Another biomineralising protostome with an *msp130* gene and conservation of *msp130* gene structure across Bilateria.

Réka Szabó and David E.K. Ferrier.

Affiliation: The Scottish Oceans Institute, Gatty Marine Laboratory, University of St Andrews, East Sands, St Andrews, Fife, KY16 8LB, UK.

e-mails: rs386@st-andrews.ac.uk

dekf@st-andrews.ac.uk

Corresponding author: David E.K. Ferrier

Tel.: +44 (0)1334 463480

e-mail: dekf@st-andrews.ac.uk

MSP130 proteins form a distinctive family of cell surface glycoproteins associated with mineralised skeletal elements in larval and adult echinoderms (Anstrom et al. 1987; Illies et al. 2002; Mann et al. 2008). In a recent paper in Evolution and Development, Ettensohn (2014) investigated the phylogenetic distribution of the MSP130 family, finding that it is only present in prokaryotes and a handful of disparate eukaryotic clades. Among animals, only deuterostomes and certain molluscs possessed *msp130* genes. Based on this patchy distribution, Ettensohn proposed that animals acquired *msp130* genes by two events of horizontal gene transfer, once in the deuterostome stem and once in the molluscan lineage. He also hypothesised that these acquisitions were related to the evolution of biomineralisation in these clades. However, he recognised the need for more data from other animals, such as calcifying annelids. Here we report the presence of an *msp130* gene in such an animal (the keelworm, *Spirobranchus* (formerly *Pomatoceros*) *lamarcki*), thus providing the first evidence for protostome *msp130* genes outside the phylum Mollusca, enhancing Ettensohn's hypothesised link between *msp130* genes and biomineralising animals. However, we question the hypothesis of multiple origins, based on comparisons of gene structure between deuterostomes and molluscs. Ettensohn allows for the possibility that bilaterian *msp130* genes originated in a single HGT event followed by extensive losses. Our findings of conserved introns lend support to a single origin, either through vertical inheritance or via a single horizontal transfer from a eukaryotic source.

Annelids possess MSP130

We have obtained transcriptome data from intact and regenerating specimens of the calcified head appendage (operculum) of the serpulid annelid *S. lamarcki*. Two contigs from our combined assembly likely represent two halves of a single *msp130* family transcript. This transcript is reasonably abundant in both intact and regenerating opercula, but with more reads in the regenerating (pre-calcifying and strongly calcifying) datasets. The transcript appears to contain a complete open reading frame (ORF) encoding a 604-amino acid protein with an N-terminal signal peptide predicted by TargetP v1.1 (Emanuelsson et al. 2000) with high confidence. We successfully amplified and cloned the 3’ end of the transcript from cDNA derived from mixed-stage opercular regenerates. The PCR fragment is predicted to span two conserved intron locations (see next section). Thus, an *msp130* gene is present in *S. lamarcki* and expressed in a structure with calcareous components.

Alignment of the *S. lamarcki* protein with selected eukaryotic MSP130 proteins and one of the partial sequences we cloned is shown in supplementary Figure S1. We used the full list of MSP130 proteins provided by Ettensohn (2014) to generate neighbour-joining and maximum likelihood trees (Fig. 1). In our analysis, the *S. lamarcki* transcript groups among other metazoan sequences, but support values at nodes separating protostome from deuterostome genes within Metazoa are not significant, making it difficult to establish whether protostome and deuterostome sequences form distinct clades. However, the tree features a well-supported branch uniting all metazoan sequences, consistent with a single origin for metazoan *msp130* genes.

Conserved gene structure across Bilateria

To further address the question of multiple origins, we compared the exon-intron structures of *msp130* genes in the chordate *Branchiostoma floridae*, the sea urchin *Strongylocentrotus purpuratus* and the mollusc *Lottia gigantea* using available whole genome sequences. Where possible, exon boundaries were also confirmed using publicly available EST data. We discovered several intron locations that appear conserved across most or all genes in these three bilaterians; alignments of these regions are shown in Figure 1B. While conservation between bilaterians and algae is more limited, at least one site is predicted to be in a highly conserved sequence motif in all examined genes, and is supported by EST evidence in both *Lottia* and the brown alga *Ectocarpus siliculosus*.

The high level of conservation in bilaterian intron positions makes convergent evolution an unlikely explanation. In our opinion, these data are most consistent with the presence of *msp130* genes in the urbilaterian, with secondary gene loss in some lineages such as the annelids *Capitella teleta* and

Helobdella robusta, and several arthropods (as surveyed by Ettensohn, 2014) for which whole genome sequences are available. Additionally, we did not find an *msp130* gene in the whole genome sequence of the biomineralising cnidarian, *Acropora digitifera*, in addition to the other supposedly non-biomineralising non-bilaterian taxa surveyed by Ettensohn (2014). Thus, this cnidarian data does not resolve whether the *msp130* genes in bilaterians stem from a HGT event early in bilaterian evolution, before the origin of protostomes and deuterostomes, or instead the gene was present at the origin of animals and has been secondarily lost, not only from some bilaterian lineages, but also several non-bilaterian lineages. Given the preponderance of gene loss across animal evolution (Putnam et al. 2007; Takahashi et al. 2009; Maeso et al. 2012) such extensive secondary loss of *msp130* genes is perhaps not so surprising. Ettensohn (2014) hypothesised that the acquisition of *msp130* genes was related to the acquisition of mineralised skeletons. While our findings about gene structure call for a revision of this hypothesis, it would be equally interesting to see whether the *retention* of this family is associated with biomineralisation (skeletal or non-skeletal), although the link between *msp130* and animal biomineralisation is clearly not absolute, given the absence of the gene from a biomineralising cnidarian.

References

Anstrom, J.A., Chin, J.E., Leaf, D.S., Parks, A.L., and Raff, R.A. 1987. Localization and expression of msp130, a primary mesenchyme lineage-specific cell surface protein in the sea urchin embryo. Development *101*, 255–265.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol *300*, 1005–1016.

