Network Evolution

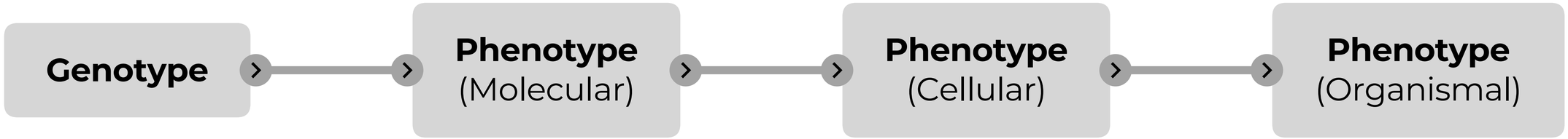
Evolution of Protein Networks and Genomic Conservation

2025/11/21
Chinmay P. Rele

Departmental Seminar Series
Reed Lab — Biological Sciences

Regulation is Complex

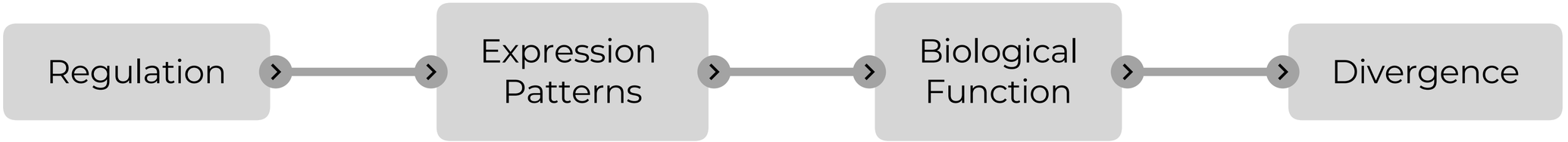
Why



- Genotype can affect phenotype at different scales.
- Evolutionary forces act on these different phenotypes differently.

Regulation is Complex

Why



Network Evolution

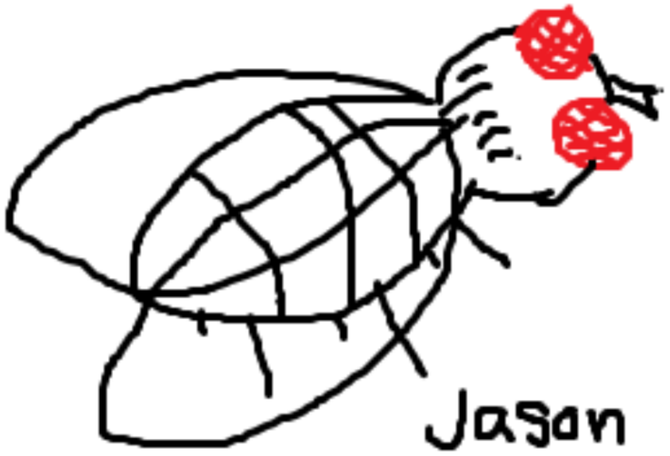
Why

- Predicts proteome rewiring
- Reveals functional adaptation
- Allows for understanding of disease mechanisms
  - Cancer
  - Neurogenerative
  - Infectious (viruses/bacteria)

Network Evolution

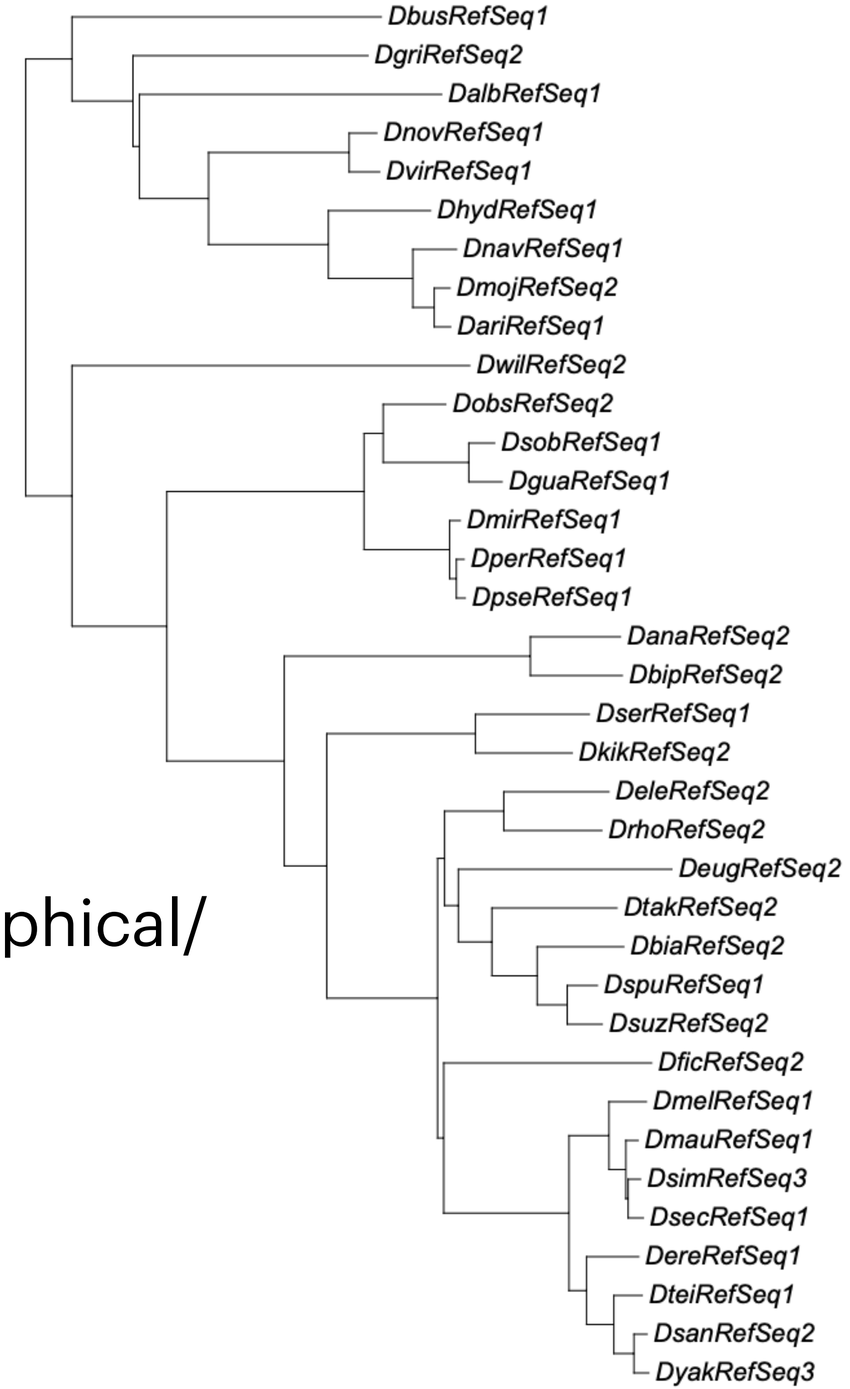
What

- 1990s → Yeast 2 Hybrid
- 2000/2003 → High throughput screening
- 2010s → Genomics and disease
- NOW → Machine Learning/Evolutionary protein-protein interaction mapping



Drosophila

Why



- Well annotated across clades by NCBI
- Evolutionary relationships known through geographical/ morphological/genetic assays
- “Small” genomes easy to compute

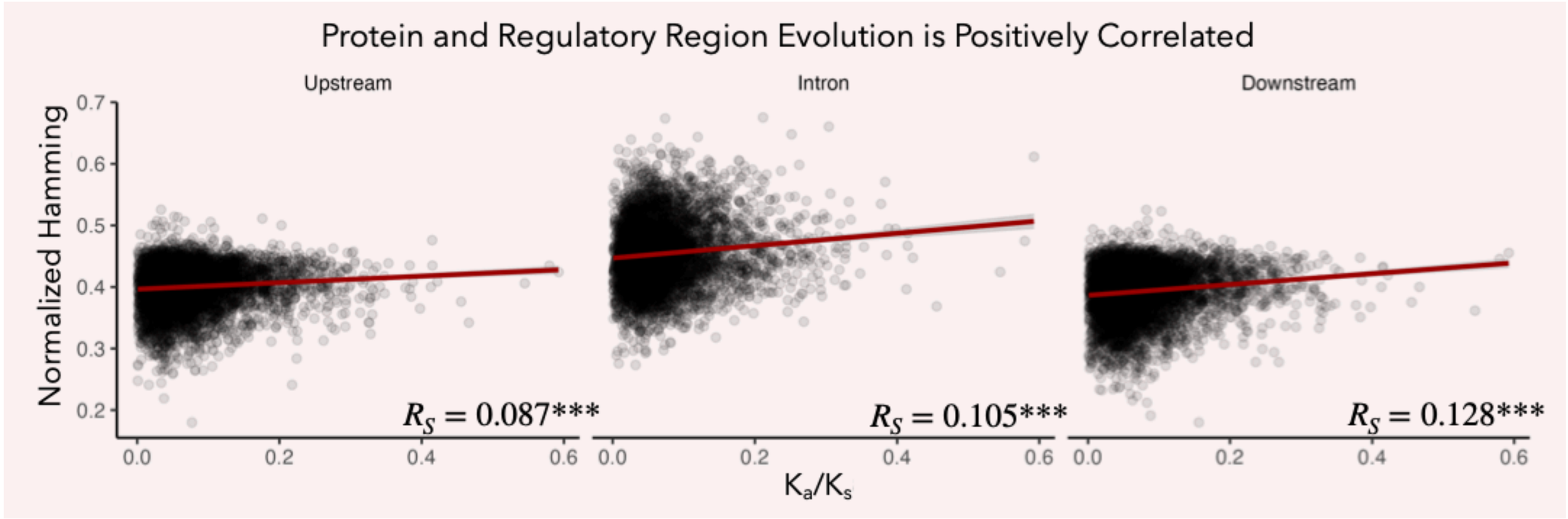
Chapters

1. Evolution of protein coding sequence to regions that regulate their expression
2. Evolution of protein coding sequence and regions that regulate their expression relative to network architecture
3. How the network architecture evolves across the clade

Sequence Conservation is Correlated

Coding Sequence vs. Regulatory Region

MORE CONSERVED



MORE CONSERVED

Sequence Conservation is Correlated

Measures of Conservation

$$\frac{K_a}{K_s} = \begin{cases} < 1 & \rightarrow \text{purifying selection/constraint} \\ 1 & \rightarrow \text{neutral evolution} \\ > 1 & \rightarrow \text{adaptive evolution} \end{cases}$$

Met	Asp	Thr	Ala	Val
ATG	GAC	ACA	GCG	GTT
ATG	GCC	ACT	TCG	GTT
Met	Ala	Thr	Ser	Val

$$K_a = \frac{2}{5} = 0.4$$

$$K_s = \frac{1}{5} = 0.2$$

$$\frac{K_a}{K_s} = \frac{0.4}{0.2} = 2$$

- Rate of synonymous/non-synonymous mutations (accounting for neutral evolution)
- Only for coding sequences

Sequence Conservation is Correlated

Measures of Conservation

Hamming

"**karolin**" and "**kathrin**" is **3**.
"**karolin**" and "**kerstin**" is **3**.
"**kathrin**" and "**kerstin**" is **4**.
0000 and **1111** is **4**.
2173896 and **2233796** is **3**.

