Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/45/20240587

Exploring the expression of *CG18269* under Zelda's regulation in early embryonic development of *Drosophila melanogaster*

Yiwi Chen

Shanghai High School International Division, Shanghai, 200231, China

2731291275@qq.com

Abstract. As a master regulatory protein, Zelda (Zld) plays a significant role in gene expression of Drosophila's embryonic development. It targets both genes, including *sna*, that are studied in previous researches, and genes that even their molecular function is still unknown, such as *CG18269*. Through examining both qualitative and quantitative data in this research, the absence of Zelda has proved to significantly down regulate the gene's expression. The gene is also observed to have multiple Zelda-binding sites, which may or may not have correlation with its neighboring genes. Future experiments, such as CRISPR-Cas9, may be conducted to further decipher the gene.

Keywords: Zelda, *CG18269*, *Drosophila melanogaster*, early embryonic development, gene regulation

1. Introduction

The study of genetics has long been a crucial focus in the study of biology. In early 21st century, a protein Zelda (Zld), short for Zinc-finger early Drosophila activator, is found [1]. Zelda is known for being a master regulatory protein that regulates the early embryonic development of Drosophila, activating both ubiquitous and patterning genes. Zelda also has a significant role in maintaining hotspots, or high occupancy transcription factor binding regions [2]. Without the presence of Zelda, its target genes may decrease in levels of expression or not express at all, affecting the embryo in multiple aspects. A fair amount of Zelda target genes had been studied and named, including *sna*, *brk*, *twi*, and many more, but a greater portion of these gene still remain mysterious to the world. Just like *CG18269*, a nameless gene being regulated by Zelda. It is a short gene with no introns and only one isoform, but its molecular function still remains unknown. Therefore, this research aims to dig into one of these unknown genes, *CG18269*, and study its relationship with Zelda as well as its neighboring genes. Hope this research can give insights to future studies on the gene's regulation, function, etc.

2. Materials & methods

2.1. Materials

No laboratory equipments were involved in this study. Software used for data collection and processing include Integrated Genome Browser, Snapgene, Flybase, and Chopchop.

DOI: 10.54254/2753-8818/45/20240587

2.2. In Situ Hybridization

Embryos are stained based on the gene sequence, allowing the expression to be visualized under a fluorescence microscope. This experiment for this study is performed by Rushlow lab.

2.3. Zelda & RNA Polymerase II ChIP

The Zld ChIP and PolII ChIP profile data in this study is credited to Nien et al. [2] and Blythe and Wieschaus [3] respectively.

2.4. Gene Region Analysis

The extended gene sequence of *CG18269* is downloaded from Flybase and loaded into Snapgene. The gene sequence is then cut to leave the gene itself and 1.5kb upstream of the gene, it is annotated based on the ChIP data in the Integrated Genome Browser.

2.5. CRISPR-Cas9

To observe the importance of Zelda to the gene, CRISPR-Cas9 is planned to perform to mutate the Zelda-binding sequences, CAGGTAG (into cTCAtag) and CAGGTAA (into cTCAtaa), upstream of *CG18269*. The guide RNAs and primers are all designed with the assistance of Chopchop and Snapgene. Their exact sequences are shown below in Tables 1 and 2 (Zelda-targeting sites underlined in orange), and their positions are shown in Figure 1.

Table 1. Designed Sequences for Mutating CAGGTAG

Guide RNA	5’ CGACAGGTAGTCGATGTGGGTGG 3’
Primer (F)	5’ TATAGTGCGACTCATAGTCGATGTGGG 3’
Primer (R)	5’ CCCACATCGACTATGAGTCGCACTATA 3’
Homology arms + cTCAtag	5’ AATTCGTGGACTGAGCCATGGCTATCGGCGGTTGGAATGTGA GTGATAATGATGAGGTTGATTTGGTGGGTTTATATAGTGCGACT CATAGTCGATGTGGGTGGTGCTCTGACCTGTTTGTGATTATGAG CAGGTCAAAGGCTGGGTTTTTTCGAAGCTTTAAATTCTTATGGC CGATTTTATACCTGCGATGCGACAGA 3’

