Leveraging Statistical Distributions for RNA sequencing across time

Presented By Rianna Santra and Ho Tin (Alex) Fan
Mentored by Prof. Gil Alterovitz

Outline

- Background and introduction to the topic
- Methodology and Objectives
- Evaluations
- Conclusion & Future Directions
- Acknowledgements

Introduction

- **Topic**: Baby Feeding Behaviors
- **Context**: Trainers (placebo and real groups) are assigned to babies to study the effectiveness of the treatment

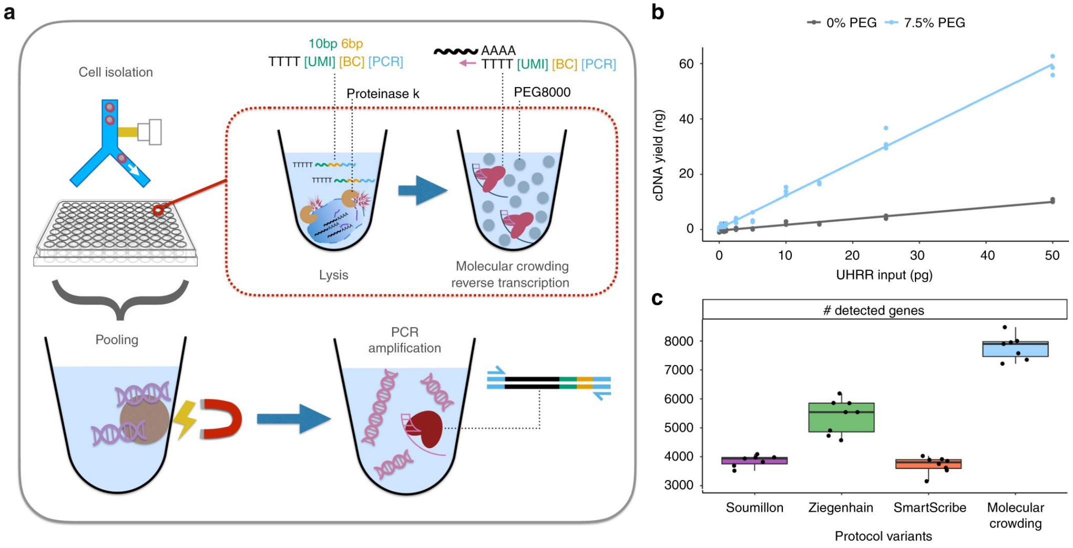
Methodology and Objectives

- **RNA Sequencing** is used on various genes of interest to measure the change in gene expression
- We leverage **statistical distributions** to both
  - **determine the effectiveness** of the treatments and
  - **devise a prediction model** of whether individual babies will respond to the treatment

Methodology and Objectives

Genes of interest that we studied include:

- CDH13
- FOXP2
- NPHP4
- NPY2R
- PLXNA1
- WNT3



Methodology and Objectives

We combined the time-series data (1st, 2nd, 3rd, 4th week of PCR 2^(cycle threshold) for each gene) and performed myriad statistical tools and models on them such as

- ANOVA Tests (Analysis of Variance)
- Bonferroni corrections
- Weighted gene progression time average

Evaluations

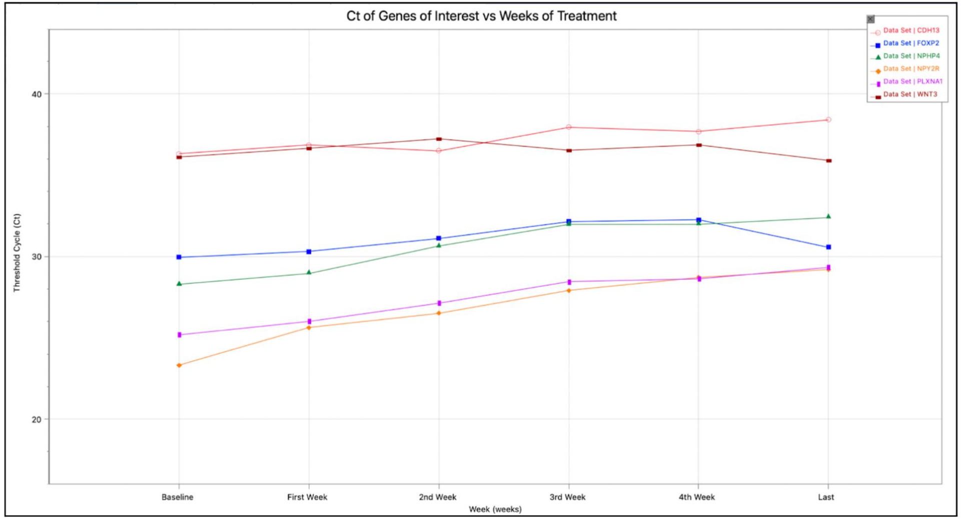
In total, we approached our research project through 6 individual questions

Evaluations - Question 1

Did the expression patterns of the genes change over time with advancing postmenstrual age and feeding status? Is there a distinct gene expression pattern in a non-feeder v. successful feeder? Can you see a maturing pattern or is it simply a random pattern?

