Computational methods to investigate the origins of bioluminescence in a charismatic non-model system: California Sea Firefly

Gigi Minsky, Lisa Mesrop, Michael Drummond, Jessica Goodheart, Todd Oakley
University of California, Santa Barbara, EEMB

Introduction

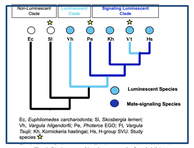


Fig. 1, Phylogeny of luminescence in Cyprinidid ostracods

Fig. 2 Bioluminescent Ostracod, Elliot Lowndes

- **Bioluminescence** is the production of light in living organisms via the reaction of enzyme luciferase and substrate luciferin
- Although bioluminescence biochemistry has been extensively studied, **the genes comprising the bioluminescent pathway, are largely unknown.**
- There is a single evolutionary point of origin in Cyprinidid ostracods (Fig 1*)* which makes them a unique model to study as all species have a similar bioluminescent "toolkit".
- We successfully reared and cultured bioluminescent ostracod *V. tsujii* in the lab, a first for any luminous organism. **This culturing allows us to better investigate the evolutionary origins of bioluminescence.**
- We propose to use **a WGCNA (weighted gene co-expression network analysis)** based approach to **identify genes responsible for the production of luciferin and elucidate the luciferin metabolic pathway.**

Background and Research Aim

We hypothesize that genes responsible for the synthesis of luciferin, are co-regulated with the luciferase gene.

Luciferin/Luciferase Biochemistry

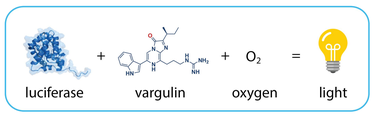


Fig. 3, Luciferase + ostracod luciferin (vargulin) + oxygen reaction to produce bioluminescence

What is a biosynthetic pathway and why are we interested in the one for bioluminescence?

- A biosynthetic pathway is a metabolic sequence of molecular/enzymatic events that lead to a byproduct in living organisms
- Bioluminescence is novel trait that has evolved independently over eighty times across species (Fig 3).

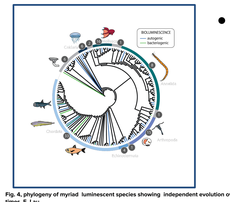
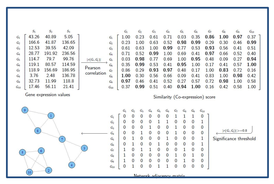


Fig. 4, phylogeny of myriad luminescent species showing independent evolution over eighty times. E. Lau

- **The biosynthetic pathway of luciferin is not well known in any species. Characterizing the biosynthetic pathway of luciferin leads us closer to understanding the relationship between biological design and function of luminescence.**

How Does WGCNA help us understand the biosynthetic pathway?

- WGCNA, a data mining package in R, produces groups or "modules" of genes involved in other molecular functions that are highly co-expressed with putative genes like luciferase, potentially crucial to the biosynthesis of the byproduct of interest luciferin,.



- This sheds light on the modular steps involved in the overall bioproduction of luminescence.

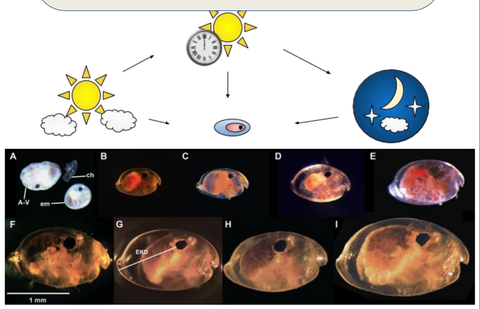
Fig. 5, Generalized workflow of a weighted gene co-expression network analysis (source)

Methods

Wet Lab

RNeasy kits from QIAGEN were utilized to extract and isolate transcriptomes from both the tissue of the various stages of ostracod instar development and of the light secreting upper lip organ.

The extractions from the instars were performed at three time points with animals induced to luminesce and non induced in order to add varying experimental conditions to the samples in the n X M gene sample matrix (shown under *in Silica* methods)



Figs 6 & 7 : 6) the three time points morning (8 am) noon and evening (6 pm) that the transcriptomes were extracted 7) stages of development of *V.tsujii*

in Silica

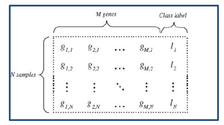


Fig 8 : general skeleton of a n by M gene expression matrix

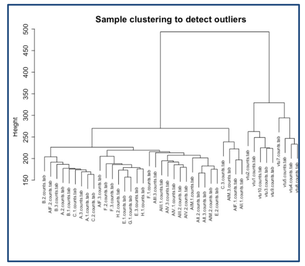


Fig 9, clustering of samples from WGCNA pipeline to detect Sample outliers; shows upper lip tissue sample clustering together as the extracted tissue for those samples were all from the same organ

Data Prep/Formatting

The transcriptomes were sequenced and reformatted with the TagSeq pipeline developed in PERL into an n X M gene matrix (n samples X M genes) with the expression counts as the matrix values.

