Ancient gene linkages support ctenophores as sister to other animals

A review and reproducibility study by: Samantha Finkbeiner, Hector Benitez, Curie Cha, and Dhruv Rokkam

BINF6310 Fall 2024

Background – Evolutionary Question

- Who diverged first from other animals: Sponges or ctenophores?
  - Important for neural and muscular system origins

- 2 competing hypotheses:
  - Sponge-Sister hypothesis
    - Neurons evolved once
  - Ctenophore-Sister hypotheses
    - Independent evolution or neuron loss in sponges

Background – Novel Analytical Approach

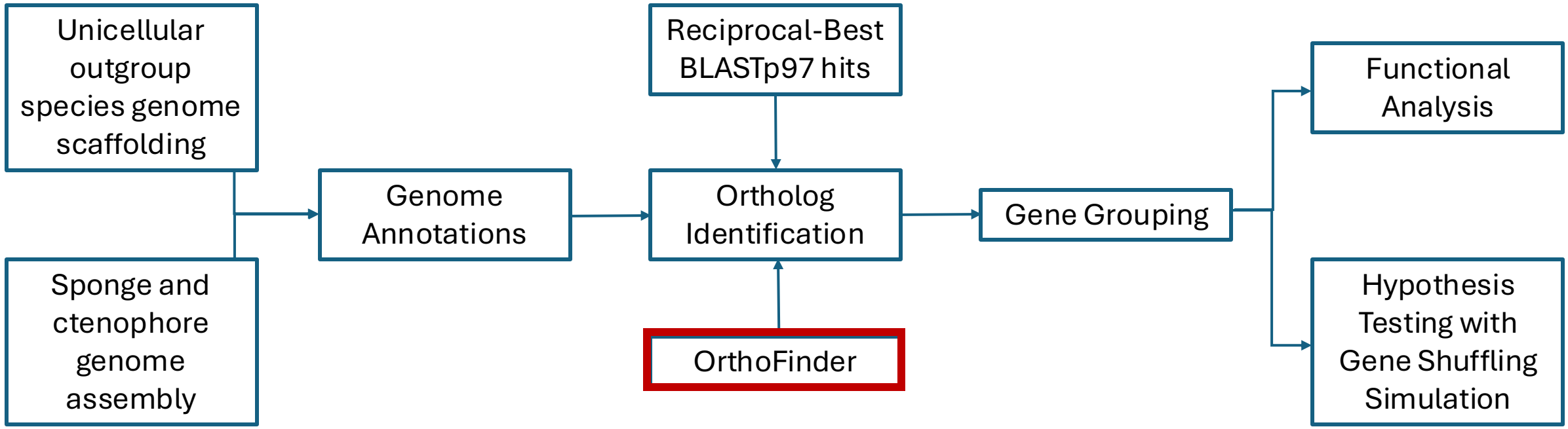
- What is Synteny?
  - Patterns of conserved chromosomal gene linkages
  - Evolve slowly and included irreversible changes
  - Synteny analysis provides robust phylogenetic markers

- Broad Study Methodology
  - Comparative analysis of chromosome-scale genomes:
    - Ctenophores
    - Sponges
    - Unicellular relatives