Table 2. Designed Sequences for Mutating CAGGTAA

Guide RNA	5’ ATAGGGGCTTTTACCTGGCGAGG 3’
Primer (F)	5’ AAGTCGGACTCATAAAGCTCCTCGCCTCATAAAAGCCCCT 3’
Primer (R)	5’ AGGGGCTTTTATGAGGCGAGGAGCTTTATGAGTCCGACTT 3’
Homology arms + cTCAtaa	5’ TTAATTTTAGTTTGTGGCTTGTAGCATTTATTATAACTTTGACA ATTTTGTGTTAATCCGTACCAGGTGTTTGGTTCGGGCCATAAAC AGCGACGCATTTGGGCCCAGCAAGTCGGACTCATAAAGCTCC TCGCCTCATAAAAGCCCCTATTGCAGAGACCCTTGTTTTTGGT GTGCCATATAAGACAGAAGGATGAGAGATTGCAGCTGTCAGAG AGAATCGAGTCTTCAGATCG 3’

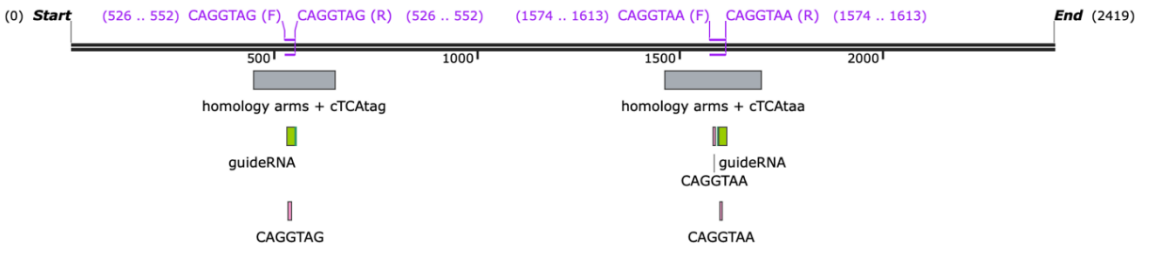

Figure 1. Positions of GuideRNAs and Primers for CRISPR-Cas9

DOI: 10.54254/2753-8818/45/20240587

Homology arms marks in grey, guide RNAs marked with green, Zelda binding sites marked with pink, and primers marked with purple.

3. Results

3.1. Expression of CG18269 in Early Embryos

The Drosophila embryos' expression of gene *CG18269* in both wt and Zld$^-$ (experiment performed by Rushlow lab) is shown in Figure 2. A younger wild type embryo in pre-blastoderm stage (Figure 2A) demonstrates high levels of *CG18269* expression all throughout the embryo, with slightly less expression in its anterior and posterior ends. Later when the embryo steps into cellularization (Figure 2B), it shows largely decreased levels of gene expression that concentrates in the center of the embryo. In comparison, the Drosophila embryos with Zelda knocked out (Figure 2C&2D) demonstrated barely any signs of gene expression during cellularization.