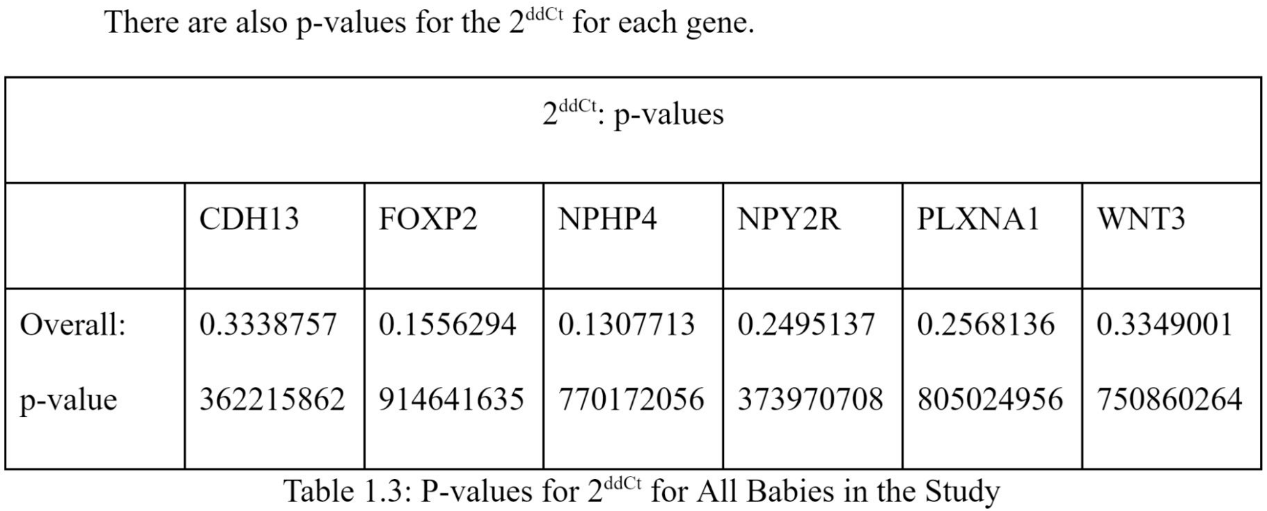
Examine as 1.) a whole, 2.) based on treatment status, and 3.) by sex

Evaluations - Question 1



Graph 1.1: Ct for All Babies in the Study

Evaluations - Question 1



There are also p-values for the 2^{ddCt} for each gene.

2^{ddCt}: p-values						
	CDH13	FOXP2	NPHP4	NPY2R	PLXNA1	WNT3
Overall: p-value	0.3338757362215862	0.1556294914641635	0.1307713770172056	0.2495137373970708	0.2568136805024956	0.3349001750860264

Table 1.3: P-values for 2^{ddCt} for All Babies in the Study

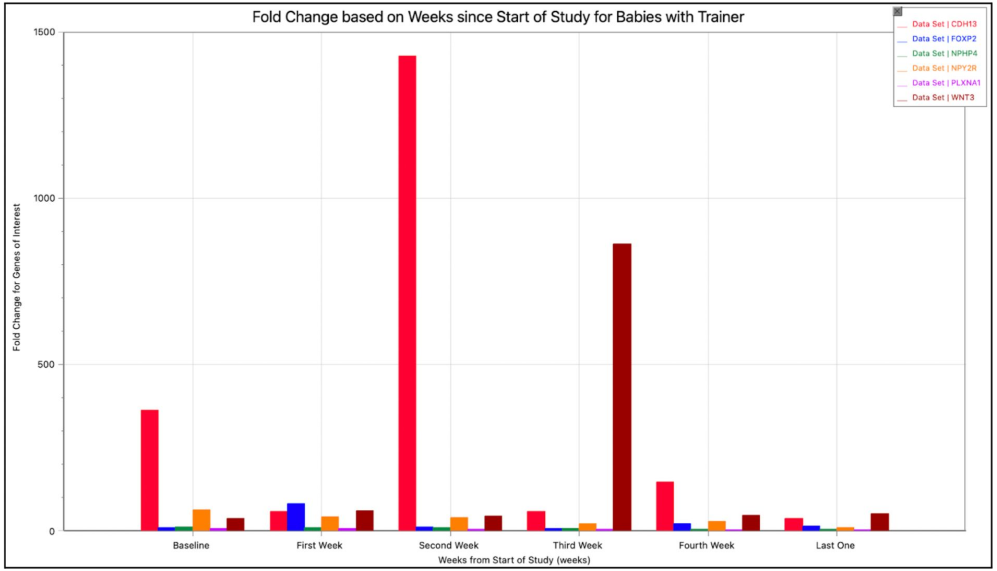
Evaluations - Question 1

Results: Some genes did change over the course of study, but not all of them. The genes NPHP4, NPY2R, and PLXNA1 are the genes that mature over the course of the babies learning how to feed. There is a little difference between babies who trained with the trainer and babies who did not, and a difference between male babies and female babies.

