and -187.99 respectively.

The final model that we choose is the bent cable model. We choose this because it has a low log-likelihood and also makes biological sense. The bend represents the point of time in which eDNA is no longer detectable due to dilution. The RSS of the bent cable model is also not much larger than the RSS of the other three models.

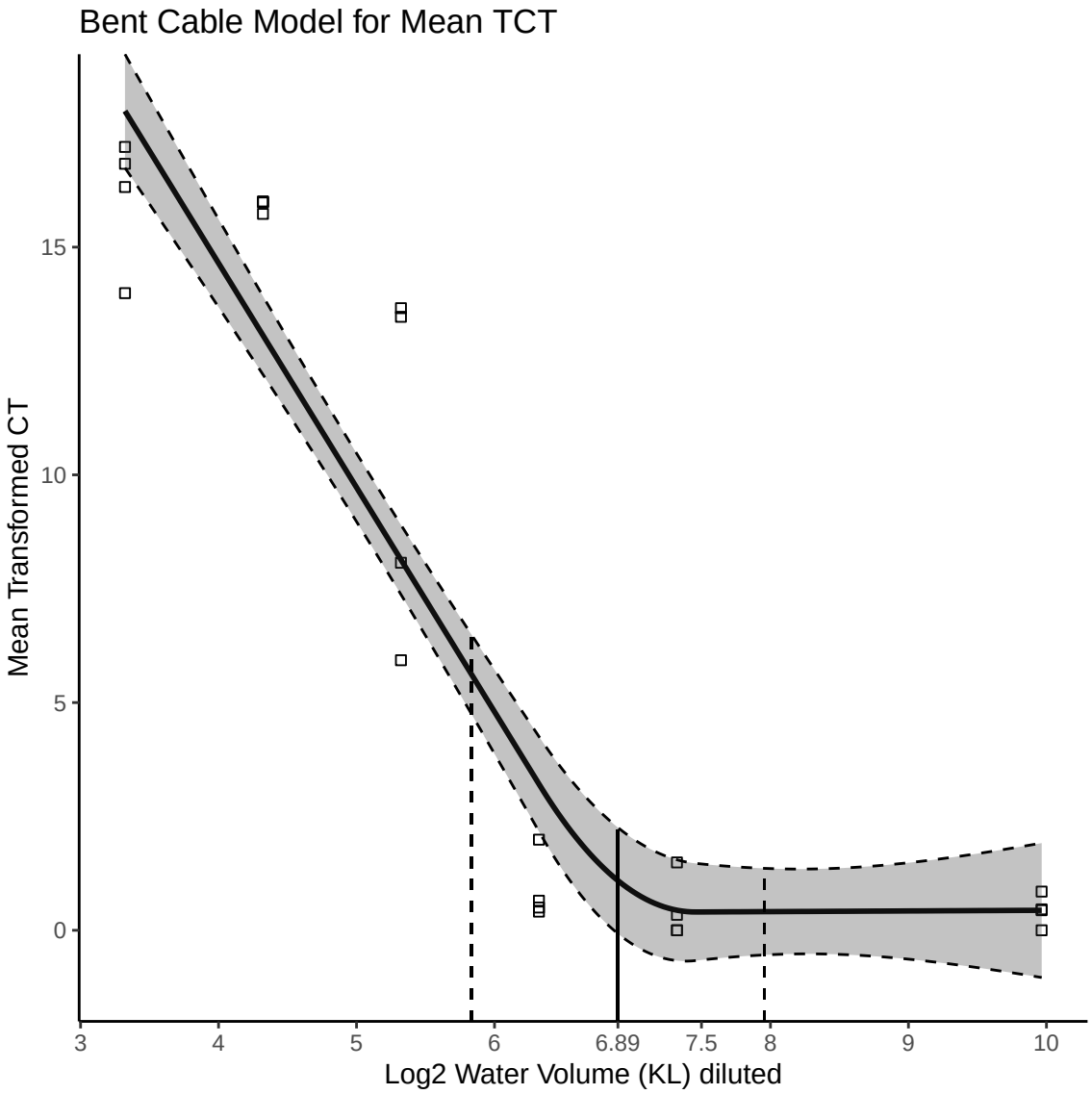

Figure 4.12: A bent cable model for mean TCT.

Figure 4.12 is a plot of our bent-cable model. The α value (estimated breakpoint) is 6.89 and is represented by the solid vertical line, the γ value is 0.565 and the log-likelihood is -189.24. Included are the confidence regions for the regression line. Also included is the confidence interval for the breakpoint represented by the vertical

dashed lines. The γ value is a parameter that controls the length of the quadratic portion of the curve.

4.4 Dilution Conclusions

The Dilution experiment allowed us to study the impact of flow on the collection of Coho eDNA samples. This study provided a controlled environment to study the loss of eDNA signal via degradation after the Coho had been removed from the tanks. The collection of plots organized in Table 4.5 illustrate the general trend. At the time of fish removal, TCT values were highest and sample replicates frequently had mean TCT values exceeding 15. As seen in Figure 4.5, at dilution levels of 160 kL and 1000 kL, TCT levels were not significantly different than background hatchery signals (Table 4.2).

In the dilution experiment, Coho eDNA was detected at the 20kL dilution with 100% reliability at the technical replicate level. At the 40kL level of dilution, Coho eDNA was detected in all but two sample replicates. Technical replicate non-detections occurred in eight of twelve sample. At 80kL dilution, only one sample replicate level detection was made, with 4 out of 8 qPCR replicates successfully amplifying target DNA. For dilutions greater than 80kL, we were not able to confirm presence with perfect detection, although several of the samples had 1 to 2 of 8 technical replicate detections.

Because there appeared to be a point at which the impact of dilution made Coho eDNA measurement negligible, we used a non-linear, bent cable model to fit a model for mean TCT. This model accounted for a 'break-point'. As seen in Figure 4.12. Our breakpoint of $\log 2(x) = 6.89$ would represent a dilution of 118 kL. This could be thought of as the point at which Coho eDNA collection is no longer differentiable from the hatchery background signal. The confidence region for the breakpoint was computed by using the standard error of the model. The 95 % confidence interval for the breakpoint ranged from 5.832 to 7.956 on the log2 scale. These points cover the range of dilutions from 56.96 kL to 247.28 kL.

Chapter 5

Field Data

While Chapter 3 and Chapter 4 analyzed controlled experiments, one goal of eDNA technology is its usage in the field. To study how eDNA analysis can be used to detect Coho Salmon in the wild, several field studies were conducted. In particular, four streams in British Columbia were studied (steams AAA, BBB, CCC and DDD). Water samples were collected and associated environmental and physical covariates were recorded. The main fish species of interest was Coho Salmon (*Oncorhynchus kisutch*), although we also considered Cutthroat Trout (*Oncorhynchus clarkii*) and Rainbow Trout (*Oncorhynchus mykiss*). Ecofish personnel also took biomass measurements of each of the species caught using electrofishing. All biomass measurements were obtained using electrofishing, a method that involves using electric current to temporarily shock and stun fish. Researchers recorded the weight of Coho Salmon, Cutthroat Trout and Rainbow Trout and also recorded the overall weight of All Fish that were caught.

The dataset consists of 432 observations. From each of the four streams, distinct sample locations were chosen. For each sample location, sample replicates were taken each of which consisted of eight observations or technical replicates. Hence there were 432/8= 54 sample replicates as each sample replicate consisted of 8 technical replicates. Nine sample replicates were from stream AAA, nine sample replicates were from stream BBB, eighteen of the sample replicates were from stream CCC and the remaining eighteen sample replicates were from stream DDD. The reason eighteen samples were collected from stream CCC and stream DDD is because two distinct reaches in these streams were studied. Nine replicates were chosen from each reach within these two streams. Samples were collected form streams AAA, BBB and DDD

in 2017, while samples from Stream CCC were collected in 2018.

In this data analysis we considered the possible impact of covariates on TCT measurements. Firstly, we include the total biomass for each species that was caught in the transect (CO.Total.Biomass.g for Coho). We also included the total biomass per meter squared of transect (CO.Biomass.g.m2 for Coho) and total biomass per meter cubed of transect (CO.Biomass.g.m3 for Coho) calculated using the dimensions of the transect. The environmental covariates that we included were water temperature (Water.Temperature.C), pH, flow rate (Transect.Flow.cms), the distance from shore in meters from which the eDNA sample was taken (eDNA.Distance.from.Shore.m) and the depth of water in meters from which the sample was taken (eDNA.Total.Water.Depth.m). Moreover, we included the possibility of interactions between flow rate and all biomass variables.

We attempt to show that by accounting for common covariates, researchers may save time and obtain comparable estimates of biomass/abundance to those obtained using costly standard sampling methods such as electrofishing. We highlight the need for continued research into the topic. We include plots and models for all Fish and for Coho in the main thesis. Results for Rainbow Trout and Cutthroat Trout can be found in the appendix. In the following pages, we include summary plots for each of the four streams. We include the total biomass of All Fish and Coho Salmon caught in each stream and their associated TCT scores for each sample replicate.

5.1 Streams

5.1.1 Stream AAA

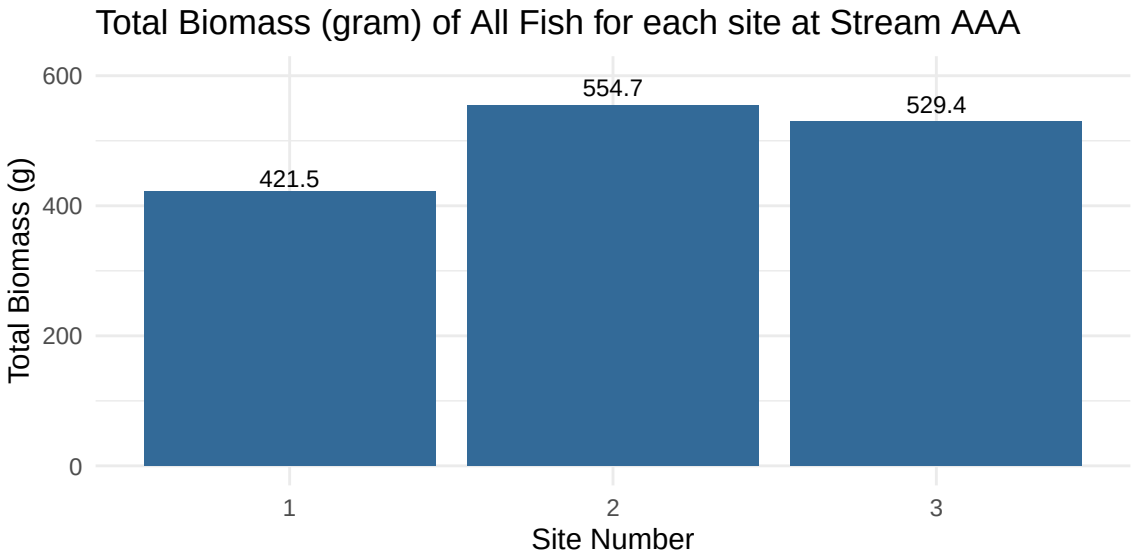

Figure 5.1: Total Biomass for All Fish at Stream AAA.

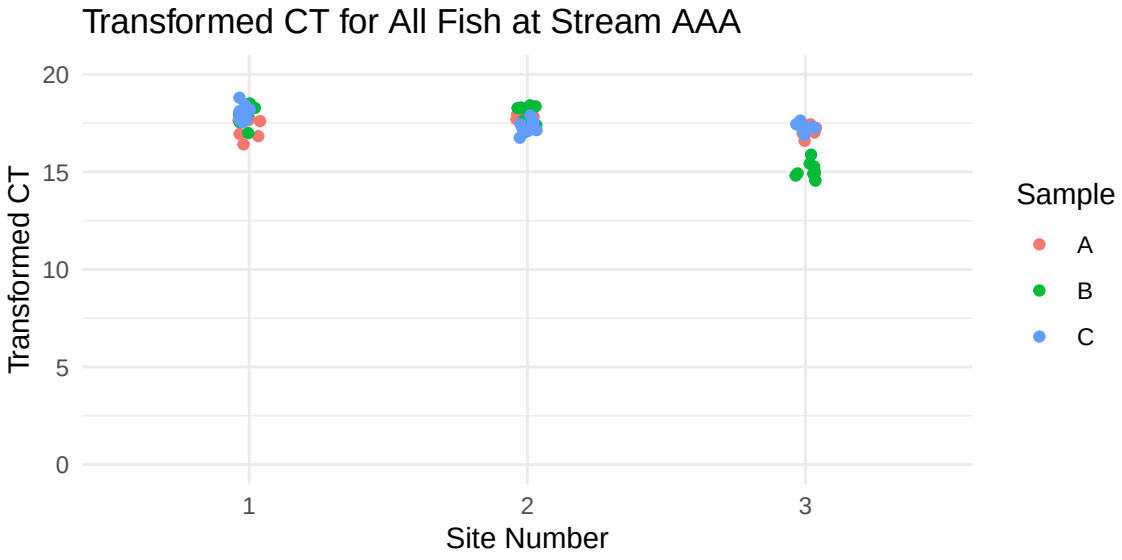

Figure 5.2: Transformed CT values of the technical replicates for All Fish at Stream AAA.

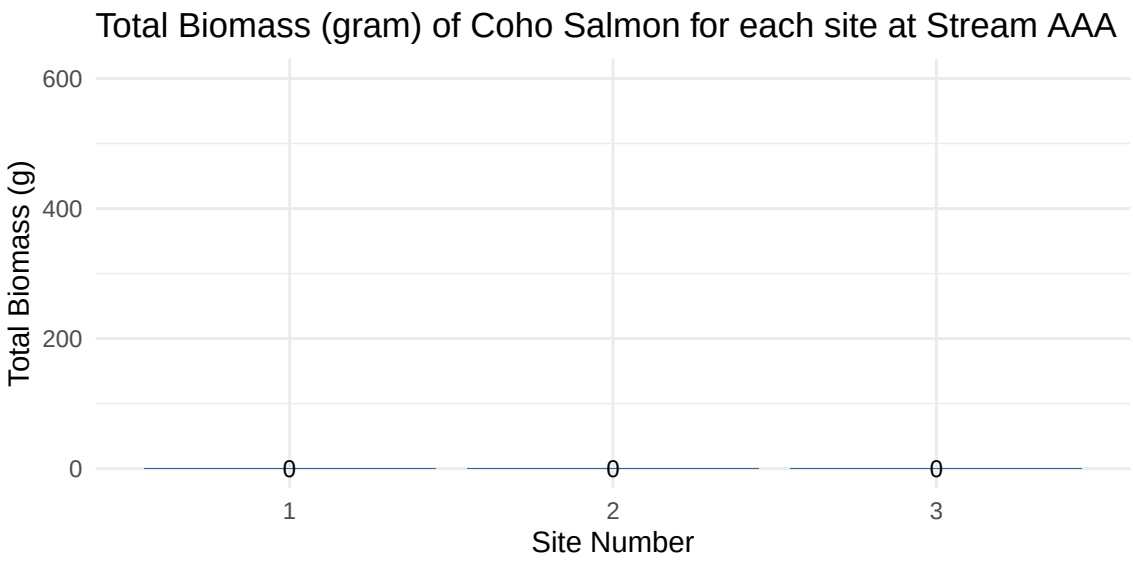

Figure 5.3: Total Biomass for Coho Salmon at Stream AAA.

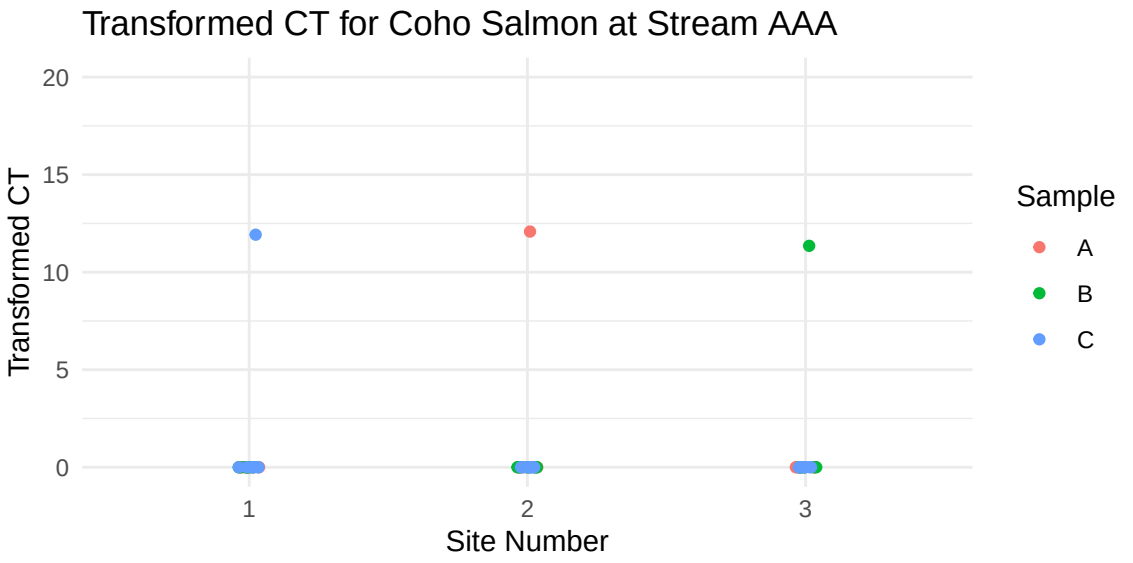

Figure 5.4: Transformed CT values of the technical replicates for Coho Salmon.

Figure 5.1 is a plot of the total biomass for All Fish species collected from the three sites at stream AAA. At each site, fish were caught in the transects. Figure 5.2 is a plot of the associated TCT values obtained from each site using a primer that amplifies all fish. Figure 5.3 shows that no Coho Salmon were caught at stream AAA.

Although Figure 5.4 indicates that there was trace detection of Coho at each of the three sites.

5.1.2 Stream BBB

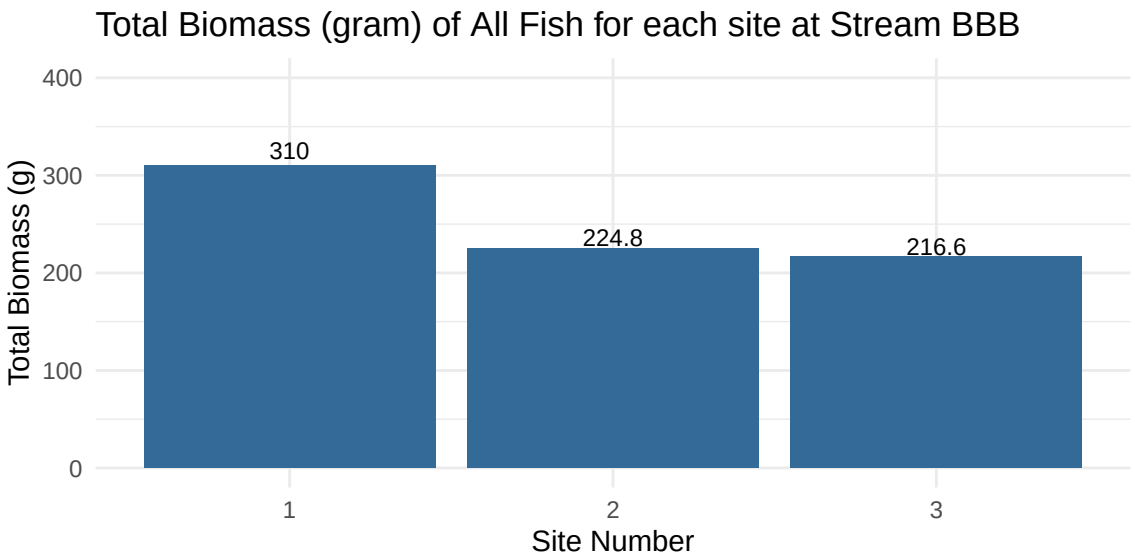

Figure 5.5: Total biomass of All Fish at each site for Stream BBB.

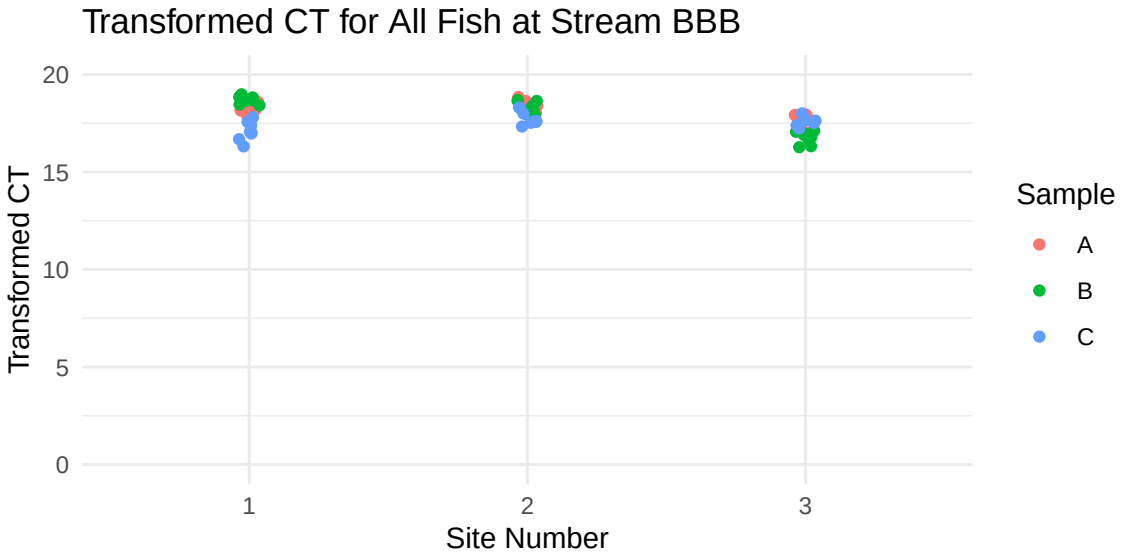

Figure 5.6: Transformed CT values of the technical replicates for all Fish.