Ettensohn, C.A. 2014. Horizontal transfer of the *msp130* gene supported the evolution of metazoan biomineralization. Evol Dev *16*, 139-148.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol *59*, 307–321.

Illies, M.R., Peeler, M.T., Dechtiaruk, A.M., and Ettensohn, C.A. 2002. Identification and developmental expression of new biomineralization proteins in the sea urchin *Strongylocentrotus purpuratus*. Dev Genes Evol *212*, 419–431.

Maeso, I., Roy, S.W., and Irimia, M. 2012. Widespread recurrent evolution of genomic features. Genome Biol Evol *4*, 486–500.

Mann, K., Poustka, A.J., and Mann, M. 2008. The sea urchin (*Strongylocentrotus purpuratus*) test and spine proteomes. Proteome Sci *6*, 22.

Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science *317*, 86–94.

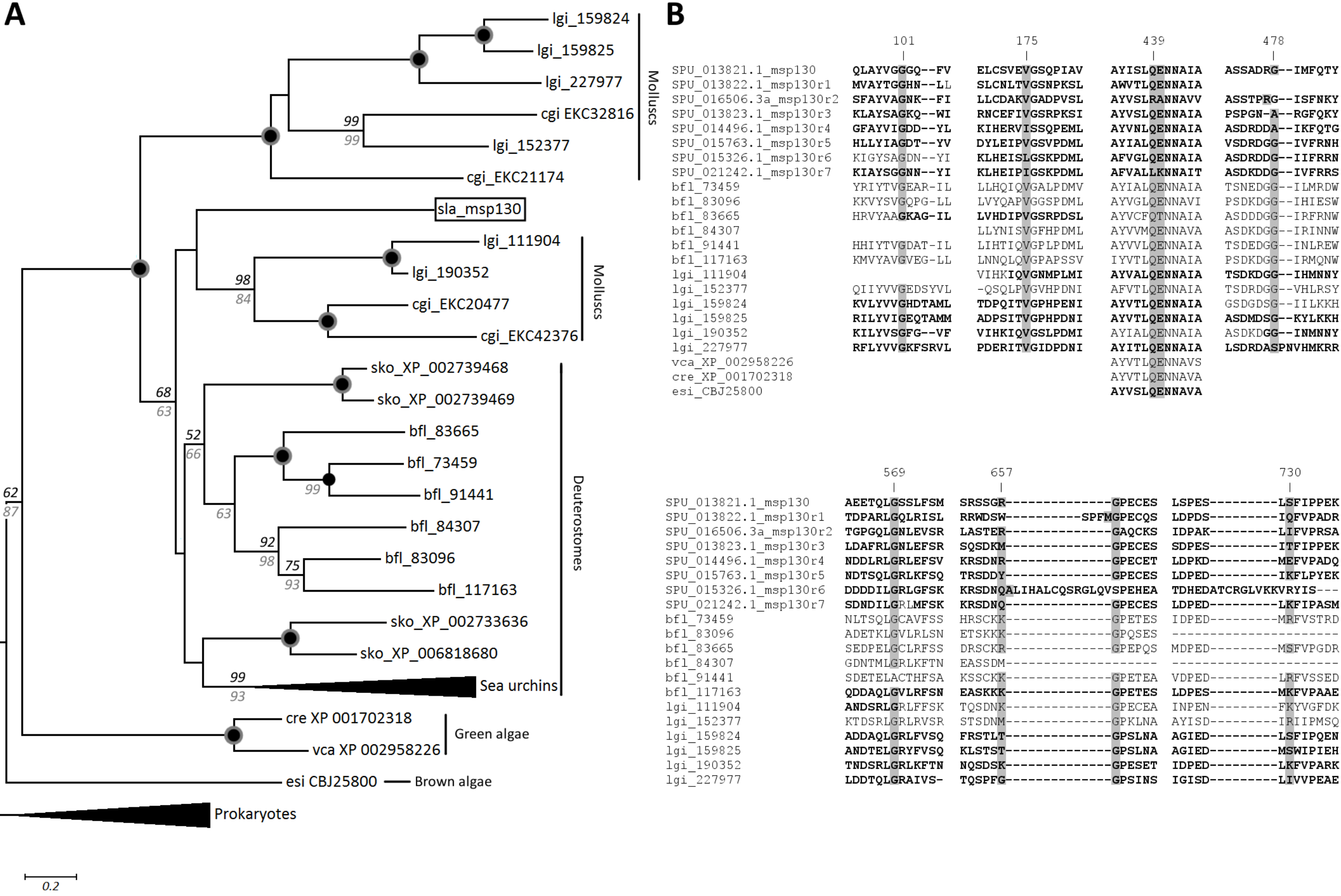
Takahashi, T., McDougall, C., Troscianko, J., Chen, W.-C., Jayaraman-Nagarajan, A., Shimeld, S., and Ferrier, D.E.K. 2009. An EST screen from the annelid *Pomatoceros lamarckii* reveals patterns of gene loss and gain in animals. BMC Evol Biol *9*, 240.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol mst197.

Tu, Q., Cameron, R.A., Worley, K.C., Gibbs, R.A., and Davidson, E.H. 2012. Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. Genome Res. *22*, 2079–2087.

