Chapter 4

Phylogenomics Using Transcriptome Data

Johanna Taylor Cannon and Kevin Michael Kocot

Abstract

This chapter presents a generalized protocol for conducting phylogenetic analyses using large-scale molecular datasets, specifically using transcriptome data from the Illumina sequencing platform. The general molecular lab bench protocol consists of RNA extraction, cDNA synthesis, and sequencing, in this case via Illumina. After sequences have been obtained, bioinformatics methods are used to assemble raw reads, identify coding regions, and categorize sequences from different species into groups of orthologous genes (OGs). The specific OGs to be used for phylogenetic inference are selected using a custom shell script. Finally, the selected orthologous groups are concatenated into a supermatrix. Generalized methods for phylogenomic inference using maximum likelihood and Bayesian inference software are presented.

Key words Phylogenomics, Transcriptomes, RNAseq, cDNA, Illumina, Phylogeny

1 Introduction

Over the last 10 years, phylogenomics has dramatically revised our understanding of metazoan relationships [1, 2]. In the broadest sense, phylogenomics refers to the inference of phylogenetic relationships based on large-scale molecular datasets. Although the original meaning of the term phylogenomics referred to the study of gene family evolution [3], popular usage now generally indicates the use of high-throughput sequencing of transcriptome or genome data for phylogenetic reconstruction. Most phylogenomic studies have used a shotgun sequencing approach, although a few have targeted specific genes [4], and there is a growing interest in 'anchored phylogenomics' that uses probes designed from diverse lineages within the target clade for targeted enrichment of loci [5–7]. Shotgun sequencing approaches tend to recover constitutively expressed 'housekeeping' genes no matter the source tissue that is sequenced, because these genes are vital to the function of the cell and are found across tissue types. Furthermore, these functionally important genes tend to be evolutionarily conserved, making them useful for inference of deep relationships.

Sarah J. Bourlat (ed.), *Marine Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1452,
DOI 10.1007/978-1-4939-3774-5_4,

Early phylogenomic studies of animal relationships made use of expressed sequence tag (EST) data collected via Sanger-based methods [8–11]. Sanger-based EST methods required cloning randomly sheared cDNA fragments into bacterial vectors. The advent of massively parallel pyrosequencing methods such as 454 facilitated the collection of data from a broader subset of non-model organisms [12–14]. At present, Illumina sequencing offers lowest cost per base pair and has become the sequencing platform of choice for most phylogenomic studies [15–23]. The method presented later uses Illumina technology, although it can be modified to accommodate sequences generated using other methods.

After obtaining sequence data, phylogenomic dataset assembly consists of a series of bioinformatics steps. The essential steps are (1) de novo assembly of raw sequencing reads, (2) determination of orthologous groups of sequences, (3) selection of orthologous groups to be used in downstream analyses, (4) multiple sequence alignment, (5) concatenation, and (6) phylogenetic inference. To accomplish these steps, there are a myriad of phylogenomics programs available, many of which have similar functionalities, making choosing the most appropriate program for a given project a challenge. Several consistently updated pipelines such as Agalma [24], Osiris [25], and the unnamed pipeline of Yang and Smith [26] provide wrapper scripts for other existing software, offering a more seamless means to take raw reads through the stages of phylogenomic dataset assembly. These software options may be preferable for users with less bioinformatics experience, although these pipelines are open source and encourage user development and modification. Here we provide modified versions of scripts used in our previous publications (e.g., Kocot et al. [12] and Cannon et al. [15]), which can easily be adapted for other systems. The following steps represent a standard workflow that can be conducted on a local computer or remote cluster using the Linux operating system. Assembly, orthology determination, and phylogenetic inference will likely need to be performed on a high-performance computing cluster. This is one approach out of many possibilities, and we have pointed out alternatives where appropriate in the Subheading 4. New programs are released all the time, so it is important to check for updates and to read program manuals to make informed choices about the best approach for your particular system.

2 Materials

2.1 RNA Extraction

1. Solution for RNA stabilization and storage, or liquid nitrogen for tissue preservation.
2. Nuclease-free 1.5 ml Eppendorf tubes.

3. Trizol.
4. Homogenizer or liquid nitrogen and mortar and pestle.
5. Chloroform.
6. 100 % isopropyl alcohol.
7. 75 % ethanol.
8. RNase-free H_2O.
9. 4 °C centrifuge.
10. Quantification equipment—e.g., Nanodrop, Qubit, Agilent Bioanalyzer.
11. Gel electrophoresis apparatus.
12. RNA cleanup kit with DNase I.
13. Optional: commercial RNA extraction kit for small tissue samples.

