Available online at www.sciencedirect.com
ScienceDirect
Zoologischer Anzeiger xxx (2011) xxx–xxx
Zoologischer Anzeiger
www.elsevier.de/jcz

Directed next generation sequencing for phylogenetics: An example using Decapoda (Crustacea)

Seth M. Bybee[a,*], Heather D. Bracken-Grissom[a], Russell A. Hermansen[a], Mark J. Clement[b], Keith A. Crandall[a], Darryl L. Felder[c]

[a]*Department of Biology, 401 WIDB, Brigham Young University, Provo, UT 84602, USA*
[b]*Computer Science Department, 3370 TMCB, Brigham Young University, Provo, UT 84602, USA*
[c]*Department of Biology, PO Box 42451, University of Louisiana at Lafayette, Lafayette, LA 70504, USA*

Received 11 September 2010; received in revised form 30 May 2011; accepted 31 May 2011

Abstract

We propose a method using next generation sequencing technology for phylogenetics. The method is PCR based, requires little training beyond basic lab skills and is both cost and time effective. With this method we generated data for and produced a phylogeny of Decapoda that demonstrates this method's potential, the quality of the data, and the ability for the method to fit or even replace current Sanger based methods of generating DNA data for phylogenetic reconstruction. Finally, we discuss advantages and current challenges of the directed next generation sequencing approach.

Key words: Crustaceans; Phylogenetics; Systematics; Amplicon; Targeted sequencing

1. Introduction

Encompassing approximately 15,000 extant taxa (De Grave et al., 2009), the order Decapoda represents a diverse and species-rich group of crustaceans. Familiar representatives include the crabs, lobsters, shrimp, crayfish, and hermits, which display a broad array of body forms and functions. The morphological diversity of decapods coupled with their economic importance in world fisheries and aquaculture make them an important focus for evolutionary and ecological studies.

Phylogenetic relationships among decapod crustaceans have been studied extensively for several decades (Scholtz and Richter, 1995; Ahyong and O'Meally, 2004; Schram and Dixon, 2004; Porter et al., 2005; Tsang et al., 2008; Bracken et al., 2009; Toon et al., 2009). Competing phylogenetic hypotheses have been generated from morphological and molecular studies, in part due to the variable quality of data used to generate phylogenies. Although reproductive, adult, and larval characters have been scarcely applied in earlier studies, molecular genetic markers have gained recent popularity because of challenges presented by more traditional methods (e.g., finding and coding homologous morphological characters that are phylogenetically informative across the order).

In recent years molecular studies have contributed greatly to our knowledge of decapod evolution, though they are often limited by scope of taxon and/or molecular character sampling. The amount of time, money, and effort required to generate sequence data across such a diverse and ancient

[*] Corresponding author.
E-mail address: seth.bybee@gmail.com (S.M. Bybee).

0044-5231/$ – see front matter © 2011 Published by Elsevier GmbH.
doi:10.1016/j.jcz.2011.05.010

group poses a challenge to decapod biologists, and finding a set of universal molecular markers that can be applied across such a large, old, and morphologically disparate group is daunting. Over the last five years, the Decapod Tree of Life project has drawn upon the combined efforts of several labs and many collaborators to define an established set of genetic markers (18S, 28S, H3, 16S, 12S, COI) for study of the group. Varied combinations of these nuclear and mitochondrial genes have proven informative at fine versus coarse evolutionary scales (taxonomic levels). The success of large-scale projects, such as the Tree of Life Decapoda, depends on the ability to generate sequence data from museum material and older ethanol preserved tissues. With the application of next generation sequencing techniques (hereafter referred to as next-gen sequencing), we can generate massive amounts of data for these "established" genes. This method is fully scalable for both EToH preserved and older museum specimens in a faster and cheaper fashion than ever before. As a result, the time, effort, and money needed to perform large-scale phylogenetics is likely to be diminished and access to the benefits of next-generation technology for nearly all labs currently doing molecular phylogenetics is possible.

Next-gen sequencing has the potential to revolutionize evolutionary biology. New advancements in molecular tools and techniques are allowing researchers to more economically generate vast amounts of data in a relatively small amount of time. The basic premise of 454 pyrosequencing is to take short segments of gDNA, PCR products, BACs, and cDNA and add an adapter to both ends via standard molecular biology techniques (i.e., ligation). These adapters then serve to help purify, amplify and sequence the desired segments. Following adapter attachment the dsDNA is separated into ssDNA segments and each individual ssDNA fragment is attached to a single DNA capture bead (i.e., one unique ssDNA fragment per one DNA capture bead) via the attached adapter. Each bead containing its unique ssDNA fragment is then emulsified in amplification reagents and a water-in-oil mixture (emulsion PCR; emPCR) for amplification. During emPCR the ssDNA fragment is amplified in parallel, resulting in several million copies of the ssDNA fragment bound to the bead. These beads are then loaded, one per well, on a PicoTiterPlate device where they are sequenced using chemiluminescent signals reflecting (i.e., pyrosequencing) recorded with by a CCD camera. Next-gen sequencing has been used extensively to construct entire genomes from a wide array of organisms but the ability to use it as a standard sequencing tool in place of traditional Sanger methods (using fluorescently labeled dideoxynulceotide triphosphates as chain terminators) has been less researched. Directed or targeted sequencing represents a tool that uses a next-gen platform (e.g., Ilumina, 454, etc.) to sequence a given product resulting from a polymerase chain reaction (PCR) or amplicon. Several approaches have been put forward to harness the power of next-gen for directed sequencing (e.g., Binladen et al., 2007; Crosby and Criddle, 2007; Meyer et al., 2007, 2008; Pertoldi et al., 2009). Here we apply a PCR based method for directed sequencing using the 454 pyrosequencing platform (http://www.454.com/products-solutions/how-it-works/index.asp) and discuss the implications for phylogenetic analyses. Decapod crustaceans represent an ideal group for this study and will highlight the benefits of applying this method across a species-rich group where sets of universal markers are proven to be diagnostic and informative.

