2013-12-01

Application of Next-Generation Transcriptomic Tools for Non-Model Organisms: Gene Discovery and Marker DevelopmentWithin Plecoptera (Insecta)

Nicholas Gregory Davis
Brigham Young University - Provo

Application of Next-Generation Transcriptomic Tools for Non-Model Organisms: Gene Discovery and Marker Development Within Plecoptera (Insecta)

Nicholas G. Davis

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Dennis Shiozawa, Chair
Joshua Udall
Paul Evans

Department of Biology

Brigham Young University

December 2013

ABSTRACT

Application of Next-Generation Transcriptomic Tools for Non-Model Organisms: Gene Discovery and Marker Development Within Plecoptera (Insecta)

Nicholas G. Davis
Department of Biology, BYU
Master of Science

Phylogenetic research on non-model organisms has been hindered by limited marker availability. Next generation sequencing techniques are eliminating that barrier. Using Illumina sequencing technology, Trinity assembly software, custom Perl reciprocal BLAST scripts, and Primer3 primer prediction software, we produced and analyzed 7 Plecopteran transcriptomes, representing 7 of the 16 total families, in an attempt to identify and develop conserved orthologous genetic markers. The transcriptomes were used to reconstruct a gene content phylogeny using a simple distance matrix generated from reciprocal blastn data. By producing and filtering a reciprocal blast network we identified and aligned over 450 putative orthologs. Out of these, 25 primer pairs were selected that showed 100% conserved primer sites across all the transcripts from which they were created. Of those 25, 3 loci (PlecSK1, Perl534, and PvC2190) show very positive phylogenetic potential. These 3 markers may also be suitable and even highly useful in population genetic studies in which the populations have had sufficient time to develop significant genetic separation. The rapid and affordable nature of this study demonstrates the ease by which non-model organism phylogenetics can be expanded and made more robust.

Keywords: transcriptome, phylogenetic, plecoptera, insects, non-model, BLAST, ortholog

ACKNOWLEDGMENTS

I would like to acknowledge the support and direction of my adviser Dennis Shiozawa, and my committee members Joshua Udall and Paul Evans, as well as a former committee member, Michael Whiting. I would like to thank Justin Page for his invaluable assistance and expertise with programming. I thank Derek Houston for his assistance and training in molecular genetics. I would like to thank C. Riley Nelson and Richard Baumann for training and assisting with specimen collection and identification. I thank Josh Raney for training in RNA extraction and Illumina RNA sequencing protocols. I would especially like to thank Roger and Victoria Sant for their charitable donations in the form of a grant which made this research possible.

Table of Contents

List of Tables

List of Figures

Introduction

Phylogenetic research with non-model organisms has been hampered by the limited number of markers. Due to the difficulty in developing additional markers, genes and markers are often chosen for a study based on availability rather than suitability for the phylogenetic hypothesis to be tested [1]. This problem, of less than ideal markers, has contributed to the continual revisions to specific evolutionary trees and taxonomic nomenclature. Adding data from a single gene can result in major relational changes, indicating that not incorporating enough markers to allow the congruence and incongruence between the individual gene trees to balance, prevents generation of an accurate and robust species tree, which is stable to the inclusion of additional data [2,3]. Adding more sequences to a dataset does not resolve all of the challenges of systematics, which include long-branch attraction, lineage sorting, well supported erroneous trees, bad alignments, proper ortholog identification, taxon sampling, and evolutionary model selection [4,5]. However, while adding sequence data is clearly not an end all solution, it will increase the tree's resistance to change. Thus marker selection is crucial to avoid reliance on an incorrect topology.

Nuclear protein coding genes are often well conserved markers. They can significantly increase the ability to recover deep rooted evolutionary relationships [6]. Messenger RNA transcript sequencing is an efficient and effective way to sample the protein coding sequence of the nuclear genome for evolutionary studies [7-10]. RNA extraction techniques, including silica matrix (SM) and guanidinium thiocyanate-phenol-chloroform (GTPC), successfully isolate RNA from all other genomic material [11,12]. RNA can then be converted into complementary DNA (cDNA) that can be sequenced. The ability to sequence large portions of genomes at reasonable costs is facilitated by advances in next generation sequencing technology [13]. Many platforms exist, but at this time none produce as many reads at the fidelity and low price per base as the Illumina platform.

Illumina sequencing technology relies on bridge amplification that clones specific sequences in clusters (solid-phase amplification) [14,15]. The base identification is achieved through a cyclic process of washing polymerase, reversible terminator nucleotides, and primers across the sample, followed by laser excitation to fluoresce and read the base addition; reversible terminators must then be unblocked [14,15]. Each sequencing cycle uses fresh chemicals, thus increasing sequencing cycles in a sequencing run raises the total sequencing cost. Advantages of Illumina's sequencing methods over other currently available technologies include no issues with homopolymer runs, reading all A, G, C, or T nucleotides within the same flow cycle, and a very high throughput (600 Gb for 100 bp paired end sequencing) (www.illumina.com) resulting in a lower cost per base (Roche/454 gives only 700 Mb) (www.454.com) [14,15]. Disadvantages include a somewhat quick dephasing that results in shorter reads (150 bp) (www.illumina.com) compared to Roche/454 (650 bp) (www.454.com) and a relatively low capacity for multiplexing samples [14,15]. Overall the much higher throughput and ease of sequencing homopolymers has made Illumina an obvious choice for many research applications.

A plethora of computer algorithms and software packages have been developed to deal with next generation sequence data. Early shotgun sequencing assembly methods relied on Overlap Layout Consensus (OLC) methods, which is optimal for the longer reads provided by Sanger and Roche/454 [16,17]. Yet, Illumina, which produces much shorter reads, requires a very different assembly approach. De Bruijn Graph (DBG) assembly methods are ideal for short read data [16]. The process relies on breaking sequencing reads into small chunks called k-mers, of length k, aligning k-mers that most likely go together meeting the designated criteria or parameters, eliminating k-mers that are likely false due to sequencing errors, and assembling the reads into transcripts or contiguous sequences [16]. DBG software choices are abundant, and which one to use can depend on the data set itself. Trinity is a software package that was developed to specifically handle de novo transcriptome assembly [18]. Its strengths over other DBG software include extremely low-base error rates (detects 99% of errors), handling small and large data sets equally well and across a range of conditions, and higher rates of gene and isoform detection [18,19]. However, Trinity software requires significantly longer run times [19], but with adequate planning this should never be a problem. Being that our RNA-seq datasets will be small, full of complex alternative splicing, and require extremely low error rates for accurate ortholog detection, Trinity is the clear choice for our data-analysis.

Homology is the underlying principle to all phylogenetics and systematics. It is the concept that a character, gene, or nucleotide base in two different organismal groups is derived from the same character, gene, or nucleotide base in a common ancestor. Homology in historical phylogenetics can only be inferred from evidence such as position, likeness, and function. An ortholog is a homologous genetic sequence or gene that is directly descended from the same copy of an ancestral gene as opposed to paralogs that are related as different copies of the same gene from a duplication event [20]. Inferring genetic orthology is complicated by many non-trivial factors. They include paralogy, gene loss, gene fusion, gene fission, horizontal gene transfer, insertions, deletions, gene duplication, back mutations (multiple hits), and incomplete lineage sorting [21,22].

Ortholog inference software programs are based on two main approaches. The first approach is a tree method that compares the proposed ortholog tree to a reliable species tree and then evaluates the evolution of the ortholog in most the parsimonious way to verify probable orthology [21,22]. The second approach is a graph-based or heuristic method that utilizes a BLAST or best match algorithm. The best-matches among all compared sequences are determined from a "pairwise sequence similarity search" and are set aside as putative homologs [21,22]. Tree-based methods tend to be less sensitive to the exact software used but rely on the accuracy of a multiple sequence alignment (MSA) and the tree reconstruction which both of which can have bias and error, causing incorrect ortholog prediction [21,22]. While tree based methods tend to be more accurate than heuristic approaches, especially when gene losses are present, they are substantially more intense computationally [21,22]. The advantages of heuristic approaches include being faster, easier to run, and not influenced by MSA and tree errors [21,22]. Ortholog inference software has been solely applied to nuclear genomic data, not to cDNA transcript data. It is very unlikely that any of the available software packages would work for our datasets. Therefore we will develop our own that combines several of the techniques used in other ortholog detection software [23] such as self-blasting and reciprocal blasting [24].