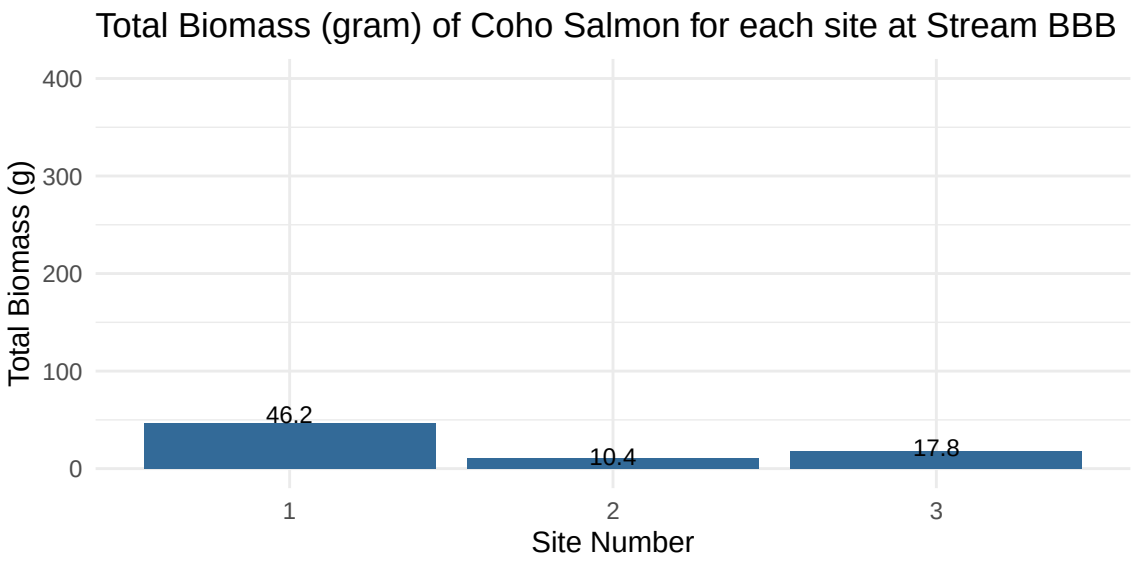

Figure 5.7: Total biomass for Coho Salmon at each site for Stream BBB.

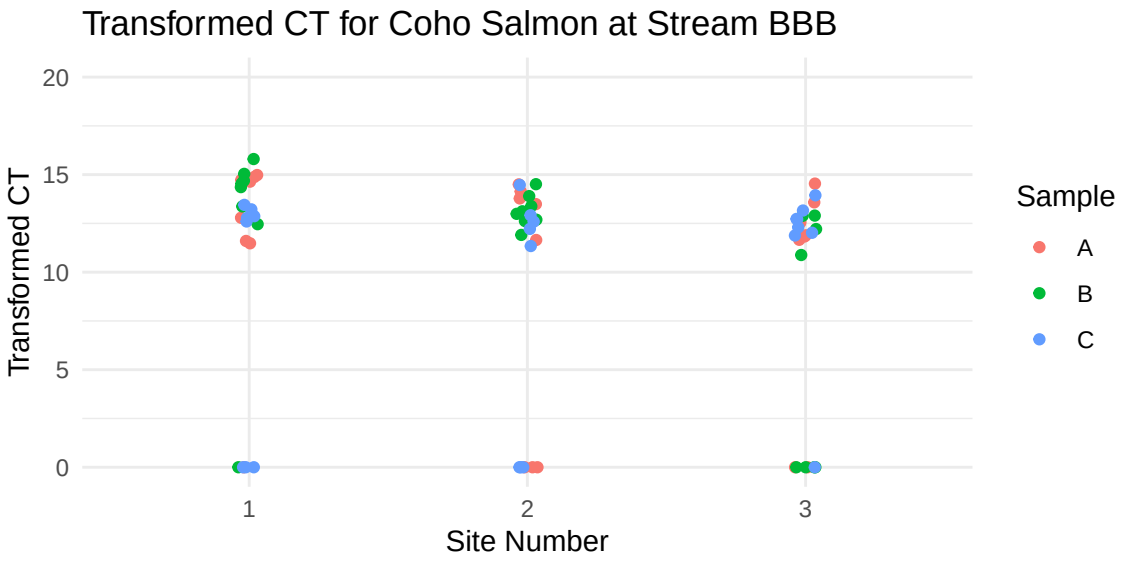

Figure 5.8: Transformed CT values of the technical replicates for Coho Salmon.

Figure 5.5 is a plot of the total biomass for All Fish species collected from the three sites at stream BBB. At each site, fish were caught in the transects. Site number 1 had the highest recorded biomass, although each site had less biomass than those caught in stream AAA.

Figure 5.6 is a plot of the associated TCT values obtained from each site at stream BBB using a primer that amplifies all fish. We see that each site detected relatively high TCT values indicating the presence of fish.

Figure 5.7 is a plot of the Coho Salmon that were caught at stream BBB. At each site in stream BBB, Coho were caught. Figure 5.8 confirms the presence of Coho by the moderate TCT values.

5.1.3 Stream CCC

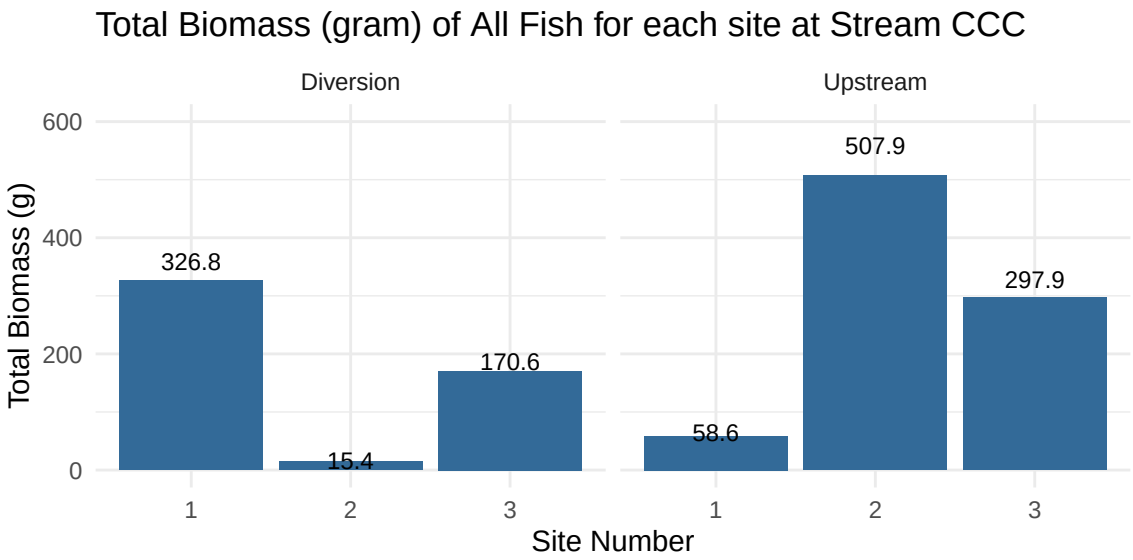

Figure 5.9: Total biomass of All Fish at each site for Stream CCC.

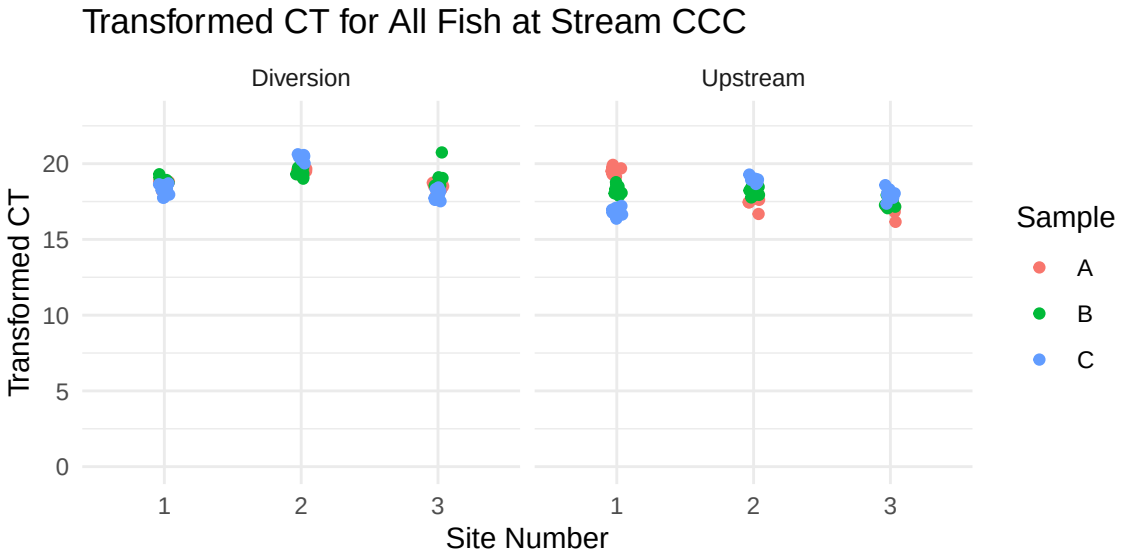

Figure 5.10: Transformed CT values of the technical replicates for all Fish.

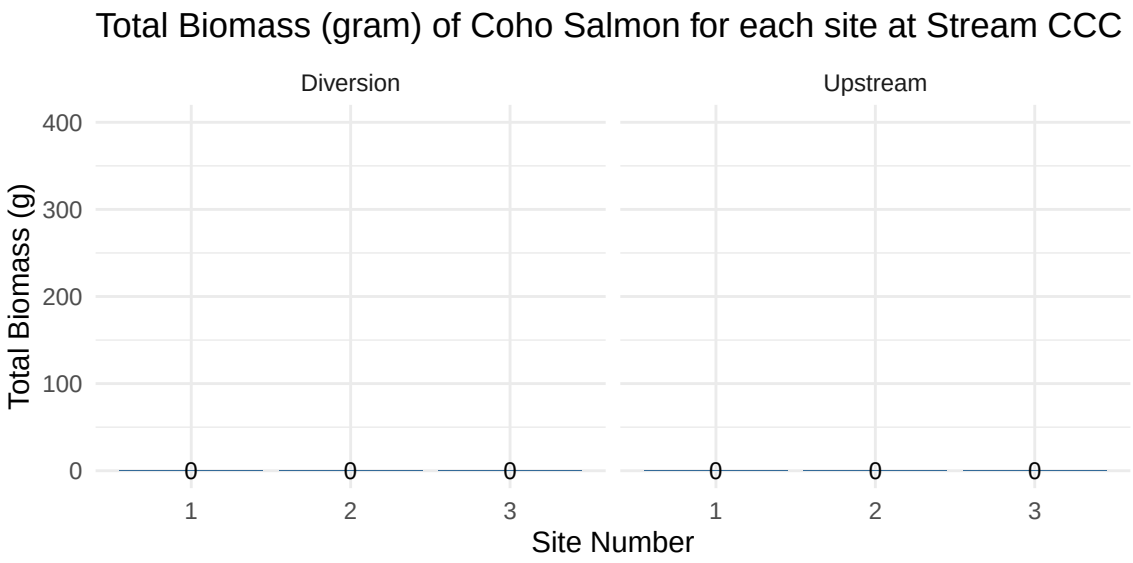

Figure 5.11: Total biomass for Coho Salmon at each site for Stream CCC.

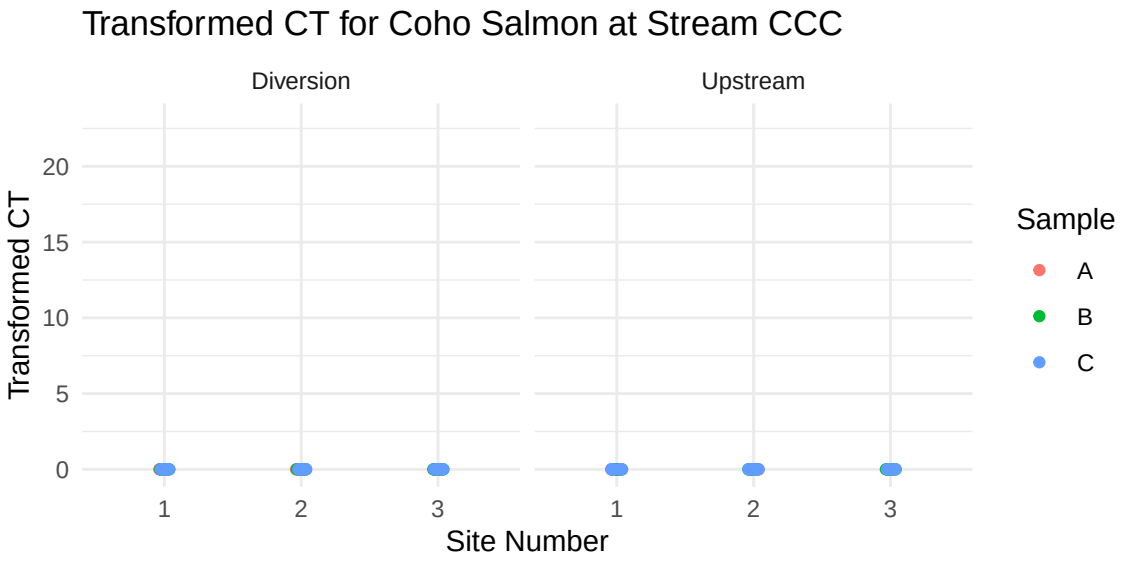

Figure 5.12: Transformed CT values of the technical replicates for Coho Salmon.

Stream CCC had two distinct sections, a diversion and a upstream portion. Measurements were thus taken from three distinct sites in each of these distinct portions. Figure 5.9 shows the biomass of All Fish caught at stream CCC. Figure 5.10 shows the associated TCT values that were obtained from the collected samples. We see from Figure 5.11 that no Coho were caught at any portion of this stream. Moreover,

Figure 5.12 had no detection of any Coho in any sample.

5.1.4 Stream DDD

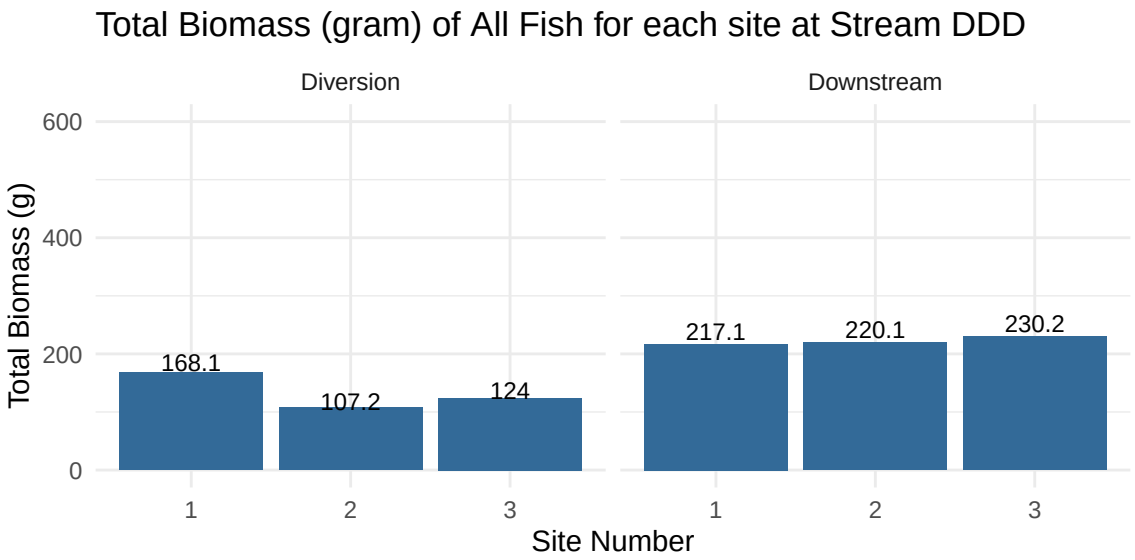

Figure 5.13: Total biomass for All Fish at each site for Stream DDD.

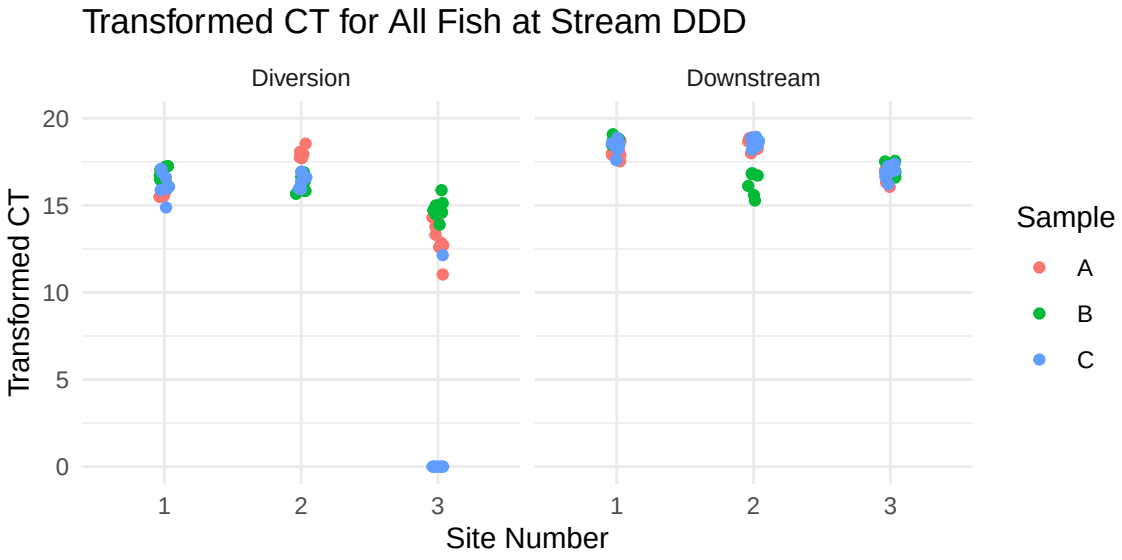

Figure 5.14: Transformed CT values of the technical replicates for All Fish.

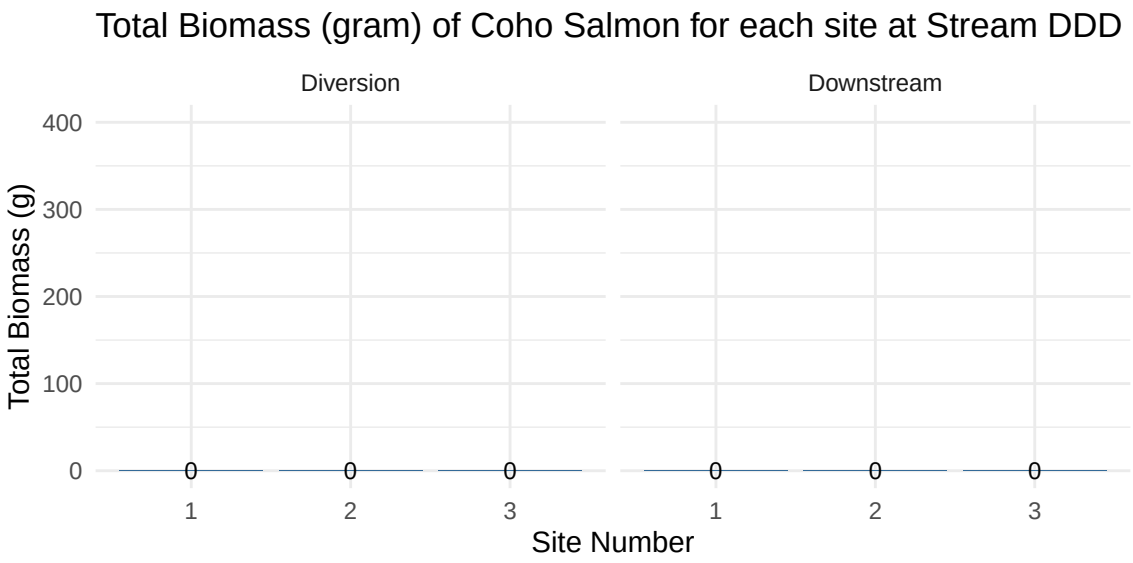

Figure 5.15: Total biomass for Coho Salmon at each site for Stream DDD.

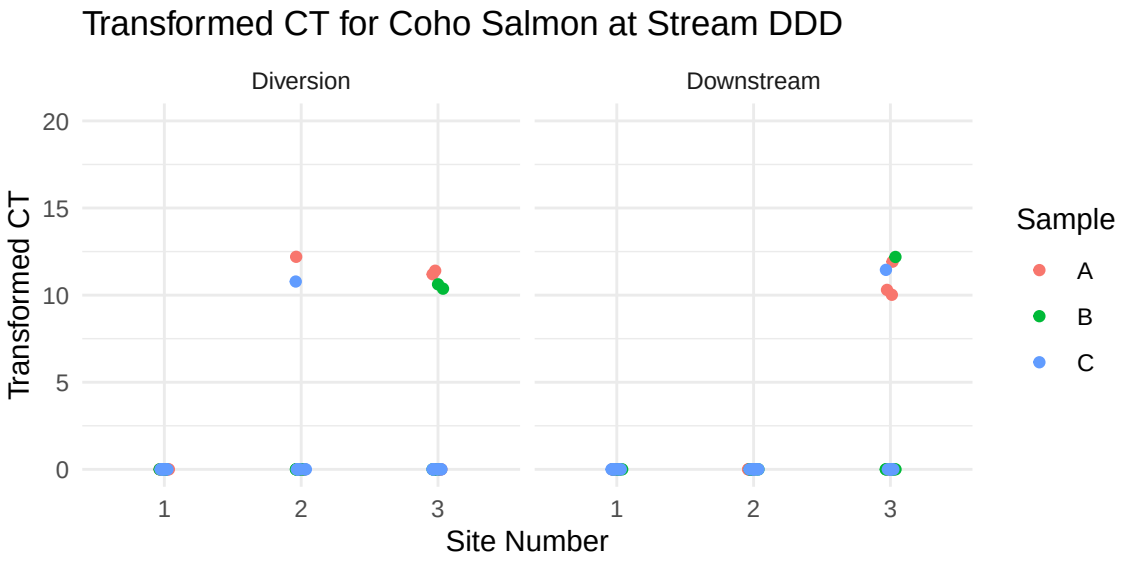

Figure 5.16: Transformed CT values of the technical replicates for Coho Salmon.

Stream DDD had two distinct sections, a diversion and a downstream portion. Measurements were thus taken from three distinct sites in each of these distinct portions. Figure 5.13 shows the biomass of all fish caught at stream DDD. Figure 5.14 shows the associated TCT values that were obtained from the collected samples. We see from Figure 5.15 that no Coho were caught at any portion of this stream. How-

ever, Figure 5.16 indicates trace detection of Coho Salmon at both the diversion and downstream.

5.2 Pairs Plots

We now examine how several covariates are associated with one another. We visualize these possible relationships using a 'pairs plot'. These plots help illustrate how several variables are possibly correlated. We consider the density and flow terms, as well as previously identified important environmental covariates. In the case of all Fish, one outlier was removed after failing to pass the DNA integrit-E test (this observation was from stream DDD, diversion). The outlier can be seen in the first pairs plot as a blue dot in the bottom left of the first image, in the portion comparing Fish.Total.Biomass.g with MeanTCTEf. The pairs plot compares each predictor with one-another. For example, the first row contains a scatter plot of mean TCT versus each of the other covariates.