2.2 cDNA Library Preparation

1. Clontech SMART cDNA Library Construction Kit.
2. Clontech Advantage2 PCR Kit.
3. 5′Primer,12μM(5′-AAGCAGTGGTATCAACGCAGAGT-3′) (*see* **Note 1**).
4. RNase-free tubes and tips.
5. RNase inhibitor.
6. Thermal cycler.
7. PCR Purification kit.
8. 3M sodium acetate.
9. Quantification equipment—e.g., Nanodrop, Qubit, Agilent Bioanalyzer.
10. Gel electrophoresis apparatus.
11. Optional: vacuum centrifuge.

2.3 Sequencing

1. A sequencing facility with access to Illumina sequencing machines.

2.4 Bioinformatics: Dataset Assembly and Phylogenetic Inference

1. Linux computer or access to a remote server with the following software installed: Trinity [27], TransDecoder (http://transdecoder.sf.net), HaMStR [28], Mafft [29], Aliscore [30], and Alicut (https://www.zfmk.de/en/research/research-centres-and-groups/utilities), FastTreeMP [31], PhyloTreePruner [32], FASconCAT [33], RAxML [34], PhyloBayes [35].
2. Custom bash scripts available on GitHub at: https://github.com/kmkocot/springer_methods_chapter.

3 Methods

3.1 Extraction of Total RNA

Tissue to be used for RNA extraction should be fresh, preserved in RNA stabilization solution and stored in the freezer, or frozen at −80 °C. Numerous alternative protocols and kits exist for extraction of total RNA, and the best method for a given sample will depend on the size and composition of the tissue to be extracted. Useful discussion of RNA preparation methods can be found at the RNA-seqlopedia (rnaseq.uoregon.edu). In general, standard TRIzol-based methods work well for most macroinvertebrates, while for meiofauna or larvae, it may be necessary to use a commercial kit specifically designed for extracting RNA from cells or very small tissue samples. Cleanup of RNA extracted using TRIzol using a silica spin column-based kit that integrates removal of genomic DNA carryover using DNase I is recommended to reduce the carryover of phenol and genomic DNA that can negatively affect assembly. Final RNA should be resuspended in nuclease-free water and evaluated with available equipment, e.g., NanoDrop, Qubit, Bioanalyzer, and gel electrophoresis (*see* **Note 2**). RNA should be kept on ice while the quantity and quality are being checked, followed immediately by first-strand cDNA synthesis.

3.2 cDNA Synthesis

Again, several options are available for synthesis of complementary DNA from RNA. Illumina TruSeq library preparation kits incorporate Illumina library preparation steps including adding adaptors and indexing, eliminating the need for additional library preparation steps at the sequencing center. TruSeq kits currently require 0.1–4 μg input RNA; these kits may be preferred for large tissue samples. For microorganisms that yield smaller quantities of RNA, the SMART cDNA Library Construction Kit from Clontech can start with as little as 50 ng total RNA (*see* **Note 3**). Following is a suggested protocol using the SMART cDNA synthesis kit with slight modifications.

1. For very low amounts of starting RNA, samples may need to be concentrated in a vacuum centrifuge. Thoroughly clean the vacuum centrifuge before beginning, and as an added precaution, RNase inhibitor may be added to the sample before concentrating. Do not heat the sample during vacuum centrifugation.
2. Follow the manual of the SMART cDNA synthesis kit through first-strand synthesis.
3. For each first-strand cDNA product, perform an amplification test to determine the optimal number of PCR cycles. Volumes listed as follows are sufficient for the amplification test only; final amplification of cDNA will be completed in a subsequent amplification reaction. For each library combine the following in a 0.2 ml PCR tube on ice:

3.0 μl Diluted first-strand cDNA (from **step 2**)

21.0 μl PCR-grade H2O.

3.0 μl 10× Advantage 2 PCR Buffer.

0.75 μl dNTP mix.

1.4 μl 5′ PCR primer (12 μM).

0.6 μl 50× Advantage 2 Polymerase Mix.

Mix gently and then briefly spin down using a microcentrifuge. Place tube(s) in a thermal cycler that has been preheated to 95 °C and run the following program:

94 °C for 5 min (1 cycle).

94 for 40 s, 65 °C for 1 min, 72 °C for 5 min (15 cycles).