Orthologous sequences can be used in a traditional phylogenetic framework, but genomic data has provided opportunities for the development of novel approaches to phylogenetic investigations. Gene content phylogeny [25], while relatively computationally inexpensive, is an entirely different approach to inferring systematic relationships. It relies on the idea that the more similar two taxa are, the more genes or gene content of the genome will be shared. In general, the concept is applied by performing BLAST on whole genomic datasets, between all of the taxa being evaluated. The higher the number of blasts between taxa the smaller the distance of relatedness is between them, thus a distance matrix is employed in tree reconstruction. Many studies have employed this method [26,27], some modifying it by applying ideas of weighting and e-values [28,29]. E-values reflect how likely a blast result would be by chance, calculated based on a number of factors including percent identity and the length of the matching, also called sequence coverage.

The intent, of this project is to develop these methods and expand the genetic marker set, identifying genes that can be used as phylogenetic markers in non-model organisms. We selected an aquatic insect order, Plecoptera. This order contains roughly 3500 species, 286 genera, and 16 families [30]. The first published molecular phylogeny of Plectoptera utilized a single nuclear marker, 28S, approximately 1,300 nucleotides in length and incorporated 30 taxa. All families were represented [31]. The most recent, and only other molecular systematic study of Plecoptera was much more extensive, utilizing 138 binary characters, 6 genes (5 loci), approximately 6000 total nucleotides, incorporating 179 taxa with all families represented [32]. All other research on the phylogenetic relations among the families of Plecoptera have been based on morphological and behavioral characteristics [32-37]. The use of phenotypic characters can be significantly biased by the researcher's experience and presuppositions [38]. It is well accepted that the best approach to inferring systematic relationships incorporates both molecular and morphological data. However, being that morphological approaches have been the major focus for most insect systematics, including Plecoptera, the expansion of the molecular data set appears to be the most promising source of additional informative data [39].

Non-model organism phylogenetics have been restricted by a lack of phylogenetic markers [40]. Current arthropod systematics is still only using a few genes in most analyses and most sequence data are limited to the mitochondrial genome, which has been well demonstrated as a relatively poor marker for deeper phylogenetic relationships [41]. Several research teams have leveraged RNA-seq data to develop nuclear markers specific to their orders [8-10]. The objectives of this study are to identify conserved, yet informative, nuclear markers suitable for designing robust PCR primers by using RNA sequence data from many different taxa within Plecoptera, and to employ genome content phylogenetic ideas to transcriptomes.

Materials and Methods

RNA Sequencing

Specimens representing seven different stonefly families were collected (See Table 1) from February to April 2012: *Capnia nana*, *Hesperoperla pacifica*, *Megarcys Sp*., *Pteronarcys californica*, *Sweltsa Sp*, *Taenionema Sp.*, *Zapada cinctipes.* Upon collection, specimens were placed live in bottles filled with stream water and stored in coolers with ice for transport. RNA extractions were performed within 48 hours of collection. For next day extractions the specimens were kept overnight in a chilled, well oxygenated, artificial stream. RNA extractions were conducted using the Qiagen RNeasy Plant Mini Kit (Qiagen Group, Valencia, CA) following their "Purification of Total RNA from Animal Tissues" protocol. Successful extraction was verified with a NanoDrop spectrometer. Failed extractions were repeated until successful. Extracted total RNA was stored at -80°C until all samples were finished and ready for cDNA library construction.

Immediately prior to cDNA library construction RNA concentrations were again quantified a second time using a NanoDrop spectrometer. One microgram of total RNA was used. cDNA libraries were made using Illumina TruSeq RNA Sample Prep Kit V2 (Illumina Inc., San Diego, CA) following the "Low Throughput" (LT) protocol. Success of cDNA library construction was verified using the Agilent DNA 7500 Kit (Agilent Technologies Inc., Santa Clara, CA) to determine overall range of fragment size. A PicoGreen assay was used to determine cDNA concentrations for each sample. cDNA libraries were submitted to the

Huntsman Cancer Institute for sequencing on an Illumina HiSeq 2000 (Illumina, San Diego, CA). The seven stonefly libraries and 6 other insect libraries were sequenced on a single lane of a flow cell at 50 bp single end reads. Reads obtained from this sequencing run were then filtered using the Sickle (najoshi, GitHub.com) quality trimmer software and assembled de novo into contigs, representative of transcripts, using Trinity assembly software [18].

Ortholog Inference

Each transcriptome or RNA sequencing dataset was compared and evaluated by a set of custom Perl scripts (runBlast.pl, blast.pl, mergeBlast.pl). These scripts filter out contigs through a combination of self-blasting and a reciprocal blast network. runBlast.pl (see appendix A) designates the files to be analyzed, sets e-value thresholds (i.e. best hit at 1e-50, and better than second best hit by a factor of 1e-20 or more) prepares a blast index for each file, and then calls blast.pl. blast.pl (see appendix B) submits a blastn job for all file combinations to the computer, removes blastn results not meeting the set e-value threshold requirements from further analysis, and prints acceptable data to an outfile. Finally, in running mergeBlast.pl (see appendix C), outfiles from blast.pl are used to create a hash tree which connects all of the file combinations blast results in a network. This network is evaluated to determine the reciprocity of blastn results and to designate apparently non-paralogous reciprocal best hits. Every contig of every file is checked for self-blast and to insure that it is also the sole significant blast for each contig it blast-ed to within the other files (each file being the blast.pl hit data of an individual species' transcriptome). Networks of contigs passing these criteria are then aligned using MUSCLE [42] and saved as alignment text files. mergeBlast.pl also allows for "missing data" in the sense that the user indicates how many of the taxa must participate in the reciprocal blast network. To evaluate the effect of missing data, up to 5 of the 7 taxa, that were RNA-sequenced, were permitted to not participate in the reciprocal blast networks.

Transcriptome Content Phylogeny

A separate reciprocal blast analysis was conducted, in which each species' transcriptome was blast-ed against that of every other species in the study. The total number of blasts that were reciprocated and significant at an e-value of 1e-40 or greater with no other hits greater than 1e-20, were counted as reciprocal best hits. The resulting hit count for each combination of species was put into a distance matrix in which the distance (number of reciprocal best hits) was interpreted inversely so that the greater the number of reciprocal best hits the closer two species were related, or in other words, the smaller the distance between them. The distance matrix was used to infer phylogenetic relationships between the seven taxa used for RNA sequencing.

Primer Development

Contig alignments produced during the mergeBlast.pl step of the ortholog inference were evaluated by eye in GeneiousPro version 6.0.5 (Biomatters; Auckland, NZ). Alignments that contained two or more regions that were conserved at 100% identity and long enough (≥18bp) as well as at least 100 bp apart from one another were used to make primers in Primer3 online [43]. Primer3 parameters were left on default.

Marker Validation

Primers were tested using template DNA from five different freshly collected stonefly specimens representing five families (See Table 2): *Capnia utahensis*, *Claassenia Sp.*, *Isogenoides Sp.*, *Pteronarcella badia*, *Zapada cinctipes*. DNA extractions were done using the

Qiagen DNeasy kit using the animal tissue protocol (Qiagen Group, Valencia, CA). One of the five species used was one of the RNA-sequenced species (*Zapada cinctipes*). PCR reactions were run using the following reaction mixture: 2.25 μL nuclease-free water, 0.5 μL forward primer, 0.5 μL reverse primer, 6.25 μL Taq polymerase, and 3 μL DNA for a total reaction volume of 12.5 μL. Cyclic PCR reactions consisted of 3 min at 95˚C; 35 cycles of 1 min at 95˚, 1 min at 55˚C, and 90 sec at 72˚C, followed by a final extension step of 7 min at 72˚C. Amplification success was verified by standard gel electrophoresis.