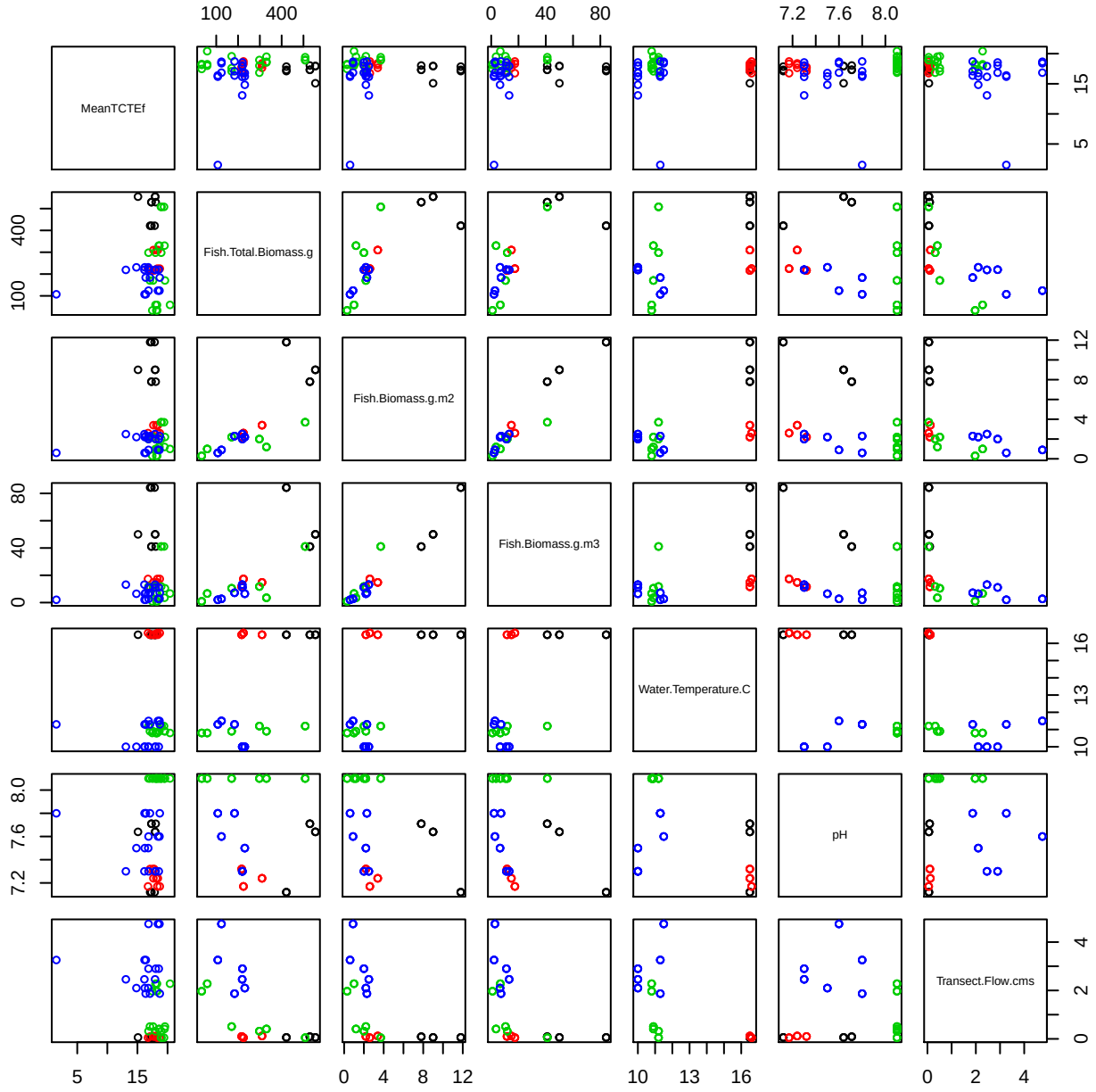

Figure 5.17: Pairs plots for All Fish with outlier included. We consider several suspected key covariates. Black corresponds to Stream AAA, red corresponds to Stream BBB, green corresponds to Stream CCC and blue corresponds to Stream DDD.

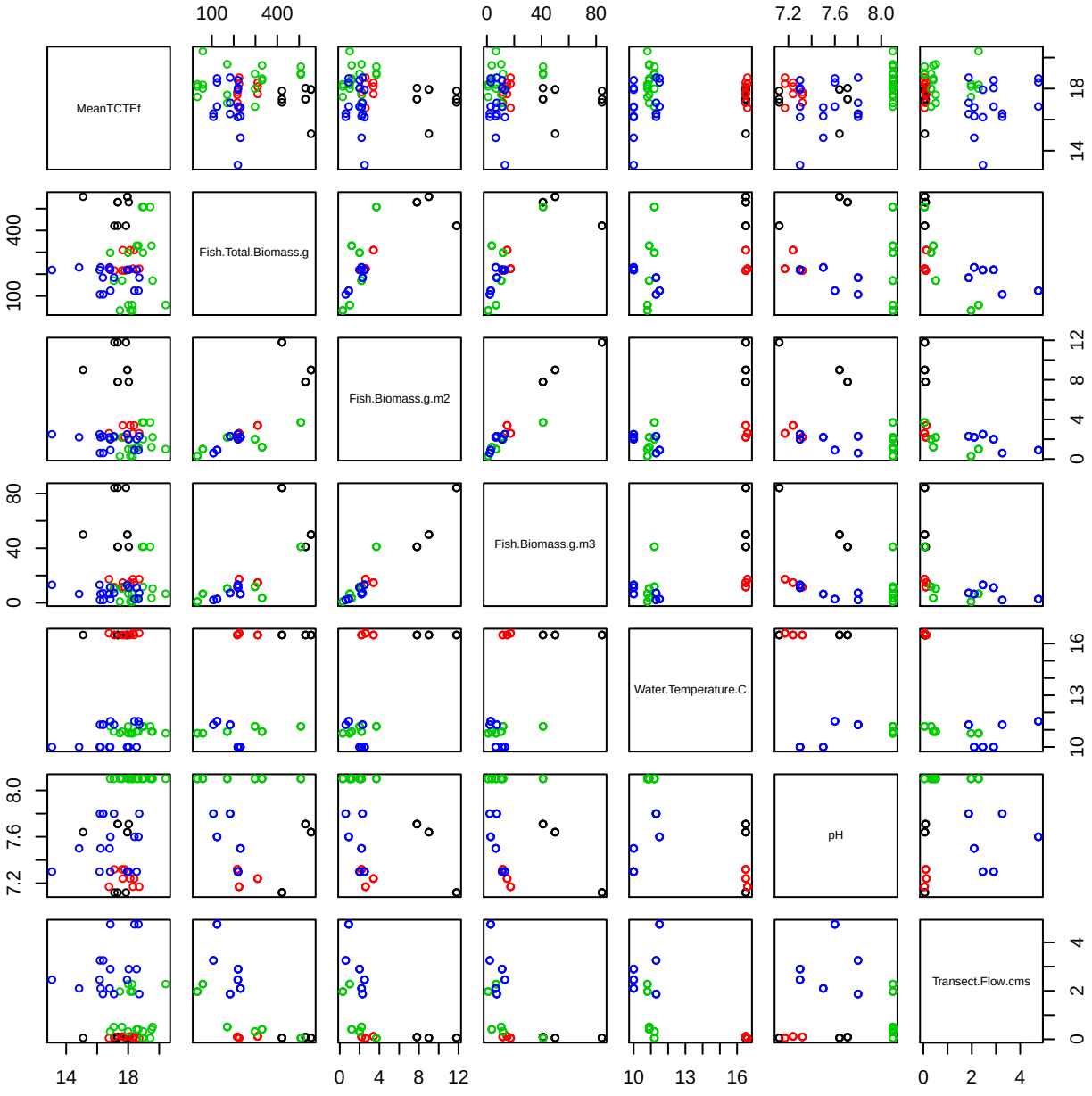

Figure 5.18: Pairs plots for All Fish with outlier removed. We consider several suspected key covariates. Black corresponds to Stream AAA, red corresponds to Stream BBB, green corresponds to Stream CCC and blue corresponds to Stream DDD.

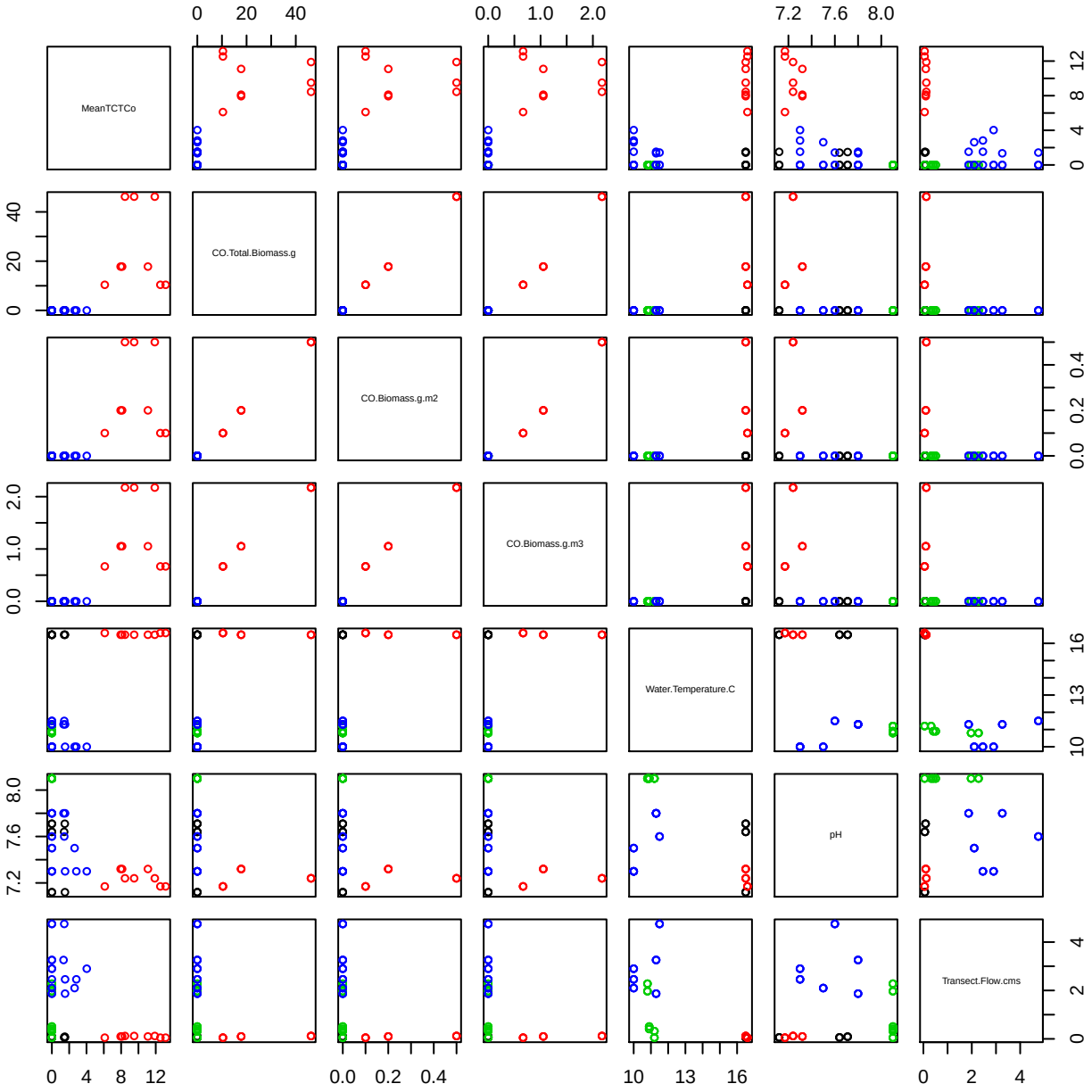

Figure 5.19: Pairs plots for Coho Salmon. Black corresponds to Stream AAA, red corresponds to Stream BBB, green corresponds to Stream CCC and blue corresponds to Stream DDD.

The Pairs Plots confirm several of our assumptions. MeanTCT tends to increase linearly as biomass increases. MeanTCT also tends to decrease linearly as Transect.Flow.cms increases. Figure 5.17 is the pairs plot for all fish in which the outlier was included. Figure 5.18 is the same pairs plot but the outlier from stream DDD diversion was removed. Not much can be analyzed from these plots as there does not appear to be any obvious patterns. However, we see from the colors that samples from the same stream tend to cluster together. Figure 5.19 is the pairs plot for Coho. We see from observing the first row that MeanTCT appears to increase linearly with CO.Total.Biomass.g, CO.Biomass.g.m2 and CO.Biomass.g.m3. MeanTCT also appears to possibly increase with higher water temperatures (Water.Temperature.C), and possibly decreases with higher levels of pH. MeanTCT also appears to decrease as Transect.Flow.cms increases.

5.3 TCT versus Biomass

Although several covariates likely influence mean transformed CT scores, we expect a relationship between biomass and TCT. We visualize this one variable (mean over eight technical replicates) using a simple linear regression model before the addition of covariates.

Firstly, we consider linear models with biomass as the only predictor for Mean TCT. Biomass refers to the weight in grams of the fish caught in the transect using electrofishing.

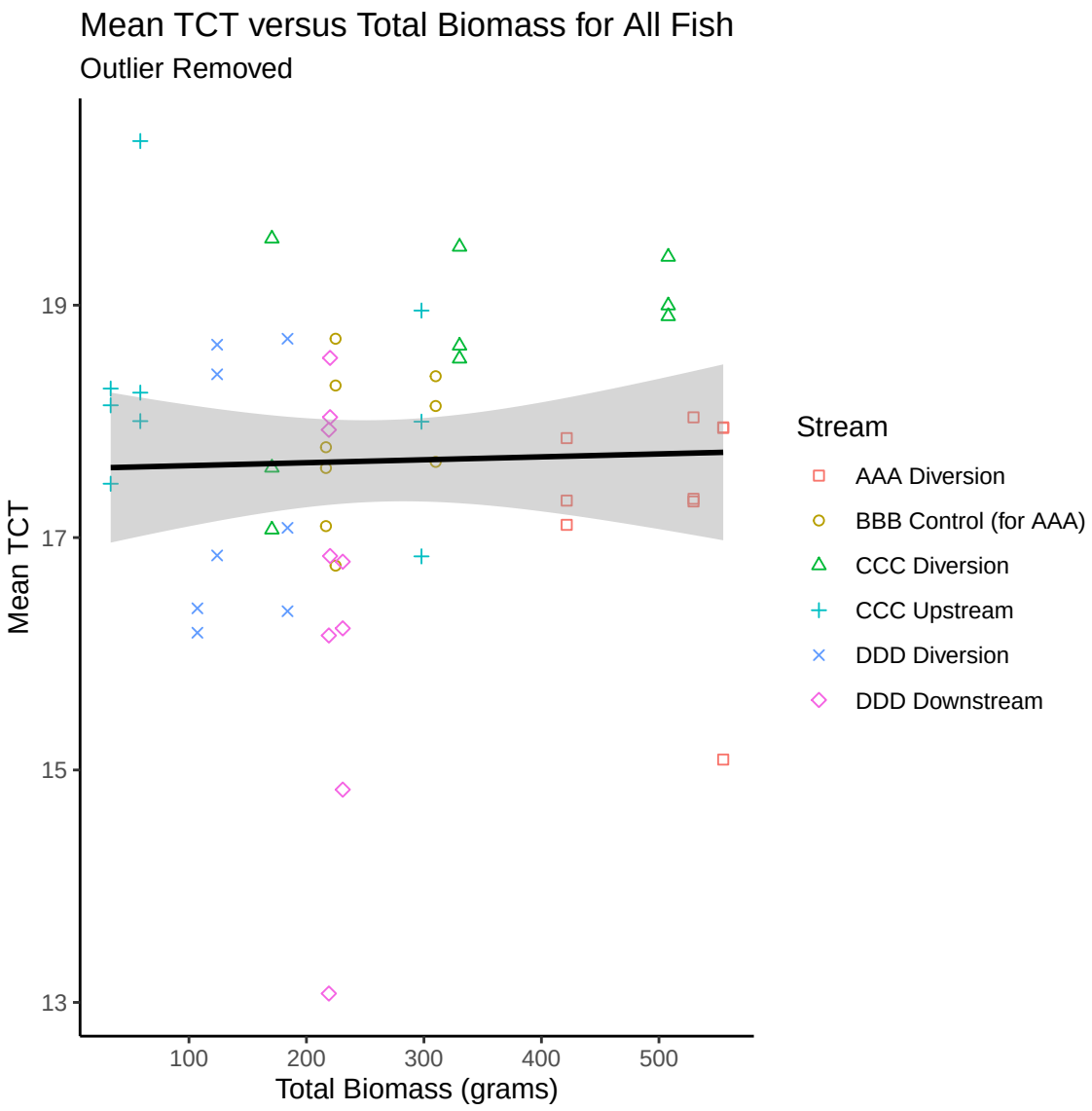

Figure 5.20: Mean TCT over each set of eight technical replicates versus Total Biomass for All Fish, outlier removed. Included is the simple linear regression model and the 95% Confidence limits for the regression line.

```
Call:
lm(formula = MeanTCTEf ~ Fish.Total.Biomass.g,
data = field.removeef)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5728 -0.5903  0.2099  0.7799  2.8036

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          17.593875   0.354609  49.615   <2e-16 ***
Fish.Total.Biomass.g  0.000252   0.001159   0.217    0.829
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.272 on 51 degrees of freedom
Multiple R-squared:  0.0009267,Adjusted R-squared:  -0.01866
F-statistic: 0.0473 on 1 and 51 DF,  p-value: 0.8287
```

Table 5.1: A model that considers biomass for All Fish. Model: model.fish.

Figure 5.20 is the plot of the model, model.fish and the associated confidence limits. The line appears flat, and no obvious relationship is clear from a simple visual consideration. Table 5.1 provides a summary of the simple linear model (model.fish) fit for All Fish. The R^2 is only 0.00093. Although the estimate for the coefficient of biomass is still positive, the estimate is not statistically significant.

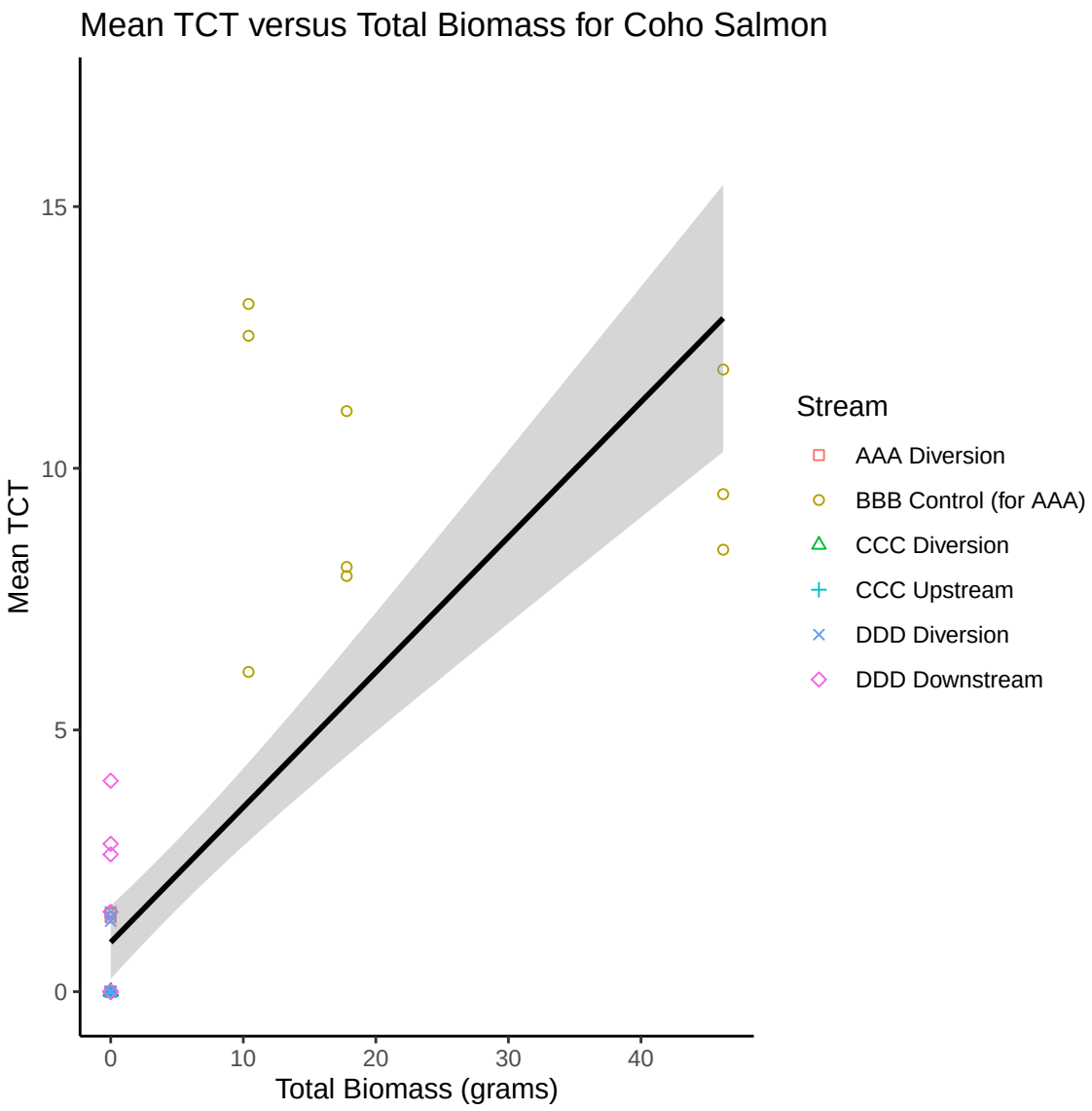

Figure 5.21: Mean TCT over each set of eight technical replicates versus Total Biomass for Coho Salmon. Included is the simple linear regression model (Biomass as the only predictor) and the 95% Confidence limits for the regression line.

```
Call:
lm(formula = MeanTCTCo ~ CO.Total.Biomass.g,
data = field.collapse)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4231 -0.9424 -0.9424  0.4860  9.5130

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.94242    0.34847   2.704  0.00923 **
CO.Total.Biomass.g  0.25813    0.02922   8.833 6.29e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.402 on 52 degrees of freedom
Multiple R-squared:  0.6001,Adjusted R-squared:  0.5924
F-statistic: 78.02 on 1 and 52 DF,  p-value: 6.294e-12
```

Table 5.2: Model: model.co.

For Coho Salmon, our estimated positive of coefficient of +0.258 on biomass is statistically significant. Figure 5.21 makes the relationship clear. Mean TCT increases as CO.Total.Biomass.g increases. Table 5.3 provides the summary of the simple linear model for mean TCT for Coho. The simple linear model only considered the predictor CO.Total.Biomass.g and achieved an R^2 of 0.600

5.3.1 Environmental Factors

We now consider the possible impact of covariates on the Mean TCT value. We include summary statistics of several of those covariates that have been shown to impact eDNA degradation such as pH and water temperature. We also include several biomass related terms such as total biomass, total biomass per meter squared of transect and total biomass per meter cubed of transect. We also include a ‘flow’ term by considering the flow through the transect.

