Hemichordata and Echinodermata GenBank accessions, Table S2, Cannon et al (2014) Curr. Bio. (Antarctic Inverts project)

Website: https://www.bco-dmo.org/dataset/671521
Data Type: Cruise Results
Version: 1
Version Date: 2016-12-22

Project
» Genetic connectivity and biogeographic patterns of Antarctic benthic invertebrates (Antarctic Inverts)

Contributors	Affiliation	Role
Halanych, Kenneth M.	Auburn University	Principal Investigator
Mahon, Andrew	Central Michigan University	Co-Principal Investigator
Copley, Nancy	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
This dataset was published as Table S2 from Cannon et al (2014). It contains transcriptome Sequence Read Archive GenBank accession numbers and associated matrices of hemichordate and echinoderm specimens collected globally, with emphasis on the Southern Ocean, from 2001 to 2013.

Table of Contents

- Coverage
- Dataset Description
  - Methods & Sampling
  - Data Processing Description
- Data Files
- Related Publications
- Related Datasets
- Parameters
- Instruments
- Deployments
- Project Information
- Funding

Coverage

Temporal Extent: 2001 - 2013

Methods & Sampling

From Cannon et al 2014):
Taxon Sampling
We sampled transcriptomic data from representatives of all recognized hemichordate families and all echinoderm classes except Echinoidea, for which data were already available (Tables S1 and S2). At least two species of each taxonomic group were sampled, except for Spengelidae. Samples from which novel transcriptomic data were obtained were collected in various locations, transported live to the laboratory, transferred to RNAlater, frozen at -80C, or kept in ethanol at -20C.

Sequencing and Assembly
Total RNA was extracted from fresh or preserved samples, and cDNA libraries were prepared using the SMART cDNA Library Construction Kit (Clontech; see Supplemental Experimental Procedures for further details). Samples were sequenced with 454 FLX, 454 Titanium, or Illumina HiSeq 2000. Generated sequence data were augmented with publicly available data. Taxa sequenced by 454 or Sanger methods were assembled and processed using the EST2uni pipeline [Forment et al, 2008], while those sequenced by Illumina were digitally

normalized using khmer [Brown et al, 2012] and assembled using Trinity [Ebersberger, et al, 2009]. Contigs (for Illumina libraries) or contigs + high-quality singletons (for 454 and Sanger libraries) were translated using TransDecoder (http:// sourceforge.net/projects/transdecoder/). Table S2 provides the number of unigenes (contigs for Illumina libraries and contigs + singletons for 454 and Sanger libraries) obtained for each taxon.

Identification of putative ortholog groups (OGs) was conducted with HaMStR [Ebersberger et al, 2009] using the ‘‘model organisms’’ reference taxon set. After orthology determination, sequences from two 454 libraries from Ophionotusvictoriae and four libraries from Rhabdopleura species were combined into chimeric OTUs in order to reduce the amount of missing data per taxon.

Orthology groups were filtered following the approach of Kocot et al. [2011] First, only sequences greater than 100 aa in length, and OGs with at least 15ambulacrarian taxa and including at least one pterobranch sequence, were retained for further analyses. Sequences less than 100 aa in length were deleted because they are likely to be incorrectly aligned. To remove mistranslated sequence ends, we trimmed amino acid sequences when stop codons (marked by X) were present in either the first or last 20 characters. Each OG was aligned with MAFFT [Katoh et al, 2005] and then trimmed with Aliscore and Alicut [Misof and Misof, 2009] to remove columns with ambiguous alignment.

Next, individual alignments were manually evaluated for partially mistranslated sequences, which were deleted or trimmed as appropriate. Single-OG trees were then constructed for each OG using RAxML v7.3.8 [Stamatakis, 2006] with the PROTGAMMALGF model. Individual gene trees were manually evaluated for sequence contamination as described in Supplemental Experimental Procedures.After manual screening of alignments and individual OG trees, 299 alignments remained, constituting the basis for the 299/33 data set. To screen for potential paralogs, we used PhyloTreePruner [Kocot et al, 2013] (details in Supplemental Experimental Procedures). Alignments passing screening via PhyloTreePruner constituted the 185/33 and 185/31 data sets. To test the effect of low-coverage taxa and missing data on our phylogenetic reconstruction, we employed MARE [Meyer et al, 2010] to generate the 165/20 data set (weighting of information content parameter alpha = 3).

