Supplementary Information

Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells

Pinar Önal[1,#], Dominic Grün[1,#], Catherine Adamidi[1,#], Agnieszka Rybak[1], Jordi Solana[1], Guido Mastrobuoni[2], Yongbo Wang[3], Hans-Peter Rahn[4], Wei Chen[3], Stefan Kempa[2], Ulrike Ziebold[1] and Nikolaus Rajewsky[1,*]

[1]Laboratory of Systems Biology of Gene Regulatory Elements,

[2]Integrative Proteomics and Metabolomics Platform,

[3] Laboratory for Novel Sequencing Technology, Functional and Medical Genomics

[4]Flow Cytometry Platform

Max-Delbrück-Center for Molecular Medicine,

Robert-Rőssle-Strasse 10, 13125 Berlin, Germany

[#]These authors contributed equally to this work

***Corresponding author:**

rajewsky@mdc-berlin.de, phone +49 30 9406 2999

Supplementary Information includes:

Supplementary Methods for Experimental Procedures

FACS and Flow Cytometric Analysis

Sample preparation for FACS was adapted from Hayashi et al (2006). Planaria were chopped on ice into small pieces. The fragments were then dissociated at room temperature in Phosphate Buffered Saline (PBS) solution containing 1% w/v Bovine Serum Albumin (BSA) (PBS/BSA) and 0.1 % w/v Trypsin for 10-20 min. After gentle pipetting to dissociate the fragments completely, the cells were washed twice with PBS/BSA The resulting cell suspension was enriched in neoblasts by sequential filtration through a 40 µm cell strainer (Becton-Dickinson) and a 20 µm pore nylon net filter (Millipore). The cell filtrate was incubated 1 hour at RT in PBS/BSA containing Calcein-AM (BD Biosciences, at a final concentration of 0.5 µg/ml) and Hoechst 33342 (Fluka Biochemika, at a final concentration of 20 µg/ml) to differentially stain cytoplasm and nucleus. Finally, Propidium Iodide (PI) at a final concentration of 1 µg/ml was added to detect and eliminate dead cells from the sample.

12 FACS runs with 20 animals each were performed and ~10^6 cells from each fraction were pooled.

For flow cytometric analyses of *BRG1L*, *SMARCC2* and *CTR9(RNAi)* and control experiments, 3 biological replicates (3 dissociated animals each) were used. For *SETD8* and *SSRP1(RNAi)* experiments 9 animals from each group were dissociated and analysed at once.

Shotgun Proteomics

Total proteins were extracted by using TriZol method according to manufacturer's protocol. Extracted proteins from each fraction were loaded onto a monodimensional 10% SDS gel. Each lane was cut into 15 bands and each band was washed with ammonium bicarbonate (50 mM). Proteins were then *in gel* digested as described in Shevchenko *et al*, 2006. Briefly, cystein residues were reduced with DTT and free sulphidrils groups alkylated with iodoacetamide; between each step gel pieces were washed by shrinking and rehydrating with acetonitrile and ammonium bicarbonate 50 mM. After alkylation, trypsin (12.5 ng/ml) was added to gel pieces to perform digestion. Samples were incubated for 16 h at 37°C under gentle shaking. Peptides were extracted the following day and desalted on Stage Tip (Rappsilber *et al*, 2003) and the eluates dried and reconstituted to 25 µl of 0.5% acetic acid in water. 5 microliters were injected in duplicate on a LC-MS/MS system (Agilent 1200 [Agilent] and LTQ-Orbitrap Velos [Thermo]) and eluted using a 155 minutes gradient ranging from 5% to 60% of solvent B (80% acetonitrile, 0.1% formic acid; solvent A= 5% acetonitrile, 0.1% formic acid). For the chromatographic separation a 20 cm long capillary (75 µm inner diameter) packed with 3 µm C18 beads (ReprosilPur C18 AQ, Dr. Maisch) was used. On one end of the capillary nanospray tip was generated using a laser puller (P-2000 Laser Based Micropipette Puller, Sutter Instruments), allowing fretless packing. The

nanospray source was operated with a spray voltage of 1.8 kV and an ion transfer tube temperature of 275°C. Data were acquired in data dependent mode, with one survey MS scan in the Orbitrap mass analyzer (resolution 60000 at m/z 400) followed by up to 20 MS/MS scans in the ion trap on the most intense ions (intensity threshold = 1000 counts). Once selected for fragmentation, ions were excluded from further selection for 40 seconds, in order to increase new sequencing events.

Resulting raw data were analyzed using the MaxQuant proteomics pipeline v1.1.1.14 (Cox and Mann, 2008) and the built in the Andromeda search engine (Cox *et al*, 2011) with an in-house database. Carbamidomethylation of cysteines was chosen as fixed modification, oxidation of methionine and acetylation of N-terminus were chosen as variable modifications. The search engine peptide assignments were filtered at 5% false discovery rate and the feature match between runs was not enabled. The spectral data were searched against an in-house database, which was created by extracting the six longest open reading frames (ORFs) from the 6-frame translations of each transcript. We considered multiple ORFs, since frame shifting errors occur in our genome independent transcript annotation with a low frequency (Adamidi *et al*, 2011).

RNAi

Depending on the efficiency of RNAi, mRNA turnover rate, protein stability and gene function, the phenotype onset varies. Hence, we stopped monitoring each experiment at the point when ~80% of the animals showed homeostatic defects: 10 days after RNAi injection for *SMARCC2(RNAi)*, 18 days for *BRG1L(RNAi)*, 20 days for *CTR9(RNAi)* animals, and 23 days for both *SETD8(RNAi)* and *SSRP1(RNAi)* (Supplementary Figure S4A).

Whole-mount In Situ Hybridization (ISH) and Immunostaining

Planaria starved for one week were treated in 2% HCl/5/8 Holtfreter's and fixed in freshly prepared Carnoy solution (60% ethanol, 30% chloroform, 10% acetic acid). Animals were then bleached with 6% H_2O_2 o/n, rinsed twice with 100% methanol and stored in -20°C if they were not used immediately.

For ISH ,antisense-riboprobes were *in vitro* synthesized from PCR-templates using T7 RNA-polymerase (Promega). Probe synthesis was carried out for 2 h at 37°C using DIG-labeling mix (Roche) and ~500 ng of DNA template. DNA was then degraded with DNAse (Promega) for 15 min at 37°C. Probes were precipitated by adding 40 µg Glycogen (20mg/ml stock solution, Roche) and ice-cold 100% ethanol, and then centrifuged at high speed at 4°C for 40 min. The resulting pellet was dried briefly and dissolved in 100 µl of 1:1 Formamide and 10 mM Tris-HCl (pH 7.5) mix. Riboprobe stocks were stored at -80°C. For use, riboprobes were diluted up to 0.05-0.02 ng/µl in hybridization solution (50% formamide, 5xSSC, 0.1 mg/ml yeast tRNA, 0.1 mg/ml yeast tRNA, 0.1% Tween, 10 mM DTT, 10% Dextran Sulfate). Planaria were rehydrated and treated with Proteinase K, postfixed for 1h and washed 3 times with PBST (1X PBS with 0.1% Tween), treated with TEA (0.1M) + acetic anhydride and washed 4 times with PBS. Specimens were prehybridized in prehybridization solution (50% formamide, 5xSSC, 0.1 mg/ml yeast tRNA, 0.1 mg/ml yeast tRNA, 0.1% Tween, 10mM DTT) for 1h, at 56°C, followed by hybridization in riboprobe mix (prehybridization buffer + 0.05-0.02 ng/µl riboprobes) o/n, 56°C. Planaria were then washed as followed:

100% Hybe solution (50% Formamide, 5xSSC), 10 min at 56°C;

75% Hybe solution + 25% (2xSSC, 0.1% TritonX100), 10 min at 56°C;

50% Hybe solution + 50% (2xSSC, 0.1% TritonX100), 10 min at 56°C;

25% Hybe solution + 75% (2xSSC, 0.1% TritonX100), 10 min at 56°C;

2xSSC, 0.1% TritonX100, 2x30min at 56°C;

0.2xSSC, 0.1% TritonX100, 2x30min at 56°C.

Washing solution was replaced with Buffer I [0.1 TritonX100 + 1xMAB (0.1M Maleic Acid+ 0.2M NaCl pH 7.5)], followed by blocking for 1h in buffer II (Buffer I+1xBoehringer Blocking solution for nucleic acids), and with anti-DIG antibody solution 3h, RT. Specimens were washed twice 10 min, RT and then o/n in Buffer I, followed by washing with AP buffer and developed using nitro blue tetrazolium/5-bromo-4-chloro-3-indolyl phosphate colorimetric assay.

Planarians fixed in Carnoy's solution and bleached in 6% H_2O_2 o/n were rehydrated by serial methanol washes and then immunostained as described in Cebrià and Newmark 2005. Mitotic nuclei were labeled with anti-phosphohistone H3 (Ser10) antibody (αH3P) (Cell Signaling, 1:500) To-pro3 Iodide (Invitrogen, 1:500) was used to label DNA.

Imaging

Brightfield pictures of were taken on a Zeiss SteREO Discovery V12 from CarlZeiss using an AxioCam MRC from CarlZeiss. Confocal laser scanning microscopy was performed with a LeicaSP2 confocal laser scanning microscope (CLSM) (Leica Lasertechnik, Heidelberg).

Supplementary Methods for Computational Analyses

Transcript and Protein Annotation

The gene models used in this study were produced by a genome independent *de novo* transcriptome assembly (Adamidi *et al*, 2011), we refer to these predictions as BIMSB transcripts, and supplemented with genome based transcript predictions annotated by using CUFFLINKS (Trapnell *et al*, 2010). More precisely, we mapped all BIMSB transcripts to the genomic contigs available for *S. mediterranea* (Robb *et al*, 2008) as outlined in Adamidi *et al* (2011). We then combined Solexa reads sequenced from whole animals used in Adamidi *et al* (2011) with Solexa reads sequenced from the three cell fractions in this study and ran TopHat (Trapnell *et al*, 2009) to map these reads to the genome. These mappings were passed over to CUFFLINKS in order to infer transcript predictions. Since CUFFLINKS predicts a large number of short and lowly expressed isoforms we ensured confidence of the new predictions by applying the following filtering criteria: the 5%-quantile of read coverage and transcript length was determined for all BIMSB transcripts recovered by CUFFLINKS and novel CUFFLINKS predictions with coverage or length lower than the respective threshold were discarded. Finally, we merged the filtered CUFFLINKS predictions with the BIMSB transcript and discarded redundant predictions by eliminating transcripts which differed by less than 35bp from another longer transcript. The minimum number of 35 unique nucleotides corresponds to the 5%-quantile of the BIMSB transcript exon length distribution. In summary, 17,546 of the 18,619 BIMSB transcripts could be aligned to the genome. After merging with novel CUFFLINKS predictions and redundancy filtering, we retained 10,701 BIMSB transcripts and 15,057 CUFFLINKS predictions. Transcripts with overlapping genomic coordinates were merged to gene loci. In total, we identified 24,239 distinct gene loci.

Quantification of Transcript Expression

The paired-end 120bp reads of all samples were mapped to the transcriptome sequences using the read alignment software bwa (Li *et al*, 2009). Prior to read mapping, we removed consecutive strings of basecalls with lowest Phred quality score from the 3' end of the reads and maintained only those with a minimum remaining length of 30 bases after trimming. We ran bwa (Li *et al*, 2009) with a minimum seed length of 30 and default parameters otherwise. We counted the number of reads mapping to each gene by summing up the read counts of all isoforms. Reads, mapping to multiple loci with equal quality were assigned to all of these loci with reduced weight given by one over the number of loci. A pseudocount of one was added if no reads were mapped to a locus. In order to quantify transcript expression for all gene loci, we calculated reads per kilobase of transcript sequence per million mapped reads (RPKM) (Pepke *et al*, 2009), dividing by the total amount of kilobases of all transcripts originating from a gene locus. The uncertainty of transcript quantification was estimated for each transcript i by the theoretical sampling error, which corresponds to the standard deviation σ_i of a binomial distribution $f(n_i;N,p_i)$ with N given by to the total number of reads and n_i given by the number of reads mapping to transcript i. Based on the estimated sampling error we defined for each transcript i the confidence $c_i=1-\sigma_i/n_i$ a number between zero and one reflecting the confidence of transcript quantification. Values close to one indicate high confidence. If only confident transcript quantifications are used, a filtering cut-off of $c_i=0.8$ is applied.