Hold at 6 °C.

4. After 15 cycles, remove and save 3 μl of the reaction mix, and subject the remaining mix to two additional cycles. Repeat this process until the reaction mix has been subjected to 25 cycles (*see* **Note 4**).
5. After cycling, analyze the reserved aliquots on an agarose gel. Estimate product concentration and size distribution in order to determine the optimal number of cycles. The cDNA should appear as a smear mostly between 500 bp and 3 kb, often with strong distinct bands representing abundant transcripts.
6. To ensure yield of >1 μg cDNA (required by most Illumina sequencing centers as of late 2015), run a final cDNA amplification as a series of multiple smaller reactions. Volumes are given as follows for 12 reactions per sample, although fewer reactions may be needed to reach 1 μg.
   For each library combine the following in a 1.5 ml tube on ice:

   36 μl Diluted first-strand cDNA (from step 2).

   252 μl PCR-grade H2O.

   36 μl 10× Advantage 2 PCR Buffer.

   12 μl dNTP mix.

   16.8 μl 5′ PCR primer (12 μM).

   7.2 μl 50× Advantage 2 Polymerase Mix.
   Aliquot 29.75 μl of this master mix into each of twelve 0.2 ml PCR tubes on an ice block.
   Mix gently and then briefly spin down using a microcentrifuge. Place tubes in a thermal cycler that has been preheated to 95 °C and run the following program:

   94 °C for 5 min (1 cycle).

   94 for 40 s, 65 °C for 1 min, 72 °C for 5 min (*n* cycles).

Hold at 6 °C.

n = the optimal number of cycles for each library determined in step 3

7. Pool the 12 reaction products generated in **step 6**, and purify the amplified cDNA using a PCR purification kit following manufacturers protocols (*see* **Note 5**).
8. Analyze 3 μl purified cDNA on an agarose gel. Quantify cDNA concentration and purity using available equipment.

3.3 Sequencing

Prepared cDNA can be submitted as is to an Illumina sequencing facility. Depending on the sequencing depth required, multiple samples may be sequenced on a single lane of Illumina HiSeq (*see* **Note 6**). When following the protocol outlined earlier, it will be necessary for the sequencing center to perform Illumina library preparation steps, including adding adaptors and indexing samples.

3.4 Dataset Assembly

1. Download and make a secure backup of the raw sequencing data, which is typically provided in FASTQ format. For the commands listed as follows, we assume you are working in your home folder, have your data in a subfolder called "data," and the scripts in a separate subfolder called "scripts." Software listed in the materials section should be installed in your path. Please note that changes to this structure will require modifications of the commands.
2. Run Trinity to assemble reads into contigs, selecting appropriate memory and CPU options for your system (*see* **Note 7**).

```
Trinity.pl --seqType fq --max_memory 50G
--CPU 8 --left
file_name_for_forward_reads_1.fastq.gz
--right
file_name_for_reverse_reads_2.fastq.gz
```

Trinity will produce a subdirectory with output files for each library, containing the completed assembly in fasta format. This file will be used for subsequent steps and should be moved to the home directory.

3. Translate assembled contigs using Transdecoder. The location of the Pfam-AB.hmm.bin file may vary depending on your system and installation of Transdecoder. Transdecoder will produce several output files, the translations with .pep file extensions (containing peptide sequences of predicted open reading frame regions in fasta format) should be carried forward to orthology determination.

```
TransDecoder -t Trinity_Output.fasta
--search_pfam ~/bin/TransDecoder/pfam/Pfam-
AB.hmm.bin
```

4. Perform steps 2 and 3 on all raw RNAseq libraries to be included in your phylogenomic analysis, using unique file names.
5. Clean up intermediate files either by deleting or compressing and archiving, such as only the final translated Transdecoder .pep files are in the home directory.
6. Collect any additional translated amino acid sequences from sources other than raw Illumina data (e.g., predicted proteins from genome projects, publically available assemblies) that are to be used in the phylogenomic dataset, and place them in the home directory, using .pep file extensions. Nucleotide data from other sources must be translated as in step 6 prior to orthology determination.
7. Prepare translated sequences for orthology assignment. The script batch_prep_sequences.sh will remove line breaks from all translated fasta files using a script called nentferner.pl that is packaged with the HaMStR orthology determination software (*see* **Note 8**). This script will also remove special characters from fasta sequence headers that will cause errors in future steps, and move the unedited .pep files to a new directory titled "original_pep_ files" that can be archived for future reference or discarded.