Stream	Site	Reach	All Fish	Coho	Cutthroat	Rainbow
AAA	1	Diversion	422	0	301	0
AAA	2	Diversion	555	0	261	0
AAA	3	Diversion	529	0	373	0
BBB	1	Control (for AAA)	310	46	190	0
BBB	2	Control (for AAA)	225	10	214	0
BBB	3	Control (for AAA)	217	18	198	0
CCC	1	Diversion	330	0	0	327
CCC	1	Upstream	33	0	0	15
CCC	2	Diversion	171	0	0	171
CCC	2	Upstream	59	0	0	59
CCC	3	Diversion	508	0	0	508
CCC	3	Upstream	298	0	0	298
DDD	1	Diversion	184	0	0	168
DDD	1	Downstream	219	0	0	217
DDD	2	Diversion	107	0	0	107
DDD	2	Downstream	220	0	0	220
DDD	3	Diversion	124	0	0	124
DDD	3	Downstream	231	0	0	230

Table 5.3: Table showing the total biomass (g) of each species captured in the transect for each site.

Table 5.3 summarizes the total biomass of Coho, Cutthroat, Rainbow and all Fish that were caught in the transects in each site at each stream. Coho was only caught in stream BBB.

Stream	Site	Reach	All Fish	Coho	Cutthroat	Rainbow
AAA	1	Diversion	11.8	0.0	8.4	0.0
AAA	2	Diversion	9.0	0.0	4.2	0.0
AAA	3	Diversion	7.8	0.0	5.5	0.0
BBB	1	Control (for AAA)	3.4	0.5	2.1	0.0
BBB	2	Control (for AAA)	2.6	0.1	2.4	0.0
BBB	3	Control (for AAA)	2.2	0.2	2.0	0.0
CCC	1	Diversion	1.2	0.0	0.0	1.2
CCC	1	Upstream	0.3	0.0	0.0	0.1
CCC	2	Diversion	2.2	0.0	0.0	2.2
CCC	2	Upstream	1.0	0.0	0.0	1.0
CCC	3	Diversion	3.7	0.0	0.0	3.7
CCC	3	Upstream	2.0	0.0	0.0	2.0
DDD	1	Diversion	2.3	0.0	0.0	2.1
DDD	1	Downstream	2.5	0.0	0.0	2.4
DDD	2	Diversion	0.6	0.0	0.0	0.6
DDD	2	Downstream	2.0	0.0	0.0	2.0
DDD	3	Diversion	0.9	0.0	0.0	0.9
DDD	3	Downstream	2.2	0.0	0.0	2.2

Table 5.4: Table showing the total biomass (g) of each species captured in the transect per meter squared of transects for each site.

Stream	Site	Reach	All Fish	Coho	Cutthroat	Rainbow
AAA	1	Diversion	84.3	0.0	60.0	0.0
AAA	2	Diversion	50.0	0.0	23.3	0.0
AAA	3	Diversion	41.1	0.0	28.9	0.0
BBB	1	Control (for AAA)	14.8	2.2	9.1	0.0
BBB	2	Control (for AAA)	17.3	0.7	16.0	0.0
BBB	3	Control (for AAA)	11.6	1.1	10.5	0.0
CCC	1	Diversion	3.5	0.0	0.0	3.5
CCC	1	Upstream	0.9	0.0	0.0	0.3
CCC	2	Diversion	10.5	0.0	0.0	10.5
CCC	2	Upstream	6.7	0.0	0.0	6.7
CCC	3	Diversion	41.1	0.0	0.0	41.1
CCC	3	Upstream	11.8	0.0	0.0	11.8
DDD	1	Diversion	7.2	0.0	0.0	6.6
DDD	1	Downstream	13.2	0.0	0.0	12.6
DDD	2	Diversion	2.1	0.0	0.0	2.1
DDD	2	Downstream	11.1	0.0	0.0	11.1
DDD	3	Diversion	2.7	0.0	0.0	2.7
DDD	3	Downstream	6.5	0.0	0.0	6.5

Table 5.5: Table showing the total biomass (g) of each species captured in the transect per meter cubed of transects for each site.

Stream	Year	Site	Reach	Temperature	pH	Flow	Depth	Distance
AAA	2017	1	Diversion	16.5	7.12	0.06	0.15	0.1
AAA	2017	2	Diversion	16.5	7.64	0.06	0.30	2.0
AAA	2017	3	Diversion	16.5	7.71	0.09	0.45	1.0
BBB	2017	1	Control (for AAA)	16.5	7.24	0.12	0.20	6.0
BBB	2017	2	Control (for AAA)	16.6	7.17	0.05	0.30	2.0
BBB	2017	3	Control (for AAA)	16.5	7.32	0.10	0.20	1.5
CCC	2018	1	Diversion	10.9	8.10	0.41	0.30	1.0
CCC	2018	1	Upstream	10.8	8.10	1.97	0.40	2.0
CCC	2018	2	Diversion	10.9	8.10	0.51	0.40	1.0
CCC	2018	2	Upstream	10.8	8.10	2.28	0.30	2.0
CCC	2018	3	Diversion	11.2	8.10	0.05	0.30	4.5
CCC	2018	3	Upstream	11.2	8.10	0.32	0.25	2.0
DDD	2017	1	Diversion	11.3	7.80	1.87	0.70	3.0
DDD	2017	1	Downstream	10.0	7.30	2.46	0.45	3.0
DDD	2017	2	Diversion	11.3	7.80	3.26	0.25	1.0
DDD	2017	2	Downstream	10.0	7.30	2.90	0.45	4.0
DDD	2017	3	Diversion	11.5	7.60	4.75	0.50	1.0
DDD	2017	3	Downstream	10.0	7.50	2.10	0.30	3.0

Table 5.6: Table showing a variety of summary statistics taken over each site. Temperature is in Celsius, pH is a scale and Transect Flow is in (cm/s), Depth is the depth of the sample area in meters, and Distance is the distance in meters from shore in which the sample was taken.

Table 5.4 summarizes the weight of each species caught per meter squared of transect. Table 5.5 shows the weight of each species caught per volume of transect, calculated by using the depth of the transect. The reason that the weight per volume is larger than the weight per area is due to the fact that the transect depth was often very shallow and less than 1 meter. Table 5.6 summarizes a variety of covariates that were recorded at each stream. These environmental covariates are taken into account in our statistical models.

5.3.2 Models with Covariates

We first use *Backward-elimination* to determine a suitable model for mean TCT for Coho. Other than biomass per meter cubed of transect (CO.Biomass.g.m3), the original covariates that we consider are the water temperature in Celsius (Water.Temperature.C), the pH, the flow rate (Transect.Flow.cms), and the interaction between Transect.Flow.cms and CO.Biomass.g.m3. We also consider sampling covariates such as the sampling depth (eDNA.Water.Sampling.Depth.cm) and the distance from shore in which the sample was collected (eDNA.Distance.from.Shore.m).

The initial 'full' model is named model.coho.full;

```
Call:
lm(formula = MeanTCTCo ~ Water.Temperature.C + pH + CO.Biomass.g.m3 +
    CO.Biomass.g.m3 * Transect.Flow.cms + eDNA.Distance.from.Shore.m +
    eDNA.Total.Water.Depth.m,
 data = field.collapse)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7698 -0.5790 -0.0741  0.2321  2.8319

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.151e+01  7.059e+00   1.631   0.1097
Water.Temperature.C              -6.760e-02  1.376e-01  -0.491   0.6256
pH                               -1.343e+00  7.305e-01  -1.839   0.0724 .
CO.Biomass.g.m3                   2.251e+01  2.136e+00  10.540 7.44e-14 ***
Transect.Flow.cms                -2.655e-02  2.079e-01  -0.128   0.8990
eDNA.Distance.from.Shore.m       -4.268e-03  1.947e-01  -0.022   0.9826
eDNA.Total.Water.Depth.m          5.742e-01  1.690e+00   0.340   0.7355
CO.Biomass.g.m3:Transect.Flow.cms -1.515e+02  1.834e+01  -8.263 1.21e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.264 on 46 degrees of freedom
Multiple R-squared:  0.902,Adjusted R-squared:  0.887
F-statistic: 60.46 on 7 and 46 DF,  p-value: < 2.2e-16
```

Table 5.7: Model: model.coho.full

Table 5.7 is a summary of the full model, model.coho.full. The fitted model indicates high levels of significance for CO.Biomass.g.m3, and a high significance estimate for the interaction between CO.biomass.g.m3 and Transect.Flow.cms.

Backward elimination is performed using 'stepAIC' (Venables and Ripley, 2002) with the direction argument specified as 'backward'. We first feed a full model, with all possible covariates into the algorithm and it proceeds as follows; R determines which predictor (if any) when removed will decrease the AIC the most. Once R determines that there are no predictors left that can be removed (i.e. there is no predictor that can be removed that results in a decreased AIC), the algorithm terminates. Akaike Information Criteria (AIC) allows us to compare similar models. One method to select a suitable model would be to select the model with the lowest AIC value. Note that this is because $AIC = 2 * p - 2\log(\hat{L})$ where p is the number of predictors and $\hat{L}$ is the maximum value of the likelihood function. AIC accounts for overfitting by penalizing the addition of more predictors. After comparing three biomass parameters (CO.total.Biomass.g, CO.Biomass.g.m2 and CO.Biomass.g.m3) we decided to work with CO.Biomass.g.m3 as this quantity can be interpreted as a fish 'density' parameter. We also include a scope argument to the stepAIC algorithm which ensures that CO.Biomass.g.m3 is always included in the model.

```
Call:
lm(formula = MeanTCTCo ~ pH + CO.Biomass.g.m3 + Transect.Flow.cms +
    CO.Biomass.g.m3:Transect.Flow.cms,
data = field.collapse)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7450 -0.5107 -0.0927  0.3443  2.9848

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          9.00093    4.49860   2.001    0.051 .
pH                                  -1.11550    0.57236  -1.949    0.057 .
CO.Biomass.g.m3                     22.29860    2.04182  10.921 9.97e-15 ***
Transect.Flow.cms                    0.06463    0.13448   0.481    0.633
CO.Biomass.g.m3:Transect.Flow.cms -150.50218   16.96964  -8.869 9.25e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.232 on 49 degrees of freedom
Multiple R-squared:  0.9009,Adjusted R-squared:  0.8928
F-statistic: 111.4 on 4 and 49 DF,  p-value: < 2.2e-16
```

Table 5.8: Model: model.co.step.

For Coho, stepwise selection of covariates indicates we should include several covariates as seen from Table 5.8. CO.Biomass.g.m3 is a significant predictor, as well as the interaction between Transect.Flow.cms and CO.biomass.g.m3. We see as CO.biomass.g.m3, mean TCT also increases. The interaction between CO.biomass.g.m3 and Transect.Flow.cms indicate that higher levels of flow may result in lower levels of mean TCT, as we have a highly negative estimate for the interaction term. pH also appears to have an impact; whereby higher pH water may result in lower estimates of TCT. By including Transect.Flow.cms and the interactions between Transect.Flow.cms and CO.Biomass.g.m3, we have greatly increased the R^2 to 0.901 and we have an adjusted R^2 of 0.893.

The highly negative estimated coefficient for the interaction can be explained by an examination of Figure 5.19 and Table 5.6. The interaction term between CO.Biomass.g.m3 and Transect.Flow.cms . This may imply that larger fish are more greatly impacted by flow. That is, at higher levels of flow, the biomass of a species

may become less important in determining mean TCT. Moreover, the large magnitude of the estimate is partially due to the fact that the term Transect.Flow.cms is in units of cm/s.

Overall, these models indicate the significance of a few key covariates. When attempting to model mean TCT, researchers may wish to first include covariates such as pH and flow rate, as well as possible interactions between the flow rate and biomass. In our analysis, higher levels of Transect.Flow.cms tended to result in lower estimates of mean TCT (likely due to the 'washing away' of the eDNA). Higher levels of pH appear to result in lower TCT, possibly due to degradation of eDNA. Researchers may wish to start with a model that includes all predictors of interest and remove ecological covariates one at time using backward-elimination.

5.3.3 Model Averaging

We now explore the concept of 'Model Averaging'. In particular, we make use of the R package 'MuMIn: Multi-Model Inference' (Barton, 2020). This package also assists in model selection. MuMIn also provides us with confidence intervals for our parameters. We use the R function 'dredge' to generate a model selection table (ma.coho) using the full model, model.co.full. dredge creates and evaluates multiple subsets of the full global model. Note that 'Adjusted AIC' or 'AICc' is defined as $AICc = AIC + \frac{2p^2+2*p}{N-p-1}$. Adjusted AIC is often preferred when the sample size, N, is small.

For Coho, the results are listed in Table 5.9 and Table 5.10.

```
Call:
model.avg(object = ma.coho, subset = delta < 4)

Component model call:
lm(formula = MeanTCTCo ~ <8 unique rhs>,
data = field.collapse)


Term Codes:
CO.Biomass.g.m3:  1
eDNA.Distance.from.Shore.m: 2
eDNA.Total.Water.Depth.m: 3
pH: 4
Transect.Flow.cms: 5
Water.Temperature.C: 6
CO.Biomass.g.m3:Transect.Flow.cms: 7


Component models:
      df logLik AICc delta weight
1457   6  -85.2  184  0.00   0.35
157    5  -87.3  186  1.49   0.17
14567  7  -85.0  187  2.21   0.12
12457  7  -85.1  187  2.42   0.11
13457  7  -85.2  187  2.48   0.10
1567   6  -87.2  188  3.83   0.05
1257   6  -87.2  188  3.85   0.05
1357   6  -87.2  188  3.98   0.05
```

Table 5.9: Model Average object for Coho.

Table 5.9 lists the associated term codes and components for the model average object. The term codes are used to save space and are used as a numeric reference to our predictors. The component models section is a list of the variety of differing

models in which we are averaging over. The highest weighted model (with a weight of $w = 0.35$) is the model containing CO.Biomass.g.m3, pH, Transect.Flow.cms and the interaction between CO.Biomass.g.m3 and Transect.Flow.cms. This is indeed the same model chosen using backward selection.The next highest weighted model (with a weight of $w = 0.17$) is simply the first model but with pH removed. We specify a delta of less than 4. That means R only returns to us component models in which the adjusted AIC does not differ by more than 4. The delta term represents the difference in adjusted AIC each model is from the best model. For example, the second best model which contains CO.Biomass.g.m3, Transect.Flow.cms and CO.Biomass.g.m3:Transect.Flow.cms differs in adjusted AIC from the highest weighted model by 1.49. Our components are thus all combinations of predictors such that the adjusted AIC is within 4 of the best performing model.

```
Model-averaged coefficients:
(full average)
                   Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept)          6.4396     5.9483      6.0174    1.07     0.28
Biomass.g.m3        22.8829     2.1605      2.2085   10.36   <2e-16 ***
pH                  -0.7849     0.7296      0.7380    1.06     0.29
Flow                 0.0748     0.1502      0.1536    0.49     0.63
Biomass.g.m3:Flow -154.6327    17.8121     18.2250    8.48   <2e-16 ***
Temperature         -0.0057     0.0493      0.0502    0.11     0.91
Distance.from.Shore 0.0100     0.0627      0.0641    0.16     0.88
Total.Water.Depth.m 0.0778     0.6357      0.6506    0.12     0.90

(conditional average)
                     Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept)           6.4396     5.9483      6.0174    1.07    0.285
Biomass.g.m3         22.8829     2.1605      2.2085   10.36   <2e-16 ***
pH                   -1.1549     0.5966      0.6117    1.89    0.059 .
Flow                  0.0748     0.1502      0.1536    0.49    0.626
Biomass.g.m3:Flow  -154.6327    17.8121     18.2250    8.48   <2e-16 ***
Temperature          -0.0337     0.1157      0.1181    0.29    0.776
Distance.from.Shore.m  0.0637     0.1469      0.1506    0.42    0.672
Total.Water.Depth.m    0.5155     1.5659      1.6059    0.32    0.748
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 5.10: Coho Model Average Estimates.

Table 5.10 contains the table of estimates for the coefficients. MuMIn works by ranking models according to their adjusted AIC. MuMIn reports two tables of estimates, included in Table 5.10. The first estimates are denoted as ‘full average’ are the estimates obtained via averaging over all possible models (including models in which

this predictor is not present. The second set of estimates, denoted as 'conditional average' are the estimates obtained via averaging over all models in which that predictor is present. The 'full' average coefficient considers models that are more than 4 units away in adjusted AIC.

	2.5 %	97.5 %
(Intercept)	-5.2632496	18.8154064
CO.Biomass.g.m3	18.5243463	27.1781654
pH	-2.3851181	0.0509008
Transect.Flow.cms	-0.2392690	0.3776501
CO.Biomass.g.m3:Transect.Flow.cms	-190.1817103	-118.6624010
Water.Temperature.C	-0.2777040	0.1951246
eDNA.Distance.from.Shore.m	-0.2525418	0.3650015
eDNA.Total.Water.Depth.m	-2.6597126	3.6782683

Table 5.11: 95% Confidence Interval for parameter estimates.

Table 5.11 lists 95 % confidence intervals for the coefficient estimates provided by the conditional-average coefficients. The estimates are similar to those obtained using backward elimination.

5.3.4 Best possible Subset

We now use the R 'leaps' package (based on Fortran code by Alan Miller, 2020) to compare models using the 'Best Subset Method'. This algorithm (regsubsets in R) fits the best possible subset of each subset size. We compare models using three criteria, 'Mallow's Cp' or 'Cp', adjusted R^2, and the Bayesian Information Criterion (BIC). One could also return the 'nbest' models for each subset size if we wished to obtain a list of candidate models. The BIC is defined as $BIC = log(N) * p - 2\log(\hat{L})$ where p is the number of predictors, $\hat{L}$ is the maximum value of the likelihood function and N is sample size. The BIC is similar to the AIC; however it more heavily penalizes complexity compared to the AIC.

Cp is another model selection metric and is defined as

$$Cp = \frac{SSres}{s^2} + 2(p+1) - N$$

Where SSres is the sum of squared residuals for the linear model fit using p parameters. N is the sample size and s^2 is an estimate of the Mean Squared Error.

```
Subset selection object
Call: regsubsets.formula(MeanTCTCo ~ CO.Biomass.g.m3 + Transect.Flow.cms +
    CO.Biomass.g.m3 * Transect.Flow.cms + Water.Temperature.C +
    pH + eDNA.Distance.from.Shore.m + eDNA.Total.Water.Depth.m,
    data = field.collapse, nvmax = 7)
7 Variables  (and intercept)
```

p	AdjR^2	Cp	BIC
1	0.67	100.45	-53.4
2	0.89	3.02	-107.8
3	0.89	0.72	-108.6
4	0.89	2.13	-105.3
5	0.89	4.02	-101.5
6	0.89	6.00	-97.5
7	0.89	8.00	-93.5

Table 5.12: Table summarizing model comparison metrics for Coho Salmon.

Best Subset Method

p	Biomass	Flow	Biomass*Flow	Temp	pH	Shore	Depth
1	*						
2	*		*				
3	*		*		*		
4	*		*	*	*		
5	*		*	*		*	
6	*	*	*	*	*	*	

Table 5.13: Table clarifying which predictors are included when using the best subset method.

Table 5.13 is a chart of which predictors are included in the best model of each specified size (size p). The possible predictors are CO.Biomass.g.m3, Transect.Flow.cms, the interaction between CO.Biomass.g.m3 and Transect.Flow.cms, the pH, the Water.Temperature.C, the distance from shore in meters in which the sample was collected (eDNA.Distance.from.Shore.m) and the depth in meters that the sample was taken (eDNA.Water.Sampling.Depth.m).

The subset size that maximizes the adjusted R^2 and minimizes both the CP and BIC value is 3 and the variables included are CO.Biomass.g.m3, the interaction between CO.Biomass.g.m3 and Transect.Flow.cms and the pH (those are the 3 predictors that when included result in the best model according to adjusted AIC). This model performs well on all three metrics, compared to other models. From the table, we can see what other top models may have been. Compared to the other predictors, the least important predictor appears to be the depth in meters for which the sample was taken. These selection metrics provides further evidence in support of our models.

5.4 Principal Component Analysis

Since Coho were only found to be present in Stream BBB (Table 5.3), we investigate the environmental features that distinguish the streams from one another. Principal Component Analysis ('PCA') is a statistical technique that is used to study a dataset by reducing the number of dimensions. We use the R package 'factoextra' (Kassambara and Mundt, 2020) to compute the principal components of our stream data. PCA works by creating uncorrelated variables that maximize variance. The problem of finding principal components reduces to a simpler problem of solving an eigenvalue/eigenvector problem. PCA preserves much variance as possible while simultaneously providing uncorrelated variables to explain the variation in the data (Jolliffe and Cadima, 2016).