Quantification of Protein Expression

Aminoacid sequences for quantification of protein expression were produced by translating a given transcript sequence in all three open reading frames (ORFs) on the genomic strand of the gene. We retained the three longest ORFs for each transcript, since we are aware of

possible frame shifting errors in our genome independent transcript annotation (Adamidi *et al*, 2011). To estimate expression of a given protein, intensities for all peptides belonging to this protein were aggregated. Only peptides with intensities measured in three unique experiments were retained and protein expression was computed by summing up the intensities of all peptides. Finally all proteins with less than three identified unique peptides were discarded.

Clustering

To obtain normalized RPKM values for each gene locus such that expression values of all cell fractions sum up to one, RPKM values were divided by the sum of RPKM values over all cell fractions. For each gene locus a vector was defined with components corresponding to normalized expression in the X1, X2 and Xins cell fractions, representing the normalized expression profile of the gene. For each pair of genes i and j, Pearson's correlation coefficient ρ_{ij} for the normalized expression profiles was computed and the ensemble of all genes was subjected to hierarchical clustering with a gene-to-gene distance measured by 1- ρ_{ij}. In order to identify the level of the cluster hierarchy with distinct clusters covering all clearly distinct observable expression profiles, the cluster hierarchy was screened by eye.

GO Analysis

GO term analysis was performed using the DAVID functional annotation tool (Huang *et al*, 2009). Human homologs were identified for each cluster by BLAST protein alignments ($E < 10^{-10}$) and enriched human GO terms were inferred comparing to GO term annotation of the ensemble of human homologs for all clusters. Only GO terms with multiple testing corrected (Benjamini) $P < 0.05$ are reported in Supplementary Table S4.

Screen for Planarian Homologs of NANOG, OCT4 and SOX2

To search for homologs of OCT4, NANOG and SOX2 with high sensitivity we employed a method based on profile Hidden Markov Models (HMMs) implemented in the HMMER 3.0 software suite (Eddy, 2009). We first produced multiple alignments of the human protein sequence to proteins of all metazoan organisms using the PSI-BLAST algorithm (Altschul *et al*, 1997) provided as an online toolkit (Biegert *et al*, 2006), which was run with default parameters. A profile HMM was computed for this multiple alignment using the HMMER3 routine hmmbuild with default parameters. The two longest open reading frames of each planarian transcript were then searched against the profile HMM for each of the three human genes and top scoring candidates were subsequently further analyzed manually.

Searching for NANOG homologs revealed a number of significant hits containing homeobox domains. We analysed the four top-scoring candidates in more detail (Supplementary Table S8A). The highest ranking gene (CUFF.32446.1) was clearly a homolog of the known planarian homeobox protein MSX-1 (Mannini *et al*, 2008). The second ranking putative NANOG homolog (isotig05838) is most similar to a homolog of human Nkx2.2, called DTH-1 in the planaria species *Dugesia tigrina* (Garcia-Fernàndez *et al*, 1993), and which belongs to the NK-like (NKL) subclass of Homeobox proteins. Both of these two top scoring homologs, however, are not up-regulated in the neoblast enriched fraction and are consequently found in clusters 5 and 6. Ranks three and four of our homology search were occupied by two BARH-like proteins, which also belong to the NKL subclass of homeobox proteins (CUFF.114724 and isotig24657), but fall into cluster 1 and thus display an expression pattern similar to known neoblast markers. Since the mammalian NANOG protein shows little defining features apart from the homeodomain, we aligned the *Homo sapiens* homeodomain (aa 98-154) to the *S. mediterranea* genome/proteome using BLAST. The most similar proteins were the above

mentioned NK-like planarian genes, arguing that there is no convincing planarian NANOG homolog found in the current gene annotation.

Our OCT4 homology search yielded six POU-domain encoding transcripts containing both the POU specific domain (POUs) and POU homeodomain (POUh) (Supplementary Table S8B). Out of these six *bona fide* POU-like genes, three were found to be highly expressed in neoblast enriched fractions X1 and X2 compared to the differentiated cell fraction Xins (isotig26184, isotig23172, and isotig21311). Isotig21311 encodes a previously described Smed-POU2/3 factor, which is required for nephridia differentiation (Scimone *et al*, 2011) and shows elevated expression in X2. Isotig21311 is therefore a candidate for a gene expressed in neoblast progenitors. Only one of the six transcripts (isotig26184), which we named *Smed-POU-P1*, showed an expression profile characteristic of neoblast enriched genes (Figure 9A). *Smed-POU-P1* thus represents our best candidate for a POU5-like factor expressed in planarian stem cells. However, although the *Smed-POU-P1* protein sequence shows high conservation of the OCT4 like POUs and POUh domains ($E < 8.7 \cdot 10^{-20}$), the presence of an unusual long linker (~50 nucleotide) between the POUs and POUh domains renders the sequence hard to classify within POU family subclasses. To refine our homology search, we aligned POU5F1/OCT4 known orthologs from divergent phyla with the planarian sequence using the MUSCLE (3.7) software (Edgar, 2004). Along with OCT4 sequences from vertebrate species, we included in our alignment a POU3/5 homolog expressed in cnidarian stem cells (Millane *et al*, 2011). Our results show extensive conservation within the POUs and the POUh domain, but also within a subsequence of the linker region (Figure 9B, Supplementary Figure S7C). In summary, the *Smed-POU-P1* expression profile and its sequence conservation suggest it to be a POU factor ortholog specifically up-regulated in planarian stem cells. Further studies will be needed to conclude on its biological function.

From the list of candidate genes with significant HMMER alignment to human SOX2, we analyzed the seven top-ranking candidates (Supplementary Table S8C). This list includes the recently annotated Sox-family homologs *Smed-soxP-1* (isotig23820), *Smed-soxP-2,* (isotig25189), and *Smed-soxP-3* (isotig17583) (Wagner *et al*, 2012). Except for isotig22872, all of the candidates have enriched expression in X1 and/or X2 compared to Xins. We performed a phylogenetic analyses of the seven genes together with mouse and human Sox2 orthologs using the PhyML tool from Phylogeny.fr (Dereeper *et al*, 2008) in order to determine, which candidate is most highly related to mammalian Sox2 (Supplementary Figure S8A). We found that among all seven genes CUFF.244913.1 is the closest homolog of human SOX2 based on sequence similarity (Supplementary Figure S8B). Moreover, among all candidates the expression profile of CUFF.244913.1 is most similar to a cluster 1 profile (Figure 9C) and it is the only candidate yielding a reciprocal best hit with human SOX2 in a BLAST search. Since SOX2 belongs to Sox class B transcription factors and genes with this class display a high sequence similarity, we annotated this gene as *Smed-SOXB-1*. We finally also validated enhanced neoblast expression of *Smed-SOXB-1* by ISH (Figure 2D). In contrast, the candidate CUFF.51041.1, which displayed best homology to another human Sox class B factor (SOX14) and was therefore annotated as *Smed-SOXB-2,* was only very lowly expressed in each of the cell fractions.

Supplementary Discussion

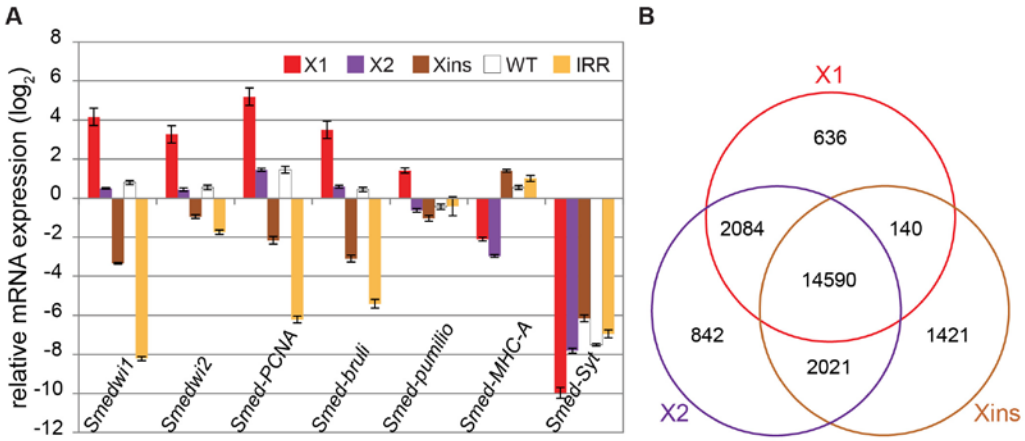
Plasticity of gene regulatory networks during development

We presented in the discussion text one important example for a replacement of a transcription factor with central function during development without affecting the downstream regulatory network controlled by this factor (Schröder, 2003). Another example

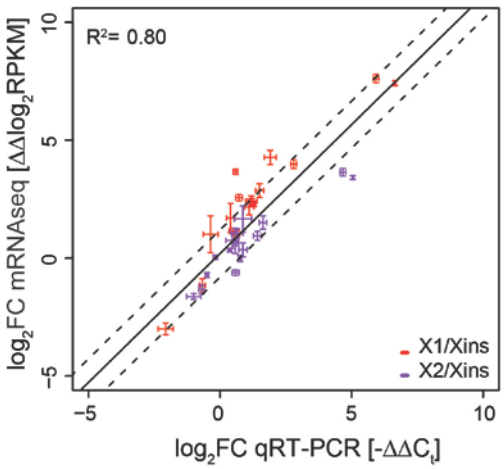
of this kind is the regulator of the transcription factor *otx*, which is crucial for establishing the regulatory state of the endomesoderm in two divergent echinoderms, the sea urchin *Strongylocentrotus purpuratus* and the sea star *Asterina miniata* (Hinman *et al*, 2007). The transcription factor *tbrain* is co-expressed with and required for expression of *otx* in the sea star and this regulatory role has been lost in the sea urchin while the downstream core regulatory network is conserved (Hinman *et al*, 2007). Gene networks of pluripotency control could have undergone similar changes during evolution, which, for instance, could have led to the inclusion of Nanog or other pluripotency regulators in the vertebrate lineage without changing the expression of downstream genes.