Figure legends

Figure 1. A. Maximum likelihood phylogenetic tree of available MSP130 protein sequences. The tree was generated in PhyML 3.0 (Guindon et al. 2010; http://atgc.lirmm.fr/phyml/)using the LG+G model with 500 bootstrap replicates. Support values over 50% are indicated; dots represent 100% support. Support values (percentage of 500 replicates) from a neighbour-joining tree of the same dataset generated in MEGA 6 (Tamura et al. 2013) are indicated in grey. Prokaryotic and echinoid sequences have been collapsed to reduce tree size. The *Spirobranchus lamarcki* sequence is highlighted with a box. **B.** Alignment of the residues surrounding conserved intron positions in bilaterian or eukaryotic MSP130 sequences. Sequences are based on gene models; bolded sequences have EST support or, in case of *Strongylocentrotus purpuratus*, mapped RNA-seq data (Tu et al. 2012). Highlighted residues indicate the position of the intron; numbers above the alignment indicate the position of the intron in *Strongylocentrotus purpuratus* MSP130. Algal sequences are only included where they show conservation with animals. Spu-msp130r1 has been omitted from the alignment at position 478 because of a large insertion in this region. Bfl_84307 and bfl_83096 are incomplete. Species abbreviations: bfl *Branchiostoma floridae*, cgi *Crassostrea gigantea*, cre *Chlamydomonas reinhardtii*, esi *Ectocarpus siliculosus*, lgi *Lottia gigantea*, sla *Spirobranchus lamarcki*, sko *Saccoglossus kowalevskii*, spu *Strongylocentrotus purpuratus*, vca *Volvox carteri*.



```
                                  101            175               439               478
                                   |              |                 |                 |
SPU_013821.1_msp130        QLAYVGGGQ--FV    ELCSVEVGSQPIAV    AYISLQENNAIA    ASSADRG--IMFQTY
SPU_013822.1_msp130r1      MVAYTGGHN--LL    SLCNLTVGSNPKSL    AWVTLQENNAIA
SPU_016506.3a_msp130r2     SFAYVAGNK--FI    LLCDAKVGADPVSL    AYVSLRANNAVV    ASSTPRG--ISFNKY
SPU_013823.1_msp130r3      KLAYSAGKQ--WI    RNCEFIVGSRPKSI    AYVSLQENNAIA    PSPGN-A--RGFQKY
SPU_014496.1_msp130r4      GFAYVIGDD--YL    KIHERVISSQPEML    AYVNLQENNAIA    ASDRDDA--IKFQTG
SPU_015763.1_msp130r5      HLLYIAGDT--YV    DYLEIPVGSVPDML    AYIALQENNAIA    VSDRDGG--IVFRNH
SPU_015326.1_msp130r6      KIGYSAGDN--YI    KLHEISLGSKPDML    AFVGLQENNAIA    ASDRDDG--IIFRRN
SPU_021242.1_msp130r7      KIAYSGGNN--YI    KLHEIPIGSKPDML    AFVALLKNNAIT    ASDKDDG--IVFRRS
bfl_73459                  YRIYTVGEAR-IL    LLHQIQVGALPDMV    AYIALQENNAIA    TSNEDGG--ILMRDW
bfl_83096                  KKVYSVGQPG-LL    LVYQAPVGGSPDML    AYVGLQENNAVI    PSDKDGG--IHIESW
bfl_83665                  HRVYAAGKAG-IL    LVHDIPVGSRPDSL    AYVCFQTNNAIA    ASDDDGG--IRFRNW
bfl_84307                                   LLYNISVGFHPDML    AYVVMQENNAVA    ASDKDGG--IRINNW
bfl_91441                  HHIYTVGDAT-IL    LIHTIQVGPLPDML    AYVVLQENNAIA    TSDEDGG--INLREW
bfl_117163                 KMVYAVGVEG-LL    LNNQLQVGPAPSSV    IYVTLQENNAIA    PSDKDGG--IRMQNW
lgi_111904                                  VIHKIQVGNMPLMI    AYVALQENNAIA    TSDKDGG--IHMNNY
lgi_152377                 QIIYVVGEDSYVL    -QSQLPVGVHPDNI    AFVTLQENNAIA    TSDRDGG--VHLRSY
lgi_159824                 KVLYVVGHDTAML    TDPQITVGPHPENI    AYVTLQENNAIA    GSDGDSG--IILKKH
lgi_159825                 RILYVIGEQTAMM    ADPSITVGPHPDNI    AYVTLQENNAIA    ASDMDSG--KYLKKH
lgi_190352                 KILYVSGFG--VF    VIHKIQVGSLPDMI    AYIALQENNAIA    ASDKDGG--INMNNY
lgi_227977                 RFLYVVGKFSRVL    PDERITVGIDPDNI    AYITLQENNAIA    LSDRDASPNVHMKRR
vca_XP_002958226                                              AYVTLQENNAVS
cre_XP_001702318                                              AYVTLQENNAVA
esi_CBJ25800                                                  AYVSLQENNAVA
```

```
                                 569          657                                        730
                                  |            |                                          |
SPU_013821.1_msp130        AEETQLGSSLFSM   SRSSGR--------------GPECES   LSPES---------LSFIPPEK
SPU_013822.1_msp130r1      TDPARLGQLRISL   RRWDSW----------SPFMGPECQS   LDPDS---------IQFVPADR
SPU_016506.3a_msp130r2     TGPGQLGNLEVSR   LASTER--------------GAQCKS   IDPAK---------LIFVPRSA
SPU_013823.1_msp130r3      LDAFRLGNLEFSR   SQSDKM--------------GPECES   SDPES---------ITFIPPEK
SPU_014496.1_msp130r4      NDDLRLGRLEFSV   KRSDNR--------------GPECET   LDPKD---------MEFVPADQ
SPU_015763.1_msp130r5      NDTSQLGRLKFSQ   TRSDDY--------------GPECES   LDPED---------IKFLPYEK
SPU_015326.1_msp130r6      DDDDILGRLGFSK   KRSDNQALIHALCQSRGLQVSPEHEA   TDHEDATCRGLVKKVRYIS---
SPU_021242.1_msp130r7      SDNDILGRLMFSK   KRSDNQ--------------GPECES   LDPED---------LKFIPASM
bfl_73459                  NLTSQLGCAVFSS   HRSCKK--------------GPETES   IDPED---------MRFVSTRD
bfl_83096                  ADETKLGVLRLSN   ETSKKK--------------GPQSES   ----------------------
bfl_83665                  SEDPELGCLRFSS   DRSCKR--------------GPEPQS   MDPED---------MSFVPGDR
bfl_84307                  GDNTMLGRLKFTN   EASSDM--------------------   ----------------------
bfl_91441                  SDETELACTHFSA   KSSCKK--------------GPETEA   VDPED---------LRFVSSED
bfl_117163                 QDDAQLGVLRFSN   EASKKK--------------GPETES   LDPES---------MKFVPAAE
lgi_111904                 ANDSRLGRLFFSK   TQSDNK--------------GPECEA   INPEN---------FKYVGFDK
lgi_152377                 KTDSRLGRLRVSR   STSDNM--------------GPKLNA   AYISD---------IRIIPMSQ
lgi_159824                 ADDAQLGRLFVSQ   FRSTLT--------------GPSLNA   AGIED---------LSFIPQEN
lgi_159825                 ANDTELGRYFVSQ   KLSTST--------------GPSLNA   AGIED---------MSWIPIEH
lgi_190352                 TNDSRLGRLKFTN   NQSDSK--------------GPESET   IDPED---------LKFVPARK
lgi_227977                 LDDTQLGRAIVS-   TQSPFG--------------GPSINS   IGISD---------LIVVPEAE
```