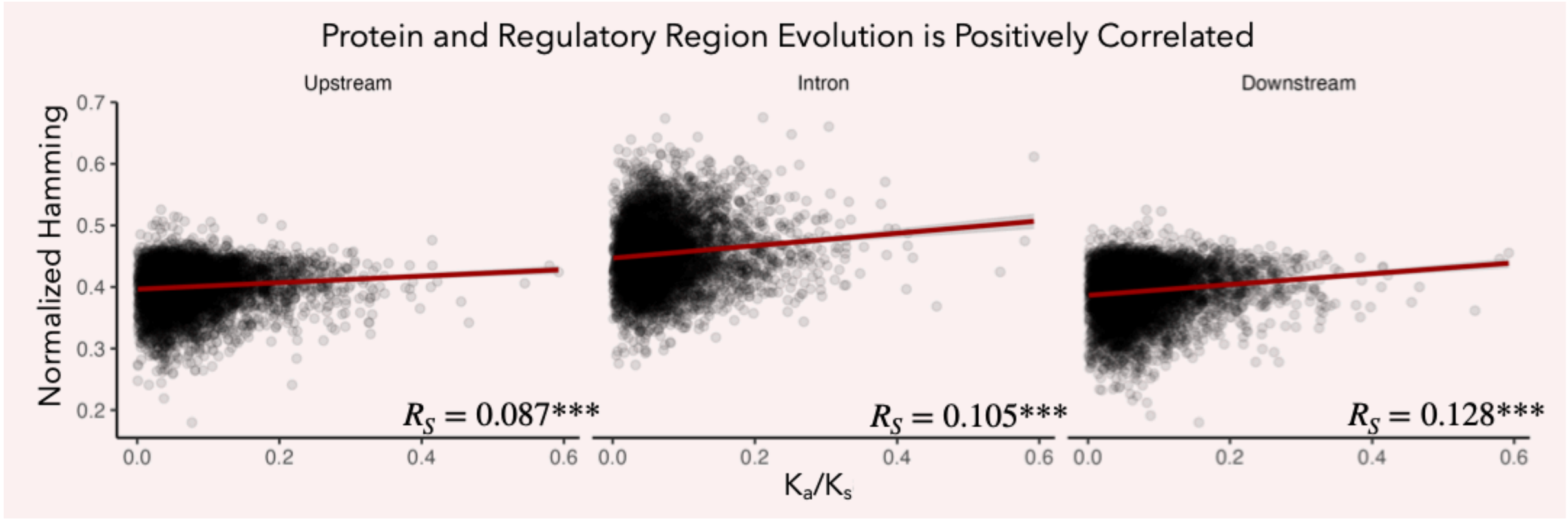
$$Hamming_{normalized} = \frac{Hamming}{Length} = \frac{4}{7}$$

Total number of changes in the sequence of identical length

Normalized Hamming takes into account the hamming score relative to the total length of the sequences.

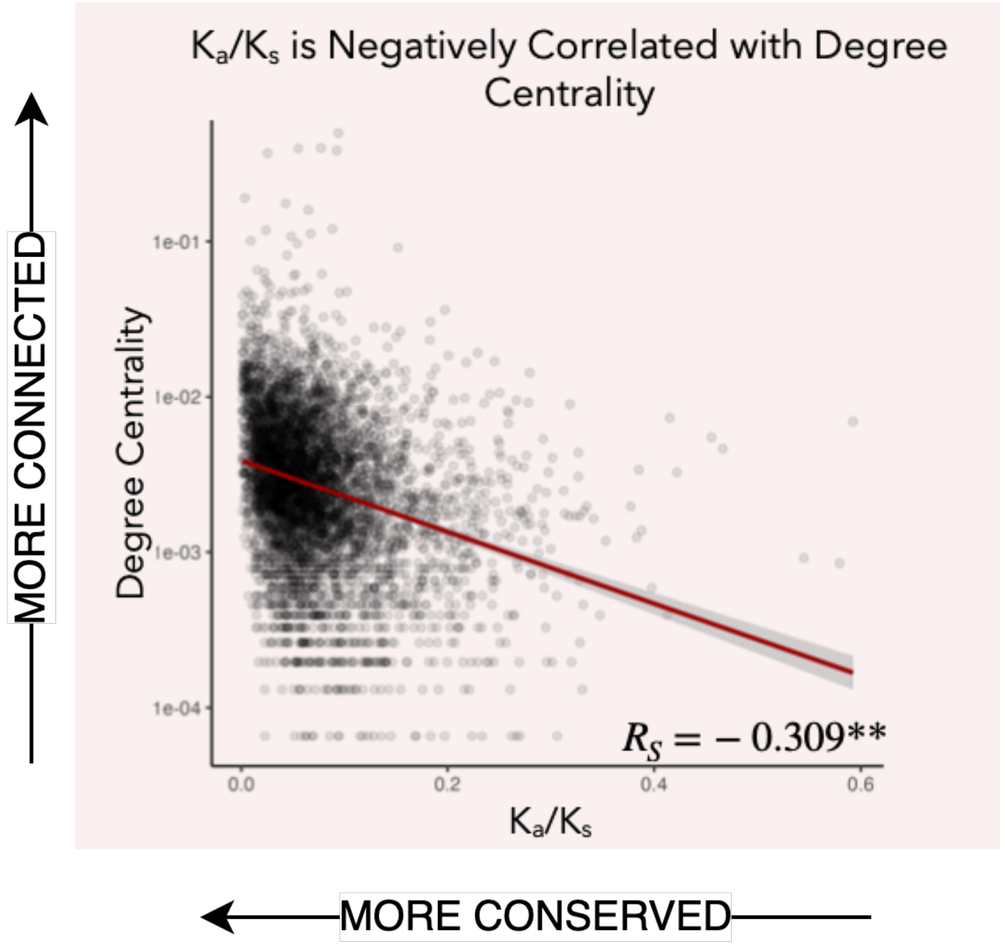
Sequence Conservation is Correlated

Coding Sequence vs. Regulatory Region



Network Position is Correlated with Sequence Divergence

Coding Sequence/Regulatory Region vs. Network

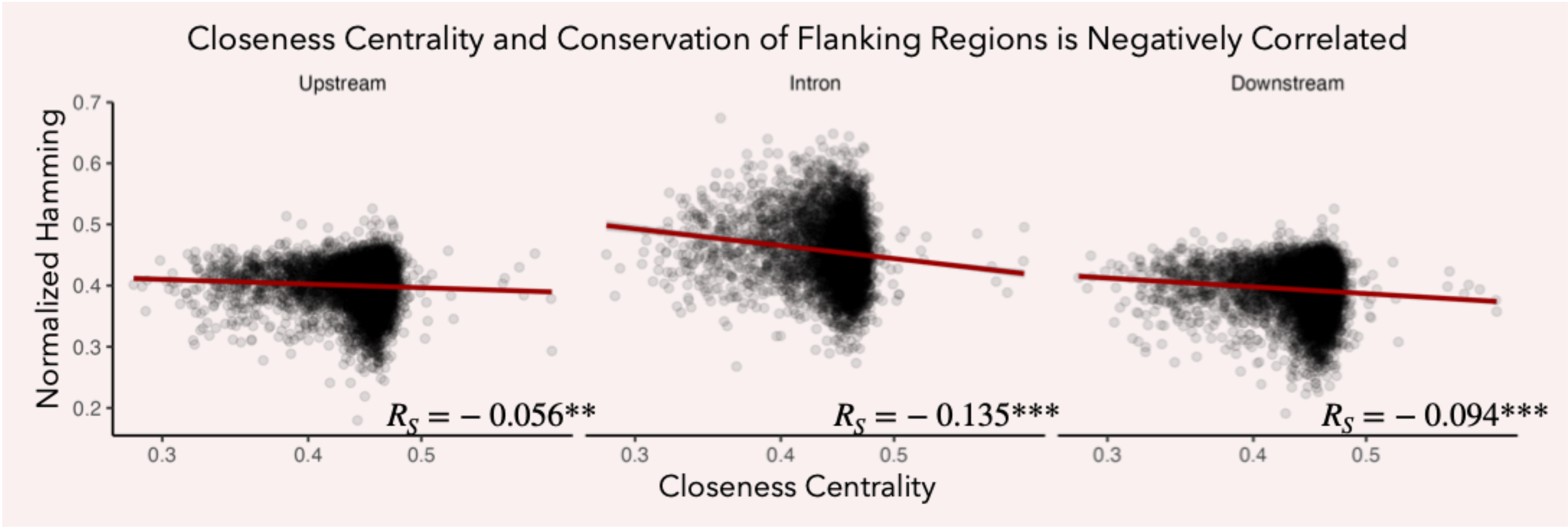


A higher degree centrality for a node indicates that it is more connected to other nodes in the network.

Genes with more conserved coding sequences tend to be connected with more genes.

Network Position is Correlated with Sequence Divergence

Coding Sequence/Regulatory Region vs. Network

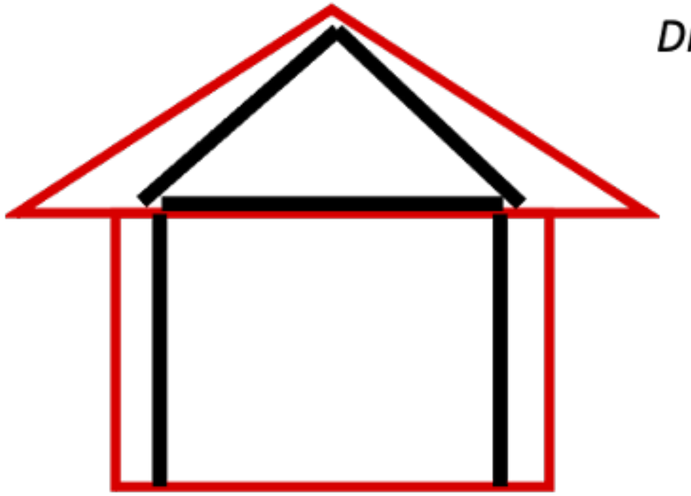


Network Architecture

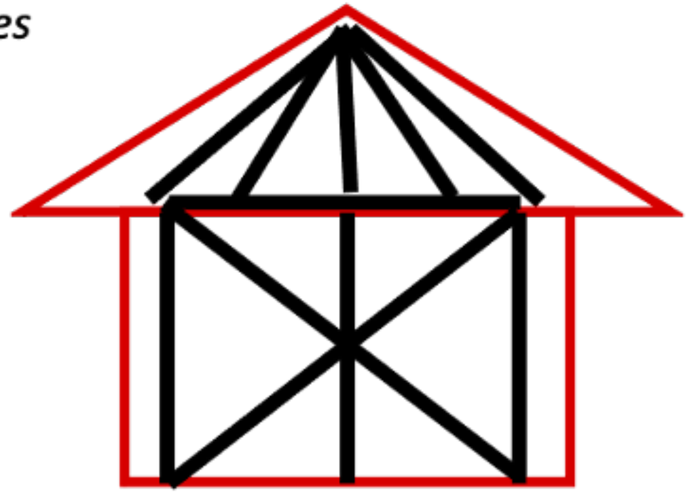
How it Affects Evolution

Same Phenotype (house)