```
Standard deviations (1, .., p=5):
[1] 1.4720592 1.0703764 0.9632790 0.7769547 0.3946781

Rotation (n x k) = (5 x 5):
                     PC1         PC2        PC3         PC4         PC5
Temperature  0.608469460  0.03838818 -0.2252912  0.36284560  0.66774111
pH          -0.357279171 -0.62625821  0.4996590  0.12029873  0.46478092
Flow (cm/s) -0.520550551  0.25356268 -0.4512178 -0.41838266  0.53487524
Depth       -0.480680595  0.22566750 -0.1813422  0.82349396 -0.08362444
Shore       -0.009476261  0.70079038  0.6805161 -0.02617118  0.21216960
```

Table 5.14: Coefficients from Principal Component Analysis on Stream Data.

Table 5.14 is a table of direction vectors for our principal components.

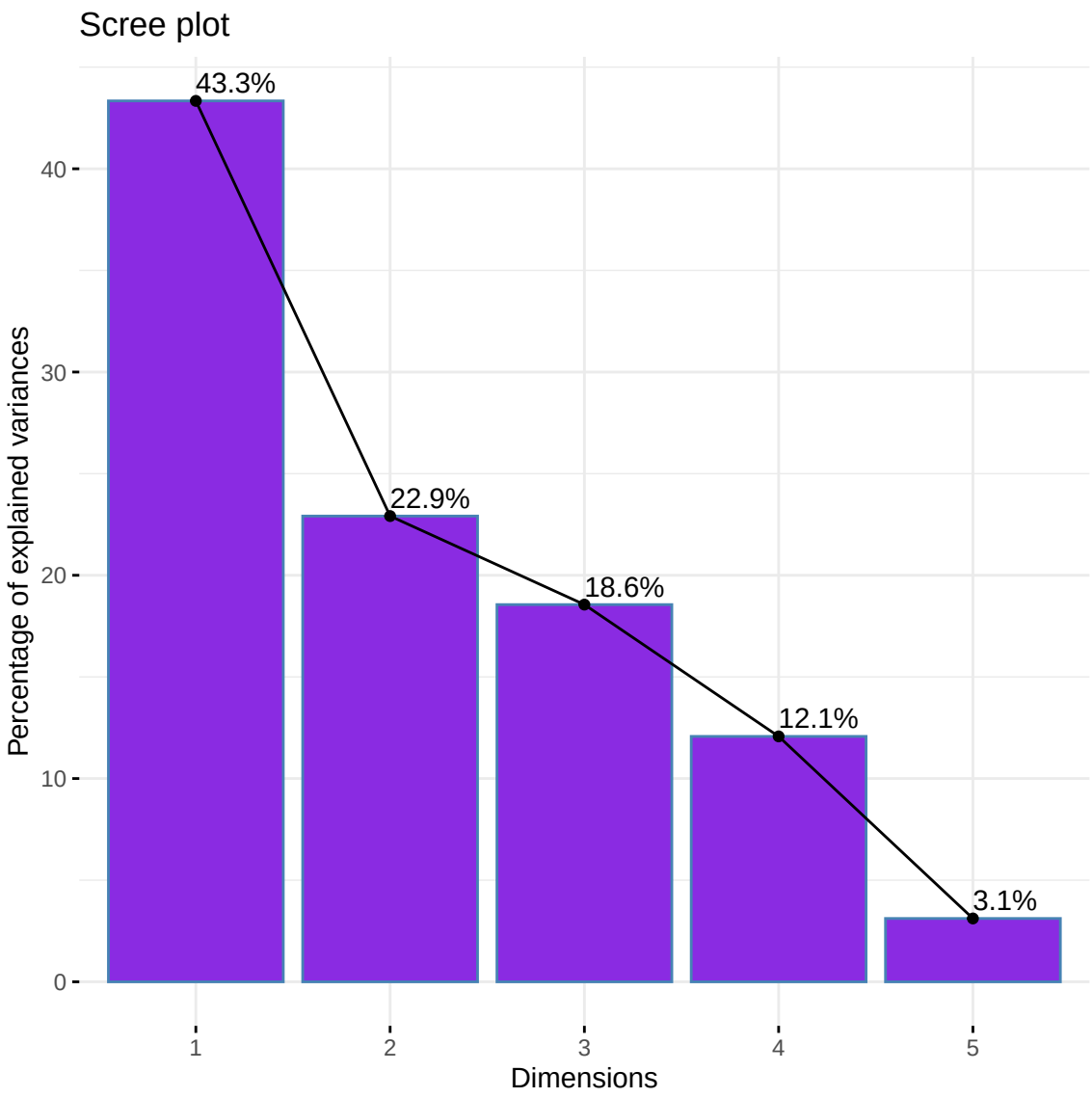

Figure 5.22: Scree Plot for our Principal Component Analysis.

Figure 5.22 is the Scree plot. The Scree plot summarizes the amount of variation explained by each principal component. The first two components alone explain more than 50% of the variance in the dataset.

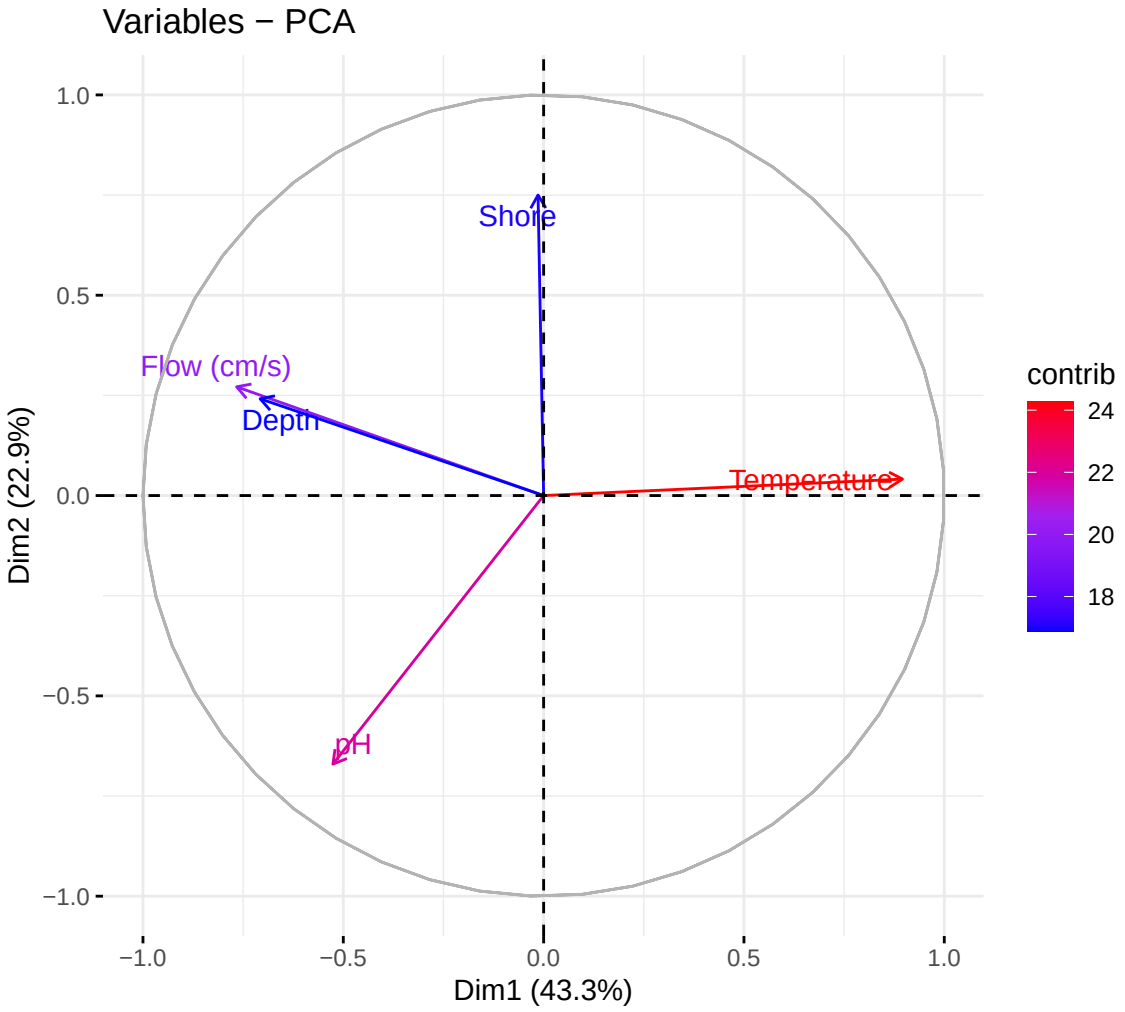

Figure 5.23: Contributions of predictors towards explaining variance. Note temperature and pH have particularly strong influence.

Figure 5.23 is a visualization of the contribution that each predictor has towards the principal components. Using the function prcomp and plotting gives us a way to visualize the dataset as projections on the principal components. From our visualization plot we can see that Temperature and pH have the greatest contributions to the first two dimensions.

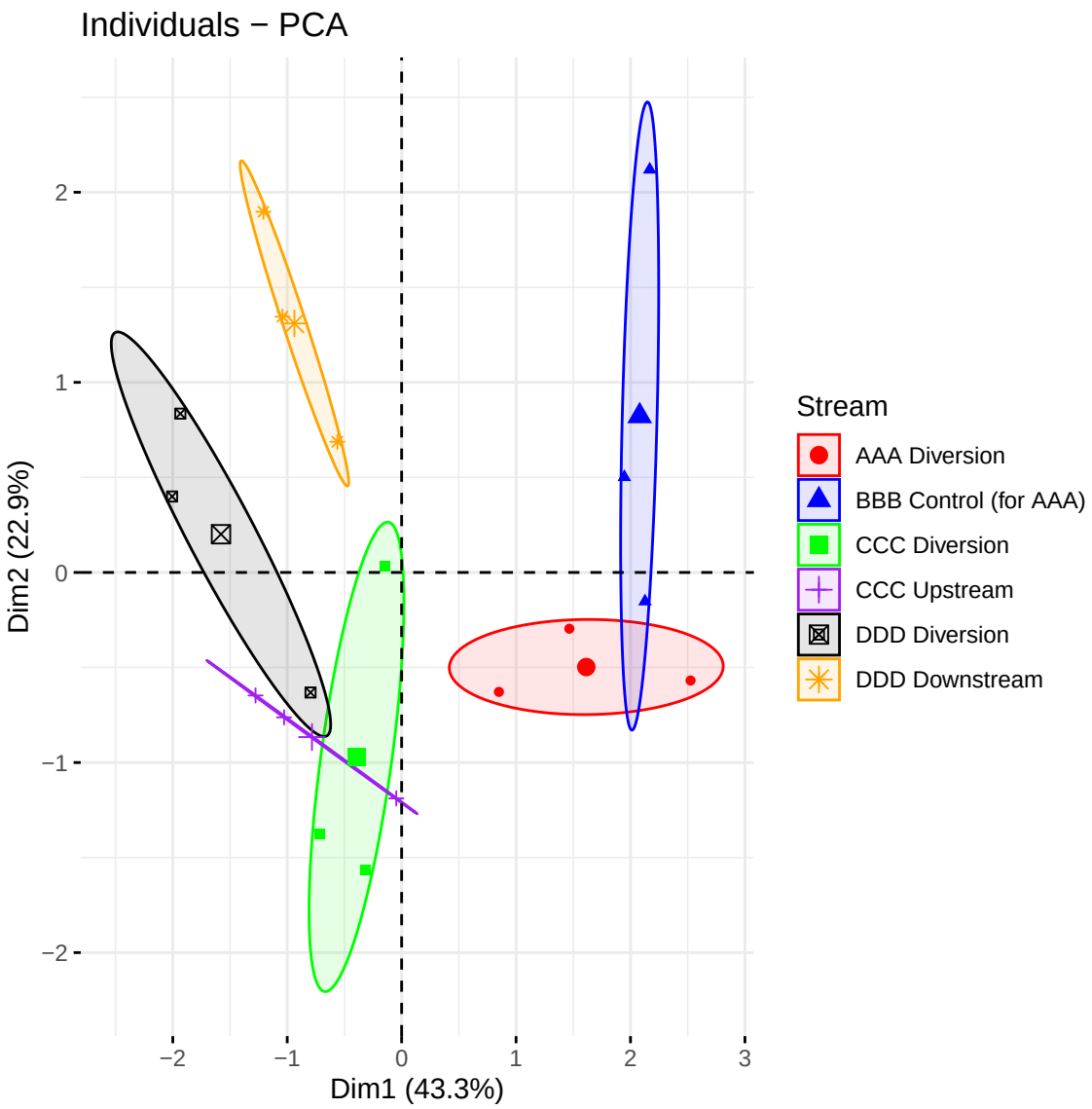

Figure 5.24: PCA on streams and reach.

Figure 5.24 is the result of applying PCA on the stream data, we can fit ellipsoids to each stream. Our plot by stream and reach helps confirm that this makes sense. As temperature has a large impact on the principal components, we see streams AAA and BBB separated from the other streams, and we can see they also have higher temperature compared to the other streams. pH also has a large contribution, with constant pH in stream CCC. This can be seen from the constant line that makes up CCC contribution. Note that Coho only appear in stream BBB which is significantly warmer than streams CCC and DDD.

5.5 Field Conclusions

The field study provided an excellent opportunity to bridge knowledge gained in Chapters 3 and 4 to a real-world study. Four streams in British Columbia were studied for the presence of several species of fish with a focus on Coho salmon. Using transects, researchers were able to obtain biomass measurements of each species in multiple sites throughout the streams, the results of which are summarized in Table 5.3.

As seen in Table 5.3 Coho were only caught using conventional methods in stream BBB. Coho in stream BBB (Figure 5.7) were further confirmed to be present using eDNA methodology (Figure 5.8). The observed biomass of Coho in Stream AAA was zero (Figure 5.3), although Coho eDNA was detected at slight levels (Figure 5.4). The observed biomass of Coho in Stream DDD was also zero (Figure 5.3), but eDNA analysis indicated the presence of Coho in 3 out of 6 sites at stream DDD. These detections may indicate that eDNA analysis is a more sensitive sampling approach and provides the possibility of detecting species which could not manually be confirmed. Overall, in the streams in which Coho was confirmed to be present using transects (stream BBB), we obtained 100 % detection for Coho. Moreover, in this stream we obtained perfect sample replicate detection (8 out of 8).

In order to create statistical models for mean TCT of Coho we first created a simple linear model, model.co (Table 5.3) which already explained much of the variation in the data (evidenced by the R^2=0.6). This model only considered a single predictor, CO.Total.Biomass.g. Plotting of this regression line made the result clear, Coho biomass and mean TCT are highly correlated (Figure 5.21).

After fitting a simple linear model, we investigated a model in which numerous predictors were included, the so called 'full model' (model.coho.full) as seen in Table 5.7. We chose to work with the predictor CO.Biomass.g.m3 as this quantity more closely resembles the density parameter discussed in Chapter 3. Using the full model as an initial model, we performed backward elimination. By considering a larger basis of possible predictors and their associated interactions in the full model, we were able to remove several predictors until our algorithm terminated on a final model. This model, model.co.step, summarized in Table 5.8, ended up being a model that contained CO.Biomass.g.m3, as well as the interaction between Transect.Flow.cms and CO.biomass.g.m3 and the pH.

Model averaging further validated our stepwise model as the highest weighted model (with a weight of $w = 0.35$) was the model containing those exact predictors. Table 5.9 lists the associated term codes and components for the model average object. Best subset modelling provided further evidence in support of theses predictors, this, as the best performing model with three parameters as seen in Table 5.13 is the model containing CO.Biomass.g.m3, pH and the interaction between Transect.Flow.cms and CO.Biomass.g.m3.

For All Fish, as seen in Table 5.1, our assumptions and testing failed to produce any meaningful results. Mean TCT and fish biomass showed no significant relationship.

In the end, multiple model selection techniques converged on the same few features for Coho. CO.Biomass.g.m3, pH, and the interaction between CO.Biomass.g.m3 and Transect.Flow.cms. It appears that environmental covariates such as pH may indeed impact eDNA concentrations. Researchers wishing to obtain accurate and reliable TCT measurements, may wish to consider the impact of environmental covariates on their sampling routines.

Chapter 6

Conclusions and Future Work

6.1 Overview

This research and analysis brought together the disciplines of statistics, biochemistry, microbiology and genetics to provide further validation of several key concepts regarding environmental DNA. Involvement in this study also provided an excellent opportunity to study a real-world topic, with an enormous number of practical and theoretical implications in the field of environmental and biological assessment.

The first experiment we analyzed was the 'density' experiment. Varying numbers of Coho salmon were placed in large tanks and water samples were collected after allowing some time for the fish to swim in the tanks. Water samples were stored and transferred to the lab where they were tested for DNA integrity. Samples that passed integrity tests were then analyzed using Real-time polymerase chain reactions. This process used a thermal cycler to monitor the amplification of Coho specific target DNA. Detection of target DNA in the thermal cycler results in the release of fluorescence, visible to the researcher. When target molecule concentrations are very low, we expect that we will need several 'cycles' in order to detect the DNA (and hence several cycles needed to produce fluorescence). On the other hand, when target DNA is highly concentrated, we expect many less cycles to detect DNA. The number of cycles taken to produce fluorescence is referred to as the Cycle threshold, or simply 'CT'. We choose to work with a monotone transformed version, TCT which was simply 50-CT. TCT should always be positive, as beyond 50 cycles is equivalent to a 'non-detection'.

Using the transformed CT ('TCT') values obtained via qPCR, we established a significant relationship between Coho biomass and detectable eDNA. This experiment confirms what we would expect, as clearly more or larger fish would be expected to shed or release more DNA into their environments. One highlight of this experiment is we were able to create a figure that summarizes many concepts in a clear and concise manner, Figure 3.14. This figure showed a strong relationship between the number of fish in a tank and the mean TCT measurements obtained. The R^2 for the model was nearly 0.750, indicating we explained much of the variation in the dataset with the model. Residual analysis was conducted that provided further validation that the models were performing well.

