# Table of Contents

## S1: Genomic DNA Isolation
- Genomic DNA preparation ................................................................. 3
- Mucus removal from planarian surface ..................................................... 4
- HMW DNA isolation for PacBio sequencing and Chicago libraries ................. 4
- Post purification of DNA using CTAB ....................................................... 5

## S2: PacBio Long Read Sequencing ......................................................... 5
- Large insert libraries ............................................................................ 6

## S3: Assembly Pipeline ........................................................................... 7
- MARVEL assembler ................................................................................ 7
- MARVEL - Determining read quality ...................................................... 8
- MARVEL - Annotation of repeat regions .................................................. 8
- MARVEL - Read correction .................................................................. 9
- MARVEL - Computational requirements .................................................. 9

## S4: Chicago/HiRise Scaffolding ............................................................... 10
- Chicago library preparation and sequencing .......................................... 10
- HiRise scaffolding of PacBio-MARVEL assembly ..................................... 10
- Analysis of HiRise changes to PacBio-MARVEL assembly ....................... 11
- Scaffold gap closure using PBjelly ......................................................... 11

## S5: Assembly Polishing ......................................................................... 11
- Final Error Correction with Quiver ......................................................... 11
- Redundancy filtering ............................................................................. 12
- Consensus polishing using Pilon ............................................................ 12

## S6: RNAseq and Transcriptomes .............................................................. 13
- RNA extraction and sequencing .............................................................. 13
- Smed_v6 transcriptome assembly: Raw data, assembly parameters, filtering, annotations ................................................................. 13
S6.3 Smes_v1 transcriptome assembly: Raw data, assembly parameters, filtering, annotations ................................................................. 13

S7: Assembly Quality Control .................................................................................................................. 14
S7.1 Transcriptome back-mapping ........................................................................................................ 14
S7.2 Genome completeness analysis ..................................................................................................... 14
S7.3 Transcript-based genome assembly quality control ...................................................................... 15
S7.4 Contamination check ................................................................................................................... 17

S8: UCSC Genome Browser ..................................................................................................................... 18
S8.1 GEM - Mappability track ............................................................................................................. 18
S8.2 RepeatMasker - Repbase track .................................................................................................. 18
S8.3 RepeatMasker - RepeatModeler track ....................................................................................... 19
S8.4 Alternative Scaffolds track ......................................................................................................... 19
S8.5 Quiver track ................................................................................................................................ 19
S8.6 PacBio coverage track ................................................................................................................. 19
S8.7 Transcripts track ......................................................................................................................... 19
S8.8 Links to PlanMine ........................................................................................................................ 20

S9: Repetitive Element Prediction ...................................................................................................... 20
S9.1 De novo prediction of repetitive elements ................................................................................... 20

S10: LTR Analysis .................................................................................................................................. 21
S10.1 LTR consensus construction ..................................................................................................... 21
S10.2 Gypsy LTR element annotation ............................................................................................... 22
S10.3 The Burro family ....................................................................................................................... 22
S10.4 LTR element expression analysis ............................................................................................ 23
S10.5 LTR element divergence analysis ............................................................................................. 23
S10.6 Solo/paired LTR element abundance analysis .......................................................................... 23
S10.7 Scaffold end analysis ................................................................................................................ 24

S11: AAT-Repeat Analysis ...................................................................................................................... 24
S11.1 AAT-microsatellite prediction and read alignment break analysis .............................................. 24
S11.2 Microsatellite repeat length estimation by Circular Consensus Sequencing (CCS) ........................ 24
S11.3 Overlap length variation in CCS and P6/C4 reads ...................................................................... 25
S11.4 Genomic context of AT-rich regions ......................................................................................... 26

S12: Assembly Heterozygosity ............................................................................................................... 26
S12.1 Drosophila assembly .................................................................................................................. 26
S12.2 Dot plot analysis ........................................................................................................................ 27
S12.3 Coverage analysis alt vs. main contigs .................................................................................... 27
S1: Genomic DNA Isolation

S1.1 Genomic DNA preparation

Our genomic DNA isolation protocol improves upon a previously developed method for DNA isolation procedure for planarians¹, making it compatible with single molecule real-time (SMRT sequencing)² on a PacBio RSII and the Chicago method³.
Important modifications include planarian external mucus removal, the choice of salt and alcohol for DNA precipitation enabling pigment removal and a post-purification step that removes likely remaining mucopolysaccharides, which tend to co-isolate with gDNA.