   ```
   ./scripts/batch_prep_sequences.sh
   ```

8. Categorize sequences into putatively orthologous groups (OG). Many software options are available for orthology determination (*see* **Note 9**). The following steps use HaMStR (Hidden Markov Model based Search for Orthologs using Reciprocity) version 13.2.3, with the "modelorganisms" core ortholog set and *Drosophila melanogaster* as the selected 'reference taxon' (*see* **Note 10**). It may be necessary to include the full path to the hamstr program and/or the hmmset, depending on your installation.

   ```
   hamstr -protein -strict -hmmset=modelorganisms_
   hmmer3 -refspec=DROME -sequence_file=Sequence_
   name.fa.nent -taxon=NAME
   ```

   Run HaMStR for each operational taxonomic unit (OTU) to be included in your dataset. Read the HaMStR manual for discussion of all flags. The -taxon flag gives each OTU a unique identifier to be supplied by the user for each OTU (here we have used the generic NAME, but you should select a unique four or five letter identifier for each species in your dataset). We advocate against the use of the -representative flag as it picks one sequence per taxon when two or more are present and can result in a final dataset including paralogs. We use a phylogenetic tree-based approach to select the best sequence from each taxon in these cases (see later).

9. Execute the bash script HaMStR_v13_concatenate.sh. HaMSrR_v13_concatenate.sh renames the files output by HaMStR into a format appropriate for orthology determination. Organisms included in the core ortholog set can be added or removed from each OG (see end of script). This script relies heavily on the Linux program rename, which works differently on different versions of Linux and may need to be modified (*see* **step 12**).

```
./scripts/HaMStR_v13_concatenate.sh
```

10. Execute the bash script phylogenomics_dataset_assembly.sh while in the folder containing the output of HaMStR_v13_concatenate.sh. The dataset assembly script takes the output of HaMStR and performs several steps to remove groups and sequences that are not suitable for phylogenomic analysis (*see* **Note 11**). The final product of this script is a set of trimmed amino acid alignments representing putatively orthologous groups suitable for phylogenomic analysis. The script requires GNU parallel be installed on your machine. There are several variables that must be modified for your purposes within the bash script. We suggest you examine the entire script carefully and modify it as needed. Input fasta file headers must be in the following format: >orthology_group_ID|OTU_abbreviation|annotation_or_sequence_ID_information (Example: >0001|LGIG|Contig1234). Fasta headers may not include spaces or nonalphanumeric characters except for underscores (pipes are OK as field delimiters only). If you have followed the earlier steps, your fasta headers should already be in this format.

```
./scripts/phylogenomics_dataset_assembly.sh
```

11. The output of the earlier script can be concatenated using FASconCAT. Before FASconCAT can be used, the fasta headers for each OTU in each OG alignment file must be made to match exactly. The simplest way to do this is to use the unique OTU identifier that was used in HaMStR. After executing the phylogenomics_dataset_assembly.sh script, the first field delimiter in your fasta headers should now be an @ symbol. If this is the case, type the following in the folder containing the individual orthogroup alignments, which will remove all characters following the first @ found on each line (*see* **Note 12**):

```
sed -i 's/\@.*//' *.fas
```

12. FASconCAT.pl will only work on files with the extension .fas, not .fa. You may need to rename .fa files to .fas. On Ubuntu Linux the command for this would be:

```
rename 's/.fa/.fas/g' *.fa
```

On Scientific Linux and some other distributions, the command would be:

```
rename .fa .fas *.fa
```

13. Create a concatenated total alignment matrix (*see* **Note 13**) using the program FASconCAT, which is an interactive program that offers many options for input and output files. To start FASconCAT, type the following in the folder containing the output sequences of the earlier script.

    ```
    perl FASconCAT.pl
    ```

    Select relaxed phylip output by typing "p" twice in the program menu. Once you have selected all the options that suit your downstream analysis, enter "s" in the program menu to start the concatenation (*see* **Note 14**).

14. Perform maximum likelihood phylogenetic inference with RAxML version 8. The following command executes a partitioned data analysis using the best-fitting model for each partition and the appropriate number of rapid bootstrap replicates. The partition data file should list "AUTO" as the model to use for each partition (*see* **Note 15**).

