O. kimflemingiae and *N. crassa*: A Case of Clock Copycats?

Erel Amit, Aidan Batista, Sarah Gorbatov, Justin Morrill

Introduction

In many organisms, the proteins responsible for important physiological and behavioral functions are expressed periodically in a clock network. In the case of a circadian clock network, the period is approximately 24 hours and is reinforced by zeitgebers like light (Pilorz, Helfrich-Förster, & Oster, 2018).

One intriguing organism with behaviors controlled by a circadian clock network is the recently sequenced parasitic fungus *Ophiocordyceps kimflemingiae*, a part of the *O. unilateralis* complex. Often referred to as the "zombie ant fungus," *O. unilateralis* infects *Camponotus leonardi,* a species of Carpenter ant, and manipulates their behavior. When an *O. unilateralis* spore makes contact with a Carpenter ant while it is out foraging, the spore attaches to the cuticle and spreads throughout the body to the muscles. Typically within 16-24 days, after the infection grows substantially in mass and the muscles atrophy, the ant's behavior abruptly changes; whereas before it wandered aimlessly, it now clamps down on a leaf vein in a "death grip" and dies. This transition from wandering to death grip is synchronized to solar noon. Within a few days after mortality, a stalk grows from the ant's head, and new spores are born (Hughes et al., 2011).

Underlying this timed parasite-host interaction is a circadian clock network of proteins and their transcription factors. This network was first studied by Bekker et al. in 2017. In the study, *O. cordyceps* spores were harvested at 4-hour intervals in either 24-hour cyclic light-dark (LD) or free-running dark (DD) conditions. RNA-Seq was performed on 48-hour time courses,

which yielded transcription-level data for 8,629 genes. With a cutoff of $p \leq 0.057$, 333 periodic genes in LD (on a 24-hr cycle) and 154 in DD (on a 26-hour cycle) were identified with the JTK_CYCLE periodicity algorithm. Since the LD data resulted in the detection of more periodic genes and lower p values, the study primarily proceeded with the LD genes, of which 14 were identified as transcription factors. However, since the study only analyzed genes through a periodicity algorithm and not also an edge-finding algorithm, the most the study could conclude is the presence of oscillating mRNA levels, i.e. the potential nodes of the *O. cordyceps* network. The edges, i.e. which nodes activate or repress other nodes, remain unexplored.

Fortunately, the clock networks of other fungi are well-established, namely that of the model fungus *Neurospora crassa*. In *N. crassa*, there are only a handful of genes largely responsible for the organism's clock mechanism. They are the White Color Complex (WCC) consisting of White Collar-1 (WC-1) and White Collar-2 (WC-2), the frequency gene (FRQ), and an FRQ-interacting RNA helicase (FRH). The network works as follows: light activates WCC, WCC activates FRQ, FRQ forms a complex with FRH (2 FRQ:1 FRH), and this complex then represses WCC to close the negative feedback loop, the transcriptional signature of periodicity. FRQ is post-translationally phosphorylated by the CK1 and CK2 kinases, ubiquitinated by the ubiquitin ligase FWD-1, and degraded by the proteasome, which then results in rising WCC levels, starting the cycle anew (Bekker et al., 2017). See the transcription and translation flow in *Figure 1*.