The second experiment we analyzed was the 'dilution' experiment. In this experiment, three juvenile Coho salmon were again placed in large tanks and left to equilibrate for several days. The fish were then removed, and samples were taken as the tank water drained and was replaced by fresh water (via an inflow pipe). We found that instead of a linear relationship, a more sophisticated model was needed. In particular, we needed a model that accounted for the point in which the water was completely diluted (the so called 'breakpoint'), where qPCR no longer detected any eDNA.

Still water ponds or lakes are only a fraction of all bodies of water. In reality, researchers will be taking measurements from all sorts of water systems, including those with strong currents and fast flows. We confirmed that increased flow rates result in lower amounts of detectable eDNA in the experiment. We were able to fit several niche models to explain the impact of flow on our TCT measurements. In the end we had success fitting a bent cable model, Figure 4.12 that captured much of the variation in the dataset.

The last set of data analyzed was actual field data obtained from several streams in British Columbia. Here we attempted to use what was found in the first two experiments (the impact of biomass and flow) to study these streams.

In addition, the field data allowed us to implement covariate analysis, whereby we were able to 'tease' out the most important factors impacting eDNA collection and analysis. As expected, Coho density and flow rate were strongly more significant than other covariates such as eDNA sample depth. The analysis of field data provided an

opportunity to connect the result of controlled experiments into the actual environment. Several statistical models were used and compared to validate field results. We used stepwise selection to validate the important covariates, and we further confirmed the importance using best subset models and model averaging.

6.1.1 Future of eDNA technology

The possible applications of an in depth understanding of eDNA are seemingly endless. One could imagine a day where invasive techniques for biodiversity monitoring are discarded all together, in favor of non-invasive and less expensive eDNA methodologies. Much work is still required, especially regarding transferring knowledge gained in research labs to field data. Indeed, one of the most needed areas of continued research is to compare the performance of differing eDNA analysis methods. This includes a wide variety of collection (different types and sizes of filters) and storage methods (such as freezing samples or analyzing immediately). Additional studies on a variety of species with differing characteristics and environmental factors will help to create a standard procedure of analysis. Even within a single species, weather and environmental conditions can greatly impact behavior and metabolic processes. Studies have already confirmed that eDNA detection probabilities can be greatly altered by 'seasonal' activity of organisms (De Souza et al., 2016).

Currently, there is a need for more independent comparative studies, and eventually a method that can account for multiple species. The current lack of independent and quantitative comparisons of techniques makes it difficult to provide advice on which methods are best for a given species. While our results were promising for Coho, they were not as accurate for the other fish species tested (Beja-Pereira et al., 2009b). Moreover, it is not clear how well the methods discussed in this paper would work on extremely low concentration streams, or areas with extreme weather and fast-moving currents. Further research on these extreme conditions is still needed.

The study of eDNA is also now being utilized in the field of forensics. Indeed, there is currently a large growing field of research is focused on human identification without using human DNA, but instead exploiting the microbial signature left behind by individuals in their environment and possessions (Allwood et al., 2019).

Another goal should be to involve the community members. eDNA technologies and the associated analysis should not be limited to those in research labs. Indeed, eDNA technology has the potential to improve lives and improve communities. In order for eDNA technology to be utilized by the public or average citizens, several researchers have pointed out a need for a more user-friendly method of eDNA exploration and analysis (Ruppert et al., 2019).

Regardless of the limitations, eDNA analysis is quickly becoming a valuable technique in the study of aquatic species in particular. Work is needed on creating a standardized method of collection and analysis (Rees et al., 2015). This could include future work on creating an 'all in one' rapid on-site analysis, such as the eDNA backpack.

In summary, eDNA monitoring provided an environmentally friendly alternative to biodiversity monitoring. Work is still needed to improve the speed and accuracy in which environmental samples can be analyzed. Another future goal is to put eDNA technology into the hands of community members, even those without technical training. The rapid development of DNA sequencing technology and assessment will likely accelerate the rate at which studies into eDNA are conducted. The speed and cost at which DNA is sequenced improves yearly and this will no doubt positively impact studies into environmental DNA. Moreover, as computers improve and can store more data, we can imagine a future whereby researchers can access and store important DNA sequences with a simple keystroke (Yoccoz, 2012).

Appendix A

Appendix

A.1 Stream Data plots

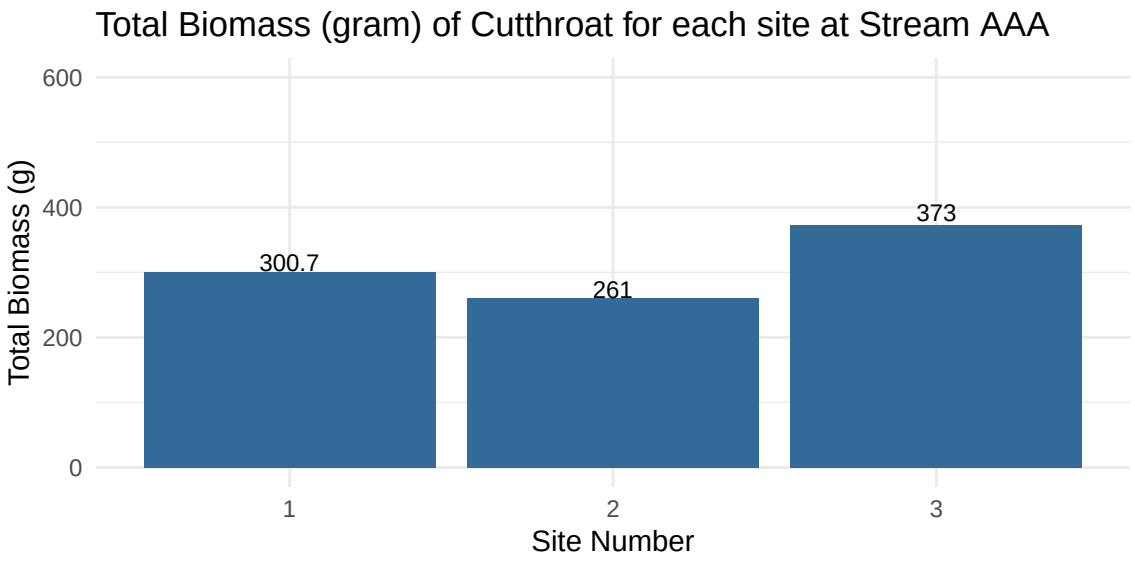

Figure A.1: Total Biomass of Cutthrout Trout at Stream AAA.

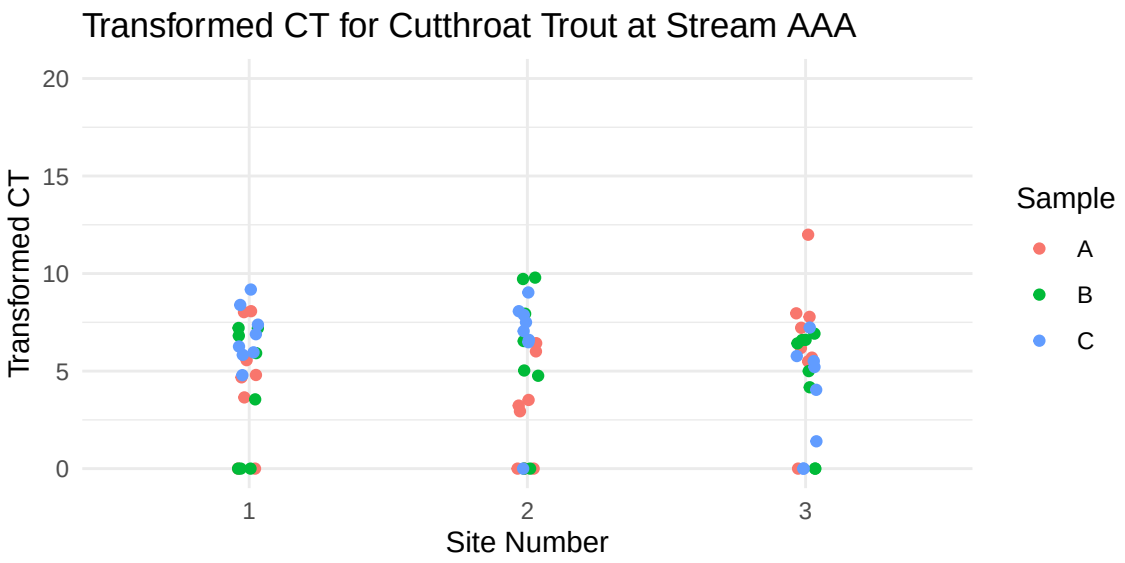

Figure A.2: Transformed CT values taken over technical replicates for Cutthroat Trout.

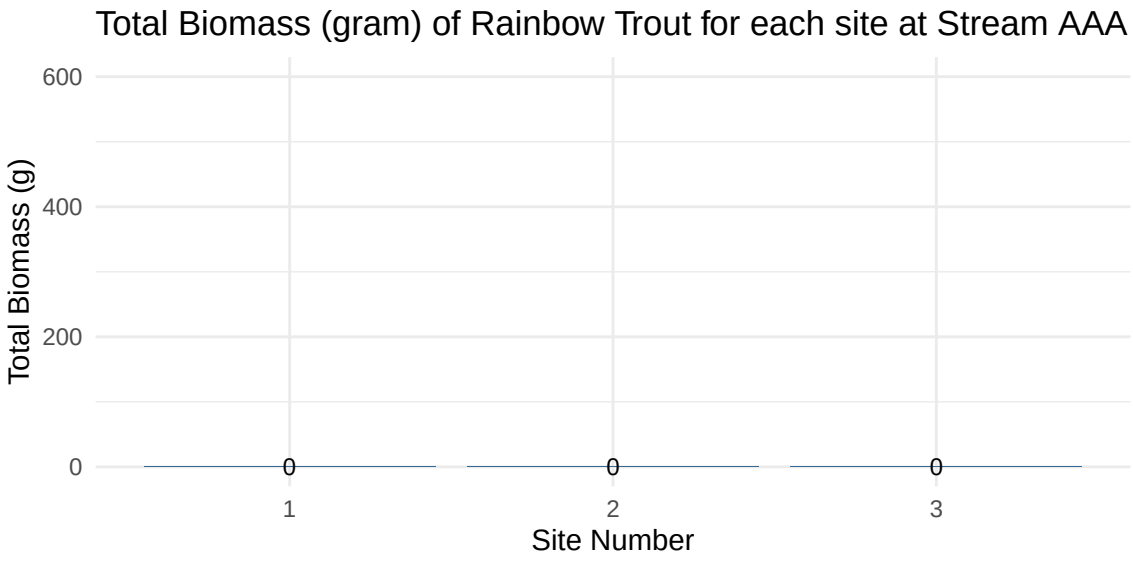

Figure A.3: Total Biomass for Rainbow Trout at Stream AAA.

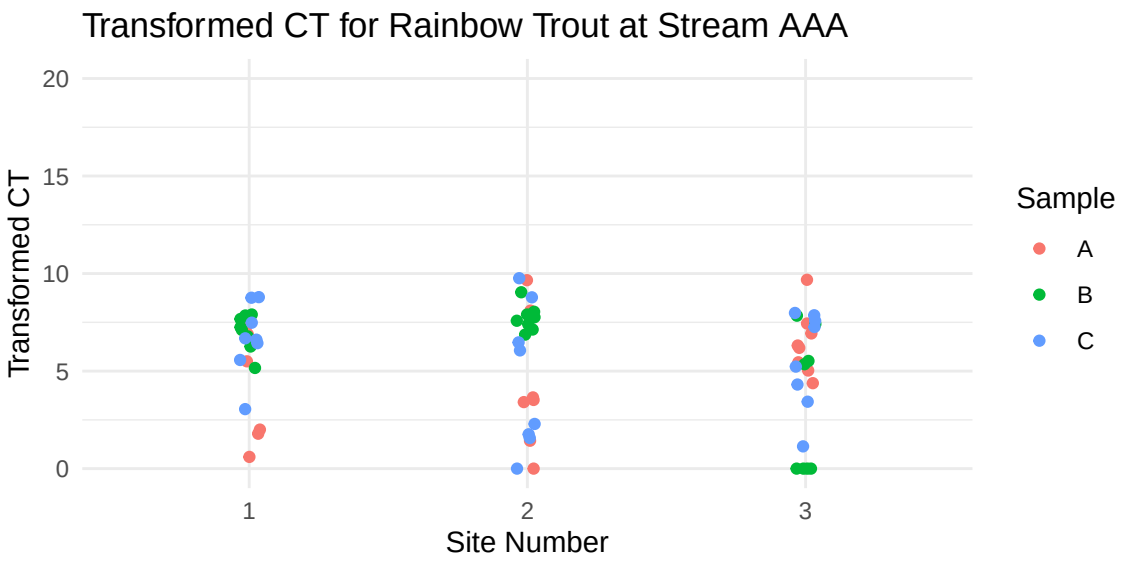

Figure A.4: Transformed CT values taken over technical replicates for Rainbow Trout.

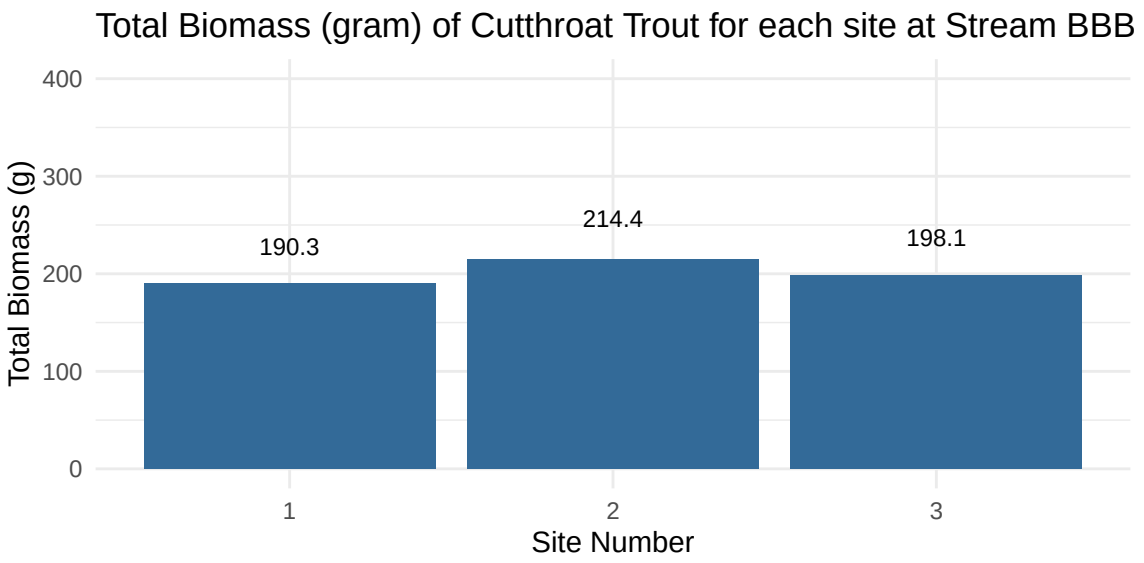

Figure A.5: Total biomass for Cutthroat Trout at each site for Stream BBB.

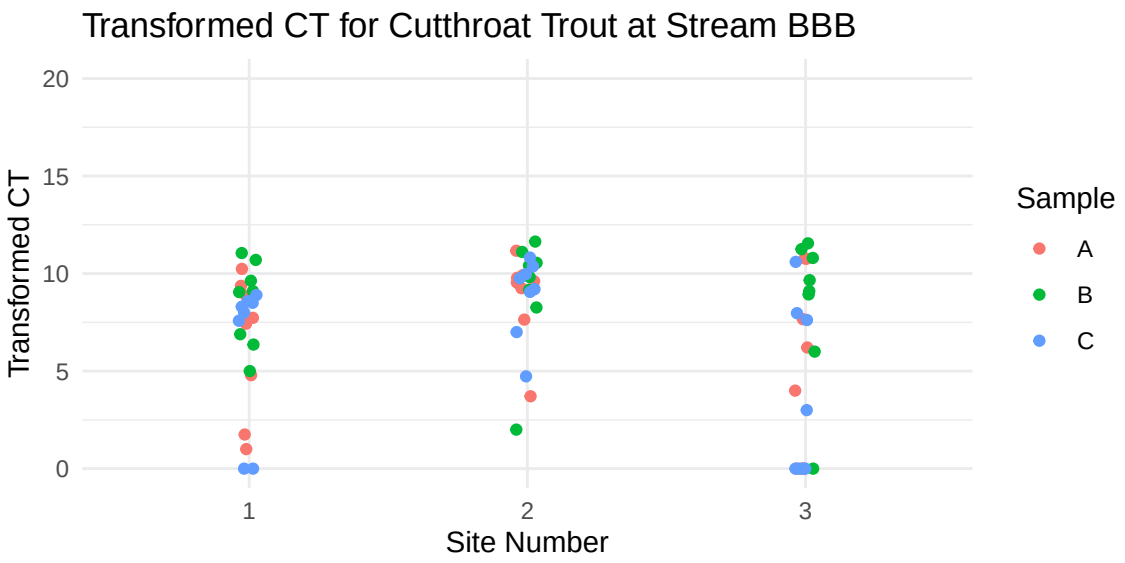

Figure A.6: Transformed CT values taken over technical replicates for Cutthroat Trout.

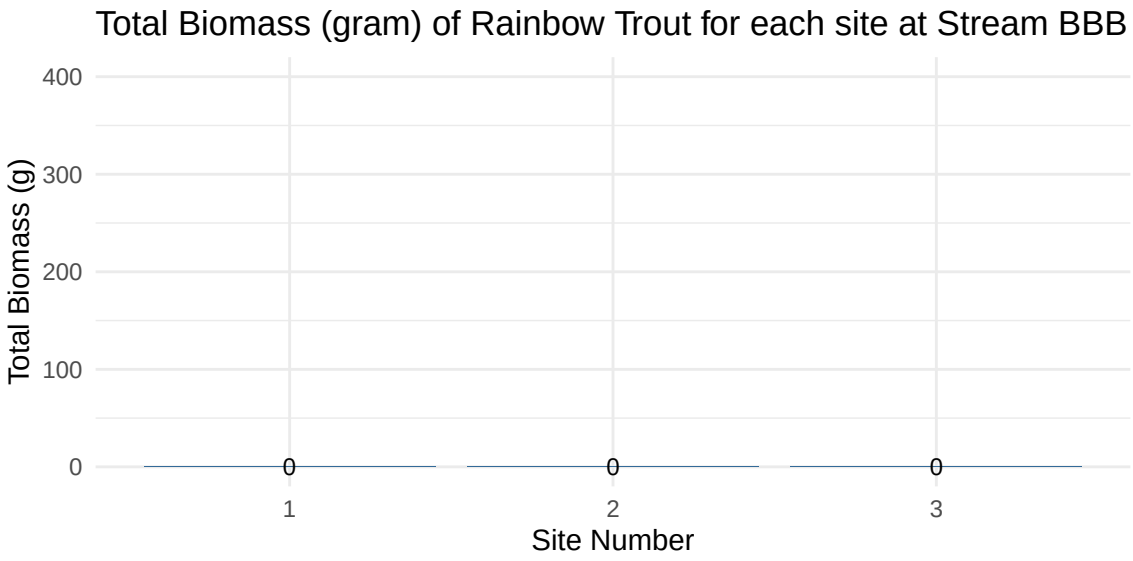

Figure A.7: Total biomass for Rainbow Trout at each site for Stream BBB.

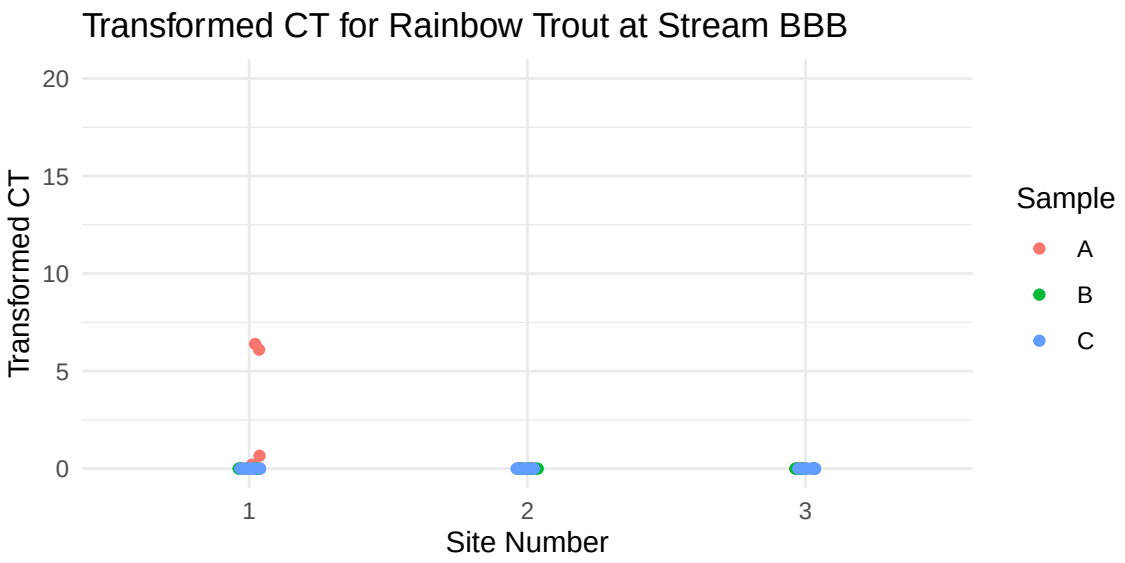

Figure A.8: Transformed CT values taken over technical replicates for Rainbow Trout.

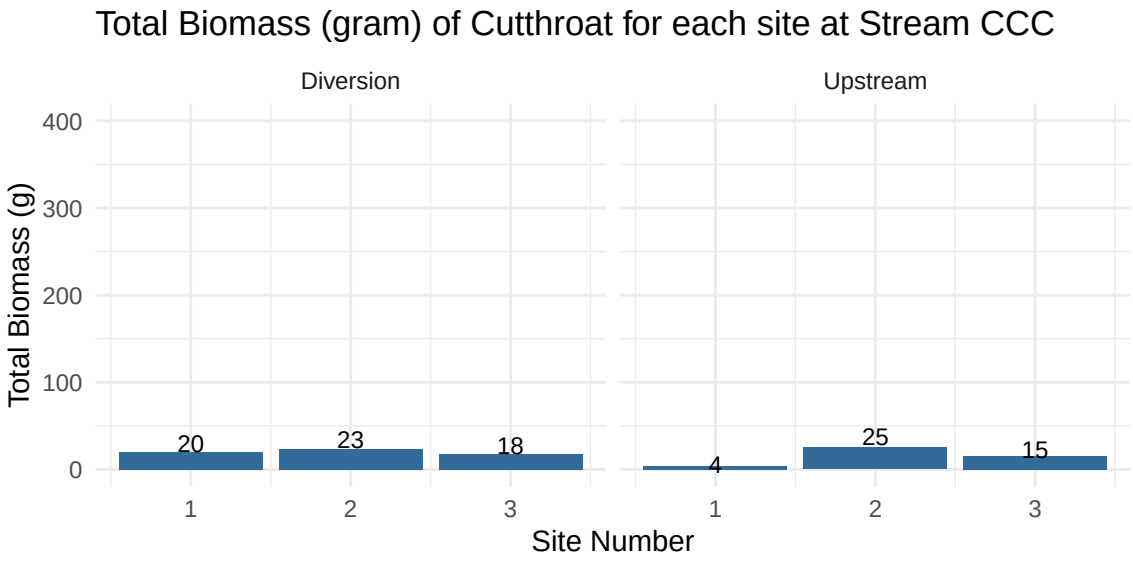

Figure A.9: Total biomass for Cutthroat Trout at each site for Stream CCC.

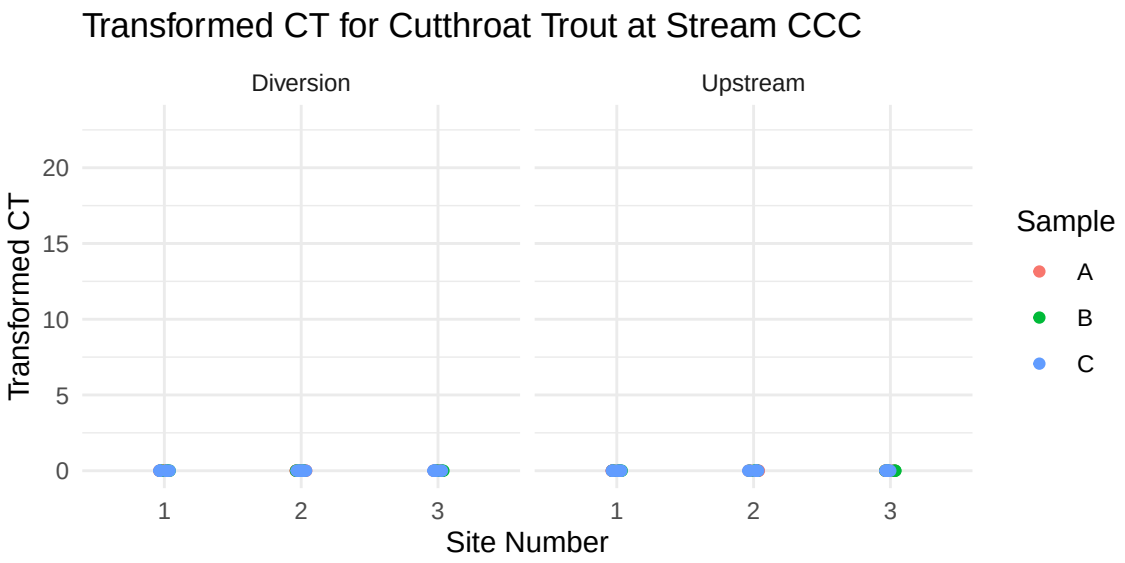

Figure A.10: Transformed CT values taken over technical replicates for Cutthroat Trout.

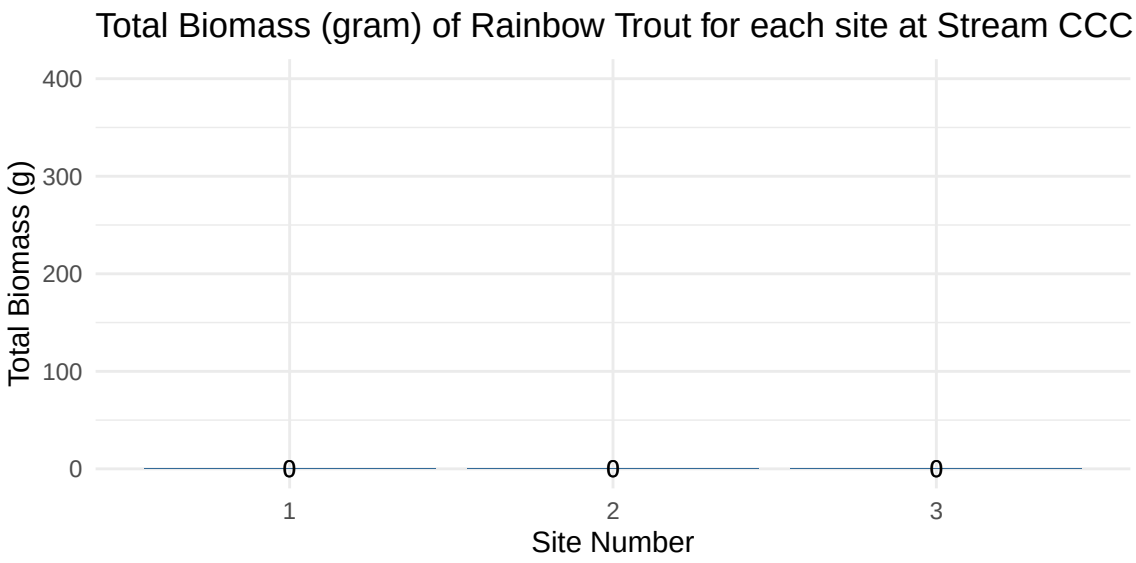

Figure A.11: Total biomass for Rainbow Trout at each site for Stream CCC.

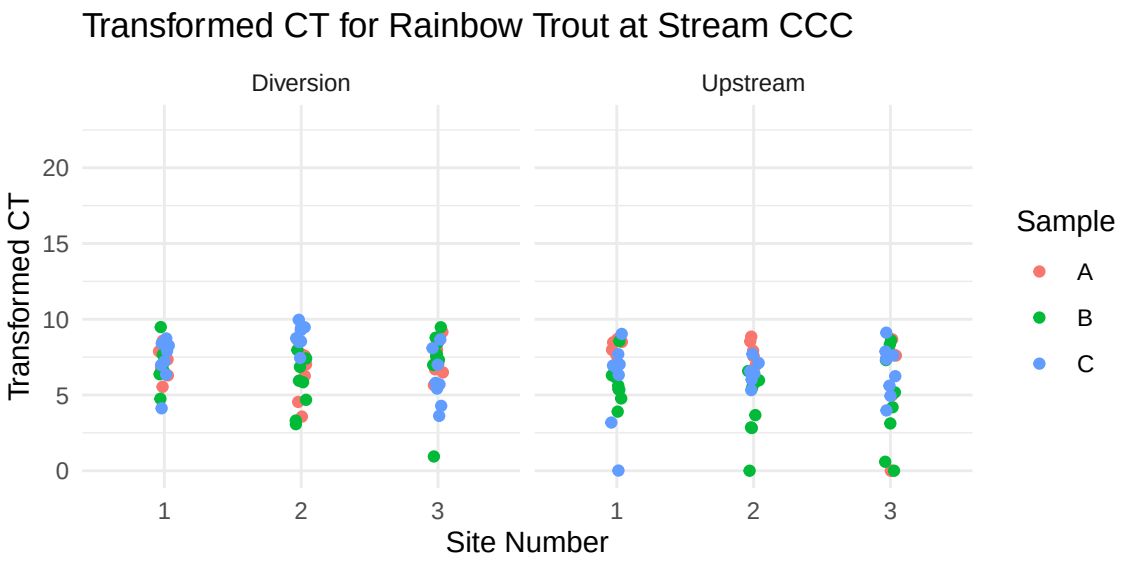

Figure A.12: Transformed CT values taken over technical replicates for Rainbow Trout.

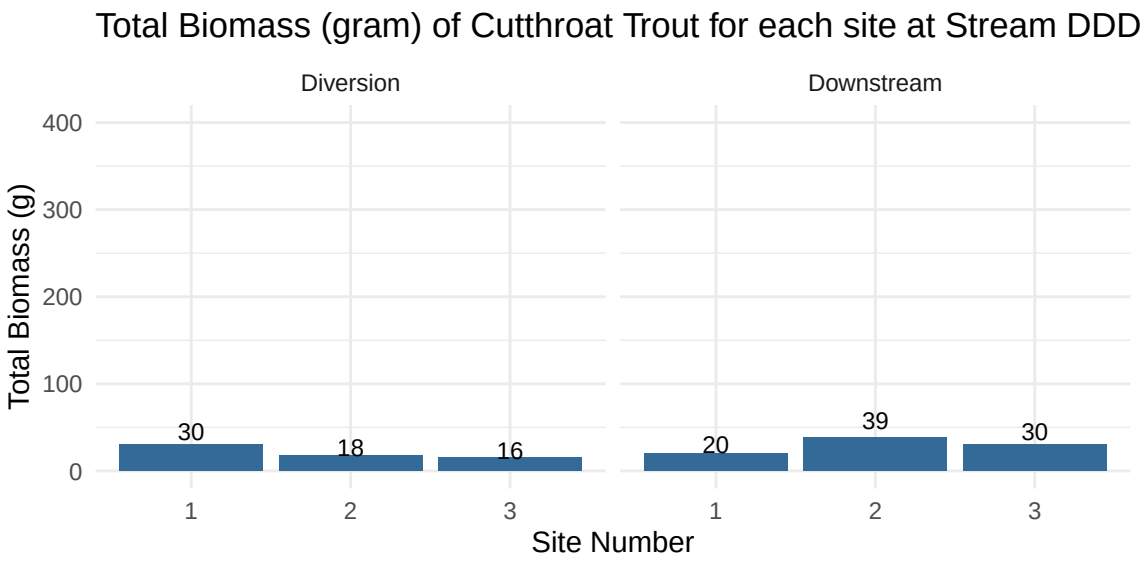

Figure A.13: Total biomass for Cutthroat Trout at each site for Stream DDD.

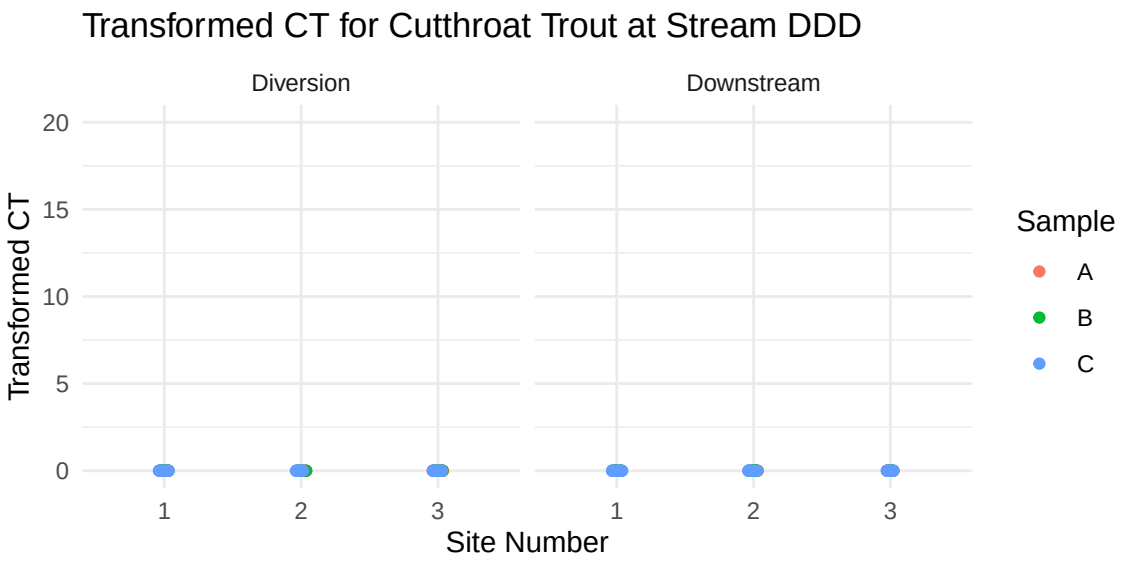

Figure A.14: Transformed CT values taken over technical replicates for Cutthroat Trout.

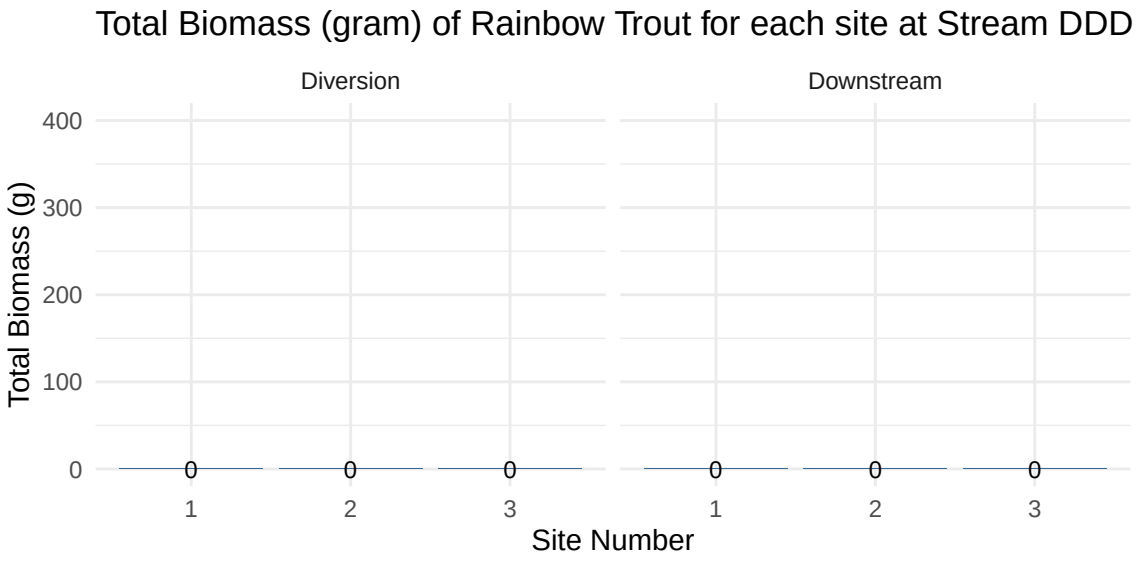

Figure A.15: Total biomass for Rainbow Trout at each site for Stream DDD.

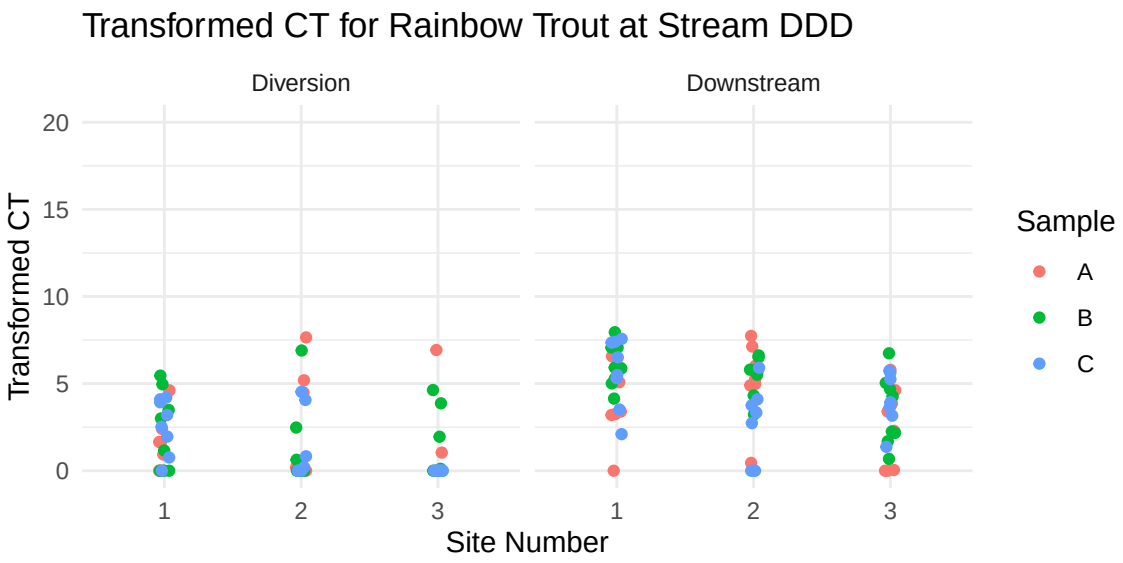

Figure A.16: Transformed CT values taken over technical replicates for Rainbow Trout.

A.2 Pairs Plots

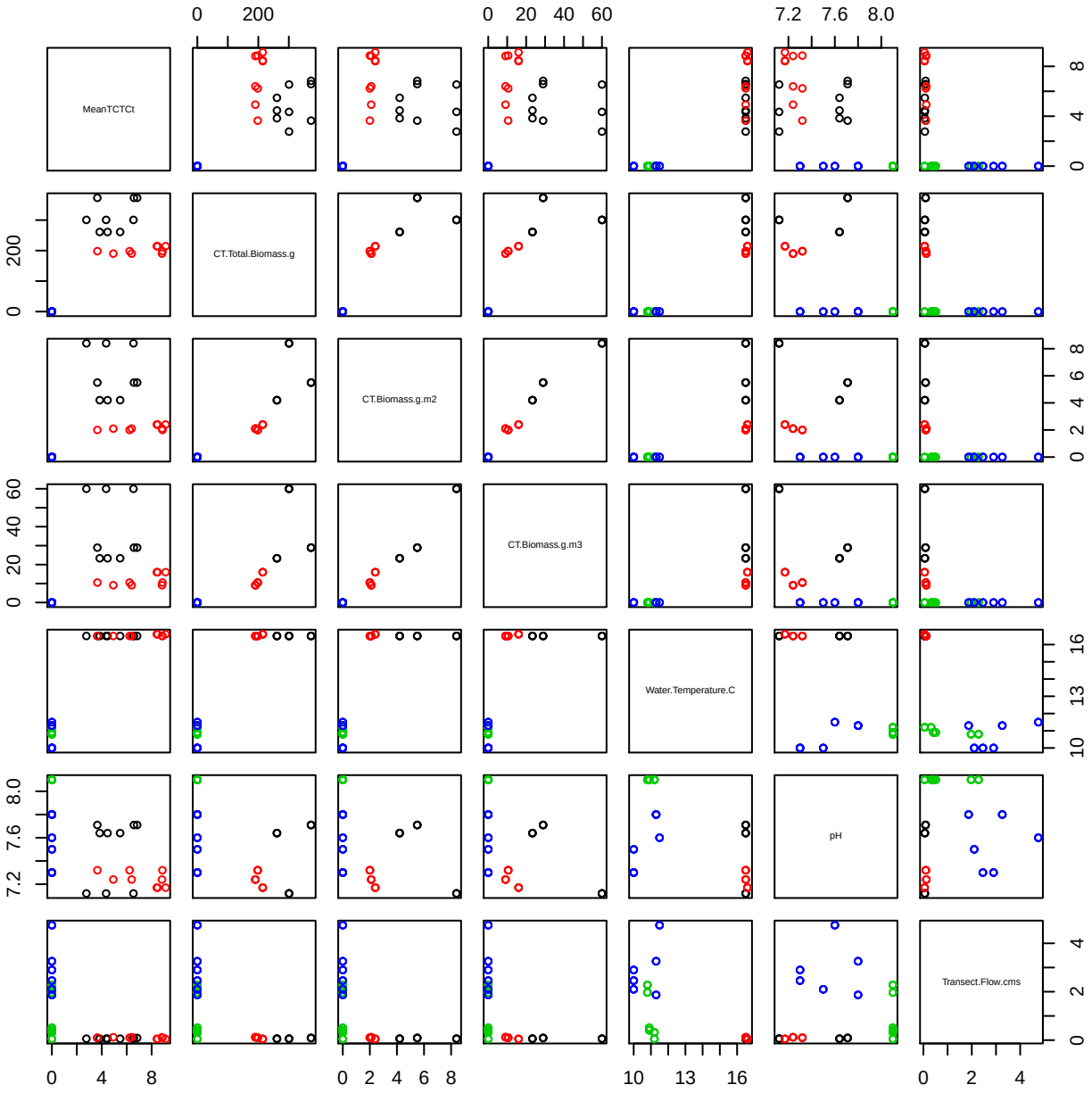

Figure A.17: Pairs plots for Cutthroat Trout. Black corresponds to Stream AAA, red corresponds to Stream BBB, green corresponds to Stream CCC and blue corresponds to Stream DDD.

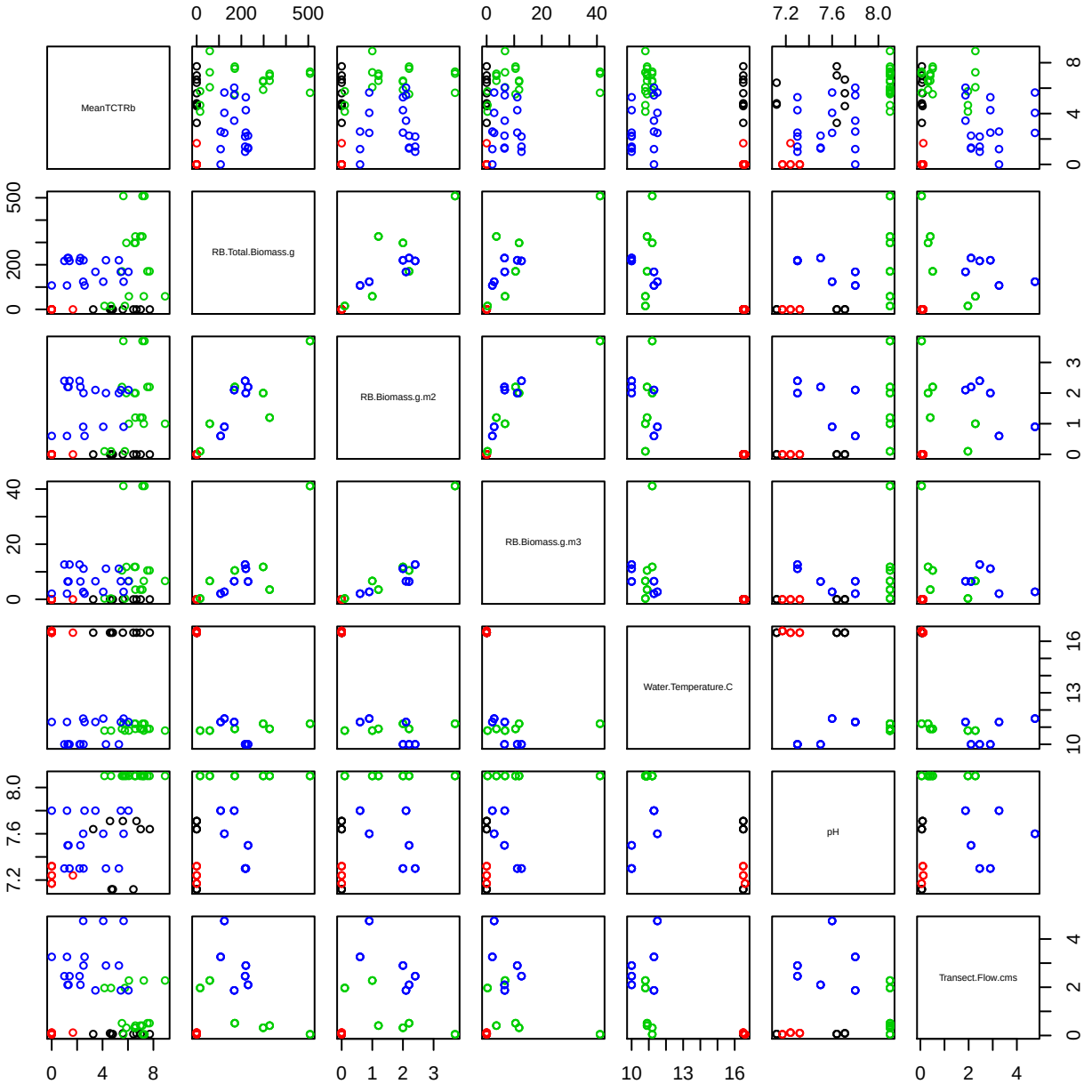

Figure A.18: Pairs plots for Rainbow Trout. Black corresponds to Stream AAA, red corresponds to Stream BBB, green corresponds to Stream CCC and blue corresponds to Stream DDD.

A.3 Field Models

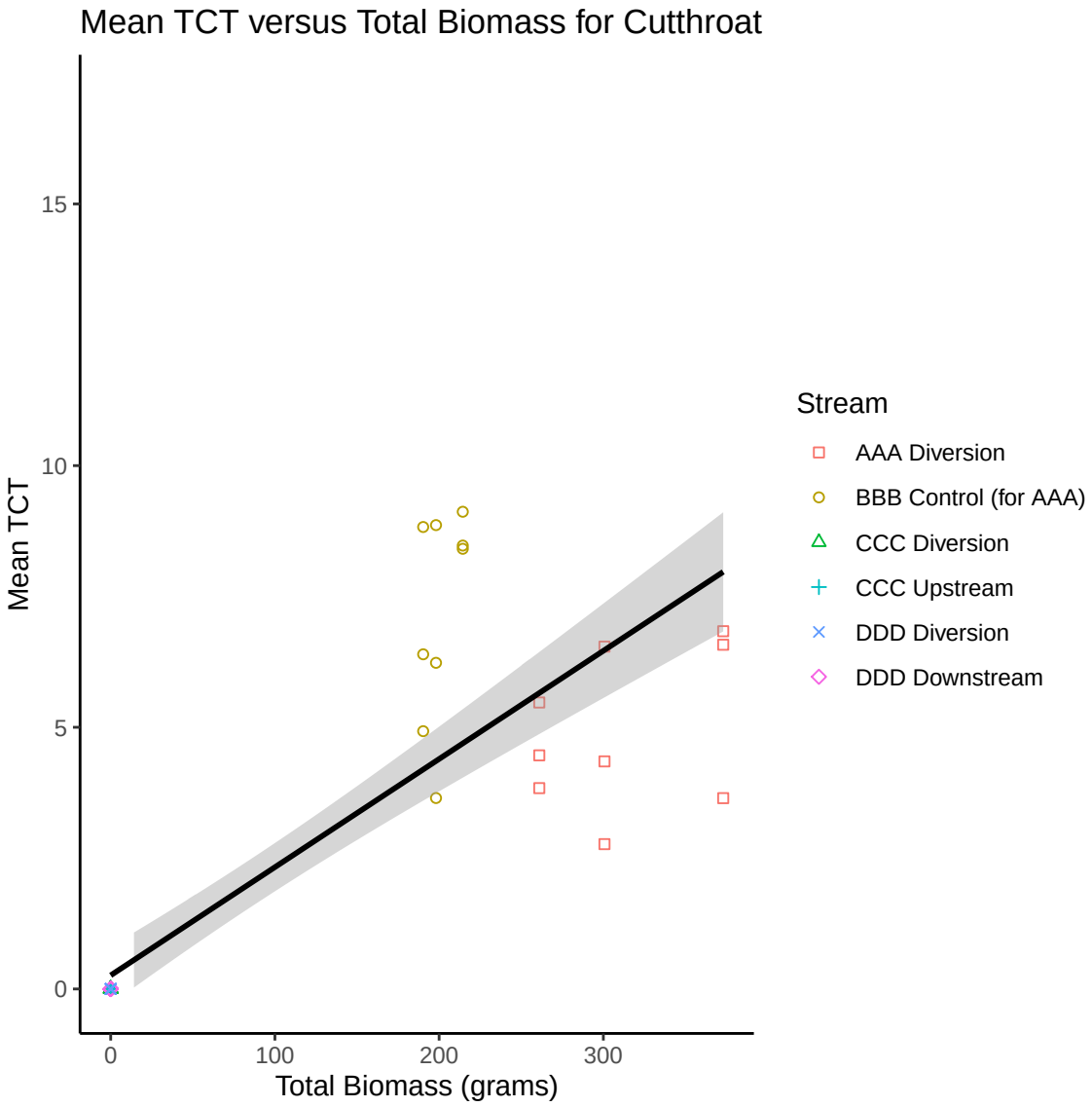

Figure A.19: Mean TCT versus Total Biomass for Cutthroat Trout. Included is the simple linear regression model and the 95% Confidence limits for the regression line.

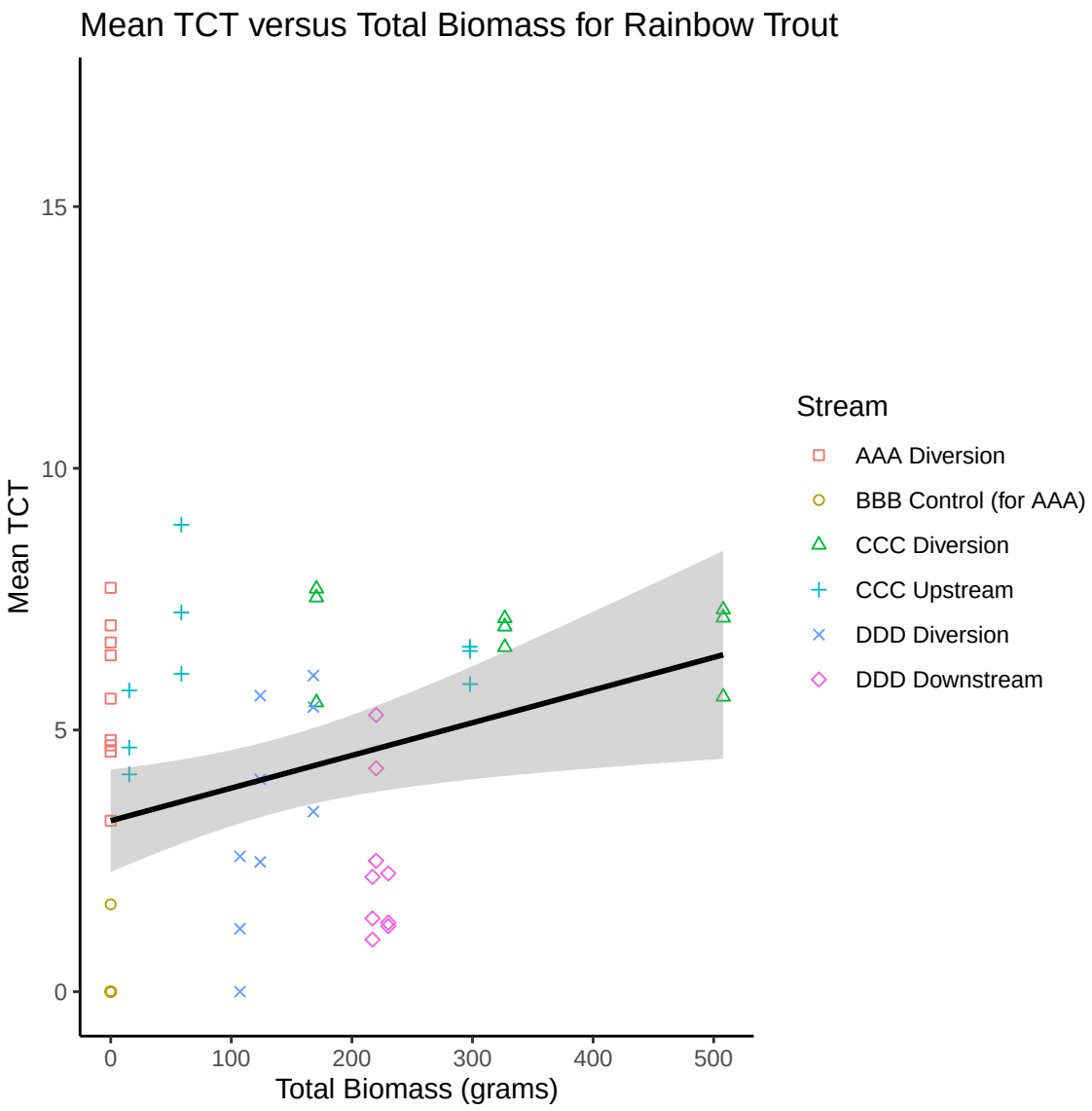

Figure A.20: Mean TCT versus Total Biomass for Rainbow Trout. Included is the simple linear regression model and the 95% Confidence limits for the regression line.

Table A.6 is the simple linear model for Cutthroat Trout (model.ct), our model achieves an R^2 of 0.716.