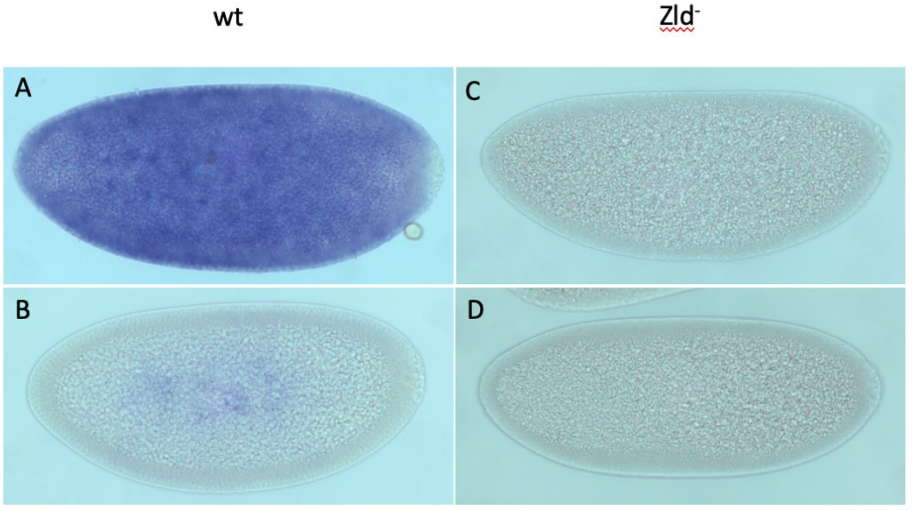

Figure 2. Drosophila Embryos Expressing Gene *CG18269*

(A)(B) wildtype embryos
(C)(D) embryos with Zelda knocked out

3.2. Zelda & RNA Polymerase II ChIP Profile

Figure 3 illustrates the Zelda ChIP [2] and RNA polymerase binding ChIP data [3] using Integrated Genome Browser. The Zld ChIP data shows three Zelda peaks (marked by orange lines in Figure 3A). Two peaks merge into a larger peak right upstream of *CG14014*, a neighbor of *CG18269*, whereas the other peak exist in the untranslated region of *CG18269*. Furthermore, similar peaks are observed in the RNA polymerase binding profile, but instead of decreasing gradually like Zld ChIP, the degree of RNA polymerase binding remains high, extending all the way into its neighboring genes (for example, *CG14013*). In contrast, the ChIP profile is greatly decreased when Zelda is knocked out, though short regions of slightly increased expression is still observed at approximately the beginning of translation.