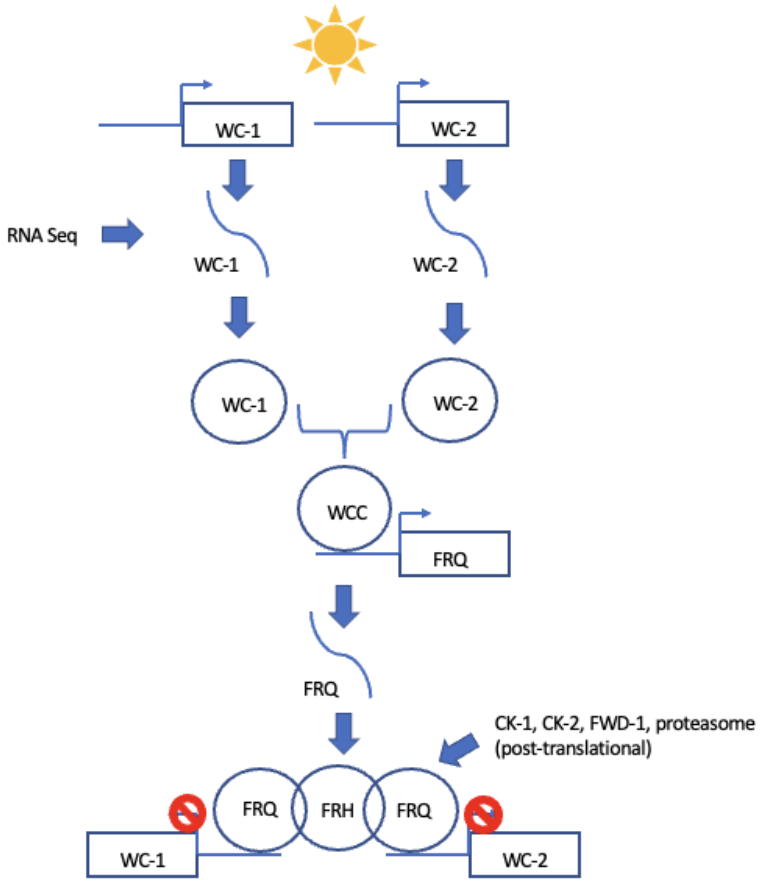

Figure 1. Transcription and translation flow of N. crassa network.

The Bekker paper's chromosomal alignments revealed 8 homolog genes with low E values between *O. cordyceps* and *N. crassa*, which included FRQ and WC-1 and 2 (*Figure* 2).

Gene ID	*Neurospora crassa* OR74A Homolog	E Value NCBI BlastP
Ophio5|6046	Frequency (*frq*) NCU02265	0
Ophio5|4975	White Collar 1 (*wc-1*) NCU02356	0
Ophio5|889	White Collar 2 (*wc-2*) NCU00902	2E-174

Figure 2. N. crassa and O. kimflemingiae homologs and their BlastP alignment E values

We hypothesize, therefore, that the *O. cordyceps* clock network closely resembles the *N. crassa* network with the same three core nodes and edges of the negative feedback loop (*Figure 3*).

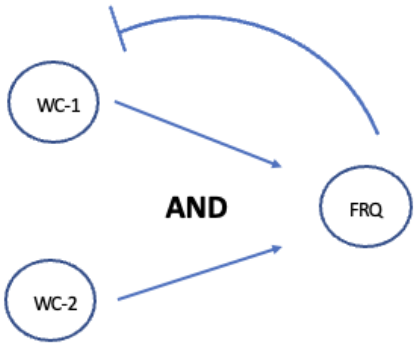

Figure 3: Proposed O. cordyceps network, which mirrors the negative feedback loop of the N. crassa network. It remains uncertain whether the inhibition arrow points to WC-1, WC-2, or both.

It is uncertain, however, whether FRQ transcriptionally represses WC-1, WC-2, or both. The following analysis seeks to (a) find evidence for the proposed three-node network and (b) confirm which of the two White Collar promoters the FRQ transcription factor binds. This will be accomplished with an edge-finding algorithm, a boolean model, ordinary differential equations, the periodicity algorithm utilized in Bekker et al., and several statistical tests.

Methods

Local Edge Machine:

The Local Edge Machine (LEM) was run on the LD data for *O. cordyceps* in order to determine if the data from Bekker et al. supports the hypothesized clock network. Since the primary genes in the network were WC-1, WC-2, and FRQ, the homologs of those genes were used as the targets for LEM. Taking the five edges with the lowest loss resulted in a network similar to what was hypothesized.

The only differences were that WC-1 and WC-2 had arrows activating themselves, which is an issue often found with LEM. Since LEM evaluates each edge individually without considering the network as a whole, it makes sense that a node's expression would be highly correlated with itself, and thus a node activating itself would be very appealing to an algorithm like LEM.