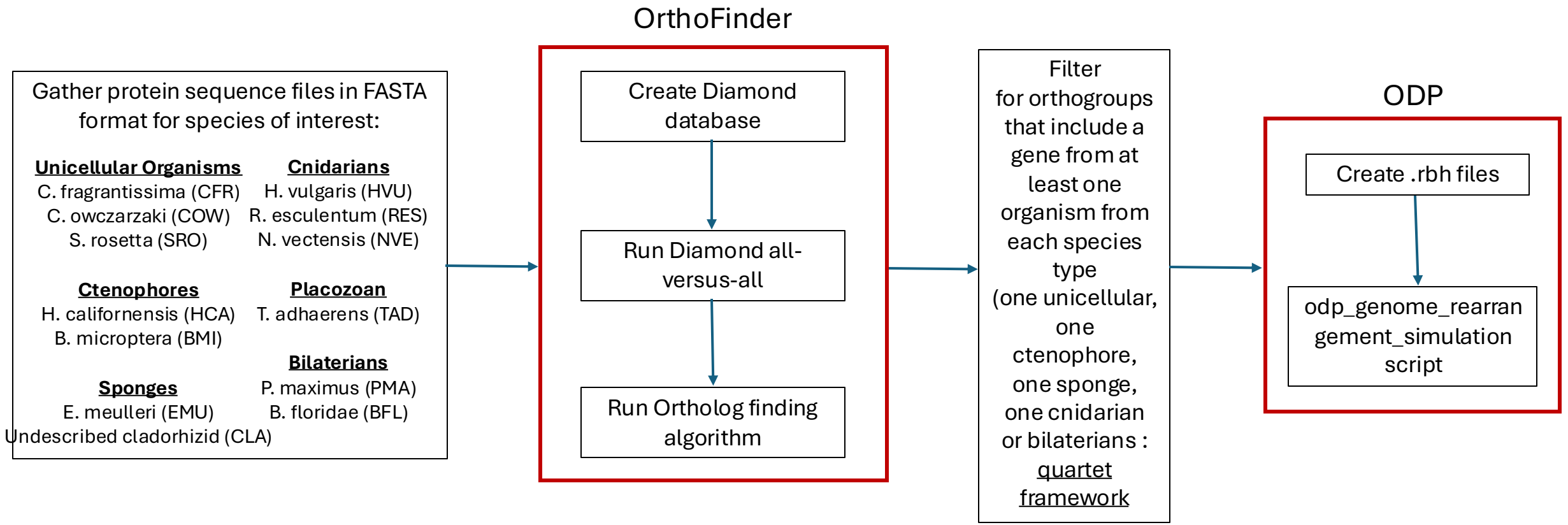
Findings and Broader Implications

- Key Findings:
  - Conserved synteny patterns link sponges, bilaterians, and cnidarians
  - Supports the Ctenophore-Sister Hypothesis
    - Suggesting that ctenophores diverged first and that neural systems evolved after this divergence.
  - What does this mean for Evolutionary Biology?
    - Synteny offers new ways to resolve phylogenetic controversies
    - Allows for greater understanding of evolution of early animal traits

Methods


Unicellular outgroup species genome scaffolding
Sponge and ctenophore genome assembly
Genome Annotations
Reciprocal-Best BLASTp97 hits
Ortholog Identification
OrthoFinder
Gene Grouping
Functional Analysis
Hypothesis Testing with Gene Shuffling Simulation