DOI: 10.54254/2753-8818/45/20240587

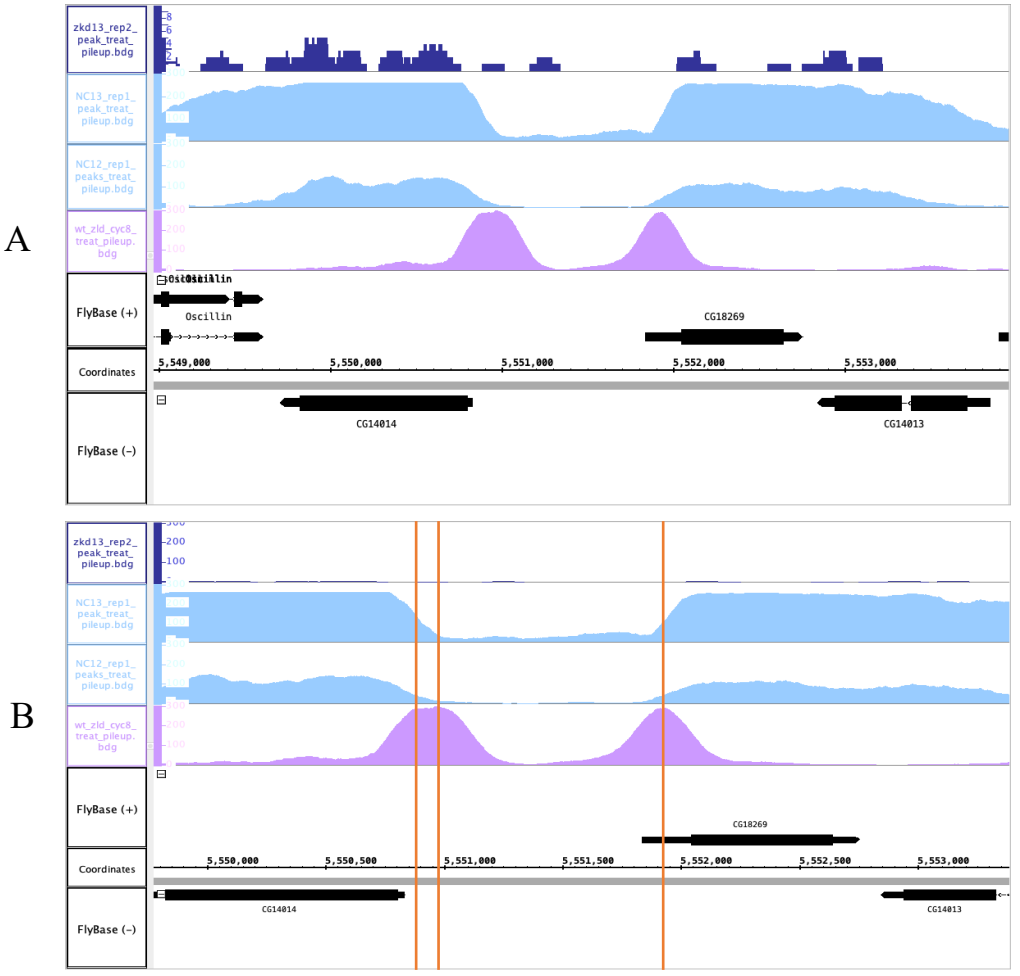

Figure 3. Zld & PolII ChIP Profile

Broad view (B) Close up view

Lane 1 (dark blue) represents PolII ChIP data without Zld, lane 2 & 3 (light blue) represents PolII ChIP data in wildtype embryos in nc13 and 12 respectively, and lane 4 (purple) represents Zld ChIP data in a wild type embryo.

Figure 3B offers a close up view in the expression data when Zelda is knocked out. As shown, short, segmented regions of expression exist in the translation region of both genes, but the peaks are relatively higher and more continuous in *CG14014* compared to *CG18269*.

3.3. Gene Region Analysis

The gene region analysis was done based on the results from the Zld and PolII ChIP data and Flybase. As shown in figure 4, *CG18269* has a longer 5' UTR while a total of 16 TATAA sites and 5 CAGGTA sites exist within the gene sequence and 1500bp upstream of the gene. One CAGGTAG site and two CAGGTAT sites occur closer upstream of *CG14014* while the other two CAGGTAA sites are present in the 5' UTR of *CG18269*. Within the 16 TATAAs, it is observed that 11 exist upstream of *CG18269*, 1 within the gene sequence, and 4 downstream of the coding region. There are also three sites with TATAA on both strands at the same time, labeled TTATAA with orange in the figure, and one exists right before +1.

DOI: 10.54254/2753-8818/45/20240587

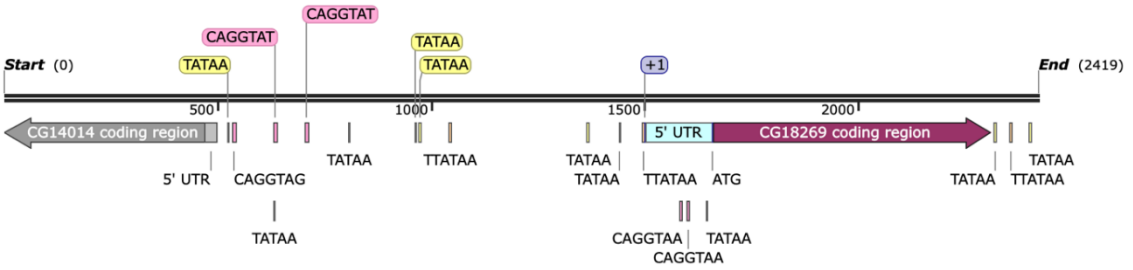

Figure 4. Different Sequences Identified Within and Up/downstream of the Gene

Gene *CG14014* marked in grey, Zld binding sites marked in pink, TATAA sites marked in yellow (TTATAA in orange), 5' UTR marked in light blue, and coding region marked in purple.

3.4. High-throughput Expression Data

As the published high-throughput expression data from Flybase [4] (Figure 5) shows, the expression of gene *CG18269* largely occurs within 0 to 4 hours of embryonic development. Embryos aged 6 to 8 hours also demonstrates a fair amount of expression, though much less than before, but little to no expression is shown in older stages.