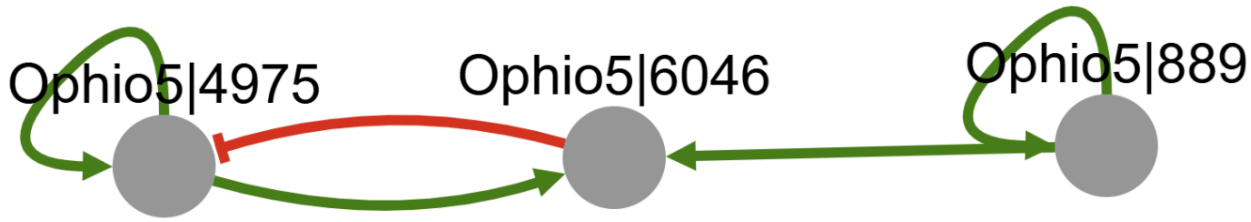

Figure 4. LEM network of WC-1 (Ophio5|4975), WC-2 (Ophio5|889), and FRQ (Ophio5|6046)

Boolean & ODE Modeling:

The next step in verifying this network is to model it. There are two possible ways of doing this: a discrete (boolean) model and a continuous model. Since the boolean model would be simpler and less prone to noise, it was tested first. This model used BooleSim, an open-source software for simulating boolean networks. The BooleSim model also did not include the self-regulation arrows, since those were likely an artifact of how LEM is designed and not true to the network. Additionally, AND logic was used for the interaction of WC-1 and WC-2 on FRQ, since in *N. crassa*, WC-1 and WC-2 must both be present to form WCC before interacting with FRQ.

The continuous model was made from a system of ODEs. There are two equations used for each form of regulation, one for activation and one for repression. The activation equation was:

$$f_j' = \gamma - \beta x_j + \alpha\left(\frac{x_i^n}{x_i^n + k^n}\right)$$

f_j' represents the change in activation of the jth gene over time, γ is the basal transcription rate, β is the degradation rate, x_i is the gene activating x_j, and α, n, and k are all coefficients that change how x_i acts. The repression equation is similar:

$$f_j' = \gamma - \beta x_j + \alpha\left(\frac{k^n}{x_i^n + k^n}\right)$$

LEM uses these equations to model each edge, so the output for LEM includes possible values for each of these parameters. These parameters were used in Scipy's "odeint" ODE solver to create a model for the proposed network.

Periodicity Algorithms:

In order to further verify the nodes in the hypothesized network, three different periodicity algorithms were applied to the LD time series data for *O. cordyceps*. The first was JTK_CYCLE, which compares the increasing & decreasing patterns of the data to that of sine or cosine curves. It takes the following as inputs: the LD dataset, a period range of (22, 26) hours, and a period step of 24 hours. It returns the original list of genes, ordered by their periodicity scores. The second periodicity algorithm used was DL, which works similarly to JTK_CYCLE, except it places more emphasis on the amplitude of the gene expression data profiles. With several other factors held constant, DL takes the average period (24 hours) and the dataset as inputs, and it returns a list similar in structure to that of JTK_CYCLE, now ordered by its own criteria. To ensure maximum reliability of candidate genes, the final algorithm used was DLxJTK, which, in essence, combines the criteria of both JTK_CYCLE and DL. DLxJTK takes the dataframes from the two other algorithms as inputs, returning an ordered list that accounts for both the regulation scores from DL and the periodicity scores from JTK_CYCLE.