Supplementary Figures

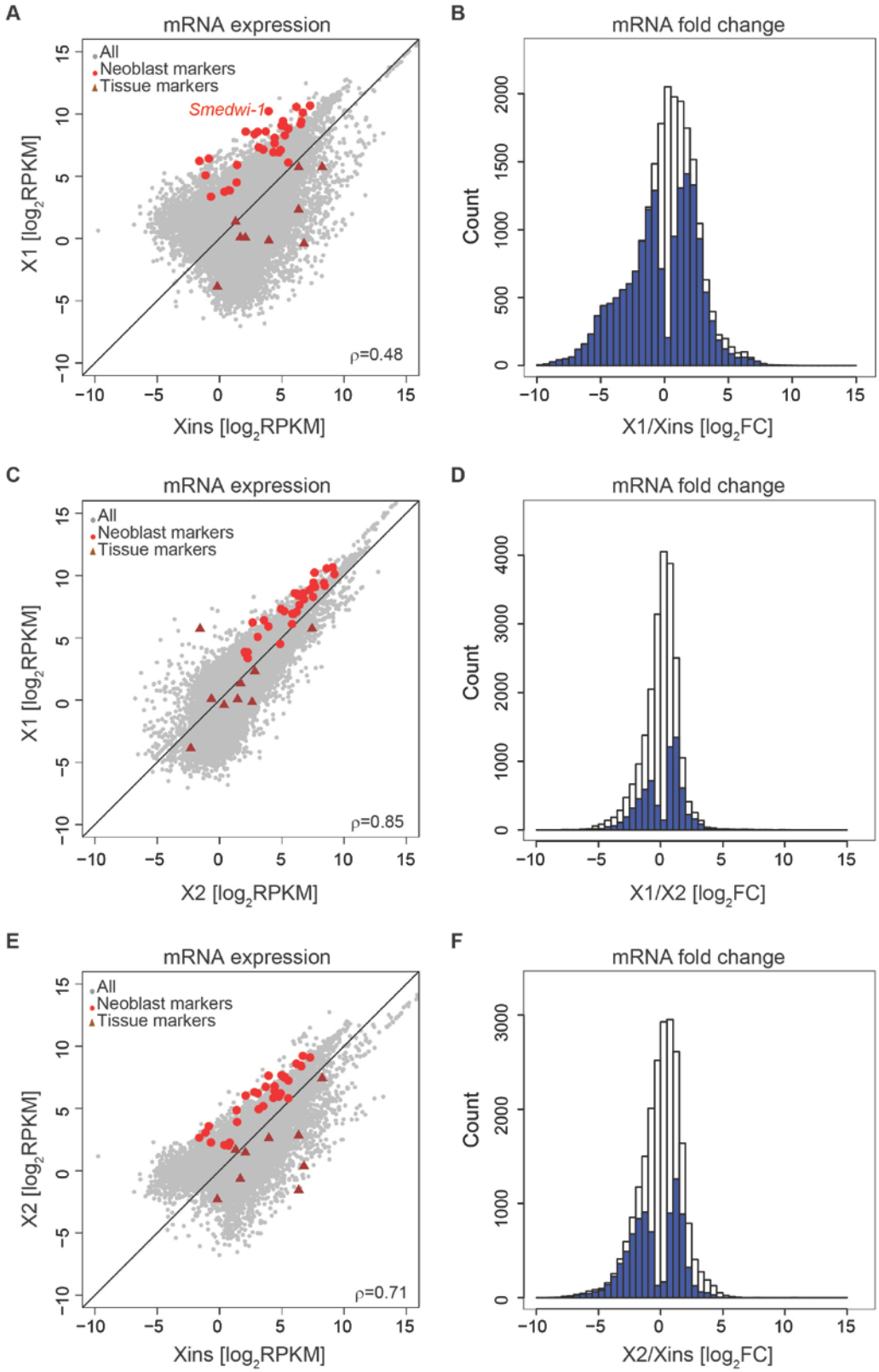
Supplementary Figure S1



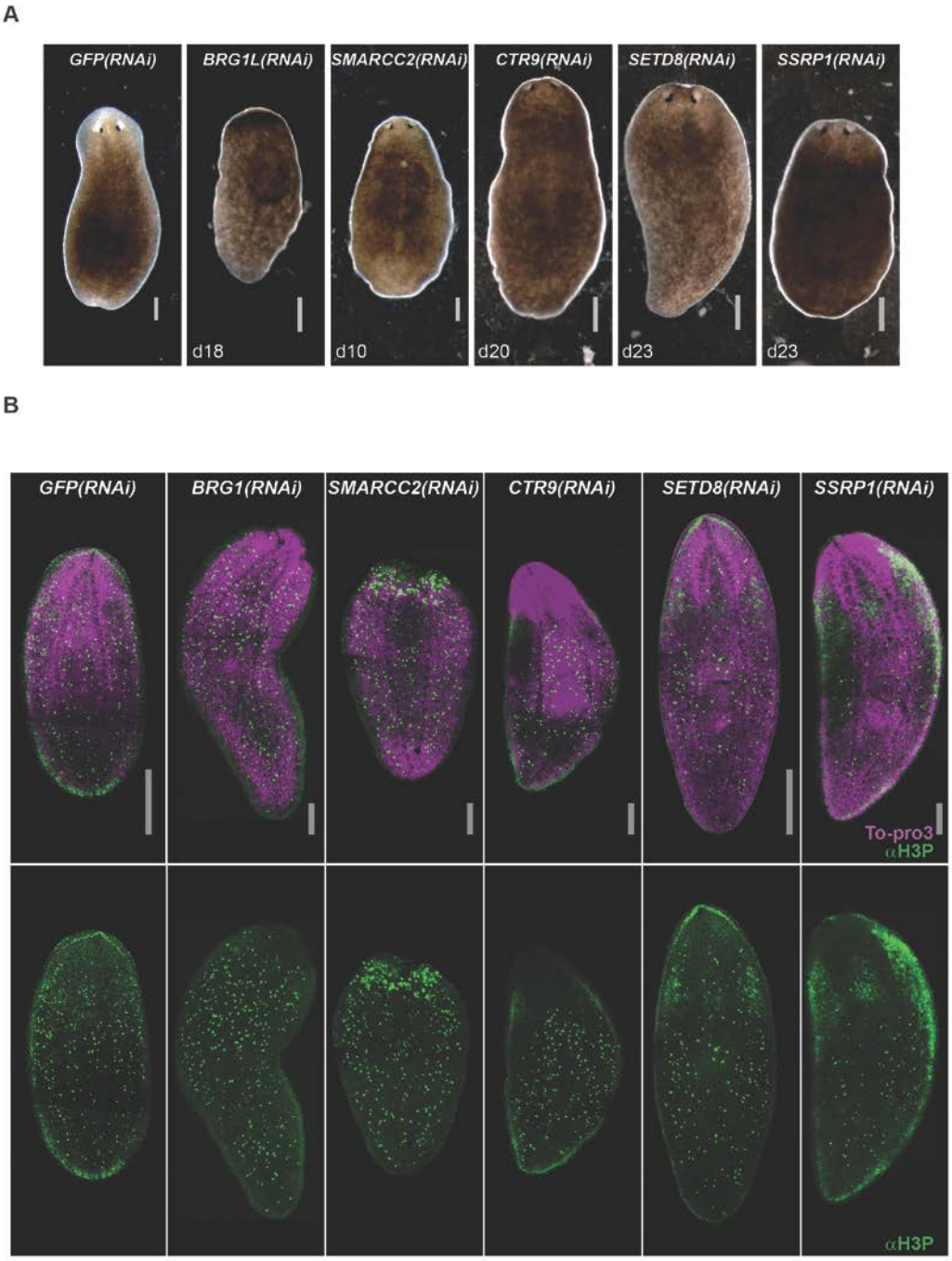
Supplementary Figure S2



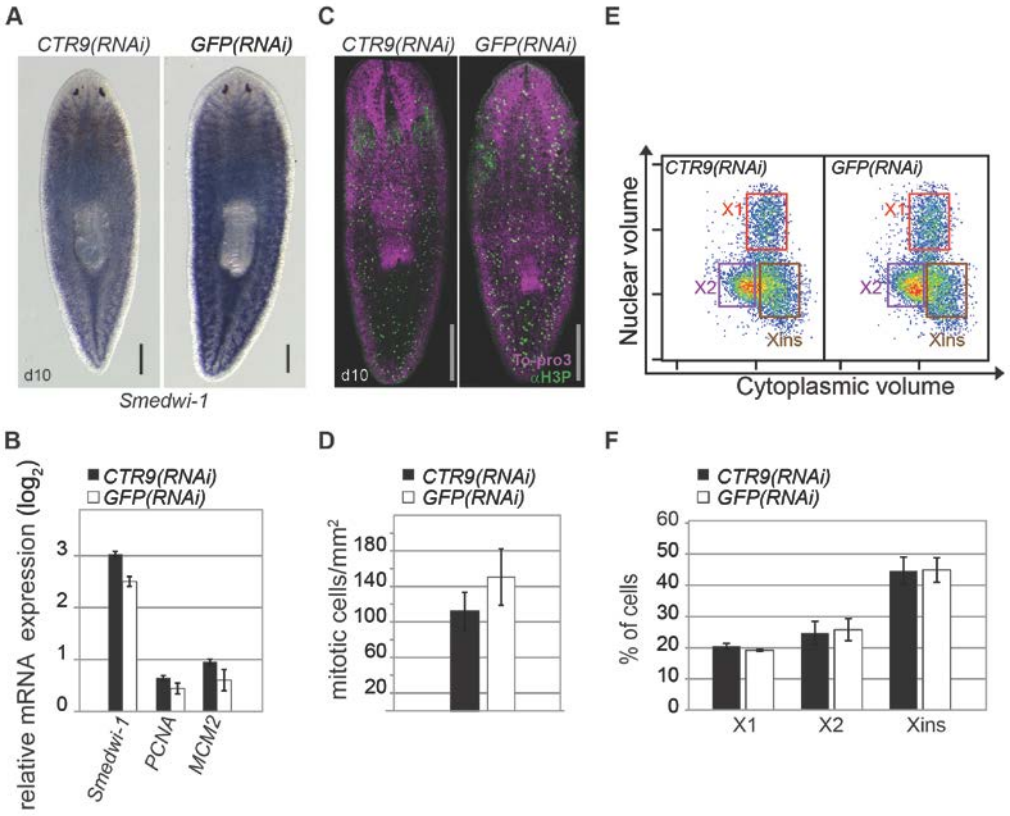
Supplementary Figure S3



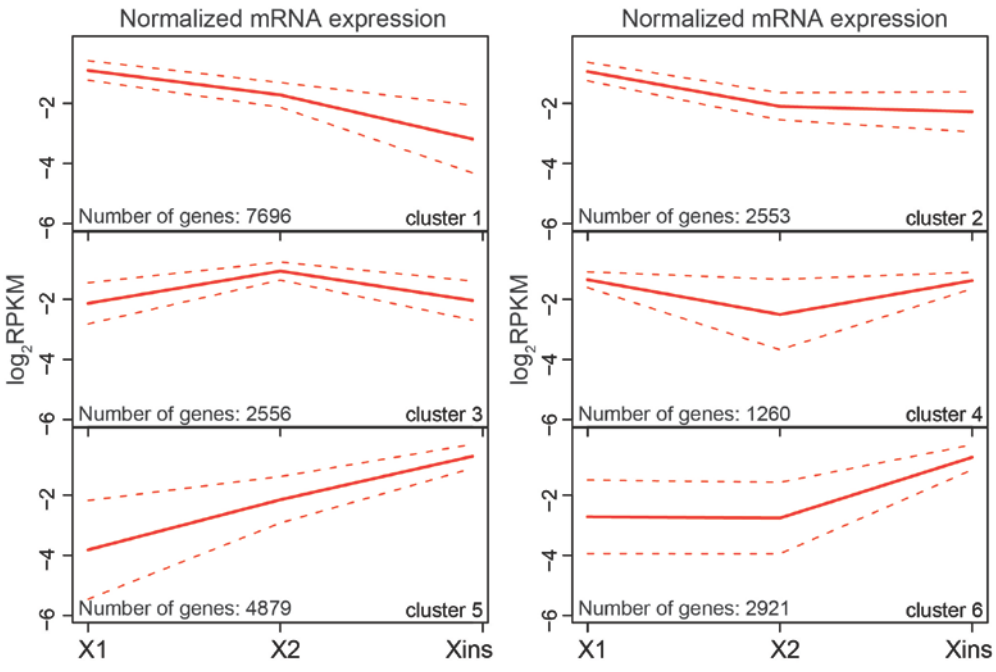
Supplementary Figure S4



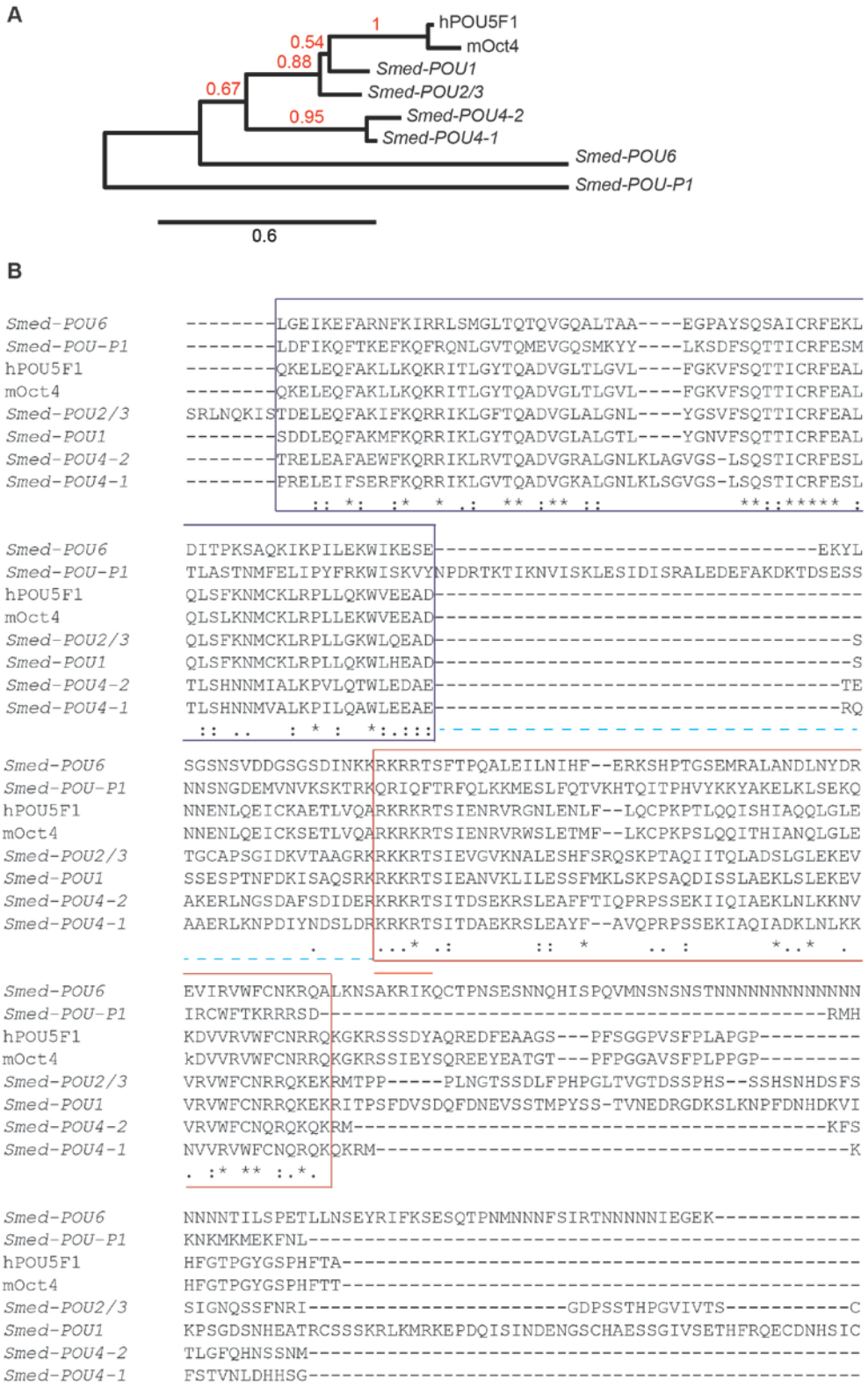
Supplementary Figure S5



Supplementary Figure S6



Supplementary Figure S7

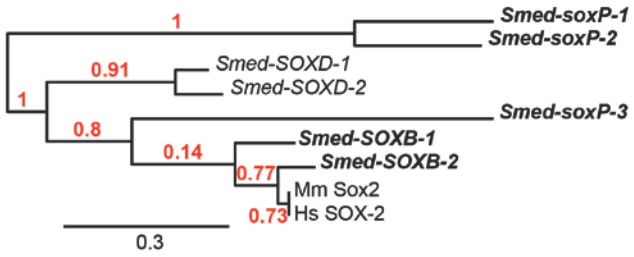


C


Smed-POU-P1 TASEQESL-NILLNKSLVFD----------LDFIKQFTKEFKQFRQNLGVTQ
HecPLN SSSGCNSV-INPADEAAIENL--------PFDESAAITTDDLEAFAKEFKRKRIKLGFTQ
DrPOU5F1 FSGGTAQNIPSAQAQSAPRSSGSSSGGCSDSEEEETLTTEDLEQFAKELKHKRITLGFTQ
AxOct4 STASSSSASPDLGAGGAPRDL--------DSGDEEGGTSADLEQFAKELKQKRITLGFTQ
hPOU5F1 PCTVTPGA-VKLEKEKLEQNP--------EESQDIKALQKELEQFAKLLKQKRITLGYTQ
mOct4 PCADRPNA---VKLEKVEPTP--------EESQDMKALQKELEQFAKLLKQKRITLGYTQ
XOct60 PSAISSRA----ERGLCSPSPNNASCG--PGTEEDGMTLEEMEEFAKELKQKRVALGYTQ
XOct91 SSATSSSS-GGTNVGTPRSLSRGASDGLSSDSEEEAPNSGEMEQFAKDLKHKRITMGYTQ
XOct25 SSTNSPNG-AINERATTIPNGEMLDGGQSSDNEEEVPSESEMEQFAKDLKHKRVSLGYTQ
Smed-POU-P1 MEVGQSMKYYLKSDFSQTTICRFESMTLASTNMFELIPYFRKWISKVYNPDRTKTIKNVI
HecPLN SDVGLGLGSLYGNIFSQTTICRFEALQLSFKNMCKLQPLLVKWLDEMD----------
DrPOU5F1 ADVGLALGNLYGKMFSQTTICRFEALQLSFKNMCKLKPLLQRWLNEAE----------
AxOct4 ADVGLALGALYGKMFSQTTICRFEALQLSFKNMCKLRPLLQRWLVEAD----------
hPOU5F1 ADVGLTLGVLFGKVFSQTTICRFEALQLSFKNMCKLRPLLQKWVEEAD----------
mOct4 ADVGLTLGVLFGKVFSQTTICRFEALQLSLKNMCKLRPLLEKWVEEAD----------
XOct60 GDIGHALGILYGKMFSQTTICRFESLQLTFKNMCKLKPLLEQWLGEAE----------
XOct91 ADVGYALGVLFGKTFSQTTICRFESLQLSFKNMCKLKPLLRSWLHEVE----------
XOct25 ADVGYALGVLYGKMFSQTTICRFESLQLSFKNMCQLKPFLERWVVEAE----------
