Supporting Information

Full Methods

Animal culture and regeneration

M. lignano was kept in petri dishes with nutrient-enriched f/2 medium (1) and fed ad libitum with diatom algae (*Nitzschia curvilineata*). Climate chamber conditions were set at 20°C, 60% humidity and a 14/10 h day/night cycle. For regeneration, worms were cut at the post-pharyngeal level in order to completely remove the gonads. The anterior part was kept under normal conditions with diatoms. 100 worms were collected for further processing at 0h, 3h, 6h, 12h, 24h, 48h and 72h after cutting.

Sequencing library preparation, DNA and RNA isolation

For DNA extractions whole worms (*M. lignano*) – starved for 3 days (to reduce diatom contamination from the gut) – or flies (*D. melanogaster*) were incubated in Proteinase K buffer (10 mM Tris·Cl; 25 mM EDTA; 100 mM NaCl; 0.5% SDS, pH=8.0) and digested with 1.5ug/ml of Proteinase (K) overnight at 50°C. DNA was extracted with phenol and chloroform, precipitated using 70% EtOH and resuspended in TE buffer. For Illumina sequencing DNA was sonicated to ~180bp size using Covaris and the standard manufacturer’s protocol. DNA-Seq libraries were prepared using the Ovation Ultralow Library Systems (Nugen) and sequenced on the Illumina GAII or Hiseq 2000 (PE100) platforms.

For PacBio sequencing we used 10µg of DNA per library. To ensure good DNA quality we ran a pulse-field gel (Pippin Pulse, Sage Sciences) before each library preparation. We sheared the DNA to ~10 Kbp using the g-TUBE (Covaris), according to the manufacturer's specifications. The libraries were prepared using the Pacbio library preparation kit, RS II, according to the manufacturer’s instructions. Ligation was extended to 16 hours. Following ligation, we performed the size selection (Blue Pippin, Sage Sciences) in 0.75% dye-free agarose and

0.5X TBE. The selected size was 6-15 Kbp and the equipment was set to resolve in 1-100 Kbp size range, according to the manufacturer's manual. The libraries were sequenced using either the p4c2 or p5c3 chemistry and standard run parameters. The movie time in each case was set to the longest time possible.

For RNA-Seq libraries 200-400 worms were resuspended in TRIzol reagent (Ambion) for RNA extraction according to manufacturer's instruction. For transcriptome assembly 3 Script seg V2 libraries were constructed according to manufacturer's specifications. One library was prepared from total RNA, one from rRNA-depleted RNA (Ribo-Gold Epibio, according to the manufacturer's specifications), and one from polyA-selected RNA (Poly(A)Purist MAG kit, Life Technologies, according to the manufacturer's specifications). Two additional sequencing libraries were generated using Encore Complete RNA-Seq DR Multiplex System according to the manufacturer's instructions

RNA-Seq libraries for the regeneration studies were generated using the Encore Complete RNA-Seq DR Multiplex System according to manufacturer's instructions. Samples were sequenced on Illumina Hiseq 2000 (PE100).

Transcriptome assembly and annotation

The transcriptome assembly was done using the Trinity package provided by the Broad Institute (2, 3), with the following parameters --SS_lib_type FR --normalize_reads --trimmomatic. The libraries included in the assembly were: total RNA prepared from 100 worms, polyA- selected RNA, ribo-depleted RNA (see above).

The transcriptome annotation was performed using Trinotate, the Trinity annotation pipeline (2). Transcripts were first blasted against SwissProt and Uniref90 and then analyzed with HMMER v3.1b2 (http://hmmer.janelia.org/) using the Pfam-A hmm. The results were loaded into a sqlite database and consolidated by Trinotate.

Alignment of transcripts to the genome was done using BLASTn with an e-value cutoff of 1e-5. The best HSPs were filtered using the LIS algorithm to find the best non overlapping set for each transcript.

The set of putative miRNAs was established from the small RNA sequencing library by selecting sequences that were 22 or 23 nucleotides in length and supported by at least 3 reads. The sequences of this size were shown to be miRNAs in *M.lignano* (4). This set was further refined by keeping only those sequences that had a BLAST hit in miRBase. This set was then mapped to the ML2 genome and the highest scoring HSPs were selected to determine the miRNA locations.

Genome Assembly

The Illumina Assembly (ML1) was built using SGA (github https://github.com/jts/sga 9/8/14) using 115x coverage of 101bp paired-end Illumina HiSeq data. Only contigs that were greater than 200bp in length were kept in the final assembly. This cutoff was chosen because is the average length of the fragments sequenced in the experiment. The salient parameters used were: sga index -a ropebwt --no-reverse; sga correct -k --discard --learn; sga index -a ropebwt; sga filter -x 2 --homopolymer-check --low-complexity-check; sga fm-merge -m 55; sga index -a ropebwt; sga rm-dup; sga overlap -m 55; sga assemble -m 55 -g 0.05 -r 10

Pacbio data was self-corrected using HGAP, obtained from github https://github.com/PacificBiosciences/HBAR-DTK on 11/21/14. Only reads greater than 10kb were used in the correction process. After correction, reads were assembled using the Celera Assembler v8.2beta generating the ML2 assembly. The salient Celera parameters used for assembly were: frgMinLen =

5000; ovlErrorRate = 0.03; utgGraphErrorRate = 0.02; ovlMinLen=1500; utgGenomeSize = 700000000; unitigger=bogart

A sample of 81665 contigs from the Illumina assembly (~10%) were aligned to all of the contigs in the Pacbio assembly using Mummer v. 3.23. The subprogram 'nucmer' was run with the flags --maxmatch -l 100 followed by 'dnadiff' on the resulting delta file, The pipeline produced a report file containing the per-base identity.

In order to exclude the possibility of contamination of our assembly with other species (i.e. diatoms) contigs were blasted using BLASTn against the non-redundant nucleotide database from NCBI. Only hits passing the e-value cutoff of 1e-10 were kept. Results were then filtered using the LIS algorithm to find the best set of hits for each contig. Database hits were then counted, the most common match being to *Caenorhabditis remanei*. These are likely *M. lignano* sequences that have orthologous sequences in other worms.

Genome Annotation

Genome annotation was performed using Maker v2.31.8 (Dec 2014). The Trinity assembly of the transcriptome (2) was used as the EST dataset and the Uniprot_Sprot database was used for the protein homology search. The initial run used the est2genome module to predict gene models directly from the transcript and protein evidence. Snap was then used to refine the gene models in a bootstrap fashion - Maker was run 2 additional times each time supplying the updated hmm generated by Snap.

Transposon analysis

RepeatScout version 1.0.5 was run on both the Illumina and Pacbio assemblies (5). Only repeats that occur at least 10 times in the genome were kept for further analysis. Repeats were annotated using a custom non-redundant library from

NCBI entries (keywords: retrotransposon, transposase, "reverse transcriptase", gypsy, copia) obtained from O. Simakov and colleagues.

K-mer analysis and peak modeling

A K-mer is a substring of length K. When counting the occurrences of these equal length substrings, the choice of K is a trade off between sensitivity and specificity. Shorter K are more robust to sequencing error and heterozygosity while longer K have a lower chance of occurring by random chance. We chose a K size of 23 nucleotides, which is suitable for a genome the size of *Macrostomum*.K-mers were counted in the Illumina data using Jellyfish 1.1.10 (Marcais and Kingsford 2011) with the -C parameter. Peak modeling was performed by fitting a mixture model composed of 4 Poisson distributions and calculating their composite in R.

Differential Expression

Reads were aligned to the transcriptome using RSEM (Li and Dewey 2011) by means of the wrapper script provided by Trinity abundance_estimates_to_matrix.pl. Differentially expressed genes (FDR <=0.001, with a minimum 4-fold change) were identified using DESeq (6). DESeq was run using the wrapper script run_DE_analysis.pl with default parameters. Heatmaps were generated using the perl script analyze_diff_expr.pl also provided by the Trinity package. The clustering methods were left at their default values of --gene_clust complete --gene_dist euclidean.

Gene Ontology Analysis

Gene Ontology terms were summarized from the output of Maker. Biological Process terms were extracted from the gff file and counted to find an overall estimation of their abundance.

Analysis of the transcripts conserved between *H. sapiens*, *M. lignano*, *D. melanogaster*, *S. mediterranea*, or *C. elegans*

Control script (reciprocalblast_allsteps.py) for running reciprocal BLASTp search was obtained from Warren *et el.* (7). Evalue cutoff was set to 1e-10. The trascriptomes/proteomes were obtained from: *C. elegans* (wormbase) ftp://ftp.wormbase.org/pub/wormbase/releases/WS247/species/c_elegans/PRJNA13758/; *D. melanogaster* (flybase) ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.02_FB2014_05/fasta/;
H. sapiens (uniprot) http://www.uniprot.org/help/human_proteome; *S. mediterrantea* form Kao *et al*.(8).

BAC library preparation and sequencing

RxBiosciences (http://www.rxbiosciences.com/) constructed the *M. lignano* bacterial artificial chromosome (BAC) library as previously described (9). We generated 60,000 BACs with an insert size of ~20 KB and 60,000 BACs with an insert size of ~50Kb. We used a modified P[acman] vector (http://pacmanfly.org/images/pacman-bw.jpg), where we replaced the *Drosophila* white gene with *mCherry* driven by an *M. lignano*-specific *Ef1a* promoter and terminated by an *M. lignano* 3'UTR (both sequences were provided to us by Dr. Eugene Berezikov, University of Groningen). The library was cloned into the BamHI restriction site, disrupting the *lacZ* gene. The library was prepared from DNA extracted from 20,000 *M. lignano* individuals.

Individual BACs were grown on 96-well plates as previously described (10). DNA was extracted using the NucleoBond BAC 100 kit (Clontech) according to manufacturer's protocol.

In order to test the completeness of our assembly we pooled individual BACs; 2X 48, 2X 96, 2X 240, and 1X 480 and we sequenced the separate pools using Illumina (PE 100). We removed all the reads mapping to the BAC backbone and

E. coli genome. The remaining reads were mapped to the ML2 assembly using Bowtie 2 (v2.2.3).

Sequence Complexity Analysis

Sequence complexity was calculated on a per read basis using a previously described algorithm (11). If reads were longer than 76 base pairs they were truncated to adjust all the samples to the length of the shortest library. A single complexity number was calculated for each read and the histograms built on a sample of 1 million reads per organism. *C. elegans* sequencing data is public under the SRA run ID SRR1797354. *S. mansoni* data was obtained from ddbj under the accession ERR582487. Human data is from Illumina's public datasource; resequencing of NA12878.

Tandem Repeat Finder masking for low complexity

1 million short reads were obtained from each organism described in the **Sequence Complexity Analysis** method section. Tandem Repeat Finder (12) was run on each sample with the following parameters: 2 7 7 80 10 50 500 -f -d -m -ngs -h. The percentage of bases masked was divided by the total bases found in the sample to get the ratio of low complexity sequence to high complexity sequence.

Estimating CpG content

CpG histograms were built using a previously described method (13). The whole genome was binned into windows of 100 bp and scanned in single nucleotide steps. Only windows with a GC content of at least 50% were considered. The ratio of observed CpG versus expected CpG, CpG[obs/exp] is defined as (Num of CpG/(Num of C × Num of G)) × Total number of nucleotides in the sequence (Gardiner-Garden and Frommer 1987). As a control [obs/exp] ratios of all remaining dinucleotides were calculated using the same method.

Bisulfite genomic DNA sequencing and analysis

M. lignano genomic DNA was sonicated to 200bp fragments in 10mM Tris-HCl, pH 8.0 using Covaris S-series and manufacturer's protocol. 500ng of the fragmented DNA was mixed with 2.5 µl 10x T4 DNA ligase buffer with 10mM ATP (NEB), 1 µl 10mM dNTPs (Roche), 1 µl T4 DNA polymerase (NEB), 1 µl T4 PNK (NEB) and 1 µl Taq DNA polymerase (Roche) in a 25 µl reaction for end-repair and A-tagging. Mix was incubated at 25ºC for 20min followed by 72ºC for 20min. 1 µl of 25 µM pre-annealed methylated forked Illumina TruSeq adaptor with 1 µl T4 DNA ligase (Roche) was added to the mix and brought to a total volume of 30 µl before incubation at 25ºC for 15min. The ligated DNA was purified by Agencourt AMPure XP beads and bisulfite converted using Zymo EZ methylation gold kit following manufacturer's instructions. Illumina-ready library was generated by PCR with annealing temperature of 65ºC using Expand High Fidelity Plus PCR system (Roche) for a minimum of 15 cycles. Reads were mapped to the ML2 assembly and analyzed as previously described (14).