2. Materials and methods

2.1. Taxon sampling

We selected 16 taxa from within the Decapoda that represented the major lineages (Table 1). To stabilize and root the phylogenetic tree the stomatopod, *Lysiosquillina maculata* (Fabricius, 1793) was used as an outgroup during phylogenetic reconstruction. Selected taxa had associated sequence data generated via traditional Sanger DNA sequencing. The inclusion of Sanger sequences allowed for (1) a direct comparison to next-gen data in order to assess the quality of DNA data produced by directed sequencing using the next generation platform, (2) an alignment "anchor" for short sequences generated by next-gen sequencing, and (3) an examination of how sequences generated with Sanger and next-gen technology would act together within a single phylogenetic estimation. While we do not infer that Sanger sequences are necessary to generate an accurate phylogeny, they do serve as a diagnostic tool to comparatively test the performance of our method.

2.2. Genes

The decapod community commonly works varied combinations of six genes that are relatively easy to isolate, amplify, and sequence (via Sanger methods) across diverse groups within the order. These genes represent a range of phylogenetic utility and can be used to resolve infraordinal to species level relationships (for a few examples see Porter et al., 2005; Bracken et al., 2009; Felder and Robles, 2009; Robles et al., 2009; Thoma et al., 2009; Toon et al., 2009). They include: 16S, large mitochondrial ribosomal subunit (~550 bp, Crandall and Fitzpatrick, 1996); 12S, small mitochondrial ribosomal subunit (~400 bp, Buhay et al., 2007); 18S, small nuclear ribosomal subunit (~1900 bp, Whiting et al., 1997; Whiting, 2002); 28S, large nuclear ribosomal subunit (~2500 bp, Whiting et al., 1997; Whiting, 2002; Toon et al., 2009); H3, nuclear protein-coding gene (~330 bp, Colgan et al., 1998); and COI, mitochondrial protein-coding gene (~600 bp, Folmer et al., 1994). We amplified each of these gene regions via directed sequencing and/or Sanger sequencing protocol below (see Table 1 for GenBank accession numbers).

Table 1. Table of taxon sampling.

	Voucher Number	12S	16S	COI	18S	28S	H3
Outgroup							
Lysiosquillina maculata (Fabricius, 1793)	KC3832	x	x		x	x	x
Ingroup							
Dendrobranchiata							
Deosergestes corniculum (Krøyer, 1855)	ULLZ11598/KC6206	x	x		x	x	
Deosergestes corniculum[a]	ULLZ11598/KC6206				x	x	x
Litopenaeus setiferus (Linnaeus, 1767)	ULLZ11629/KC6204	x	x		x	x	
Gennadas scutatus Bouvier, 1906[a]	ULLZ11476/KC6203	x			x		x
Pleoticus robustus (Smith, 1885b)	ULLZ10956/KC6205	x	x		x	x	
Pleoticus robustus[a]	ULLZ10956/KC6205				x		x
Caridea							
Notostomus gibbosus (A. Milne-Edwards, 1881)[a]	ULLZ11481/KC6197				x	x	x
Crangon crangon (Linnaeus, 1758)	KC3052	x	x		x	x	x
Glyphocrangon nobilis (A. Milne-Edwards, 1881)	ULLZ11024/KC6196	x	x		x	x	
Glyphocrangon nobilis[a]	ULLZ11024/KC6196	x	x		x		x
Axiidea							
Pseudogourretia sp.	ULLZ11472/KC6177	x	x	x	x	x	
Callianassa aqabaensis (Dworschak, 2003)[a]	ULLZ7924/KC5826				x	x	x
Stenopodidea							
Stenopus hispidus (Olivier, 1811)	KC4276	x	x		x	x	x
Stenopus hispidus[a]	KC4276	x			x	x	x
Polychelida							
Stereomastis sculpta (Smith, 1880)[a]	ULLZ8022/KC5840				x		x
Polycheles typhlops (Heller, 1862)	ULLZ8051/KC5846	x	x	x	x	x	
Polycheles typhlops[a]	ULLZ8051/KC5846	x		x	x	x	x
Achelata							
Projasus bahamondei (George, 1976)[a]	KC3207				x		x
Scyllarus americanus (Smith, 1869)[a]	ULLZ8500/KC5845	x	x		x	x	x
Scyllarus depressus (Smith, 1881)	ULLZ8168/KC5850	x	x	x	x	x	
Astacidea							
Nephrops norvegicus (Linnaeus, 1758)	KC2163	x	x		x	x	x
Homarus americanus (H. Milne Edwards, 1837)	KAChoam		x		x	x	x
Astacus astacus (Linnaeus, 1758)	JF134		x		x	x	x
Procambarus clarkii (Girard, 1852)	KC1497	x	x	x	x	x	x
Procambarus liberorum (Fitzpatrick, 1978)[a,b]	USNM260016				x		x
Procambarus spiculifer (LeConte, 1856)[a]	KC4054				x	x	x
Anomura							
Albunea gibbesii (Stimpson, 1859)	ULLZ7316/KC4753	x	x		x	x	
Xylopagurus cancellarius (Walton, 1950)	ULLZ9443/KC4783	x	x		x	x	
Galacantha valdiviae (Balss, 1913)	KC3102	x	x		x	x	x
Brachyura							
Ala cornuta (Stimposn, 1860)[a]	ULLZ9065/KC5791				x	x	x
Cancer pagurus (Linnaeus, 1758)	KC2158		x		x	x	x
Dyspanopeus sayi (Smith, 1869)[a]	ULLZ7227/KC5851		x		x	x	x
Cycloes granulosa (De Haan, 1837)	KC3082	x	x		x	x	x
Garthiope spinipes (A. Milne-Edwards, 1880)[a]	ULLZ7840/KC5857				x		x

[a] Sequences generated via 454.

