MADA Data Analysis Project

Monica Chan

2021-10-30

# 1 Summary/Abstract

*Write a summary of your project.*  
Applying MADA learned techniques to a RT-PCR dataset to determine the effectiveness of sample enhancement strategies for environmental surface sampling of low bio-burden viruses.

# 2 Introduction

## 2.1 Background

The global Coronavirus Disease 2019 (COVID-19) pandemic continues to persist due to the timing of an available vaccine, varying adherence to COVID-19 public health safety recommendations, and the emergence of SARS-CoV-2 (SC2) variants. Environmental reservoirs are a reported, indirect transmission route for viral pathogens. Although many viruses have a low infectious dose, they can be difficult to detect due to the low bioburden in the environment. The ability to detect viral pathogens is critical to identifying hidden reservoirs which can inform infection control. Culture-based methods can be labor and time intensive. The use and limitations of commercial “target capture” (TC) technologies are varied compared to molecular approaches utilizing automated extraction. The objective of this study is improve outbreak response through rapid pathogen detection by evaluating a TC technology’s ability to capture SC2 from complex matrices.

### 2.1.1 Questions/Hypotheses to be addressed

There are several major aims of this:

1. Do the sample enhancement strategies work?
2. How well do they work compared to each other?
3. Are there any environmental variables recorded that have a positive impact on detection?
4. Can this data be modeled to back calculate a sample’s original concentration based on the Ct values recovered?

Testing several sample enhancement strategies (Nanoparticles and Bait & Capture) for molecular detection of heat-inactivated SARS-CoV-2 from stainless-steel surfaces. The primary aim of this work is to determine if any of the enhancement strategies increase the concentration of RNA copies in a sample that will express itself as a lower Ct value from the Rt-PCR output. Lower Ct values equate to more copies being present in the sample meaning less time needed to confirm presence.

In the case the sample enhancements work, we would then compare how well they worked against an automated extraction method. The degree of “wellness” will be determined by how low the Ct value results are produced after a sample has going through a specific method.

In addition to the comparison of enhancement strategies the effects of the tool used for sampling (macrofoam or knit-fiber), time held, and other environmental variables will be explored any significant effects, with interest in positive impacts to detection.

## 2.2 Methods and Results

### 2.2.1 Experiment setup and conditions

Using established standard methods for environmental sampling with a self-contained polyurethane swab from stainless steel surface, a TC approach (NTP; Nanotrap® Magnetic Virus Particles, Ceres Nanosciences) was compared to an automated extraction method (MX48; Maxwell® RSC Viral Total Nucleic Acid Purification Kit, Promega) for recovery of heat-inactivated SC2 from the polyurethane swab as detected by rRT-PCR

## 2.3 Data aquisition

Reports per each rRT-PCR run were generated and quality checked on their respective machines before transfer to a Teams repository. Summary data was selected from the reports and added to additional environmental data gathered on based on the extraction date of the sample.

Completed run data is then collated into one file and read into R./A composited data set was compiled and saved as a csv and used as the base data for the project

### 2.3.1 Data import and cleaning

processingscript.R was modified to do initial cleaning and preparation of the dataset for further exploration.

Data was imported and cleaned. Data was then split between the two different experimental types (P1 and P2).

### 2.3.2 Data exploration

*Use a combination of text/tables/figures to explore and describe your data. You should produce plots or tables or other summary quantities for the most interesting/important quantities in your data. Depending on the total number of variables in your dataset, explore all or some of the others. FIgures produced here might be histograms or density plots, correlation plots, etc. Tables might summarize your data.*

*Continue by creating plots or tables of the outcome(s) of interest and the predictor/exposure/input variables you are most interested in. If your dataset is small, you can do that for all variables. Plots produced here can be scatterplots, boxplots, violinplots, etc. Tables can be simple 2x2 tables or larger ones.*

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. t-tests, simple regression model with 1 predictor, etc.) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

#### 2.3.2.1 All Data

Table 2.1 shows a table summarizing the data.

#### 2.3.2.2 Phase 1

### 2.3.3 Phase 2

### 2.3.4 Data Analysis

### 2.3.5 Data Modeling

## 2.4 Exploratory analysis

Table 2.1: Data summary table.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | EXP | Target | Dilution.Factor | Method | Type | n | Mean | SD |
| Min. | 34 | 32 | -5.000000 | 28 | 2 | 2.00000 | 19.27845 | 0.0106066 |
| 1st Qu. | 30 | 32 | -4.000000 | 34 | 6 | 5.00000 | 23.30454 | 0.7214929 |
| Median | 34 | 32 | -3.000000 | 2 | 24 | 6.00000 | 26.67273 | 0.9764277 |
| Mean | 30 | 32 | -2.862069 | 28 | 32 | 10.65625 | 27.03186 | 1.4500559 |
| 3rd Qu. | 34 | 32 | -2.000000 | 34 | 2 | 15.00000 | 30.53858 | 2.1052618 |
| Max. | 30 | 32 | -1.000000 | 2 | 6 | 42.00000 | 36.18193 | 4.2336573 |
| NA’s | 34 | 32 | 6.000000 | 28 | 24 | 2.00000 | 25.00000 | 25.0000000 |

Figure 2.1 shows a scatterplot figure produced by one of the R scripts.