Ortholog Identification Pipeline



OrthoFinder Terminal Output

```
(binf6310) [cha.c@c0747 final_project]$ orthofinder -f Proteomes -og -t 8
```

```
Dividing up work for BLAST for parallel processing
--------------------------------------------------
2024-12-11 20:35:16 : Creating diamond database 1 of 13
2024-12-11 20:35:16 : Creating diamond database 2 of 13
2024-12-11 20:35:17 : Creating diamond database 3 of 13
2024-12-11 20:35:17 : Creating diamond database 4 of 13
2024-12-11 20:35:17 : Creating diamond database 5 of 13
2024-12-11 20:35:18 : Creating diamond database 6 of 13
2024-12-11 20:35:18 : Creating diamond database 7 of 13
2024-12-11 20:35:19 : Creating diamond database 8 of 13
2024-12-11 20:35:19 : Creating diamond database 9 of 13
2024-12-11 20:35:19 : Creating diamond database 10 of 13
2024-12-11 20:35:20 : Creating diamond database 11 of 13
2024-12-11 20:35:20 : Creating diamond database 12 of 13
2024-12-11 20:35:21 : Creating diamond database 13 of 13
```

```
Running diamond all-versus-all
------------------------------
Using 8 thread(s)
2024-12-11 20:35:21 : This may take some time....
2024-12-11 20:35:21 : Done 0 of 169
2024-12-11 20:48:27 : Done 10 of 169
2024-12-11 21:08:07 : Done 20 of 169
2024-12-11 21:16:36 : Done 30 of 169
2024-12-11 21:22:02 : Done 40 of 169
2024-12-11 21:25:24 : Done 50 of 169
2024-12-11 21:28:11 : Done 60 of 169
2024-12-11 21:32:13 : Done 70 of 169
2024-12-11 21:35:16 : Done 80 of 169
2024-12-11 21:38:41 : Done 90 of 169
2024-12-11 21:40:55 : Done 100 of 169
2024-12-11 21:43:40 : Done 110 of 169
2024-12-11 21:46:07 : Done 120 of 169
2024-12-11 21:47:50 : Done 130 of 169
2024-12-11 21:49:38 : Done 140 of 169
2024-12-11 21:51:08 : Done 150 of 169
2024-12-11 21:52:36 : Done 160 of 169
2024-12-11 21:54:19 : Done all-versus-all sequence search
```

```
Running OrthoFinder algorithm
-----------------------------
2024-12-11 21:54:20 : Initial processing of each species
2024-12-11 21:54:41 : Initial processing of species 0 complete
2024-12-11 21:54:59 : Initial processing of species 1 complete
2024-12-11 21:55:05 : Initial processing of species 2 complete
2024-12-11 21:55:33 : Initial processing of species 3 complete
2024-12-11 21:55:39 : Initial processing of species 4 complete
2024-12-11 21:56:12 : Initial processing of species 5 complete
2024-12-11 21:56:22 : Initial processing of species 6 complete
2024-12-11 21:56:44 : Initial processing of species 7 complete
2024-12-11 21:56:55 : Initial processing of species 8 complete
2024-12-11 21:57:24 : Initial processing of species 9 complete
2024-12-11 21:57:35 : Initial processing of species 10 complete
2024-12-11 21:57:43 : Initial processing of species 11 complete
2024-12-11 21:57:51 : Initial processing of species 12 complete
2024-12-11 21:58:12 : Connected putative homologues
2024-12-11 21:58:17 : Written final scores for species 0 to graph file
2024-12-11 21:58:21 : Written final scores for species 1 to graph file
2024-12-11 21:58:22 : Written final scores for species 2 to graph file
2024-12-11 21:58:28 : Written final scores for species 3 to graph file
2024-12-11 21:58:30 : Written final scores for species 4 to graph file
2024-12-11 21:58:37 : Written final scores for species 5 to graph file
2024-12-11 21:58:39 : Written final scores for species 6 to graph file
2024-12-11 21:58:45 : Written final scores for species 7 to graph file
2024-12-11 21:58:48 : Written final scores for species 8 to graph file
2024-12-11 21:58:55 : Written final scores for species 9 to graph file
2024-12-11 21:58:58 : Written final scores for species 10 to graph file
2024-12-11 21:59:00 : Written final scores for species 11 to graph file
2024-12-11 21:59:02 : Written final scores for species 12 to graph file
```

Filtering Orthogroups Gathered from OrthoFinder

- The OrthoFinder generated Orthogroups Gene Count file was imported as a Pandas DataFrame.

```
[1]: import pandas as pd
```

```
[13]: orthogroup_gene_count = pd.read_csv('/Users/curiecha/Desktop/OrthoFinder/Results_Dec11/Orthogroups/Orthogroups.GeneCount.tsv', sep='\t')
```

```
[14]: orthogroup_gene_count
```

[14]:

	Orthogroup	BFL	BMI	CFR	CLA	COW	EMU	HCA	HVU	NVE	PMA	RES	SRO	TAD	Total
0	OG0000000	0	0	0	0	0	0	0	752	0	0	1	0	0	753
1	OG0000001	0	45	0	228	0	269	2	37	25	22	1	0	0	629
2	OG0000002	7	2	1	228	36	291	13	2	2	15	10	19	1	627
3	OG0000003	1	0	0	0	0	0	0	549	0	3	0	0	0	553
4	OG0000004	1	0	0	149	1	172	0	1	0	36	14	0	1	375
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
37852	OG0037852	0	0	0	0	0	0	0	0	0	0	0	0	2	2
37853	OG0037853	0	0	0	0	0	0	0	0	0	0	0	0	2	2
37854	OG0037854	0	0	0	0	0	0	0	0	0	0	0	0	2	2
37855	OG0037855	0	0	0	0	0	0	0	0	0	0	0	0	2	2
37856	OG0037856	0	0	0	0	0	0	0	0	0	0	0	0	2	2

37857 rows × 15 columns

Filtering Orthogroups Gathered from OrthoFinder

- The filtering criteria was implented in a Python script as conditionals and applied to the orthogroup gene count DataFrame to filter the groups.

[21]:
```
# (COW | SRO) & (HCA | BMI) & (EMU | CLA) & (RES | NVE | BFL | PMA)

mask1 = (orthogroup_gene_count['COW'] >= 1) | (orthogroup_gene_count['SRO'] >= 1)
mask2 = (orthogroup_gene_count['HCA'] >= 1) | (orthogroup_gene_count['BMI'] >= 1)
mask3 = (orthogroup_gene_count['EMU'] >= 1) | (orthogroup_gene_count['CLA'] >= 1)
mask4 = (orthogroup_gene_count['RES'] >= 1) | (orthogroup_gene_count['NVE'] >= 1) | (orthogroup_gene_count['BFL'] >= 1) | (orthogroup_gene_cou

final_mask = mask1 & mask2 & mask3 & mask4
filtered_df = orthogroup_gene_count[final_mask]
print(filtered_df)
```

```
      Orthogroup  BFL  BMI  CFR  CLA  COW  EMU  HCA  HVU  NVE  PMA  RES  SRO  \
2      OG0000002    7    2    1  228   36  291   13    2    2   15   10   19
5      OG0000005   59   44    2   11    3   15   15   32   58   71   40    0
7      OG0000007   65   29    2   11    0   11   21   24   34   48   21    2
9      OG0000009   42   19    2   27   24   31   11   32   18   24   16   14
15     OG0000015   29   12    1    6    1    8   12   47   42   28   42    1
...          ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
9522   OG0009522    0    1    1    1    1    1    0    0    0    1    0    0
9528   OG0009528    0    1    0    2    1    2    0    0    0    0    1    0
9761   OG0009761    0    0    0    1    1    1    1    0    0    1    1    0
10223  OG0010223    1    1    0    1    1    1    0    0    0    0    0    0
10472  OG0010472    0    1    0    1    0    1    1    0    0    1    0    1

       TAD  Total
2        1    627
5       21    371
7       11    279
9        5    265
15       7    236
...    ...    ...
9522     1      7
9528     0      7
9761     1      7
10223    1      6
10472    0      6

[3645 rows x 15 columns]
```

Analyzing Filtered Orthogroups

- Gathered from previous methodological reciprocal best hit step, the A1a, B1, C1, Ea, F, G, L, and N ancestral linkage groups were used to analyze the filtered orthogroups.
- For each orthogroup, the count of genes of each linkage group was found and aggregated across the groups to get the total count of relevant genes for each group.
- For all 32 possible combinations of quartets from the criteria, .rbh files were created (via script from the authors) to represent recipricol best hits.
- Using the .rbh files and the aggregated gene count values, the provided odp_genome_rearrangement_simulation script was run to get the genes that support the sponge sister hypothesis as well as the Ctenophore sister hypothesis.
- The genes were used to get the associated quartets that have these genes.

Results – Orthofinder analysis

Reproduced results

- Our results revealed a total of 37,857 orthogroups.
- The authors' Orthofinder analysis resulted in a total of 35,102 orthogroups.
- For both our reproduction and the authors' results, these orthogroups are not guaranteed to contain a gene from all of the species.

