A computational approach to investigate the role of zinc finger protein, ZF30C, in epigenetic repression of *Drosophila engrailed* gene

Jessica Escobar, Marharyta Petukh and Payal Ray
Department of Biology, Presbyterian College, Clinton, SC

SCINBRE South Carolina IDeA Networks of Biomedical Research Excellence

INTRODUCTION

- Polycomb group proteins (PcG) are conserved epigenetic regulators that serve to sustain the transcriptional repression of target genes. They are able to form. multiprotein complexes known as Polycomb Repressive Complexes 1 and 2 (PRC 1 and PRC2)
- Polycomb Repressive Elements (PREs) are DNA elements necessary to act as binding sites for *Drosophila melanogaster* PcG proteins to regulate target genes. DNA elements recruit specific sets of DNA-binding proteins which then bind to PRCs via protein-protein interaction (PPI)
- Studies have shown important roles of PcGs in neurogenesis and brain development. Mutations in PcGs have also been linked to various forms of cancer

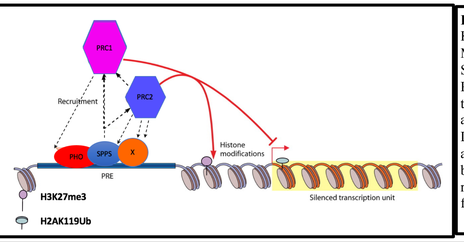

Figure 1: Adapted from Kasis and Kennison, Molecular Cell Biology 2010. Schematic representation of PRC1 and PRC2 recruited to the PREs by cooperative action of the PhoRC and DNA-binding proteins. The as yet unidentified DNA binding proteins are represented by 'X' on the figure.

- Drosophila *engrailed* gene expression occurs in the posterior part of wing disks and is silenced in the anterior part of wing disks by PcG repression of Histone 3 at lysine 27
- *Engrailed* contains PREs with DNA elements which bind to DNA-binding proteins that in turn interact with PcG proteins, some are yet to be identified
- A DNA-IP coupled mass spectrometric screen identified Zf30C as a candidate protein. Same screen identified, Combgap – a bonafide PRE binding protein
- ZF30C contains 13 zinc fingers and was previous implicated in controlling receptor gene transcription in sensory neurons and wing pattern formation
- ZF30C contains 777 amino acids and is a C2H2 zinc finger protein, meaning it contains the consensus sequence (F/Y)-X-C-X2-5-C-X3-(F/Y)-X5-Ψ-X2-H-X3-4-H, where X is any residue and Ψ is any hydrophobic residue. C2H2 zinc finger proteins also obtain their name from the binding of a zinc ion to cysteine and histidine residues

AIM OF THIS STUDY: The aim of this study is to carry out an evolutionary analysis to identify key residues and regions of ZF30C that may be important in protein binding, build a 3D model of ZF30C and compare to other known proteins with similar structures that may be involved in DNA binding, and using an artificially modeled enzyme to view how Zf30C binds to DNA or other proteins.

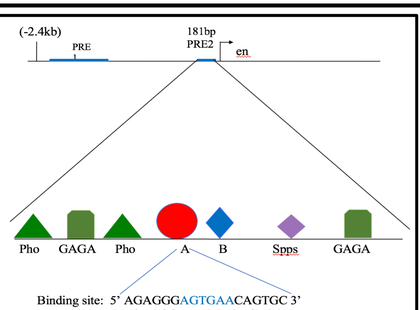

Figure 2: The upstream region of engrailed contains two PREs. When the binding site was previously mutated during prior experimentation, the PRE was no longer functional. The magnified region was used as a bait to identify proteins that bind to Site A, of en PRE in the DNA-IP coupled Mass Spectrometric analysis . A mutated oligo, as shown on the bottom was used as a control. Candidate proteins were identified as those binding specifically to the wild type oligo (AGTGAA).