Besides the timing of expression, the graph also shows that the start of high level expression occurs shortly downstream of 5,552,000bp (marked with red line) and extends all the way to the end of the gene. A little expression is also shown about 50bp upstream of the red line.

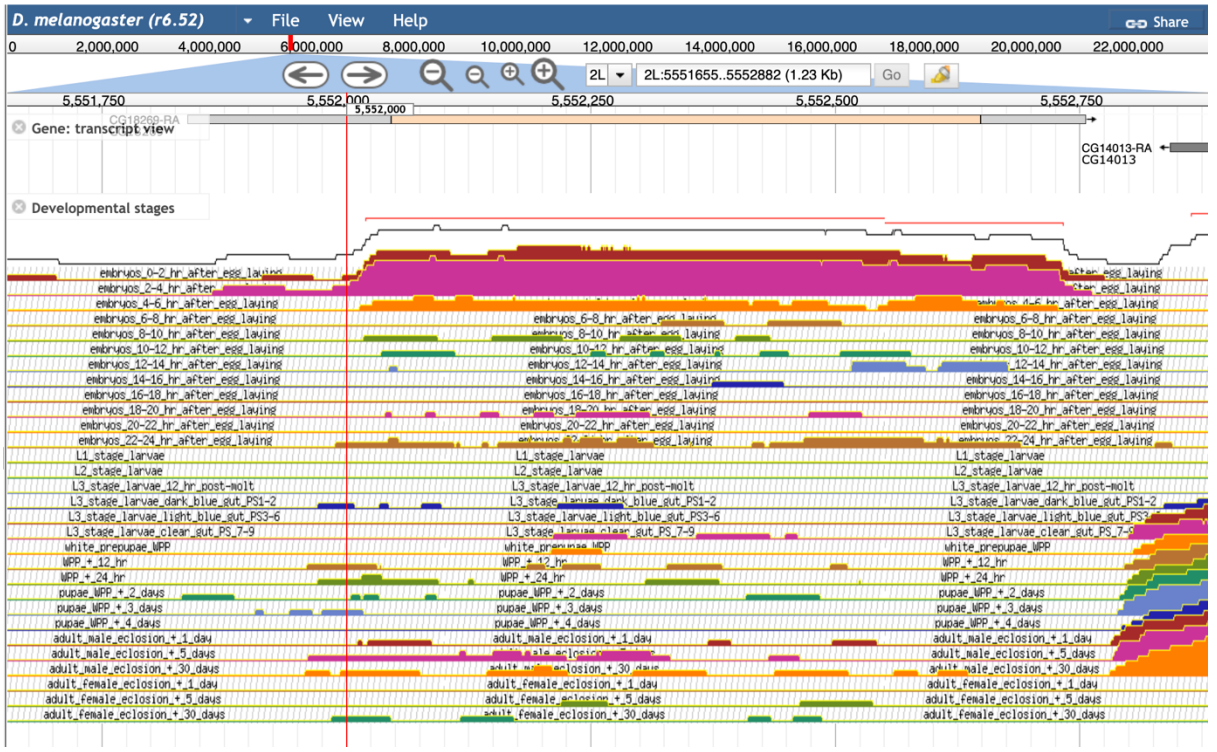

Figure 5. Screenshot of High-Throughput Expression Data from FlyBase

DOI: 10.54254/2753-8818/45/20240587

3.5. Revised Gene Region Analysis

Revising the various regions of *CG18269* using the published high-throughput expression data from Flybase and getting rid of the unused TATAAs obtains figure 6. Now *CG18269* has a shorter 5' UTR that is about the same length as that of *CG14014*. The TATAA is 30bp upstream of the gene, and the CAGGTAA sites exist at approximately another 50bp upstream.