Supplementary information

Methods summary

Identification of *sla_msp130*

Three total RNA samples (derived from unoperated, 2-day and 6-day regenerating opercular filaments) were sequenced on the Illumina HiSeq2000 platform as 100-bp paired-end reads. The three sets of reads were pooled and assembled de novo using Trinity. Two msp130 contigs were identified in the assembly using spu-msp130 as a TBLASTN query; the four-codon overlap between these contigs was confirmed using the raw reads. To confirm the expression of this gene in the regenerating opercular filament, Primer3web was used to design primers to amplify the 3' end of the predicted transcript including two conserved splice junctions (see Figure S1) and part of the 3' UTR. A mixed-stage cDNA sample containing material from 8-hour and 1-4 day regenerates (kindly donated by Mr Tom Barton-Owen) was used as the template. PCR bands of the expected size were extracted and transformed into competent cells using the pGEM-T Easy® vector. Eight clones were sequenced, all of which were close matches (90-98% nucleotide level identity) to the contig sequence. Clone 5, which produced the closest match, was uploaded to Genbank under accession KM588349.

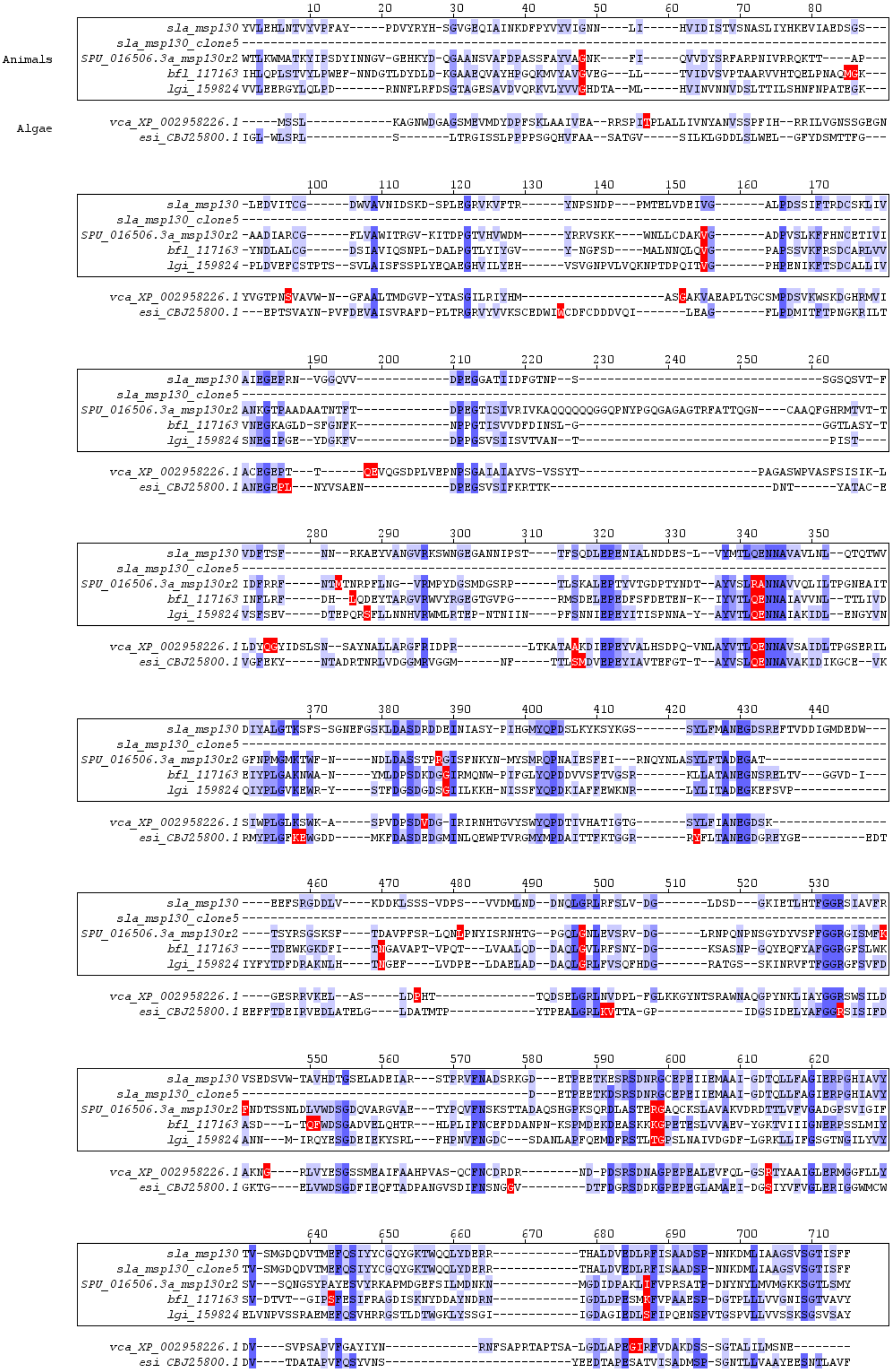
Alignment and phylogeny

Sequences used for phylogenetic reconstruction were based on the list provided in Ettensohn's (2014) supplementary information. Species, databases and accessions are given in Table S1, along with notes on corrections to automatic gene models. The *Acropora digitifera* genome (version 1.1, available at http://marinegenomics.oist.jp/) was also searched, but no good matches were found. Sequences were aligned with MAFFT and edited manually to remove poorly aligned regions and long repetitive stretches such as the polyG tracts in several sea urchin sequences. A neighbour-joining tree was constructed in MEGA 6 using the default settings, and a maximum likelihood tree was built in PhyML3.0 with the subtree pruning and regrafting search method, 500 bootstrap replicates, and the LG+G model selected by all three information criteria in Modelgenerator v0.85.