METHODS

Evolutionary Analysis

- Identifying key residues or regions of the protein that are conserved across species and therefore, may be critical for DNA or protein binding
- The analysis was carried out using Blast of ZF30C and its N and C terminus (N: 1-400, C: 401-777). This split was made because the identified conserved domains in NCBI revealed the region of most conservation to be around residue 400

Structural Analysis

Building 3D models of ZF30C and Cg and comparing these models to other known protein models with similar structures to identify residues involved in DNA binding

- Using I-Tasser, a server that makes protein structures and function predictions, we made the structure of ZF30C using its fasfa sequence

Molecular Dynamic Simulation

Using an artificially modeled system and monitoring the change of atom location to see how ZF30C attaches to DNA or other proteins

- The crystallized structure of 2KMK was used for the Molecular Dynamic (MD) Simulation by substituting its DNA and protein structures with ZF30C's structures from the modeling along with inserting the sequence AGTGAA into the DNA helix
- MD Simulation monitors the change of location of all atoms and binding activity through Root Mean Square Deviation and Root Mean Square Fluctuation
- The goal of the MD Simulation is to figure how much confidence there is that ZF30C is being bound to the DNA sequence AGTGAA

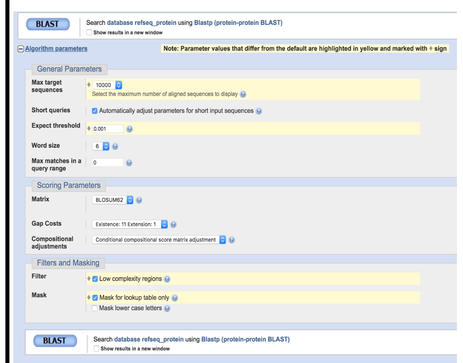

Figure 3: The Blast parameters for the evolutionary analysis.

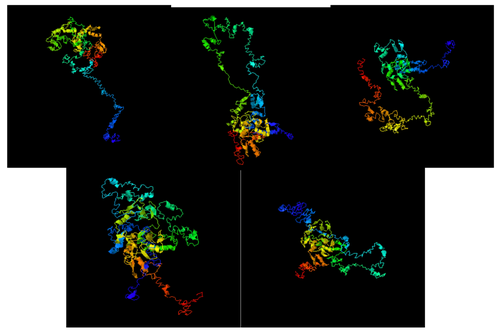

Figure 4: Five models of ZF30C made by threading through I-Tasser. Model 4 (bottom left) was chosen because it had the best match and lowest RMSD.

RESULTS

- **Evolutionary Analysis:** The analogs from the Blast after filtering were assessed and we realized that the "hits" left were comparable to ZF30C on the basis of containing zinc finger domains rather than having a specific comparable sequence, meaning the hits were not highly conserved. While the full length zinc finger and N terminus had a minimal amount of analogs and all belonged to the Drosophila genus (10-70), we did see that the C terminus had nearly 2,000 analogs from a variety of genera, thus indicating that C terminus may be partly conserved in the metazoans.
- **Structural Analysis:** Given that we did not find any homologous sequences or structures from the evolutionary analysis, we used I-Tasser as opposed to Swiss Model or Phyre 2 (homology-based modeling) because it does "threading" which looks at regions of a sequence to compare. A Blast was also run using the PDB setting to analyze proteins of similar structure to ZF30C. The protein "Gfi-1 Zinc Fingers 3-5 complexed with DNA" came up as a hit with 14% query coverage; however, a 40% percent identity. "Gfi-1 Zinc Fingers 3-5 complexed with DNA" is a transcriptional repressor involved in the differentiation of blood cell formation. It also binds to DNA using C2H2 zinc fingers. Most interestingly, the protein binds the DNA sequence AGTGAT which is only one base different than AGTGAA, the region of enPRE that pulled out Zf30C.
- **Molecular Dynamic Simulation:** Three helices interacted with DNA: Helix 1: amino acids 670-680 (containing a zinc ion), Helix 2: amino acids 585-596 (containing zinc ion), Helix 3: amino acids 641-651 (containing zinc ion), Helix 4: amino acids 557-567 (no zinc ion). The incorporation of ZF30C's characteristics involved adding zinc ions in order to keep the model intact and the N and C terminus.