Evaluations - Question 2

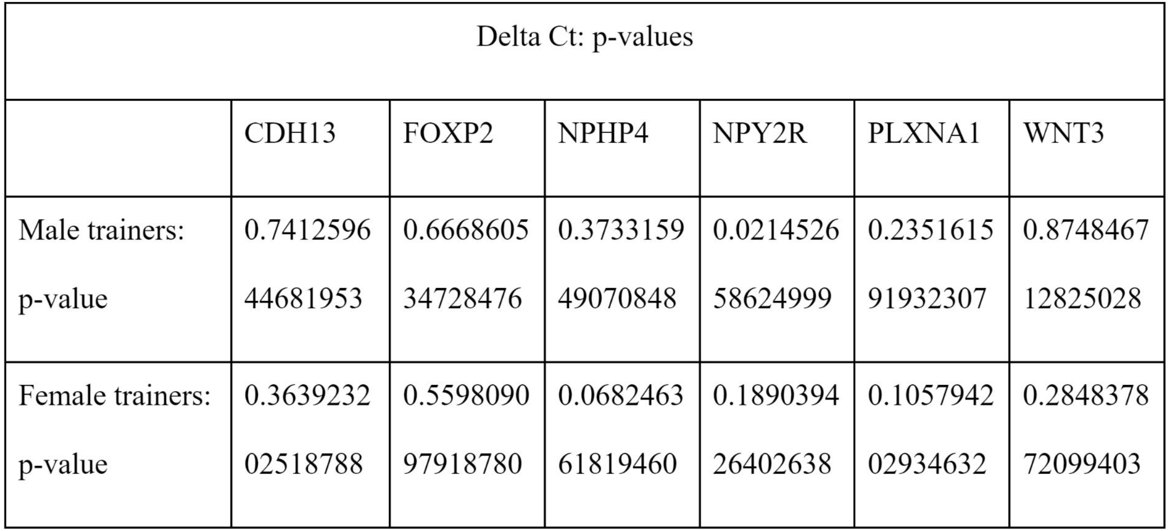
Do infants who undergo therapy have a more rapidly maturing gene expression feeding pattern compared to infants who did not? Importantly are there certain genes that seemed to change in response to therapy—NOT all genes may have been impacted. Sex may play a role here.

Evaluations - Question 2



Graph 2.1: Fold Change for Babies without Trainer for All Genes

Evaluations - Question 2



Delta Ct: p-values						
	CDH13	FOXP2	NPHP4	NPY2R	PLXNA1	WNT3
Male trainers: p-value	0.741259644681953	0.666860534728476	0.373315949070848	0.021452658624999	0.235161591932307	0.874846712825028
Female trainers: p-value	0.363923202518788	0.559809097918780	0.068246361819460	0.189039426402638	0.105794202934632	0.284837872099403

Table 2.1: P-values for Delta Ct for Babies with Trainer Separated by Sex

Evaluations - Question 2

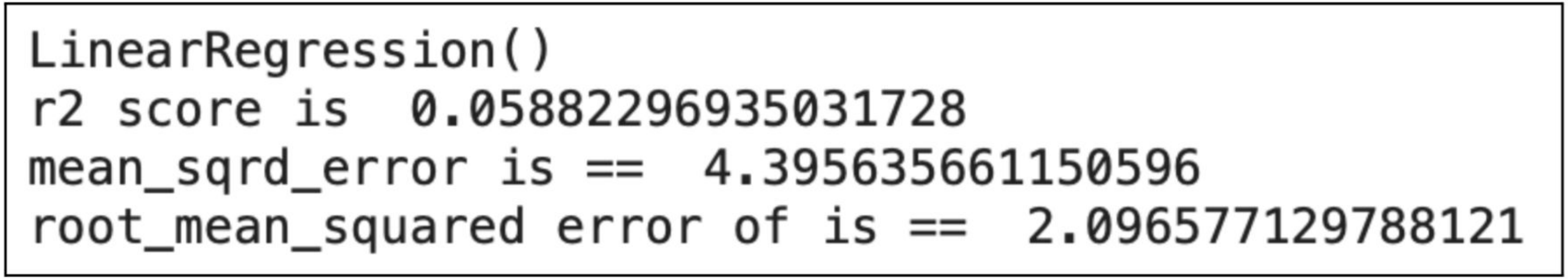
Results: If fold change steadily decreasing means there is a more rapidly maturing gene expression, then yes, but only for certain genes. The gene NPY2R changes a lot, while the genes NPHP4 and PLXNA1 change only a little bit. The other genes simply have random fold changes.

Evaluations - Question 3

Could you predict at the beginning of therapy based on gene expression patterns which babies would be more likely to respond to the therapy and learn to feed sooner?

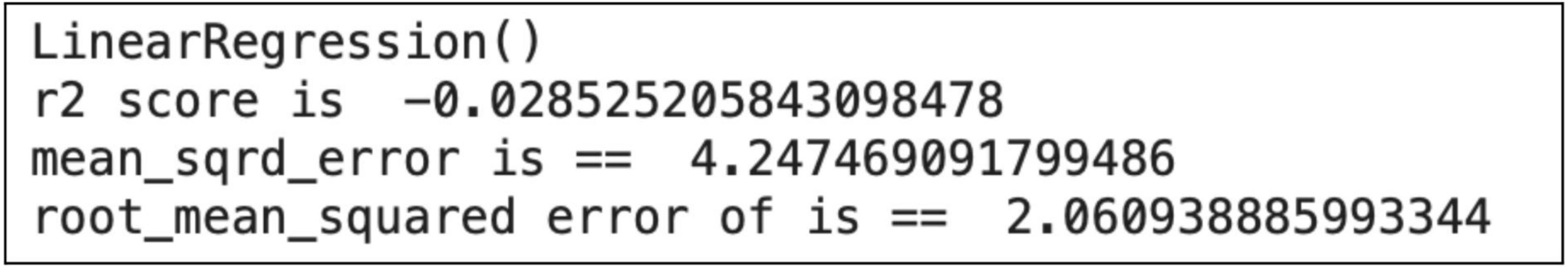
Evaluations - Question 3

3.1 Based on gene expression patterns at the beginning of therapy.



```
LinearRegression()
r2 score is  0.05882296935031728
mean_sqrd_error is ==  4.395635661150596
root_mean_squared error of is ==  2.096577129788121
```