Gene structure in bilaterians and algae

Gene models were examined in six eukaryotes with sequenced genomes, representing the breadth of the distribution of *msp130* genes in this domain based on Ettensohn's (2014) findings. Species and genome databases were as follows: the deuterostomes *Strongylocentrotus purpuratus* (SpBase, version 3.1 assembly) and *Branchiostoma floridae* (JGI, Brafl1 and 2), the protostome *Lottia gigantea* (JGI, Lotgi1), the green algae *Chlamydomonas reinhardtii* (JGI, Chlre4) and *Volvox carteri* (Phytozome v10), and the brown alga *Ectocarpus siliculosus* (http://bioinformatics.psb.ugent.be/genomes/view/Ectocarpus-siliculosus). To confirm predicted splice junctions, the NCBI EST database was searched for matches to each predicted protein except those in the sea urchin, for which the RNA-seq data available through SpBase were used; in addition, ESTs were sought from the amphioxus cDNA resource (http://amphioxus.icob.sinica.edu.tw/) for *B. floridae*, and the aforementioned genome resource for *E. siliculosus*.

Figure S1. Alignment of sla_msp130 with the best matching PCR clone derived from cDNA from a mixed sample of 8-hour and 1- to 4-day opercular regenerates, and selected bilaterian and algal sequences. The animal sequences are indicated by boxes. Columns are shaded according to the level of conservation, and intron positions where known are highlighted in red. Non-conserved N- and C-termini have been trimmed off; no other columns were removed. Species abbreviations: sla *Spirobranchus lamarcki*, spu *Strongylocentrotus purpuratus*, bla *Branchiostoma floridae*, lgi *Lottia gigantea*, vca *Volvox carteri*, esi *Ectocarpus siliculosus*.



```
Animals
                              10        20        30        40        50        60        70        80
          sla_msp130 YVLEHLNTVYVPFAY-----PDVYRYH-SGVGEQIAINKDFPYVYVIGNN----LI-----HVIDISTVSNASLIYHKEVIAEDSGS---
   sla_msp130_clone5 ------------------------------------------------------------------------------------------
SPU_016506.3a_msp130r2 WTLKWMATKYIPSDYINNGV-GEHKYD-QGAANSVAFDPASSFAYVAGNK----FI-----QVVDYSRFARPNIVRRQKTT----AP---
          bfl_117163 IHLQPLSTVYLPWEF-NNDGTLDYDLD-KGAAEQVAYHPGQKMVYAVGVEG---LL-----TVIDVSVPTAARVVHTQELPNAQMGK---
          lgi_159824 VVLEERGYLQLPD-------RNNFLRFDSGTAGESAVDVQRKVLYVVGHDTA--ML-----HVINVNNVDSLTTILSHNFNPATEGK---

Algae
 vca_XP_002958226.1 -----MSSL------------KAGNWDGAGSMEVMDYDPFSKLAAIVEA--RRSPITPLALLIVNYANVSSPFIH--RRILVGNSSGEGN
      esi_CBJ25800.1 IGL-WLSRL------------S-------LTRGISSLPPPPSGQHVFAA--SATGV-----SILKLGDDLSLWEL--GFYDSMTTFG---

                             100       110       120       130       140       150       160       170
          sla_msp130 -LEDVITCG------DWVAVNIDSKD-SPLEGRVKVFTR------YNPSNDP---PMTELVDEIVG-------ALPDSSIFTRDCSKLIV
   sla_msp130_clone5 ------------------------------------------------------------------------------------------
SPU_016506.3a_msp130r2 -AADIARCG------FLVAWITRGV-KITDPGTVHVWDM------YRRVSKK----WNLLCDAKVG-------ADPVSLKFFHNCETIVI
          bfl_117163 -YNDLALCG------DSIAVIQSNPL-DALPGTLYIYGV------Y-NGFSD----MALNNQLQVG-------PAPSSVKFRSDCARLVV
          lgi_159824 -PLDVEFCSTPTS--SVLAISFSSPLYEQAEGHVILYEH------VSVGNPVLVQKNPTDPQITVG-------PHPENIKFTSDCALLIV

 vca_XP_002958226.1 YVGTPNSVAVW-N--GFAALTMDGVP-YTASGILRIYHM--------------------ASGAKVAEAPLTGCSMPDSVKWSKDGHRMVI
      esi_CBJ25800.1 ---EPTSVAYN-PVFDEVAISVRAFD-PLTRGRVYVVKSCEDWIWCDFCDDDVQI-------LEAG-------FLPDMITFTPNGKRILT

                             190       200       210       220       230       240       250       260
          sla_msp130 AIEGEPRN--VGGQVV-------------DPEGGATIIDFGTNP--S----------------------------------SGSQSVT-F
   sla_msp130_clone5 ------------------------------------------------------------------------------------------
SPU_016506.3a_msp130r2 ANKGTPAADAATNTFT-------------DPEGTISIVRIVKAQQQQQQGGQPNYPGQGAGAGTRFATTQGN----CAAQFGHRMTVT-T
          bfl_117163 VNEGKAGLD-SFGNFK-------------NPPGTISVVDFDINSL-G----------------------------------GGTLASY-T
          lgi_159824 SNEGIPGE--YDGKFV-------------DPPGSVSIISVTVAN--T-----------------------------------PIST----

 vca_XP_002958226.1 ACEGEPT---T------QEVQGSDPLVEPNPSGAIAIAYVS-VSSYT-------------------------PAGASWPVASFSISIK-L
      esi_CBJ25800.1 ANEGEPL---NYVSAEN------------DPEGSVSIFKRTTK-------------------------------DNT------YATAC-E