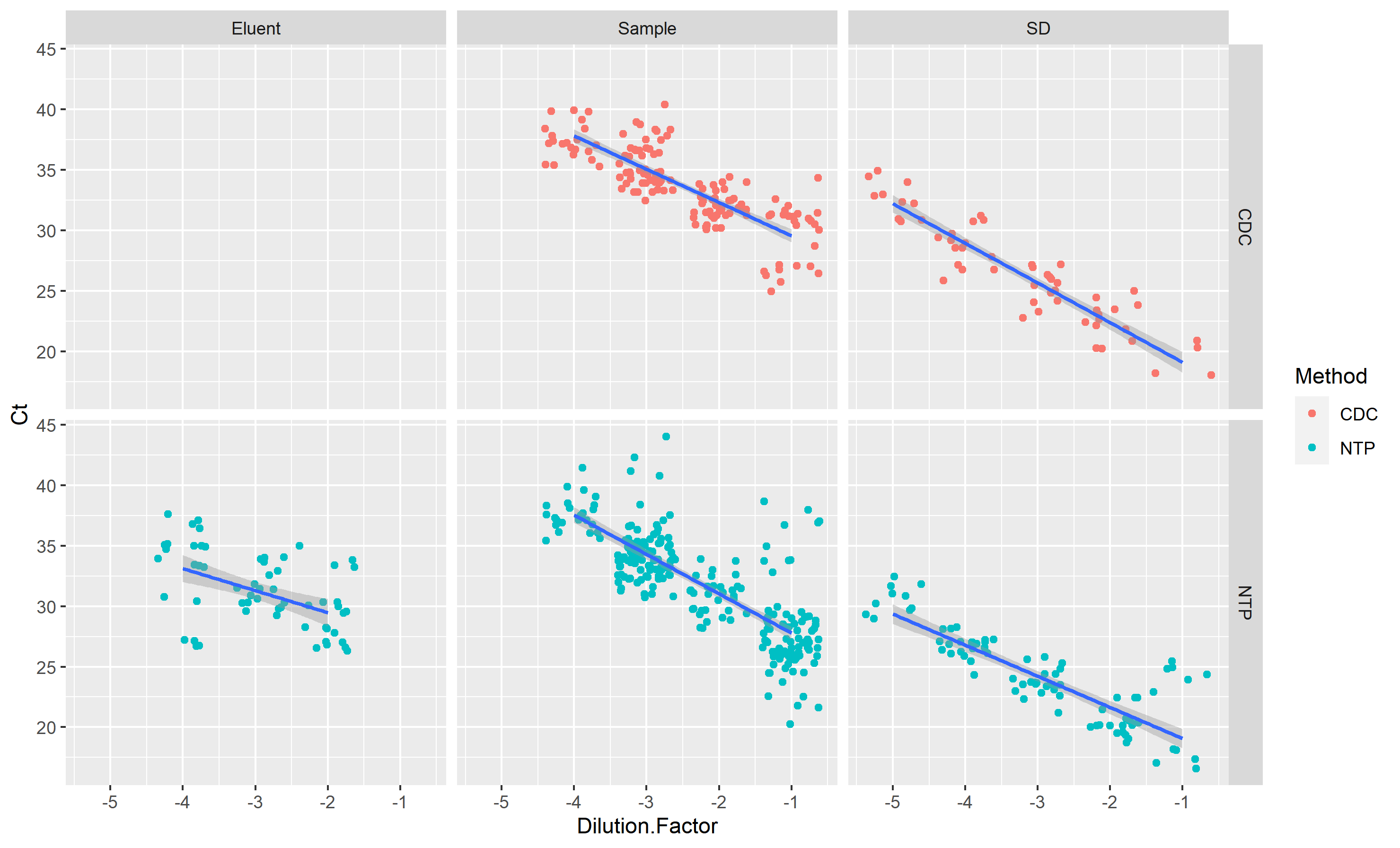


Figure 2.1: Analysis figure.

## 2.5 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example table 2.2 shows a table summarizing a linear model fit.

Table 2.2: Linear model fit table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.2758699 | 0.2595347 | 1.06294 | 0.2882583 |
| Ct | -0.0937034 | 0.0083682 | -11.19757 | 0.0000000 |

# 3 Discussion

## 3.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 3.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 3.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*, 1314–1315. <https://doi.org/10.1126/science.aaa6146>