[b] Museum specimen.

2.3. Sanger sequencing

Total genomic DNA was extracted from the abdomen, gills, pereopods or pleopods using the Qiagen DNeasy® Blood and Tissue Kit (Cat. No. 69582). Gene regions were amplified by means of PCR using one or more sets of primers (see references above). Reactions were performed in 25 μl volumes containing 10 μM forward and reverse primer for each gene, 2.5 μM each dNTP, PCR buffer, magnesium chloride, 1 unit HotMasterTaq polymerase (5 PRIME), and 30–100 ng/μL extracted DNA. The thermal cycling profile conformed to the following parameters: Initial denaturation for 1–2 min at

94 °C followed by 25–40 cycles of 1 min at 94 °C, 1 min at 46–58 °C (depending on gene region), 1 min 30 s at 72 °C and a final extension of 10 min at 72 °C. PCR products were purified using filters (PrepEase^{TM} PCR Purification 96-well Plate Kit, USB Corporation) and sequenced with ABI BigDye$^{®}$ terminator mix (Applied Biosystems, Foster City, CA, USA). An Applied Biosystems 9800 Fast Thermal Cycler (Applied Biosystems, Foster City, CA, USA) was used in PCR and cycle sequencing reactions, and sequencing products were run (forward and reverse) on an ABI 3730xl DNA Analyzer 96-capillary automated sequencer. Sequences were assembled, cleaned, and edited using the computer program Sequencher 4.8 (GeneCodes, Ann Arbor, MI, USA).

2.4. Directed sequencing

The process for preparing amplicons (targeted DNA region) for directed next-gen sequencing required two PCRs and a template DNA (in our case gDNA) (Fig. 1). The first PCR (PCR I) used a locus specific primer coupled with a 22 base pair (bp) adapter (total primer length of ∼40–45 bp). A standard three step PCR protocol was used [(94 °C: 2:00 (94 °C: 1:00; 50 °C: 1:00; 72 °C: 1:15 × 25) 72 °C: 7:00)] with HotMasterTaq polymerase (5 PRIME) as outlined above but with only 10 μL reactions. After the initial PCR, successful amplicons had the known adapter incorporated into the 5′ and 3′ ends. These amplicons were gene cleaned with a PrepEase^{TM} PCR Purification 96-well Plate Kit (USB Corporation) and vacuum manifold. One microliter of cleaned PCR product was used as the template for the second PCR (PCR II). The PCR II protocol and reagents followed PCR I. The primers used in PCR II consisted of a complimentary 3′ end adapter coupled with a 10 bp "barcode" or multiplexing identifier (MID), 4 bp key, and 21 bp 454 Titanium primer at the 5′ end. Each taxon is assigned a unique 10 bp MID that is attached during PCR II to all PCR I products generated for that taxon (e.g., Taxon A is assigned MID 3 and all PCR II products for all loci amplified for taxon A will have MID 3 incorporated). Following PCR II, samples were again gene cleaned using the millipore system. Directly following the gene clean of PCR II the quantity of DNA (in ng) was estimated via a ladder and gel electrophoresis. Every effort was made to use PicoGreen to get the most accurate estimate of DNA quantity but a large ∼200 bp primer dimer, present in nearly every PCR product prevented an accurate reading. Samples were pooled according to targeted amplicon and quantity to a total of more than 500 ng of total DNA and subsequently submitted to the DNA sequencing facility at Brigham Young University, where samples were further purified in an effort to remove primer dimers and size-select the proper amplicon via gel purification and AMPure beads. Following purification, amplicons were combined in an emPCR (emulsion PCR) and subsequently sequenced via 454 GS FLX Titanium pyrosequencing technology (Roche) at Brigham Young University.

2.5. Computation of raw next-gen DNA data

In order to deal with the large amount of data generated from the next-gen sequencing run, we developed a program pipeline to computationally clean, trim, group and analyze the data dubbed the BarcodeCruncher Pipeline (available from the authors upon request). The BarcodeCruncher Pipeline began by extracting the raw DNA sequence reads from the SFF output files generated from the 454 sequencer and separating each read into a fasta file by its accompanying barcode. The separation process was performed by using the "ssffile" program found in the Genome Sequence Data Analysis Software package (http://www.genome-sequencing.com/). An error rate of one nucleotide was allowed in the barcode to diminish the amount of lost read data. Each read was then separated into individual fasta files, for further contaminant control and for adapter and primer removal. This process left the reads clean of any unwanted sequence data that were a result of the PCR and sequencing processes. To further reduce sample contamination a BLAST (Altschul et al., 1990) search was performed on each read. Because the genes being used in the experiment were known, a list of query sequences gathered from GenBank was created for BLAST comparison. All reads scoring an *e*-value greater than $1e{-}3$ were discarded, under the assumption that these were either poor reads or contaminant reads. Since read length varied considerably and short reads stood the possibility of randomly matching in BLAST comparison against the query samples, only reads of 50 bp or longer were used in the analysis.

Once contamination was identified and excluded, contigs were assembled from the individual sequence reads according to barcode/MID. The contigs were created using the GS *De Novo* Assembler found in the Genome Sequence Data Analysis Software package (http://www.genome-sequencing.com/). Once all of the contigs were created, each contig was then mapped against the query sequences to determine the gene region. This process was done using the Genomic Next-generation Universal Mapper (GNUMAP) (Clement et al., 2009); a next-generation sequence mapping program developed by Brigham Young University. To improve mapping results, each query subject was elongated by placing "N's" at the beginning and end of the sequence. This was done to enhance GNUMAP's ability to accurately map the longer barcode contigs to the query sequences. Contigs that did not map against any of the query sequences were discarded from further analysis. Following contig mapping, the final function of BarcodeCruncher was to compress each mapped contig into a consensus sequence and deposit each into a fasta file for each of the corresponding genes selected for the analysis (e.g., 28S mapped consensus sequences were all combined into a single fasta file). The name of each barcode was then replaced by the actual name of the corresponding taxon, outputting the final named fasta files into a directory for alignment and phylogenetic analyses.