    ```
    raxmlHPC-THREADS-AVX -T 16 -s Total_Alignment.phylip -n RaxML.out -f a -N autoMRE -x 12345 -p 12345 -m PROTGAMMAAUTO -q partition_data.txt
    ```

15. Perform Bayesian inference phylogenetic analysis using PhyloBayesMPI. The following commands execute four independent chains of 15,000 generations sampling one tree per generation under the site heterogeneous CAT + GTR model (*see* **Note 16**). More than 15,000 generations may be necessary for some datasets.

    ```
    pb -x 1 15000 -cat -gtr -d Total_Alignment.phy Chain1
    pb -x 1 15000 -cat -gtr -d Total_Alignment.phy Chain2
    pb -x 1 15000 -cat -gtr -d Total_Alignment.phy Chain3
    pb -x 1 15000 -cat -gtr -d Total_Alignment.phy Chain4
    ```

16. Assess convergence of the four chains using the bpcomp program packaged with PhyloBayes.

    ```
    bpcomp -x 5000 Chain1 Chain2 Chain3 Chain4
    ```

    This command discards one-third of all trees produced by the chains as burn-in, and compares the remaining lists of trees and outputs "maxdiff," a discrepancy index measuring how different the consensus trees produced by the four chains are.

The PhyloBayes manual recommends that the maxdiff value should be 0.1 or less, but 0.3 or less may be acceptable. bpcomp may be executed on currently running chains, so it is possible to check on progress of a run without stopping the analysis. bpcomp also produces a majority rule consensus tree.

4 Notes

1. The 5′ PCR primer is packaged with the Clontech SMART cDNA Library Construction kit at a concentration of 12 μM. We have found that the supply provided in the kit is often not sufficient to carry out the multiple amplification reactions recommended in our modified protocol, thus we recommend purchasing an additional supply and reconstituting it to 12 μM. Reconstituting to a more standard 10 μM will require modification to reaction volumes.
2. When working with very small samples (e.g., meiofaunal animals or marine invertebrate larval samples), visualization of RNA by gel electrophoresis will not be feasible. Synthesis of cDNA is usually successful even when measured quantities of RNA are extremely low or below the recommended starting amounts for the cDNA synthesis kit. If your samples are precious, proceed with cDNA synthesis and you will likely be rewarded. We have generated a successful cDNA library from an RNA sample that gave a negative reading on a Nanodrop.
3. We have had much success with the Clontech SMART cDNA Library Construction Kit with a variety of marine invertebrate samples. This kit can be used for as little as 50 ng total RNA up to 1 μg total RNA, so a single kit can be used if specimens in a range of sizes are to be processed. Keep in mind that indexing and Illumina library preparation steps will still need to be done at the sequencing center if submitting cDNA generated via the SMART kit. Clontech also manufactures kits that can start with as little as 100 pg RNA called "SMARTer Stranded RNA-Seq Kits—Strand-Specific Library Construction for Transcriptome Analysis on Illumina Platforms" that incorporate library preparation steps including indexing and adding adaptors, eliminating the need for downstream library preparation kits. We have no direct experience using this kit, but it may be a good option.
4. For most samples, 25 cycles will be sufficient. However, for some very tiny organisms, more cycles may be required. Fewer cycles will generally result in fewer nonspecific PCR products.
5. Most common PCR purification kits have a maximum yield of 10 μg per spin column, making it efficient to purify the replicate PCR products by pooling them and running the pooled products over a single silica spin column, loading multiple times. Double-check the maximum yield for your PCR

purification kit of choice before using this approach. The larger volumes of PCR master mix added to the purification kit buffers can affect pH, so we recommend that you use pH indicators included with your kit for all buffers to ensure efficient binding. If pH is too high, 3 M sodium acetate can be added to the buffer in order to lower pH to the optimal range.