Evaluations - Question 3



```
LinearRegression()
r2 score is  -0.028525205843098478
mean_sqrd_error is ==  4.247469091799486
root_mean_squared error of is ==  2.060938885993344
```

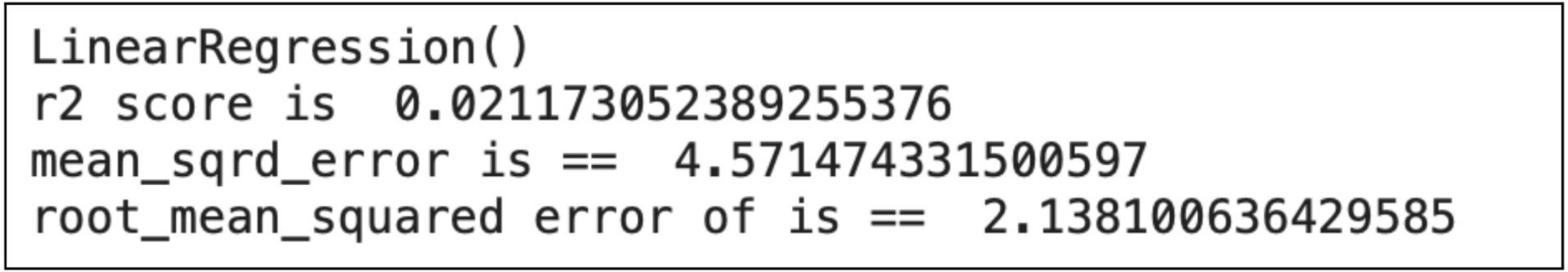
Evaluations - Question 3

Results: We keep on getting very low scores correct, likely due to the amount of noises in the dataset. Therefore, we do not think we can predict at the beginning of therapy if a baby can respond to therapy and learn to feed sooner given the current dataset.

Evaluations - Question 4

Can you generate a prediction model based on initial gene expression pattern, sex, and gestational age to predict responders?

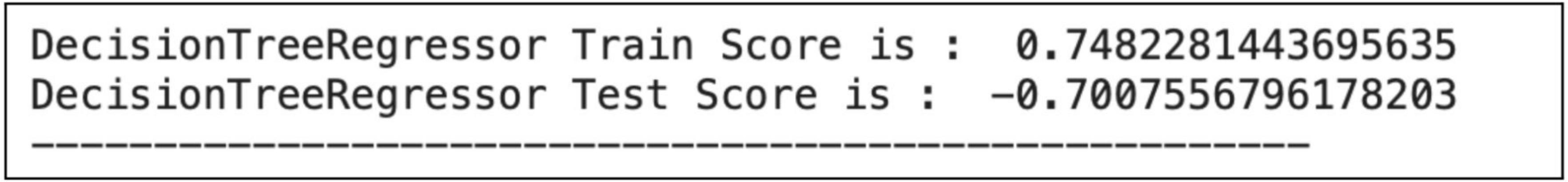
Evaluations - Question 4



```
LinearRegression()
r2 score is  0.021173052389255376
mean_sqrd_error is ==  4.571474331500597
root_mean_squared error of is ==  2.138100636429585
```

Evaluations - Question 4

We also used a Decision Tree Regressor Model, and although the training score was very good, the test score was negative.



```
DecisionTreeRegressor Train Score is :  0.7482281443695635
DecisionTreeRegressor Test Score is :  -0.7007556796178203
----------------------------------------------------
```

Evaluations - Question 4

Results: Similar to the previous prediction models, most of our statistical techniques yield extremely low scores, partly due to the low number of samples and large amount of noises. Even more advanced techniques such as Deep Learning yield no better results. Thus we can not generate an accurate prediction model given the current dataset

Evaluations - Question 5

Can you generate a prediction model based on gene expression patterns over time to predict when a baby became a successful feeder? Can you do it independent of therapy? Based on sex?