The expression counts were then normalized; based on those expression values we can see how **similar** the **samples** are before clustering the genes via calculating the Euclidean distance in principal component analysis space (Fig, 9)

Network Analysis/Module Creation

A single block and multi-block (blockwise) network analysis was run on the gene expression matrix in our optimized WGCNA pipeline.

The **single block network analysis** takes the entire gene matrix and transforms it into a **signed** similarity matrix. The **signed** matrix considers both positive and negative correlations. A network of modules, or groups of highly correlated/co-expressed genes is then created.

A **blockwise network analysis** was also conducted; blockwise network analyses are useful for large datasets and low computing power; the initial normalized gene expression matrix is broken up into "blocks" of specified amounts of genes randomly, and the same transformation processes as the single block analysis are carried out on each of the blocks including a network analyses for each individual block. After this, the algorithm produces gene modules that are closely related from each network using topological overlap.

Network Analysis Conclusion

The results of these analyses were similar- the single network analysis found 123 modules and the blockwise found 97, but luciferase and its relatives were found in the same module for both analyses (Fig 13,) which the algorithm labels turquoise.

The tree dendrograms are displayed with each of the modules labeled by a color on the bottom, and the branches represent a distance between genes representing similarity. Fig 12 represents all the modules from the single block analysis while Fig 13 represents just one module from the blockwise analysis

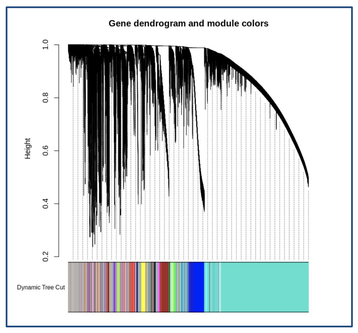


Fig 12, Singe block analysis dendrogram

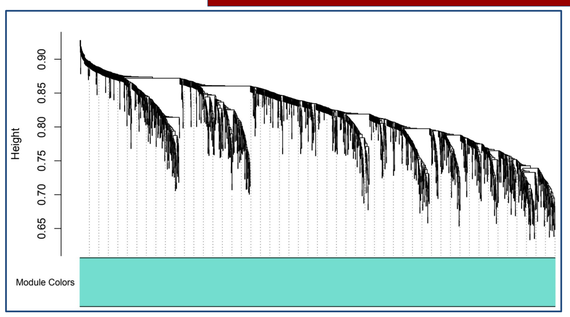


Fig 13, Turquoise module (where luciferase is) dendrogram from blockwise analysis

Results/Discussion

Finding the module that contains luciferase

The turquoise module (Fig 13) contained geneID Node_10321 (highlighted by the yellow star) shown to be luciferase based on the MAFFT sequence alignment (Fig 14). In this alignment we used the known sequence of *Vargula tsujii* luciferase and aligned it to the *Vargula tsujii* transcriptome and found a few closely related genes. In addition to luciferase, we used these paralogs as additional 'bait' sequences to find candidate genes in the luciferin pathway. These genes were also found in the turquoise module (Fig 14)

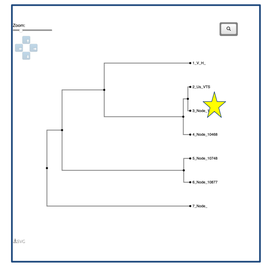


Fig 14, MAFFT sequence alignment showing luciferase and closely related (phylogenetically) genes to luciferase. Node_10321 from the expression matrix aligns closest to the known sequence of luciferase in *tsujii* . The other geneIDs are paralogs.

Network graph illustrating the number and strength of connections within the turquoise module. The luciferase gene is highlighted by the yellow dot. We found 36 genes connected to luciferase and this included VWD domain proteins and toxin-like peptides.'

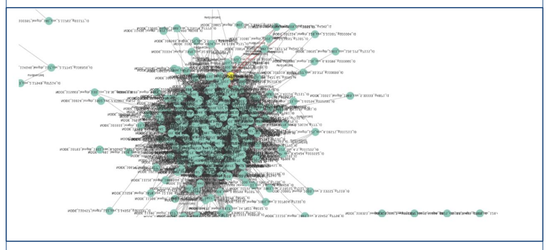


Fig 15, cytoscape visualization of the turquoise module, respective to luciferase

Future Directions

-Functional Enrichment
Calculates proportion of genes in a module that play a role in a given molecular or biological function

-CRISPR Validation
Knockout genes we hypothesize are in the luciferin pathway to see if the luciferin substrate is affected

-Consensus Networks
Better highlight modules across multiple data sets, upper lip vs whole animal tissue

Contact Information

Geetanjali "Gigi" Minsky
UCSB, CCS Biology
617-620-3287
ggminsky@ucsb.edu

References

1. Fig, 8 Dey, Lopamudra & Mukhopadhyay, Anirban. (2020). Microarray Gene Expression Data Clustering using PSO based K-means Algorithm. International Journal of Computer Science and its Applications. 232.
2. Fig 5, creative commons licensed image
3. Figs 3 and 4, Emily Lau
4. Fig 1, Jessica Goodheart

Acknowledgements

Emily Lau
Nikolai Hensley
Dr. Steve Horvath and Dr. Peter Langfelder