6. We typically pool 6–8 transcriptomes in one lane of an Illumina HiSeq 2000 for phylogenomics. Use caution when combining samples across a single lane of Illumina HiSeq, as bleed-through has been demonstrated to occur. When multiple samples are sequenced in the same lane of an Illumina instrument, the data are sorted after sequencing by sequence 'barcodes' or 'indices' with a different code for each sample. Sometimes the barcode is misread. Usually, the misread barcode doesn't correspond to any of the samples being sequenced and that read is discarded. However, sometimes by random chance the barcode is misread as having the sequence of one of the other samples being processed so it gets put in the wrong 'bin.' If one of the samples has one or more really highly expressed genes (mitochondrial genes, nuclear ribosomal RNA, or other tissue-specific highly expressed genes), there might be so many reads from that transcript that end up incorrectly 'binned' that this gene ends up showing up in the assemblies of the other samples that were sequenced in parallel.
7. This command takes raw reads and assembles them directly. In many cases, it may be advisable to trim low quality reads and adaptor sequences prior to assembly. This can be accomplished using the Trimmomatic [36] program packaged with Trinity. Trinity can now also conduct digital normalization, which can significantly speed up assembly times. Normalization is not recommended if you have not used DNase treatment on your RNA prior to cDNA synthesis. Check the Trinity manual for details.
8. Many bioinformatics programs that manipulate sequence data in fasta format require that each sequence be listed on a single line (in other words, there are no line breaks within the sequence string). The perl script nentferner.pl is an extremely useful tool for removing line breaks in fasta files that is packaged with the HaMStR orthology determination program. This can also be accomplished with the fasta_formatter tool bundled with the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), and we highly recommend that you install one of these in your path.
9. There are several commonly used programs for orthology assignment. The program used here, HaMStR (Hidden Markov Model based Search for Orthologs using Reciprocity), generates profile hidden Markov models (pHMMs), each representing a set of orthologous genes for selected reference taxa from the InParanoid database [37] for which whole genomes are available. Sequences are searched against a reference taxon set, the

"model organisms" set in this example, which includes 1032 orthologous groups (OG) with sequences from *Homo sapiens, Ciona intestinalis, Drosophila melanogaster, Caenorhabditis elegans*, and *Saccharomyces cerevisiae*. Translated contigs are scanned for significant hits to each OG's pHMM. Matching sequences are then compared to the proteome of a selected primer taxon (*Drosophila melanogaster* in this example) using BLASTP (-strict option). If the *Drosophila melanogaster* amino acid sequence that contributed to the pHMM was the best BLASTP hit, then the sequence was assigned to that OG. If this reciprocity criterion is not met, the sequence is discarded. Other popular programs include OMA [38], FastOrtho [39] (a reimplementation of OrthoMCL [40]), and ab initio methods starting with all-by-all BLASTP searches followed by phylogenetic identification of orthologous sequences [13], implemented in programs such as ProteinOrtho [41] and Agalma [24].

10. HaMStR currently offers several precompiled core ortholog sets. The model organism set used here includes *Homo sapiens*, *Ciona intestinalis*, *Drosophila melanogaster*, *Caenorhabditis elegans,* and *Saccharomyces cerevisiae,* which works well in studies with broad taxon sampling across Metazoa. Also available are ortholog sets for Amniota, Arthropoda, basal metazoans, Chordata, Fungi, Insecta, Lophotrochozoa, and plants. It is also possible to use available genomic and transcriptomic data to build core ortholog sets from scratch for your taxonomic group of interest, although this process is arduous, and if none of the available core ortholog sets are appropriate for your study, it may be preferable to use an alternative orthology determination software program.
11. The input of this script is the putative orthologous groups generated by HaMStR. The script uses several other programs to produce individual trimmed alignments for each OG and to remove groups and sequences that are less suitable for phylogenomic analysis. The script is made up of a series of intermediate steps that are commented inside the script. A backup of all starting fasta files is created and placed into a new directory called "unedited_sequences." Next, newlines are removed from all files as described in **Note** 7. This process is repeated several times throughout the script. Sequences shorter than a set threshold are removed. This cutoff value is set in the program header using the variable MIN_SEQUENCE_LENGTH. OGs containing fewer taxa than a set threshold are removed and placed in a new directory called "rejected_few_taxa_1." This cutoff value is also set in the program header using the variable MIN_TAXA. Next, OGs are aligned using the program MAFFT [29]. Each OG is trimmed with the perl scripts Aliscore and Alicut [30] to remove columns with ambiguous alignment or little phylogenetic signal. Note that one recent study advocated against aggressive use of such alignment trimming software, particularly if it is trimming >20%

of aligned regions [42]. After trimming, spaces and gap only columns are removed, short alignments are discarded, and OGs containing too few taxa are removed and placed in a new directory called "rejected_few_taxa_2." Individual OG trees are generated using FastTreeMP [31] and the utility PhyloTreePruner [32] is used to screen for potential paralogs. PhyloTreePruner screens trees for instances where multiple sequences from the same OTU do not form monophyletic clades. Suspected paralogs are trimmed from the data matrix, leaving the maximally inclusive subtree in which sequences from each OTU form monophyletic clades or are part of the same polytomy. If an OG still possesses more than one sequence for an OTU (inparalogs), PhyloTreePruner is set to select the longest sequence for inclusion in the final concatenated alignment (-u option).