Smed-POU-P1 SKLESIDISRALEDEFAKDKTDSESSNNSNGDEMVNVK-----SKTRKQRIQFTRFQLKK
HecPLN ----------NNFLGTENATSRNLFP-ARKRKKRTSIDLTLKET
DrPOU5F1 ----------NSENPQDMYKIERVFVDTRKRKRRTSLEGTVRSA
AxOct4 ----------TNENLQELCNLENALQQARKRK-RTSIENSVKDN
hPOU5F1 ----------NNENLQEICKAETLVQ-ARKRK-RTSIENRVRGN
mOct4 ----------NNENLQEICKSETLVQ-ARKRK-RTSIENRVRWS
XOct60 ----------NNDNLQEMIHKAQIEEQNRKRKMRTCFDTVLKGQ
XOct91 ----------NNKNLQEIISRGQIIPQVQKRKHRTSIENNVKCT
XOct25 ----------NNDNLQELINREQVIAQTRKRKRRTNIENIVKGT
Smed-POU-P1 MESLFQTVKHTQITPHVYKKYAKELKLSEKQIRCWFTKRRRSDRMHKNKM----------
HecPLN LEMYF--IKHQKPSGHDITEIAMQLNLEKEVVRVWFCNRRQKEKRLLASTGFVKSGDASM
DrPOU5F1 LESYF--VKCPKPNTLEITHISDDLGLERDVVRVWFCNRRQKGKRLALPF----DDECVE
AxOct4 LEAFF--LKCPKPTHQEIAHISEDLNLEKDVVRVWFCNRRQKGKRSICR----EEYDGFQ
hPOU5F1 LENLF--LQCPKPTLQQISHIAQQLGLEKDVVRVWFCNRRQKGKRSSSDYAQREDFEAAG
mOct4 LETMF--LKCPKPSLQQITHIANQLGLEKDVVRVWFCNRRQKGKRSSIEYSQREEYEATG
XOct60 LEGHF--MCNQKPGARELTEIAKELSLEKDVVRVWFCNRRQKEK-SKFRMSKGHEFVGGA
XOct91 LENYF--MQCSKPSAQEIAQIARELNMEKDVVRVWFCNRRQKGKRQVYPYIRENGGEPYD
XOct25 LESYF--MKCPKPGAQEMVQIAKELNMDKDVVRVWFCNRRQKGKRQGMPTVEENDGEGYD
Smed-POU-P1 ----------KMEKFNL------
HecPLN VSPSLSYPISFNHVPQEQSVID----------QLHSQQPYYQVDSFETTKS---
DrPOU5F1 AQYYEQ-SPPPPPHMGGTVLPGQGYPGPAHPGGAP---ALY-MPSLHRPDVFKNGLHPGL
AxOct4 QYPGMQ-PGPPALSHLPTSYIAQGYNG-----AAAAFAAVY-MQPFHDSEMYSQ----TV
hPOU5F1 SPFSGG-PVSFPLAPGPHFGTP-GY-------GSPHFTALYSSVPFPEGEAFPPVSVTTL
mOct4 TPFPGG-AVSFPLPPGPHFGTP-GY-------GSPHFTTLY-SVPFPEGEAFPSVPVTAL
XOct60 SPGSIQ-SEHISFTPIPANSQDYGLAS-----LHPNRAPFY-PPPFPRNELFPH-MAPGI
XOct91 APQTLT-PPSQGPFPLPQVMPSQVF-PTVPLGANP---TIY-VPTYHKNDMFPQAMHHGI
XOct25 VAQTMG-SPPVGHYALQQVVTPQGYM------AAP---QIY-ASAFHKNDLFPQTVPHGM

