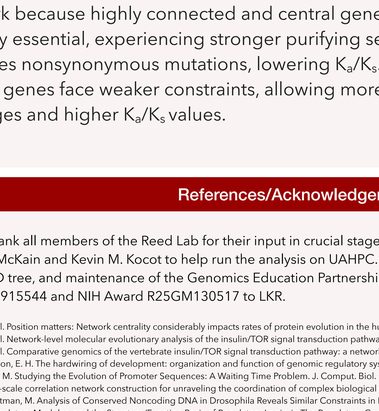
Network Topology is Correlated with Evolutionary Rates of Genes and their Regulatory Regions in *Drosophila*

Chinmay P. Rele & Laura K. Reed
Department of Biological Sciences, The University of Alabama, Tuscaloosa AL 35401

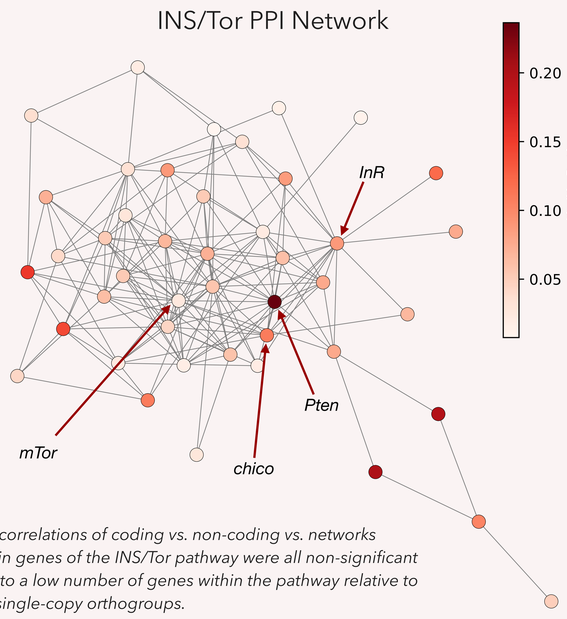
Introduction

- Important genes are often highly conserved.
- Gene regulation may also be conserved.
- We studied 8,473 single copy orthologous genes and their regulatory regions in 10 *Drosophila* species, focusing on the Insulin Signaling (**INS/Tor**) pathway[3].
- OrthoFinder[9] was used to identify single-copy orthogroups (**OGs**).
- We found correlation of the amount of conservation between genes, regulatory regions, and network architecture.
- Genes and their regulatory regions have correlated rates of evolution.
- This study helps us understand genetic and regulatory conservation and how they shape gene function across species[4-6,13].

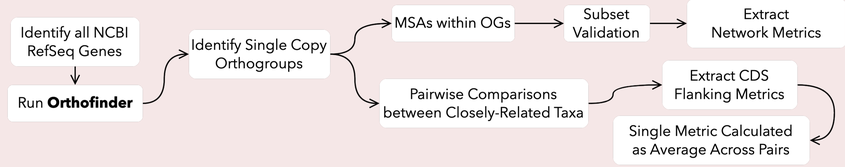


Five species pairs were identified for **P**hylogenetically **I**ndependent **C**ontrast

Method

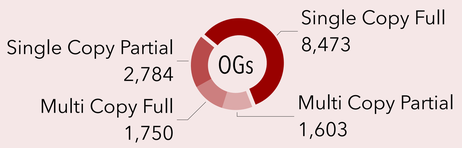


The correlations of coding vs. non-coding vs. networks within genes of the INS/Tor pathway were all non-significant due to a low number of genes within the pathway relative to the single-copy orthogroups.

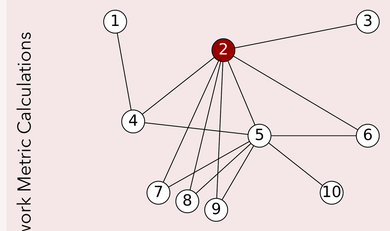


General Orthology Inference and Metrics Extraction

Normalizing Hamming Divergence Scores



OrthoFinder Called OrthoGroup Sets



Degree centrality for node **2** would be calculated by comparing how many edges it has relative to the total number of edges in the network ($\frac{7}{16} = 0.4375$).

Closeness centrality for node **2** is calculated as the reciprocal of the average length of the shortest paths from node 2 to all other nodes in the graph.

- Removing 5%, 10%, and 25% of the lower confidence edges did not greatly affect the distribution of network centralities.

Combine DroID PPI Networks → Assign Confidence Values to Each Edge → Omit 5% of the Low Confidence Nodes in the Combined Network

Identifying Network Metrics

Conclusions

More conserved genes tend to have more conserved regulatory regions and also tend to be more centralized in networks.