12. The sed -i flag will modify the file itself. To test the command prior to executing it, simply remove the -i option from the command and the output will appear in the terminal only.
13. The approach outlined here will generate a "total alignment" of all the OGs that pass through paralogy screening in PhyloTreePruner. In addition to conducting analyses of this total alignment, a number of approaches may be worth considering in an attempt to remove various sources of systematic error or "noise" from the data. Among others, MARE (matrix reduction) [43] maximizes information content of genes, taxa, and the overall alignment. BMGE (Block Mapping and Gathering with Entropy) [44] conducts trimming and recoding of alignments aimed at reducing artifacts due to compositional heterogeneity. TreSpEx [45] and BaCoCa [46] perform a variety of statistical calculations on individual taxa, OGs, or the total alignment to identify possible biases in phylogenomic datasets from sources such as long branch attraction, saturation, missing data, and rate heterogeneity. Combining these tools to generate multiple alignments can provide valuable insights into potential sources of bias in your data and strengthen your overall analysis.
14. By default, FASconCAT generates an .xls file containing single range information of each sequence fragment and a checklist of all concatenated sequences. The information in this file may easily be adapted to use in phylogenetic analyses to partition the concatenated matrix into gene regions for model specification, etc.
15. Model choice in phylogenomic analysis has been the subject of debate [47]. RAxML implements traditional site-homogenous models, or more recently developed LG4X and LG4M models [48] that integrate four substitution matrixes to improve modeling of site heterogeneity. The newest version of RAxML allows the user to choose to have the program select the best-fitting model for each partition in the concatenated matrix. We recommend either partitioning data by OG and selecting the best model of evolution for each group using RAxML or other

model selection software such as ProtTest [49], or partitioning sites using software such as PartitionFinder [50] over selecting a single substitution model across the concatenated matrix.

16. PhyloBayes implements the site-heterogeneous CAT model [51], which does not assume homogenous substitution patterns across an alignment. This assumption is likely to be violated in large, concatenated data matrices, so these models have been preferred over site homogenous models for phylogenomic datasets [47]. Bayesian inference under such complex models is extremely computationally expensive and will need to be carried out on a remote high performance computing cluster.

References

1. Giribet G (2015) New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. Org Divers Evol 2015:1–8. doi:10.1007/s13127-015-0236-4
2. Telford MJ, Budd GE, Philippe H (2015) Phylogenomic insights into animal evolution. Curr Biol 25:R876–R887. doi:10.1016/j.cub.2015.07.060
3. Eisen JA, Fraser CM (2003) Phylogenomics: intersection of evolution and genomics. Science 300:1706–1707. doi:10.1126/science.1086292
4. Regier JC, Shultz JW, Zwick A et al (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463:1079–1083. doi:10.1038/nature08742
5. Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst Biol 61:727–744. doi:10.1093/sysbio/sys049
6. Peloso PLV, Frost DR, Richards SJ et al (2015) The impact of anchored phylogenomics and taxon sampling on phylogenetic inference in narrow-mouthed frogs (Anura, Microhylidae). Cladistics 32:113–140. doi:10.1111/cla.12118
7. Prum RO, Berv JS, Dornburg A et al (2015) A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573. doi:10.1038/nature15697
8. Philippe H, Lartillot N, Brinkmann H (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22:1246–1253. doi:10.1093/molbev/msi111
9. Dunn CW, Hejnol A, Matus DQ et al (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749. doi:10.1038/nature06614
10. Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965–968
11. Bourlat SJ, Juliusdottir T, Lowe CJ et al (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444:85–88. doi:10.1038/nature05241
12. Kocot KM, Cannon JT, Todt C et al (2011) Phylogenomics reveals deep molluscan relationships. Nature 477:452–456
13. Smith SA, Wilson NG, Goetz FE et al (2011) Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480:364–367. doi:10.1038/nature10526
14. Telford MJ, Lowe CJ, Cameron CB et al (2014) Phylogenomic analysis of echinoderm class relationships supports Asterozoa. Proc Biol Sci 281(1786):pii: 20140479. doi:10.1098/rspb.2014.0479
15. Cannon JT, Kocot KM, Waits DS et al (2014) Phylogenomic resolution of the hemichordate and echinoderm clade. Curr Biol 24:2827–2832. doi:10.1016/j.cub.2014.10.016
16. Whelan NV, Kocot KM, Moroz LL, Halanych KM (2015) Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci 112:5773–5778. doi:10.1073/pnas.1503453112
17. Struck TH, Golombek A, Weigert A et al (2015) The evolution of annelids reveals two adaptive routes to the interstitial realm. Curr Biol 25:1993–1999. doi:10.1016/j.cub.2015.06.007