Different architectures



easily perturbed but highly evolvable



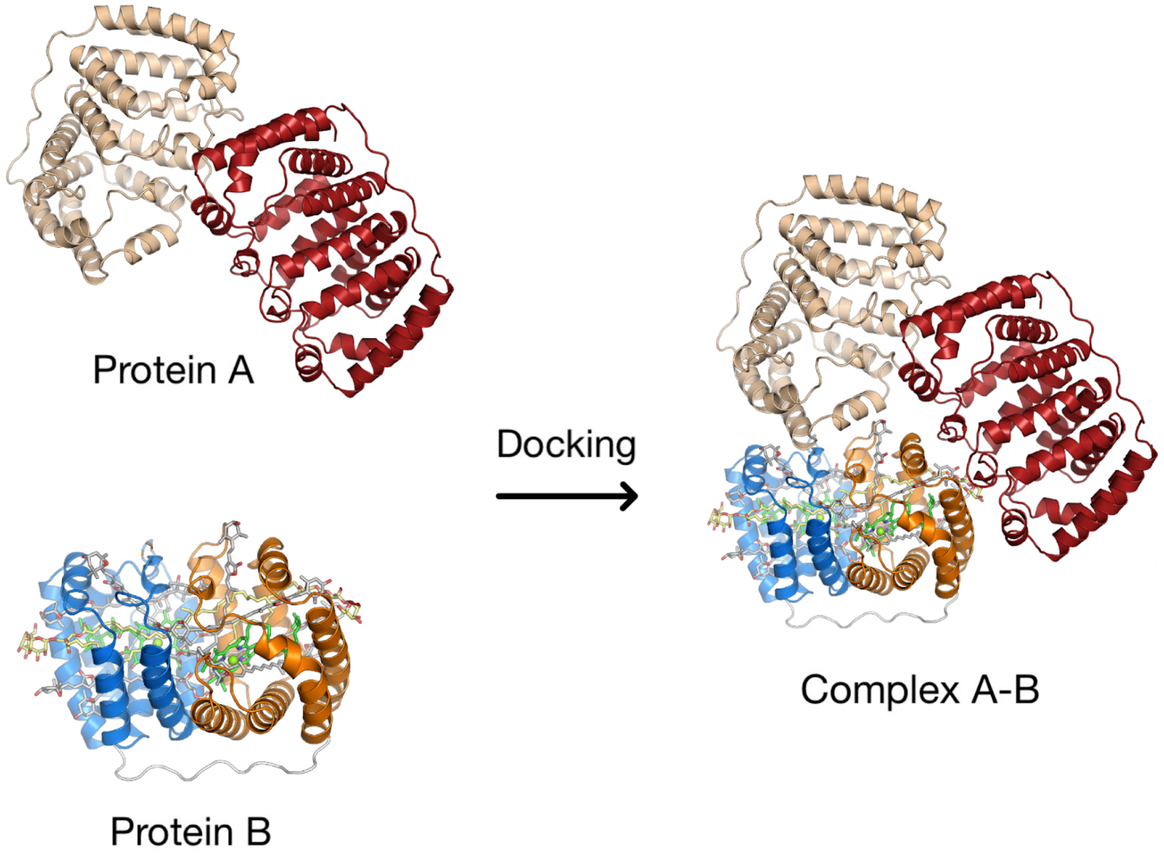
robust but resistant to evolution

Reed L.K., Personal Communication

Weitz et al. (2007) Evolution, Interactions, and Biological Networks. PLoS Biol 5(1): e11. https://doi.org/10.1371/journal.pbio.0050011

Protein-Protein Interactions Inference

Connections



A connection/interaction is inferred if two proteins can interact with each other to form a stable complex.

https://neurosnap.ai/blog/post/protein-protein-docking-simplified-illuminating-the-mechanics-of-protein-interactions/64b5c256f78f3ad5845d8ac0

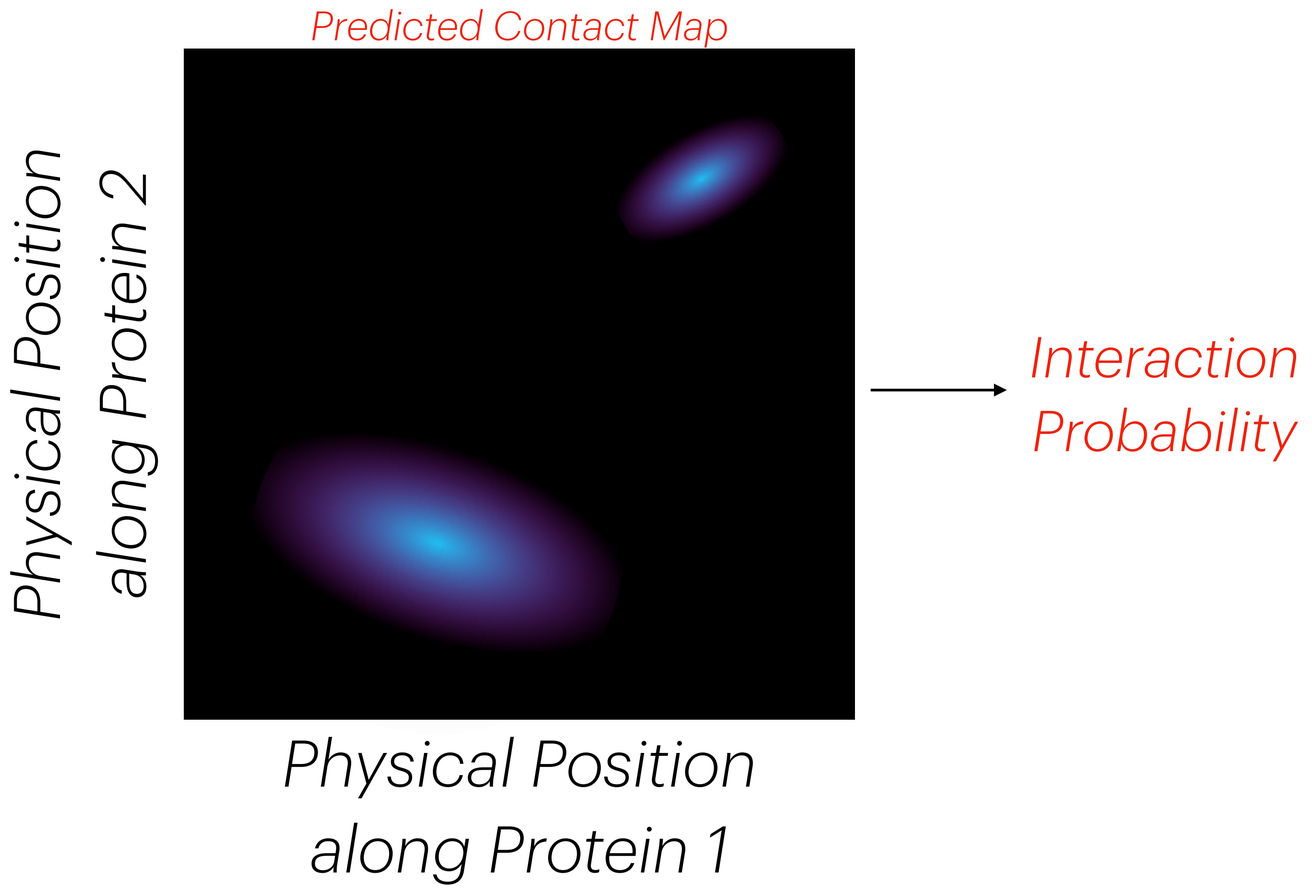
Protein-Protein Interactions Inference

Interaction Assay Types

Empirical Studies	***in silico* Assays**
Biologically-derived interactions	Generally sequence-derived interactions
Confirms real interactions	Predicts probabilities of interactions
Protein of interest to see all interacting partners	Can take any proteins and identify probability of interaction
High Confidence Biological	Can be run on multiple genes/species easily