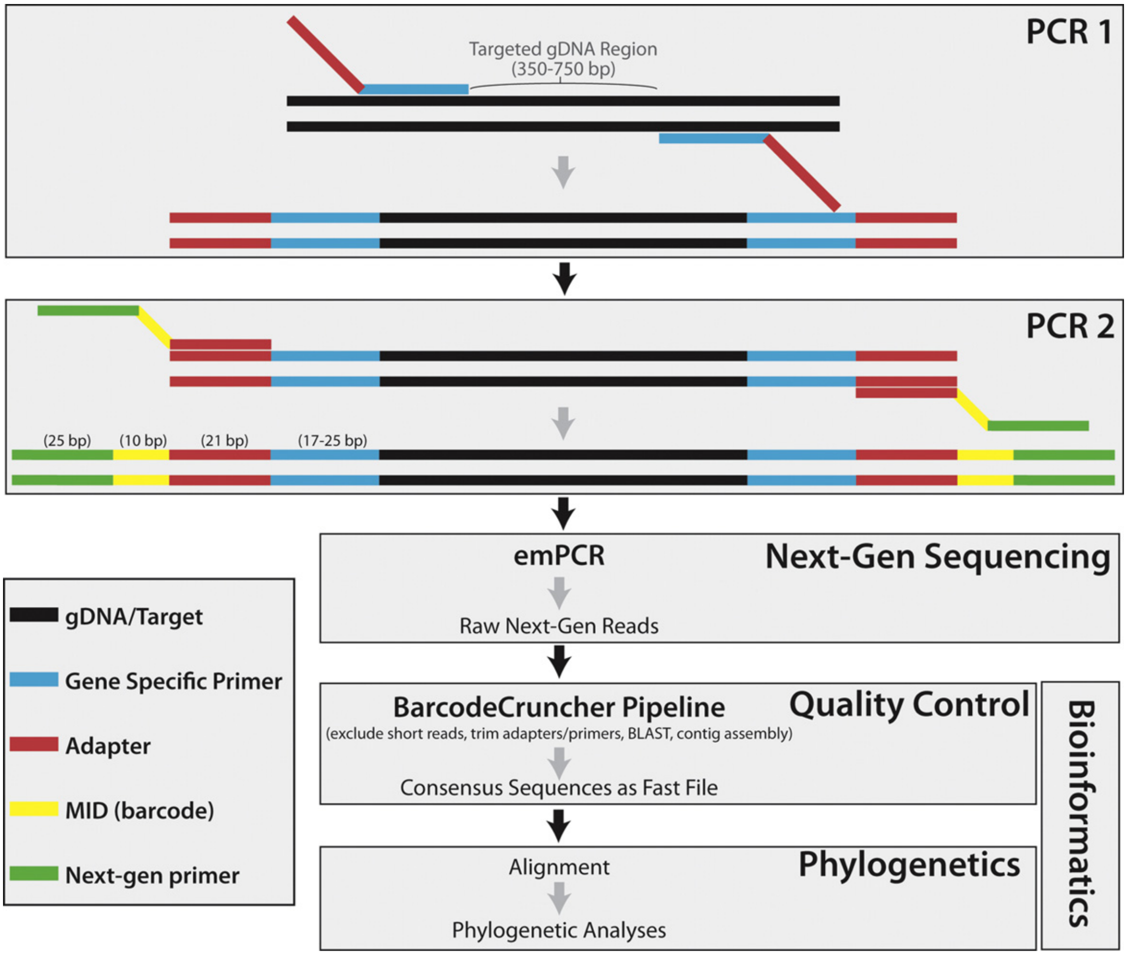

Fig. 1. Flow diagram of laboratory and bioinformatic methods of directed DNA sequence generation for phylogenetic analyses. Two-step PCR method used to attach the adapter (PCR I) and MID and Titanium 454 primers (PCR II). Bioinformatic analyses used to produce a consensus sequence for phylogenetic reconstruction.

To streamline the use of BarcodeCruncher and to allow easy access of data into the program, BarcodeCruncher was configured to use a control file. Use of the "-control" option on the command line automatically created the template for the control file. The control file accepted all of the data necessary to correctly run the experiment and create usable barcode data. All analyses were performed at the Fulton Supercomputing Lab at Brigham Young University on the marylou5 supercomputer. Completion time for the pipeline from beginning to end was approximately 5 h using four processors on the BYU core supercomputer.

2.6. Phylogenetic methods

The fasta file output from the BarcodeCruncher pipeline contained taxa in the study found to correctly match against a specific gene. To take our study to its end-point, a phylogenetic analysis was performed. To demonstrate that the barcoding procedure had correctly mapped the reads to the correct gene, sequences created using Sanger sequencing (either downloaded from GenBank or new; Table 1) were added to the analysis. Each gene file was then aligned using MAFFT v6.713b (Katoh et al., 2005). The "E-INS-i" alignment option was used for all alignments, since it was suspected that the relatively small sized barcode sequence reads could potentially align into multiple conserved domains with long gaps in between. To further enhance the quality of the alignments and to eliminate regions of poor arrangement, each alignment was curated using Gblocks v0.91b (Castresana, 2000). Options allowing for a less stringent blocking of the alignments were selected to decrease the amount of data lost to curation. All gene alignments were then concatenated with one another to create a partitioned dataset.