CONCLUSIONS

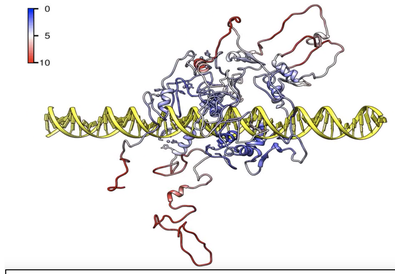

Figure 5: Simulation of Zf30C binding to AGTGAA sequence in the presence of Zn ions. The simulation was run for 40 nanoseconds and revealed regions of ZF30C stably bound to DNA. The model shows an interaction of 4 Zinc finger between amino acids 557-680 with the DNA sequence.

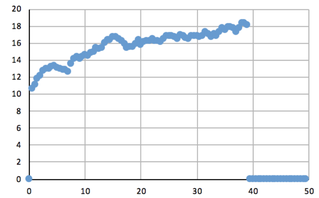

Figure 6 Graph showing RMSD (Root Mena Square Deviation), which reveals the model's deviation from the initial position. Increased peaks show increased deviation. The graph eventually plateaus to show when the model is most stable and has equilibrated. This equilibrated region reveals when the model is most stable.

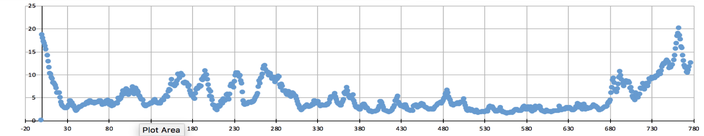

Figure 7: Graph showing RMSF (Root Mean Square Fluctuation) for Zf30C structure. RMSF measure the position change of each residue in the sequence. The region with little fluctuation is where DNA binding most likely occurs, whereas the fluctuating regions is where Protein Protein Interaction may occur. From this data, it is likely that the protein binds to the DNA between residues 530-680 and may be involved in PPI in the other fluctuating regions.

Limitations: I-Tasser and other tools such Swiss Model or Phyre 2 generate five models for each FASTA sequence entered. For the MD simulation, it was up to us to pick the best structure of ZF30C out of these five models. Additionally, the MD simulation works best on smaller ligands as there is less room for error when building complex models or interactions. ZF30C is a large ligand with 777 residues, meaning the simulation is less reliable. Gfi-1 Zinc Fingers 3-5 complexed with DNA (2KMK) only had a 14% query coverage and 40% percent identity in comparison to the structure of ZF30C.

FUTURE DIRECTIONS:

1. Based on structural modeling, it is highly likely that amino acids 557-680 interact with enPRE, Site A. Biochemical analysis will be performed via gel shift assay to validate this direct interaction using purified Zf30C protein.
2. Computational analysis of ZF30C interaction with Combgap and other PRE DNA binding proteins
3. Presence of Zf30C at PREs will be tested using polytene staining in conjunction with validated PRE DNA binding proteins such as Pho and Combgap.