Statistical Methods:

We began by normalizing the WC-1, WC-2, and FRQ time series data by subtracting from each data point that gene's mean and dividing by its standard deviation. This data was used to graph *figure 7* and *figure 8*.

$$x_{normalized} = \frac{x - mean(x)}{\sigma(x)}$$

We also sought to confirm the negative correlation between WC-1 and FRQ observed in *figure 8* and *figure 9*. To this end, an OLS regression was run on WC-1 (called W in the formula below) and FRQ (called F in the formula below) with the following form. The full results can be found in *Appendix 1*.

$$W_i = \beta_0 + \beta_1 \cdot F_i + \varepsilon_i$$

A one-tailed F-test was also run with $\alpha = 0.05$ in an attempt to confirm that WC-1 levels were best predicted by FRQ (and vice versa, since regression relationships are symmetrical). The genes used in the unrestricted model were WC-1, FRQ, and the top 9 most periodic genes found by DLxJTK. The full results and genes used can be found in *Appendix 2*, but to summarize, a p-value of 0.469 was generated by the test. In the context of this one-tailed F-test, this score means that we are unable to reject the null hypothesis that the restricted linear relationship: $W_i = \beta_0 + \beta_1 \cdot F_i + \varepsilon_i$ is the correct model.

Results

The boolean model resulted in a periodic cycle. Due to the simplicity of the small network, consisting only of three nodes and three edges, and the inherent simplicity of a boolean model, the resulting data was also simplistic. The network had both WC-1 and FRQ oscillating, with WC-2 staying activated. If WC-2 started "off," the network would stagnate due to the "AND" logic. The primary goal of this boolean model is to show whether a periodic clock cycle is probable with this network, not to create a realistic model of gene expression over time.

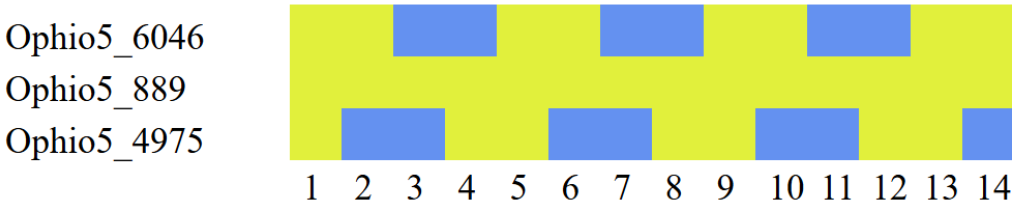

Figure 5. Expression over time of FRQ, WC-2, and WC-1 (top to bottom) in the boolean model

The continuous model was less conclusive, without any cyclic or interesting behavior occurring. The model predicted WC-1 sharply decreasing and both WC-2 and FRQ increasing until plateauing at a single value, strongly differing from the periodic behavior measured from these genes in Bekker et al.

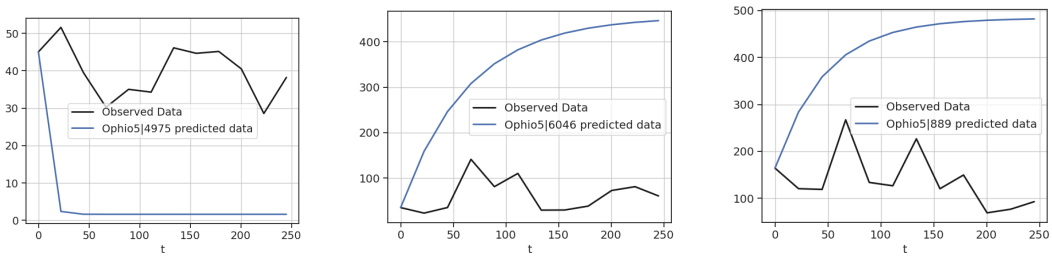

Figure 6. ODE model results for the WC-1, FRQ, and WC-2 homologs (left to right)

The three periodicity algorithms returned mixed results. FRQ was the seventh most periodic gene according to JTK_CYCLE and the sixth most periodic gene according to DLxJTK.

It was not found in the list of the top 100 genes for DL. Meanwhile, according to DLxJTK, WC-1 was ranked #1649 and WC-2 was ranked #2888. Neither were found in the top 100 for either DL or JTK_CYCLE.