PHILHARMONIC/D-SCRIPT

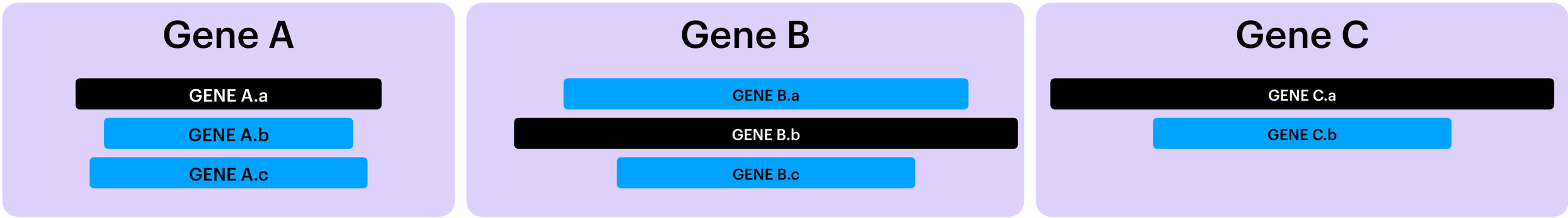
in silico Network Prediction



DOI: 10.1101/2024.10.25.620267; DOI: 10.1016/j.cels.2021.08.010

PHILHARMONIC/D-SCRIPT

in silico Network Prediction



PHILHARMONIC/D-SCRIPT

in silico Network Prediction

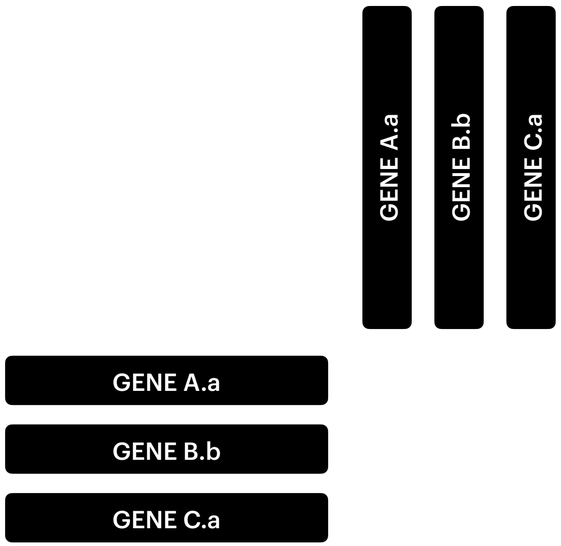
GENE A.a

GENE B.b

GENE C.a

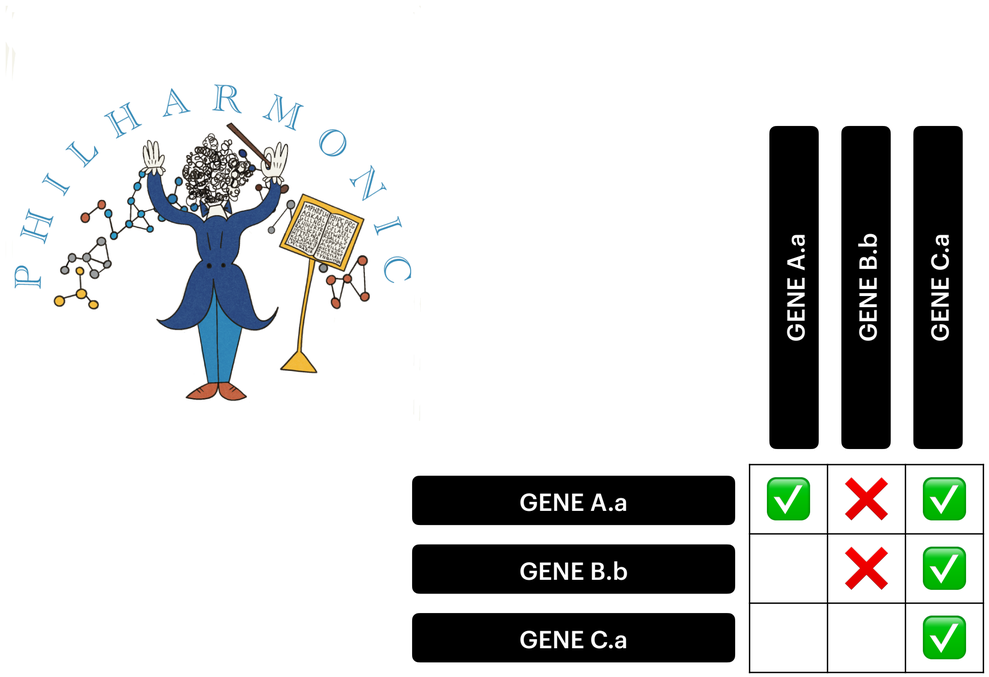
PHILHARMONIC/D-SCRIPT

in silico Network Prediction



PHILHARMONIC/D-SCRIPT

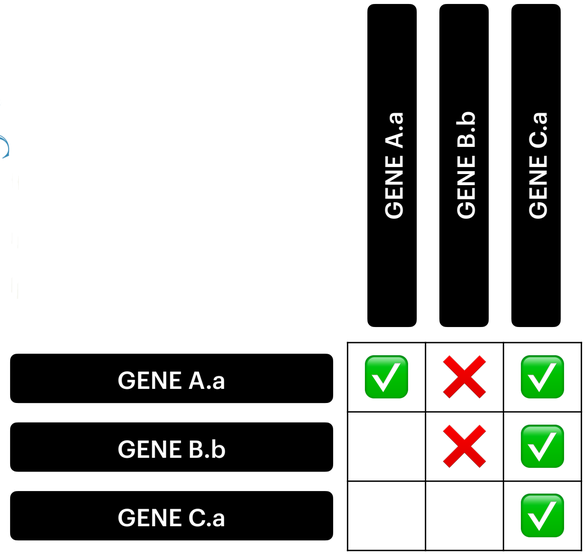
in silico Network Prediction

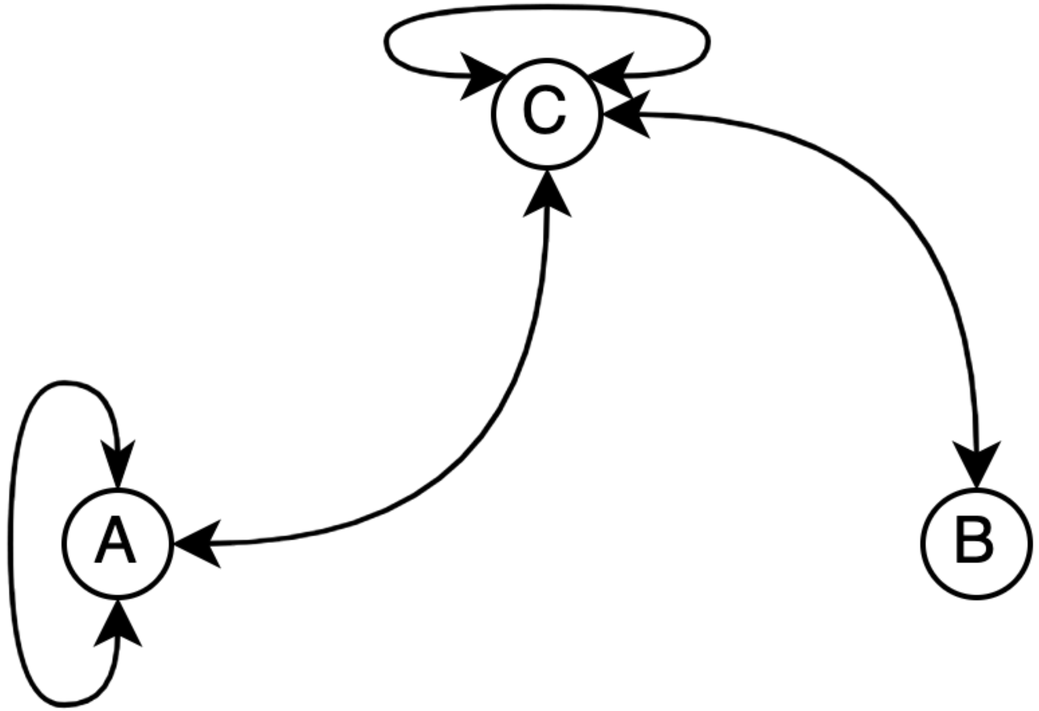


DOI: 10.1101/2024.10.25.620267; DOI: 10.1016/j.cels.2021.08.010

PHILHARMONIC/D-SCRIPT

in silico Network Prediction





DOI: 10.1101/2024.10.25.620267; DOI: 10.1016/j.cels.2021.08.010

Network Summaries

Network Architecture

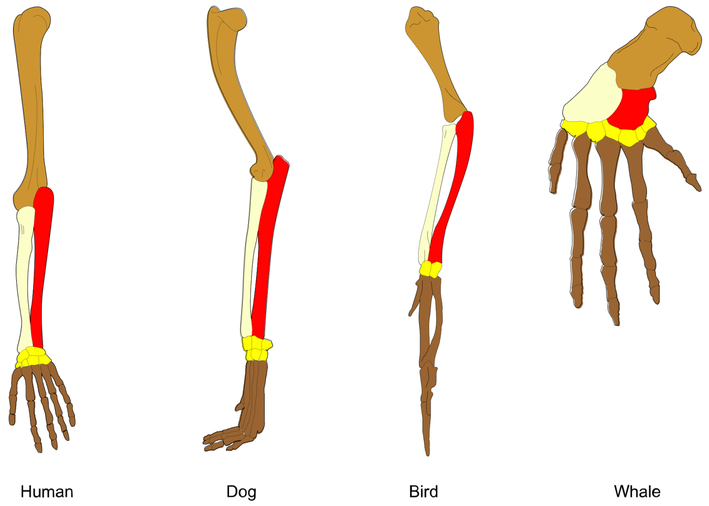
Each species have similar _____ in their networks as each other.

- number of proteins
- number of interactions
- density $\rho = \dfrac{\text{number of present edges}}{\text{number of possible edges}}$

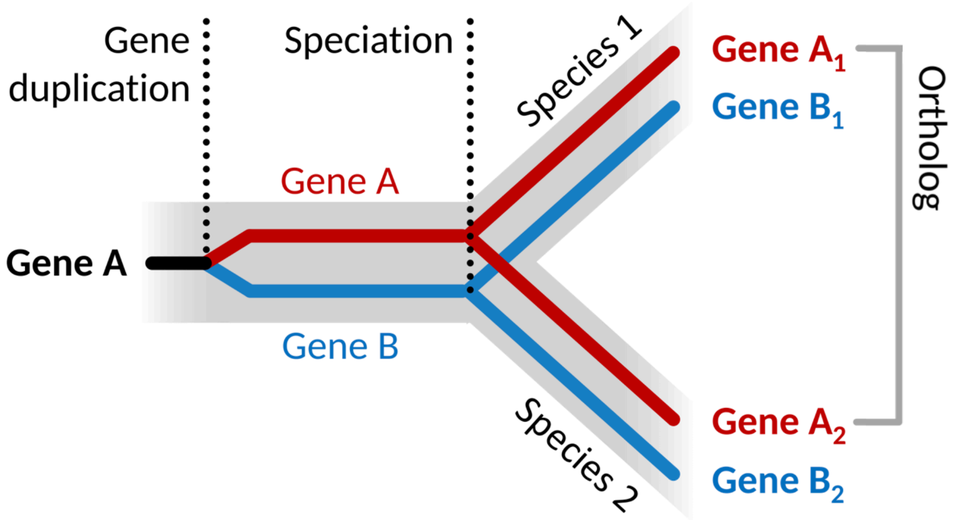
Neighborhood Consistency

Orthogroup Assignment

Orthology is a specialized version of homology.



Any gene that has the same ancestry in 2+ species.