Phylogenetic trees were created using RAxML 7.0.4 (Stamatakis et al., 2005), a fast maximum-likelihood phylogenetic program. The algorithm used in the analysis was the "-f a" option, for a rapid bootstrap analysis and search for the best tree in a single pass. Likelihood settings followed the General Time Reversible Model (GTR) with a gamma distribution and invariable sites. RAxML estimated all free

Please cite this article in press as: Bybee, S.M., et al., Directed next generation sequencing for phylogenetics: An example using Decapoda (Crustacea). Zool. Anz. (2011), doi:10.1016/j.jcz.2011.05.010

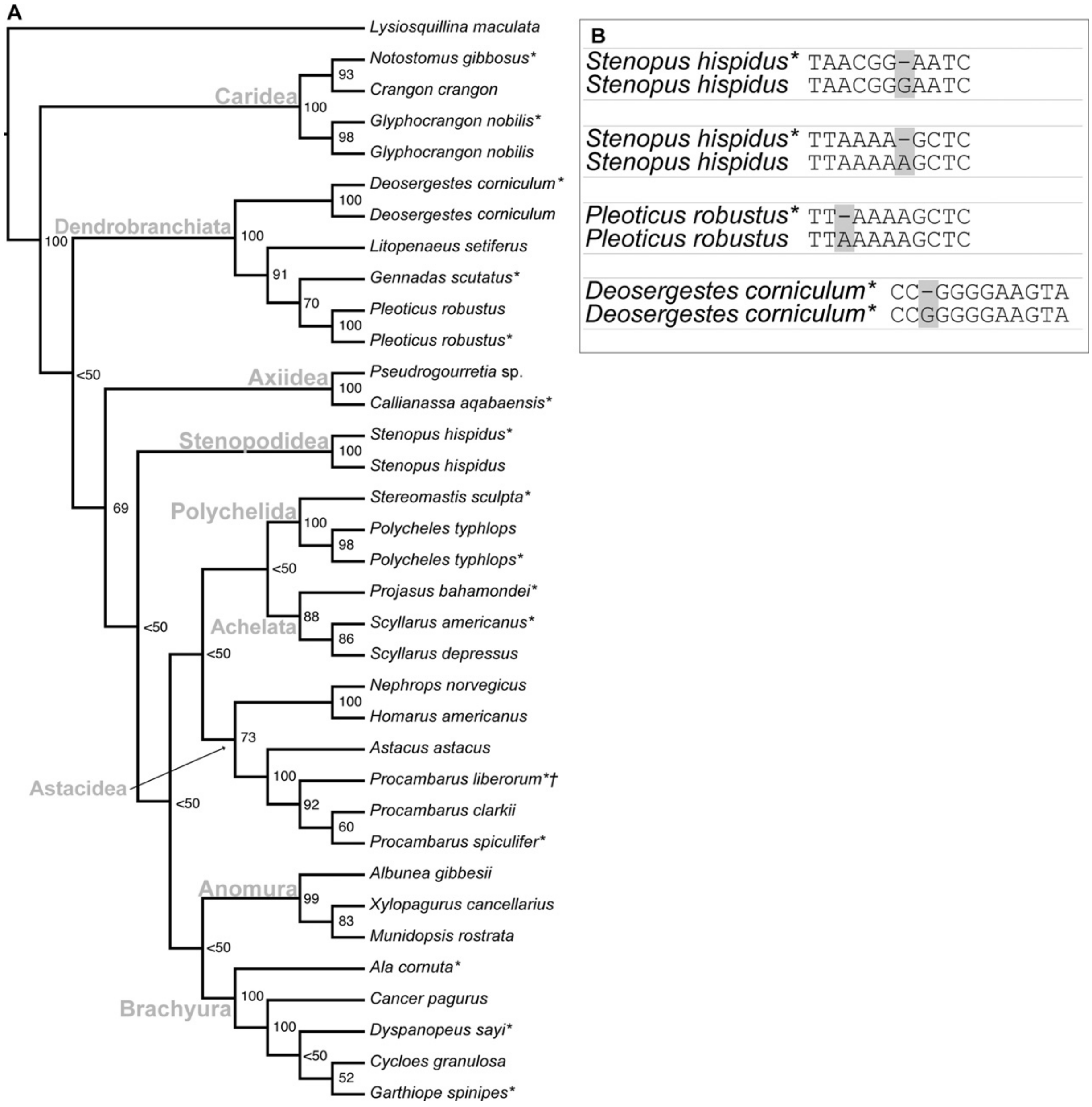

Fig. 2. (A) Phylogeny of Decapoda generated via RAxML under the GTR + G model. Support values are bootstraps. Sequence data generated via next-gen sequencing denoted with a star. The dagger denotes a museum specimen stored in 70% ethanol at room temperate for ~18 years. (B) Represents examples taken from snippets of the alignment demonstrating the errors associated with 454 pyrosequencing technology.

parameters following a partitioned dataset. Confidence in the resulting topology was assessed using non-parametric bootstrap estimates (Felsenstein, 1985) with 10,000 bootstraps (bs) values >50% are presented on the resulting phylogeny (Fig. 2A).

3. Results

3.1. Phylogenetic relationships

Since our study is not intended to robustly explore the relationships among the Decapoda nor test the monophyly of decapod clades, sampling was limited to representatives from major lineages (Fig. 2A). Even so, there is statistical support for all pleocyemate infraorders: Caridea (bs = 100); Axiidea (bs = 100); Stenopodidea (bs = 100); Polychelida (bs = 100); Achelata (bs = 88); Astacidea (bs = 73); Anomura (bs = 99); Brachyura (bs = 100). Additionally, the suborder Dendrobranchiata (bs = 100) and all major lineages within this clade are significantly supported: Penaeoidea (bs = 91), and Sergestoidea (bs = 100). Infraordinal relationships are unresolved in our phylogeny, while there is some support for a clade uniting Axiidea, Stenopodidea, Polychelida, Achelata, Astacidea, Anomura, and Brachyura (bs = 69). A monophyletic Reptantia and Pleocyemata, by current definition, were not recovered. Although with low support, the lobster-like decapod lineages form a clade (Polychelida, Astacidea, Achelata) and the anomurans (hermit, porcelain, king crabs) fall as the sister group to the brachyurans (true crabs) (anomuran + brachyuran = Meiura). These findings are in accord with recent morphological and molecular studies (Ahyong and O'Meally, 2004; Tsang et al., 2008; Bracken et al., 2009). However, our results must be interpreted with caution as bootstrap values are low and taxon sampling is limited.

