# Comparative transcriptomics of *Oenothera*

Jun 26, 2022

Eunice Kariñho-Betancourt[1], David Carlson[2]

[1]Department of Biology, University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada;

[2]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA

1 Works for me	Share

dx.doi.org/10.17504/protocols.io.bwnzpdf6

Eunice Kariñho-Betancourt

ABSTRACT

We present the comparative transcriptomics protocol employed for evolutionary analysis in the genus Oenothera. We examined genome-wide patterns and functional diversification by searching for orthologous genes and employed phylogenetic inference methods to run gene family evolutionary analysis.

DOI

dx.doi.org/10.17504/protocols.io.bwnzpdf6

EXTERNAL LINK

https://doi.org/10.1371/journal.pone.0269307

PROTOCOL CITATION

Eunice Kariñho-Betancourt, David Carlson 2022. Comparative transcriptomics of Oenothera. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.bwnzpdf6

FUNDERS ACKNOWLEDGEMENT

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Kariñho-Betancourt E, Carlson D, Hollister J, Fischer A, Greiner S, Johnson MTJ (2022) The evolution of multi-gene families and metabolic pathways in the evening primroses (*Oenothera*: Onagraceae): A comparative transcriptomics approach. PLoS ONE 17(6): e0269307. doi: 10.1371/journal.pone.0269307

CREATED

Jul 17, 2021

LAST MODIFIED

Jun 26, 2022

PROTOCOL INTEGER ID

51641

Orthology, phylogenomics and gene family evolution

1 Transcriptome assembly and functional annotation

RNA-seq reads of 32 *Oenothera* taxa (63 samples that included replicates for some species)
were trimmed and filtered for quality using fastp v.20.0, set to an average PHRED quality score
of 20. We employed Trinity v2.11.0 for *de novo* assembly of 63 transcriptomes using default
settings, except CPU and memory allocation. Each assembly was functionally annotated and
enriched using Trinoatate v3.2.1. We used blastp v2.5.0 (e-value of 1e-5) against the Swiss-
Prot, Pfam SignalIP and TMHMM databases to transcriptome annotation (each set of the
transcriptome coding sequences was previously translated into amino acid sequences using
TransDecoder v5.5.0).For every *Trinity gene*, we retained the longest isoform for downstream
analysis.

2 Ortholog identification and phylogenetic inference

We employed the Orthofinder program v2.4.0 to construct orthologs for a subset of 30 taxa (30
transcriptomes). The program allows to identify orthologous and paralogous genes, and
identify orthogroups, ie., gene families of conserved protein domains (which feed the
subsequent evolutionary analysis). At the same time, the program can run multiple alignments
for gene and species phylogenetic inference. Orthofinder was run with DIAMOND v.0.9.14, for
protein alignment, we used MAFFT v7.471 and IQ-TREE v.2.0.3 for gene tree inference. The
analysis was conducted using 28 threads and run on the Stony Brook University SeaWulf HPC
cluster using the Slurm workload manager.
Note that each proteome of each species must be in fasta format

We build a concatenated phylogeny using a subset of 1,017 orthogroups of single-copy genes,
which were aligned using MAFFT v7.471, and IQ-TREE v.2.0.3 for maximum likelihood
phylogenetic inference. Node confidence was assessed using 1000 ultrafast bootstrap
replicates by specifying the '-B 1000 -bnni' flags.

3   Constructing an ultrametric tree

The ultrametric tree (time-calibrated) can be done with diverse programs such as the R package Ape and its Cronos function. We employed the r8s v1.8.1 module of the Computational Analysis of gene Family Evolution (CAFE) program for making the ML tree of single-copy genes ultrametric.

4   Gene family analysis

We employed the CAFE programs to estimate protein family expansion and contractions. CAFE needs two files to run; the ultrametric species tree (netwick format) and the gene count of each orthogroup, which is an output from Orthofinder (Orthogroups.GeneCount.cvs). From the last file the column TOTAL needs to be removed and a first column with the header "descriptions" should be added. Columns on this file should be separated by TAB. Before running the program the GeneCount database needs to be filtered. Gene families with lower counts (e.g. < three genes per protein family) should be removed. In addition, a parameter model for lambda should be provided. The parameter model represents the species topology without branch lengths.

