# Conservation of known vertebrate cyclooxygenase functional domains and residues, from Havird et al (2015) J Mol Evol., Table 3 (Antarctic Inverts project)

**Website**: https://www.bco-dmo.org/dataset/671770
**Data Type**: document
**Version**:
**Version Date**: 2016-12-27

**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 |

## Table of Contents


- Dataset Description
  - Methods & Sampling
  - Data Processing Description
- Data Files
- Parameters
- Deployments
- Project Information
- Funding


## Dataset Description

This dataset was published as Table 3 from Havird et al (2015). It contains a list of mostly invertebrate taxa and whether or not they possess known vertebrate cyclooxygenase functional domains and residues.

Notes:
- AA residue: based on sheep COX-1 numbering (GenBank accession# NP_001009476)
- Many features in Brachiopoda are unknown (?) due to an incomplete (20 %) sequence being the only representative from this phylum.
- Annelida excludes Sipuncula sequences.
- Cephalochordata: see Havird et al. (2008) for conservation of domains and residues in Cnidaria and other Chordata.


**Related Reference:** Havird, J.C., K. M. Kocot, P. M. Brannock, J. T. Cannon, D. S. Waits, D. A. Weese, S. R. Santos, K. M. Halanych. 2015. Reconstruction of cyclooxygenase evolution in animals suggests variable, lineage-specific prostaglandin synthesis pathways. Journal of Molecular Evolution. 80:193-208. DOI 10.1007/s00239-015-9670-3


**Related Datasets:**
Havird_2015_T1: NCBI accessions: COX and PXDN
Havird_2015_T2: NCBI accessions: MPO and PXDN

### Methods & Sampling

From Havird et al (2015):

***Data Acquisition-Publically Available Data***

*The basic local alignment search tool (BLAST, Altschul et al. 1990) was used to screen publically available genomic and transcriptomic datasets (Online Resource 1) for COX homologs. Remote BLASTs were performed for genomic datasets either via the Ensembl Genome Browser (Flicek et al. 2014), NCBI's GenomeBLAST, or to the Joint Genome Institute (JGI). Specifically, BLASTp searches were performed by utilizing previously described COX sequences as queries: COX-1b from mummichog (Fundulus heteroclitus, Chordata: Vertebrata, GenBank accession ACH73265.1), COX-a from sea squirt (Ciona intestinalis, Chordata: Urochordata, Ensembl accession ENSCINP00000013352), COXa1 from oyster (Crassostrea gigas, Mollusca, GenBank accession ACP28169.2), COX from the water flea (Daphnia pulex, Arthropoda, GenBank accession EFX85708.1), COX-A from Arctic soft coral (Gersemia fruticosa, Cnidaria, GenBank accession AAF93168.1), and COX (15R specific) from sea whip (Plexaura homomalla, Cnidaria, GenBank accession AAF93169.1). This panel, which encompasses the known phylogenetic diversity of animal lineages possessing COX, was assembled to maximize the probability that COX homologs spanning varying levels of divergence would be recovered. BLAST searches were performed in a taxondirected manner across all major metazoan lineages and complemented cases where new genetic resources were generated (see below). Publically available transcriptomic data were either downloaded as raw reads and assembled de novo (as for soft corals, see below) or as assembled sets of contigs (i.e., hard corals) and then searched via BLAST as above. The approximately top five (5) BLAST 'hits' with e-values of 10^-20 or less were retained from each search and used in preliminary phylogenetic analyses with previously characterized COX sequences to determine if candidates were in fact members of COX or closely related gene families (see Sequence alignment and phylogenetic analyses section).*

*In addition to the above searching of genomic and transcriptomic datasets, previously described COX sequences (Havird et al. 2008 and Table 1 of Kawamura et al. 2014) were also included in subsequent phylogenetic analyses. GenBank's protein (i.e., nr) database was also searched for COX sequences, with ones from chordate lineages being discarded (except for lineages of special interest such as coelacanth and Epaulette shark) and nonchordate lineages retained.*

***Data Acquisition-Novel Data***

*Novel transcriptomic data, generated from a range of invertebrate lineages and covering ~250 taxa (Online Resource 2), were also searched for COX homologs using either tBLASTn (which is based on translated protein similarity; Altschul et al. 1990) as described above or by searching for several specific COX-related terms (e.g., cyclooxygenase, PG, COX, PGHS, and arachidonic acid) among annotation records of assembled contigs. While these transcriptomes will be described in their entirety elsewhere, methods for RNA extraction, cDNA library generation and sequencing as well as assembly and contig annotation generally followed those for previously described molluscan, crustacean, and hemichordate transcriptomes (Kocot et al. 2011; Genomic Resources Development Consortium et al. 2014; Cannon et al. 2014). When searching novel transcriptomes, top BLAST hits with e-values of 10^-20 or less were retained from each transcriptome for preliminary phylogenetic analyses. A listing of transcriptomes from which COX candidates was queried is given in Online Resource 2.*