Purified PCR product was used as template for cycle sequencing reactions with Big Dye chemistry (Applied Biosystems, Inc. Foster City, CA) using the same primers for PCR, using the following reaction mixture: 2.75 μL nuclease-free water, 1.75 μL 5x buffer, 0.5 μL Big Dye, 0.5 μL primer (~10 pmoles), and 5.0 μL of purified PCR product for a total reaction volume of 10.5 μL. Separate sequencing reactions were used for the forward and reverse primers. Products of cycle sequencing were purified with Sephadex (G-50; Sigma-Aldrich Co., St. Louis, MO) spin columns. Dry samples were submitted to the Brigham Young University DNA Sequencing Center to be Sanger sequenced using a 3730xl automated sequencer (Applied Biosystems, Inc. Foster City, CA).

Sequences were aligned and edited in GeneiousPro. Individual alignments were used to create a single consensus sequence. All consensus sequences for a single gene were aligned together as nucleotide sequences and were translated using "Codon Align", on the HIV Sequence Database website (http://www.hiv.lanl.gov/content/sequence/CodonAlign/codonalign.html), which infers the best reading frame aligned as protein sequences. Each consensus sequence was also blast-ed against NCBI's non-redundant nucleotide database using blastn [44] as well as the protein database using blastx [44]. Alignment and blast results were evaluated for consistency.

Genes that were amplified and sequenced as the same single copy gene were then sequenced in additional species and in two cases, individuals from the same species but from different populations (see Table 3). Identifications of the specimens used in this part of the protocol and in previous steps were performed by the author and Charles R. Nelson (Department of Biology, BYU) using the appropriate manuals ([45-48]. The resulting sequence data were then aligned to the data from the original test PCR taxa as well as the transcripts from which the primers were created. The alignments were trimmed in BioEdit [49] so that the same nucleotides were evaluated/compared for each taxon. A maximum likelihood phylogeny was created by performing a maximum likelihood search and GTR+G model of rate heterogeneity at 100 bootstraps using RAxML online version 7.7.1 [50] and PhyML [51] at 1000 bootstraps [52]. Neighbor-Joining [53] trees were created using HKY [54] and Jukes-Cantor [55] genetic distance models and bootstrap values were calculated from 1000 replicates. Trees were re-rooted separating the two stonefly suborders systellognathan and euholognathan. The figures were made using FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Results

RNA Sequencing

RNA extraction concentrations ranged from 60 to 2500 nanograms/microliter, the average being 1047 ng/µl. The cDNA library concentrations ranged from 12 to 66 nanograms/microliter, the average being 42 ng/µl. All 7 Plecopteran Illumina libraries sequenced and assembled, ranging from 15,305 to 26,526 contigs.

Ortholog Inference

Applying the custom BLAST scripts to the 7 stonefly transcriptomes, not allowing for any “missing” data, created about 30 alignments (at 70% shared identity) of unique orthologs. When 1, 2, or 3 transcriptomes were allowed to be missing from the reciprocal blast network, the set of putative ortholog alignments increased to ~120. Permitting up to 4 to be missing, generated a total of ~170 alignments were generated. Two hundred and thirty alignments (at 80% shared identity) were output when the reciprocal blast network was constrained to just the hypothesized super-family Perloidea (see Fig. 1). When evaluating the probable sister families of Perlodidae and Chloroperlidae, 465 alignments (at 85% shared identity) were found.

Transcriptome Content Phylogeny

The transcriptome content phylogenetic reconstruction recovered the major relationships found in the leading Plectopteran morphological systematics research. Those major relationships include the monophyly of Perloidea, Systellognatha, and Euholognatha. The resulting tree differs in that the relationships between Perloidea families are resolved (Perlidae as the outgroup to Perlodidae and Chloroperlidae) as is a rearrangement of relationships between the Euholognathan taxa (Nemouridae as the outgroup to Capniidae and Taeniopterygidae). Overall, the differences in the total numbers of genes shared between different nodes in the tree are high, with the exception of the Nemouridae and the Capniid and Taeniopterygid sister relationship (52 reciprocal blasts). This is much lower in comparison to the difference between Perlidae and the sister relationship of Perlodidae and Chloroperlidae (594 reciprocal blasts) as well the difference between Pteronarcyidae and Perloidea (184 reciprocal blasts).

Primer Development

Of the ~170 alignments, when evaluating all of the stonefly transcriptomes together, 22 appeared suitable for developing primers, in that they contained regions which were conserved >18 bp in a row. In designing the primers, 14 of those 22 produced suitable primer pairs, suitable meeting standard primer optimization criteria (i.e. melting temperature, homopolymers, dimerization, etc.). Of the ~230 Perloidea alignments, 22 appeared suitable with only 9 of the 22 adequately meeting the primers generation parameters. Thirty-nine primer pairs were found suitable from the Perlodid-Cholorperlid alignments, 35 of those were conserved enough for primers. All of the Plecoptera and Perloidea primers were tested, while only 2 from Perlodidae vs. Chloroperlidae were tested, due to the fact that it was poor subsampling of order level variation.

Marker Validation

Overall, successful, consistent, and intentional PCR amplification was rare. Eight of the 25 tested markers amplified in at least a single taxon. Of those 8 loci, 5 produced sequence from the intended target region. However, only three (PlecSK1, Perl534, and PvC2190) of the 5 markers met the criteria of consistent amplifications, clean sequencing of all amplified taxa, and clean alignment (Table 4).

PlecSK1 sequence data were produced in four out the seven transcriptomes, five out of the five primer testing taxa, and for 11 additional individuals. Total taxonomic representation consisted of 7 families, 15 genera, and 17 species, two species having two individuals from two different populations. PlecSK1 nucleotide and protein sequence returned the same best BLAST results using blastn and blastx respectively; identifying the sequences generated from the PlecSK1 primers as a muscle-specific actin of the sugar kinase HSP70 actin superfamily. The trimmed nucleotide alignment (870 bp) of 19 of the 20 nucleotide sequences (*Claassenia Sp.* sequence data quality was too poor to include) had 175 variable sites (20.1%). The protein alignment of the same sequence data showed five non-synonymous mutations (98.3% identical, 16/19 taxa being 100% identical). Two of those non-synonymous mutations only occurred in both of the Pteronarcyid species. One occurred in both Pteronarcyid species and the Neoperla species. One occurred in only the two species of Agnetina. The last one only occurred in the two Agnetina species and the two Pteronarcyid species.

The maximum likelihood (Fig. 2) and neighbor-joining (Fig. 3) phylogenies, based on the PlecSK1 nucleotide alignment, display, similar results to each other. In both methods, all of the families represented by more than one species, remained monophyletic. In both methods, the monophyly of Systellognatha and Euholognatha are strongly supported. Although nearly all of the systematic relationships are conserved between the two methods, eight of the 17 nodes have relatively weak bootstrap support.

Perl534 sequence data is present from three (from Perloidea) of the seven transcriptomes and was sequenced in three of the primer testing taxa; *Claassenia Sp.*, *Isogenoides Sp.*, and *Pteronarcella badia.* The trimmed nucleotide alignment of these six sequences was 179 bases, 60 of which were variable (66.5% conserved identity). The translated nucleotide alignment is 50% variable, however, the variation exists between two regions with 100% identity on the end of the locus, the regions being eight and ten amino acids long. The variable areas tend to vary among all taxa and the more conserved regions, tend to be conserved among all taxa, with only three gaps of one amino acid spacing. The sequences for this locus did not blast to any known nucleotide or protein sequences using both blastn and blastx.

PvC2190 was sequenced for a total of six species, two sequences from the transcriptomes of *Megarcys Sp*. and *Sweltsa Sp*.; the rest being *Claasenia* Sp., *Isogenoides* Sp., *Capnia utahensis*, and *Pteronarcella badia*. The nucleotide alignment is 427 base pairs long with 172 variable sites (40.3%). The protein sequence alignment is only 37.2% conserved. However, the conserved regions are in series of 17-26 residues. Blastn and blastx against NCBI's database do not return any results. Specifics, including the specific primer sequences, for PlecSK1, Perl534, and PvC2190 are found in Table 5.

Discussion

Inferring the orthology of any trait, genetic or morphological, carries an inseparable burden, one that affects all of phylogenetics. That burden is the fact that in spite of a methodical and well-reasoned approach, just as you cannot directly observe the history of evolution, you cannot know that the marker you are using conveys correct phylogenetic signal. The goal and expectation of this study was to develop and proof a relatively inexpensive methodology for finding and developing a plethora of genetic markers which would allow enough correct phylogenetic signal to overshadow incorrect or misleading similarities.