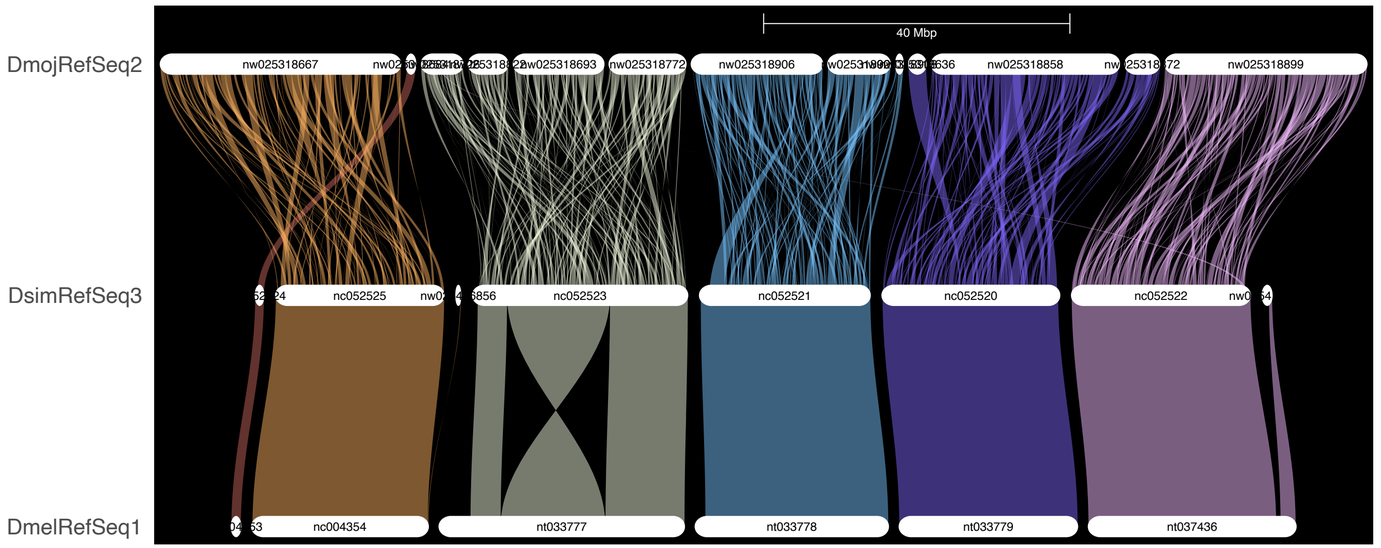
GENESPACE

Neighborhood Consistency

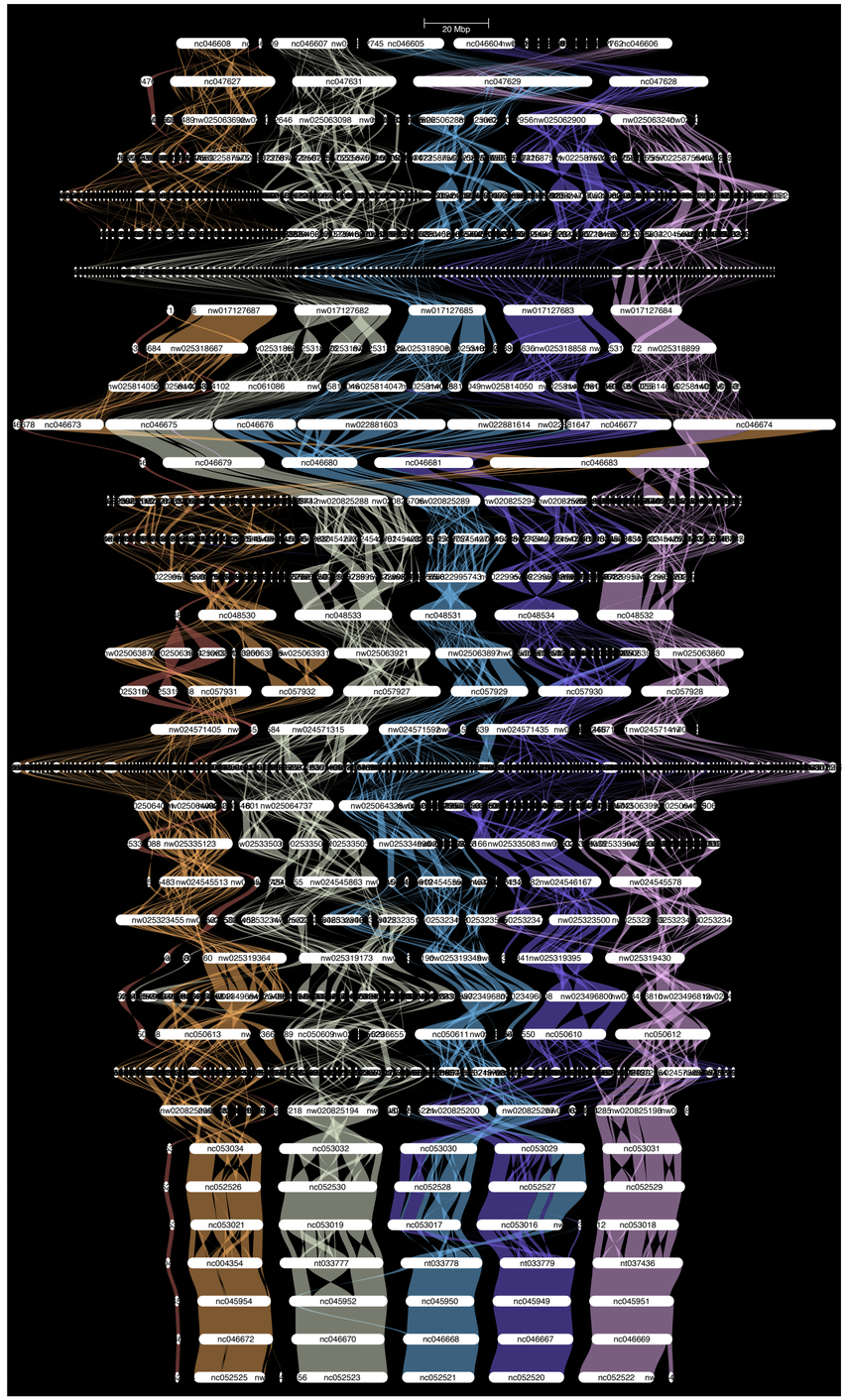
Specialization within Clade

Uses sequence identity to identify orthology.

Refines orthology using position within genome.

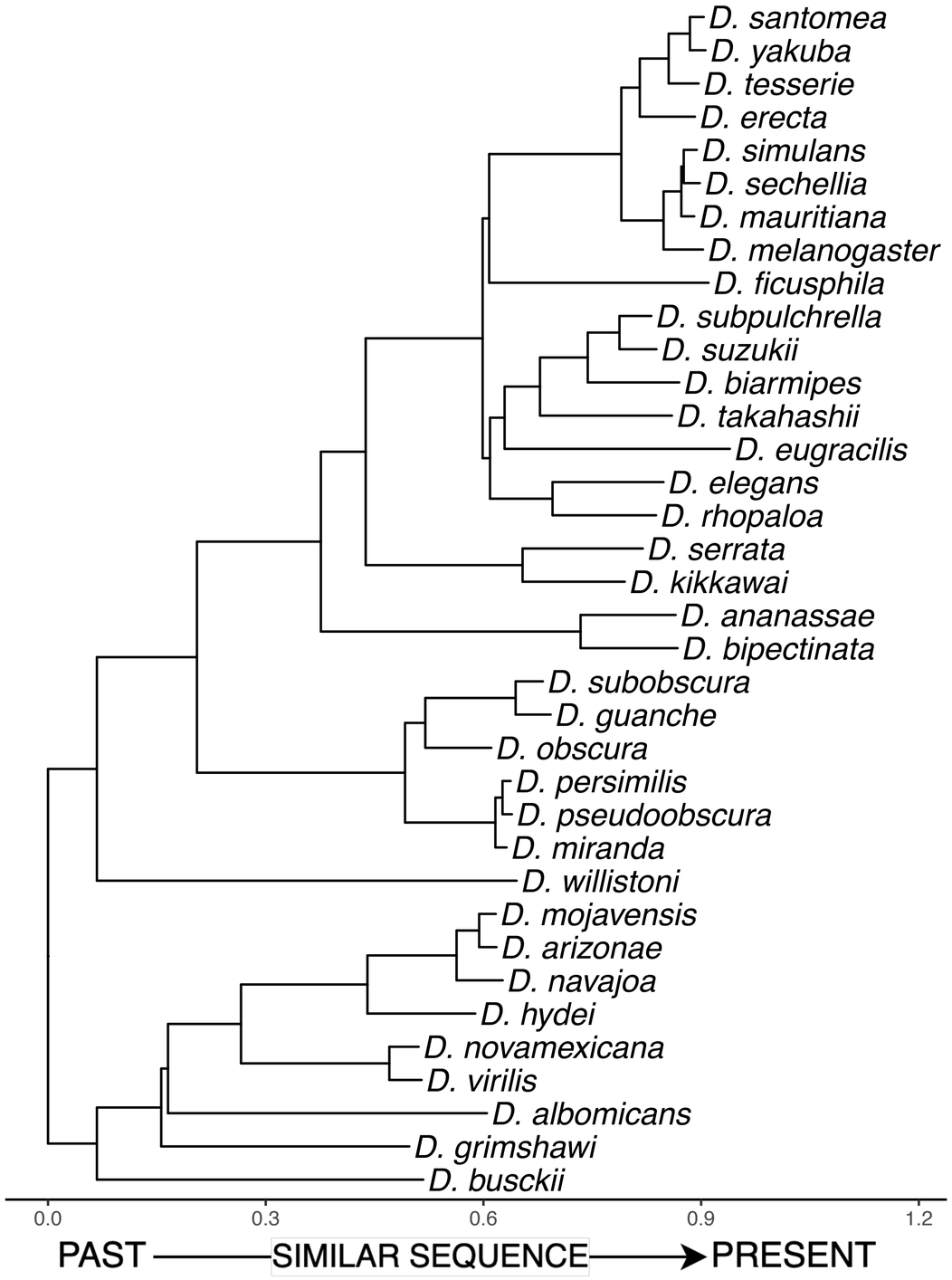


Chromosomes scaled by physical position



Chromosomes scaled by physical position

John T Lovell, Avinash Sreedasyam, M Eric Schranz, Melissa Wilson, Joseph W Carlson, Alex Harkess, David Emms, David M Goodstein, Jeremy Schmutz (2022) GENESPACE tracks regions of interest and gene copy number variation across multiple genomes eLife 11:e78526



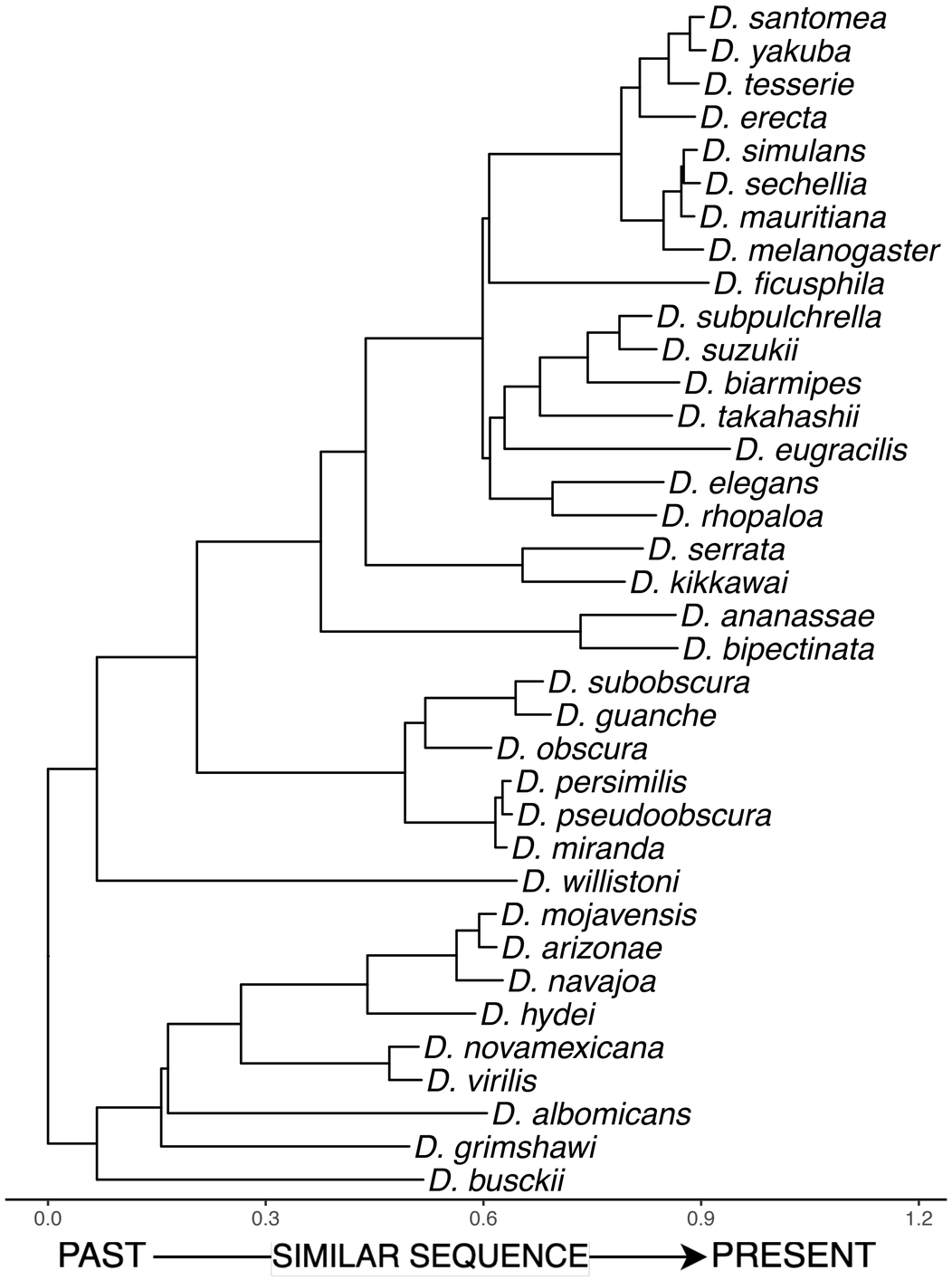
Neighborhood Consistency

Specialization within Clade

Closely related species have similar protein sequence.

Species with distant ancestry have more dissimilar sequence.