Immunofluorescence and labeling of S-phase cells

The polyclonal Macpiwi1 antibody was produced by PrimmBiotech by rabbit immunization with peptide RPAPPPGLSAQAG (amino acid positions 44-56). Antibodies were purified from serum using synthetic peptides and the sulfolink immobilization kit (Thermo Scientific) according to the manufacturer's instruction. Macpiwi1 staining was performed as previously described (Pfister et al. 2008; De Mulder et al. 2009). For double staining of S-phase cells and Macpiwi1, worms were soaked in 5mM EdU (Life Technologies) for 30min. EdU-positive cells were labeled using the click-iT cell reaction buffer kit (Life Technologies) and Alexa Fluor 594 azide (Life Technologies) according to the manufacturer's instruction, after secondary antibody reaction. Nuclei were stained with DAPI (5µg/ml) at room temperature for 15min. Specimens were mounted with ProLong Gold antifade reagent (Life Technologies) for imaging. Images were captured using a Zeiss LSM 710 confocal microscope.

S-phase cell sorting

At least 10,000 worms (after EDU - secondary antibody staining) were collected and relaxed in a mix of f/2 and 7.14% $MgCl_2$ (1:1) at room temperature for 10min. Relaxed worms were washed in CMFM (88mM NaCl, 1mM KCl, 2.4mM $NaHCO_3$, 7.5mM Tris-HCl (pH 7.6)) on ice (3*5min). Worms were trypsinized with 1% Trypsin in CMFM at 37°C for 20min with agitation. An equal volume of maceration solution (glacial acetic acid: glycerol: H_2O 1:1:13, 9% sucrose) was added, and samples were incubated at room temperature for 1min. Cells were spun down at 5,000g, at 4°C for 10min, resuspended in PBS. Cells were blocked with 2% BSA on ice for 5min and allowed to recover in 500µl 2% FBS in PBS for 10min at 4°C. Hoechst (20µg/ml) was added to cell suspensions, and these were incubated on ice for 30min. Cells were sorted using an Aria IIU cell sorter (BD biosciences) directly into Proteinase K buffer for DNA extraction.

Homeobox survey

We used the complete homeobox inventories from amphioxus (*Branchiostoma floriade*, Deuterostomia) and the red flour beetle (*Tribolium castaneum*, Protostomia) as queries for a comprehensive and saturated search of the transcriptome of *M. lignano*. The choice of using these species is due to the following: (i) their homeobox sequences are less divergent than other members of these groups, (ii) they have not undergone whole genome duplication events as this precludes precise orthology assignment, and (iii) to recover the maximum diversity of homeoboxes as they have the majority of the families well represented. The candidate searches implemented BLASTp searches using as queries the inventories described above, both outcomes were merged and redundancies were removed. Homeodomains were aligned using MAFFT (v7.130b, (15)) and visualized using JALVIEW (v.2.8, (16)) to detect regions of ambiguity, remove them, and remain with the homeodomain region. This alignment was used to produce (i) a neighbour-joining tree (PHYLIP v.3.696, (17)) using the evolutionary model, JTT, and 1000 bootstrap replicates and (ii) a maximum likelihood phylogenetic inference tree (PhyML v.3.0, (18)) using the sequence evolution model, LG+G (gamma = 0.79), using the prediction of the BIC criteria from Modelgenerator (v.851, (19)). The positions within the genome assembly of each homeobox gene were noted to detect some instances of clustering.

Myc Analysis

Mycs and Maxs gene candidates were retrieved based on reciprocal best BLASTp for Myc helix-loop-helix domain from the available platyhelminthes' sequences, (chordates (*Homo sapiens*), poriferans (*Amphimedon queenslandica*), ecdyzosoans (*Drosophila melanogaster*, *Caenorhabditis elegans, Priapulus caudatus*), cnidarians (*Hydra vulgaris*, *Rhabdopleura sp*.) and other lophothrochozoans (*Lottia gigantea* , *Capitella telleta, Golfingia vulgaris, Celebratulus sp.*)). In order to catalogue and infer the history the platyhelminthes putative candidates of Mycs and Maxs, we performed phylogenetic analysis composed of a distance tree inferred using neighbor-joining based on JTT sequence evolution model (1000 bootstrap replicates). Human USF proteins with similar to Myc helix-loop-helix domain are used as an outgroup. Transcriptomes of 24 lophotrochozoan species were assembled from publicly available data using Trinity assembler version 2014-07-17 with parameters --SS_lib_type FR --trimmomatic. Accession numbers: SRX871300, SRX871445, SRX872404, SRX871533, SRX872327, SRX872365, SRX871508, SRX872321, SRX872403, SRX872314, SRX883021, SRX872398, SRX872347, SRX872356, SRX872362, SRX872414, SRX872416, SRX879690, SRX872410, SRX874324, SRX872402, SRX875881, SRX875739, SRX875742. Publicly available Myc and Max sequences: myc_pdu GenBank: AGS55451.1; 166474_cte; GenBank: ELT88315.1; diminutive_dme, GenBank: ABW87508.1; 88480_lgi GenBank: ESO88258.1; MXL3_cel GenBank: CAA94125.1; MXL1_cel GenBank: AAB40926.1, Myc2_hvu GenBank: ADA57607.1; 118760_cte GenBank: ELT88674.1; max_hsa GenBank: AAH25685.1; max_hvu GenBank: ACX32069.1; 133235_lgi GenBank: ESO83519.1; Max_dme GenBank:

AAL90428.1; max_aqu NCBI Reference Sequence: XP_011402619.1; myc_aqu NCBI Reference Sequence: XP_003390966.1; cmyc_hsa GenBank: BAG64849.1; nmyc_hsa GenBank: AAA36370.1; lmyc_Hsa GenBank: CAA30249.1; mycl_hvu NCBI Reference Sequence: XP_002170328.3, mycAl_hvu NCBI Reference Sequence: XP_012556510.1; myc1_hvu GenBank: ACX32068.1; USF2_hsa NCBI Reference Sequence: NP_003358.1; USF1_hsa NCBI Reference Sequence: NP_001263302.1; 785741_scma GenBank: CCD78575.1; 785751_scma GenBank: CCD78574.1; S000209_sma SMU15000209 http://smedgd.stowers.org/cgi-bin/genePage.pl?ref=SMU15000209; S35429_sma SMU15035429 http://smedgd.stowers.org/cgi-bin/genePage.pl?ref=SMU15035429; Max_pdu GenBank: CCK33027.1; Max_dme GenBank: AAL90428.1

SL RNA analysis

Sequences in EST libraries (20) were aligned to the genome using BLASTn. Alignments that were split within the first 100bp were selected (to ensure that the leader is derived from a different genomic location). The sequence that was shared by the majority of these split EST alignments was selected as a candidate leader sequence. Putative SL RNA was identified using BLASTn of identified Leader sequence, followed by GTAAGNATCG, a sequence conserved in other flatworm SL RNAs *(21).* SL RNAs from different flatworm species were aligned using ClustalW *(21)*. Phylogenetic tree of sequence relationships was generated by ClustalW.

Supplementary Figure Legends

Figure S1

A. Sequence complexity comparison across five organisms. *D. melanogaster* has an abundance of very low complexity sequence not found in the other species. *M. lignano* has a sizable amount of moderately complex sequence that are not found in other species and that do not appear to be expressed. **B.** Different populations of dissociated *M. lignano* cells. Cells were analyzed according to a set of criteria including side scatter, forward scatter, Hoechst incorporation (DNA dye) and EdU incorporation (marks the DNA of proliferating S-phase cells). Different populations are marked. EdU-positive cells (EdU+) are the presumptive stem cells. EdU-negative populations divide into Hoechst 4N (Ho+) (Cells that entered S-phase before or after EdU treatment) and Hoechst 2N (Ho-) – enriched in differentiated cells. Cells were sorted based on Hoechst and EdU incorporation. **C.** Tandem Repeat Finder was run on five species to assess their low complexity sequence composition. *M. lignano* had far more bases masked by Tandem Repeat Finder than the other organisms in the test set.

Figure S2

Histogram of the annotated repeats found by RepeatMasker. GA-Rich repeats were the most common repeats found. The frequency was calculated based on the number of bases annotated as a particular type of repetitive element.

Figure S3

A. Distribution of repeat element sizes. Tandem Repeat Finder was run on six genomes and the frequency of each element size was binned. *M. lignano* has a larger number of repetitive elements than other genomes in the sample. The top panel depicts the frequency of repetitive elements normalized by genome size. The bottom panel has the same information log-transformed to highlight the longer elements. **B.** The repeat unit frequency for 10 random samples of 2.6% of the genome. This is compared to the 50 largest contigs which also make up 2.6% of the genome (Figure 3). The data is normalized by the total number of repeats reported in each region. Repeat distribution is similar throughout the genome.

Figure S4

A. Whole genome distribution of CpG observed/expected dinucleotides in *M. lignano*. The ratio was computed using a sliding window of 100bp. **B**. CpG dinucleotide ratio observed/expected. Depletion of CpGs is an indication of genomic methylation. **C**. Dinucleotide ratio (observed/expected) for all dinucleotides in four species, with known and varying whole genome methylation rates.

Figure S5

A. A summary of the predicted genomic features of *M. lignano*. **B.** Distribution of the number of exons per gene. The majority of annotated genes are comprised of 3 exons. **C**. Size distribution of the annotated exons. **D**. Size distribution of the annotated genes.

Figure S6

Pie chart representation of the relative frequency of elements annotated as transposons in the *M. lignano* genome.

Figure S7

A. Assembled transcripts length distribution. The number of transcripts is plotted (Log2 scale). **B.** Gene ontology analysis of *M. lignano* RNA-Seq libraries prepared from whole worms.

Figure S8

A. Alignment between first 130nt of *Macrostomum lignano's* putative SL RNA and SL RNAs from other flatworms. The conserved splice junction is indicated by an arrowhead. Spliced leader sequences are labeled in blue. The potential initiator AUG (last three nucleotides of the spliced leader) is labeled in green. S.med - Schmidtea *mediterranea*. **B.** Phylogenetic tree of sequence relationships of flatworm SL RNAs generated by ClustalW. This is a neighbor-joining tree without distance correlations.

Figure S9

A. The number of reciprocal blast hits between the *M. lignano* and *S. mediterranea* translated transcriptomes. Only the hits passing the E-value cutoff of ≤1e-10 were counted **B.** The number of reciprocal blast hits against the *H. sapiens* transcriptome for four different species. Only the hits passing the E-value cutoff of ≤1e-10 were counted.

Figure S10

A. A diagram representing transcripts that were found in *H. sapiens* as well as in one, two, three, or all of the other species analyzed. Reciprocal blast found 10427 *H. sapiens* genes that were present in at least one of the 4 species analyzed. 1747 genes were present in all four species analyzed. **B.** Gene ontology analysis of the 1949 genes shared with *H. sapiens* that were found in *M. lignano*, but neither in *D. melanogaster* nor *C. elegans*. The selected ontologies were: molecular function, biological process, and protein class.

Figure S11

Known pluripotency pathways from *H. sapiens* and *M. musculus* were adapted from the Kyoto Encyclopedia of Genes and Genomes (22, 23) (http://www.genome.jp/kegg-bin/show_pathway?hsa04550). Factors that had potential homologues in *M. lignano* are labeled.