It is potentially more challenging to develop markers for all of Plecoptera and other old insect orders like it, simply because, they share a more distantly related common ancestor than more recent orders such as Diptera or Lepidoptera [8]. It could be the case that age has less of an effect on genetic divergence than the biology and behavior of a particular lineage's genome or life history, however, realistically it is an interaction of many factors, including time. As part of this same consideration, molecular evolution will vary from locus to locus and even base pair to base pair. 3rd base pair codon position typically evolves faster than its two other counter parts which tend to cause changes in the protein sequence. The study marker PlecSK1 appears to demonstrate this principle in that nearly all of the variable nucleotides between taxa are 3rd codon and synonymous. Being that the variability is synonymous, it stands to reason that those nucleotide positions should each represent near neutral phylogenetic data points. One major drawback, however, is that in comparing taxa at an ordinal level and of an order as old as Plecoptera, neutral data points may mutate so fast that they give useless and even misleading information. Transcriptomes, representative of the "active" parts of the genome may have inherent advantages and disadvantages relative to marker development.

Transcriptomes represent the RNA being actively transcribed in cells at a particular moment in time. Expression varies over an organism's lifespan as well as in response to environmental inputs. In this study, for many of the samples, individuals were pooled together for extractions, especially for the taxa with relatively low mass. Whether an extraction was performed on pooled individuals or not, the RNA generally only represented one life stage's response or status with regard to regulatory processes for that life stage and environmental conditions. By incorporating individuals from a large variety of the life stages of an organism, the likelihood of capturing a greater proportion of the species' exome should increase. Doing this for all of the species sequenced would also likely increase the number of, or at least confidence in, predicted orthologs. That being said, there were still hundreds of putative orthologs identified without performing a more robust sampling of the exome.

While suitable primers could be not be made for all of the putative orthologs we identified, the markers that we were able to amplify and evaluate are promising. The fact that the marker PvC2190, which amplified in multiple families, was created by comparing two species from two different families of the same superfamily. This shows that it may not be necessary in all cases to sequence RNA for many taxa for both a large and old groups. While the marker

Perl534 does not amplify a particularly large section of DNA, it does have a high amount of variability. It too was created from a significantly narrowed subset of the transcriptomes. PlecSK1, created only from four of the seven transcriptomes is the longest and most conserved when it comes to the protein sequence, the variable nucleotides being functionally neutral third codon positions.

The maximum likelihood and neighbor-joining phylogenies, constructed using the PlecSK1 alignment, have only minor incongruences with the most accepted systematic hypotheses of Pleopteran systematics [36]. One incongruence being that Pteronarcyidae was not recovered as the out-group to the Superfamily Perloidea, rather that Perlidae is the out-group to a monophyletic clustering of Chloroperlidae, Perlodidae, and Pteronarcyidae (See Figures 2 and 3). The congruence between the current major Plecopteran systematic studies and the PlecSK1 phylogeny, may indicate, through the topology and support values, that the bases that are informative for resolution of deeper nodes may be able to overshadow the noise provided by the bases that may be helping to resolve the shallower nodes. While PlecSK1 sequence is unlikely to be sufficient to reconstruct or represent the real history of all relationships within Plecoptera, its greatest importance would be to analyze it in conjunction to the datasets of Zwick et al 2000 and Terry et al 2003.

Additional sequencing, similar to that in investigating the phylogenetic potential of PlecSK1, should be completed for both Perl534 and PvC2190. It is interesting to note in both PlecSK1 phylogenies, that the support value for the monophyly of *Hesperoperla pacifica* is lower than would be expected (ML 58, NJ 93) (see Figures 2 and 3). These lower bootstrap values are low due to the amount of intraspecific variation found between the two specimens used. This may indicate that PlecSK1 could be useful in population genetic studies in some Plecopteran species in addition to *Hesperoperla pacifica*.

There are several points of possible modification and improvement for this work. Two of those factors include generating higher quality sequence data and using an improved application of BLAST comparisons. The Illumina sequencing in this study was conducted using single end reads. Paired end sequencing can result in 10 times the number of assembled contigs (Shiozawa unpublished data), which would likely increase the ability of our Perl scripts to infer more orthologs and paralogs. In addition to generating more complete transcriptomes by changing the BLAST software to run reciprocal blast networks based on translated nucleotide sequences (tblastx), cross-comparing translated transcriptome sequences and performing protein sequence alignments, proteins would give a much more accurate picture of orthology. Suitable alignments would then be back-translated to visualize the potential for making primers.

The transcriptome content phylogeny we produced, akin the genome content phylogeny concept [25], is apparently its first application to RNA sequence data. It assumes that the more related two genomes are to one another, the more they will share reciprocal best blasts or reciprocal best hits between them. It generated a phylogeny (Fig.1) which is nearly identical to the leading Plecopteran systematic hypotheses, supporting the monophyly of Perloidea, Euholognatha, and Systellognatha. In addition, Perlidae was placed as the out-group to a sister relationship between Perlodidae and Chloroperlidae, which is the prominent hypothesis among Plecopteran morphologists. While not ideally sampled, both for life stages, taxonomic breadth, and lack of replicate independent sequencings, it demonstrates the application's potential. Its accuracy should be quantitatively compared to traditional base pair to base pair comparisons, as it uses a broad source of evidence, based on the comparisons of thousands of loci. This approach may even benefit from running the analysis based on both blastn and tblastx, doing both may increase robustness and accuracy.

From this research many advantages and disadvantages to working with transcriptomes for phylogenetic or population genetics work have become clear. Some of these advantages or disadvantages could be seen the other way based on the specifics of the study but they are worth considering. Positive attributes include that transcriptomes are relatively conserved. Transcriptome sequencing is a very effective form of genome reduction. The molecular protocols involved are relatively easy and can be accomplished with standard lab equipment. Transcriptomes can be assembled De Novo. The overall process is relatively affordable (<$5000 for extractions, cDNA library construction, and Illumina sequencing). Illumina transcriptome data are relatively bioinformatically simple to evaluate. Disadvantages to RNA sequencing include the rapid degradation of the RNA itself, requiring fresh carefully processed tissue. The Illumina library construction protocol is somewhat time consuming. Without a reference genome, predicting introns and primer mispriming issues for creating optimized primers, are not possible. There is a bigger upfront investment in the sequencing process compared to Sanger methods, however, over all costs are low. It simply implies that the focus should shift to the planning stages of the sequencing project. It requires bioinformatics skills and the processing power of a supercomputer, especially for transcriptome assembly.

Overall, this study demonstrates the rapid and relatively simple process of generating hundreds of orthologous sequence alignments for a group of non-model organisms. While genes may very well be orthologs, that does not imply that they have long enough regions of identical base pairs to be able to create primers for successful Sanger sequencing. However, investing in paired end sequencing, running the reciprocal blast as a tblastx, and pooling individuals together that represent a robust sampling of the life cycle, should dramatically improve results. This is especially the case when working with sub-ordinal relationships or with groups that are more recently diverged. While Sanger sequencing will generally not facilitate marker discovery, it still remains a very high quality sequencing method and is very useful in validating genetic markers as well as being the method of choice for smaller scale target sequencing studies.

Citations

1. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798-804.
2. Guidetti R, Schill RO, Bertolani R, Dandekar T, Wolf M (2009) New molecular data for tardigrade phylogeny, with the erection of Paramacrobiotus gen. nov. Journal of Zoological Systematics and Evolutionary Research 47: 315-321.
3. Yi ZZ, Song WB, Clamp JC, Chen ZG, Gao S, et al. (2009) Reconsideration of systematic relationships within the order Euplotida (Protista, Ciliophora) using new sequences of the gene coding for small-subunit rRNA and testing the use of combined data sets to construct phylogenies of the Diophrys-complex. Molecular Phylogenetics and Evolution 50: 599-607.
4. Kumar S, Filipski AJ, Battistuzzi FU, Pond SLK, Tamura K (2012) Statistics and Truth in Phylogenomics. Molecular Biology and Evolution 29: 457-472.
5. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, et al. (2011) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. Plos Biology 9.

6. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, et al. (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463: 1079-U1098.

7. Hittinger CT, Johnston M, Tossberg JT, Rokas A (2010) Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proceedings of the National Academy of Sciences of the United States of America 107: 1476-1481.

8. Papanicolaou A, Joron M, McMillan WO, Blaxter ML, Jiggins CD (2005) Genomic tools and cDNA derived markers for butterflies. Molecular Ecology 14: 2883-2897.

9. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, et al. (2009) Benchmarking Next-Generation Transcriptome Sequencing for Functional and Evolutionary Genomics. Molecular Biology and Evolution 26: 2731-2744.

10. Wahlberg N, Wheat CW (2008) Genomic outposts serve the phylogenomic pioneers: Designing novel nuclear markers for genomic DNA extractions of lepidoptera. Systematic Biology 57: 231-242.

11. Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, et al. (2011) Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. Molecular Ecology Resources 11: 650-661.

12. Tan SC, Yiap BC (2009) DNA, RNA, and Protein Extraction: The Past and The Present. Journal of Biomedicine and Biotechnology.

13. Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. Trends in Ecology & Evolution 24: 192-200.

14. Kircher M, Heyn P, Kelso J (2011) Addressing challenges in the production and analysis of illumina sequencing data. Bmc Genomics 12.

15. Metzker ML (2010) Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation. Nature Reviews Genetics 11: 31-46.

16. Li ZY, Chen YX, Mu DS, Yuan JY, Shi YJ, et al. (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Briefings in Functional Genomics 11: 25-37.

17. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nature Reviews Genetics 12: 671-682.

18. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29: 644-U130.

19. Zhao QY, Wang Y, Kong YM, Luo D, Li X, et al. (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. Bmc Bioinformatics 12.

20. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19: 99-113.

21. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV (2011) Computational methods for Gene Orthology inference. Briefings in Bioinformatics 12: 379-391.

22. Kuzniar A, van Ham R, Pongor S, Leunissen JAM (2008) The quest for orthologs: finding the corresponding gene across genomes. Trends in Genetics 24: 539-551.

23. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, et al. (2011) Proteinortho: Detection of (Co-)orthologs in large-scale analysis. Bmc Bioinformatics 12.

24. Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics 19: 1710-1711.

25. Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nature Genetics 21: 108-110.

26. Fuchsman CA, Rocap G (2006) Whole-genome reciprocal BLAST analysis reveals that Planctomycetes do not share an unusually large number of genes with Eukarya and Archaea. Applied and Environmental Microbiology 72: 6841-6844.

27. Trost B, Haakensen M, Pittet V, Ziola B, Kusalik A (2010) Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera. Bmc Microbiology 10.

28. Rosenfeld JA, DeSalle R (2012) E value cutoff and eukaryotic genome content phylogenetics. Molecular Phylogenetics and Evolution 63: 342-350.

29. Zhang H, Zhong Y, Hao B, Gu X (2009) A simple method for phylogenomic inference using the information of gene content of genomes. Gene 441: 163-168.

30. Fochetti R, de Figueroa JMT (2008) Global diversity of stoneflies (Plecoptera : Insecta) in freshwater. Hydrobiologia 595: 365-377.

31. Thomas MA, Walsh KA, Wolf MR, McPheron BA, Marden JH (2000) Molecular phylogenetic analysis of evolutionary trends in stonefly wing structure and locomotor behavior. Proceedings of the National Academy of Sciences of the United States of America 97: 13178-13183.

32. Terry MD, Whiting MF (2003) Phylogeny of Plecoptera: molecular evidence and evolutionary trends. Entomologische Abhandlungen (Dresden) 61: 130-131.

33. Maketon M, Stewart KW (1988) PATTERNS AND EVOLUTION OF DRUMMING BEHAVIOR IN THE STONEFLY FAMILIES PERLIDAE AND PELTOPERLIDAE. Aquatic Insects 10: 77-98.

34. Nelson CH (1984) NUMERICAL CLADISTIC-ANALYSIS OF PHYLOGENETIC-RELATIONSHIPS IN PLECOPTERA. Annals of the Entomological Society of America 77: 466-473.

35. Zwick P (1973) Insecta: Plecoptera. Das Tierreich No. 94: 1-465.

36. Zwick P (2000) Phylogenetic system and zoogeography of the plecoptera. Annual Review of Entomology 45: 709-746.

37. Zwick P (2003) Morphological support of the major clades of Plecoptera. Entomologische Abhandlungen (Dresden) 61: 128-130.

38. Wiens JJ (2001) Character analysis in morphological phylogenetics: Problems and solutions. Systematic Biology 50: 689-699.

39. Zwick P (2003) Plecoptera research today: questions to be asked in the new millennium; Gaino E, editor. 245-251 p.

40. Sanderson MJ (2008) Phylogenetic signal in the eukaryotic tree of life. Science 321: 121-123.

41. Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. Molecular Ecology 13: 729-744.

42. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.

43. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, et al. (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Research 35: W71-W74.

44. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

45. Stark BP, Armitage BJ (2000) Stoneflies (Plecoptera) of Eastern North America. Volume I. Pteronarcyidae, Peltoperlidae, and Taeniopterygidae. Bulletin of the Ohio Biological Survey 14: i-viii, 1-99.

46. Stark BP, Armitage BJ (2004) Stoneflies (Plecoptera) of Eastern North America. Volume II. Chloroperlidae, Perlidae, and Perlodidae (Perlodinae). Bulletin of the Ohio Biological Survey 14: i-vi, 1-192.

47. Baumann RW, Gaufin AR, Surdick RF (1977) The stoneflies (Plecoptera) of the Rocky Mountains. Memoirs Am ent Soc No. 31: 1-208,illust.

48. Merritt RW, Cummins KW, Berg MB (2008) An Introduction to Aquatic Insects of North America. Dubuque, IA: Kendall/Hunt Publishing Company. 1158 p.

49. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series: 95-58.

50. Stamatakis A, Hoover, P., Rougemont, J. (2008) A Rapid Bootstrap Algorithm for the RAxML Web-Servers. Systematic Biology 75: 758-771.

51. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696-704.

52. Guindon S, Delsuc F, Dufayard J-F, Gascuel O (2009) Estimating Maximum Likelihood Phylogenies with PhyML. In: Posada D, editor. Bioinformatics for DNA Sequence Analysis. pp. 113-137.

53. Saitou N, Nei M (1987) THE NEIGHBOR-JOINING METHOD - A NEW METHOD FOR RECONSTRUCTING PHYLOGENETIC TREES. Molecular Biology and Evolution 4: 406-425.

54. Hasegawa M, Kishino H, Yano TA (1985) DATING OF THE HUMAN APE SPLITTING BY A MOLECULAR CLOCK OF MITOCHONDRIAL-DNA. Journal of Molecular Evolution 22: 160-174.

55. Jukes TH, Cantor CR (1969) Evolution of protein molecules. Mammalian Protein Metabolism: 21-132.

Table 1: RNA Sequenced Taxa

Family	**Taxon**	**State**	**County**	**Location**	**Collection Date**	**Life Stage**
Capniidae	Capnia nana	UT	Utah	South Fork Lower Provo River	1-Mar-12	Adult
Chloroperlidae	Sweltsa Sp.	UT	Utah	South Fork Lower Provo River	5-Mar-12	Nymph
Nemouridae	Zapada cinctipes	UT	Utah	South Fork Lower Provo River	1-Mar-12	Adult
Perlidae	Hesperoperla pacifica	UT	Utah	South Fork Lower Provo River	23-Feb-12	Nymph
Perlodidae	Megarcys Sp.	UT	Utah	South Fork Lower Provo River	23-Feb-12	Nymph
Pteronarcyidae	Pteronarcys californica	UT	Utah	Diamond Fork River	4-Mar-12	Nymph
Taeniopterygidae	Taenionema Sp.	UT	Utah	South Fork Lower Provo River	5-Mar-12	Nymph

Table 2: Primer Testing Taxa

Family	**Taxon**	**State**	**County**	**Location**	**Collection Date**
Capniidae	*Capnia utahensis*	UT	Utah	Hobble Creek	12-Mar-13
Nemouridae	*Zapada cinctipes*	UT	Utah	Hobble Creek	12-Mar-13
Perlidae	Claassenia	UT	Utah	Diamond Fork River	12-Mar-13
Perlodidae	*Isogenoides Sp.*	UT	Utah	Soldier Creek	12-Mar-13
Pteronarcyidae	*Pteronarcella badia*	UT	Utah	Soldier Creek	12-Mar-13