Sequence → Structure → Connections → Function

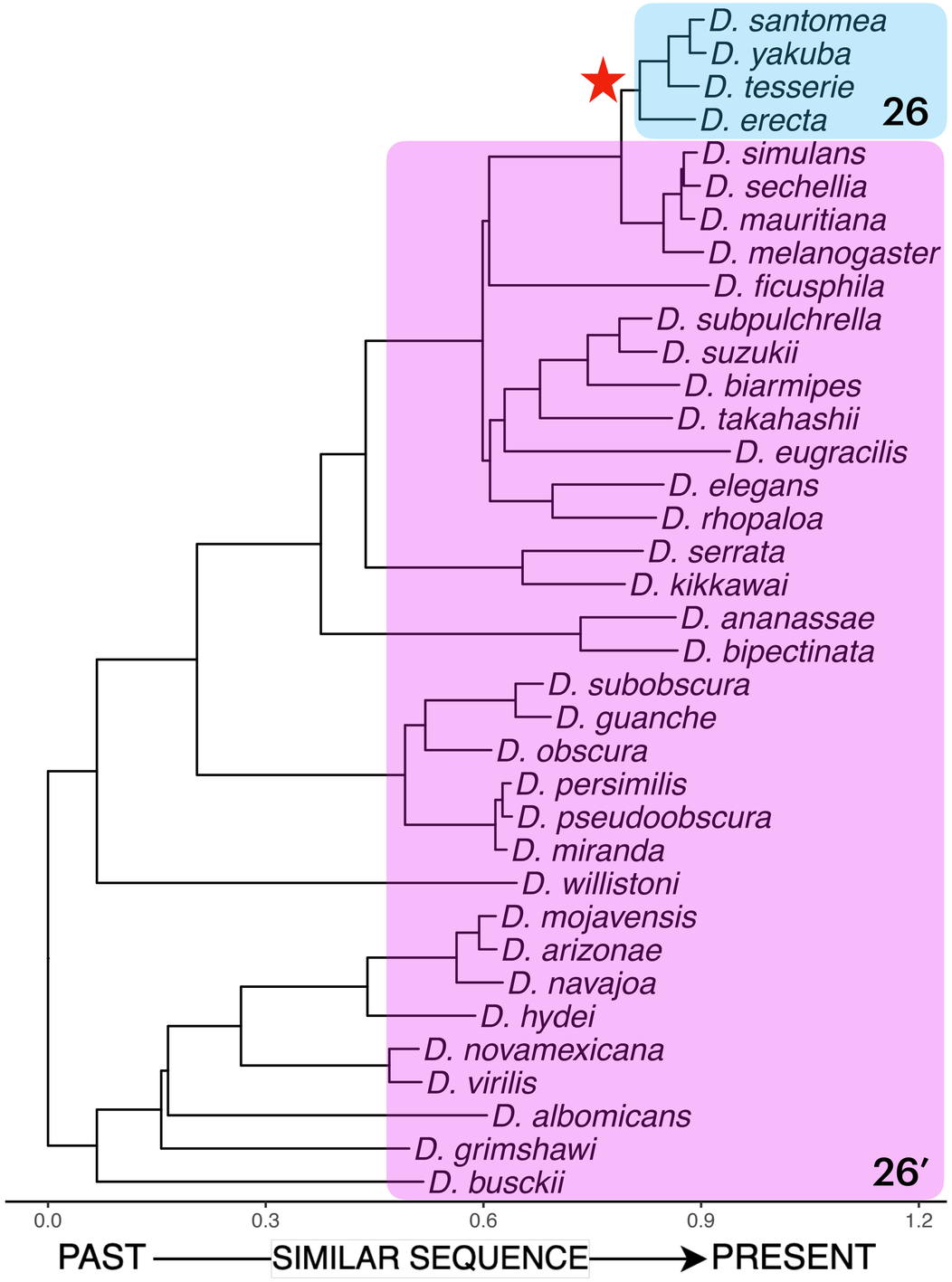


Neighborhood Consistency

Specialization within Clade

Sequence variation has already been studied across clade.

Variation in function is derived through variation in connections to other proteins.

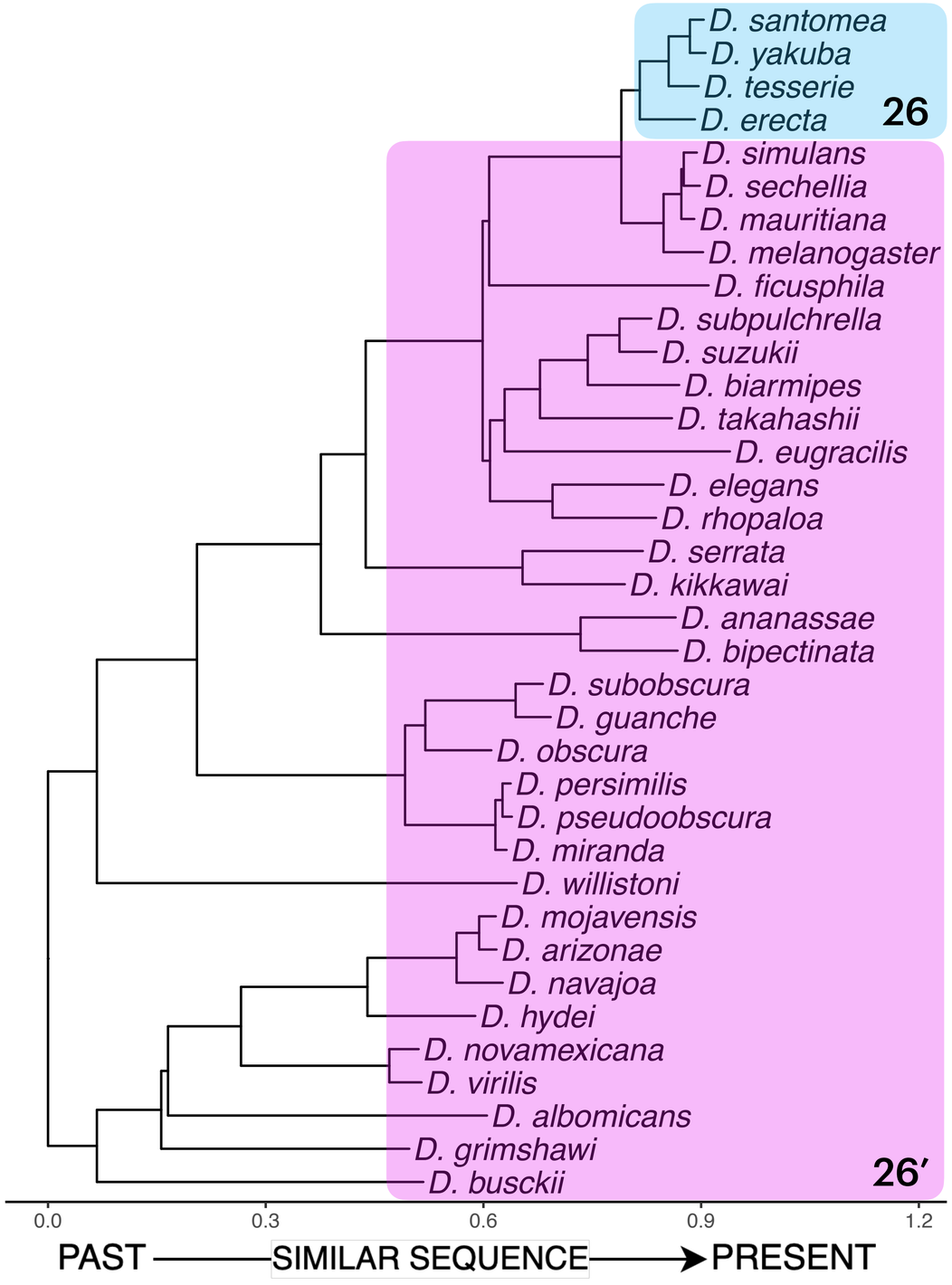


Neighborhood Consistency

Specialization within Clade

How similar the neighborhood of a specific protein within a node.

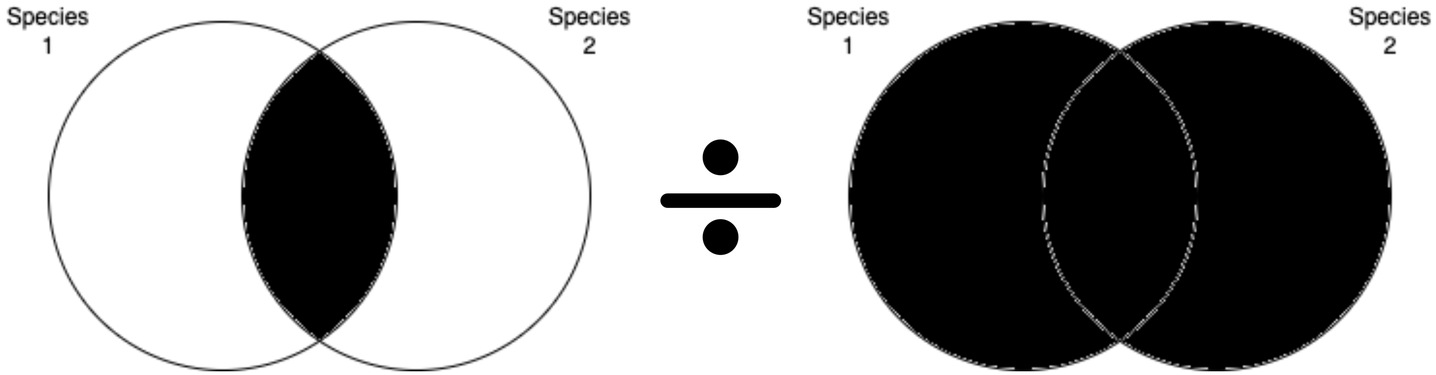
Compare for all species in one one node (26) vs all species not in that node (26′).



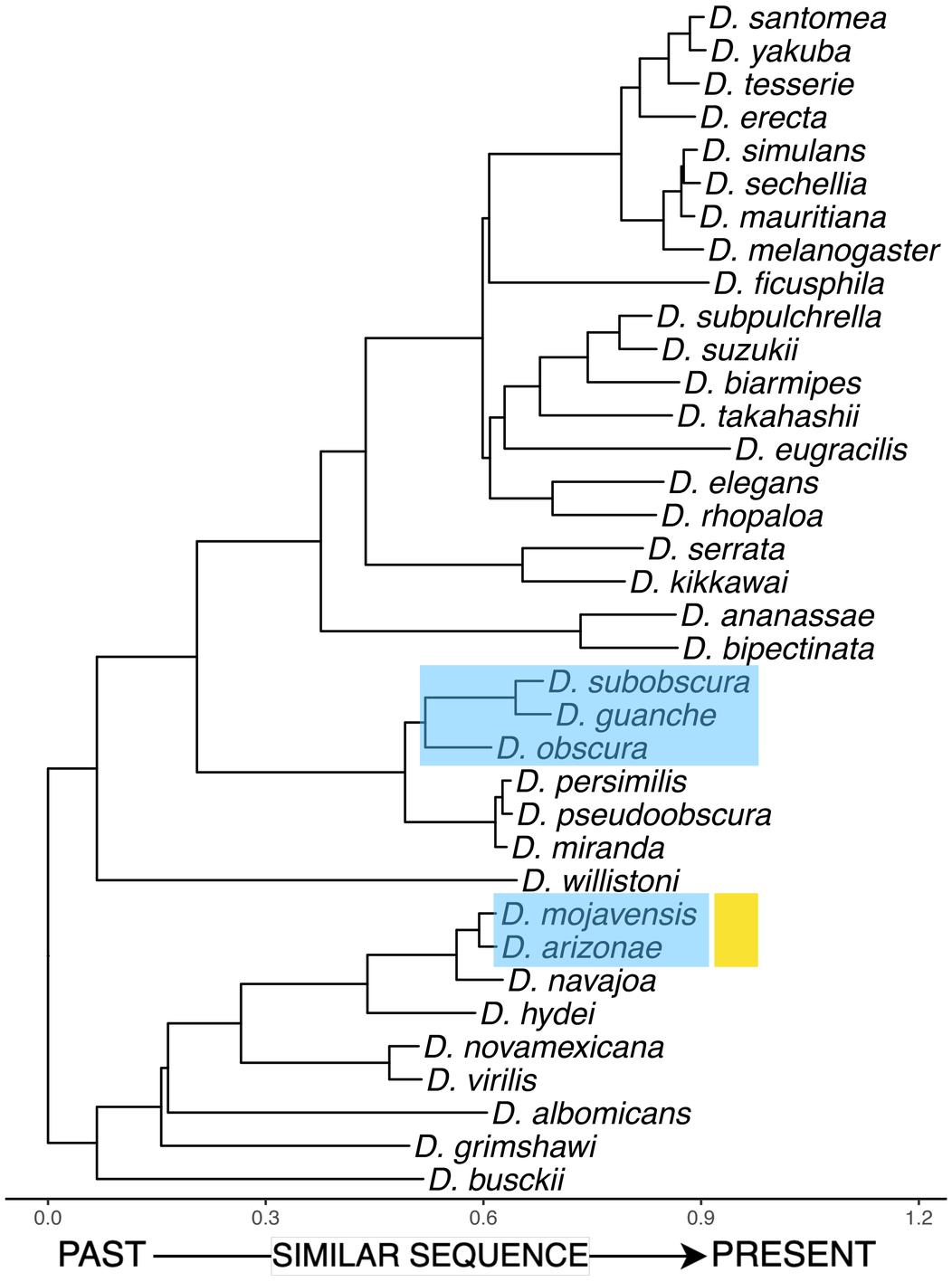
Neighborhood Consistency

Specialization within Clade

How similar the neighborhood of a specific protein within a node.