Figure S12

Evolution of the of Myc and Max gene families across different representatives of the animal phyla. Mycs and Maxs gene candidates are retrieved based on reciprocal best BLASTp from the available transcriptomes. The distance tree was inferred using neighbor-joining based on JTT sequence evolution model (1000 bootstrap replicates). Human USF proteins are used as an outgroup. The Myc branch is labeled in green, the Max branch is labeled in blue. dme – *Drosophila melanogaster*, hsa – *Homo sapiens*, lgi – *Lottia gigantea*, cte – *Capitella teleta*, hvu – *Hydra vulgaris*, aqu – *Amphimedon queenslandica*, cel – *Caenorhabditis elegans*, mli – *Macrostomum lignano*, mfu – *Microdalyellia fusca*, mosp – *Monocelis sp*., psi – *Prosthiostomum siphunculus*, ltr – *Leptoplana tremellaris*, ece – *Echinoplana celerrima*, meli – *Mesostoma lingua* , msc – *Microdalyella schmidtii*, mili – *Microstomum lineare*, nco – *Nematoplana coelogynoporoides*, rsp – *Rhabdopleura sp.*, gvu – *Golfingia vulgaris*, csp – *Cerebratulus sp*., pca – *Priapulus caudatus*, sst – *Stenostomum sthenum*, cle – *Catenula lemnae*, pdu – *Platynereis dumerilli*, sma – *Schmidtea mediterranea*, scma – *Schistosoma mansoni*. Transcript ID is next to each phylum name. For phylogenetic reference see Egger *et al*. (24).

Figure S13

Homeobox gene diversity observed in *M. lignano* in a comparative context with *Tribolium castaneum* and *Branchiostoma floridae* homeobox complements. Phylogenetic analysis is a distance tree inferred using neighbor-joining with a JTT sequence evolution model using the homeodomain sequences (60 aminoacids) from *M. lignano*. Gene classes are indicated by different branch colors and genes with no associated classes are colored in grey. *M. lignano* genes are colored in red. As one could observe there are some classes that are not recovered as monophyletic groups however the majority of the families within the classes are shown to be monophyletic. As no branch support values are shown in here, this tree should be used only to show the diversity of homeodomain sequences.

Figure S14

Classification of all *M. lignano* homeodomain genes using phylogenetic analysis with branch support values using *Tribolium castaneum* and *Branchiostoma floridae* homeodomain complements using the homeodomain sequences (60 aminoacids) from *M. lignano*. This phylogenetic analysis is an aggregation of the support values of the branches inferred upon neighbor-joining with a JTT sequence evolution model (1000 bootstrap replicates) and maximum likelihood LG+G (gamma=0.79). Black asterisks denote branch support based on bootstrap over 70% and blue asterisks denote branch support based on SH-like aLRT over 80%. Gene classes are indicated by different branch colors and genes with no associated classes have grey branches. *M. lignano* genes are colored in red. Majority of the gene families are well supported allowing classifying these homeodomains into *bona-fide* families.

Figure S15

Schematic representation of the experimental design: 200 worms (per replicate) underwent amputation at a level between the brain and the gonads. The heads were allowed to regenerate, and regenerating animals were collected at different timepoints post amputation (0, 3, 6, 12, 24, 48, 72 hours). RNA-Seq libraries from each timepoint were analyzed for differentially expressed genes. At each time point cells were immunostained with Macpiwi1 (green) antibody (raised against RPAPPPGLSAQAG peptide, PrimmBiotech) and for EdU incorporation (Click-iT, EdU imaging kit, Thermofisher) (representing stem cell and dividing cell markers, respectively). Nuclei were labeled with DAPI (blue), h - head, rt - regenerating posterior segment (tail), asterisks denote eyes.

Figure S16

Six synexpression classes of transcripts differentially expressed at different time points after tail amputation were generated by DESeq analysis. Two independent biological replicates are plotted. Grey lines show transcript abundance at different timepoints.

Supplementary Tables

Table S1

Sequence of an abundant 150-mer found in the *M. lignano* genome

Table S2

Sequence of *Macrostomum lignano* putative spliced leader RNA. The spliced leader is labeled in blue. Potential initiator AUG is indicated in bold.

Table S3

Homeobox gene localization in *Macrostomum lignano* genome. Column A) Gene family to which the homeobox gene belongs. Column B) Transcript identifier of the homeobox gene. Column C) Scaffold where the homeobox gene is located.

Supplementary Datasets

SI Dataset 1

Genomic coordinates of the putative DNA-methyltransferase (DNMT) homologs found in the *M. lignano* genome and transcript IDs of the putative Methyl Binding Proteins (MBDs) found in the *M. lignano* transcriptome.

SI Dataset 2

The 25 most abundant transcripts annotated as transposons from RNA seq libraries prepared from 100 whole worms. IDs and annotations are listed.

SI Dataset 3

Analysis of the transcripts conserved between *H. sapiens* and *M. lignano*, *D. melanogaster*, *S. mediterranea*, or *C. elegans* (one worksheet per comparison). Results from reciprocal BLASTp after transcriptome translation. Hsa – *Homo sapiens*, Mlig – *Macrostomum lignano*, Cel – *Caenorhabditis elegans*, Dmel – *Drosophila melanogaster*.

SI Dataset 4

List of transcripts conserved only between *H. sapiens* and *M. lignano*, but not between *M. lignano* and *D. melanogaster* or *C. elegans*. Annotations are based on BLASTp search of translated transcriptomes.

SI Dataset 5

Transcript IDs and annotations of putative homologs of key human and mouse pluripotency factors.

SI Dataset 6

Differentially expressed transcripts from six different synexpression classes. Transcript IDs and annotations, as well as Log2 fold change in expression at seven different timepoints in two replicates are shown (one worksheet per class).

References

1. Andersen RA, Berges JA, P.J. H, & Watanabe MM (2005) *Recipes for freshwater and seawater media* (Elsevier, Amsterdam).
2. Grabherr MG*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29(7):644-652.
3. Haas BJ*, et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 8(8):1494-1512.

4. Zhou X*, et al.* (2015) Dual functions of Macpiwi1 in transposon silencing and stem cell maintenance in the flatworm, Macrostomum lignano. , RNA Published in Advance August 31, 2015, doi:10.1261/rna.052456.115

5. Price AL, Jones NC, & Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21 Suppl 1:i351-358.

6. Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome biology* 11(10):R106.

7. Warren IA*, et al.* (2014) Extensive local gene duplication and functional divergence among paralogs in Atlantic salmon. *Genome biology and evolution* 6(7):1790-1805.

8. Kao D, Felix D, & Aboobaker A (2013) The planarian regeneration transcriptome reveals a shared but temporally shifted regulatory program between opposing head and tail scenarios. *BMC genomics* 14:797.

9. Venken KJ*, et al.* (2009) Versatile P[acman] BAC libraries for transgenesis studies in Drosophila melanogaster. *Nature methods* 6(6):431-434.

10. Wild J & Szybalski W (2004) Copy-control pBAC/oriV vectors for genomic cloning. *Methods in molecular biology* 267:145-154.

11. Gabrielian A & Bolshoy A (1999) Sequence complexity and DNA curvature. *Computers & chemistry* 23(3-4):263-274.

12. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 27(2):573-580.

13. Simakov O*, et al.* (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature* 493(7433):526-531.

14. Dos Santos CO, Dolzhenko E, Hodges E, Smith AD, & Hannon GJ (2015) An epigenetic memory of pregnancy in the mouse mammary gland. *Cell reports* 11(7):1102-1109.

15. Katoh K, Misawa K, Kuma K, & Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30(14):3059-3066.

16. Waterhouse AM, Procter JB, Martin DM, Clamp M, & Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189-1191.

17. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. *. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

18. Guindon S*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* 59(3):307-321.

19. Keane TM, Creevey CJ, Pentony MM, Naughton TJ, & McLnerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol* 6:29.

20. Morris J*, et al.* (2006) The Macrostomum lignano EST database as a molecular resource for studying platyhelminth development and phylogeny. *Development genes and evolution* 216(11):695-707.

21. Zayas RM, Bold TD, & Newmark PA (2005) Spliced-leader trans-splicing in freshwater planarians. *Molecular biology and evolution* 22(10):2048-2054.
22. Nakaya A*, et al.* (2013) KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic acids research* 41(Database issue):D353-357.
23. Kanehisa M (2013) Molecular network analysis of diseases and drugs in KEGG. *Methods in molecular biology* 939:263-275.
24. Egger B*, et al.* (2015) A Transcriptomic-Phylogenomic Analysis of the Evolutionary Relationships of Flatworms. *Current biology : CB*.

Table S1. Sequence of an abundant 150-mer found in the *M. lignano* genome

Repeat sequence	# of bp covered	% of the entire genome	# of contigs containing the repeat
TTTTCGAAACGCCTGTGCATGCGCGAGACTGCTTGGTGTAGTTATTAGAAGGCAACTGCGCCTCTAGCTTAAACCGTGTCTATTTGTCTAAAGAAACTGACTGCGTGCAAAATACAAGTGCACGGAAGCAGAATAACACGCCGTGAGCGTATTTT	11 Mb	1.57	369

Table S2. *Macrostomum lignano* putative SL RNA

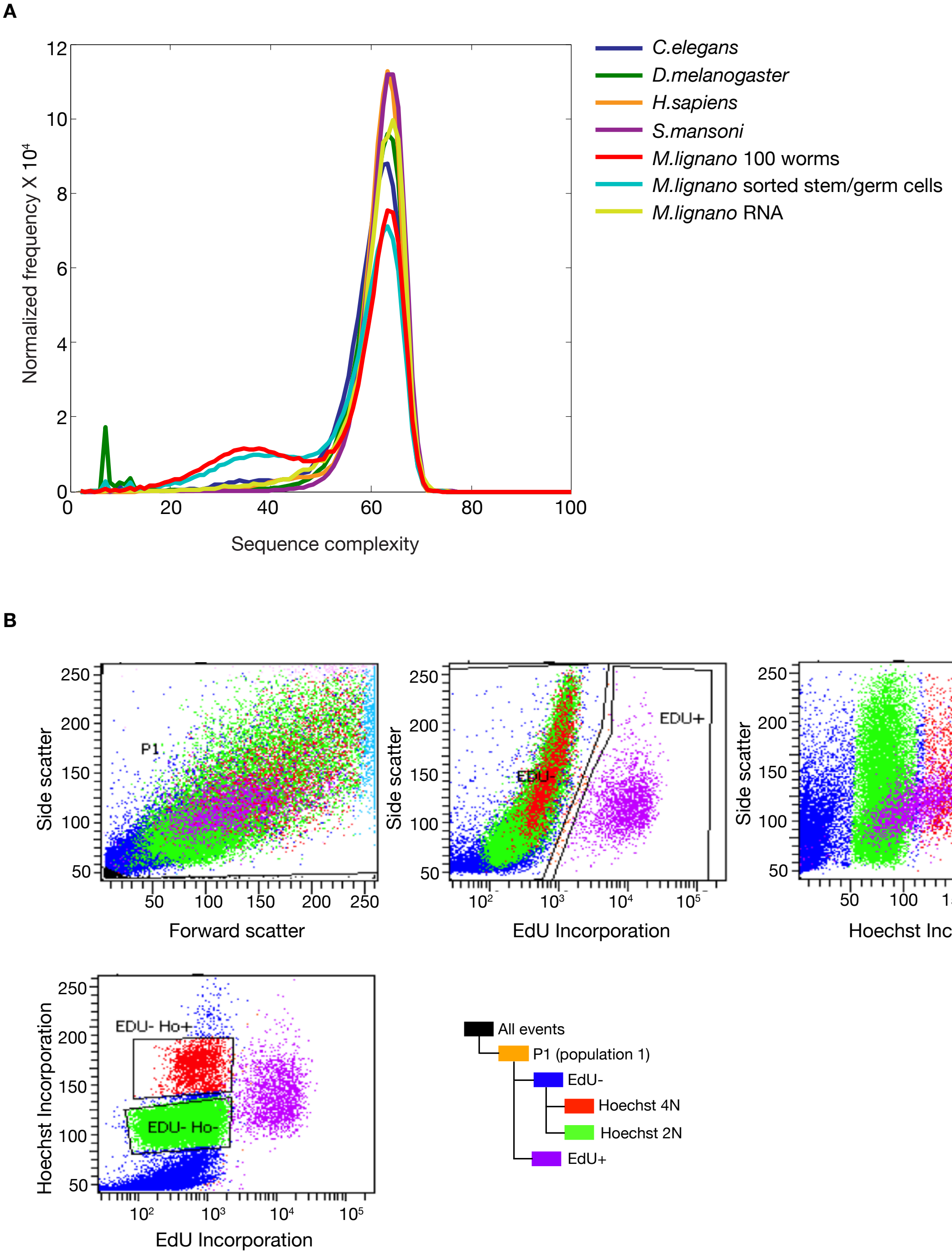
Putative SL RNA sequence	**Sequence length**
5’GCCGTAAAGACGGTCTCTTACTGCGAAGACTCAATTTATTGC**ATG**CTCAGTATCGACCCAGCTTCATCAAATAAAAGAATGCGAATCGAATATACAGCCGAGCCCGACAACTCGGCACTGTCTGCTCCGTTTATTGTGTACTCAGATGCTGATTTGTTGATTTCTAATTCCGATAGTGATAATGTACCCGAATCTGAGAAGCAATTGCTGCCTAATTTGTTGGAACAGGGCTGGCTGGTGAGATATTTTCTAAGTAGCACTTTCTAAGTATGAACCGTT3’	279 nt