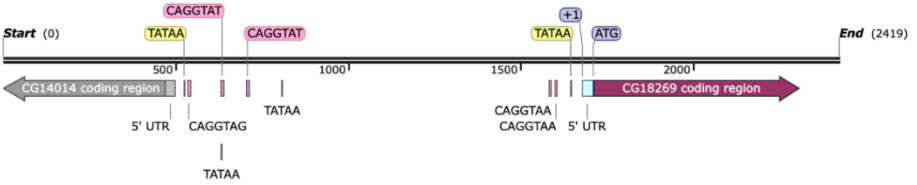

Figure 6. Identified Sequences Revised (All color codings remain the same as Figure 4)

4. Discussion

Nien et al. stated that Zelda levels increase significantly during the second hour of development but can already be detected in nc 2. Without Zld, its target genes' expression may be delayed or absent [2]. *CG18269* seems to be the latter situation, for the Zld^- embryo is in cellularization (third hour of development) but no signs of Zld is observed. This highlights Zld's significance to Drosophila as a master activator while also suggests that *CG18269*'s function has nothing to do with cellularization and that its function is not as essential that it prevents the embryo from cellularizing. Furthermore, the wildtype embryos hint the gene's function may be relevant with the center area of the embryo, but further experiments and observations would be needed to find out the specific function, such as observing the embryos again after they move into adult stage.

Drosophila's early development is characterized by 13 nuclear cycles, which are rapid cycles that skips gap phases [5]. Looking at the primary gene region analysis, it seems unusual for *CG18269*, as a short gene, to have a 5' UTR of this size, for a longer 5' UTR will only decrease the replication and transcription efficiency, which violates Drosophila's ideals for rapid cycles in early development. The positions of Zelda binding sites also seems odd—they either are very far away from the gene, almost reaching the previous gene, or it appears within the 5' UTR of *CG18269*. Thus, it is hypothesized that Flybase has been wrong for the gene regions on *CG18269*, for it is only a prediction and no one has ever closely studied the gene before. Then looking at the high-throughput expression data, the high expression area represents the position of the gene while the little expression upstream likely represents where Zld binds. The data shows that the gene starts after 5,552,000bp, but the same position appears to be in the 5' UTR looking at the Flybase (+) lane in the Integrated Genome Browser (Figure 3)—another evidence that FlyBase is flawed.

Taking the CAGGTAAs as enhancer, located the closest TATAA downstream and identified it as -30. Using this method, a new +1 is located, and the closest ATG found downstream marks the start of the coding region. In this revised gene analysis (Figure 6), *CG18269* has a much shorter 5' UTR, which is about the same size as its neighboring gene, *CG14014*.

However, the work of Pfeiffer et al. [6] proved the efficiency of translational enhancers located in untranslated regions. Even though their work is based on combined elements in both the 5' UTR and 3' UTR, it still opens up the possibility that the Zld-binding site may exist in the 5' UTR and the original gene region analysis turns out to be correct.

Besides the gene regions, the positions of Zld-binding sites and TATAAs are also worth noting. As previously mentioned, the gene has 16 TATAA sites and 3 CAGGTA sites (both CAGGTAA sites are regarded as one site, same goes for CAGGTAT). The CAGGTAG site, which is the most ideal binding sequence [2], exists all the way upstream to CG14014. In fact, it appears to be the Zld-binding site for CG14014. However, it is also reasonable to hypothesize these to be the enhancer sequences for

DOI: 10.54254/2753-8818/45/20240587

CG18269, for the research of Nien et al. concluded 83% enhancer sites to be within 2kb upstream of the gene [2]. Even the furthest site, CAGGTAG, is only about 1kb apart from the gene.

Furthermore, previous research from Herr and Harris [7] identified a sufficient amount of genes share a common upstream region in a head-to-head orientation. This orientation is especially favored in Drosophila, and the genes seem to have greater relevance with each other compared to other orientations, including correlated expression patterns. The genes *CG18269* and *CG14014* happens to be in this head-to-head orientation, meaning they share the same upstream sequences. Considering the short distance in between the neighbors (about 1kb), it can be hypothesized that the genes share Zld-binding sites and are co-regulated by the sites. Thus, a CRISPR-Cas9 can be designed to further test out the genes' correlation in the future. The CAGGTAG and/or CAGGTAAs can be knocked out, and the changes in expression for both genes can be observed with in situ hybridization. The CAGGTAT is ignored in this experiment because the work of Nien et al. [2] showed that it is a much weaker binding site than CAGGTAG or CAGGTAA. But since two peaks are observed in the Zld ChIP data of CG14014 (Figure 3), CAGGTAT seems to also have contributed to the expression though the degree of contribution in questionable considering its weak binding ability. Nevertheless, CAGGTAT can also join the mutation experiment, testing out the effects of each Zld-binding site to both genes. Just that the work load would greatly increase.