***Sequence Alignment and Phylogenetic Analyses***

*Amino acid sequences from COX candidates identified from BLAST searches were aligned with previously characterized COX sequences taken from Havird et al. (2008) and Kawamura et al. (2014) using MUSCLE (Edgar 2004) as implemented in SeaView version 4 with default values (Gouy et al. 2010). Poorly aligned and/or divergent regions of resulting alignments were trimmed using Gblocks version 0.91b (Castresana 2000) with the following parameters: the minimum number of 50 % of sequences was selected to identify conserved and flanking positions, maximum number of contiguous non-conserved sequences in a block set to 10, and gap positions allowed in all blocks. This approach resulted in final alignments of ~450 amino acids (roughly 25 % of amino acids were discarded). Preliminary phylogenetic trees were then generated from trimmed alignments using the parallelized form of FastTreeMP version 2.1.7 (Price et al. 2010). FastTreeMP was used in preliminary analyses instead of PhyML or RAxML because these latter programs were too computationally intensive for the numerous preliminary analyses that were performed (Price et al. 2010). Thirteen sequences from major animal lineages that were not COX homologs, but instead members of the related gene families myeloperoxidase (MPO) andperoxidasin (PXDN), were retained as outgroups, consistent with a previous phylogenetic analysis indicating MPO and PXDN sequences are appropriate outgroups for COX (Daiyasu and Toh 2000). If candidate sequences clustered with previously characterized COX genes, they were retained as possible homologs, while sequences were discarded if they clustered with MPO/PXDN sequences to the exclusion of the COX clade. This approach was iterated several times (as opposed to one large, computationally intensive preliminary analysis) for subsets of candidate sequences (inferred via BLAST) and known COX homologs (totaling ~200 sequences per subset), with candidate sequences either being discarded or retained as potentially 'true' homologs for each subset.*

*Candidate sequences passing the above screenings were then combined with previously described COX sequences (defined as the 'complete dataset') for additional phylogenetic analyses using Maximum Likelihood (ML) and Bayesian Inference (BI). ML analyses were performed using the pthreads implementation of RAxML*

*version 8.0.24 (Stamatakis 2006; Ott et al. 2007; Stamatakis 2014), with runs (performed in duplicate) consisting of 1000 rapid bootstrap replicates (Stamatakis et al. 2008) using the PROTGAMMALG4X model of protein evolution (as specified with the -m flag). PROTGAMMALG4X uses the Gamma model of rate heterogeneity (Yang 1996) and the LG4X model of protein substitution, which has been shown to significantly outperform single matrix substitution models (Le et al. 2012). BI analyses were performed with PhyloBayes version 3.3 (Lartillot and Philippe 2004; Lartillot et al. 2009) and utilized two runs of four chains to infer phylogeny based on the same trimmed amino acid alignment as in ML analyses. The CAT model of protein evolution was used (as an alternative to the LG4X model, which is not available in PhyloBayes), with 15,000 cycles (corresponding to ~15,000,000 generation) and a 10 % burn-in (when likelihood scores had reached stability). Maximum discrepancy (i.e., max diff) between bipartitions was 0.128, suggesting that independent chains had reached a similar, stable point in treespace, and the number of generations was 'acceptable' to give a good overview of the posterior consensus (Lartillot and Philippe 2004). The ML and BI analyses were conducted at the Alabama Supercomputer Center (ASC) in Huntsville, Alabama. In an effort to encourage future studies, all sequences and alignments are available publically via http://www.auburn. edu/~santosr/sequencedatasets.htm.*

*Based on the examination of preliminary trees, nodal support values, and phylogenetic affinities from the FastTreeMP analyses, 'rogue' sequences, or ones decreasing statistical support for otherwise robust clades, were suspected in the dataset. 'Rogues', which can be due to incomplete sequences (e.g., 53 % of novel COX sequences identified here were incomplete), incomplete taxon sampling, accelerated substitution rates leading to long branches, or the possible misclassification of non-COX sequences as COX homologs (Sanderson and Shaffer 2002), were identified via the RogueNaRok webservice (Aberer et al. 2013; http://rnr.h-its.org/) under default parameters. Additional ML and BI phylogenetic analyses were then conducted using RAxML and PhyloBayes under the same parameters as above with these sequences excluded (designated the 'rogues removed' dataset).*

*Phylogenetic placement of cnidarian COX sequences was particularly interesting based on the final topology (see below). Given this, alternate topological placements of cnidarian COX sequences were compared to the most likely topology using Shimodaira-Hasegawa (SH) tests (Shimodaira and Hasegawa 1999) as implemented in RaxML via the -f h flag.*

***Characterization of Novel COX Sequences***
*To further characterize novel genes identified in this study, those sequences used in final phylogenetic analyses (i.e., the 'complete dataset') were searched for conserved protein domains, motifs, and residues characteristic of COX functionality. Domains were characterized using InterProScan version 5 (Jones et al. 2014) and all features were manually checked by eye. Predicted protein features indicative of COX were derived from previous descriptions of known vertebrate COX homologs (Kulmacz et al. 2003; Simmons et al. 2004; Ishikawa et al. 2007; Havird et al. 2008) including residues for COX activity (tyrosine-385, histidine-388, and serine-530), peroxidase activity (glutamine- 203 and histidine-207), and substrate binding (arginie-120), as well as heme-binding, membrane-binding, and dimerization domains.*

**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
- replaced special characters: ± with +/- and ? with 'unknown'

[ table of contents | back to top ]

## Data Files

| File |
|---|
| **Havird_2015_T3.csv** (Comma Separated Values (.csv), 1.45 KB) MD5:2dfea201375ff1862e0869c78821d385 Primary data file for dataset ID 671770 |

[ table of contents | back to top ]

## Parameters

| Parameter | Description | Units |
|---|---|---|
| Domain_or_residue | vertebrate cyclooxygenase functional domains and residues | unitless |
| AA_Residue | amino acid peroxidase residue | unitless |
| Amphimdeon | poriferan species | unitless |
| Insecta | invertebrate taxon | unitless |
| Brachiopoda | invertebrate taxon | unitless |
| Nemertea | invertebrate taxon | unitless |
| Annelida | invertebrate taxon | unitless |
| Mollusca | invertebrate taxon | unitless |
| Sipuncula | invertebrate taxon | unitless |
| Myriapoda | invertebrate taxon | unitless |
| Crustacea | invertebrate taxon | unitless |
| Cephalochordata | vertebrate taxon | unitless |

[ table of contents | back to top ]

## Deployments

**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 ]