Table S3. Homeobox gene localization in *Macrostomum lignano* genome

Gene family	Transcript ID	Scaffold ID
NK1	c105535_g1_i1	uti_cns_0010298,uti_cns_0015914, uti_cns_0006972
Barh	c69451_g2_i2	uti_cns_0000842, uti_cns_0000521, uti_cns_0000322
AbdB	c45137_g1_i1	uti_cns_0017005, uti_cns_0012343, uti_cns_0011893
Hox1 A	c96574_g2_i6	uti_cns_0019113, uti_cns_0002978
Hox1 B	c96574_g2_i3	uti_cns_0005133, uti_cns_0015616
Hox1 C	c96574_g2_i1	uti_cns_0004142, uti_cns_0019113, uti_cns_0002978
Hox3	c88646_g1_i2	unitig_9280, uti_cns_0008583, uti_cns_0005254, uti_cns_0004013
Hox6-8	c72978_g3_i1	uti_cns_0006754, uti_cns_0005919, uti_cns_0001747
Mnx	c46612_g1_i1	uti_cns_0046004, uti_cns_0046039, uti_cns_0000322, uti_cns_0046038
Mox	c72821_g1_i1	uti_cns_0002595
Dlx	c52228_g2_i1	uti_cns_0006526, unitig_44276, uti_cns_0020363, uti_cns_0018478, unitig_26478, uti_cns_0012561, uti_cns_0019380, uti_cns_0005922, unitig_21520, unitig_20458
NK6	c73067_g2_i1	uti_cns_0015575, uti_cns_0005742, uti_cns_0047200, uti_cns_0011175, uti_cns_0000855
Nk2.1 A	c96574_g2_i5	uti_cns_0007608, uti_cns_0004779, uti_cns_0002758, uti_cns_0014120, uti_cns_0047531, unitig_43096, uti_cns_0015353
Nk2.1 B	c96574_g2_i4	uti_cns_0007608, uti_cns_0004779, uti_cns_0002758, uti_cns_0014120, uti_cns_0015353
NK2.2 A	c5173_g1_i1	uti_cns_0015282, unitig_1446, uti_cns_0010045, uti_cns_0005451, uti_cns_0009925
NK2.2 B	c70639_g2_i1	unitig_1446, uti_cns_0000473, uti_cns_0015282, uti_cns_0010045, uti_cns_0009925, uti_cns_0005451
NK2.2 C	c70639_g1_i1	unitig_1446, uti_cns_0000473, uti_cns_0015282, uti_cns_0010045, uti_cns_0009925, uti_cns_0005451
Dbx	c70882_g1_i1	uti_cns_0008573, uti_cns_0015274, uti_cns_0001045
Lbx	c95603_g6_i1	uti_cns_0002945, uti_cns_0002173, uti_cns_0003087
Evx	c67516_g1_i2	uti_cns_0007714, uti_cns_0006606
CdxA	c90244_g2_i1	uti_cns_0045842, uti_cns_0001195, uti_cns_0015309, uti_cns_0014346
CdxB	c98507_g2_i3	uti_cns_0014346, uti_cns_0004813, uti_cns_0012706, uti_cns_0045872, uti_cns_0045842, uti_cns_0001195, uti_cns_0015309
Phox	c6079_g1_i1	uti_cns_0001095, uti_cns_0010675, unitig_43241, uti_cns_0000622, uti_cns_0000695, uti_cns_0002425
Hbn	c28718_g1_i1	uti_cns_0016694, uti_cns_0008099, uti_cns_0003228, uti_cns_0046592, uti_cns_0000500, uti_cns_0001334
Prrx	c77097_g2_i1	unitig_11652, uti_cns_0049053, uti_cns_0048959, unitig_28355
Otx	c69274_g2_i3	uti_cns_0014992, uti_cns_0000361, uti_cns_0000324, uti_cns_0010582, uti_cns_0013964, uti_cns_0011381, uti_cns_0017607, uti_cns_0003136,
Pitx	c91337_g3_i3	uti_cns_0015271, uti_cns_0007372, uti_cns_0006726
Isl	c67586_g2_i2	uti_cns_0000848, uti_cns_0046165, uti_cns_0004103, uti_cns_0000491, uti_cns_0003421, unitig_42300, unitig_19646, uti_cns_0003919, uti_cns_0012499
Zfhx	c110639_g1_i1	uti_cns_0006803, uti_cns_0015968, uti_cns_0011950, unitig_42974

POU4	c119645_g1_i1	uti_cns_0010869, uti_cns_0008027, unitig_25277, unitig_6057, uti_cns_0048108, uti_cns_0007852
POU6	c19464_g1_i1	uti_cns_0008028, uti_cns_0007843, uti_cns_0005053, uti_cns_0046235,
Six3/6	c91070_g2_i1	uti_cns_0017479, unitig_40733, uti_cns_0045711, unitig_29385, uti_cns_0015916, uti_cns_0014557, uti_cns_0013523, uti_cns_0016673, uti_cns_0045710, uti_cns_0000288, uti_cns_0000539
IrxA	c23130_g1_i1	uti_cns_0011437, uti_cns_0000129, uti_cns_0001112, unitig_22191, unitig_30688, unitig_26789, unitig_39949
IrxB	c31631_g1_i1	unitig_22191, uti_cns_0011437, uti_cns_0001112, uti_cns_0000129, unitig_30688, uti_cns_0010873, uti_cns_0003580
IrxC	c25448_g1_i1	uti_cns_0018966, uti_cns_0016154, uti_cns_0016905, uti_cns_0008098
IrxD	c40553_g1_i1	uti_cns_0000129, unitig_22191, uti_cns_0011437, uti_cns_0001112, unitig_30688, uti_cns_0018966, unitig_39949, uti_cns_0016154, uti_cns_0016905, uti_cns_0008098
IrxE	c52568_g1_i1	uti_cns_0047159, uti_cns_0000491
MeisA	c119444_g1_i1	uti_cns_0005222,uti_cns_0003580,
MeisB	c95914_g6_i2	uti_cns_0005222, uti_cns_0003580
MeisC	c95605_g1_i1	uti_cns_0005222, uti_cns_0003580
PknoxA	c21802_g1_i1	uti_cns_0007214, uti_cns_0045779
PknoxB	c87910_g1_i2	uti_cns_0005416, uti_cns_0003530,
Exd/Pbx	c95819_g4_i1	uti_cns_0001820, uti_cns_0046527, uti_cns_0046098, uti_cns_0000370, uti_cns_0004945, uti_cns_0001083
Cux	c98994_g5_i2	uti_cns_0047989, uti_cns_0047990, uti_cns_0048022, uti_cns_0003933
Onecut	c82826_g1_i1	uti_cns_0007903, uti_cns_0007101, uti_cns_0046589
Cers	c84636_g1_i2	uti_cns_0006228
Six4/5	c115999_g1_i1	unitig_31168, uti_cns_0002637, uti_cns_0045438, uti_cns_0003263, unitig_20335, unitig_35453, uti_cns_0013307, uti_cns_0005718, uti_cns_0010333
Pros	c37734_g1_i1	uti_cns_0012883, uti_cns_0010764, uti_cns_0009012,
Pax6	c97812_g1_i1	uti_cns_0015886, uti_cns_0013927, uti_cns_0013738, uti_cns_0004692, uti_cns_0013395