Supplementary Figure S8

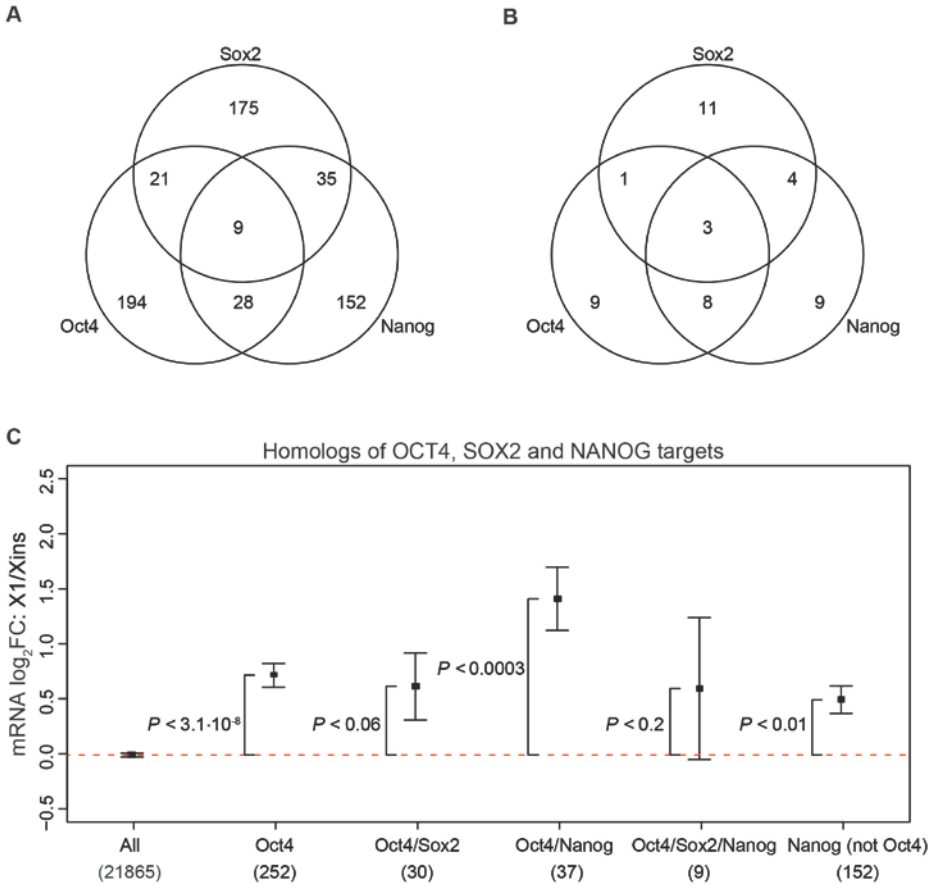
A



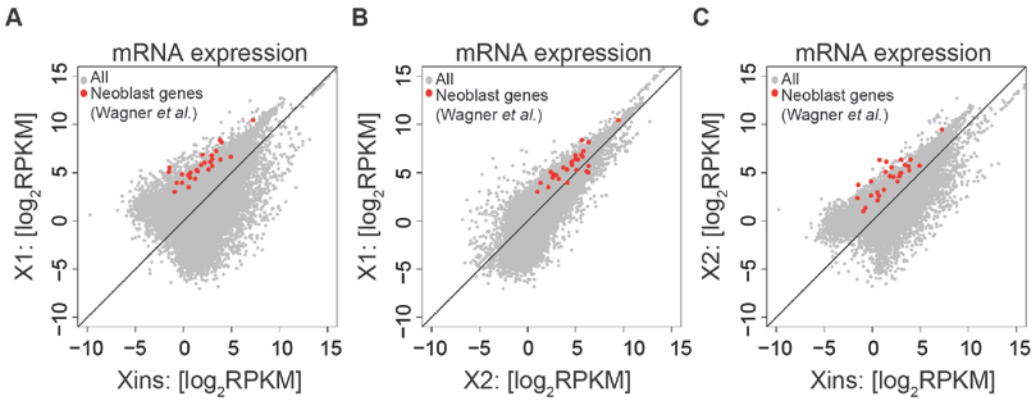
B

```
Smed-soxP-3     --------------------CENTENY----------------------------LCRQF
Smed-SOXD-1     SMNNSIMPRAGNSPITPSYNSSVSSDY--------ADFNERKINSSELNNVHSNNNNSNN
Smed-SOXD-2     HRLSAHQINQGLAIPSSNPQITSPTGV---IQINPPLETKLLNSENKNSNIESNVKCAET
Smed-soxP-1     --------------------CSNCSNN----------------QDTLKLWMVHLASCHP
Smed-soxP-2     ------------------DLCQKCSAY----------------QNSLRDWMKHLVDRHP
Smed-SOXB-2     ------------------SPTDESNNQY-------------------------SNQQ
Smed-SOXB-1     SIDGNNVEPNHKSDSSPNPPTTQCQGAHSFPTLNNSSCQQRPGLRTPKSEDIDRNQNDYF
Sox2 [Hsa]      ----------------PGPQQTSGGGG---------------------------------
Sox2 [Mmu]      ----------------PGPQQASGGGG---------------------------------

Smed-soxP-3     EDYPTIKDKIQ---VNREMKEFKDITKKSVKNSNHIK-IPMNAFMVWAKINRKLMASQYP
Smed-SOXD-1     CHSNTNNKSEQPLGLLNARIIRANRPGEIENRTNHVK-RPMNAFMVWSREERSKILKACP
Smed-SOXD-2     DDSQNSNGKILGARIVRSQRNPDKSDNMSRSSSSHIK-RPMNAFMVWAREERRKILKACP
Smed-soxP-1     LPPEWLSEKAKVLNDDENQRPYTIVTDSRKRRRRPTTPRPLNTFMIFSRHIRRNILSVFP
Smed-soxP-2     LPIQWMSDLANNLTDDKLSKPYSTVTESRKKRKKHITPRPLNSFMIFSRHLRRNVLIFFS
Smed-SOXB-2     HSPEIGYSDEDKSDIVEETEKINVSSSNASIKEDHVK-RPMNAFMVWSRGQRRKMAQENP
Smed-SOXB-1     DDCEFSKDQTQNVTTANSQTSQSNITSSSQKKDDRVK-RPMNAFMVWSRGQRRRMAQDNP
Sox2 [Hsa]      ----------------GNSTAAAAGGNQKNSPDRVK-RPMNAFMVWSRGQRRKMAQENP
Sox2 [Mmu]      ----------------GGGNATAAATGGNQKNSPDRVK-RPMNAFMVWSRGQRRKMAQENP
                                                    .  *:*:**:::.  *  :    .
                Sox-TCF HMG-box superfamily domain
Smed-soxP-3     NFSNLDISKLLGKKWNEMSAEERTPFINEAKQLRIEHLNQYPEFRYQPRRRQKKSKKSKP
Smed-SOXD-1     DMHNSNISKLLGKKWKEMSALDKRPYYEEQSRLSRQHMEEHPDYRYKPRPKRTCIINGRK
Smed-SOXD-2     DMHNSNISKLLGQKWKSMTAADKQPYYEEQSRLSRQHMEQHPEYRYRPRPKRTCIVDGRR
Smed-soxP-1     DASNSQLSKELGILWKKVTNEIKSLYEEEAKRLAQLHALEFPDYKYQPKKRVRRF-----
Smed-soxP-2     DASNSEISKLLGKLWNNIPATIKSMYDNEAKKLAILHSKEFPEYKYQP------------
Smed-SOXB-2     KMHNSEISKRLGAEWKHLTELQKRPFIDEAKRLRAIHMKDHPDYKYRPRRKPKSIFRKDK
Smed-SOXB-1     KMHNSEISKRLGGMWKNLSEEDKKPFIDEAKRLRANHMTQYPDYKYRPRRKHKAMDKQKK
SOX2 [Hsa]      KMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKTKTLMKKDK
Sox2 [Mmu]      KMHNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYRPRRKTKTLMKKDK
                .  * ::** **  *: :.   .  : :* ..*   *  :.*::.*.*

Smed-soxP-3     I--------------------KTDLASPTNQCSADFHSFENLN---------------LP
Smed-SOXD-1     L-----------------------------RISEYKDLIKLR---------------KS
Smed-SOXD-2     L-----------------------------RISEYKDLMRARKGPDG----------NR
Smed-soxP-1     ----------GRAIYDRQPCTTERSNSTSATDEENGKDNNSVQDGRRSSDIENQ----YI
Smed-soxP-2     ------------------------------------------------------------
Smed-SOXB-2     YTIP---LVYPWSNLPLIPSCSRSSLLSSEKLLGTLPSLSSLCSSSRPF---------MM
Smed-SOXB-1     FL--------GSAAMANVTALSHQLGREMG-YFNSGASGNGYLASGYASLAHSIDSTPYS
SOX2 [Hsa]      YTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQDQLG---YP
Sox2 [Mmu]      YTLPGGLLAPGGNSMASGVGVGAGLGAGVNQRMDSYAHMNGWSNGSYSMMQEQLG---YP
```

Supplementary Figure S9



Supplementary Figure S10



Supplementary Figure Legends

Supplementary Figure S1 Validation of sample quality for mRNA sequencing. **(A)** RNA from each sorted fraction was tested for the enrichment or depletion of validated neoblast markers (*Smedwi-1*, *Smedwi-2*, *Smed-PCNA*, *Smed-bruli* and *Smed-pumilio*) and tissue-specific genes (*Smed-MHC-A* and *Smed-syt*) by qRT-PCR. mRNA quantity relative to ubiquitously expressed *ura4* was plotted. WT and irradiated samples were used as control. qRT-PCR was performed in triplicates and error bars represent the standard deviation. **(B)** Venn diagram showing the number of transcripts obtained from each fraction by mapping the reads to the transcriptome.

Supplementary Figure S2 Validation of the reproducibility of the mRNA sequencing from FACS fractions. qRT-PCR was performed on 17 selected genes with samples from a different FACS experiment (biological replicate). After normalizing by the expression of the ubiquitously expressed gene *ura4*, fold changes of X1 (red symbols) and X2 (blue symbols) versus Xins ($-\Delta\Delta Ct$) were compared to sequencing derived values ($\Delta\Delta \log_2 RPKM$) by a linear regression ($R^2 = 0.80$). qRT-PCR was performed in triplicates and error bars represent the standard deviation.

Supplementary Figure S3 Gene expression analysis of neoblast enriched (X1, X2) and neoblast depleted (Xins) cell fractions. We computed Spearman's correlation coefficient (ρ) for each pair of fractions. Comparison of transcript expression for X1 versus Xins **(A)**, X1 versus X2 **(C)** and X2 vs Xins **(E)**. Consistent with the presumed cell composition of each fraction, the highest correlation is observed between fractions X1 and X2 ($\rho = 0.85$), which are both enriched in neoblasts. In comparison, the correlation of fraction X2 and Xins is reduced ($\rho = 0.71$), but still substantially higher than the correlation between fraction X1 and

Xins (ρ = 0.48), reflecting the presence of differentiated cells in X2 but not in X1. The distribution of fold changes reveals that, in comparison to Xins, the majority of genes is significantly down-regulated in fraction X1 **(B)** and X2 **(F)**. The fold change distribution of X1 versus X2 **(D)** exhibits a pronounced tail for negative log-fold changes reflecting the more complex transcriptome composition of the X2 fraction. Blue colored bars represent significant fold changes ($P < 0.01$).

Supplementary Figure S4 Genes whose depletion blocked regeneration are also required for tissue turnover. **(A)** Representative images of RNAi injected intact planaria. Pictures are taken on the last day of the RNAi experiments (see Materials and methods and Supplementary Methods for Experimental Procedures). Days after first RNAi injection is shown on each RNAi picture. A representative *GFP(RNAi)* animal is shown as control. **(B)** Representative images of αH3P staining from all RNAi experiments and control RNAi performed in this study. To-pro3 was used as background staining for DNA. Scale bar is 0.5 mm.

Supplementary Figure S5 Analysis of *CTR9(RNAi)* animals 10 days after first RNAi injection shows that neoblasts are still present and proliferating at an early time point. **(A)** Neoblast presence was detected by *Smedwi-1* ISH on animals fixed 10 days after first RNAi injection. **(B)** Relative mRNA levels of neoblast markers *Smedwi-1*, *Smed-pcna*, and *Smed-MCM2* in *CTR9(RNAi)* and *GFP(RNAi)* control animals. qRT-PCR was performed in triplicates and error bars represent the standard deviation. **(C, D)** RNAi animals were assayed for mitoses. Mitotic cells labeled with αH3P were counted and the number per surface area was calculated. **(C)** Representative image of *CTR9(RNAi)* and *GFP(RNAi)* animals labeled with αH3P are shown. To-pro3 was used as background staining for DNA. **(D)** Quantification of the mitotic cell count per surface area. More than eight animals were used per RNAi

experiment and average count numbers are shown. Error bars represent the standard deviation**. (E, F)** RNAi animals were assessed for the presence of proliferating neoblasts by flow cytometry. **(E)** Representative images of flow cytometry profile of cells dissociated from RNAi animals and the corresponding control RNAi. **(F)** Mean percentage of X1, X2 and Xins cell fractions (in the total population of Hoechst stained cells) of three biological replicates was determined by flow cytometry. Error bars represent the standard deviation. Scale bars are 0.5 mm. *$P < 0.05$, **$P < 0.001$ (*t*-test)

Supplementary Figure S6 Hierarchical clustering of all genes based on the correlation of their expression profiles in X1, X2 and Xins fractions (Supplementary Methods for Computational Analyses). Six clusters were identified with different expression profiles across the three fractions and each cluster contains more than 1,000 genes. Enriched GO-terms (Supplementary Methods for Computational Analyses) for each cluster are shown in Supplementary Table S5.

Supplementary Figure S7 Identification of OCT4 (POU5 domain transcription factor) homologs in planaria. Six candidates were selected based on sequence homology (Supplementary Information; Supplementary Table S8B). **(A)** Maximum likelihood phylogenies were estimated using the PhyML tool from Phylogeny.fr (Dereeper *et al*, 2008). **(B)** Alignment of POU domain homologs with human OCT4 (hPOU5F1) and mouse Oct4 (mOct4) using MUSCLE (3.7) multiple sequence alignment (Edgar *et al*, 2004; Deereper *et al*, 2008). **(C)** Alignment of the protein sequence of *Smed-POU-P1* with OCT4 orthologs. The open reading frame (ORF) of *Smed-POU-P1* was aligned with Oct4 orthologs from mammals (mouse mOct4 and human hPOU5F1), frog (*Xenopus* XOct-25, -91 and -60; Hinkley *et al*,

1992), axolotl (AxOct-4; Bachvarova *et al*, 2004), fish (*Danio rerio* DrPOU5F1; Takeda *et al*., 1994) and cnidarian (hydra HecPLN; Millane *et al*, 2011) using MUSCLE (3.7).

Supplementary Figure S8 Identification of SOX2 (Sox class B) homologs in planaria. Seven candidates were selected based on sequence homology (Supplementary Methods for Computational Analyses; Supplementary Table S8C) including the recently annotated Sox homologs *Smed-soxP-1, -2* and *-3* (Wagner *et al*, 2012). **(A)** Maximum likelihood phylogenies were estimated using the PhyML tool from Phylogeny.fr (Dereeper *et al*, 2008). Marked in bold are the three recently reported Sox family member proteins (*Smed-soxP-1,-2 and -3*) (Wagner *et al*, 2012) and the two closest SOX2 homologs in our planarian transcriptome assembly, which we annotated as *Smed-SOXB-1* and *Smed-SOXB-2*. **(B)** Aligment of the planarian SOX homologs with human and mouse Sox2 using MUSCLE (3.7) multiple sequence alignment (Edgar 2004). The conservation is restricted to the SOX-TCF HMG-box superfamily domain. Outside the HMG-box domain, Sox family proteins are quite variable.

Supplementary Figure S9 Coregulation of genes by Oct4 and Nanog in mouse ESCs is reflected by enhanced up-regulation of planarian homologs in neoblasts. **(A)** Overlap of planrian homologs of physical Oct4, Nanog and Sox2 targets in mouse. Direct targets were identified by ChIP-Seq (Chen *et al*, 2008). Overlap of genes homologous to the top 10% targets of Oct4, Nanog and Sox2 identified by Chen *et al*, 2008. **(B)** Homologs were determined as in (A), but only genes were retained, that are at least four fold up-regulated in ESCs and ICM outgrowths and are therefore considered pluripotency maintenance genes according to Tang *et al*, 2010. **(C)** Comparison of average transcript fold changes between X1 and Xins for homologs of physical Oct4 targets, and Oct4 targets co-regulated by Nanog and

Sox2, respectively. In addition, average up-regulation is shown for homologs of mouse genes regulated by Nanog but not by Oct4. While, in comparison to homologs of Oct4 targets, co-regulation of Oct4 and Nanog leads to increased fold changes of planarian homologs, co-regulation by Sox2 does not increase the average up-regulation of planarian homologs in neoblasts. Data are shown for the top 10% of ChIP-Seq targets from Chen *et al* (2008). Error bars represent the standard error of the mean. The hypergeometric *P* reflects the significance of observing a population of genes with an average fold change higher or lower, respectively, than the average of all genes (Fisher's exact test). The number of mammalian genes with planarian homologs is shown in parentheses.