Table 3: SK1 Phylogenetic Reconstruction Taxa

Family	Taxon	State	County	Location	Collection Date	Transcripts
Capniidae	*Capnia nana*	UT	Utah	South Fork Lower Provo River	1-Mar-12	✓
	Capnia utahensis	UT	Utah	Hobble Creek	12-Mar-13	
	Utacapnia logana	UT	Utah	Hobble Creek	12-Mar-13	
Chloroperlidae	*Alloperla thalia*	UT	Utah	South Fork American Fork River	20-Jul-13	
	Sweltsa Sp.	UT	Utah	South Fork Lower Provo River	5-Mar-12	
Leuctridae	*Paraleuctra Sp.*	UT	Duschesne	Yellowstone Creek	18-Sep-10	
Nemouridae	*Zapada cinctipes*	UT	Utah	South Fork Lower Provo River	1-Mar-12	✓
		UT	Utah	Hobble Creek	12-Mar-13	
Perlidae	*Agnetina capitata*	PA	Perry	Juniata River	11-Jun-13	
	Agneticna flavescens	PA	Perry	Juniata River	11-Jun-13	
	Hesperoperla pacifica	UT	Utah	South Fork Lower Provo River	23-Feb-12	✓
		UT	Washington	Leeds Creek	14-Oct-10	
	Neoperla Sp.	PA	Perry	Juniata River	11-Jun-13	
Perlodidae	*Isogenoides Sp.*	UT	Utah	Soldier Creek	12-Mar-13	
	Isoperla sobria	UT	Utah	South Fork American Fork River	20-Jul-13	
	Megarcys signata	UT	Utah	South Fork American Fork River	23-Feb-13	✓
	Skwala Sp.	UT	Summit	Upper Provo River	18-Sep-10	
Pteronarcyidae	*Pteronarcella badia*	UT	Utah	Soldier Creek	12-Mar-13	
	Pteronarcys proteus	PA	Clinton	Bear Creek.	9-Jun-13	

Table 4: Primer Testing (Amplification) Results

Locus	Relatioship	% Identity	# Individuals	Test PCR	Sequenced	Target	Variable
Plec761	Order	82	4	0/5			
PlecSK1	Order	86	4	5/5	5/5	5/5	Yes
Plec12	Order	98	5	1/5	1/1	Yes	
Plec600	Order	100	3	0/5			
Plec100	Order	100	7	0/5			
Plec628	Order	100	2	1/5	1/1	No	
Plec1790	Order	100	2	0/5			
Plec1810	Order	100	2	0/5			
PlecEF1	Order	70	4	1/5	1/1	Yes	
Plec57	Order	71	5	0/5			
Plec37	Order	74	3	0/5			
Plec73	Order	77	5	0/5			
Plec98	Order	77	4	0/5			
PlecMLC	Order	82	3	0/5			
Perl674	Superfamily	93	3	2/5	1/2	No	
Perl648	Superfamily	91	3	2/5	2/2	0/2	
Perl1205	Superfamily	99	3	0/5			
Perl1243	Superfamily	92	3	0/5			
Perl1552	Superfamily	91	3	0/5			
Perl845	Superfamily	91	3	0/5			
Perl534	Superfamily	91	3	3/5	3/3	3/3	Yes
Perl337	Superfamily	90	3	0/5			
Perl245	Superfamily	90	3	0/5			
PvC1026	Sister Families	94	2	0/5			
PvC2190	Sister Families	97	2	4/5	4/4	4/4	Yes

Table 5: Primer Specifications

Locus	Forward (5'-3')	Reverse (5'-3')	Target Length (bp)	Unedited Length (bp)	Edited Length (bp)	Tm ºC
PlecSK1	GTGGGCATGGGACAGAAG	TAGAAGCACTTGCGGTGGAC	995	900-1300	~870	55
Perl534	TGATTGCTTTTCGCCATGT	AGGTCGTCCTTCATATCTCCAC	264	200-500	~180	55
PvC2190	TTTGGCCTAGTGCATTTTAGTG	TGTTTGATTTTACAAACGGGAAG	520	550-1000	~400	55

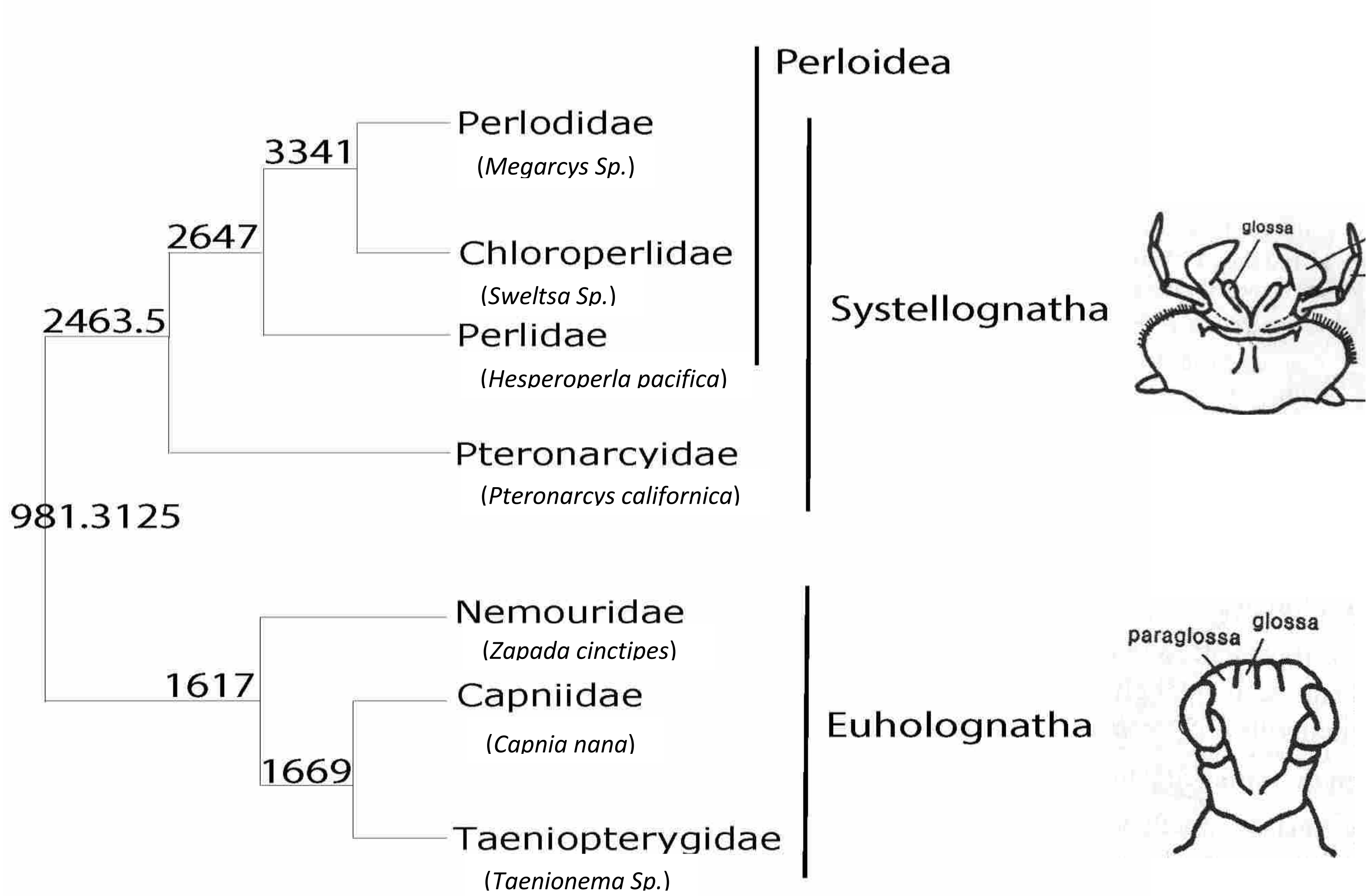

Figure 1: Transcriptome Content Phylogeny. The phylogeny is based on a distance matrix generated by applying a reciprocal BLAST algorithm in which each of the above taxon was BLASTed against every other taxon. For a gene or locus to be considered in common, the best blast hit must receive a score of 1e-40, be 1e-20 higher than the next highest hit, and the corresponding sequence must reciprocate and meet the same criteria. Images depicting the glossae and paraglossae of stoneflies from Merritt, Cummins, and Berg 2008.