                             280       290       300       310       320       330       340       350
          sla_msp130 VDFTSF-----NN--RKAEYVANGVRKSWNGEGANNIPST----TFSQDLEPENIALNDDES-L--VYMTLQENNAVAVLNL--QTQTWV
   sla_msp130_clone5 ------------------------------------------------------------------------------------------
SPU_016506.3a_msp130r2 IDFRRF-----NTMTNRPFLNG--VRMPYDGSMDGSRP------TLSKALEPTYVTGDPTYNDT--AYVSLRANNAVVQLILTPGNEAIT
          bfl_117163 INFLRF-----DH--LQDEYTARGVRWVYRGEGTGVPG------RMSDELEPEDFSFDETEN-K--IYVTLQENNAIAVVNL--TTLIVD
          lgi_159824 VSFSEV-----DTEPQRSFLLNNHVRWMLRTEP-NTNIIN----PFSNNIEPEYITISPNNA-Y--AYVTLQENNAIAKIDL--ENGYVN

 vca_XP_002958226.1 LDYQGYIDSLSN--SAYNALLARGFRIDPR----------LTKATAAKDIEPEYVALHSDPQ-VNLAYVTLQENNAVSAIDLTPGSERIL
      esi_CBJ25800.1 VGFEKY-----NTADRTNRLVDGGMRVGGM------NF-----TTLSMDVEPEYIAVTEFGT-T--AYVSLQENNAVAKIDIKGCE--VK

                             370       380       390       400       410       420       430       440
          sla_msp130 DIYALGTKSFS-SGNEFGSKLDASDRDDEINIASY-PIHGMYQPDSLKYKSYKGS-------SYLFMANEGDSREFTVDDIGMDEDW---
   sla_msp130_clone5 ------------------------------------------------------------------------------------------
SPU_016506.3a_msp130r2 GFNPMGMKTWF-N-----NDLDASSTPRGISFNKYN-MYSMRQPNAIESFEI---RNQYNLASYLFTADEGAT-----------------
          bfl_117163 EIYPLGAKNWA-N-----YMLDPSDKDGGIRMQNW-PIFGLYQPDDVVSFTVGSR-------KLLATANEGNSRELTV---GGVD-I---
          lgi_159824 QIYPLGVKEWR-Y-----STFDGSDGDSGIILKKH-NISSFYQPDKIAFFEWKNR-------LYLITADEGKEFSVP-------------

 vca_XP_002958226.1 SIWPLGLKSWK-A-----SPVDPSDVDG-IRIRNHTGVYSWYQPDTIVHATIGTG-------SYLFIANEGDSK----------------
      esi_CBJ25800.1 RMYPLGFKEWGDD-----MKFDASDEDGMINLQEWPTVRGMYMPDAITTFKTGGR-------RYFLTANEGDGREYGE---------EDT

                             460       470       480       490       500       510       520       530
          sla_msp130 ----EEFSRGDDLV----KDDKLSSS-VDPS---VVDMLND--DNQLGPLRFSLV-DG-------LDSD---GKIETLHTFGGRSIAVFR
   sla_msp130_clone5 ------------------------------------------------------------------------------------------
SPU_016506.3a_msp130r2 ----TSYRSGSKSF----TDAVPFSR-LQNLPNYISRNHTG--PGQLGNLEVSRV-DG------LRNPQNPNSGYDYVSFFGGRGISMFK
          bfl_117163 ----TDEWKGKDFI----TNGAVAPT-VPQT---LVAALQD--DAQLGVLRFSNY-DG-------KSASNP-GQYEQFYAFGGRGFSLWK
          lgi_159824 IYFYTDFDRAKNLH----TNGEF----LVDPE--LDAELAD--DAQLGRLFVSQFHDG-------RATGS--SKINRVFTFGGRGFSVFD

 vca_XP_002958226.1 ----GESRRVKEL--AS-----LDPHT--------------TQDSELGRLNVDPL-FGLKKGYNTSRAWNAQGPYNKLIAYGGRSWSILD
      esi_CBJ25800.1 EEFFTDEIRVEDLATELG----LDATMTP------------YTPEALGRLKVTTA-GP------------IDGSIDELYAFGGRSISIFD

                             550       560       570       580       590       600       610       620
          sla_msp130 VSEDSVW-TAVHDTGSELADEIAR---STPRVFNADSRKGD---ETPEETKESRSDNRGCEPEIIEMAAI-GDTQLLFAGIERPGHIAVY
   sla_msp130_clone5 ----------------------------------------D---ETPEETKDSRSDNRGCEPEIIEMAAI-GDTQLLFAGIERPGHIAVY
SPU_016506.3a_msp130r2 FNDTSSNLDLVWDSGDQVARGVAE---TYPQVFNSKSTTADAQSHGPKSQRDLASTERGAQCKSLAVAKVDRDTTLVFVGADGPSVIGIF
          bfl_117163 ASD---L-TQFWDSGADVELQHTR---HLPLIFNCEFDDANPN-KSPMDEKDEASKKKGPETESLVVAEV-YGKTVIIIGNERPSSLMIY
          lgi_159824 ANN---M-IRQYESGDEIEKYSRL---FHPNVFNGDC---SDANLAPFQEMDFRSTLTGPSLNAIVDGDF-LGRKLLIFGSGTNGILYVY

 vca_XP_002958226.1 AKNG----RLVYESGSSMEAIFAAHPVAS-QCFNCDRDR--------ND-PDSRSDNAGPEPEALEVFQL-GSRTYAAIGLERMGGFLLY
      esi_CBJ25800.1 GKTG----ELVWDSGDFIEQFTADPANGVSDIFNSNGGV---------DTFDGRSDDKGPEPEGLAMAEI-DGSIYVFVGLERIGGWMCW

                             640       650       660       670       680       690       700       710
          sla_msp130 TV-SMGDQDVTMEFQSIYYCGQYGKTWQQLYDERR------------THALDVEDLRFISAADSP-NNKDMLIAAGSVSGTISFF
   sla_msp130_clone5 TV-SMGDQDVTMEFQSIYYCGQYGKTWQQLYDERR------------THALDVEDLRFISAADSP-NNKDMLIAAGSVSGTISFF
SPU_016506.3a_msp130r2 SV---SQNGSYPAYESVYRKAPMDGEFSILMDNKN------------MGDIDPAKLIFVPRSATP-DNYNYLMVMGKKSGTLSMY
          bfl_117163 SV-DTVT--GIPSFESIFRAGDISKNYDDAYNDRN------------IGDLDPESMKFVPAAESP-DGTPLLLVVGNISGTVAVY
          lgi_159824 ELVNPVSSRAEMEFQSVHRRGSTLDTWGKLYSSGI------------IGDAGIEDLSFIPQENSPVTGSPVLLVVSSKSGSVSAY

 vca_XP_002958226.1 DV----SVPSAPVFGAYIYN-------------RNFSAPRTAPTSA-LGDLAPEGIRFVDAKDSS-SGTALILMSNE--------
      esi_CBJ25800.1 DV----TDATAPVFQSYVNS--------------------------YEEDTAPESATVISADMSP-SGNTLLVAAYEESNTLAVF
```