REFERENCES

. Kassis, J. A., Kennison, J. A., & Tamkun, J. W. (2017). Polycomb and Trithorax Group Genes inDrosophila. Genetics, 206(4), 1699–1725.
2. Kassis, J. A., & Brown, J. L. (2013). Polycomb group response elements in Drosophila andvertebrates. Advances in genetics, 81, 83–118.
3. Moazed D, O'Farrell PH. (1992) Maintenance of the engrailed expression pattern by Polycombgroup genes in Drosophila. Development. 116(3):805-10.
4. Americo J, Whiteley M, Brown JL, Fujioka M, Jaynes JB, Kassis JA. (2002). A complex array ofDNA-binding proteins required for pairing-sensitive silencing by a Polycomb group response elementfrom the Drosophila engrailed gene. Genetics,160:1561–1571.
5. Ray, P., De, S., Mitra, A., Bezstarosti, K., Demmers, J. A., Pfeifer, K., & Kassis, J. A. (2016).Combgap contributes to recruitment of Polycomb group proteins in Drosophila. Proceedings of theNational Academy of Sciences of the United States of America, 113(14), 3826–3831.
6. Campbell, G. L., & Tomlinson, A. (2000). Transcriptional regulation of the Hedgehog effector Ci bythe zinc-finger gene combgap. Development (Cambridge, England), 127(19), 4095–4103.
7. Svendsen, P. C., Marshall, S. D., Kyba, M., & Brook, W. J. (2000). The combgap locus encodes azinc-finger protein that regulates cubitus interruptus during limb development in Drosophilamelanogaster. Development (Cambridge, England), 127(19), 4083–4093.
8. Schertel, C., Albarca, M., Rockel-Bauer, C., Kelley, N. W., Bischof, J., Hens, K., ... Deplancke, B.(2015). A large-scale, in vivo transcription factor screen defines bivalent chromatin as a key propertyof regulatory factors mediating Drosophila wing development. Genome research, 25(4), 514–523.
9. Jafari, S., Alkhori, L., Schleiffer, A., Brochtrup, A., Hummel, T., & Alenius, M. (2012). Combinatorialactivation and repression by seven transcription factors specify Drosophila odorant receptorexpression. PLoS biology, 10(3), e1001280.
10. Chan, H. L., & Morey, L. (2019). Emerging Roles for Polycomb-Group Proteins in Stem Cells andCancer. Trends in biochemical sciences, 44(8), 688–700.
11. Albert, M., & Huttner, W. B. (2018). Epigenetic and Transcriptional Pre-patterning-An EmergingTheme in Cortical Neurogenesis. Frontiers in neuroscience, 12, 359.
12. Schuettengruber, B., Bourbon, H. M., Di Croce, L., & Cavalli, G. (2017). Genome Regulation byPolycomb and Trithorax: 70 Years and Counting. Cell, 171(1), 34–57.
13. Schuettengruber, B., Oded Elkayam, N., Sexton, T., Entrevan, M., Stern, S., Thomas, A., ... Cavalli,G. (2014). Cooperativity, specificity, and evolutionary stability of Polycomb targeting inDrosophila. Cell reports, 9(1), 219–233.
14. Celniker, S.E.,Dillon, L.A.,Gerstein,M. B.,Gunsalus, K. C., Henikoff, S., Karpen, G. H., ...modENCODE Consortium (2009). Unlocking the secrets of the genome. Nature, 459(7249), 927–930.
15. Kahn, T. G., Stenberg, P., Pirrotta, V., & Schwartz, Y. B. (2014). Combinatorial interactions arerequired for the efficient recruitment of pho repressive complex (PhoRC) to polycomb responseelements. PLoS genetics, 10(7), e1004495.
16. Brown, J. L., Sun, M. A., & Kassis, J. A. (2018). Global changes of H3K27me3 domains andPolycomb group protein distribution in the absence of recruiters Spps or Pho. Proceedings of theNational Academy of Sciences of the United States of America, 115(8), E1839–E1848.
17. Laugesen, A., Højfeldt, J. W., & Helin, K. (2019). Molecular Mechanisms Directing PRC2Recruitment and H3K27 Methylation. Molecular cell, 74(1), 8–18.
18. Brown, J. L., & Kassis, J. A. (2013). Architectural and functional diversity of polycomb groupresponse elements in Drosophila. Genetics, 195(2), 407–419.
19. Orsi GA, et al. High-resolution mapping defines the cooperative architecture of Polycomb responseelements. (2014). Genome Res, 24(5):809–820.
20. Yang Zhang. I-TASSER: Fully automated protein structure prediction in CASP8. Proteins, 77 (Suppl 9): 100-113, 2009.
21. Ambrish Roy, Jianyi Yang, Yang Zhang. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Research, 40: W471-W477, 2012.
22. Jianyi Yang, Yang Zhang. I-TASSER server: new development for protein structure and function predictions, Nucleic Acids Research, 43: W174-W181, 2015.

ACKNOWLEDGEMENTS

We acknowledge funding and support from the SC-INBRE program for this research. We also thank Kobie Kirven for helping us with this project at the initial steps. Authors declare no competing financi interest.