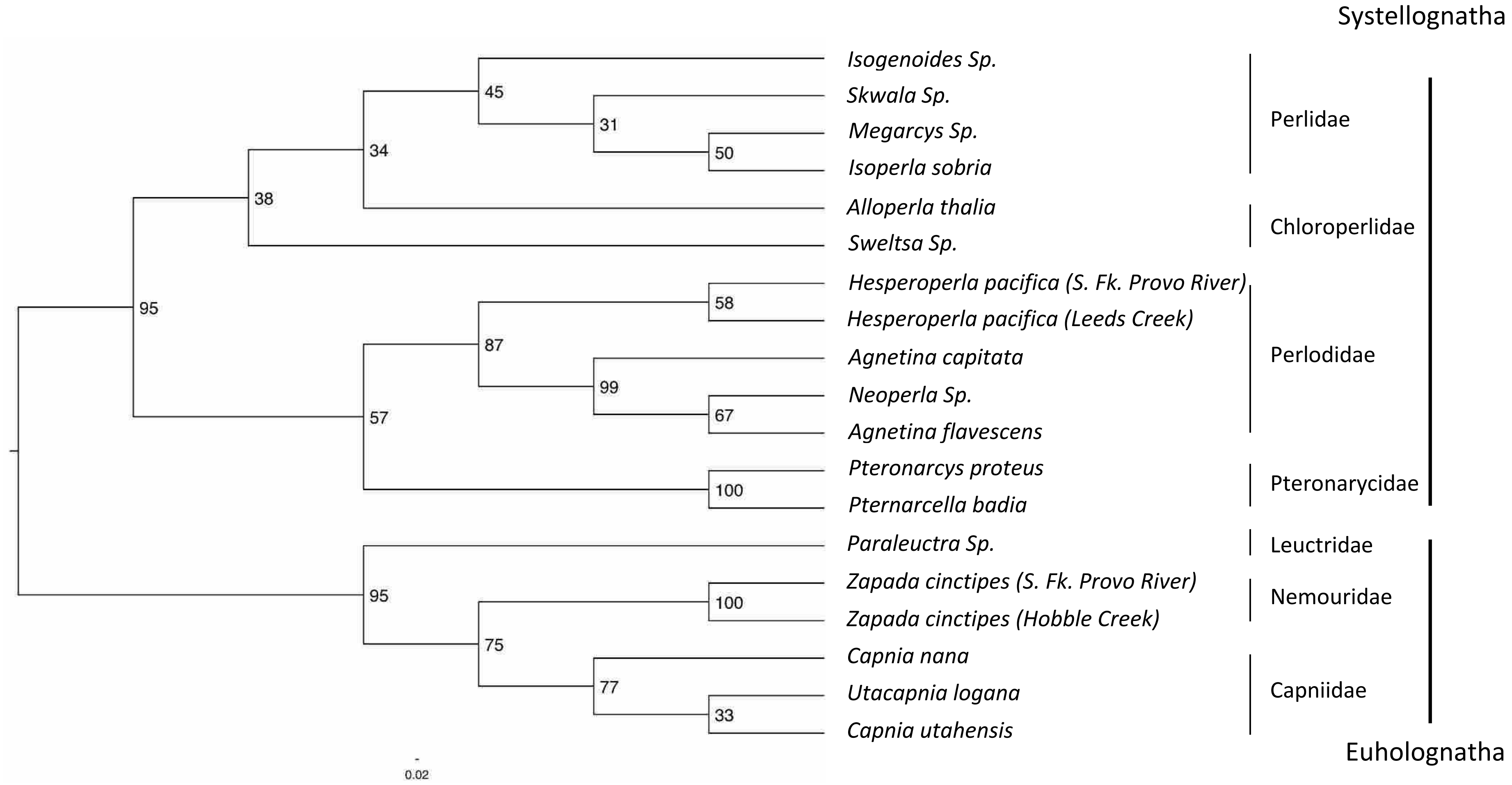

Figure 2: PlecSK1 Maximum Likelihood Phylogeny. ML phylogeny inferred from the single locus of PlecSK1 created using RAxML online at 100 bootstraps and PhyML at 1000 bootstraps both using the GAMMA+GTR rate of heterogeneity. The relationships were the same and support values nearly identical for the trees produced using the two different software.

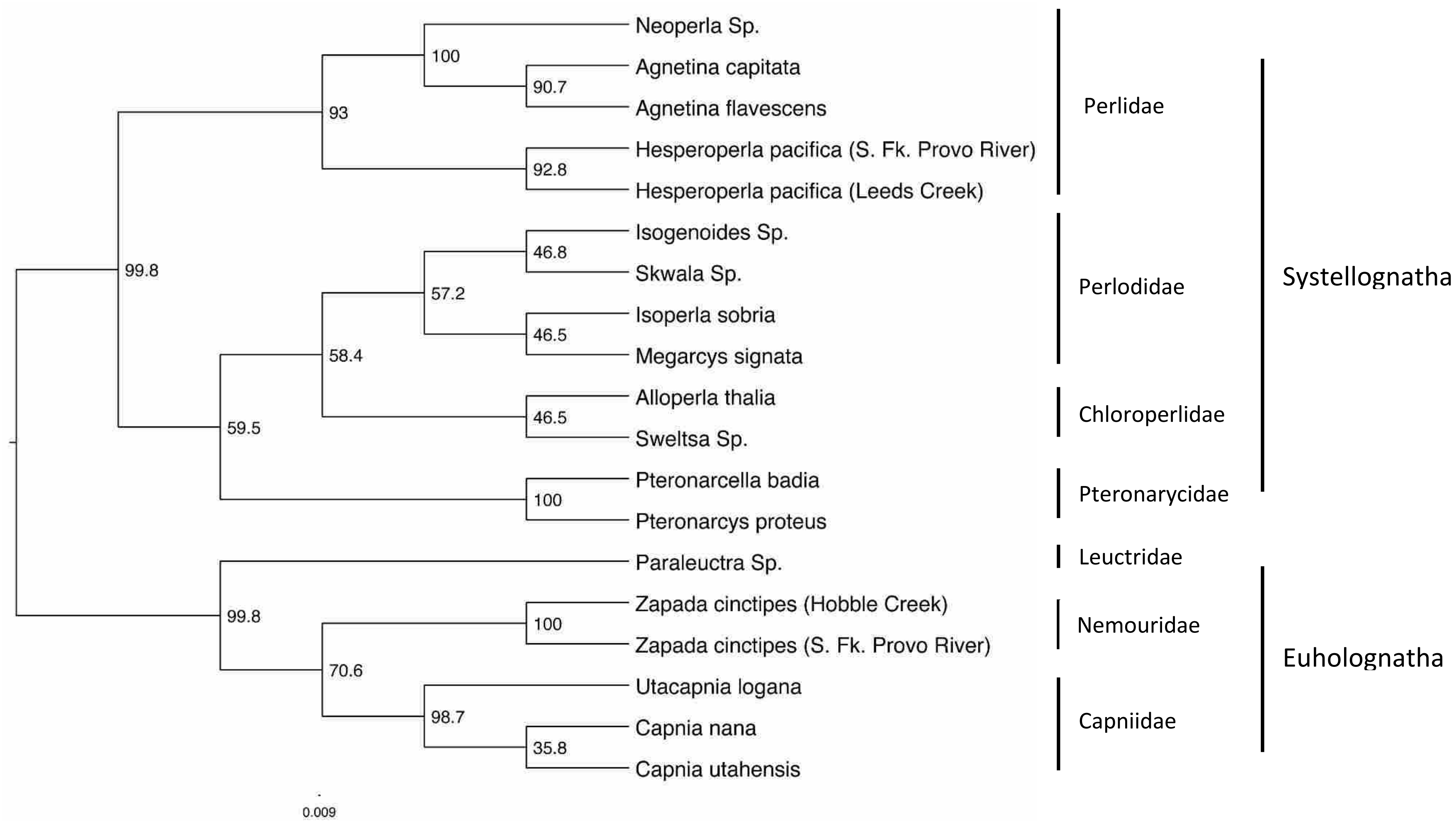

Figure 3: Neighbor Joining Phylogeny. This phylogeny was created using the Neighbor Joining method at 1000 bootstrap replicates employing the HKY and Jukes-Cantor genetic distance models. The two models produced exactly the same relationships with very minor differences in bootstrap values (the most being approximately by 3).