Supplementary Figure S10 Genes with enhanced neoblast expression, identified by irradiation experiments (Wagner *et al*, 2012) while this manuscript was under review, are highly up-regulated in the X1 fraction. Comparison of transcript expressions for X1 versus Xins **(A)**, X1 versus X2 **(B)** and X2 and versus Xins **(C)**. All genes are represented by grey dots and genes with validated up-regulation in neoblasts by Wagner *et al*, 2012 are highlighted in red. All of these genes are up-regulated in neoblast enriched fractions X1 and X2 and all but one (*Smed-soxP-3*) fall into expression cluster 1, which contains all previously known neoblast markers. *Smed-soxP-3* is most highly expressed in X2 and therefore falls into cluster 3.

Supplementary Tables

Supplementary Table S1 Planarian neoblast and tissue-specific markers

Neoblast markers			
Function	**Gene name**	**Transcript ID**	***Reference***
RNA processing, metabolism and transport	*Smedwi-1*	isotig03286	*Smedwi-1 (Reddien et al. 2005b)*
	Smedwi-2	isotig05155	*Smedwi-2 (Reddien et al., 2005b)*
	Smedwi-3	isotig00135	*Smedwi-3 (Palakodeti et al., 2008)*
	Smed-bruli	isotig00744	*Smed-bruli (Guo et al., 2006)*
	Smed-tud1A	isotig12850	*Spoltud-1 (Solana et al., 2009)*
	Smed-tud1B	isotig12304	*Spoltud-1 (Solana et al., 2009)*
	Smed-pum	CUFF.266451.1	*DjPum (Salvetti et al., 2005)*
	Smed-CBC-1	isotig04425	*DjCBC-1(Yoshida-Kashikawa et al., 2007)*
	Smed-SmB	isotig14565	*Smed-SmB (Fernandéz-Taboada et al., 2010)*
	Smed-sam68L1	isotig08321	*Dj-Sam68like1 (Rossi et al., 2007, Eisenhoffer et al., 2008)*
	Smed-VLGA1	isotig09591	*SpolvlgA and DjVLGA (Solana and Romero, 2009; Shibata et al, 2010)*
	Smed-nanos	isotig19923 CUFF.280331.2	*Smednanos and Smednos (Wang et al., 2006 and Handberg-Thorsager and Saló, 2007)*
	Smed-THOC4	isotig03534	*Smed-THOC4 (Reddien et al., 2005a; Eisenhoffer et al., 2008)*
Cell cycle and replication	*Smed-MCM2*	isotig15091	*DjMCM2 (Salvetti et al., 2000)*
	Smed-PCNA	CUFF.66956.1	*DjPCNA (Ito et al., 2001)*
	Smed-CCNBL1	isotig24777	*Smed-cyclinB (Reddien et al., 2005a; Eisenhoffer et al., 2008)*
	Smed-RRM2L1	CUFF.307465.1	*Smed-RRM2-1 (Eisenhoffer et al., 2008)*
	Smed-PP32L1	CUFF.261724.1	*Smed-pp32a (Eisenhoffer et al., 2008)*
	Smed-PHB1	isotig10463	*Smed-prohibitin-1 (Eisenhoffer et al., 2008)*
Chromatin	*Smed-CHD4*	CUFF.60081.1	*Smed-CHD4 (Scimone et al., 2010)*

modification	*Smed-RBAP48*	isotig07452	*DjRbAp48 (Rossi et al., 2007 and Bonucelli et al., 2009)*
	Smed-H2AZ	isotig05441	*DjH2AZ (Rossi et al., 2007)*
	Smed-CBX1	isotig15327	*Smed-Cbx-1 (Reddien et al., 2005a; Eisenhoffer et al., 2008)*
	Smed-HDAC1	isotig10381	*Smed-HDAC-1(Reddien et al., 2005a; Eisenhoffer et al., 2008)*
Nucleic acid binding	*Smed-PAIRBP-1*	isotig10428	*Smed-PAIRBP-1 (Eisenhoffer et al., 2008)*
	Smed-HMG-1	CUFF.40159.1	*Smed-HMG-1 (Eisenhoffer et al., 2008)*
	Smed-HMG-2	isotig06018	*Smed-HMG-2 (Eisenhoffer et al., 2008)*
Chaperones	*Smed-HSP60A*	CUFF.325473.1	*DjHSP60 (Rossi et al., 2007)*
	Smed-HSP60B	isotig19365	*DjHSP60 (Rossi et al., 2007)*
Cell cell communication	*Smed-inx11*	CUFF.236310.1	*Smed-inx11 (Oviedo and Levin, 2007)*
	Smed-egfr-3	isotig10370	*Smed-egfr-3 (Fraguas et al., 2011)*
Mitochondrial translation	*Smed-EF-TU*	CUFF.309618.2	*Smed-EF-TU (Eisenhoffer et al., 2008)*

Tissue specific markers			
Tissue	**Gene name**	**Transcript ID**	***Reference***
brain	*Smed-netrin1*	isotig19061	*Smed-netrin1 (Cebrià and Newmark, 2005)*
	Smed-netrin2	isotig12149	*Smed-netrin2 (Cebrià and Newmark, 2005)*
	Smed-syt	isotig14123	*Djsyt (Tazaki et al., 1999)*
	Smed-foxG	isotig25372	*DjFoxG (Koinuma et al., 2002)*
gatrovascular system	*Smed-egfr-2*	CUFF.232164.1	*Smed-egfr-2 (Fraguas et al., 2011)*
gland cells	*Smed-mag-1*	isotig03702	*marginal adhesive gland-1(mag-1) (Zayas et al., 2010)*
	Smed-rcn1	isotig11665	*Smed-reticulocalbin1 (rcn1) (Zayas et al., 2010)*
central secretory cells	*Smed-ZMPL*	isotig14950	*similar to CG6763 gene product (hypothetical zinc metalloproteinase) (Newmark and Sánchez Alvarado, 2002)*
muscle	*Smed-MHC-A*	isotig19061	*DjMHC-A (Kobayashi et al., 1999)*

Supplementary Table S2 Smed-genes encoding for putative chromatin proteins and associated factors are globally upregulated in neoblasts.

Smed-gene	Transcript ID (E value)	mRNA Fold Change (Log2)		Protein Fold Change (Log2)	
		X1/Xins	X2/Xins	X1/Xins	X2/Xins
Smed-CHD1L	CUFF.114734.1 (1e-173)	8.97	6.13	NA	NA
Smed-EHMT1	CUFF.218498.1 (4e-16)	7.04	3.95	NA	NA
Smed-RAD54L1	CUFF.315523.1 (6e-48) isotig13354 (9e-64)	6.54 6.14	3.79 3.54	NA NA	NA NA
Smed-RAD54L2	CUFF.167404.1 (0.0)	5	2.52	NA	NA
Smed-SETD8	isotig03820 (3e-38)	4.98	2.78	1.29	2.08
Smed-CREBBP	CUFF.15491.2 (7e-149)	4.68	2.07	NA	NA
Smed-CBX3L2	isotig14730 (1e-28)	4.62	2.95	NA	NA
Smed-CBX3L1	isotig15327 (1e-13)	4.4	2.43	1.04	1.26
Smed-SSRP1	isotig15484 (1e-118)	4.39	1.72	8.6	6.76
Smed-WDR82	isotig15484 (1e-127)	4.39	1.72	8.6	6.76
Smed-Rb-like (RBL1)	CUFF.194815.1 (3e-103) isotig05062 (1e-36)	4.39 3.56	2.78 1.87	NA NA	NA NA
Smed-SUPT16H	isotig21091 (2e-75) isotig25051 (3e-172)	4.06 3.56	1.55 1.87	NA NA	NA NA
Smed-SETD6	isotig25474 (6e-34)	3.99	1.77	-0.48	1.24
Smed-HAT1	isotig18510 (8e-65)	3.99	1.55	NA	NA
Smed-RBBP4L1	isotig10084 (0.0)	3.94	1.97	NA	4.85
Smed-suz12	isotig23175 (4e-34) isotig19342 (1e-05)	3.84 3.84	1.93 1.93	NA NA	NA NA
Smed-SETL1	CUFF.179612.1 (5e-68)	3.78	2.45	NA	NA

Smed-PRMT5	isotig23262 (5e-166)	3.71	2.04	NA	NA
Smed-BAZ1B	isotig23650 (6e-29)	3.65	2.08	NA	NA
	isotig23594 (5e-68)	3.43	1.95	NA	NA
	isotig22057 (4e-18)	2.53	1.15	NA	NA
Smed-MED17	isotig00222 (1e-16)	3.53	1.53	NA	NA
Smed-KDM1B	isotig11099 (2e-37)	3.52	2.3	NA	NA
Smed-ASF1B	CUFF.167247.1 (1e-68)	3.52	1.98	NA	NA
Smed-MED8	isotig13594 (4e-18)	3.51	2.34	-0.26	3.51
Smed-SMARCC2	CUFF.320239.1 (5e-49)	3.5	2.7	-2.24	NA
	isotig10619 (1e-163)	3.5	2.7	-2.24	NA
	CUFF.320241.1 (3e-95)	3.32	2.07	NA	NA
Smed-RUVBL1	isotig15765 (0.0)	3.48	2.08	3.06	2.53
Smed-RBBP4L2	isotig19102 (7e-100)	3.43	2	2.48	3.53
Smed-HDAC1	isotig10381 (0.0)	3.42	2.54	1.79	0.98
Smed-SETL2	CUFF.223953.1 (2e-77)	3.41	2.19	0.86	3.25
Smed-RBBP5	isotig04201 (8e-61)	3.4	1.85	NA	NA
Smed-RUVBL2	isotig14523 (0.0)	3.4	1.77	1.03	0.79
Smed-TRRAP	CUFF.194440.1 (8e-100)	3.36	2.16	NA	NA
Smed-MORF4	isotig14053 (1e-36)	3.31	1.8	NA	NA
Smed-SETL3	isotig06685 (3e-80)	3.3	1.85	0.57	0.43
Smed-CHAF1B	isotig18639 (3e-97)	3.28	2.08	NA	NA
Smed-PRMT1	isotig00993 (7e-146)	3.28	2.4	1.22	0.9
Smed- WHSC1L1	isotig23418 (6e-44)	3.18	1.35	NA	NA
Smed-EZH2	isotig21889 (4e-76)	3.17	2.03	NA	NA
Smed-RCOR2	isotig10270 (2e-19)	3.16	1.58	NA	NA
Smed-MED7	isotig18334 (5e-48)	3.15	1.85	NA	5.1

Smed-SMARCA5	CUFF.138859.1 (0.0)	3.14	1.65	NA	NA
Smed-CHD4L1	CUFF.60081.1 (0.0)	3.14	1.3	NA	NA
Smed-EMSY	CUFF.206254.1 (4e-30)	3.13	1.57	4.52	2
Smed-MTA1L1	isotig21086 (7e-159)	3.1	1.67	NA	NA
Smed-CBX3L3	CUFF.4604.1 (3e-14)	3.07	2.15	NA	NA
Smed-MBD2	isotig03836 (1e-44)	3.06	2.15	NA	NA
Smed-ISWI	isotig09198 (2e-62)	3.05	1.47	NA	NA
Smed-RBAp48 (RBBP4)	isotig07452 (0.0)	3.04	2.27	0.59	1.14
Smed-HINFP	isotig24141 (1e-27)	3.03	1.45	NA	NA
Smed-CTR9	isotig14289 (0.0)	2.96	1.3	NA	NA
Smed-SUPT5H	isotig21360 (4e-49) isotig14569 (3e-78)	2.92 2.09	1.39 0.58	NA 5.75	NA 4.39
Smed-SPT6	isotig23138 (8e-90) isotig13856 (2e-135) isotig13856 (2e-135)	2.91 2.7 2.7	1.35 1.08 1.08	0.86 -0.37 2.46	0.02 -2.37 2.73
Smed-HCF1	isotig23256 (0.0)	2.91	1.45	2.48	2.38
Smed-SMARCA1L1	CUFF.63650.1 (8e-162) isotig22339 (0.0)	2.88 2.88	1.4 1.4	NA NA	NA NA
Smed-EED	isotig15566 (3e-36)	2.72	1.49	NA	NA
Smed-LEO1	CUFF.34790.2 (1e-93) isotig25475 (2e-89)	2.66 2.66	1.83 1.83	NA NA	NA NA
Smed-SMARCD1	CUFF.70832.1 (3e-87)	2.66	2.01	0.32	3.53
Smed-PHF8	CUFF.15759.1 (8e-90) isotig00331 (6e-59)	2.66 2.63	1.81 1.81	NA NA	NA NA
Smed-KAT8	isotig22266 (6e-145)	2.63	1.59	NA	NA