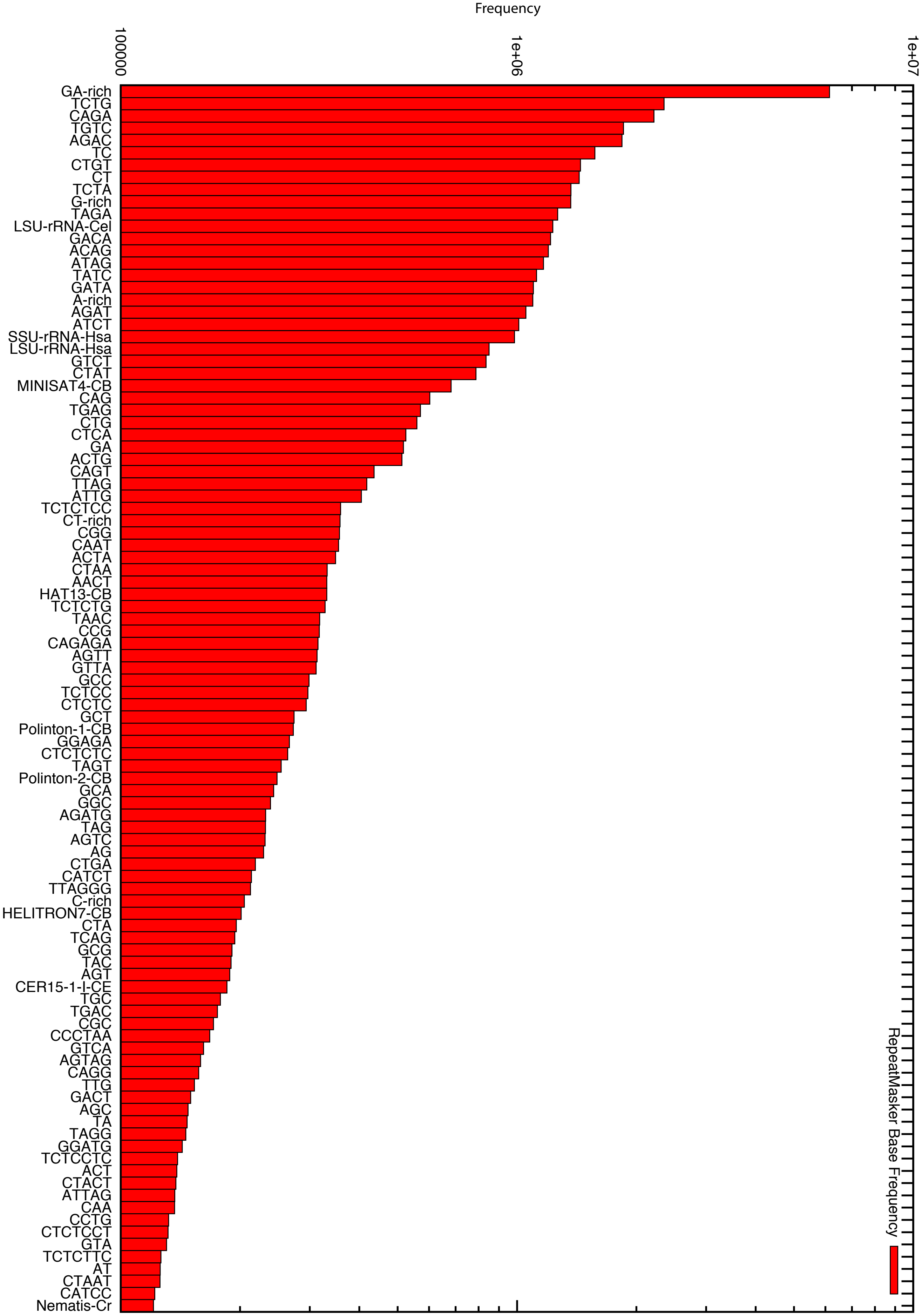
Supplementary Figure 1



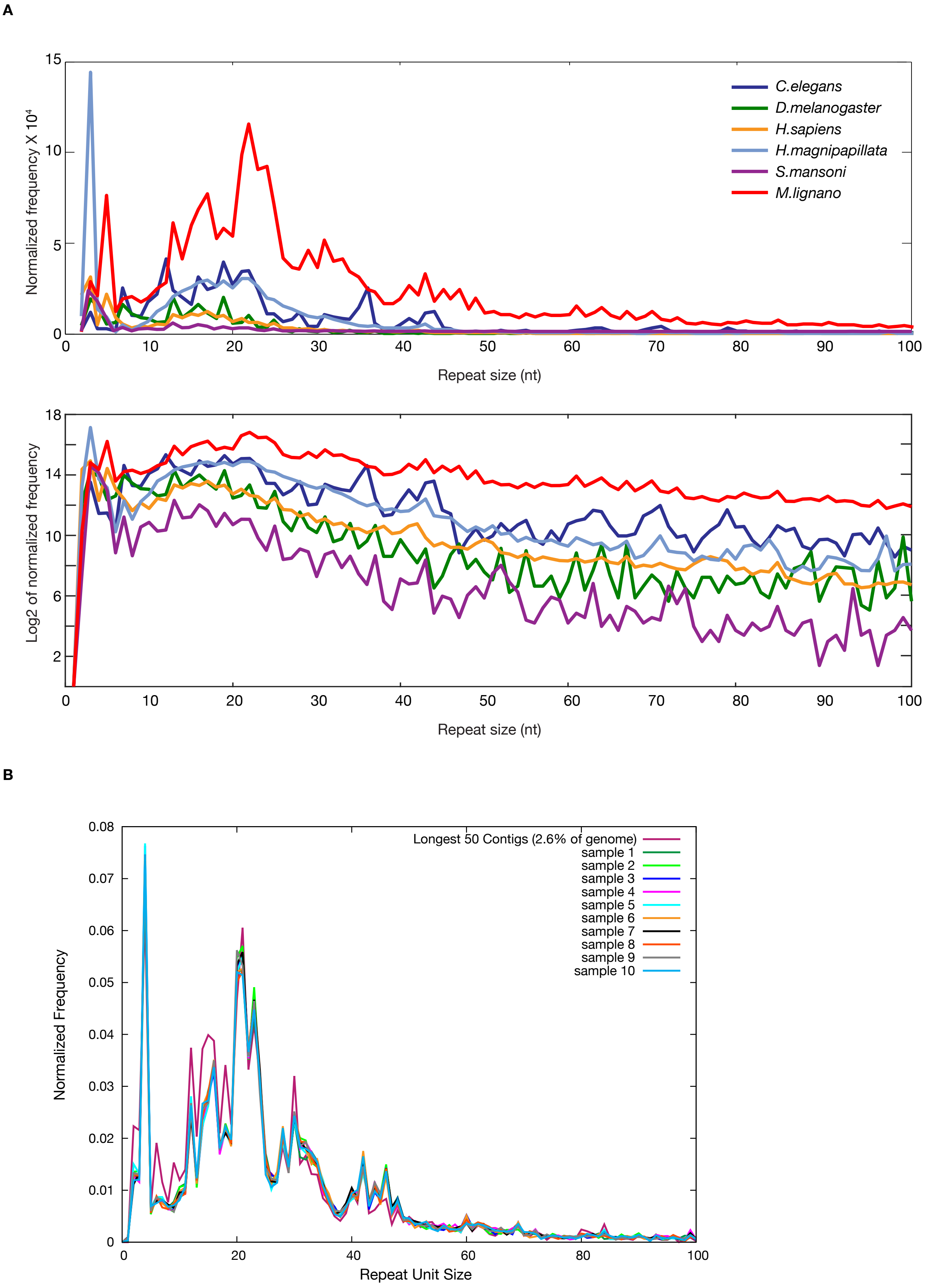
C

Genome	% of bases masked by TRF
M.lignano	24.8
C.elegans	6.8
D.melanogaster	4.7
H.sapiens	2.2
S.mansoni	0.3

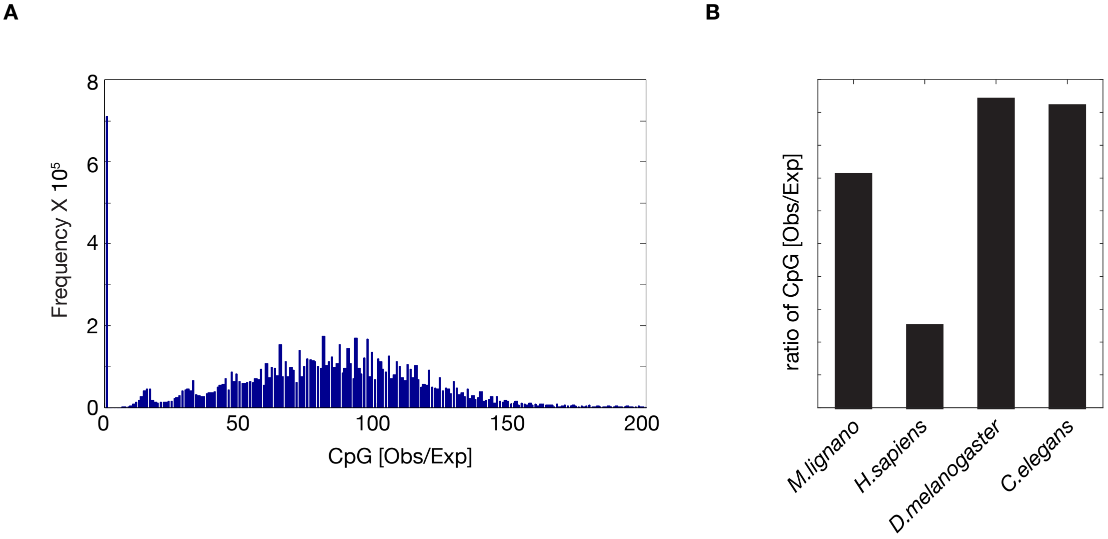
Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4



C

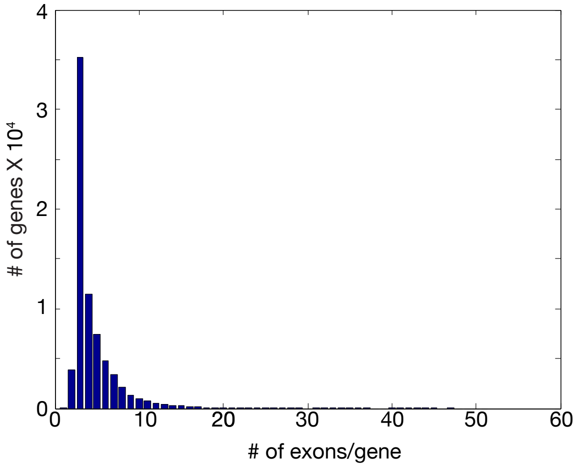
Dinucleotide	Dinucleotide ratio [Obs/Exp]			
	M.lignano	*H.sapiens*	*D.melanogaster*	*C.elegans*
AA	1.06	1.21	1.2	1.14
AC	0.84	0.9	0.87	0.86
AG	1.1	0.96	0.9	0.95
AT	0.95	1.25	0.97	1
CA	1.1	1.3	1.12	1.2
CC	0.99	1.35	1.05	0.9
CG	**0.71**	**0.26**	**0.94**	**0.93**
CT	1.15	1.25	0.9	0.95
GA	1.06	1.1	0.92	1.3
GC	1.13	1.1	1.26	0.95
GG	0.98	1.35	1.06	0.9
GT	0.84	0.9	0.87	0.86
TA	0.79	0.81	0.77	0.5
TC	1.06	1.06	0.92	1.3
TG	1.1	1.3	1.12	1.2
TT	1.06	1.2	1.22	1.14

Supplementary Figure 5

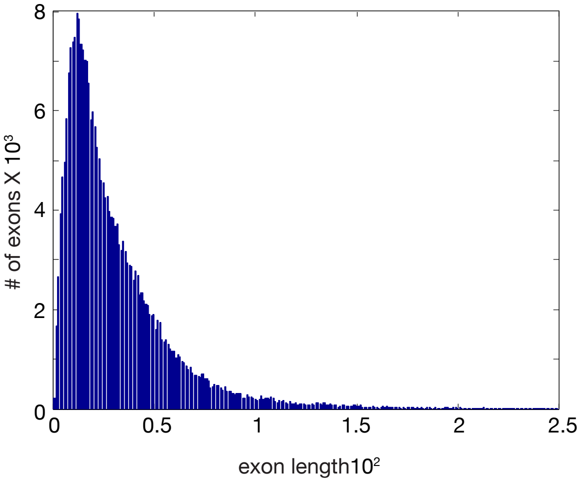
A

Genomic features in *M.lignano*	
# of genes annotated	61,257
# of genes supported by RNA-Seq	19,794
Avg. gene length	5,703bp
longest gene	118,019bp
% of CDS in the genome	13
# of exons annotated	323,393
Avg. exon length	323.17bp
Avg # of exons per gene	5.82
% GC content	46