The normalized WC-1, WC-2, and FRQ data give the following graphs:

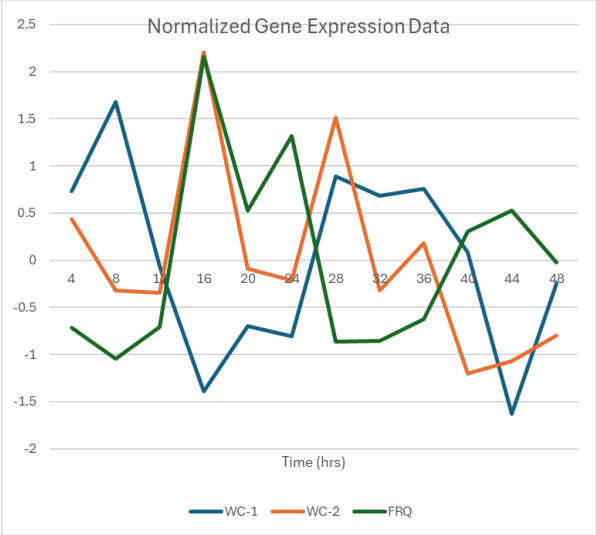

Figure 7. Normalized Gene Expression Data

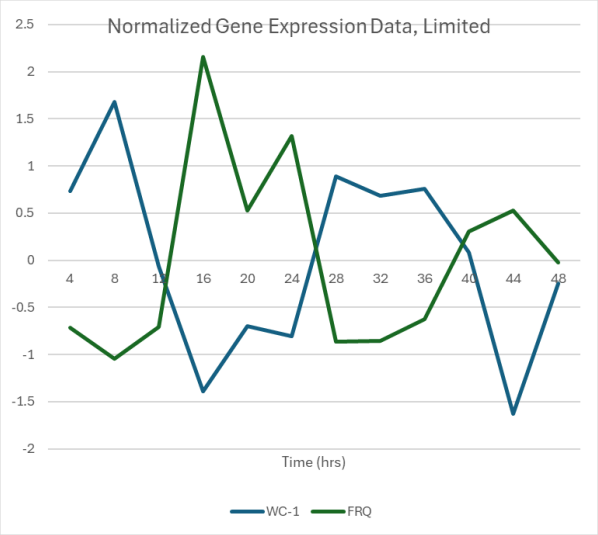

Figure 8. Normalized Gene Expression Data, WC-1 and FRQ

Our statistical examination focuses on the relationship between WC-1 and FRQ expression data. We made this decision based on both the hypothesized model and the obvious negative correlation seen between the two genes' normalized data, as shown in *Figure 8* and *Figure 9*.

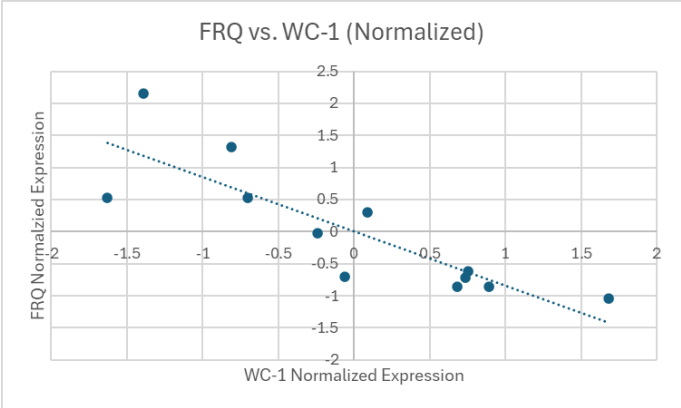

Figure 9. FRQ vs WC-1 (Normalized) Scatterplot

A full breakdown of the regression statistics can be found in *Appendix 1*. The relevant statistic to consider is the p-value for the slope, which was 0.0005. This emphasizes the overwhelmingly strong negative correlation between WC-1 and FRQ gene expression levels over the entire 48-hour observation period.