Smed-ACTL6A	isotig20196 (1e-167)	2.62	2.1	1.33	1.3
Smed-PAF1	isotig16023 (1e-123)	2.61	1.26	NA	NA
Smed-SIN3A	isotig10422 (3e-41)	2.58	1.37	NA	NA
Smed-PCGF2L1	isotig25501 (2e-52)	2.58	1.99	NA	NA
Smed-KDM1A	isotig22047 (7e-31)	2.53	1.8	NA	NA
Smed-KDM5AL1	CUFF.88356.1 (0.0)	2.5	0.94	1.07	3.57
Smed-KDM5B	CUFF.218482.1 (3e-178)	2.48	2.02	NA	NA
Smed-SMARCA1L2	isotig10391 (0.0)	2.42	1.34	NA	NA
Smed-KDM6A	isotig23268 (3e-174)	2.4	1.43	NA	NA
Smed-RNF40	isotig22285 (1e-59)	2.36	1.33	1.43	-0.44
	isotig22285 (1e-59)	2.36	1.33	3.23	0.6
Smed-EP400	CUFF.13256.2 (3e-102)	2.33	1.56	NA	NA
Smed-SETDB1L1	CUFF.187048.1 (6e-40)	2.33	0.44	NA	NA
Smed-MLL1	isotig22961 (2e-37)	2.27	1.18	NA	NA
Smed-ASH2L	isotig09869 (8e-100)	2.26	1.14	NA	NA
	CUFF.95059.1 (4e-11)	2.69	1.63	NA	NA
Smed-MTA1L2	isotig09679 (5e-133)	2.23	1.36	4.76	4.16
Smed-JMJD6	isotig21880 (8e-159)	2.21	0.86	NA	NA
Smed-NSD1	isotig23535 (2e-54)	2.19	1.37	NA	NA
Smed-SUV420H1	CUFF.89708.1 (2e-68)	2.15	0.27	NA	NA
Smed-CDC73	isotig08692 (8e-139)	2.15	1.03	NA	NA
	isotig08692 (8e-139)	2.15	1.03	2.25	NA
Smed-WDR5	isotig10299 (5e-123)	2.13	1.44	-0.07	-1.83
Smed-BPTF	isotig22812 (2e-72)	2.12	1.3	NA	NA
Smed-RNF2	isotig12399 (8e-58)	2.12	0.75	NA	NA
	isotig12400 (6e-61)	2.12	0.75	NA	NA

Smed-SUV39H1	isotig11145 (3e-83)	2.12	0.65	NA	NA
Smed-CHD4	isotig22790 (0.0)	2.08	1.38	NA	NA
Smed-ARID2	CUFF.228007.1 (1e-38)	1.99	1.3	NA	NA
Smed-KDM5AL2	isotig21389 (2e-84)	1.96	1.65	NA	NA
Smed-BRD7	CUFF.69337.1 (2e-20)	1.96	0.52	NA	NA
Smed-MCRS1	isotig16417 (2e-27)	1.95	0.74	NA	NA
Smed-KAT7	isotig15587 9e-93)	1.89	1.18	NA	NA
Smed-RING1	isotig25760 (3e-35)	1.87	1.24	NA	NA
Smed-PBX2	isotig20749 (3e-136)	1.85	1.44	NA	NA
Smed-RTF1	isotig16407 (2e-76)	1.78	0.96	3.63	0.8
Smed-TBL1XR1	CUFF.65859.3 (3e-45) isotig17012 (3e-137)	1.78 1.68	0.71 0.45	NA NA	NA NA
Smed-KAT5	isotig09000 (8e-157)	1.78	0.94	NA	NA
Smed-SMARCA4	isotig22778 (0.0)	1.68	1.18	0.38	-0.6
Smed-EPC1	isotig20325 (9e-35)	1.59	1.12	NA	NA
Smed-PHC2	isotig21205 (8e-17**)**	1.58	0.92	NA	NA
Smed-ASH1L	CUFF.57846.1 (3e-119)	1.57	1.18	NA	NA
Smed-L3MBTL	CUFF.295850.1 (5e-62)	1.51	0.16	NA	NA
Smed-SETDB1L2	CUFF.79184.1 (2e-56)	1.25	-0.34	NA	NA
Smed-SMARCB1	isotig11020 (7e-164)	1.22	0.96	0.71	3.31
Smed-SCMH1	isotig12839 (2e-46)	1.2	0.07	NA	NA
Smed-DOT1L	isotig15599 (7e-93)	1.15	0.53	NA	NA
Smed-PCGF2L2	isotig19827 (3e-39)	1.14	0.94	NA	NA
Smed-PCGF3	isotig09477 (1e-44)	1.07	0.89	NA	NA
Smed-SUV39H2	isotig20060 (2e-27)	0.76	0.31	NA	NA
Smed-HDAC11	isotig21741 (3e-106)	0.45	0.1	-0.47	-1.84

	CUFF.730.2 (9e-69)	0.45	0.1	NA	NA
Smed-WDR88	isotig21237 (2e-48)	0.19	-0.29	NA	NA
Smed-HDAC4	isotig22594 (7e-144)	-0.17	-0.21	NA	NA

Supplementary Table S3 Smed-genes encoding for putative germ granule components and interacting factors show strong up-regulated expression in neoblasts.

Smed-gene	Transcript ID (E value)	mRNA Fold Change (Log2)		Protein Fold Change (Log2)	
		X1/Xins	X2/Xins	X1/Xins	X2/Xins
Smed-EIF3C	CUFF.13203.1	6.68	4.24	1.4	-0.66
Smedwi-1	isotig03286	6.3	3.7	6.49	3.42
Smed-bruli	isotig00744	5.58	3.5	NA	NA
Smed-PAPD4 *PAPD4/GLD2*	CUFF.65828.2 (1e-56)	4.76	2.06	NA	NA
Smed-TDRD9 *TDRD9*	isotig23459 (3e-127)	4.66	2.3	NA	NA
Smed-vasa *DDX4/VASA* *Djvas-1*	isotig05106 (1e-74) (0.0)	4.6	1.86	10.74	8.07
Smed-EIF4E *EIF4E2*	isotig23956 (1e-65)	4.36	2.16	NA	NA
Smedwi-2	isotig05155	4.34	2.45	2.22	0.55
Smed-MOV10L1B *MOV10L1*	CUFF.71041.2 (5e-158)	4.28	1.5	NA	NA
Smed-TDRD1L2 *TDRD6* *Spoltud1*	isotig12304 (3e-08) (6e-73)	4.17	1.79	3.9	1.18
Smedwi-3	isotig00135	4.1	2.69	7.21	4.43
Smed-nanos	isotig19923|CUFF.280331.2	4.05	2.97	NA	NA
Smed-CBP80 *NCBP1*	isotig21130 (4e-134)	3.77	1.97	3.61	1.78

DjCBP80	(0.0)				
Smed-PRMT5 *PRMT5*	isotig23262 (5e-166)	3.71	2.04	NA	NA
Smed-TDRD1L1 *TDRD6* *Spoltud-1*	isotig12850 (6e-06) (0.0)	3.63	1.68	5.62	1.2
Smed-TDRD1L3 *TDRD1* *Spoltud-2*	isotig25555 (8e-14) (3e-96)	3.42	1.91	NA	NA
Smed-DAZL *DAZ/BOULE*	CUFF.172807.1 (3e-36)	3.35	2.15	NA	NA
Smed-THOC4 *THOC4*	isotig03532 (2e-16)	3.22	1.99	1.43	3.69
Smed-TDRKH *TDRKH*	CUFF.18165.1 (9e-28)	3.14	1.6	NA	NA
Smed-XRN1 *XRN1*	isotig23342 (2e-122)	2.98	1.57	NA	NA
Smed-TIA1 *TIA1* *DjTIA1*	isotig08477 (2e-70) (9e-157)	2.87	1.72	0.29	-0.17
Smed-SmB	isotig14565	2.83	1.81	1.9	1.97
Smed-HNRPF *HNRPF* *DjHNRPF*	isotig08476 (1e-61) (0.0)	2.82	2.36	1.48	1.03
Smed-EIF4A3 *EIF4A3* *DiEIF4A3*	isotig19249 (0.0) (0.0)	2.81	1.72	-0.43	0.45
Smed-RBM8	isotig24901	2.75	1.94	-0.62	NA

RBM8A	(2e-53)				
Smed-PUF6 *KIAA0020*	isotig12238 (2e-22)	2.73	1.4	NA	NA
Smed-PAPB1 *PAPB2*	isotig16256 (3e-48)	2.62	1.87	-2.15	-0.48
Smed-PELO1 *PELO*	CUFF.131884.1 (1e-88)	2.61	1.39	0.55	-2.36
Smed-MOV10L1A *MOV10L1*	isotig04939 (7e-161)	2.54	0.33	NA	3.06
Smed-MAGOH *MAGOH*	isotig16373 (8e-62)	2.48	1.62	NA	NA
Smed-HNRNPL *HNRNPL* *DjHNRNPL*	isotig10064 (3e-53) (0.0)	2.44	1.12	NA	NA
Smed-DCPS *DCPS* *DjDCPS*	contig23324 (4e-73) (3e-81)	2.43	0.44	0.2	1.63
Smed-FRX2 *FRX2*	isotig11061 (2e-57)	2.43	0.75	NA	NA
Smed-CBP20 *NCBP2*	isotig18382 (4e-75)	2.39	1.46	NA	NA
Smed-LSM14 *LSM14B*	isotig16507 (4e-35)	2.34	0.92	-0.7	2.22
Smed-AGO2 *EIF2C2*	isotig02996 (4e-136)	2.19	1.81	NA	NA
Smed-pum *DjPum*	CUFF.266451.1 (0.0)	2.19	1.29	NA	NA
Smed-PELO2	CUFF.293162.1	2.15	1.77	NA	NA

PELO	(4e-83)				
Smed-DDX6 *DDX6* *DjCBC1*	isotig04425 (0.0) (0.0)	2.14	1.18	0.5	-0.28
Smed-EDC4 *EDC4*	isotig03992 (5e-21)	2.12	0.85	NA	NA
Smed-DICER1 *DICER1*	isotig23598 (9e-91)	2.04	0.7	NA	NA
Smed-CNOT4 *CNOT4*	isotig13247 (2e-91)	1.66	0.48	5.26	2.96

Supplementary Table S8 Results of HMM based homology search in planarians.

A

E-value (full sequence)	E-value (best domain)	Domain score	Smed-gene	Putative Homology	Cluster
7.5e-36	1e-35	125.3	CUFF.32446.1 *Smed-MSX-1*	msh1 [*D. japonica*] (5e-139) MSX1 [*H. sapiens*] (3e-42)	5
1.8e-35	1e-35	124.5	isotig05838 *Smed-DTH-1*	Dth-1 [*D. tigrina*] (1e-152) NKX2-2 [*H. sapiens*] (2e-39)	6
4.3e-35	5.4e-35	122.9	CUFF.114724.1 *Smed-BARHL1*	BARHL2 [*H. sapiens*] (9e-46)	1
5.2e-33	5.2e-33	116.4	isotig24657 *Smed-BARHL2*	BARHL2 [*H. sapiens*] (9e-48)	1

B

E-value (full sequence)	E-value (best domain)	Domain score	Smed-gene	Putative Homology (Blastp E value)	Cluster
1.7e-95	1.9e-95	322.1	isotig25237 *Smed-POU1*	DjPOU1 [*D. Japonica*] (0.0) POU3F4 [*H. sapiens*] (2e-88)	3
1.7e-93	2.2e-93	315.3	isotig23172 *Smed-POU2/3*	POU3F2 [*H. sapiens*] (9e-82)	1
7e-64	8.9e-64	217.9	isotig21311 *Smed-POU4-1*	POU4F3 [*H. Sapiens*] (2e-83)	1
2.5e-62	3.6e-62	212.7	CUFF.259540 *Smed-POU4-2*	POU4F3 [*H. sapiens*] (4e-82)	3
1.7e-43	2.1e-43	150.9	isotig19788 *Smed-POU6*	POU6F2 [*H. sapiens*] (1e-77)	4
2.5e-17	6.8e-17	63.7	isotig26184 *Smed-POU-P1*	POU3F1 [*H. sapiens*] (8e-17)	1

C

E-value (full sequence)	E-value (best domain)	Domain score	Smed-gene	Putative Homology	Cluster
1.1e-63	2.6e-63	216.5	CUFF.51041.1 *Smed-SOXB-2*	SOX-14 [*H. sapiens*] (7e-55)	3
3.1e-62	3.1e-62	212.9	CUFF.244913.1 *Smed-SOXB-1*	SOX-2 [*H. sapiens*] (8e-43)	1
1.7e-42	2.6e-42	147.5	isotig23570 *Smed-SOX-D1*	SOX-5 [*H. sapiens*] (5e-51)	1
4.8e-41	4.8e-41	143.3	isotig22872 *Smed-SOX-D2*	SOX-6 [*H. sapiens*] (1e-53)	3
2e-38	2.4e-38	134.5	isotig17583 *Smed-soxP-3*	SOX-11 [*H. sapiens*] (2e-24)	3
1.5e-24	1.8e-24	88.9	isotig23820 *Smed-soxP-1*	SOX-5 [*H. sapiens*] (2e-18)	1
2.8e-20	3.4e-20	74.8	isotig25189 *Smed-soxP-2*	SOX-22 [*H. sapiens*] (1e-15)	1

Supplementary Table Legends

Supplementary Table S1 Known planarian neoblast and tissue specific genes. Listed are the planarian gene names, gene identifiers from our own annotation, and for each gene the functional category and a reference.