3.2. Results of data quality

Read quality and length were of high concern in the study as any major problems in data quality or length will compromise phylogenetic results. Average read length, excluding titanium primer, adapter and barcode for the raw data was 289 bp. After the raw data were processed via the BarcodeCruncher Pipline the average read length improved to 561 bp with an average coverage of 330 reads per targeted gene region. We also examined the similarity between sequences generated via 454 and those generated via Sanger sequencing to determine next-gen read quality. To do this we included five couples where existing Sanger sequence data was available to be directly compared to 454 sequences amplified from the same gDNA extraction. The error percentage rate between aligned next-gen and Sanger sequence couples was 1.09%. A close examination of the difference between next-gen and Sanger sequence data in the alignment made it clear that nearly all differences between the Sanger and 454 generated sequences can be attributed to incorrect base pair calls due to homopolymer repeats, mistakes inherent to 454 pyrosequencing technology resulting in small insertions or deletions of nucleotides (Fig. 2B). An adjusted error rate where homopolymer differences are excluded is 0.05%. By comparing the percent difference between these couples, we identified two contaminant sequences and the presence of a NUMT (nuclear mitochondrial DNA) (see Table 1). The contaminant sequences appear to be from barnacles and it is not certain where the contaminant DNA entered our samples (e.g., as a result of cross contamination or as parasites on a decapod specimen that was incidentally co-extracted during gDNA extraction or PCR). The NUMT was identified by the presence of several stop codons in the translated COI sequence.

4. Discussion

4.1. Phylogenetics

The objective of this study was to highlight the potential for next generation sequencing within a species-rich and diverse group of organisms. Although this study was not intended to deeply explore phylogenetic relationships among the Decapoda and thus included a very limited taxon sampling, many findings are in agreement with recent molecular and morphological studies. Applied to phylogenetic questions, this method provides the ability to sequence hundreds of taxa for hundreds of genes (or even thousands if scaled up using robotics), thus allowing for more efficient generation of standard sequence data than by currently used Sanger-based methods. Further, as the average read length increases for next-gen platforms, the number of universal primers needed to generate sequence data for a large gene region (e.g., 18S, ~1800 bp) across a diverse group of organisms will be greatly reduced (e.g., future technologies will be capable of up to 900 bp per read). In short, with increased read length the method outlined herein becomes even more potent, especially as the need for truly universal primers become greatly reduced.

Contamination appears to be a potential problem of the method. It seems likely that the contaminants were amplified during PCR I and during PCR II the barcode for the targeted taxon was attached prior to 454 pyrosequencing. It could also be that contaminating gDNA was in the original extracts from decapod taxa because of cross-contamination or perhaps co-extraction of a parasite associated with the decapod subject specimen. Whatever the scenario for contamination, our findings fall in line with Binladen et al. (2007) and further demonstrate the sensitivity of 454 technology, including its tendency to detect and amplify contaminant sequences. Another interesting finding was the presence of a nuclear mitochondrial DNA or NUMT. The presence of NUMTS in arthropod taxa has been well documented (Song et al., 2008) and was a concern for this protocol (as it is to any PCR based method of sequence generation). Only one NUMT was detected in the data set because of the presence of several stop codons in the translated COI sequence. Read length among next-gen sequencing platforms is a major limiting factor to studying NUMTs (pers. commun., H. Song), but again, as the technology allows for longer read lengths, a method such as here outlined could certainly be effective in elucidating NUMT evolution among plants and animals.

4.2. Potential for museum samples

One powerful potential of this method is the sensitivity of next-gen sequencing in generating quality sequence data from museum material. Often, museum samples can play critical roles in taxonomic and systematic questions with implications in conservation biology (e.g., Crandall et al., 2009). We extracted gDNA from a specimen that had been stored in 70% ethanol at room temperature for 18 years. While sequencing using Sanger method of such materials does frequently succeed, repeated attempts to amplify quality gDNA from this specimen by this method failed. By using the 454 and the two-step PCR protocol we were able to generate sequence data for this specimen. While this was the only museum specimen included in the present series of 454 pyrosequence analyses, it encourages us to apply our protocols to museum specimens with varied preservation histories. Should these methods succeed in broader application to museum specimens, they could potentially overcome one of the larger constraints on present molecular phylogenetic studies of decapod crustaceans. Thus, the potential for using next-gen sequencing coupled with our direct sequencing protocol seems promising for studies that would benefit by including genetic data from museum specimens.

4.3. Cost and time effectiveness

Because this method is based on two PCRs with the end product ready to be directly sequenced using a next-gen platform, it is possible to go from gDNA to 454 barcoded amplicon library in 8–10 h depending on lab equipment (e.g., thermocyclers) and competency of laboratory personnel. This is compared with the current 2–3 days to prepare a similar library with equivalent methods. This method is also cost effective when compared to Sanger method of sequencing. We estimate that the sequencing cost of a single read of 561 bp (our average read length) using our method is $0.07 and that 25× coverage is required to be certain of sequence quality, making the total cost per sequence $1.75 ($0.003 per bp). The internal rate for Sanger sequencing of both a forward and reverse sequence at the BYU Sequencing Center is $2.86 ($0.004 per bp), one of the lowest rates in the country, and usually results in the equivalent of two 650 bp reads. Thus, our method represents a significant saving in both time (∼1–2 days) and cost (∼$1.10 ($0.001 per bp) less expensive).