Phylogenetic Analyses
Maximum-likelihood phylogenetic analyses of all data sets were conducted using RAxML v7.7.6 [Stamatakis, 2006] using a PROTGAMMALGF model for each individual OG partition. This model was the most appropriate for the majority of genes and was among the best-fitting models for nearly all OGs as assessed by ProtTest [Darriba et al, 2011]. Previous work [Kocot et al, 2011] and preliminary analyses on these data sets indicated that employing a single model across all partitions recovers highly similar trees while proving considerably less computationally expensive than specifying individual models for all partitions. Nodal support was assessed with 1,000 replicates of nonparametric bootstrapping. Bootstrapped trees from the 185/33 data set were used to calculate leaf stability and taxonomic instability indices of each OTU using the RogueNaRok server (http://rnr.h-its.org). Competing hypotheses ofambulacrarian phylogeny were evaluated using the SH test [Shimodaira, 2002] as implemented in RAxML with the PROTGAMMALGF model for each OG partition. Bayesian inference analyses were conducted using PhyloBayes 2.3 [Lartillot et al, 2013] with the CAT model, which accounts for site specific rate heterogeneity, with two independent chains run for 11,000 cycles each and 1,100 cycles discarded as burn-in. Topology and posterior consensus support was generated using the bpcomp program within PhyloBayes; convergence of the two chains was indicated by ‘‘maxdiff’’ values below 0.2. All phylogenetic analyses were conducted on the Auburn University CASIC HPC supercomputer.

Accession Numbers
Sequence Read Archive (SRA) accession numbers for all data are given in Table S2. The four data matrices are available from the Dryad Digital Repository (http://datadryad.org) with the DOI http://dx.doi.org/10.5061/dryad.20s7c.

Data Processing Description

BCO-DMO Processing notes:
- added conventional header with dataset name, PI name, version date
- modified parameter names to conform with BCO-DMO naming conventions
- added links to NCBI accession pages

[table of contents | back to top]

Data Files

File
Cannon_2014_TS2.csv (Comma Separated Values (.csv), 7.51 KB) MD5:6bd017616d4b23c446311856e9810b7f Primary data file for dataset ID 671521

[table of contents | back to top]

Related Publications

Cannon, J. T., Kocot, K. M., Waits, D. S., Weese, D. A., Swalla, B. J., Santos, S. R., & Halanych, K. M. (2014). Phylogenomic Resolution of the Hemichordate and Echinoderm Clade. Current Biology, 24(23), 2827–2832. doi:10.1016/j.cub.2014.10.016
Results

Cannon, J. T., Kocot, K. M., Waits, D. S., Weese, D. A., Swalla, B. J., Santos, S. R., & Halanych, K. M. (2015). Data from: Phylogenomic resolution of the Hemichordate and Echinoderm clade (Version 1) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.20S7C https://doi.org/10.5061/dryad.20s7c
Different Version

Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evolutionary Biology, 9(1), 157. doi:10.1186/1471-2148-9-157
Methods

Forment, J., Gilabert, F., Robles, A., Conejero, V., Nuez, F., & Blanca, J. M. (2008). EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. BMC Bioinformatics, 9(1), 5. doi:10.1186/1471-2105-9-5
Methods

Katoh, K. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research, 33(2), 511–518. doi:10.1093/nar/gki198
Methods

Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A., … Halanych, K. M. (2011). Phylogenomics reveals deep molluscan relationships. Nature, 477(7365), 452–456. doi:10.1038/nature10382
Methods

Kocot, K. M., Citarella, M. R., Moroz, L. L., & Halanych, K. M. (2013). PhyloTreePruner: A Phylogenetic Tree-Based Approach for selection of Orthologous sequences for phylogenomics. Evolutionary Bioinformatics, 9, EBO.S12813. doi:10.4137/ebo.s12813 https://doi.org/10.4137/EBO.S12813
Methods

Lartillot, N., Rodrigue, N., Stubbs, D., & Richer, J. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Systematic Biology, 62(4), 611–615. doi:10.1093/sysbio/syt022
Methods

Meyer, B., Meusemann, K., and Misof, B. (2010). MARE v0.1.2-rc. http://mare.zfmk.de/.
Software

Misof, B., & Misof, K. (2009). A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion. Systematic Biology, 58(1), 21–34. doi:10.1093/sysbio/syp006
Methods