Supplementary Table S2 Planarian genes encoding for putative chromatin proteins and associated factors are globally upregulated in neoblasts. The selected Smed-genes encode for mammalian homologs known to be associated with chromatin organization, nucleosome assembly, transcription regulation, and DNA replication and repair. The official NCBI gene name of the human homolog is given, unless the Smed-gene has been characterized before or its sequence was entered in GenBank database. Previously characterized genes are highlighted in red (with human homolog official gene name in parenthesis). When more than one

homolog is found, L (like) followed by the homolog number is appended at the end. If long protein coding genes are fragmented into two or more transcripts, all corresponding transcripts mapping to the same gene locus are listed. The E-values of the BLAST protein alignment between planarian genes and their human homologs are indicated in parentheses. The list is ordered by decreasing fold changes between X1 and Xins. Corresponding protein fold-changes are also given (NA: protein could not be detected). In a few cases, artificial frameshift or fusion events were introduced in our transcriptome assembly (Adamidi *et al.*, 2011). In these cases, protein fold changes were detected for more than one ORF encoded within the corresponding transcript and are listed in the table.

Supplementary Table S3 Planarian genes encoding for putative germ granule components and interacting factors show strong overexpression in neoblasts. The table comprises homologs encoding for proteins that have been found to be associated with chromatoid bodies, pi-bodies, p-bodies, nuage, polar or germ granule formation and more generally with RNA localization and processing. The official NCBI gene name for the human homolog gene or, if applicable, to the closely related species *Dugesia japonica* (Dj) is entered below the corresponding Smed-gene. If the Smed-gene was previously characterized, the gene entry is highlighted in red. When long protein coding genes are fragmented into two or more transcripts, all corresponding transcripts mapping to the same gene locus are listed. The E-values of the BLAST protein alignment between planarian genes and their homologs in human or *Dugesia japonica* are indicated in parentheses. The list is ordered by decreasing fold changes between X1 and Xins. Corresponding protein fold-changes are also given (NA: protein could not be detected).

Supplementary Table S4 List of genes within each of the three expression cluster. The table indicates the internal gene identifier, best protein BLAST hits to human and mouse and log_2-fold changes between the different fractions. The list is ordered by decreasing fold changes between X1 and Xins. Download excel file Supplementary Table S4.

Supplementary Table S5 Enriched GO terms for the six expression clusters. The clustering procedure and GO annotation is described in the Supplementary Information. All GO terms with a multiple testing (Benjamini) corrected $P<0.05$ are shown for the categories biological processes (BP), cellular compartment (CC) and molecular function (MF). Download excel file Supplementary Table S5.

Supplementary Table S6 *Bona fide* set of neoblast enriched genes (see main text). The table indicates the internal gene identifier, best protein BLAST hits to human and mouse and log_2-fold changes between the different fractions. The list is ordered by decreasing fold changes between X1 and Xins. Download excel file Supplementary Table S6.

Supplementary Table S7 Planarian homologs of genes associated with pluripotency control in human and mouse. **(A, B)** Genes required for maintenance (A) or repression (B) of pluripotency in mouse ESCs (Tang *et al*, 2010). **(C, D)** Regulators of OCT4 (C) and NANOG (D) expression in human ESCs (Chia *et al*, 2010). **(E, F, G)** Direct targets of Oct4 (E), Nanog (F) and Sox2 (D) identified by ChIP-Seq in mouse ESCs. The tables contain the planarian and the mammalian gene identifier as well as expression values for the cell fractions X1, X2 and Xins. Highly overlapping transcripts (90% of transcript length) mapping to the same mammalian homolog appear as entries in the same row. Download excel file Supplementary Table S7.

Supplementary Table S8 Best hits of HMM based search for homologs of human pluripotency factors **(A)** NANOG, **(B)** OCT4 and **(C)** SOX2 in the planarian proteome. The E-value for the full sequence HMM alignment is indicated. For the best matching domain, the E-value and the domain score are shown. These numbers were computed using the HMMER3 software suite with standard parameters. The putative homology was determined by manual inspection of HMM protein alignments combined with protein BLAST. E-values of the BLAST protein alignment are indicated in parentheses.

Supplementary References

Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, Dodt M, Mackowiak SD, Gogol-Doering A, Oenal P, Rybak A, Ross E, Sánchez Alvarado A, Kempa S, Dieterich C, Rajewsky N, Chen W (2011) De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Res* **21**:1193-1200

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402

Bachvarova RF, Masi T, Drum M, Parker N, Mason K, Patient R, Johnson AD (2004) Gene expression in the axolotl germ line: Axdazl, Axvh, Axoct-4, and Axkit. *Dev Dyn* **231**: 871-880

Biegert A, Mayer C, Remmert M, Söding J, Lupas AN (2006) The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res* **34**: W335-W339

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK *et al* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**: 1106-1117

Cebrià F, Newmark PA (2005) Planarian homologs of netrin and netrin receptor are required for proper regeneration of the central nervous system and the maintenance of nervous system architecture. Development 132: 3691-3703

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**: 1367-1372

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**: 1794-805

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**: W465-W469.

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**: 205-211

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **5**: 1792-1797

Eisenhoffer GT, Kang H, Sánchez Alvarado A (2008) Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian Schmidtea mediterranea. *Cell Stem Cell* **3**: 327-339

Fernandéz-Taboada E, Moritz S, Zeuschner D, Stehling M, Schöler HR, Saló E, Gentile L (2010) Smed-SmB, a member of the LSm protein superfamily, is essential for chromatoid body organization and planarian stem cell proliferation. *Development* **137**:1055-1065

Fraguas S, Barberán S, Cebrià F (2011) EGFR signaling regulates cell proliferation, differentiation and morphogenesis during planarian regeneration and homeostasis. *Dev Biol* **354**: 87-101

Garcia-Fernàndez J, Baguñà J, Saló E (1993) Genomic organization and expression of the planarian homeobox genes Dth-1 and Dth-2. *Development* **118**: 241-253

Guo T, Peters AH, Newmark PA (2006) A Bruno-like gene is required for stem cell maintenance in planarians. *Dev Cell* **11**: 159-169

Handberg-Thorsager M, Saló E (2007) The planarian nanos-like gene Smednos is expressed in germline and eye precursor cells during development and regeneration. *Dev Genes Evol* **217**: 403-411

Hayashi T, Asami M, Higuchi S, Shibata N, Agata K (2006) Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting. *Dev Growth Differ* **48**: 371-380

Hinkley CS, Martin JF, Leibham D, Perry M (1992) Sequential expression of multiple POU proteins during amphibian early development. *Mol Cell Biol* **12**: 638-649

Hinman VF, Nguyen A, Davidson EH (2007) Caught in the evolutionary act: precise cis-regulatory basis of difference in the organization of gene networks of sea stars and sea urchins. *Dev Biol* **312**: 584-595

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc* **4**: 44-57

Ito H, Saito Y, Watanabe K, Orii H (2001) Epimorphic regeneration of the distal part of the planarian pharynx. *Dev Genes Evol* **211**: 2-9

Kobayashi C, Watanabe K, Agata K (1999) The process of pharynx regeneration in planarians. *Dev Bio*l **211**: 27-38

Koinuma S, Umesono Y, Watanabe K, Agata K (2003) The expression of planarian brain factor homologs, DjFoxG and DjFoxD. *Gene Expr Patterns* **3**: 21-27

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760

Mannini L, Deri P, Gremigni V, Rossi L, Salvetti A, Batistoni R (2008) Two msh/msx-related genes, Djmsh1 and Djmsh2, contribute to the early blastema growth during planarian head regeneration. *Int J Dev Bio* **52**: 943-952

Millane RC, Kanska J, Duffy DJ, Seoighe C, Cunningham S, Plickert G, Frank U (2011) Induced stem cell neoplasia in a cnidarian by ectopic expression of a POU domain transcription factor. *Development* **138**: 2429-2439

Newmark PA, Sánchez Alvarado A (2002) Not your father's planarian: a classic model enters the era of functional genomics. *Nat Rev Genet* **3**: 210-219

Oviedo NJ, Levin M (2007) smedinx-11 is a planarian stem cell gap junction gene required for regeneration and homeostasis. *Development* **134**: 3121-3131

Palakodeti D, Smielewska M, Lu YC, Yeo GW, Graveley BR (2008) The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. *RNA* **14**: 1174-1186

Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**: S22-32

Rappsilber J, Ishihama Y, Mann M (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**: 663-670

Reddien PW, Bermange AL, Murfitt KJ, Jennings JR, Sánchez Alvarado A (2005a) Identification of genes needed for regeneration, stem cell function, and tissue homeostasis by systematic gene perturbation in planaria. *Dev Cell* **8**: 635-649

Reddien PW, Oviedo NJ, Jennings JR, Jenkin JC, Sánchez Alvarado A (2005b) SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science* **310**: 1327-1330

Robb SM, Ross E, Sánchez Alvarado A (2008) SmedGD: the Schmidtea mediterranea genome database. *Nucleic Acids Res* **36**: D599-606

Rossi L, Salvetti A, Marincola FM, Lena A, Deri P, Mannini L, Batistoni R, Wang E, Gremigni V (2007) Deciphering the molecular machinery of stem cells: a look at the neoblast gene expression profile. *Genome Biol* **8**: R62

Salvetti A, Rossi L, Lena A, Batistoni R, Deri P, Rainaldi G, Locci MT, Evangelista M, Gremigni V (2005) DjPum, a homologue of Drosophila Pumilio, is essential to planarian stem cell maintenance. *Development* **132**: 1863-1874

Salvetti A, Rossi L, Deri P, Batistoni R (2000) An MCM2-related gene is expressed in proliferating cells of intact and regenerating planarians. *Dev Dyn* **218**: 603-614

Schröder R. (2003) The genes orthodenticle and hunchback substitute for bicoid in the beetle Tribolium. Nature **422**: 621-625.

Scimone ML, Srivastava M, Bell GW, Reddien PW (2011) A regulatory program for excretory system regeneration in planarians. *Development* **138**: 4387-4398

Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* **1**: 2856-2860

Solana J, Lasko P, Romero R (2009) Spoltud-1 is a chromatoid body component required for planarian long-term stem cell self-renewal. *Dev Biol* **328**: 410-421

Solana J, Romero R (2009) SpolvlgA is a DDX3/PL10-related DEAD-box RNA helicase expressed in blastomeres and embryonic cells in planarian embryonic development. *Int J Biol Sci* **5**: 64-73

Takeda H, Matsuzaki T, Oki T, Miyagawa T, Amanuma H (1994) A novel POU domain gene, zebrafish pou2: expression and roles of two alternatively spliced twin products in early development. *Genes Dev* **8**: 45-59

Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA (2010) Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**: 468-478

Tazaki A, Gaudieri S, Ikeo K, Gojobori T, Watanabe K, Agata K (1999) Neural network in planarian revealed by an antibody against planarian synaptotagmin homologue. *Biochem Biophys Res Commun* **260**: 426-432

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105-1111

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals

unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515

Wagner DE, Ho JJ, Reddien PW (2012) Genetic Regulators of a Pluripotent Adult Stem Cell System in Planarians Identified by RNAi and Clonal Analysis. *Cell Stem Cell* **10**: 299-311

Wang Y, Zayas RM, Guo T, Newmark PA (2007) nanos function is essential for development and regeneration of planarian germ cells. *Proc Natl Acad Sci U S A* **104**: 5901-5906

Yoshida-Kashikawa M, Shibata N, Takechi K, Agata K (2007) DjCBC-1, a conserved DEAD box RNA helicase of the RCK/p54/Me31B family, is a component of RNA-protein complexes in planarian stem cells and neurons. *Dev Dyn* **236**: 3436-3450

Zayas RM, Cebrià F, Guo T, Feng J, Newmark PA (2010) The use of lectins as markers for differentiated secretory cells in planarians. *Dev Dyn. Nov* **239**: 2888-2897