Future Directions

1. Run MEME on the flanking regions to identify putative regulatory elements.
2. Run synteny-based orthology inference in addition to OrthoFinder.

Results

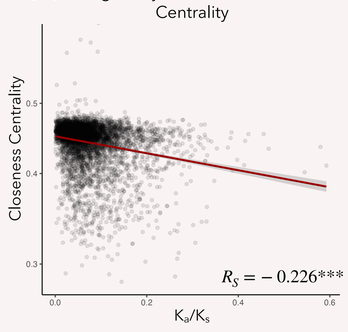
$p^{***} < 10^{-10}$ $p^{**} < 10^{-3}$ $p^{*} < 0.05$ p^{NS} : Not Significant

Network vs. Coding Sequence

Hypothesis

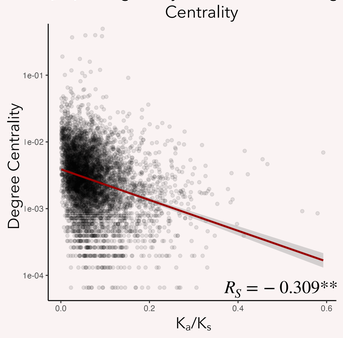
As the coding sequence of the gene is more conserved, the node will be more connected to the rest of the network (either directly or indirectly)[1,2].

K_a/K_s is Negatively Correlated with Closeness Centrality



Closeness centrality measures how close a node is to all other nodes in a network. **High** values indicate more direct access to others, while **low** values suggest more distant relationships in the network.

K_a/K_s is Negatively Correlated with Degree Centrality



Degree centrality measures a node's number of direct connections in a network. **High** values indicate many connections, while **low** values mean fewer connections within the network.

K_a/K_s negatively correlates with degree and closeness centrality in a PPI network because highly connected and central genes are more functionally essential, experiencing stronger purifying selection. This reduces nonsynonymous mutations, lowering K_a/K_s. In contrast, peripheral genes face weaker constraints, allowing more amino acid changes and higher K_a/K_s values.

$\frac{K_a}{K_s}$ measures selection on protein-coding genes.

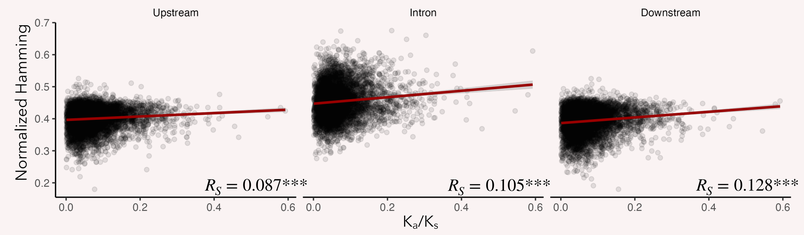
$\frac{K_a}{K_s} \begin{cases} > 1; \text{ positive/adaptive selection} \\ \sim 1; \text{ neutral selection} \\ < 1; \text{ negative/purifying selection} \end{cases}$

Coding Sequence vs. Flanking Region

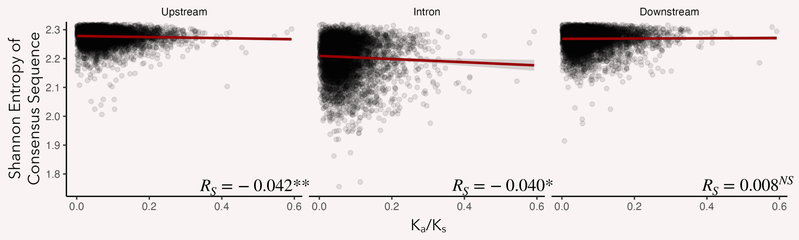
Hypothesis

When the coding sequence of the gene is more conserved, the regulatory regions of the gene will also be more conserved.

Protein and Regulatory Region Evolution is Positively Correlated



Sequence Entropy is Negatively Correlated with Ka/Ks of Upstream and Intron regions, but not Downstream



Hamming scores of regulatory regions correlate with K_a/K_s because they evolve alongside coding sequences. Strong purifying selection conserves both, while relaxed selection allows more changes.

Hamming counts differences between two equal-length sequences.
$\text{Hamming}_{\text{normalized}} \rightarrow 1$ implies **more** differences.
$\text{Hamming}_{\text{normalized}} \rightarrow 0$ implies **less** differences.

K_a/K_s negatively correlates with Shannon entropy because conserved genes need stable regulatory signals, limiting variability. In contrast, relaxed selection increases mutations, raising entropy in non-coding regions.