However, there was the concern that this data was cherry-picked, especially given the low periodicity score (outside of the top 20%) for WC-1 provided by the DLxJTK algorithm. To combat this concern, the nine most periodic (per DLxJTK) genes were additionally included in an alternate regression, these being: Ophio5|8288, Ophio5|7799, Ophio5|7767, Ophio5|4671, Ophio5|4554, Ophio5|3466, Ophio5|3328, Ophio5|2508, and Ophio5|1446. The full results of this regression can be found in *Appendix 2*. The most important statistic to note for this regression is its F-significance of 0.273. This is the equivalent to a p-score but for a multivariate regression. A score of 0.273 is very high, especially compared with the 0.0005 earned by the single-variable FRQ/WC-1 model.

This discrepancy between the expanded and original models is quantified by an F-test. This compares the two regressions to determine if there is statistical significance to the addition

of the increased number of variables to the expanded regression model. This test resulted in a p-value of 0.469, well above any reasonable threshold for statistical significance (i.e. 0.05). As before, this difference is especially stark when compared to the 0.0005 earned by the single-variable model that the expanded regression was compared against. Unfortunately, the structure of an F-test requires that the limited/original model be the null hypothesis; thus, strictly speaking, the F-test and its score of 0.469 only allow us to "fail to reject" the null hypothesis, and the test is unable to formally confirm it. However, this is strong evidence when combined with the individually strong score of the original regression, the further modeling described in this paper, and biological orthology with *N. crassa.*

Discussion

The results from the two models showed that a network in *O. cordyceps* similar to that in *N. crassa* is possible, although the evidence is still inconclusive. The small size of the proposed network makes it much simpler to analyze with both modeling tools and statistical analysis. However, this ignores the complexity of true biological systems. Additionally, the ODE model did not create any meaningful results. However, the statistical analysis of the raw data means that it is likely that there is some direct interaction between the genes studied.

Perhaps the biggest drawback of using LEM to create a network is that it assumes each node only has one edge regulating it. By creating a network that has two inputs to one node (both WC-1 and WC-2 activate FRQ), this limits the accuracy of LEM. This also may have impacted the accuracy of the ODE model. Because of this restriction in LEM, the input parameters for the ODE model were adjusted based on the assumption that no gene has more than two genes directly impacting it. This means the ODE model could not account for the entirety of the

proposed system, leading to the large discrepancy between the observed and predicted expression levels.

The results from the periodicity algorithms indeed support the claim that FRQ is a core node in the clock network. Since DLxJTK and JTK_CYCLE each place a slightly different weight on periodicity vs. regulation, the fact that FRQ was in the top 7 for both could be seen as a testament to its involvement in the network. It should also be noted that when comparing the top 100 genes for DLxJTK and JTK_CYCLE, the results were quite similar; while on the other hand, the results for DL were much different. Furthermore, it may be reasonable to say that not all relevant genes will be included in the top set of genes according to a periodicity algorithm. This is especially true when such genes act as a complex, as do WC-1 and WC-2. However, it is worth further consideration why FRQ and WC-1 ranked so differently in periodicity despite being nearly perfect mirror images of one another (as seen in *Figure 8*).

Future research could look more into other genes in the *N. crassa* network, such as VVD, FRH, CK-1, and CK-2. Also, due to the limitations of LEM, different edge-finding methods that account for genes being acted on by multiple other genes could remove much of the inconsistencies in the modeling data.

Citations

Bekker, C. de, Will, I., Hughes, D. P., Brachmann, A., & Merrow, M. (2017). Daily rhythms and enrichment patterns in the transcriptome of the behavior-manipulating parasite Ophiocordyceps kimflemingiae. *PLOS ONE*, *12*(11), e0187170. https://doi.org/10.1371/journal.pone.0187170

Bock, M., Scharp, T., Talnikar, C., & Klipp, E. (2014). BooleSim: An interactive Boolean network simulator. *Bioinformatics*, *30*(1), 131–132. https://doi.org/10.1093/bioinformatics/btt568

Deckard, A., Anafi, R. C., Hogenesch, J. B., Haase, S. B., & Harer, J. (2013). Design and analysis of large-scale biological rhythm studies: A comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics*, *29*(24), 3174–3180. https://doi.org/10.1093/bioinformatics/btt541