Appendix A: runBlast.pl

```
#!/fslhome/user/perl/bin/perl

use strict;
use warnings;

my $DIR = '/fslhome/user/directory/';
my @FILES = ('File1.fasta', 'File2.fasta', 'File3.fasta',
'File4.fasta', 'File5.fasta', 'File6.fasta', 'File7.fasta');
my $minE = 50;
my $diffE = 20;


# prepare BLAST indices
for my $fastaFile (@FILES) {
    `/fslapps/blast/blast-2.2.21/bin/formatdb -pF -i $fastaFile`;
}

# perform blasts (All files x all files)
for my $file1 (@FILES) {
    for my $file2 (@FILES) {
      `qsub -v
DIR=$DIR,FILE1=$file1,FILE2=$file2,MIN_E=$minE,DIFF_E=$diffE
blast.pl`;
    }
}

exit 0;
```

Appendix B: blast.pl

```
#!/fslhome/jtpage/perl/bin/perl

#PBS -l nodes=1:ppn=1:beta,pmem=6gb,walltime=24:00:00

use strict;
use warnings;

use Bio::SearchIO;

my $dir = $ENV{DIR} || '.';
my $file1 = $ARGV[0] || $ENV{FILE1};
my $file2 = $ARGV[1] || $ENV{FILE2};

#E-value criteria passed in from runBlast
my $minE = $ENV{MIN_E} || 30;
my $diffE = $ENV{DIFF_E} || 10;

my $blastFile = "$file1-$file2.blast";
my $outFile = "$file1-$file2.hits";

my $cutE = '1e-'.($minE - $diffE);
`/fslapps/blast/blast-2.2.21/bin/blastall -p blastn -e $cutE -d
$dir/$file2 -i $dir/$file1 -o $dir/$blastFile`;
my $blast = Bio::SearchIO->new( -file => "$dir/$blastFile", -format =>
'blast' );

open (OUT, ">$dir/$outFile");

#Utilizes blast files and generates hit files
while (my $result = $blast->next_result) {
    my $first = $result->next_hit;
    next unless (defined $first);
    my $exp1 = $first->expect;
    if ($exp1 == 0) {
      $exp1 = 255;
    }
    else {
      $exp1 =~ m/\de-(\d*)/;
      $exp1 = $1;
    }

#next blast result evaluated unless current result meets min e-value
    next unless ($exp1 >= $minE);

    my $second = $result->next_hit;
    my $exp2 = 0;
    if (defined $second) {
      $exp2 = $second->expect;
      if ($exp2 == 0) {
           $exp2 = 255;
      }
      else {
          $exp2 =~ m/\de-(\d*)/;
          $exp2 = $1;
      }
    }
#If second best blast result isn’t different enough it evaluates the
#next “gene’s” or “locus’” blast results
    next unless ($exp1 - $exp2 >= $diffE);

    print OUT join ("\t", $result->query_name, $first->name, $first-
>expect('exp'), "\n");
}

close (OUT);

exit 0;
```

Appendix C: mergeBlast.pl

```
#!/fslhome/user/perl/bin/perl

use strict;
use warnings;

use Bio::DB::Fasta;
use Bio::Tools::Run::Alignment::Muscle;
use Bio::AlignIO;

# Creates new link to MUSCLE
my $MUSCLE = Bio::Tools::Run::Alignment::Muscle->new();

# These arguments must be given at the command line with mergeBlast.pl
# Example: AllPlecoptera 1 1 File1.fasta File2.fasta File3.fasta

#This ID is used in naming all of the alignment files produced
my $ID = $ARGV[0];

# 0 means don’t perform alignments, 1 means do perform alignments
my $ALIGN = $ARGV[1];

# The number given here indicates how many species are allow to not
# participate in the reciprocal blast networks
my $MISSING = $ARGV[2];

my $FILE_START = 3;
my $MIN = scalar(@ARGV) - $FILE_START - $MISSING;

# Indicates minimum length allowed for alignments to be kept
my $MIN_LEN = 100;

my %fastas = ();
my %hits = ();

# load hits
for (my $i = $FILE_START; $i <= $#ARGV; $i++) {
    $fastas{$i} = Bio::DB::Fasta->new ($ARGV[$i], -reindex => 1);

    for (my $j = $FILE_START; $j <= $#ARGV; $j++) {
      my $hitsFile = "$ARGV[$i]-$ARGV[$j].hits";

      open (HITS, $hitsFile);
      while (<HITS>) {

          chomp;
          my ($query, $hit) = split (/\t/, $_);

# Builds a “hash” tree or web of relationships of edges and nodes
          $hits{$i}{$query}{$j} = $hit;
      }
      close (HITS);
    }
}



# check edges
my $countBlast = 0;

for (my $startPoint = $FILE_START; $startPoint <= $FILE_START +
$MISSING; $startPoint++) {
  FIRST: for my $first (keys %{ $hits{$startPoint} }) {
      my @group = ();


      # avoid redundancy: ensures that this network of genes hasn't
      # already been evaluated
      for (my $i = $FILE_START; $i < $startPoint; $i++) {
        next FIRST if (defined $hits{$startPoint}{$first}{$i});
      }


    SECOND: for (my $i = $startPoint; $i <= $#ARGV; $i++) {
      my $second = $hits{$startPoint}{$first}{$i};

      next SECOND unless (defined $second);


      # check reflexiveness and symmetry
      next SECOND unless (defined $hits{$i}{$second});
      next SECOND unless (defined $hits{$i}{$second}{$i});
      next SECOND unless ($second eq $hits{$i}{$second}{$i});
# reflexive

      next SECOND unless (defined $hits{$i}{$second}{$startPoint});
      next SECOND unless ($first eq $hits{$i}{$second}{$startPoint});
# symmetric

      # check consistency of other edges
      for (my $j = $FILE_START+1; $j <= $i; $j++) {
          my $third = $hits{$i}{$second}{$j};
          next SECOND unless (defined $third);
          next SECOND unless (defined $hits{$j}{$third});
          next SECOND unless (defined $hits{$j}{$third}{$i});
          next SECOND unless ($second eq $hits{$j}{$third}{$i});
    }

      #Retrieves sequences by name and organizes them for alignment
      my $seq1 = $fastas{$i}->get_Seq_by_id
($hits{$startPoint}{$first}{$i});
      next unless (defined $seq1);
```

```
     my $seq2 = Bio::LocatableSeq->new ( -id => "$ARGV[$i]\_\_".$seq1-
>id, -seq => $seq1->seq);
     push (@group, $seq2);
    }


      # good gene (BLAST)
      if (scalar (@group) >= $MIN) {
       $countBlast++;
       my $gene = "$ID\_$countBlast";

       my $aln;
       if ($ALIGN) {
           $aln = $MUSCLE->align (\@group);

           my $min_length = $MIN_LEN;
           for my $seq ($aln->each_seq) {
            $min_length = $seq->length if ($seq->length <
$min_length);
           }

           next unless ($min_length >= $MIN_LEN);
       }
           # if MUSCLE off, places unaligned seqs in group file
       else {
           $aln = Bio::SimpleAlign->new;
           for my $seq (@group) {
            $aln->add_seq ($seq);
           }
       }

       my $outId = "$gene.fasta";

           # incorporates alignment %ID into naming of alignments
       $outId = int ($aln->percentage_identity) ."__$outId" if
($ALIGN);
       my $alnOut = Bio::AlignIO->new( -file => ">$outId", -format =>
'fasta' );
       $alnOut->write_aln ($aln);
      }
  }
}

exit 0;
```