```
Call:
lm(formula = MeanTCTCt ~ CT.Total.Biomass.g,
data = field.collapse)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3239 -0.2602 -0.2602 -0.2602  4.6339

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.260163   0.275491   0.944    0.349
CT.Total.Biomass.g 0.020669   0.001805  11.448 7.94e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.678 on 52 degrees of freedom
Multiple R-squared:  0.7159,Adjusted R-squared:  0.7105
F-statistic: 131.1 on 1 and 52 DF,  p-value: 7.938e-16
```

Table A.1: Model: model.ct

```
Call:
lm(formula = MeanTCTRb ~ RB.Total.Biomass.g,
 data = field.collapse)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9350 -2.6604  0.7287  1.6554  5.2882

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.265738   0.487112   6.704 1.47e-08 ***
RB.Total.Biomass.g 0.006243   0.002488   2.509   0.0153 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.579 on 52 degrees of freedom
Multiple R-squared:  0.108,Adjusted R-squared:  0.09085
F-statistic: 6.296 on 1 and 52 DF,  p-value: 0.01525
```

Table A.2: Model: model.rb

Table A.2 summarizes the model that only considers Biomass for Rainbow Trout and it does not generalize well. The R^2 for this model is only 0.108. This implies that only considering the biomass of Rainbow Trout does not explain most of the variation in the data.

A.4 Stepwise Elimination and Model Averaging

```
Call:
lm(formula = MeanTCTEf ~ Transect.Flow.cms + pH + Fish.Biomass.g.m3 +
    Transect.Flow.cms:Fish.Biomass.g.m3,
data = field.removeef)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6188 -0.5725  0.1032  0.6946  2.5055

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        9.7641663  4.0772817   2.395   0.0206 *
Transect.Flow.cms                 -0.0841080  0.1810530  -0.465   0.6444
pH                                 1.0664736  0.5148347   2.071   0.0437 *
Fish.Biomass.g.m3                 -0.0001366  0.0096998  -0.014   0.9888
Transect.Flow.cms:Fish.Biomass.g.m3 -0.0199482  0.0240409  -0.830   0.4108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.195 on 48 degrees of freedom
Multiple R-squared:  0.1708,Adjusted R-squared:  0.1017
F-statistic: 2.472 on 4 and 48 DF,  p-value: 0.05694
```

Table A.3: Backward elimination for all Fish.

Table A.3 summarizes backward stepwise elimination for all Fish, backward selection results in an R^2 of 0.171.