Statistics_Overall

Number of species	**13**			
Number of genes	277315			
Number of genes in orthogroups	253700			
Number of unassigned genes	23615			
Percentage of genes in orthogroups	91.5			
Percentage of unassigned genes	8.5			
Number of orthogroups	37857			
Number of species-specific orthogroups	7940			
Number of genes in species-specific orthogroups	41264			
Percentage of genes in species-specific orthogroups	14.9			
Mean orthogroup size	6.7			
Median orthogroup size	3.0			
G50 (assigned genes)	13			
G50 (all genes)	12			
O50 (assigned genes)	4807			
O50 (all genes)	5731			
Number of orthogroups with all species present	1644			
Number of single-copy orthogroups	235			
Date	2024-12-11			
Orthogroups file	Orthogroups.tsv			
Unassigned genes file	Orthogroups_UnassignedGenes.tsv			
Per-species statistics	Statistics_PerSpecies.tsv			
Overall statistics	Statistics_Overall.tsv			
Orthogroups shared between species	Orthogroups_SpeciesOverlaps.tsv			

Statistics_Overall.tsv file generated from OrthoFinder program

Results – Filtering

- To address this, the Orthogroups were then filtered down to only include at least one of every unicellular species(i.e at least one ctenephore, sponge, cnidarain, and bilaterian)
- After filtering, we found 3,645 orthogroups that fit this criteria.
- The authors found 3,746 orthogroups that fit this criteria.

Reproduced results

[21]:
```
# (COW | SRO) & (HCA | BMI) & (EMU | CLA) & (RES | NVE | BFL | PMA)

mask1 = (orthogroup_gene_count['COW'] >= 1) | (orthogroup_gene_count['SRO'] >= 1)
mask2 = (orthogroup_gene_count['HCA'] >= 1) | (orthogroup_gene_count['BMI'] >= 1)
mask3 = (orthogroup_gene_count['EMU'] >= 1) | (orthogroup_gene_count['CLA'] >= 1)
mask4 = (orthogroup_gene_count['RES'] >= 1) | (orthogroup_gene_count['NVE'] >= 1)

final_mask = mask1 & mask2 & mask3 & mask4
filtered_df = orthogroup_gene_count[final_mask]
print(filtered_df)
```

```
      Orthogroup  BFL  BMI  CFR  CLA  COW  EMU  HCA  HVU  NVE  PMA  RES  SRO  \
2      OG0000002    7    2    1  228   36  291   13    2    2   15   10   19
5      OG0000005   59   44    2   11    3   15   15   32   58   71   40    0
7      OG0000007   65   29    2   11    0   11   21   24   34   48   21    2
9      OG0000009   42   19    2   27   24   31   11   32   18   24   16   14
15     OG0000015   29   12    1    6    1    8   12   47   42   28   42    1
...          ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
9522   OG0009522    0    1    1    1    1    1    0    0    0    1    0    0
9528   OG0009528    0    1    0    2    1    2    0    0    0    0    1    0
9761   OG0009761    0    0    0    1    1    1    1    0    0    1    1    0
10223  OG0010223    1    1    0    1    1    1    0    0    0    0    0    0
10472  OG0010472    0    1    0    1    0    1    1    0    0    1    0    1

       TAD  Total
2        1    627
5       21    371
7       11    279
9        5    265
15       7    236
...    ...    ...
9522     1      7
9528     0      7
9761     1      7
10223    1      6
10472    0      6

[3645 rows x 15 columns]
```

Results – Orthofinder analysis

- Upon running the odp_genome_rearrangement_simulation script with the generated .rbh files and the filtered orthogroups, 146 orthologs were found to support the Ctenophore sister hypothesis, while 11 were found to support the sponge sister hypothesis.
- This is accurate with the results obtained by authors.