18. Weigert A, Helm C, Meyer M et al (2014) Illuminating the base of the annelid tree using transcriptomics. Mol Biol Evol 31:1391–1401. doi:10.1093/molbev/msu080
19. Laumer CE, Bekkouche N, Kerbl A et al (2015) Spiralian phylogeny informs the evolution of microscopic lineages. Curr Biol 25:2000–2006. doi:10.1016/j.cub.2015.06.068
20. Andrade SCS, Novo M, Kawauchi GY et al (2015) Articulating "Archiannelids": phylogenomics and annelid relationships, with emphasis on Meiofaunal taxa. Mol Biol Evol 32:2860–2875. doi:10.1093/molbev/msv157
21. Andrade SCS, Montenegro H, Strand M et al (2014) A transcriptomic approach to ribbon worm systematics (Nemertea): resolving the Pilidiophora problem. Mol Biol Evol 31:3206–3215. doi:10.1093/molbev/msu253
22. Laumer CE, Hejnol A, Giribet G (2015) Nuclear genomic signals of the "microturbellarian" roots of platyhelminth evolutionary innovation. eLife e05503. doi:10.7554/eLife.05503
23. Egger B, Lapraz F, Tomiczek B et al (2015) A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. Curr Biol 25:1347–1353. doi:10.1016/j.cub.2015.03.034
24. Dunn CW, Howison M, Zapata F (2013) Agalma: an automated phylogenomics workflow. BMC Bioinformatics 14:330. doi:10.1186/1471-2105-14-330
25. Oakley TH, Alexandrou MA, Ngo R et al (2014) Osiris: accessible and reproducible phylogenetic and phylogenomic analyses within the Galaxy workflow management system. BMC Bioinformatics 15:230. doi:10.1186/1471-2105-15-230
26. Yang Y, Smith SA (2014) Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol Biol Evol 31:3081–3092. doi:10.1093/molbev/msu245
27. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652
28. Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol 9:157
29. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518
30. Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst Biol 58:21–34
31. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490
32. Kocot KM, Citarella MR, Moroz LL, Halanych KM (2013) PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol Bioinf Online 9:429–435. doi:10.4137/EBO.S12813
33. Kück P, Meusemann K (2010) FASconCAT: convenient handling of data matrices. Mol Phylogenet Evol 56:1115–1118. doi:10.1016/j.ympev.2010.04.024
34. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. doi:10.1093/bioinformatics/btu033
35. Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI. Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol 62:611–615. doi:10.1093/sysbio/syt022
36. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi:10.1093/bioinformatics/btu170
37. Östlund G, Schmitt T, Forslund K et al (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38:D196–D203
38. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2011) OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 39:D289–D294. doi:10.1093/nar/gkq1238
39. Wattam AR, Abraham D, Dalay O et al (2014) PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42:D581–D591. doi:10.1093/nar/gkt1099
40. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. doi:10.1101/gr.1224503
41. Lechner M, Findeiß S, Steiner L et al (2011) Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124. doi:10.1186/1471-2105-12-124
42. Tan G, Muffato M, Ledergerber C et al (2015) Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst Biol 64:778–791. doi:10.1093/sysbio/syv033

43. Meyer B, Meusemann K, Misof B (2010) MARE v0.1.2-rc
44. Criscuolo A, Gribaldo S (2010) BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210. doi:10.1186/1471-2148-10-210
45. Struck TH (2014) TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. Evol Bioinf Online 10:51–67. doi:10.4137/EBO.S14239
46. Kück P, Struck TH (2014) BaCoCa—a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Mol Phylogenet Evol 70:94–98. doi:10.1016/j.ympev.2013.09.011
47. Philippe H, Brinkmann H, Lavrov DV et al (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9, e1000602. doi:10.1371/journal.pbio.1000602
48. Le SQ, Dang CC, Gascuel O (2012) Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol 29:2921–2936. doi:10.1093/molbev/mss112
49. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165. doi:10.1093/bioinformatics/btr088
50. Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695–1701. doi:10.1093/molbev/mss020
51. Lartillot N, Philippe H (2004) A Bayesian Mixture Model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109. doi:10.1093/molbev/msh112