```
Call:
lm(formula = MeanTCTCt ~ Water.Temperature.C + pH + CT.Biomass.g.m3 +
    Transect.Flow.cms,
data = field.collapse)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2120 -0.3742 -0.0912  0.5151  2.1018

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.37426    5.02637   1.268 0.210732
Water.Temperature.C 1.01396    0.10145   9.994 2.05e-13 ***
pH                 -2.14087    0.54340  -3.940 0.000258 ***
CT.Biomass.g.m3    -0.05147    0.01450  -3.550 0.000863 ***
Transect.Flow.cms  -0.30831    0.14877  -2.072 0.043510 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.059 on 49 degrees of freedom
Multiple R-squared:  0.8934,Adjusted R-squared:  0.8847
F-statistic: 102.7 on 4 and 49 DF,  p-value: < 2.2e-16
```

Table A.4: Backward elimination for Cutthroat Trout.

For Cutthroat, backwards stepwise elimination results in a model with an R^2 of 0.893.

```
Call:
lm(formula = MeanTCTRb ~ pH + Transect.Flow.cms + Water.Temperature.C +
    RB.Biomass.g.m3 + eDNA.Distance.from.Shore.m + Transect.Flow.cms:RB.Biomass.g.
    data = field.collapse)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6985 -1.1142 -0.0279  1.0255  4.3280

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -58.93243   13.30507  -4.429 5.62e-05 ***
pH                                7.33900    1.34999   5.436 1.90e-06 ***
Transect.Flow.cms                -0.43015    0.29662  -1.450  0.15365
Water.Temperature.C               0.53657    0.24086   2.228  0.03072 *
RB.Biomass.g.m3                   0.04778    0.03705   1.290  0.20345
eDNA.Distance.from.Shore.m       -0.49597    0.20726  -2.393  0.02076 *
Transect.Flow.cms:RB.Biomass.g.m3  0.16648    0.05260   3.165  0.00272 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.805 on 47 degrees of freedom
Multiple R-squared:  0.6052,Adjusted R-squared:  0.5548
F-statistic: 12.01 on 6 and 47 DF,  p-value: 3.947e-08
```

Table A.5: Backward elimination for Rainbow Trout.

For Rainbow Trout, backward elimination results in a model with an R^2 of 0.605.

```
Call:
model.avg(object = ma.fish, subset = delta < 5)

Component model call:
lm(formula = MeanTCTEf ~ <37 unique rhs>,
data = field.collapse)

Term Codes:
eDNA.Distance.from.Shore.m: 1
eDNA.Total.Water.Depth.m: 2
Fish.Biomass.g.m3: 3
pH: 4
Transect.Flow.cms: 5
Water.Temperature.C: 6
Fish.Biomass.g.m3:Transect.Flow.cms: 7

Component models:
       df logLik AICc delta weight
5       3   -124  255  0.00   0.11
25      4   -123  255  0.09   0.10
45      4   -124  256  1.27   0.06
56      4   -124  257  1.50   0.05
(Null)  2   -126  257  1.52   0.05
15      4   -124  257  1.66   0.05
245     5   -123  257  1.93   0.04
125     5   -123  257  2.04   0.04
256     5   -123  257  2.08   0.04
35      4   -124  257  2.13   0.04
235     5   -123  258  2.45   0.03
145     5   -123  258  2.66   0.03
4       3   -126  258  2.85   0.03
1       3   -126  258  2.99   0.02
3       3   -126  258  3.36   0.02
156     5   -124  259  3.47   0.02
6       3   -126  259  3.53   0.02
456     5   -124  259  3.60   0.02
2       3   -126  259  3.63   0.02
345     5   -124  259  3.70   0.02
1245    6   -122  259  3.71   0.02
46      4   -125  259  3.79   0.02
356     5   -124  259  3.92   0.02
135     5   -124  259  4.02   0.01
14      4   -125  259  4.05   0.01
34      4   -125  259  4.19   0.01
1256    6   -123  259  4.25   0.01
146     5   -124  259  4.25   0.01
357     5   -124  259  4.32   0.01
2456    6   -123  259  4.41   0.01
2345    6   -123  260  4.46   0.01
1235    6   -123  260  4.56   0.01
2356    6   -123  260  4.61   0.01
13      4   -125  260  4.68   0.01
134     5   -124  260  4.78   0.01
2357    6   -123  260  4.97   0.01
16      4   -126  260  4.98   0.01


Model-averaged coefficients:
(full average)
```

```
Call:
model.avg(object = ma.cutthroat, subset = delta < 4)

Component model call:
lm(formula = MeanTCTCt ~ <5 unique rhs>,
data = field.collapse)


Term codes:
CT.Biomass.g.m3: 1
eDNA.Distance.from.Shore.m: 2
eDNA.Total.Water.Depth.m:  3
pH:  4
Transect.Flow.cms: 5
Water.Temperature.C: 6
CT.Biomass.g.m3:Transect.Flow.cms: 7


Component models:
      df logLik   AICc delta weight
1456   6 -77.07 167.93  0.00   0.43
12456  7 -76.61 169.66  1.73   0.18
146    5 -79.34 169.93  2.00   0.16
14567  7 -76.99 170.41  2.48   0.12
13456  7 -77.07 170.58  2.65   0.11


Model-averaged coefficients:
(full average)
                                Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept)                     5.904451   5.766750    5.883321   1.004 0.315576
CT.Biomass.g.m3                -0.050704   0.021740    0.022245   2.279 0.022648 *
pH                             -2.110519   0.608632    0.621673   3.395 0.000687 ***
Transect.Flow.cms              -0.271227   0.184933    0.187705   1.445 0.148468
Water.Temperature.C             1.033545   0.110504    0.112878   9.156  < 2e-16 ***
eDNA.Distance.from.Shore.m     -0.021905   0.073653    0.074794   0.293 0.769617
CT.Biomass.g.m3:Transect.Flow.cms -0.033700   0.265700    0.271792   0.124 0.901320
eDNA.Total.Water.Depth.m        0.005467   0.441637    0.453041   0.012 0.990373

(conditional average)
                               Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept)                     5.90445    5.76675     5.88332   1.004 0.315576
CT.Biomass.g.m3                -0.05070    0.02174     0.02225   2.279 0.022648 *
pH                             -2.11052    0.60863     0.62167   3.395 0.000687 ***
Transect.Flow.cms              -0.32170    0.15597     0.15985   2.013 0.044167 *
Water.Temperature.C             1.03355    0.11050     0.11288   9.156  < 2e-16 ***
eDNA.Distance.from.Shore.m     -0.12184    0.13415     0.13762   0.885 0.375974
CT.Biomass.g.m3:Transect.Flow.cms -0.27336    0.71212     0.73054   0.374 0.708265
eDNA.Total.Water.Depth.m        0.04813    1.30966     1.34352   0.036 0.971422
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table A.7: Model Averaging for Cutthroat Trout.

```
Call:
model.avg(object = ma.rainbow, subset = delta < 4)

Component model call:
lm(formula = MeanTCTRb ~ <4 unique rhs>,
data = field.collapse)


Term Codes:
eDNA.Distance.from.Shore.m: 1
eDNA.Total.Water.Depth.m: 2
pH: 3
RB.Biomass.g.m3:   4
Transect.Flow.cms: 5
Water.Temperature.C: 6
RB.Biomass.g.m3:Transect.Flow.cms: 7


Component models:
        df  logLik   AICc delta weight
134567   8 -104.76 228.71  0.00   0.58
1234567  9 -104.55 231.20  2.49   0.17
13457    7 -107.47 231.37  2.66   0.15
34567    7 -107.86 232.15  3.44   0.10


Model-averaged coefficients:
(full average)
                                Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept)                    -55.23056   15.92515    16.19530   3.410 0.000649 ***
eDNA.Distance.from.Shore.m      -0.44811    0.24969     0.25386   1.765 0.077532 .
pH                               7.00291    1.57845     1.60755   4.356 1.32e-05 ***
RB.Biomass.g.m3                  0.04071    0.03914     0.04006   1.016 0.309428
Transect.Flow.cms               -0.47127    0.31567     0.32304   1.459 0.144612
Water.Temperature.C              0.45021    0.29539     0.29993   1.501 0.133344
RB.Biomass.g.m3:Transect.Flow.cms  0.15154    0.05841     0.05965   2.540 0.011070 *
eDNA.Total.Water.Depth.m         0.23628    1.11991     1.14348   0.207 0.836298

(conditional average)
                                Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept)                    -55.23056   15.92515    16.19530   3.410 0.000649 ***
eDNA.Distance.from.Shore.m      -0.49970    0.20915     0.21468   2.328 0.019928 *
pH                               7.00291    1.57845     1.60755   4.356 1.32e-05 ***
RB.Biomass.g.m3                  0.04071    0.03914     0.04006   1.016 0.309428
Transect.Flow.cms               -0.47127    0.31567     0.32304   1.459 0.144612
Water.Temperature.C              0.53156    0.24450     0.25094   2.118 0.034155 *
RB.Biomass.g.m3:Transect.Flow.cms  0.15154    0.05841     0.05965   2.540 0.011070 *
eDNA.Total.Water.Depth.m         1.42005    2.42008     2.48544   0.571 0.567765
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table A.8: Model Averaging for Rainbow Trout.

Bibliography

Allwood, J., Fierer, N., and Dunn, R. (2019). The future of environmental dna in forensic science. *Applied and Environmental Microbiology*, 86.

Arnold, J. B. (2019). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 4.2.0.

Barton, K. (2020). *MuMIn: Multi-Model Inference*. R package version 1.43.17.

based on Fortran code by Alan Miller, T. L. (2020). *leaps: Regression Subset Selection*. R package version 3.1.

Baty, F., Ritz, C., Charles, S., Brutsche, M., Flandrois, J.-P., and Delignette-Muller, M.-L. (2015). *A Toolbox for Nonlinear Regression in R: The Package nlstools*.

Beja-Pereira, A., Oliveira, R., Alves, P., Schwartz, M., and Luikart, G. (2009a). Advancing ecological understandings through technological transformations in noninvasive genetics. *Molecular ecology resources*, 9:1279–301.

Beja-Pereira, A., Oliveira, R., Alves, P., Schwartz, M., and Luikart, G. (2009b). Advancing ecological understandings through technological transformations in noninvasive genetics. *Molecular ecology resources*, 9:1279–301.

Berger, C. S. and Aubin-Horth, N. (2018). An eDNA-qPCR assay to detect the presence of the parasite schistocephalus solidus inside its threespine stickleback host. *Journal of Experimental Biology*, 221(9).

Bergman, L. (2020). *Courtesy of L. Bergman*.

Bergman, P., Schumer, G., Blankenship, S., and Campbell, E. (2016). Detection of adult green sturgeon using environmental dna analysis. *PloS one*, 11:e0153500.

Chiu, G., Lockhart, R., and Routledge, R. (2006). Bent-cable regression theory and applications. *Journal of the American Statistical Association*, 101:542–553.

Coble, A., Flinders, C., Homyack, J., Penaluna, B., Cronn, R., and Weitemier, K. (2018). eDNA as a tool for identifying freshwater species in sustainable forestry: A critical review and potential future applications. *Science of The Total Environment*, 649.

Collins, R., Wangensteen, O., O'Gorman, E., Mariani, S., Sims, D., and Genner, M. (2018). Persistence of environmental DNA in marine systems. *Communications Biology*, 1.

Cristescu, M. and Hebert, P. (2018). Uses and misuses of environmental dna in biodiversity science and conservation. *Annual Review of Ecology, Evolution, and Systematics*, 49.

De Souza, L., Godwin, J., Renshaw, M., and Larson, E. (2016). Environmental dna (edna) detection probability is influenced by seasonal activity of organisms. *PLOS ONE*, 11.

Dunn, N., Priestley, V., Herraiz, A., Arnold, R., and Savolainen, V. (2017). Behavior and season affect crayfish detection and density inference using environmental DNA. *Ecology and Evolution*, 7.

Ficetola, G. F., Miaud, C., Pompanon, F., and Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology letters*, 4:423–5.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hindson, B. J. and Ness (2011). High-throughput droplet digital pcr system for absolute quantitation of DNA copy number. *Analytical Chemistry*, 83(22):8604–8610.

Hocking M.D, M. J., R, S., Allison M, B. L., B, K., B, S., M, L., and C, H. (2020). *The quantitation of salmonid envrionmental DNA at high and low copy number*.

Hunter, M., Oyler-McCance, S., Dorazio, R., Fike, J., Smith, B., Hunter, C., Reed, R., and Hart, K. (2015). Environmental dna (edna) sampling improves occurrence and detection estimates of invasive burmese pythons. *PLoS ONE*, 10.

Jerde, C., Mahon, A., Chadderton, W., and Lodge, D. (2011). "sight-unseen" detection of rare aquatic species using environmental dna. *Conservation Letters*, 4:150 – 157.

Jerde, C., Olds, B., Shogren, A., Allan, E., Mahon, A., Bolster, D., and Tank, J. (2016). The influence of stream bottom substrate on the retention and transport of vertebrate environmental dna. *Environmental Science and Technology*, 50.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis, a review and recent developments. *Philosophical Transactions*, 374.

Kassambara, A. and Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

Kralik, P. and Ricchi, M. (2017). A basic guide to real time PCR in microbial diagnostics: Definitions, parameters, and everything. *Frontiers in Microbiology*.

Lacoursière-Roussel, A., Rosabal, M., and Bernatchez, L. (2016). Estimating fish abundance and biomass from edna concentrations: variability among capture methods and environmental conditions. *Molecular Ecology Resources*, 16(6):1401–1414.

Lannoo, M. (2005). Amphibian declines: The conservation status of united states species. *Amphibian Declines: The Conservation Status of United States Species*.

Larionov, A., Krause, A., and Miller, W. (2005). A standard curve based method for relative real time pcr data processing. *BMC Bioinformatics*, 6:62.

Livak, K. and Schmittgen, T. (2002). Analysis of relative gene expression data using real-time quantitative pcr. *Methods (San Diego, Calif.)*, 25:402–8.

Lodge, D., Turner, C., Jerde, C., Barnes, M., Chadderton, W., Egan, S., Feder, J., Mahon, A., and Pfrender, M. (2012). Conservation in a cup of water: Estimating biodiversity and population abundance from environmental dna. *Molecular ecology*, 21:2555–8.

MacAdams, J. (2018). *Fish Forensics: Envrionmental DNA Detection of Juveinle Coho Salmon and Resident Salmonids in Pacific Coastal Streams*.

Meschiari, S. (2015). *latex2exp: Use LaTeX Expressions in Plots*. R package version 0.4.0.

Nevers, M., Byappanahalli, M., Morris, C., Shively, D., Przybyla-Kelly, K., Spoljaric, A., Dickey, J., and Roseman, E. (2018). Environmental dna (edna): A tool for quantifying the abundant but elusive round goby (neogobius melanostomus). *PLOS ONE*, 13:e0191720.

Penarrubia, L., Alcaraz, C., Bij de Vaate, A., Sanz, N., Pla, C., Vidal, O., and Viñas, J. (2016). Validated methodology for quantifying infestation levels of dreissenid mussels in environmental dna (edna) samples. *Scientific Reports*, 6:39067.

Petty, T., Thorne, D., Huntsman, B., and Mazik, P. (2014). The temperature-productivity squeeze: Constraints on brook trout growth along an appalachian river continuum. *Hydrobiologia*, 727.

Pilliod, D., Goldberg, C., Arkle, R., and Waits, L. (2013). Estimating occupancy and abundance of stream amphibians using environmental dna from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences*, 70:1123.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reed, R., Hart, K., Rodda, G., and Mazzotti, F. (2011). A field test of attractant traps for invasive burmese pythons (python molurus bivittatus) in southern florida. *Wildlife Research*, 38.

Rees, H., Gough, K., Middleditch, D., Patmore, J., and Maddison, B. (2015). Applications and limitations of measuring environmental dna as indicators of the presence of aquatic animals. *Journal of Applied Ecology*, 52.

Robinson, D., Hayes, A., and Couch, S. (2020). *broom: Convert Statistical Objects into Tidy Tibbles*. R package version 0.7.0.

Rodriguez-Lazaro, D. and Hernández, M. (2013). Real-time pcr in food science: Introduction. *Current issues in molecular biology*, 15:25–38.

Ruppert, K., Kline, R., and Rahman, M. (2019). Past, present, and future perspectives of environmental dna (edna) metabarcoding: A systematic review in methods, monitoring, and applications of global edna. *Global Ecology and Conservation*, 17:e00547.

Schmittgen, T. and Livak, K. (2008). Analyzing real-time pcr data by the comparative c(t) method. *Nature protocols*, 3:1101–8.

Selong, J., Mcmahon, T., Zale, A., and Barrows, F. (2001). Effect of temperature on growth and survival of bull trout, with application of an improved method for determining thermal tolerance in fishes. *Transactions of The American Fisheries Society - TRANS AMER FISH SOC*, 130:1026–1037.

Shogren, A., Tank, J., Allan, E., Olds, B., Jerde, C., and Bolster, D. (2016). Modelling the transport of environmental dna through a porous substrate using continuous flow-through column experiments. *Journal of The Royal Society Interface*, 13:20160290.

Shogren, A., Tank, J., Allan, E., Olds, B., Mahon, A., Jerde, C., and Bolster, D. (2017). Controls on edna movement in streams: Transport, retention, and resuspension open. *Scientific Reports*, 7:5065.

Snyder, D. (2003). Invited overview: Conclusions from a review of electrofishing and its harmful effects on fish. *Reviews in Fish Biology and Fisheries*, 13:445–453.

Sonderegger, D. (2020). *SiZer: Significant Zero Crossings*. R package version 0.1-7.

Strayer, D. (2010). Freshwater biodiversity conservation: Recent progress and future challenges. *Journal of the North American Benthological Society*, 29.

Takahara, T., Minamoto, T., Yamanaka, H., Doi, H., and Kawabata, Z. (2012). Estimation of fish biomass using environmental dna. *PloS one*, 7:e35868.

Thomsen and Willerslev (2015). Environmental DNA-an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, 183:4–18.

Tillotson, M., Kelly, R., Duda, J., Hoy, M., Kralj, J., and Quinn, T. (2018). Concentrations of environmental dna (edna) reflect spawning salmon abundance at fine spatial and temporal scales. *Biological Conservation*, 220:1–11.

Toms, J. and Lesperance, M. (2003). Piecewise regression: A tool for identifying ecological thresholds. *Ecology*, 84:2034–2041.

Tsuji, S., Yamanaka, H., and Minamoto, T. (2017). Effects of water ph and proteinase k treatment on the yield of environmental DNA from water samples. *Limnology*, 18.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., Zivot, E., Rocke, D., Martin, D., Maechler, M., and Konis., K. (2020). *robust: Port of the S+ "Robust Library"*. R package version 0.5-0.0.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilcox, T., McKelvey, K., Young, M., Sepulveda, A., Shepard, B., Jane, S., Whiteley, A., Lowe, W., and Schwartz, M. (2016). Understanding environmental dna detection probabilities: A case study using a stream-dwelling char salvelinus fontinalis. *Biological Conservation*, 194:209–216.

Willems, E., Leyns, L., and Vandesompele, J. (2008). Standardization of real-time pcr gene expression data from independent biological replicates. *Analytical biochemistry*, 379:127–9.

Yoccoz, N. (2012). The future of environmental dna in ecology. *Molecular ecology*, 21:2031–8.

Yoccoz, N., Bråthen, K., Gielly, L., Haile, J., Edwards, M., Goslar, T., Stedingk, H., Brysting, A., Coissac, r., Pompanon, F., Sønstebø, J. H., Miquel, C., Valentini, A., Bello, F., Chave, J., Thuiller, W., Wincker, P., Cruaud, C., Gavory, F., and Taberlet, P. (2012). Dna from soil mirrors plant taxonomic and growth form diversity. *Molecular ecology*, 21:3647–55.

Zhu, H. (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.1.0.