Independent Project – Phylogeny of Lycaenidae

Dana Hughes

1. Introduction

Information contained in the sequences of DNA can be used to analyze and construct phylogenies to identify relationships among distantly related taxa. Many different methodological approaches are available to produce accurate phylogeny estimations at lower costs. Here we use a simplified and reproducible anchored hybrid enrichment (AHE) assembly program to construct a phylogeny of 13 unknown Lycaenidae species. AHE is a popular method best suited for this experiment because of its ease of production and utilization of existing Lycaenidae subset references. The results from the AHE method were found to accurately evaluate phylogenetic relationships among the different taxa.

2. Methods

2.1 Sequencing

Read pairs were merged together to preserve informative reads that have been lost during the trimming process. A histogram output file is produced, containing the values of the merged read lengths. Read lengths were then plotted using R to visually represent the overall read distributions. The slope indicates a higher abundance of short read lengths and a lower abundance of long read lengths. After the merged reads were assembled, the reads were processed using Lycaenidae subset references with .fasta output files containing the sequence assembly coverages. These coverages would then be extracted and summarized in a .csv file comparing the individual loci, homolog IDs, and the number of reads mapped to the corresponding locus.

2.2 Contamination screening

Identification of contamination, hybridization, and miscellaneous assembly problems was processed to improve data analysis and minimize misassembly and inaccurate conclusions.

2.3 Orthology prediction

In order to assess orthology, homologous consensus sequences with acceptable coverage thresholds were collected and partitioned to specific loci. The consensus sequences are then aligned and measured using matrix optimization in order to define genetic distances between the sequences by counting the number of pairwise differences between the sequences. After calculating genetic distances, homologs are clustered into orthologous sets for each individual. Finally, the alignments are trimmed and concatenated to produce a RAxML file to be used to visually estimate phylogenetic relationships.

3. Results and Discussion

3.1. Sequencing and assembly

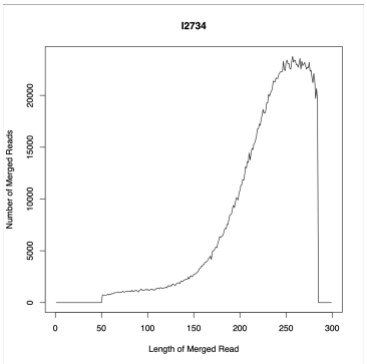

Figure 1: Summary of read lengths of single unknown loci where x = length of merged reads, y = number of merged reads. Figure based on single unknown individual.

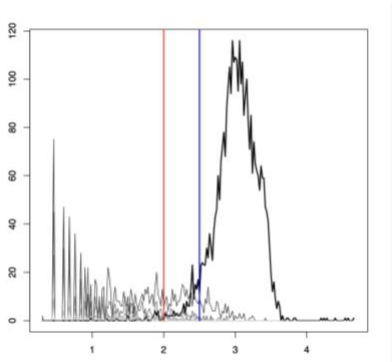

Figure 2: Summary of good sequences from the mapped reads where x = mapped reads on log10 scale, y = frequency rate.

Assembly_Summary

Individual	SampleID	Family	Genus	Epithet	Notes	TaxonSet	nRawReads	nRawBases	nLoci125	nLoci250	nLoci500	nLoci1000	AvgLocLen	%OnTarget	AvgNHomologs	DupRateAll
l2734							6534228	1329674700	494	494	490	427	1376.1307947019900	21.0375465219991	1.4403973509933800	0.4660514168810210
l2751							1580701	319298400	142	42	10	5	177.16887417218500	0.9510318086317540	0.4486754966887420	0.1321351288274700
l2777							2063948	430246200	602	601	595	503	1634.3692052980100	45.46431760968720	1.3476821192053000	0.438090158376195
l2786							5871365	1201184700	0	0	0	0	0.0	26.781839307575300	1.7980132450331100	0.4039865380487730
l2791							5916666	1175775600	601	599	587	482	1593.76821192053	18.208385281614100	1.4470198675496700	0.4132162854455200
l2792							3165909	606629100	600	597	581	420	1426.887417218540	27.492879445736200	1.3245033112582800	0.335601168935766
l2793							2739747	563071800	602	600	592	510	1615.432119205300	36.20445736925820	1.4271523178807900	0.3749693450158370
l2795							2266326	454801500	572	568	545	332	1231.8245033112600	27.742880141786000	1.0811258278145700	0.343479962467289
l2796							2075246	404157600	565	561	549	366	1309.5281456953600	41.680898749785000	1.0678807947019900	0.338192815689210
l2797							1171833	237023400	562	558	532	299	1153.2665562913900	54.34505389421960	1.0529801324503300	0.509707964210409
l2798							1278833	254579700	567	563	538	306	1171.1158940397400	48.973584798390000	1.0894039735099300	0.448026948229786

Figure 3: Assembly summary of each individual.

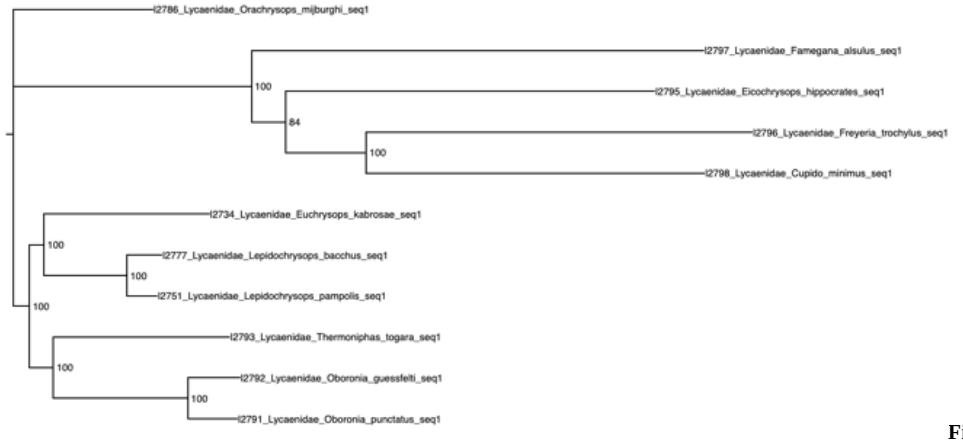

Figure 4: phylogeny of species.

Reports of the read length distribution in figure 1 indicate the read length distributions were dependent on the fragment lengths. After sequencing and assembling unknown Lycaenidae, reports indicate relationship results in good accuracy.

4. References

1. Rokyta, D.R., Lemmon, A.R., Margres, M.J. et al. The venom-gland transcriptome of the eastern diamondback rattlesnake (Crotalus adamanteus). BMC Genomics 13, 312 (2012). https://doi.org/10.1186/1471-2164-13-312

2. Young, A.D. and Gillung, J.P. (2020), Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. Syst Entomol, 45: 225-247. https://doi.org/10.1111/syen.12406

3. Jesse W. Breinholt, Chandra Earl, Alan R. Lemmon, Emily Moriarty Lemmon, Lei Xiao, Akito Y. Kawahara, Resolving Relationships among the Megadiverse Butterflies and Moths with a Novel Pipeline for Anchored Phylogenomics, Systematic Biology, Volume 67, Issue 1, January 2018, Pages 78–93, https://doi.org/10.1093/sysbio/syx048