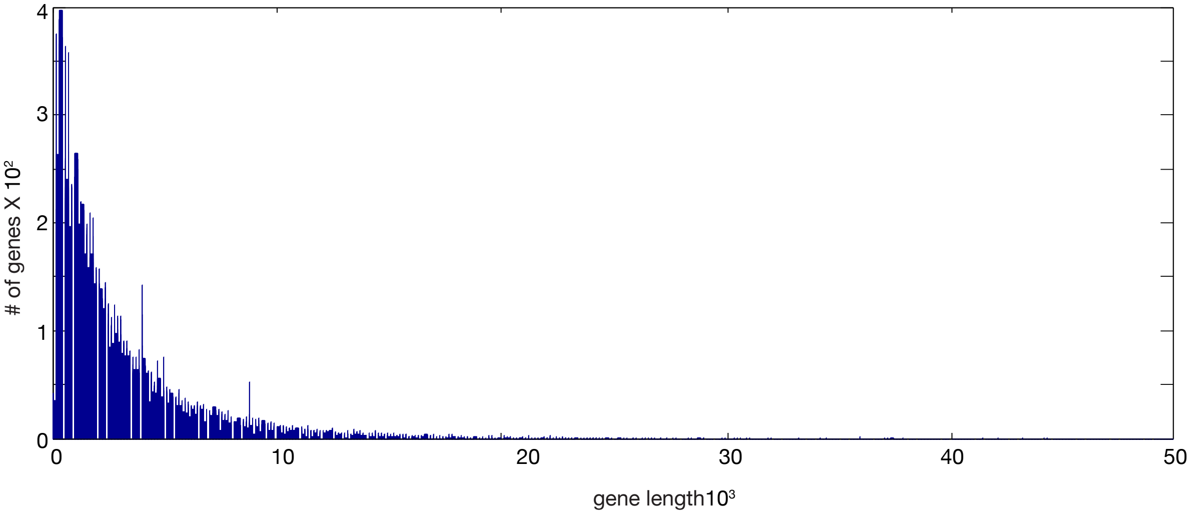
B



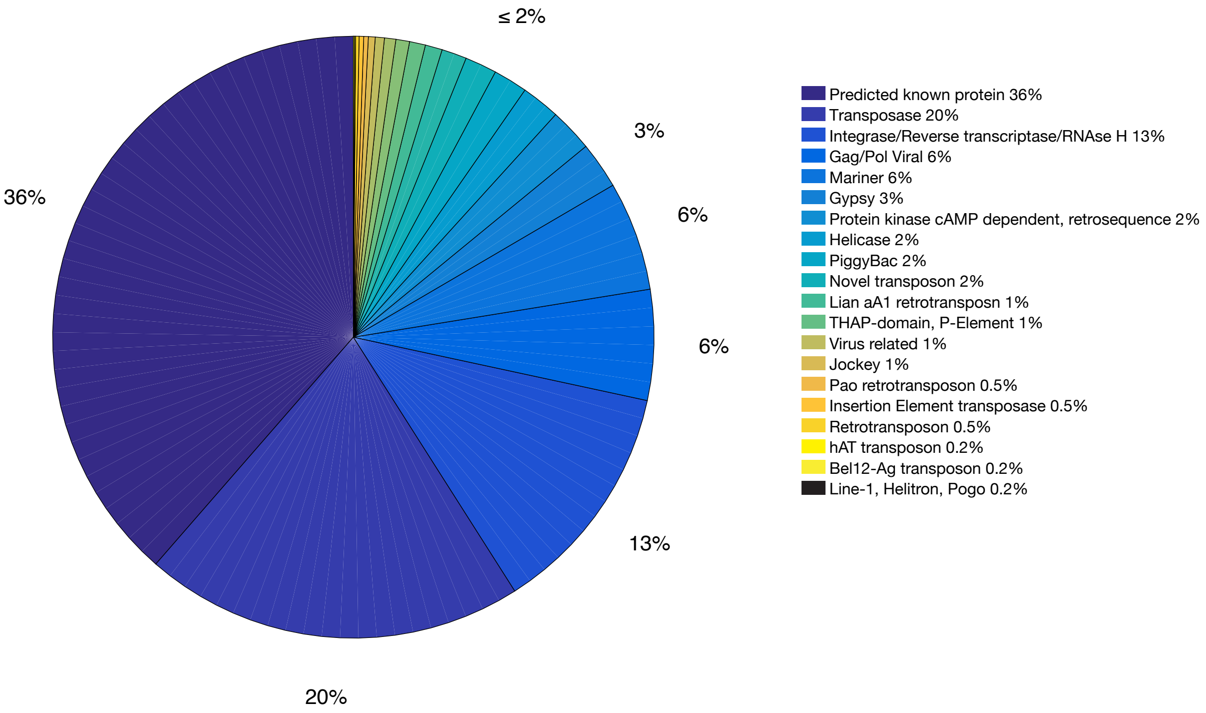
C



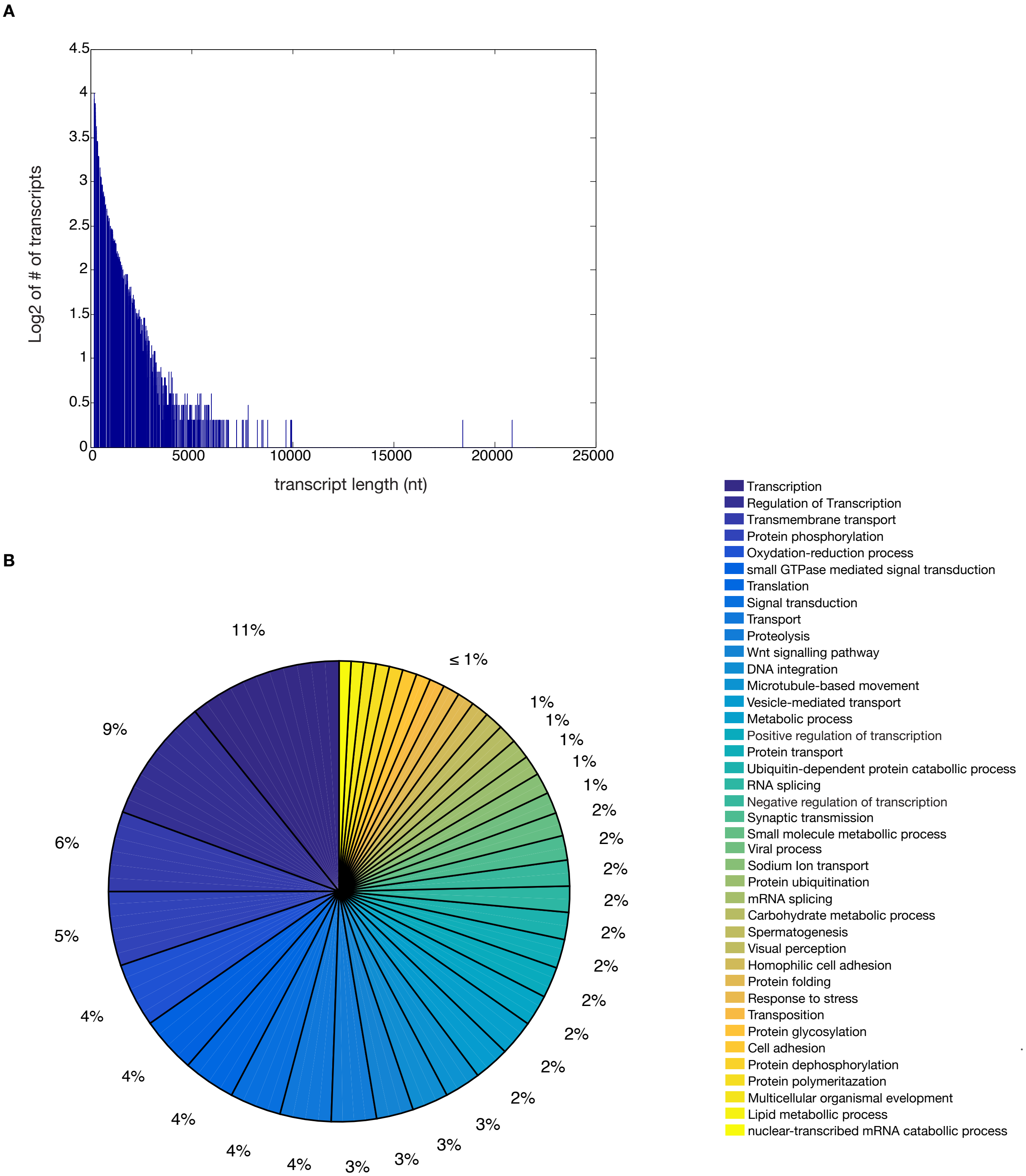
D



Supplementary Figure 6



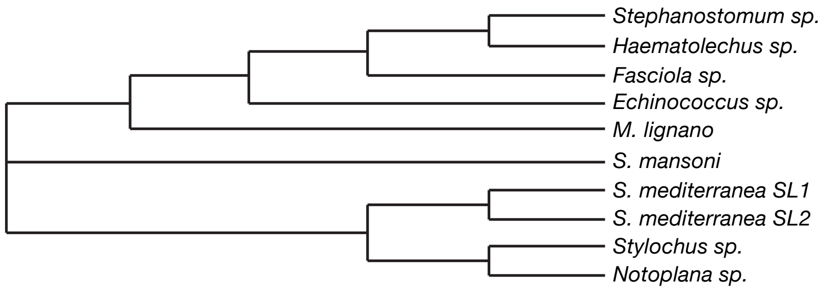
Supplementary Figure 7



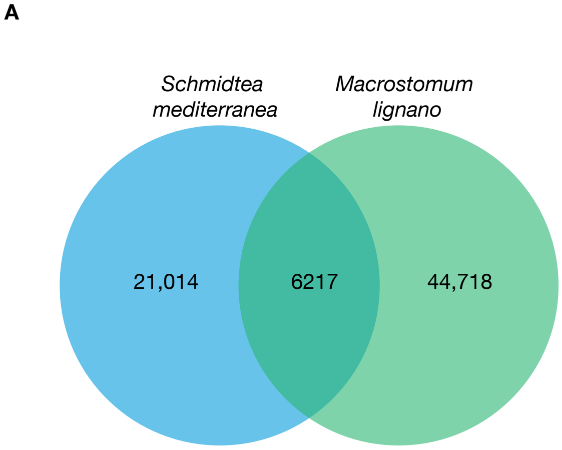
A

```
Stephanostomum sp.    -ACC----TATACGGTT---CTCT-GCCGTGTA------TATTAGT-C-ATGGT-AAGAA
Haematolechus sp.     -ACC----TATACGGTT---CTCT-GCCGTGTA------TCAGTG--C-ATGGT-AAGAA
Fasciola sp.          AACC----TTAACGGTT---CTCTTGCCCTGTA------TATTAGTGC-ATGGTAAAGAA
S.mansoni             AACC----GTCACGGTT---TTAC--TCTTGTG------ATTTGTTGC-ATGGT-AAGAA
Echinococcus sp.      CACCG--TTAATCGGTC---CTTA--CCTTGCA------ATTTTGT---ATGGT-GAGTA
M.lignano             -GCCG--TAAAGACGGT---CTCTTACTGCGAAGACTCAATTTATTGC-ATGCT-CAGTA
S.med SL1             -GCCG--TTAGACGGTC---TTATCGAAATCTATAT---AAATCTTAT-ATGGT-ACGGA
S.med SL2             -GCCG--TTAGACGGTC---TTATCGAAATCTATAT---AAAAATTAT-ATGGT-GAGGA
Stylochus sp.         TGCCGTATTTGACGGTCTCAAAAATTTCGTGTTTATTGCAATAATTGCAATGGT-AAGCA
Notoplana sp.         TGCCGTATTTGACGGTCTCAAAAATTTCGTGTTTATTGCAATAATTGCAATGGT-AAGCA
                       .**     :  . *      :                 : :       *** *  .* *

Stephanostomum sp.    TCGAA-----TTCGAC------CTATGGTCGAATAA-ATTCTTTGGCTAG-CCTCT----
Haematolechus sp.     TCGAG-----TTCGACTCACATCGTTGGTCGAATAAGATTATTTGGCTAG-CCTCCACTC
Fasciola sp.          TCG-------TTGGAC------CATCGGTCCAAACCCATTATTTGGCTAG-CCTCCATTC
S.mansoni             CCG--------TCGAC------CAAGAATCGAAGTT--TTCTTTGGCAGC-CCTAACACA
Echinococcus sp.      TCGATGCAGCTCAGGCTG-TGCCTACGGAGCTGACCCAGTATTTGGCTGGTCCTT-----
M.lignano             TCGACCCAGCTTCATCAAAT-AAAAGAATGCGAATCGAATATACAGCCGAGCCCGACAAC
S.med SL1             CCG--------TTATC------CAACATTAGTTGGTTAATTTTTGACAGTCACTTGAATC
S.med SL2             CCG--------TTTGC------CAGCATTAGTTGGCTAATTTTTGACAGTAGCTTGCAT-
Stylochus sp.         TCAAAT-------GAT------CCAGTGTGATCGTCGAGTCTTTG--ACAGGCCG-----
Notoplana sp.         TCAAA--------GAT------CCA-TGTGATCGTCGAGTCTTTGACACAGGCCG-----
                       *.                   .     :          * *: .       *

Stephanostomum sp.    ---TCGGGGGCTAA------ 96
Haematolechus sp.     TGGTCGGGGGCTA------- 108
Fasciola sp.          TG--CAGAGGCTAAGAATCC 110
S.mansoni             ----CGGGG----------- 91
Echinococcus sp.      ----CGAGGGCC-------- 105
M.lignano             TCGGCACTGTCTGCTCCGC- 130
S.med SL1             --ACAAGTGACTAT------ 107
S.med SL2             --GCAAGTGACTAT------ 106
Stylochus sp.         ----CGAGGCCTATAT---- 111
Notoplana sp.         ----CAAGGCCTATTT---- 111
                          ..  *
```