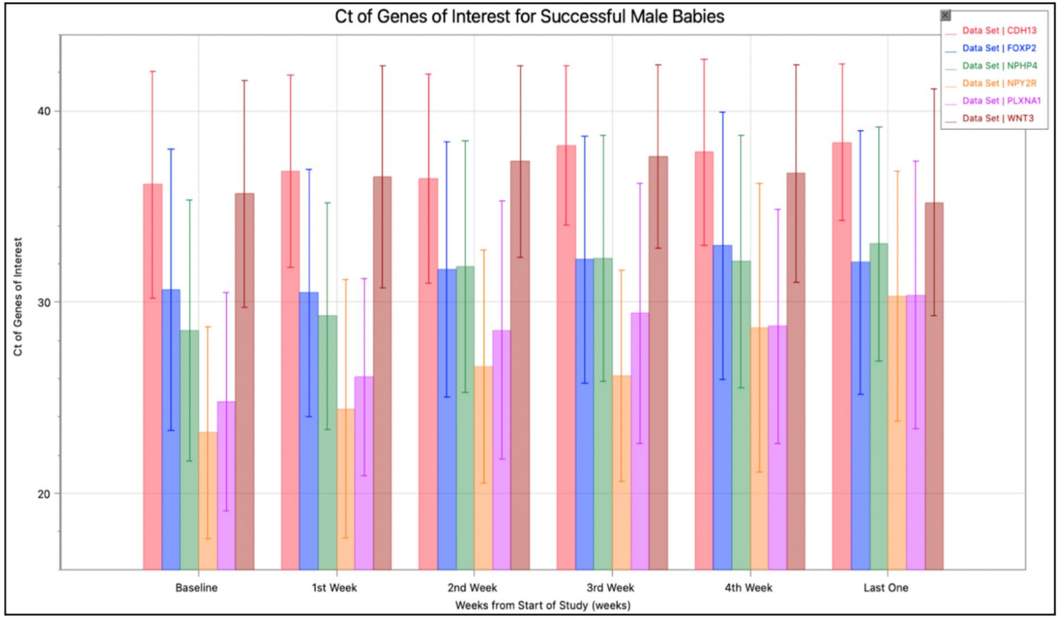
Evaluations - Question 5

Results: We were not able to generate one successfully independent of therapy because again, there wasn’t enough data. We would need much more data in order to create a successful algorithm. There is only so much a computer can do with 112 samples.

Evaluations - Question 6

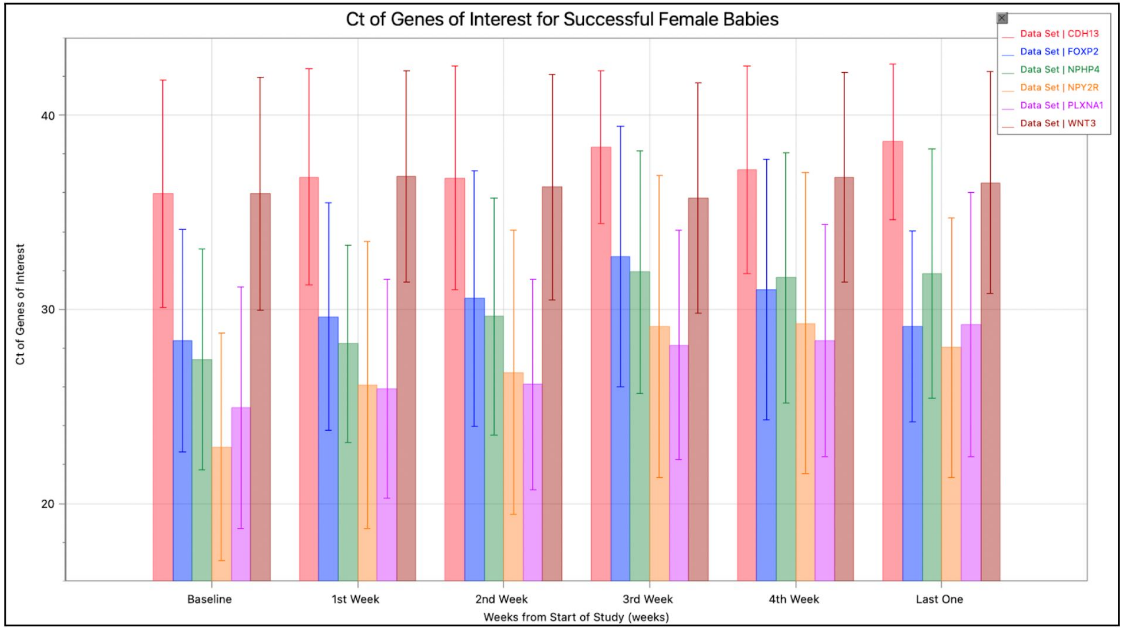
Finally, does a successful male oral feeder have a different gene expression pattern compared to a successful female oral feeder?