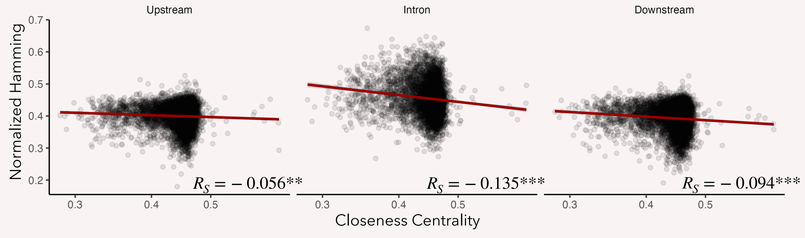
Shannon sequence entropy measures variability at a position in a sequence. **High entropy** indicates more information in the consensus sequence (high variation), while **low entropy** means less variation (less information content) across sequences.

Flanking Region vs. Network

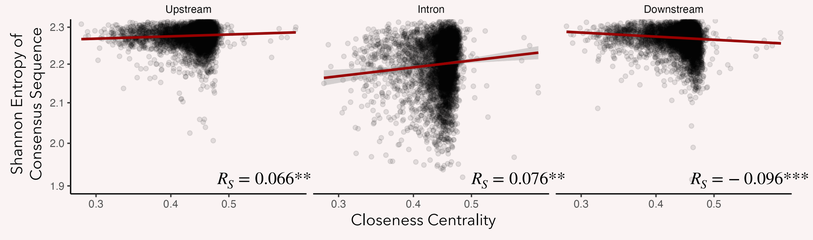
Hypothesis

As the gene is more connected to the rest of the network, its regulatory region will be more conserved/will have more islands of conservation.

Closeness Centrality and Conservation of Flanking Regions is Negatively Correlated



Closeness Centrality and Entropy of Introns and Upstream Regions is Positively Correlated



Hamming scores of **flanking regions** may negatively correlate with **closeness centrality** because genes connecting other genes face strong purifying selection, conserving regulatory regions and limiting variability[12].

String 1	String 2	Hamming	Normalized Hamming
KAROLIN	KATHRIN	3	3/7
KAROLIN	KERSTIN	3	3/7
2222	1111	4	4/4

- **Shannon entropy** of **introns** and **upstream** regions positively correlates with **closeness centrality** because these regions often contain regulatory elements that influence gene expression.
- Central genes in a network require more complex regulation, resulting in higher sequence information (entropy) in their introns and upstream regions.

References/Acknowledgements

I would like to thank all members of the Reed Lab for their input in crucial stages of development. I would also like to thank Daryl Lam, Michael R. McKain and Kevin M. Kocot to help run the analysis on UAHPC. Additional thanks to Wilson Leung for advice, providing the 4D tree, and maintenance of the Genomics Education Partnership computational infrastructure, supported by NSF Grant No. IUSE-1915544 and NIH Award R25GM130517 to LKR.

1. Alvarez-Ponce, D., et al. Position matters: Network centrality considerably impacts rates of protein evolution in the human protein–protein interaction network. Genome Biol. Evol. 9, evx117- (2017).
2. Alvarez-Ponce, D., et al. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. Genome Res 19, 234-42 (2009).
3. Alvarez-Ponce, D., et al. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. Genome Biol Evol 3, 87-101 (2011).
4. Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. Development 124, 1851-1864 (1997).
5. Behrens, S. & Vingron, M. Studying the Evolution of Promoter Sequences: A Waiting Time Problem. J. Comput. Biol. 17, 1591-1606 (2010).
6. Becker, M. et al. Large-scale correlation network construction for unraveling the coordination of complex biological systems. Nat. Comput. Sci. 3, 346-359 (2023).
7. Bergman, C. M. & Kreitman, M. Analysis of Conserved Noncoding DNA in Drosophila Reveals Similar Constraints in Intergenic and Intronic Sequences. Genome Res 11, 1335-1345 (2001).
8. Davidson, E. H. cis-Regulatory Modules, and the Structure/Function Basis of Regulatory Logic. in The Regulatory Genome (ed. Davidson"], ["Eric H.) 31-86 (Academic Press, Burlington, 2006). doi:10.1016/b978-012088563-3.50020-1.
9. Emms, D.M., Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20, 238 (2019). https://doi.org/10.1186/s13059-019-1832-y
10. Villar, D. et al. Enhancer Evolution across 20 Mammalian Species. Cell 160, 554-566 (2015).
11. Wittkopp, P. J. Evolution of cis-regulatory sequence and function in Diptera. Heredity 97, 139-147 (2006).
12. Yona, A. H., et al. Random sequences rapidly evolve into de novo promoters. Nat. Commun. 9, 1530 (2018).
13. Yu, H. & Gerstein, M. Genomic analysis of the hierarchical structure of regulatory networks. Proc National Acad Sci 103, 14724-14731 (2006).

83
CSHL – Network Biology 2025