B



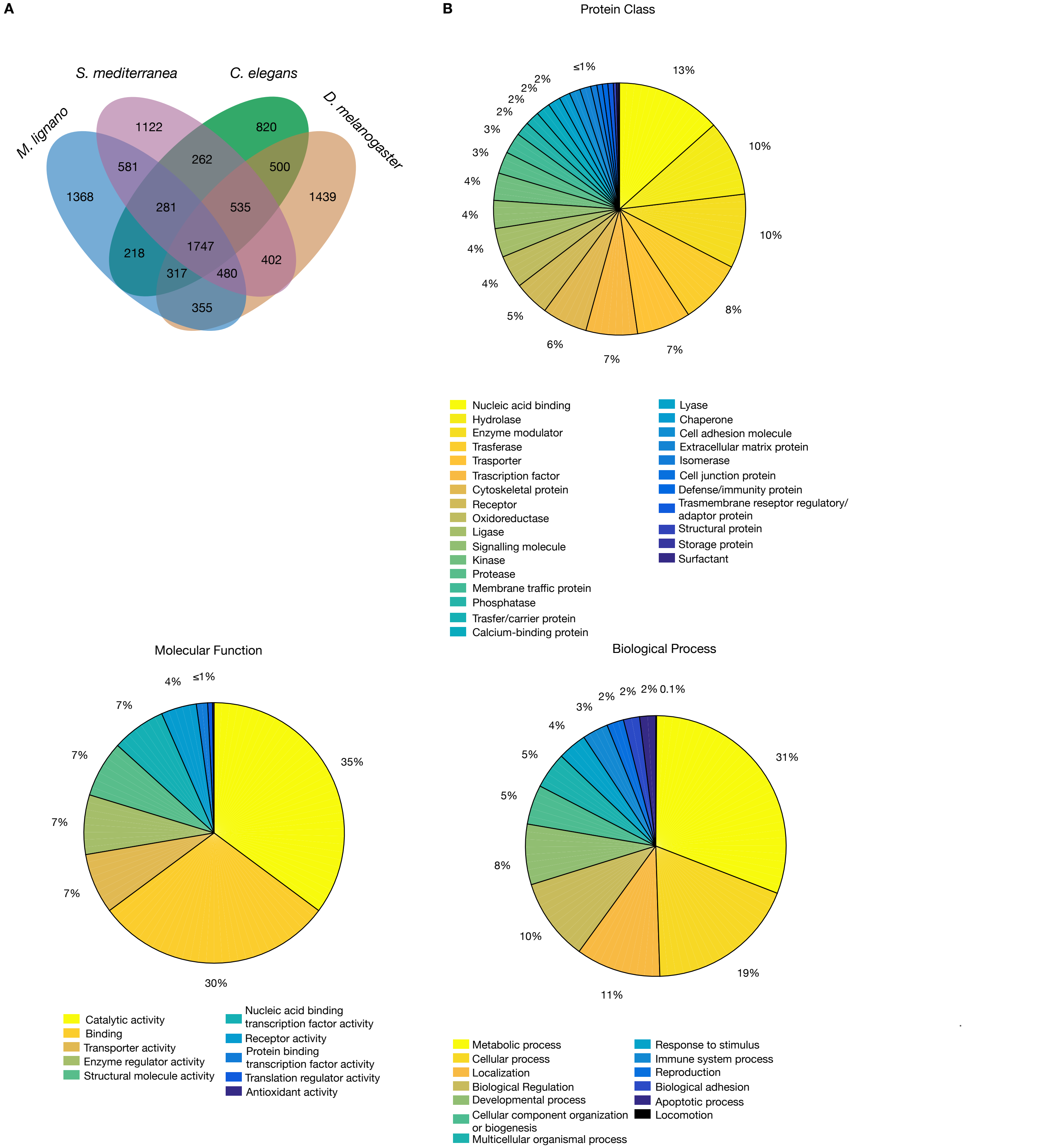
Supplementary Figure 9



B

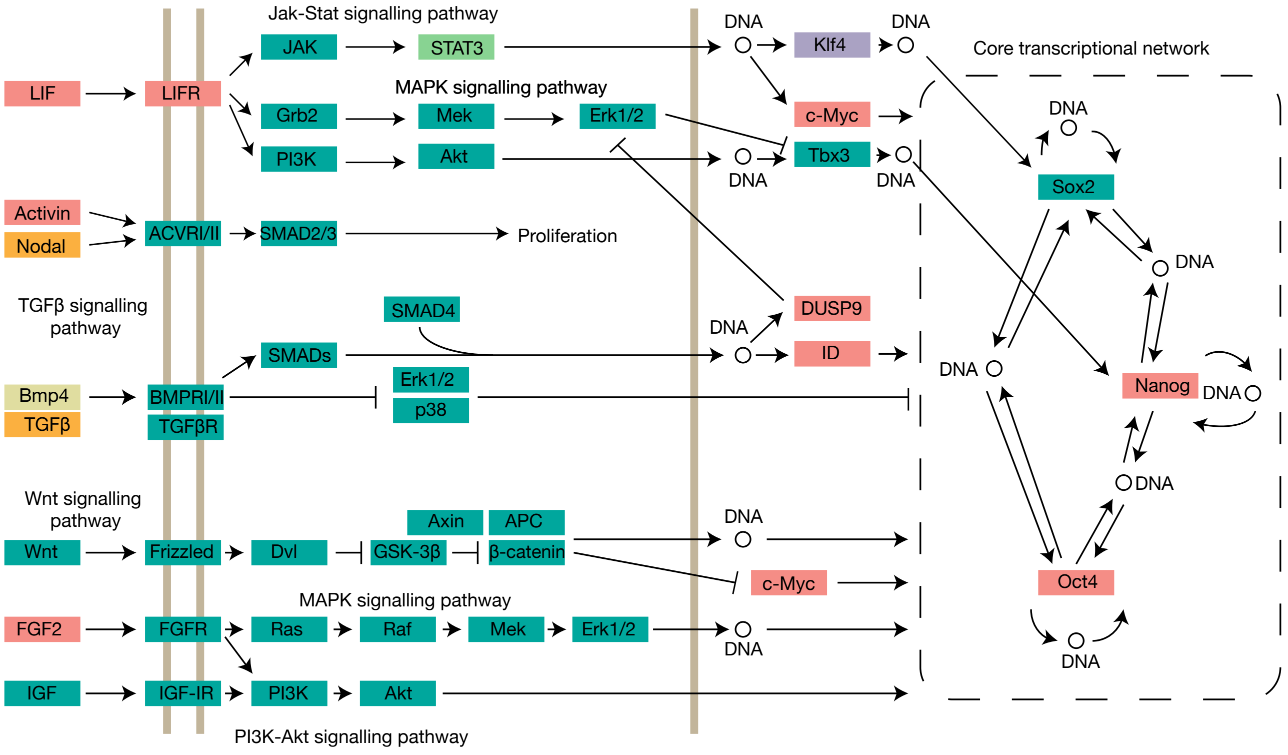
Transcriptome	**Reciprocal BLAST hits against *Homo sapiens***
M. lignano	5347
C. elegans	4680
D. melanogaster	5775
S. mediterranea	5410

Supplementary Figure 10



Suplementary Figure 11

Pluripotency pathways from human/murine stem cells:



Genes from human/murine stem cells:

🟥 Not found in *M.lignano* transcriptome

🟦 Found in *M.lignano* transcriptome

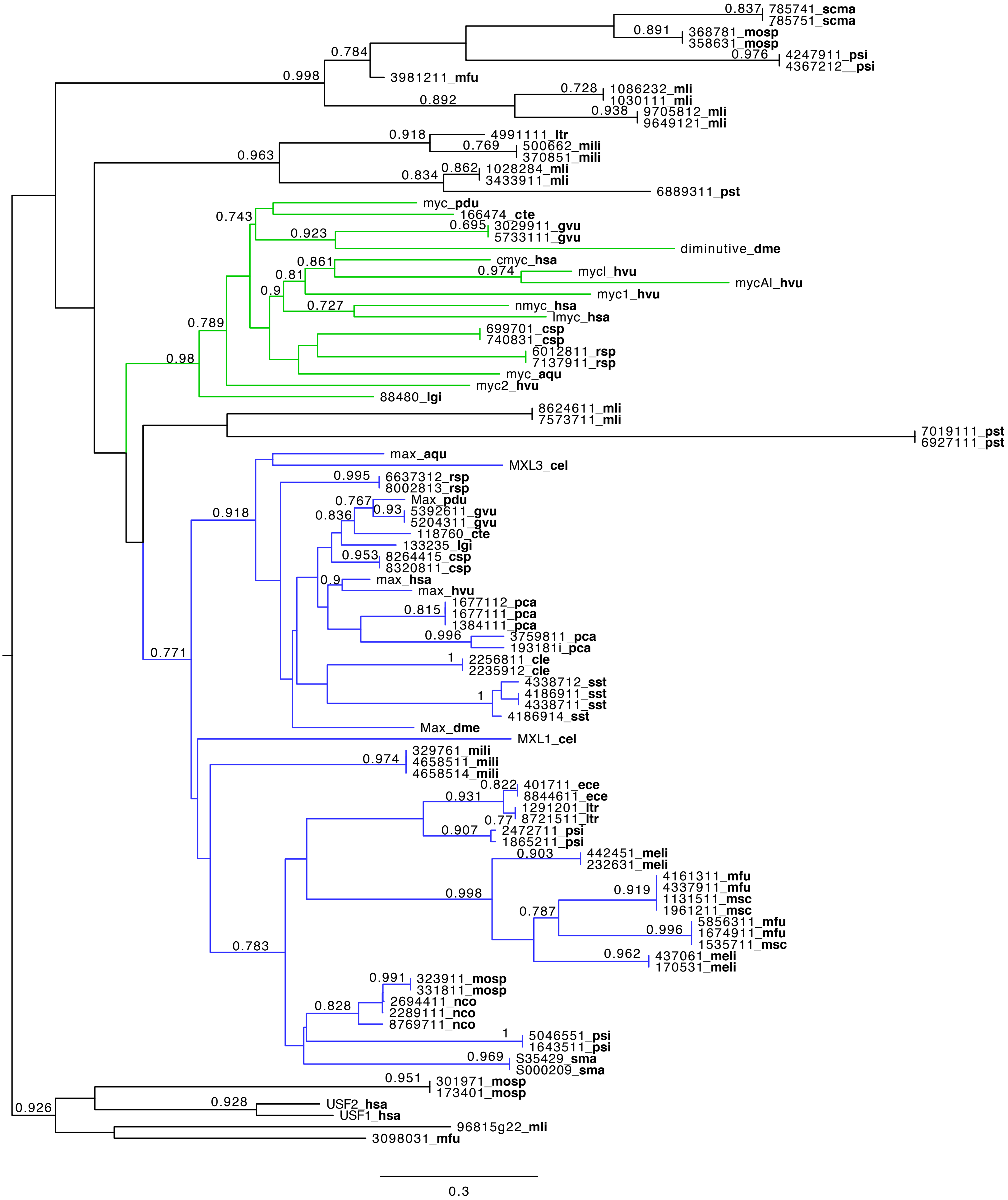
🟧 Factors with TGFβ -like domain found *M.lignano* transcriptome