The existence of the 16 TATAAs also seems peculiar to the gene, as *CG14014* does not present this phenomenon. This may simply be the result of probability since it is much more likely to have 5 exact nucleotides (1/32) than 7 (1/128). But since TATA box as the promoter is always located downstream of enhancers, the TATAAs may exist in multiple locations to make sure transcription could go both directions for both genes. However, they seem to locate in no observable pattern, which give rise to another question: do the distance between enhancers and promoters matter? The question may be tested out in the future through inserting or deleting sequences in between the Zld-binding site and TATAA, but it's best to be done on another gene with less numbers of such sites. The abundant TATAAs of *CG18269* can make it hard to eliminate outside effects. Furthermore, the three TTATAAs can potentially be extra-strong binding sites by having TATAA on both strands, but so far they don't display any function to the gene.

Now switching lenses to *CG18269*'s other neighbor, *CG14013*. As shown in Figure 3B, this is a slightly more complex gene that contains an intron, and has a tail-to-tail orientation with *CG18269*. The expression of *CG18269* in PolII ChIP extends all the way into *CG14013*, suggesting a possibility of some relationship in between the genes. However, considering Herr and Harris's research stating that genes of tail-to-tail orientations are less likely to have correlations, it may also simply be incorrect data. To test out experimentally, a probe can also be made for *CG14013* and see whether the gene would be downregulated as Zld-binding sites are knocked out.

5. Conclusion

This research achieves three main conclusions. It offers a revised labeled gene sequence of the gene *CG18269*, identifying a shorter, more reasonable 5' UTR as well as 5 CAGGTAA sites and 16 TATAA sites within the gene sequence and 1500bp upstream. It hypothesizes the gene function to not be relevant to significant processes such as cellularization, and that it may function on the center area of a drosophila embryo. It also highlights the correlation the gene has with its neighbor *CG14014*, and suggests a possibility of co-regulation among the two genes. However, all of these above are only hypotheses done based on research and some experimental data. Further lab work must be carried out to verify the conclusions in advance.

References

[1] Liang, H. L., Nien, C. Y., Liu, H. Y., Metzstein, M. M., Kirov, N., & Rushlow, C. (2008). The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, *456*(7220), 400–403.

DOI: 10.54254/2753-8818/45/20240587

[2] Nien C-Y, Liang H-L, Butcher S, Sun Y, Fu S, et al. (2011). Temporal Coordination of Gene Networks by Zelda in the Early Drosophila Embryo. *PLoS Genet*, 7(10): e1002339.

[3] Blythe, S. A., & Wieschaus, E. F. (2015). Zygotic Genome Activation Triggers the DNA Replication Checkpoint at the Midblastula Transition. *Cell,* 160(6), 1169–1181.

[4] FlyBase. (n.d.). *FlyBase Gene Report: DMEL\CG18269*. http://flybase.org/reports/FBgn0031719

[5] Harrison, M. M., & Eisen, M. B. (2015). Transcriptional Activation of the Zygotic Genome in Drosophila. *Current topics in developmental biology*, *113*, 85–112.

[6] Pfeiffer, B. D., Truman, J. W., & Rubin, G. M. (2012). Using translational enhancers to increase transgene expression in Drosophila. *Proceedings of the National Academy of Sciences, 109(17), 6626–6631.*

[7] Herr, D. R., & Harris, G. L. (2004). Close head-to-head juxtaposition of genes favors their coordinate regulation in Drosophila melanogaster. *FEBS letters*, *572*(1-3), 147–153.