Variation of Jaccard Index



Neighborhood Consistency

Specialization within Clade

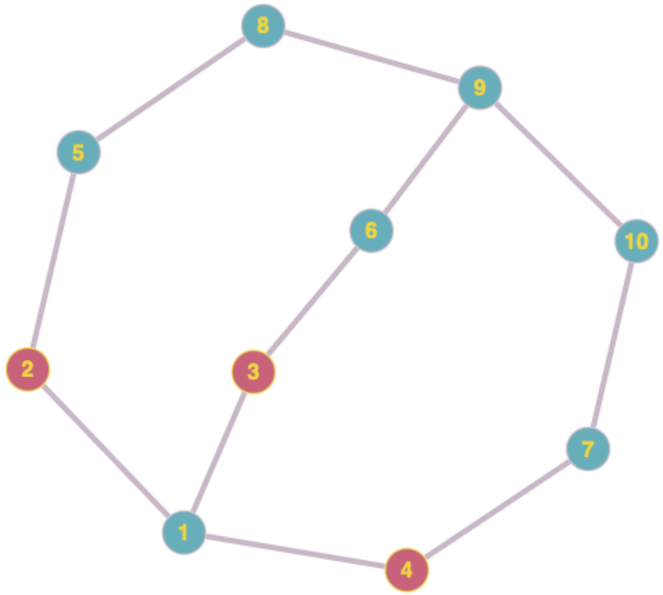
Node	Protein	Δ consistency
24	*Tankyrase*	-0.607
10		-0.531
24	*oskyddad*	-0.685

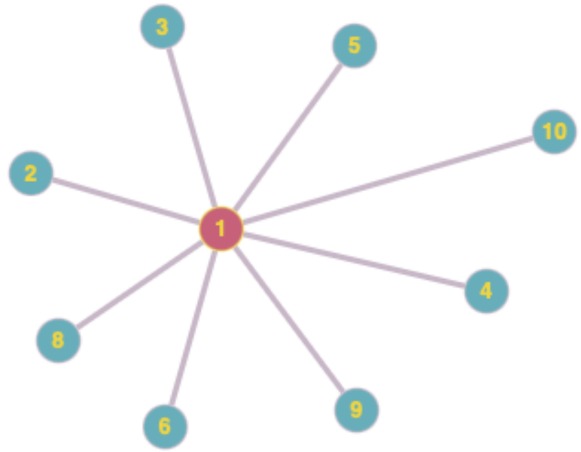
- Some genes are evolving more in certain lineages than others
- Pattern need further individual investigation

* Low statistical power due to small sample size

Neighborhood Consistency

Correlation to Genomic Sequence

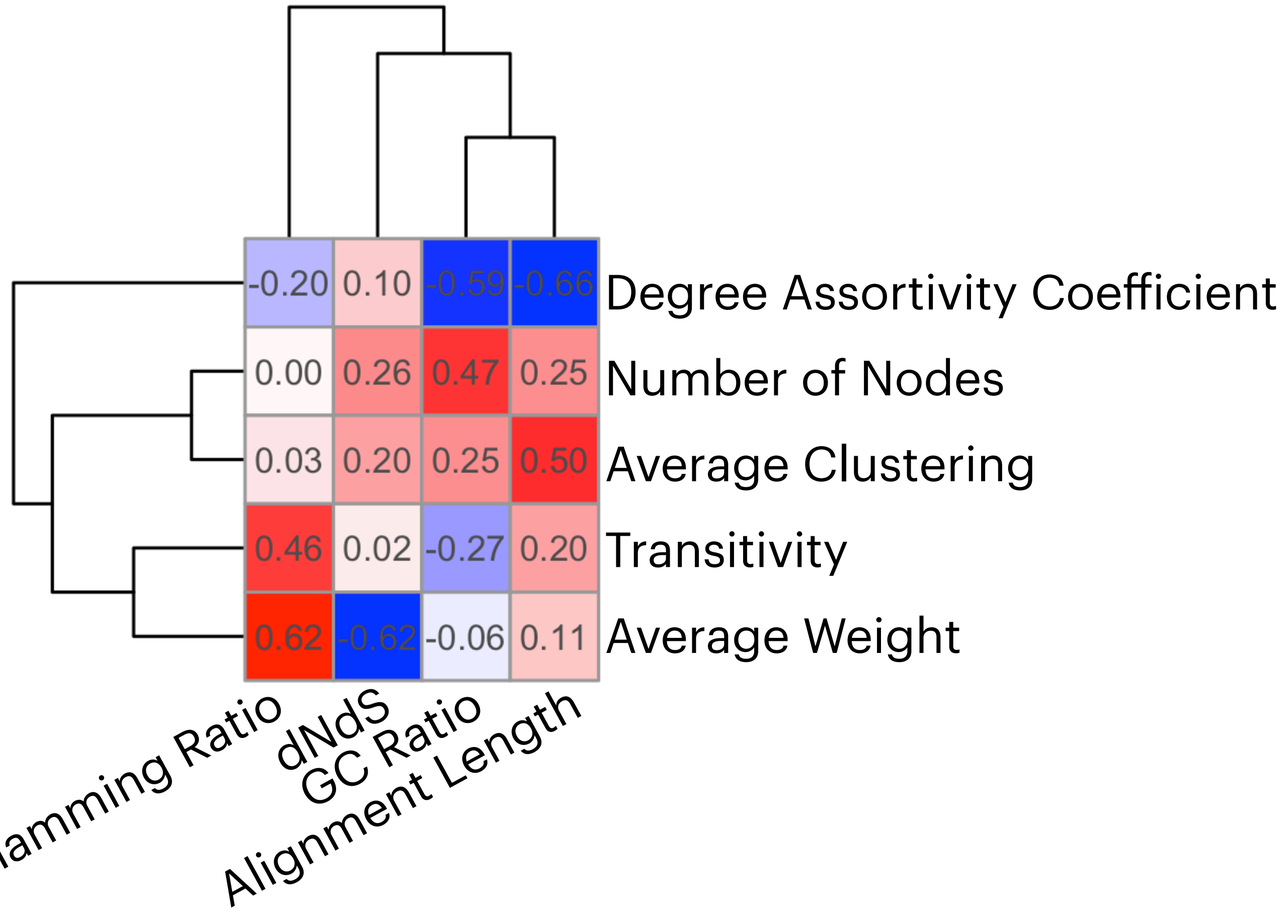




graphonline.ru

Neighborhood Consistency

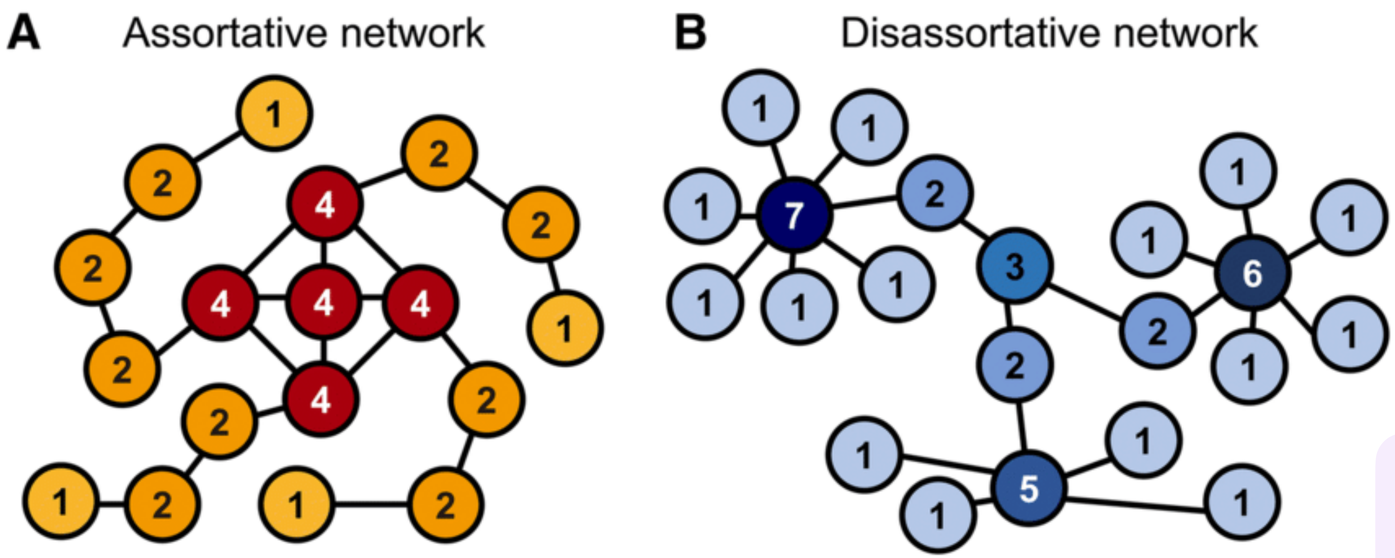
Correlation to Genomic Sequence



* Low statistical power due to small sample size

Neighborhood Consistency

Assortavity vs. Regulatory Region GC Ratio



High NEGATIVE correlation with GC ratio.