Inferred nucleotide and protein sequence of sla_msp130 based on two manually merged Trinity contigs

```
>comp389772_merge
ATGTAACTATATGTACCGAATTCTGTCTAATTTTCAGTGGTTGAGAAGAGAGTGATAAAAAATAAGTGCACGC
CATTCTCGAAAAAGTCATAAAATGTGCTTACGTATAAGTGACAGGAGACGCCGAAGTCATTCCACTTATTACA
AATGGAAAAAGGATAATTAAAAGAAGAAACGTCAATGTAATTCAAACTGGAAAAAAATCAGATTTAGAATTA
TTAAATTGTTGAATAATATTCTAATTAAAAAATCACACATAATCAACAGCAAAATGGCTTTTAAGTGTATATTTT
CTATCGTAACCCTATTGATTTGTGTTGTTCGTGGAAAGTATGTATTGGAGCATCTCAACACGGTCTATGTGCCT
TTCGCATATCCGGATGTGTATAGATACCATAGCGGTGTTGGAGAACAAATTGCTATCAATAAGGATTTTCCATA
TGTGTATGTTATAGGAAATAATCTGATACACGTTATCGATATTTCAACTGTGTCAAACGCCTCTCTCATCTATCA
TAAGGAAGTCATAGCTGAAGATAGTGGTAGCCTTGAGGATGTCATAACATGCGGGGACTGGGTAGCTGTTAA
TATTGACAGCAAAGATTCCCCACTTGAAGGGAGGGTGAAAGTATTCACGAGATACAACCCAAGCAATGACCCT
CCCATGACAGAACTAGTGGATGAAATAGTTGGAGCCCTACCCGATAGTTCGATATTCACACGGGACTGCAGCA
AGCTGATCGTGGCCATAGAGGGAGAGCCGAGGAATGTGGGCGGCCAGGTCGTGGACCCGGAAGGGGGCGC
CACAATCATCGACTTTGGTACAAATCCCTCCAGTGGTAGCCAATCAGTTACCTTTGTTGACTTCACCAGCTTTAA
TAACAGGAAGGCGGAATATGTGGCCAATGGTGTTCGCAAGAGTTGGAATGGTGAAGGAGCTAACAATATTCC
ATCAACAACATTTTCTCAAGATCTGGAGCCTGAAAATATTGCGCTGAATGATGATGAATCTTTGGTGTATATGA
CATTGCAGGAGAATAATGCAGTTGCCGTTCTCAACCTGCAAACACAAACCTGGGTTGATATATACGCGCTGGG
GACCAAGAGCTTCAGCAGTGGTAACGAATTCGGCAGCAAACTAGACGCAAGCGACAGAGATGATGAAATCA
ATATAGCAAGTTATCCTATCCATGGTATGTATCAACCGGATAGTCTGAAGTACAAGTCCTACAAAGGGAGCAG
TTATTTGTTCATGGCCAACGAGGGAGACTCGCGAGAGTTCACCGTAGATGACATTGGCATGGACGAGGACTG
GGAGGAGTTCAGCAGAGGAGATGACCTTGTGAAAGATGATAAGCTCAGTTCAAGTGTGGACCCCTCGGTGGT
AGATATGCTAAACGATGATAACCAGCTAGGTCGGTTACGCTTCAGTCTAGTGGACGGATTGGACAGTGATGG
CAAAATAGAAACTCTGCATACATTTGGTGGTCGCAGTATCGCTGTATTTAGAGTATCAGAGGACAGCGTGTGG
ACGGCGGTACATGATACGGGCAGTGAGCTGGCGGACGAGATAGCTAGGTCCACACCCAGGGTCTTCAACGC
AGACTCCAGGAAAGGAGATGAAACCCCAGAGGAAACCAAAGAAAGTAGATCAGATAACCGGGGATGCGAAC
CGGAAATAATAGAAATGGCTGCTATCGGTGACACACAACTACTATTTGCGGGTATAGAGCGCCCAGGCCACA
TCGCCGTCTACACGGTTAGCATGGGCGACCAAGATGTAACCATGGAGTTTCAGAGTATATACTACTGTGGCCA
GTACGGAAAAACATGGCAACAGCTTTATGACGAGCGCAGGACGCATGCTCTTGATGTGGAGGATTTAAGGTT
TATTAGCGCAGCGGACAGCCCTAATAATAAGGACATGTTAATTGCCGCGGGCTCGGTTAGTGGCACAATCTCC
TTCTTCCAAGTAACCGATGATGGGGAAATGCCCGTGGAAAATAATTCAGGTGAAGGAAGGCATCTGTCAGCG
GTGCTGGGAGCTCTAATTGGAGCAATAACTTTATTGCTGGTGTCTGCATGAGGACAACCTCTCAATGGACCCA
AAGAAATGAAGGCACCGCTCGAACTGCCCGATGAAACATCTGTTCGATGGC
```

```
>comp389772_merge_ORF
MAFKCIFSIVTLLICVVRGKYVLEHLNTVYVPFAYPDVYRYHSGVGEQIAINKDFPYVYVIGNNLIHVIDISTVSNASLIY
HKEVIAEDSGSLEDVITCGDWVAVNIDSKDSPLEGRVKVFTRYNPSNDPPMTELVDEIVGALPDSSIFTRDCSKLIVAI
EGEPRNVGGQVVDPEGGATIIDFGTNPSSGSQSVTFVDFTSFNNRKAEYVANGVRKSWNGEGANNIPSTTFSQDL
EPENIALNDDESLVYMTLQENNAVAVLNLQTQTWVDIYALGTKSFSSGNEFGSKLDASDRDDEINIASYPIHGMYQ
PDSLKYKSYKGSSYLFMANEGDSREFTVDDIGMDEDWEEFSRGDDLVKDDKLSSSVDPSVVDMLNDDNQLGRLR
```