Reproduced results

```
(base) curiecha@C
SRO-HCA-CLA-PMA
0
SRO-BIN-CLA-BFL
0
SRO-BIN-CLA-RES
0
COW-BIN-CLA-RES
0
SRO_HCA_CLA_BFL
0
SRO_BIN_CLA_NVE
0
COW_HCA_CLA_RES
11
SRO_HCA_CLA_NVE
0
COW_BIN_CLA_NVE
0
COW_BIN_CLA_BFL
0
SRO_HCA_CLA_RES
0
COW_HCA_CLA_PMA
11
SRO_BIN_CLA_PMA
0
COW_HCA_CLA_BFL
0
COW_HCA_CLA_NVE
11
COW_BIN_CLA_PMA
0
SRO-HCA-EMU-PMA
0
```

```
SRO-BIN-EMU-BFL
0
SRO-BIN-EMU-RES
0
COW-BIN-EMU-RES
0
SRO_HCA_EMU_BFL
0
SRO_BIN_EMU_NVE
0
COW_HCA_EMU_RES
0
SRO_HCA_EMU_NVE
0
COW_BIN_EMU_NVE
0
COW_BIN_EMU_BFL
0
SRO_HCA_EMU_RES
0
COW_HCA_EMU_PMA
0
SRO_BIN_EMU_PMA
0
COW_HCA_EMU_BFL
0
COW_HCA_EMU_NVE
0
COW_BIN_EMU_PMA
0
```

Unicell.	Cteno.	Sponge	Cnid. or Bilat.	A1a		B1		C1		Ea		F		G		L		N		Total Gene Count	Genes supporting Sponge-sister
				_x	_y	_x	_y	_x	_y	_x	_y	_x	_y	_x	_y	_x	_y	_x	_y		
COW	BMI	CLA	BFL	9	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14	-
COW	BMI	CLA	NVE	9	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14	-
COW	BMI	CLA	PMA	10	5	-	-	-	-	-	-	-	-	7	5	-	-	-	-	27	-
COW	BMI	CLA	RES	10	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-
COW	BMI	EMU	BFL	11	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18	-
COW	BMI	EMU	NVE	11	6	5	5	-	-	-	-	-	-	-	-	-	-	-	-	27	-
COW	BMI	EMU	PMA	12	7	5	5	-	-	-	-	8	5	5	5	-	-	-	-	52	-
COW	BMI	EMU	RES	11	6	5	5	-	-	-	-	-	-	-	-	-	-	-	-	27	-
COW	HCA	CLA	BFL	10	6	-	-	3	7	-	-	-	-	-	-	-	-	-	-	28	-
COW	HCA	CLA	NVE	9	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	11
COW	HCA	CLA	PMA	10	6	-	-	3	8	-	-	8	5	9	5	-	-	-	-	56	11
COW	HCA	CLA	RES	10	6	-	-	3	7	-	-	7	5	-	-	-	-	-	-	40	11
COW	HCA	EMU	BFL	13	8	-	-	5	7	-	-	8	5	-	-	-	-	-	-	48	-
COW	HCA	EMU	NVE	12	7	5	6	-	-	7	5	7	5	-	-	-	-	-	-	54	-
COW	HCA	EMU	PMA	13	8	5	6	4	8	6	5	8	6	7	6	-	-	-	-	84	-
COW	HCA	EMU	RES	12	7	5	6	4	7	-	-	7	6	8	5	-	-	-	-	69	-
SRO	BMI	CLA	BFL	8	6	-	-	-	-	-	-	-	-	6	6	5	5	-	-	36	-
SRO	BMI	CLA	NVE	8	7	-	-	-	-	-	-	-	-	7	6	5	5	-	-	38	-
SRO	BMI	CLA	PMA	7	6	-	-	-	-	-	-	-	-	7	6	5	5	-	-	36	-
SRO	BMI	CLA	RES	8	5	-	-	-	-	-	-	-	-	7	6	5	5	-	-	36	-
SRO	BMI	EMU	BFL	11	6	-	-	-	-	-	-	-	-	5	6	-	-	-	-	28	-
SRO	BMI	EMU	NVE	11	7	-	-	-	-	-	-	-	-	6	7	5	6	-	-	42	-
SRO	BMI	EMU	PMA	10	6	-	-	-	-	-	-	-	-	6	7	5	6	-	-	40	-
SRO	BMI	EMU	RES	10	5	-	-	-	-	-	-	-	-	6	7	5	6	-	-	39	-
SRO	HCA	CLA	BFL	9	6	-	-	-	-	-	-	-	-	7	5	-	-	5	6	38	-
SRO	HCA	CLA	NVE	9	7	-	-	-	-	-	-	-	-	8	5	5	6	5	5	50	-
SRO	HCA	CLA	PMA	8	6	-	-	-	-	-	-	-	-	8	5	5	6	5	6	49	-
SRO	HCA	CLA	RES	9	5	-	-	-	-	-	-	-	-	7	5	5	5	5	6	47	-
SRO	HCA	EMU	BFL	12	6	-	-	-	-	-	-	-	-	6	7	-	-	5	7	43	-
SRO	HCA	EMU	NVE	12	7	-	-	-	-	-	-	-	-	7	8	5	7	5	6	57	-
SRO	HCA	EMU	PMA	11	6	-	-	-	-	-	-	-	-	7	8	5	7	5	7	56	-
SRO	HCA	EMU	RES	11	5	-	-	-	-	-	-	-	-	6	8	5	6	5	7	53	-
Unique Counts of this column				20	13	5	6	5	8	7	5	8	7	15	15	5	12	6	7	146	11