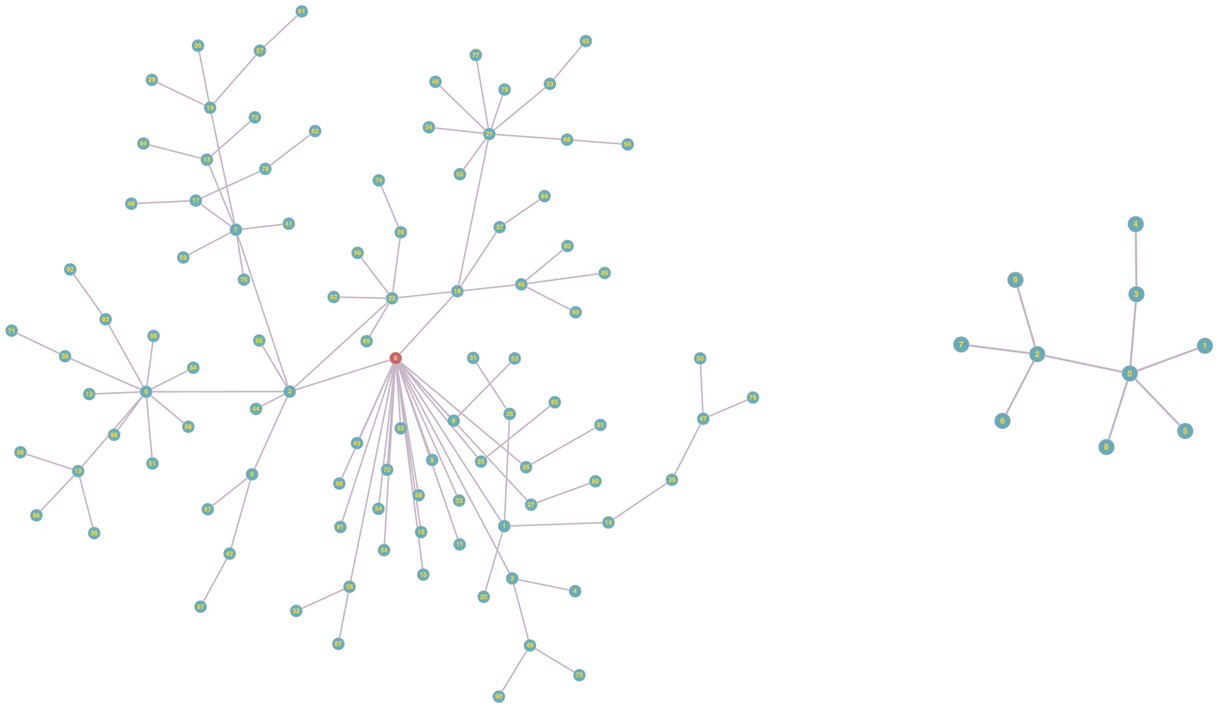
- Low assortavity means that highly connected modes are connected to lowly connected nodes.
- More **hub and spoke** architecture

Networks with more hub/spoke architecture had more GC content.

Park, Kang Min & Cho, Kyoo Ho & Lee, Ho-Joon & Heo, Kyoung & Lee, Byung & Kim, Sung. (2020). Predicting the antiepileptic drug response by brain connectivity in newly diagnosed focal epilepsy. Journal of Neurology. 267.
* Low statistical power due to small sample size

Neighborhood Consistency

Number of Nodes vs. Regulatory Region GC Ratio



High POSITIVE correlation with regulatory region GC ratio.

- Number of unique interacting partners across the entire clade

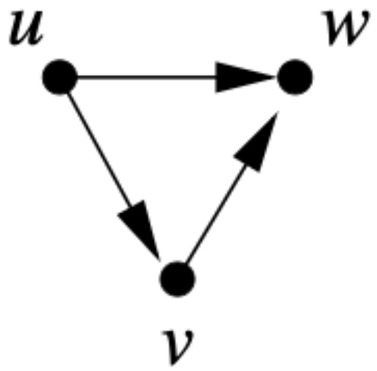
The more the interacting partners, the higher the GC content.

* Low statistical power due to small sample size

Networks (2005) *Newman;* graphonline.ru

Neighborhood Consistency

Transitivity vs. Regulatory Region Conservation



A transitive triple of nodes in a directed network.

- Transitivity is how often two connected nodes have the same neighbor
- Global metric

High POSITIVE correlation with Hamming.

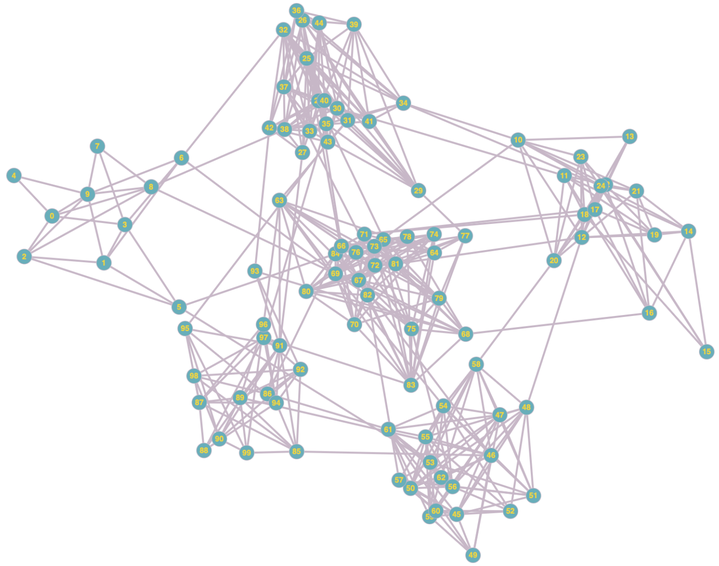
More closed triplets meant more change in regulatory region sequence.

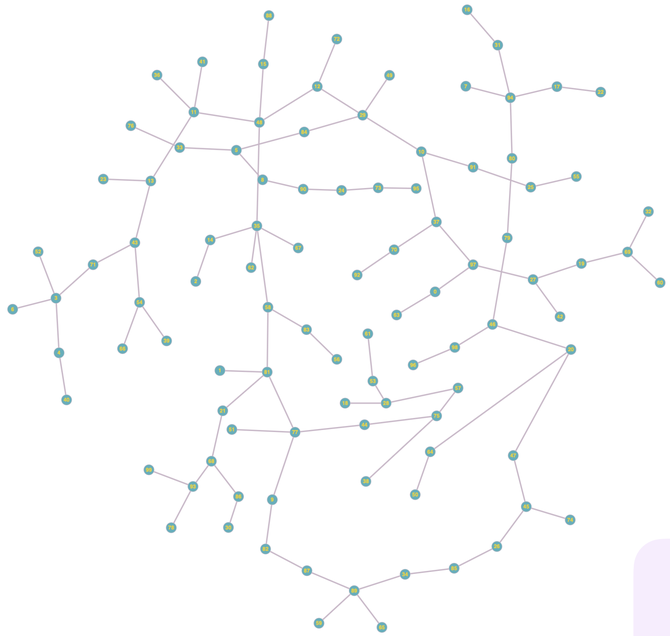
* Low statistical power due to small sample size

Networks (2005) *Newman*

Neighborhood Consistency

Clustering vs. Regulatory Region Alignment Length





High POSITIVE correlation with regulatory region alignment length.

- How clustered are neighborhoods
- Local metric

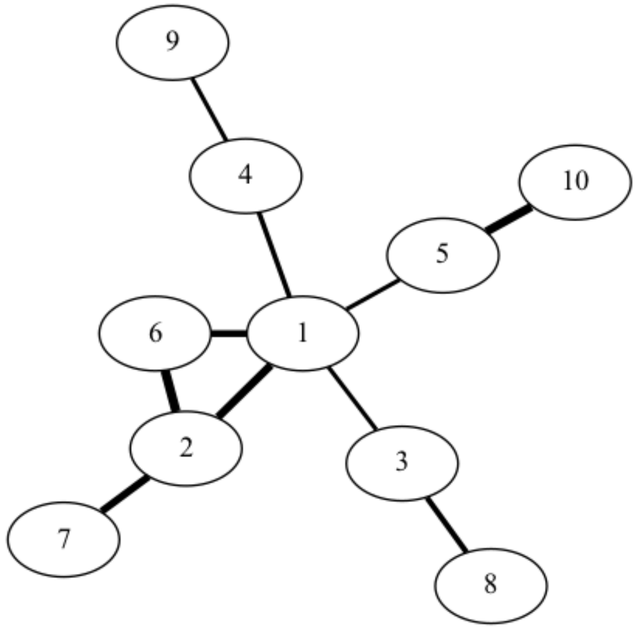
Higher the clustering, the longer the multiple sequence alignment.

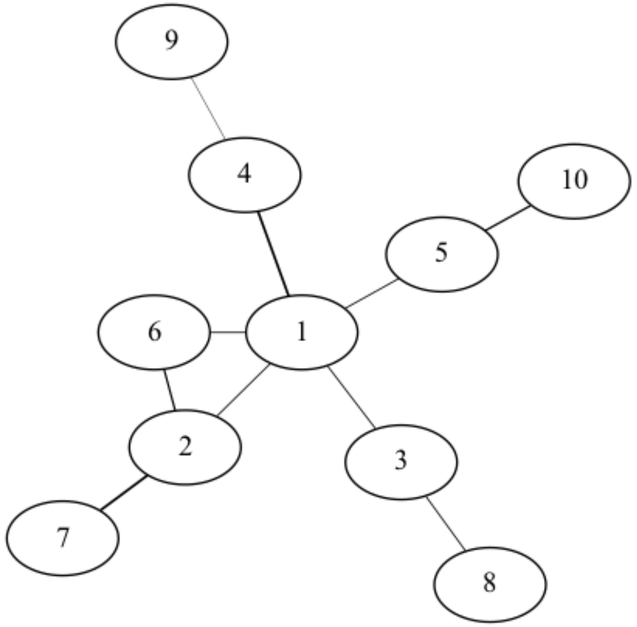
* Low statistical power due to small sample size

graphonline.ru

Neighborhood Consistency

Average Edge Weight vs. Conservation





Higher average edge weight indicates more conserved neighborhood across clade

High NEGATIVE correlation with dNdS.

High POSITIVE correlation with Hamming.

More conserved protein-coding sequence means more conserved connections. But why positive correlation with Hamming?

* Low statistical power due to small sample size

Networks (2005) *Newman*; pygraphviz

Network Evolution

Further Work

Perform **permutation testing** on edges of networks to see if patterns re-emerge

- Using scale-free model of preferential attachment

Further explore **motif distribution** in regulatory regions using MEME

- Using motif entropy as an analog of change in motif distribution

Timothy L. Bailey, James Johnson, Charles E. Grant, William S. Noble, "The MEME Suite", Nucleic Acids Research, 43(W1):W39-W49, 2015. [full text] (https://academic.oup.com/nar/article/43/W1/W39/2467905)

Network Evolution

Conclusion

- Coding sequence and regulatory regions show positively correlated correlation.
- Genomic conservation correlates with network connectivity.
- Changes in networks are correlated with protein evolution.
  - Interesting potential for exploring relationship between network architecture.

Acknowledgements

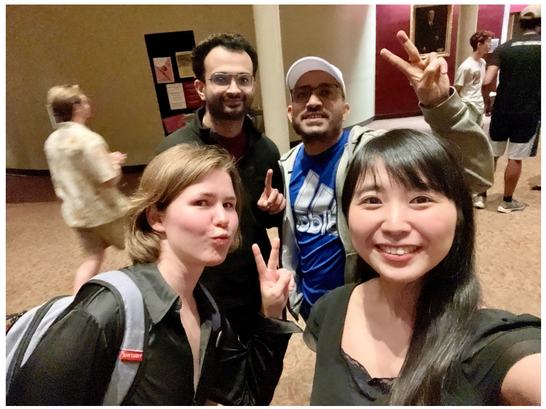
Laura K. Reed
Paige F. Ferguson
Kevin M. Kocot
Michael R. McKain
John H. Yoder
Réka Albert *ext.*

Daryl Lam
Wilson Leung
Samuel Sledzieski
Faith Ocitti
Joshua Lotfi
Richard Adkins

Jenna M. Poole
Olivia F. Raper
Seema Soti

Reed Lab Members
Logan T. Cohen
Tolulope R. Kolapo-Mabayoje

Sahar Abdollahi
Bishnu Adhikari
David “Kyle” Breault



Anna Parul
Torin A. Alter
German A. Lopez Otero
Hrishikesh S. Tupkar

Friends and Family

AA ANM BNG DH
DMP DP EM ER GK
HA HG HN ID IN *JC*
JV LAJ MH MKW
MM MS NR PC PS
RJ RN SA SK SP SS

NCTTA/USATT/Senior Games

fin