Evaluations - Question 6



Graph 6.1: Ct of Successful Male Babies for All Genes

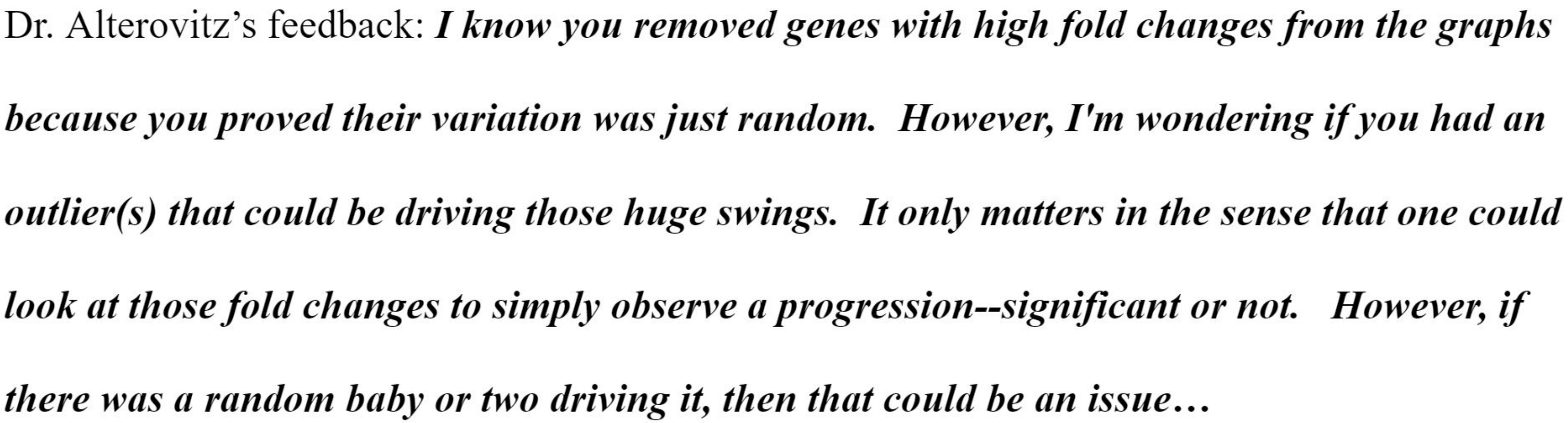
Evaluations - Question 6



Evaluations - Question 6

Results: We can see that these graphs are relatively the same, within error. Therefore, we can conclude that a successful male oral feeder does not have a different gene expression compared to a successful female oral feeder.

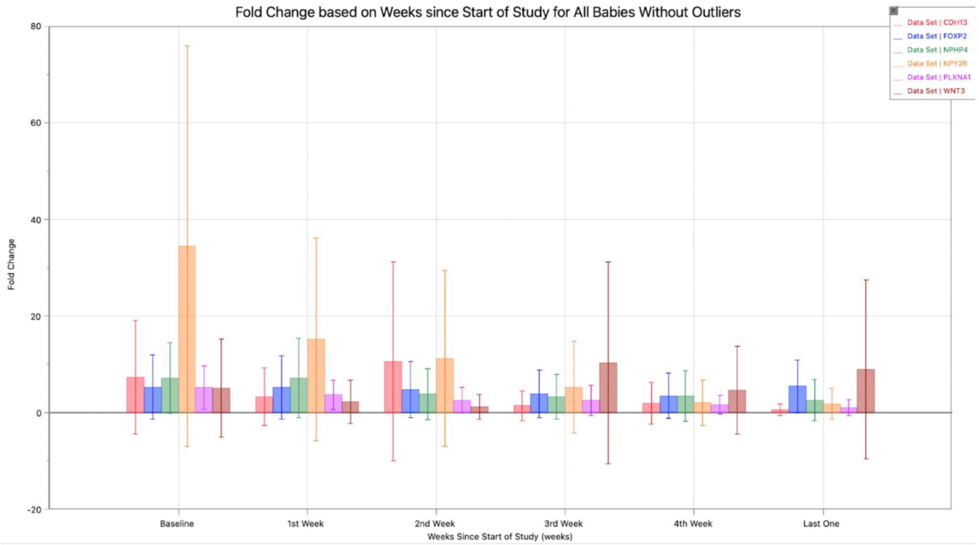
Evaluations - Additional Analysis I



Dr. Alterovitz’s feedback: ***I know you removed genes with high fold changes from the graphs because you proved their variation was just random. However, I'm wondering if you had an outlier(s) that could be driving those huge swings. It only matters in the sense that one could look at those fold changes to simply observe a progression--significant or not.   However, if there was a random baby or two driving it, then that could be an issue…***

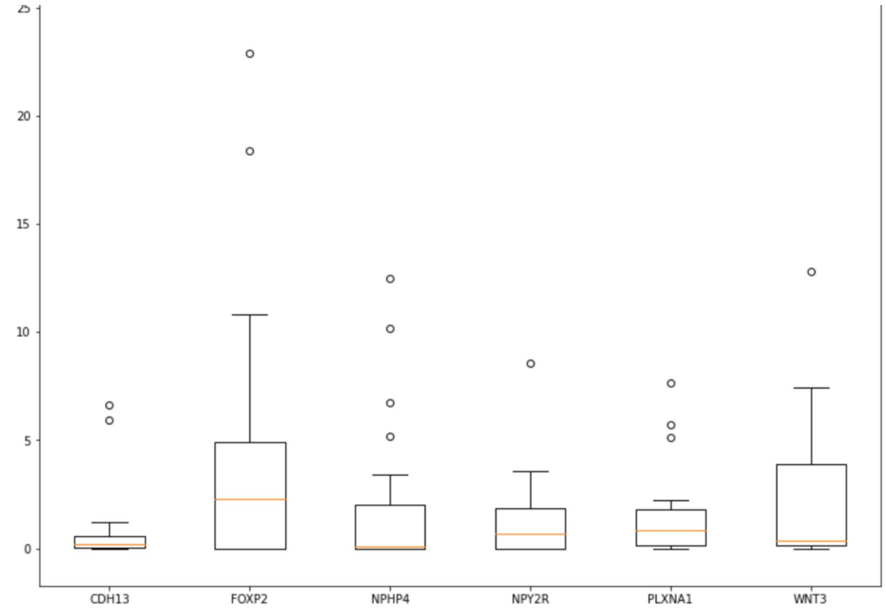
Evaluations - Additional Analysis I

Outcome from Additional Analysis:



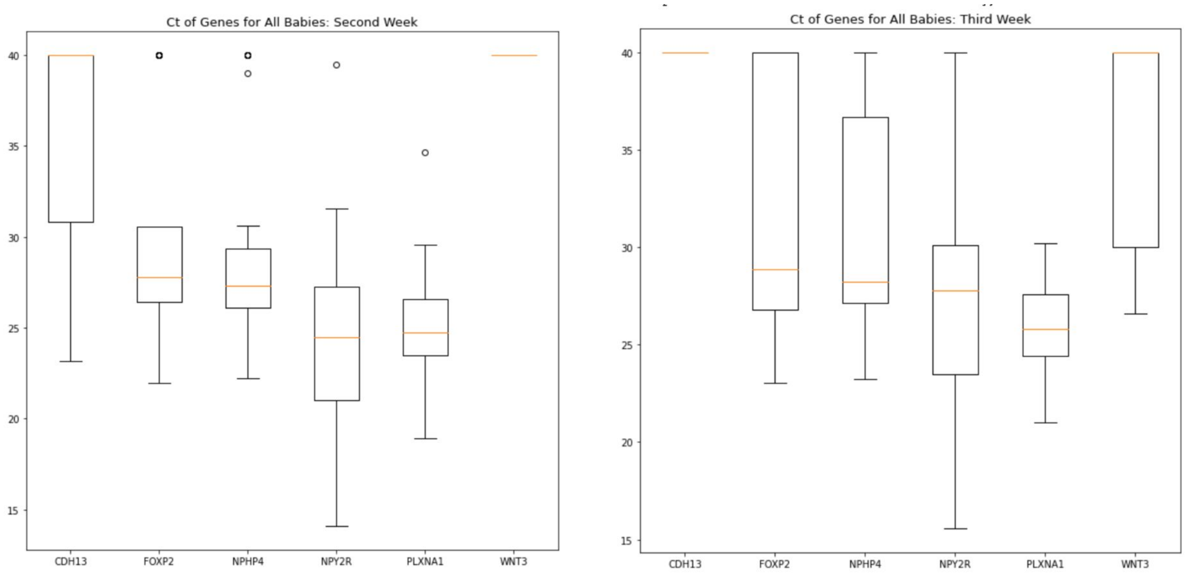
Graph 7.1: Fold Change for All Babies Without Outliers for All Genes

Evaluations - Additional Analysis I



Graph 7.9: Boxplot for All Babies with Trainers: Fourth Week

Evaluations - Additional Analysis I

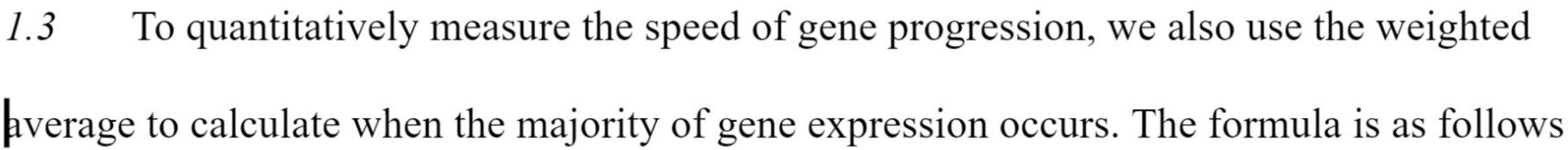


Graphs 8.1-8.4: Boxplots of Ct of Genes for All Babies: Baseline through Third Week

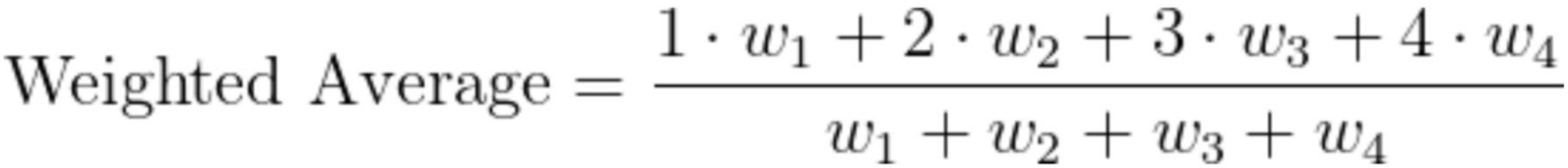
Evaluations - Additional Analysis I

Results: We can see here that FOXP2, NPHP4, NPY2R, and PLXNA1 had a higher Ct (lower gene expression) in the babies without the trainer than they had with the trainer. We conclude that the trainer does indeed change gene expression. This is a difference in these graphs without the outliers - we can see that the gene FOXP2 does change, albeit slightly.

Evaluations - Additional Analysis II

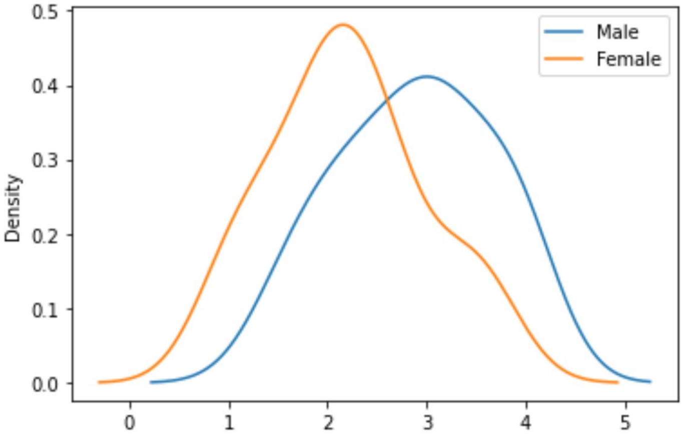


1.3 To quantitatively measure the speed of gene progression, we also use the weighted average to calculate when the majority of gene expression occurs. The formula is as follows