Hughes, D. P., Andersen, S. B., Hywel-Jones, N. L., Himaman, W., Billen, J., & Boomsma, J. J. (2011). Behavioral mechanisms and morphological symptoms of zombie ants dying from fungal infection. *BMC Ecology*, *11*(1), 13. https://doi.org/10.1186/1472-6785-11-13

McGoff, K. A., Guo, X., Deckard, A., Kelliher, C. M., Leman, A. R., Francey, L. J., Hogenesch, J. B., Haase, S. B., & Harer, J. L. (2016). The Local Edge Machine: Inference of dynamic models of gene regulation. *Genome Biology*, *17*(1), 214. https://doi.org/10.1186/s13059-016-1076-z

Motta, F. C., Moseley, R. C., Cummins, B., Deckard, A., & Haase, S. B. (2022). Conservation of dynamic characteristics of transcriptional regulatory elements in

periodic biological processes. *BMC Bioinformatics*, *23*(1), 94. https://doi.org/10.1186/s12859-022-04627-9

Pilorz, V., Helfrich-Förster, C., & Oster, H. (2018). The role of the circadian clock system in physiology. *Pflugers Archiv: European Journal of Physiology*, *470*(2), 227–239. https://doi.org/10.1007/s00424-017-2103-y

Appendix

Appendix 1: FRQ vs WC-1 Regression Statistics

FRQ vs. WC-1 (Normalized) SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.84773545
R Square	0.718655394
Adjusted R Square	0.690520933
Standard Error	0.556308428
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7.905209329	7.905209329	25.54359945	0.000496278
Residual	10	3.094790671	0.309479067		
Total	11	11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.85013E-16	0.16059241	4.26554E-15	1	-0.357822189	0.357822189	-0.357822189	0.357822189
X Variable 1	-0.84773545	0.167733302	-5.054067614	0.000496278	-1.221468537	-0.474002363	-1.221468537	-0.474002363

Appendix 2: Alternate Regression, using FRQ, WC-1, Ophio5|8288, Ophio5|7799, Ophio5|7767, Ophio5|4671, Ophio5|4554, Ophio5|3466, Ophio5|3328, Ophio5|2508, Ophio5|1446

Alternate Regression SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.993631781
R Square	0.987304116
Adjusted R Square	0.860345274
Standard Error	13.85336295
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	10	14924.4529	1492.44529	7.776568373	0.27264804
Residual	1	191.9156649	191.9156649		
Total	11	15116.36857			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	386.8066375	208.6812523	1.853576367	0.314963944	-2264.740079	3038.353354	-2264.740079	3038.353354
X Variable 1	-20.15966372	12.31702045	-1.636732179	0.34915361	-176.6622473	136.3429199	-176.6622473	136.3429199
X Variable 2	-1168.622945	779.7949452	-1.49862852	0.374602903	-11076.85717	8739.611281	-11076.85717	8739.611281
X Variable 3	-1.966781176	1.066040752	-1.844939953	0.316207933	-15.51211323	11.57855088	-15.51211323	11.57855088
X Variable 4	-0.530247734	1.604487513	-0.330477944	0.796804652	-20.91719457	19.8566991	-20.91719457	19.8566991
X Variable 5	2.493105492	1.473629564	1.691812891	0.339850729	-16.23113345	21.21734443	-16.23113345	21.21734443
X Variable 6	-21.76082016	14.27617612	-1.524275126	0.36963177	-203.1568368	159.6351964	-203.1568368	159.6351964
X Variable 7	155.2244856	85.85324175	1.808021251	0.321628663	-935.6443813	1246.093353	-935.6443813	1246.093353
X Variable 8	-0.594332638	1.406243045	-0.422638633	0.745435584	-18.46234467	17.2736794	-18.46234467	17.2736794
X Variable 9	112.8385335	102.0656849	1.105548193	0.46811383	-1184.028956	1409.706023	-1184.028956	1409.706023
X Variable 10	214.3230824	167.9623555	1.276018556	0.42317068	-1919.840995	2348.48716	-1919.840995	2348.48716