Schultz, et al.

Final Results

	Our Results	Authors' Results
Total Orthogroup Count	37,857	35,102
Filtered Orthogroup Count	3,645	3,746
Orthologs Supporting Ctenophore Sister Hypothesis	146	146
Orthologs Supporting Sponge Sister Hypothesis	11	11

Discussion

Reproduced Results:

- 146 orthologs identified support the Ctenophore-Sister hypothesis, compared to only 11 for the Sponge-Sister hypothesis.
- Minor deviations in orthogroup counts due to software version differences and algorithm updates.

Challenges and Methodological Limitations:

- Software dependencies and updates impacted reproducibility.
- Variability in orthogroup identification algorithms introduced minor inconsistencies.
- Lack of standardized computational environments may have contributed to discrepancies.

Implications for Phylogenetics:

- Synteny provides robust markers for resolving phylogenetic relationships.
- Highlights the importance of chromosomal linkages in understanding evolutionary history.
- Supports the independent evolution or loss of neural systems in early animal lineages.

Overall Insights

Value of Reproducibility:

- Validates the reliability of novel methodologies.
- Identifies areas for standardization in bioinformatics pipelines.
- Promotes transparency and rigor in evolutionary biology research.

Future Directions:

- Develop standardized computational frameworks for reproducibility.
- Investigate synteny in other major animal lineages to test broader applicability.
- Explore the functional roles of conserved genes in early neural evolution.

Retrospective analysis

- The reproduced results of the OrthoFinder pipeline were marginally different in comparison to the study of interest.

	Our Results	Authors' Results
Total Orthogroup Count	37,857	35,102
Filtered Orthogroup Count	3,645	3,746
Orthologs Supporting Ctenophore Sister Hypothesis	146	146
Orthologs Supporting Sponge Sister Hypothesis	11	11

- Upon reflection, these differences may be attributed to the version of OrthoFinder and its dependencies that were used for the reproduction.
  - Updates and differences in clustering and orthogroup finding algorithms.

Citations

Schultz, D. T., Haddock, S. H. D., Bredeson, J. V., Green, R. E., Simakov, O., & Rokhsar, D. S. (2023). Ancient gene linkages support ctenophores as sister to other animals. *Nature*, *618*(7963), 110-117. https://doi.org/10.1038/s41586-023-05936-6