FSLVDGLDSDGKIETLHTFGGRSIAVFRVSEDSVWTAVHDTGSELADEIARSTPRVFNADSRKGDETPEETKESRSD
NRGCEPEIIEMAAIGDTQLLFAGIERPGHIAVYTVSMGDQDVTMEFQSIYYCGQYGKTWQQLYDERRTHALDVED
LRFISAADSPNNKDMLIAAGSVSGTISFFQVTDDGEMPVENNSGEGRHLSAVLGALIGAITLLLVSA

Table S1. Sources of sequences used for alignment and phylogenetic reconstruction. Where the final sequence used differs from the one under the given accession, a brief note on modifications is given.

Higher taxon	**Species**	**Database**	**Accession**	**Notes**
Deuterostomes	Strongylocentrotus purpuratus	SpBase	SPU_013821.1	
		SpBase	SPU_013822.1	
		SpBase	SPU_016506.3a	
		SpBase	SPU_013823.1	
		SpBase	SPU_014496.1	
		SpBase	SPU_015763.1	
		SpBase	SPU_015326.1	Chosen as the more complete of Ettensohn's two accessions for msp130rel6.
		SpBase	SPU_021242.1	
	Heliocidaris erythrogramma	Genbank	her_CAC20358.1	
	Heliocidaris tuberculata	Genbank	htu_CAC20589.2	
	Eucidaris tribuloides	EchinoBase transcripts	etr_6853_2	
		EchinoBase transcripts	etr_38352	
		EchinoBase transcripts	etr_1749_2	
		EchinoBase transcripts	etr_21290	
		EchinoBase transcripts	etr_10262	
		EchinoBase transcripts	etr_22657_1	
	Saccoglossus kowalevskii	RefSeq	sko_XP_002739468.1	

		RefSeq, Metazome	sko_XP_002739469.1	A probably spurious region in the RefSeq sequence was replaced by the conserved exon beginning ENNAI based on the genomic sequence
		RefSeq	sko_XP_002733636.1	
		RefSeq, Metazome	sko_XP_006818680.1	This sequence was not listed by Ettensohn but appears distinct from the others. A missing conserved exon was reconstructed from genomic sequence.
	Branchiostoma floridae	JGI	73459	A conserved exon was added to the gene model from genomic sequence
		JGI	83096	
		JGI, NCBI EST	83665, FE585043.1	N-terminus of the gene model was replaced by sequence from the EST – JGI pipeline probably missed first exon
		JGI	84307	C-terminus probably spurious (poorly aligned and on far side of 17 kb sequencing gap) and thus removed from alignment
		JGI	91441	
		JGI	117163	
Protostomes	Lottia gigantea	JGI, NCBI EST	111904, FC586646 .1	The gene prediction seems to have missed the first exon and mispredicted portions of the protein due to a frame shift. Corrected based on alignment of the EST to the genome.
		JGI	152377	
		JGI	159824	
		JGI	159825	

		JGI, NCBI EST	190352, FC596551.1, FC761686.1	ESTs were used to extend the N-terminus
	Crassostrea gigas	Genbank	EKC20477.1	
			EKC42376.1	
			EKC21174.1	
			EKC32816.1	
	Spirobranchus lamarcki	own	n/a	Merged from two contigs with a slight overlap confirmed by raw read BLAST.
Green algae	Chlamydomonas reinhardtii	RefSeq	XP_001702318.1	
	Volvox carteri	RefSeq	XP_002958226.1	
Brown algae	Ectocarpus siliculosus	Genbank	CBJ25800.1	
Bacteria and archaea	Comamonas testosterone	RefSeq	WP_003075952.1	
	Corynebacterium lipophiloflavum	RefSeq	WP_006839975.1	
	Cyanothece sp	RefSeq	YP_001804430.1	
	Cyanothece sp	RefSeq	ZP_01729342.1	
	Cyanothece sp	RefSeq	YP_003887791.1	
	Cyanothece sp	RefSeq	YP_002377271.1	
	Cyanothece sp	RefSeq	YP_003138182.1	
	Desulfuromonas acetoxidans	RefSeq	WP_006002570.1	
	Gallaecimonas xiamenensis	RefSeq	WP_008482472.1	
	Haloferax gibbonsi	RefSeq	WP_004975775.1	
	Halorubrum teberiquichense	RefSeq	WP_006630545.1	
	Halosimplex carlsbadense	RefSeq	WP_006883787.1	
	Janibacter hoylei	RefSeq	WP_007928713.1	
	Ketogulonicigenium vulgare	RefSeq	YP_003964898.1	

	Leptothrix cholodnii	RefSeq	YP_001792238.1	
	Planktomyces maris	RefSeq	WP_002648981.1	
	Planktomyces limnophilus	RefSeq	YP_003631837.1	
	Plesiocystis pacifica	RefSeq	WP_006972868.1	
	Pseudoalteromonas sp	RefSeq	WP_008130733.1	