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. Systematic Biology, 51(3), 492–508. doi:10.1080/10635150290069913
Methods

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22(21), 2688–2690. doi:10.1093/bioinformatics/btl446
Methods

[table of contents | back to top]

Related Datasets

IsSupplementedBy

Halanych, K. M., Mahon, A. (2016) **Collection and locality information of Hemichordata and Echinodermata from global sites, Table S1, Cannon et al (2014) Curr. Bio. (Antarctic Inverts project).** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2016-12-22 http://lod.bco-dmo.org/id/dataset/671493 [view at BCO-DMO]

[table of contents | back to top]

Parameters

Parameter	Description	Units
species	taxonomic genus and species name	unitless
taxa_new_to_study	whether these are new sequences for this study	unitless
source	genomics services lab and instrument	unitless
num_reads	number of reads	reads
num_unigenes	number of uni-genes (contigs for Illumina libraries and contigs + singletons for 454 and Sanger libraries) obtained for each taxon	genes
HaMStR_OGs	???orthologous genes from HaMStr profile	unitless
NCBI_accession	NCBI GenBank accession number	unitless
link_NCBI_accession	link to GenBank accession page	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina MiSeq (San Diego, CA) at Auburn University
Generic Instrument Name	Automated DNA Sequencer
Generic Instrument Description	General term for a laboratory instrument used for deciphering the order of bases in a strand of DNA. Sanger sequencers detect fluorescence from different dyes that are used to identify the A, C, G, and T extension reactions. Contemporary or Pyrosequencer methods are based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step.

Dataset-specific Instrument Name	
Generic Instrument Name	Thermal Cycler
Generic Instrument Description	A thermal cycler or "thermocycler" is a general term for a type of laboratory apparatus, commonly used for performing polymerase chain reaction (PCR), that is capable of repeatedly altering and maintaining specific temperatures for defined periods of time. The device has a thermal block with holes where tubes with the PCR reaction mixtures can be inserted. The cycler then raises and lowers the temperature of the block in discrete, pre-programmed steps. They can also be used to facilitate other temperature-sensitive reactions, including restriction enzyme digestion or rapid diagnostics. (adapted from http://serc.carleton.edu/microbelife/research_methods/genomics/pcr.html)

[table of contents | back to top]

Deployments

NBP1210

Website	https://www.bco-dmo.org/deployment/568987
Platform	RVIB Nathaniel B. Palmer
Report	http://dmoserv3.bco-dmo.org/jg/serv/BCO-DMO/OA_Antarctic_organisms/727518.html0%7Bdir=dmoserv3.whoi.edu/jg/dir/BCO-DMO/OA_Antarctic_organisms/,info=dmoserv3.bco-dmo.org/jg/info/BCO-DMO/OA_Antarctic_organisms/mg_ca_ratios%7D
Start Date	2013-01-06
End Date	2013-02-09
Description	Seaglider AUV-SG-503-2012 was recovered on this cruise.

Halanych_lab_2011-16

Website	https://www.bco-dmo.org/deployment/671488
Platform	Auburn University lab
Start Date	2011-08-01
End Date	2016-07-31
Description	Invertebrate genomics

[table of contents | back to top]

Project Information

Genetic connectivity and biogeographic patterns of Antarctic benthic invertebrates (Antarctic Inverts)

Coverage: Antarctica

Extracted from the NSF award abstract:

The research will explore the genetics, diversity, and biogeography of Antarctic marine benthic invertebrates, seeking to overturn the widely accepted suggestion that benthic fauna do not constitute a large, panmictic population. The investigators will sample adults and larvae from undersampled regions of West Antarctica that, combined with existing samples, will provide significant coverage of the western hemisphere of the Southern Ocean. The objectives are: 1) To assess the degree of genetic connectivity (or isolation) of benthic invertebrate species in the Western Antarctic using high-resolution genetic markers. 2) To begin exploring planktonic larvae spatial and bathymetric distributions for benthic shelf invertebrates in the Bellinghausen, Amundsen and Ross Seas. 3) To continue to develop a Marine Antarctic Genetic Inventory (MAGI) that relates larval and adult forms via DNA barcoding.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Office of Polar Programs (formerly NSF PLR) (NSF OPP)	PLR-1043745
NSF Office of Polar Programs (formerly NSF PLR) (NSF OPP)	PLR-1043670

[table of contents | back to top]