4.4. Scalability

This method is fully scalable to meet the needs of a lab focused on molecular phylogenetics in two ways. First, the method can be used to prepare a single taxon and its associated targeted loci for the 454 or it can be done on hundreds of taxa at the same time. Second, the plates that are used by 454 to generate sequence data can be split into as many as 16 sections. For example, if one wished to generate sequence data for only 10 taxa for 3 loci it would be very possible to use this method to prepare amplicon libraries and then use 1/16th of a 454 plate to generate the sequence data.

The method is not limited by the number of MIDs. By adding a different MID to both ends of a PCR product and use both to determine the taxon to which the sequence belongs, it is possible to have ∼25,000 different MID combinations (153 MIDs × 153 MIDs). Of course, this means that all PCR products must be ∼400 bp in length so that the 454 can sequence through the entire amplicon to properly sort each read to its associated MID and Taxon.

4.5. Future challenges

There are still challenges associated with using our next-gen direct sequencing protocol. First, the quality of next-gen sequence data for each platform has been well reviewed in the literature (e.g., Wicker et al., 2006; Huse et al., 2007; Kunin et al., 2010), specifically among homopolymer sites, and thus it will not be reviewed here though it remains a challenge. In most instances we observed a single insertion or deletion within these homopolymer regions (Fig. 2B). However our data show that errors among these sites were not frequent enough to influence relationships during phylogenetic reconstruction. The DNA sequence data supporting this research generated via next-gen sequencing were overall of high quality (1.09% error rate, assuming the Sanger sequencing is 100% accurate, which is a conservative assumption with respect to the 454 error rate). Further, we show that the data align appropriately with DNA sequence data generated via Sanger sequencing to provide a compelling phylogenetic result (Fig. 2A). Given that 454 data has been used extensively in “phylogenomics”, it is difficult to find any reason they would not also be appropriate for application in higher-level phylogenetics. However, an obvious challenge arises when this method is applied to population level questions, where DNA sequence fidelity is of critical concern due to the potential for a few sequencing errors to greatly influence a topology.

Major challenges also remain in our ability to remove the ∼200 bp primer dimer from the PCR product prior to next-gen sequencing. We estimate a loss of approximately 50% of the reads to primer dimers and this is likely directly related to the low molecular weight preference of 454 technology, but this may be of less concern on other platforms where smaller reads are desired. Further research is presently underway to remove the primer dimer so that we can avoid costly and time-consuming methods (such as gel cutting and purification). An improved protocol, wherein primer dimers were no longer present, would allow us to optimize scalability of the method (e.g., the potential for use with robotics). The method as described herein is fully amenable to phylogenetic research, but for other applications, where having a high number of reads is important (e.g., NUMT detection and characterization), the method still falls short due to the loss of reads from primer dimers.

Bias among MIDs may also be influencing our samples. An examination of reads generated per barcode revealed that some barcodes produced more reads than others, leading us to believe that there could be, and very likely is, bias toward certain barcodes (see Meyer et al., 2008) during next-gen sequencing. Our data were not standard across taxa, nor were the gel and ladder estimates of DNA quantity from PCR products accurate enough to statistically test for bias among MIDs. Further, it is likely that biases among MIDs are not great enough to affect directed sequencing for a phylogenetic application (i.e., because only a single string or consensus sequence is desired in phylogenetic analyses, very little coverage is needed). However, we acknowledge it is very desirable to have non-biased MIDs so that coverage between targeted gene regions and taxa that are multiplexed on the same next-gen sequencing platform can be as high as possible. We are currently planning experiments to detect and better understand any potential biases due to MIDs.

One of the major limiting factors to this method, or any method for directed sequencing, is the lack of universal primers that are effective across large groups of diverse organisms (e.g., 18S), such as those common to any deep phylogenetic project. It is very likely that next-gen sequencing will provide an exceptional amount of data in the form of genomes and EST libraries to design and select among degenerate primes that will be "universal" across groups as large and diverse as arthropods. It is also likely that genes amplified from these universal primers will be, though perhaps not in all cases (e.g., mitochondrial genes), slowly evolving. Such genes will not only be excellent for directed sequencing because they can be isolated using universal primers, but highly useful in phylogenetics because they are slowly evolving and thus somewhat optimized for high-level phylogenetics. However, the protocol for directed sequences that we propose is not entirely dependent on universal primers. Any primer set can be used to generate amplicons that can then be barcoded via MIDs and pooled for next-gen sequencing.

As a method, we see great potential for the two-step PCR protocol described above for the preparation of samples for directed sequencing. While already cost and time efficient, the method is fully scalable and promises to become ever more useful, more economical (than it already is) and rapid, especially intensive as next-gen sequencing platforms increase their ability to produce longer and better quality read lengths.

Acknowledgements

We are indebted to J. Udall and R. Byers for their generous sharing of ideas and primers and to E. Wilcox for his ideas and processing of the 454 run. We would also like to thank Fluidigm Inc. for their adapter sequence. We would like to thank collaborators who have helped with the collection and/or loan of decapod samples over the past 15 years. We thank the US National Science Foundation (MRI-0821728, EF-0531762 and EF-0531603) and Brigham Young University for financial support of this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jcz.2011.05.010.

References

Ahyong, S.T., O'Meally, D., 2004. Phylogeny of the Decapoda Reptantia: resolution using three molecular loci and morphology. Raffles B. Zool. 52, 673–693.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Binladen, J., Gilbert, M.T.P., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R., et al., 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2, e197.

Bracken, H., Toon, A., Felder, D.L., Martin, J.W., Finley, M., Rasmussen, J., et al., 2009. The Decapod Tree of Life: compiling the data and moving toward a consensus of decapod evolution. Arthropod Syst. Phylogenet. 67, 99–116.

Buhay, J.E., Moni, G., Mann, N., Crandall, K.A., 2007. Molecular taxonomy in the dark: evolutionary history, phylogeography, and diversity of cave crayfish in the subgenus *Aviticambarus*, genus *Cambarus*. Mol. Phylogenet. Evol. 42, 435–448.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.