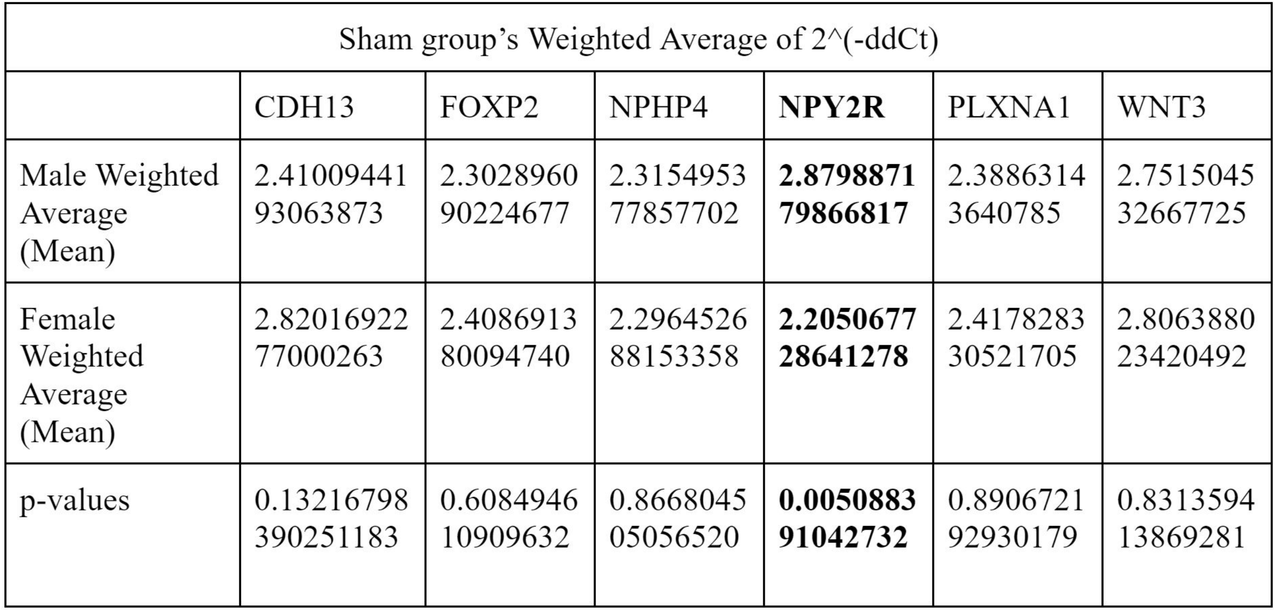
$$\text{Weighted Average} = \frac{1 \cdot w_1 + 2 \cdot w_2 + 3 \cdot w_3 + 4 \cdot w_4}{w_1 + w_2 + w_3 + w_4}$$

Evaluations - Additional Analysis II



Graph 8.17: Sham group male v. female babies' weighted average of NPY2R 2^(-ddCt)

Evaluations - Additional Analysis II



Sham group's Weighted Average of 2^(-ddCt)						
	CDH13	FOXP2	NPHP4	**NPY2R**	PLXNA1	WNT3
Male Weighted Average (Mean)	2.4100944193063873	2.302896090224677	2.315495377857702	**2.879887179866817**	2.38863143640785	2.751504532667725
Female Weighted Average (Mean)	2.8201692277000263	2.408691380094740	2.296452688153358	**2.205067728641278**	2.417828330521705	2.806388023420492
p-values	0.13216798390251183	0.608494610909632	0.866804505056520	**0.005088391042732**	0.890672192930179	0.831359413869281

Table 1.2: P-Values for Sham group male v. female babies' weighted average of 2^(-ddCt)

Evaluations - Additional Analysis II

Results: There is a statistically significant difference in NPY2R and PLXNA1's gene expression between the sham and trainer groups. Specifically, NPY2R's difference is very prominent amongst male babies, whereas PLXNA1 difference is only noticeable amongst female babies. In both cases, the trainer group displayed a smaller weighted average, indicating that the majority of gene expression occurred earlier and thus the maturation progression is quicker when babies are given therapy.

Evaluations - Additional Analysis Summary

The results of this supplementary study demonstrate that the trainer therapy does have some impact on the babies’ gene expression, and explore the extent of the effects that the trainer has on the acceleration of **NPY2R** and **PLXNA1** gene expression patterns. The discoveries in this supplementary analysis mostly align with the main paper’s conclusions, which stated that there are statistically significant fold changes for the **NPY2R** and **PLXNA1** genes.

Summary of our Results

- **The trainer does have some statistically significant impact** on the RNA gene expression patterns.
- Some genes exhibited **patterns of upregulation** of fold changes depending on sex and treatment status.
- We couldn’t generate a reliable prediction model due to the noise and small number of 112 samples.

Future Directions

- With more data, we hope to more effectively eliminate outliers and thus **generate accurate prediction models** for treatment outcome.

- **Extending our current methods to more genes of interests**, and investigate whether they have similar or perhaps even more significant patterns

Acknowledgements

We would like to recognize the following two members who have graduated at the time of our presentation but have contributed to our project:

- **Powell Zhang**
- **Arthur Hu**

Acknowledgements

In addition, we would like to thank our mentor **Gil Alterovitz** for providing the data for us to do the data analysis and the questions to guide our research. We would also like to thank his assistant **Ning Xie** for answering our questions and giving us information about this project, and the entire MIT PRIMES program for providing us the opportunity to participate in this research.

Questions?