🟩 STAT3 not found, other STATs identified in *M.lignano* transcriptome

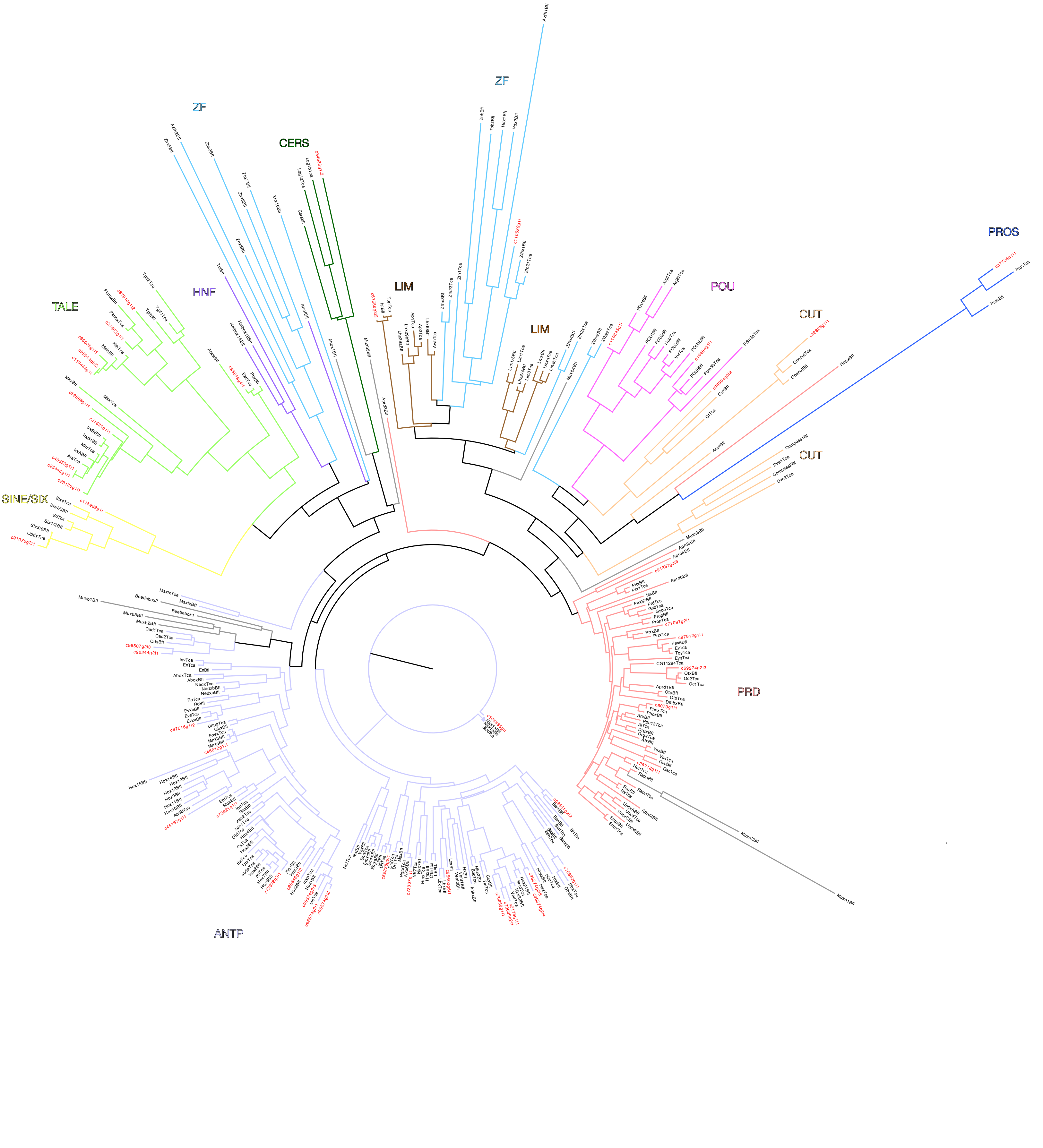
🟪 KLF4 not found, other KLFs identified in *M.lignano* transcriptome

🟨 BMP4 not found, other BMPs identified in *M.lignano* transcriptome

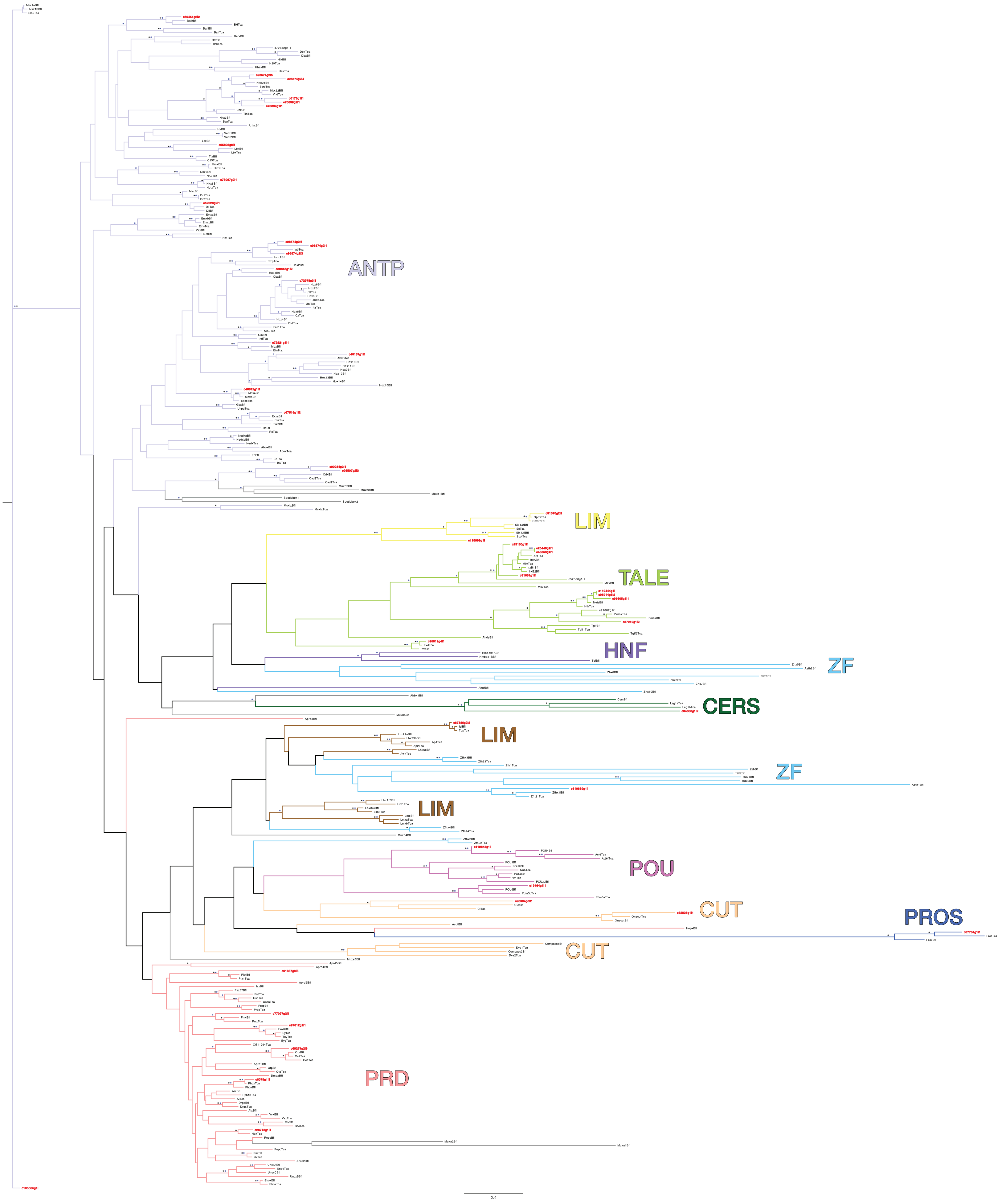
Suplementary Figure 12



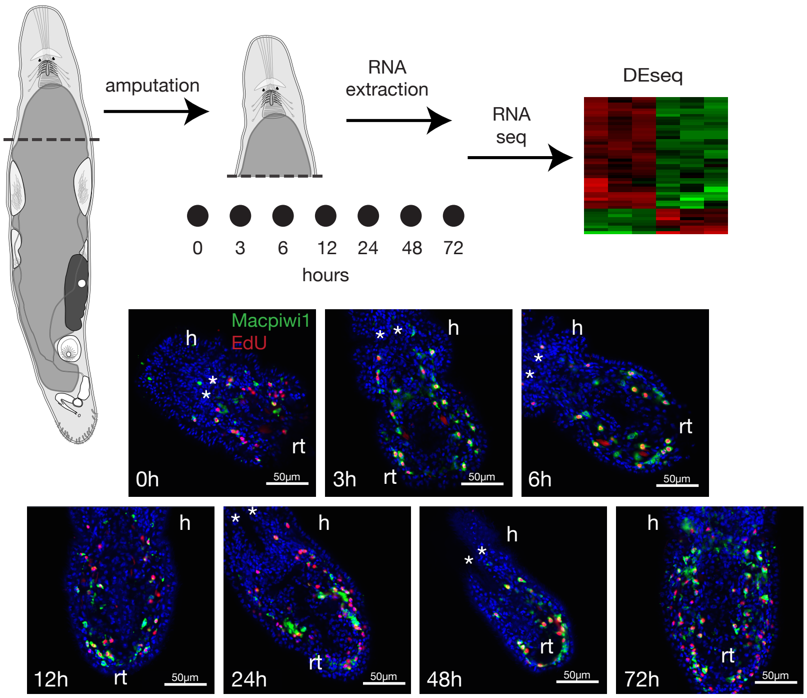
Supplementary Figure 13



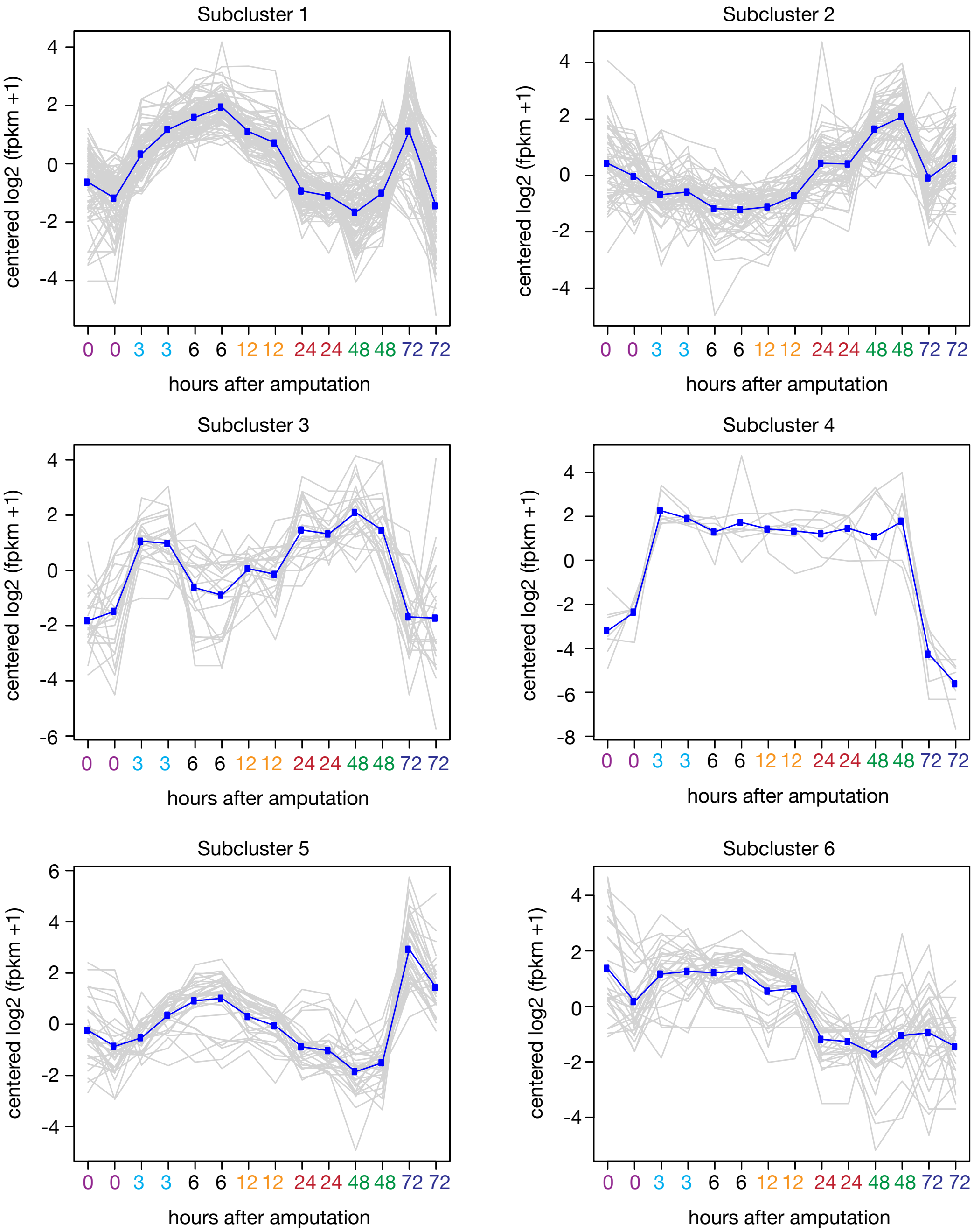
Supplementary Figure 14



Supplementary Figure 15



Supplementary Figure 16


Subcluster 1
Subcluster 2
Subcluster 3
Subcluster 4
Subcluster 5
Subcluster 6
centered log2 (fpkm +1)
hours after amputation