Clement, N.L., Snell, Q., Clement, M.J., Hollenhorst, P.C., Purwar, J., Graves, B.J., et al., 2009. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next generation sequencing. Bioinformatics 26, 38–45.

Colgan, D.J., McLauchlan, A., Wilson, G.D.F., Livingston, S.P., Edgecombe, G.D., Macaranas, J., et al., 1998. Histone 3 and U2 snRNA DNA sequences and arthropod molecular evolution. Aust. J. Zool. 46, 419–437.

Crandall, K.A., Fitzpatrick, J.F., 1996. Crayfish molecular systematics: using a combination of procedures to estimate phylogeny. Syst. Biol. 45, 1–26.

Crandall, K.A., Robinson, H.W., Buhay, J.E., 2009. Avoidance of extinction through nonexistence: the use of museum specimens and molecular genetics to determine the taxonomic status of an endangered freshwater crayfish. Conserv. Genet. 10, 177–189.

Crosby, L.D., Criddle, C.S., 2007. Gene capture and random amplification for quantitative recovery of homologous genes. Mol. Cell. Probe. 21, 140–147.

De Grave, S., Pentcheff, N.D., Ahyong, S.T., Chan, T.-Y., Crandall, K.A., Dworschak, P.C., Felder, D.L., Feldmann, R.M., Fransen, C.H.J.M., Goulding, L.Y.D., Lemaitre, R., Low, M.E.Y., Martin, J.W., Ng, P.K.L., Schweitzer, C.E., Tan, S.H., Tshudy, D., Wetzer, R., 2009. A classification of living and fossil genera of decapod crustaceans. Raffles B. Zool. Suppl. 21, 1–109.

Felder, D.L., Robles, R., 2009. Molecular phylogeny of the family Callianassidae based on preliminary analyses of two mitochondrial genes. In: Martin, J.W., Crandall, K.A., Felder, D.L. (Eds.), Decapod Crustacean Phylogenetics (Crustacean Issues 18). CRC Press, Boca Raton, FL, pp. 319–334.

Felsenstein, J., 1985. Confidence-limits on phylogenies with a molecular clock. Syst. Zool. 34, 152–161.

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotech. 3, 294–299.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., Welch, D.M., 2007. Accuracy and qaulity of massively parallel DNA pyrosequencing. Genome Biol. 8, R143.

Katoh, K., Kuma, K., Toh, H., Myata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.

Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P., 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflations of diversity estimates. Environ. Microbiol. 12, 118–123.

Meyer, M., Stenzel, U., Hofreiter, M., 2008. Parallel tagged sequencing on the 454 platform. Nat. Protoc. 3, 267–278.

Meyer, M., Stenzel, U., Myles, S., Prufer, K., Hofreiter, M., 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res. 35, e97.

Pertoldi, C., Tokarska, M., Wojcik, J.M., Demontis, D., Loeschcke, V., Gregersen, V.R., et al., 2009. Depauperate genetic variability detected in the American and European bison using genomic techniques. Biol. Direct. 4, 48.

Porter, M.L., Perez-Losada, M., Crandall, K.A., 2005. Model-based multi-locus estimation of decapod phylogeny and divergence times. Mol. Phylogenet. Evol. 37, 355–369.

Robles, R., Tudge, C.C., Dworschak, P.C., Poore, G.C.B., Felder, D.L., 2009. Molecular phylogeny of the Thalassinidea based on nuclear and mitochondrial genes. In: Martin, J.W., Crandall, K.A., Felder, D.L. (Eds.), Decapod Crustacean Phylogenetics (Crustacean Issues 18). CRC Press, Boca Raton, FL, pp. 301–318.

Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463.

Song, H., Buhay, J.E., Whiting, M.F., Crandall, K.A., 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc. Natl. Acad. Sci. U. S. A. 105, 13486–13491.

Scholtz, G., Richter, S., 1995. Phylogenetic Systematics of the Reptantian Decapoda (Crustacea Malacostraca). Zool. J. Linn. Soc. 113, 289–328.

Schram, F.R., Dixon, C.J., 2004. Decapod phylogeny: addition of fossil evidence to a robust morphological cladistic data set. Bull. Mitzunami Fossil Mus. 31, 1–19.

Thoma, B., Schubart, C., Felder, D.L., 2009. Molecular phylogeny of western Atlantic representatives of the genus Hexapanopeus (Decapoda: Brachyura: Panopeidae). In: Martin, J.W., Crandall, K.A., Felder, D.L. (Eds.), Decapod Crustacean Phylogenetics (Crustacean Issues 18). CRC Press, Boca Raton, pp. 274–300.

Toon, A., Finley, M., Staples, J., Crandall, K.A., 2009. Decapod phylogenetics and molecular evolution. In: Martin, J.W., Crandall, K.A., Felder, D.L. (Eds.), Decapod Crustacean Phylogenetics (Crustacean Issues 18). CRC Press, Boca Raton, FL, pp. 14–28.

Tsang, L.M., Ma, K.Y., Ahyong, S.T., Chan, T.Y., Chu, K.H., 2008. Phylogeny of Decapoda using two nuclear protein-coding genes: origin and evolution of the Reptantia. Mol. Phylogenet. Evol. 48, 359–368.

Whiting, M.F., 2002. Mecoptera is paraphyletic: multiple genes and phylogeny of Mecoptera and Siphonaptera. Zool. Scripta 31, 93–104.

Wicker, T., Schlagenhauf, E., Graner, A., Close, T.J., Keller, Beat., Nils, S., 2006. 454 sequencing put to the test using the complex genome barley. BMC Genomics 7, 275.

Whiting, M.F., Carpenter, J.C., Wheeler, Q.D., Wheeler, W.C., 1997. The strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Syst. Biol. 46, 1–68.