This is the example of the input files for CAFE

#specify data file, p-value threshold, # of threads to use, and log file
load -i 23526_subset.tab -filter -t 1 -l log_run_scriptCAFE_filter.txt

#*the phylogenetic tree structure with branch lengths*
Tree (((((filiformis:8.9,gaura:8.9):173.8,suffulta:182.7):531.3,
(rosea:506.4,speciosa:506.4):207.7):252.3,
((((argillicola:75.8,parviflora:75.8):30.4,oakesiana:106.2):34.1,
(((grandiflora:4.4,nutans:4.4):1,biennis:5.5):100.2,((((((elataajbk:52.1,wolfii:52.1):9,
((elataoeee:43.9,villosa:43.9):10.2,longissima:54.1):7):4.8,
(hirsuti:55.8,jamesii:55.8):10.2):14.6,glazioviana:80.7):12.9,stuchii:93.6):12.1):34.6):109.8,
(((clelandii:90.1,rhombipetala:90.1):90.7,(grandis:12,laciniata:12):168.9):44.9,
(((longituba:75.4,nana:75.4):27.2,villaricae:102.6):19.5,
(affinis:69.1,picensis:69.1):53.1):103.6):24.3):716.2):524.2,berlandieri:1490.8)

#*search for 1 parameter model*
lambda -s -t (((((((1,1)1,1)1,(1,1)1)1,((((1,1)1,1)1,(((1,1)1,1)1,(((((1,1)1,((1,1)1,1)1)1,
(1,1)1)1,1)1,1)1)1)1,(((1,1)1,(1,1)1)1,(((1,1)1,1)1,(1,1)1)1)1)1,1)

3   Constructing an ultrametric tree

The ultrametric tree (time-calibrated) can be done with diverse programs such as the R package Ape and its Cronos function. We employed the r8s v1.8.1 module of the Computational Analysis of gene Family Evolution (CAFE) program for making the ML tree of single-copy genes ultrametric.

4   Gene family analysis

We employed the CAFE programs to estimate protein family expansion and contractions. CAFE needs two files to run; the ultrametric species tree (netwick format) and the gene count of each orthogroup, which is an output from Orthofinder (Orthogroups.GeneCount.cvs). From the last file the column TOTAL needs to be removed and a first column with the header "descriptions" should be added. Columns on this file should be separated by TAB. Before running the program the GeneCount database needs to be filtered. Gene families with lower counts (e.g. < three genes per protein family) should be removed. In addition, a parameter model for lambda should be provided. The parameter model represents the species topology without branch lengths.

This is the example of the input files for CAFE

#specify data file, p-value threshold, # of threads to use, and log file
load -i 23526_subset.tab -filter -t 1 -l log_run_scriptCAFE_filter.txt

#*the phylogenetic tree structure with branch lengths*
Tree (((((filiformis:8.9,gaura:8.9):173.8,suffulta:182.7):531.3,
(rosea:506.4,speciosa:506.4):207.7):252.3,
((((argillicola:75.8,parviflora:75.8):30.4,oakesiana:106.2):34.1,
(((grandiflora:4.4,nutans:4.4):1,biennis:5.5):100.2,((((((elataajbk:52.1,wolfii:52.1):9,
((elataoeee:43.9,villosa:43.9):10.2,longissima:54.1):7):4.8,
(hirsuti:55.8,jamesii:55.8):10.2):14.6,glazioviana:80.7):12.9,stuchii:93.6):12.1):34.6):109.8,
(((clelandii:90.1,rhombipetala:90.1):90.7,(grandis:12,laciniata:12):168.9):44.9,
(((longituba:75.4,nana:75.4):27.2,villaricae:102.6):19.5,
(affinis:69.1,picensis:69.1):53.1):103.6):24.3):716.2):524.2,berlandieri:1490.8)

#*search for 1 parameter model*
lambda -s -t (((((((1,1)1,1)1,(1,1)1)1,((((1,1)1,1)1,(((1,1)1,1)1,(((((1,1)1,((1,1)1,1)1)1,
(1,1)1)1,1)1,1)1)1)1,(((1,1)1,(1,1)1)1,(((1,1)1,1)1,(1,1)1)1)1)1,1)