S1.2 Mucus removal from planarian surface
This step makes use the mucolytic agent N-acetyl-L-cysteine (NAC) to strip the animals of surface mucus, which can be excreted in copious amounts by large animals. Since NAC is acidic in aqueous solution and the highest mucolytic activity is observed between pH 7 to 9 the solution was neutralized, while the pH was monitored using the common pH indicator dye phenol red. The animals were washed in 10 ml freshly prepared NAC stripping solution (0.5 w/v N-acetyl-L-cysteine, 20 mM HEPES-NaOH, pH 7.25, 5 µl phenol-red solution (0.5 % w/v), pH to ~7 using 1 M NaOH) for 10-15 min at room temperature with strong agitation (e.g. rotator). The worms were rinsed briefly in distilled water prior to further processing.

S1.3 HMW DNA isolation for PacBio sequencing and Chicago libraries
All procedures were carried out at room temperature (RT) unless otherwise specified. For handling HMW DNA, only wide-bore tips were used and the sample was mixed only by careful inversion. For DNA isolation, 20 mucus stripped animals (~1 cm; starved for 1-3 weeks) were transferred to a 50 ml tube and residual liquid was removed. The animals were lysed in 15 ml of cold GTC buffer (4 M guanidinium thiocyanate, 25 mM sodium citrate, 0.5 % (w/v) N-Lauroylsarcosine, 7 % v/v β-mercaptoethanol) for 30 min on ice. The tube was carefully inverted every 10 min to help dissociating the tissue. Then, 1 volume of phenol/chloroform (Phenol/chloroform/isoamyl alcohol (25:24:1), 10 mM Tris (pH 8.0), 1 mM EDTA) was added to the lysate and mixed by carefully inverting the tube 10 to 15 times. Phases were separated by centrifugation for 20 min at 4,000 x g at 4 °C and the upper phase carefully transferred to a fresh tube. The phenol/chloroform extraction was repeated 1-2 times, until the interphase disappeared. Residual phenol was removed by extracting the aqueous phase once with 1 volume of chloroform and mixing by inversion 10 to 15 times and another round of centrifugation. The upper aqueous
phase was transferred to a fresh tube and 1 volume of ice-cold 5 M NaCl was added and mixed well by inverting the tube several times. The tube was placed on ice for 15 min to salt out additional contaminants, which were pelleted by centrifugation for 10 min at maximum speed at 4 °C. The supernatant containing the nucleic acids was carefully decanted or pipetted off to a fresh tube without disturbing the pellet. The nucleic acids were precipitated using 0.7-1 volumes of isopropanol and centrifugation at 2,000 x g at 30-45 min at RT. The pellet was carefully washed with 70 % ethanol and spun at 2,000 x g at 5 min at RT and briefly air-dried and re-suspended in 50 µl TE buffer over night at 4 °C.

S1.4 Post purification of DNA using CTAB
All procedures were carried out at room temperature, as cetyltrimethylammonium bromide (CTAB) precipitates at temperatures lower than 15 °C. This step allowed the removal of contaminants (e.g. polysaccharides) from DNA preparations by complex formation through tuning the NaCl concentration

The isolated and re-suspended nucleic acids were treated with 1 µl of 4 mg/ml RNase A and incubated for 1 h / 37 °C to remove RNA. The NaCl concentration was adjusted by adding 0.3 volumes of 5 M NaCl and the volume was adjusted to 500 µl with 2 % CTAB (w/v) / 1.4 M NaCl solution. After careful and thorough mixing by inversion, 1 vol of chloroform was added and carefully mixed by inversion until the sample turned milky. The phases were separated by centrifugation for 15 min at 12,000 – 16,000 x g at RT. The upper clear phase was recovered without disturbing the white interphase using a wide bore pipette tip and transferred to a new tube. The DNA was precipitated using 0.7-1 volumes of isopropanol and centrifugation at 2,000 x g at 30-45 min at RT. The pellet was carefully washed with 70 % ethanol and spun at 2,000 x g at 5 min at RT and briefly air-dried and re-suspended in 50 µl TE buffer over night at 4 °C. Quality of the DNA was checked by agarose gel electrophoresis and quantified using the Qubit™ fluorometer (dsDNA BR assay).

S2: PacBio Long Read Sequencing
Quality control of the high molecular weight genomic DNA was performed by spectrophotometric determination of 260/280 and 260/230 ratios making use of the
Nanodrop device (Thermo Scientific). The 260/280 ratios were in the range of 1.66 to 2.08 and 260/230 ratios between 1.84 and 2.01. The concentration was determined in triplicates via a Qubit™ fluorometer with the high sensitivity standard. Integrity of the HMW gDNA was determined by pulse field gel electrophoresis run using the Pippin Pulse™ device (SAGE Science). Before starting any PacBio library preparation, planarian gDNA samples were purified with 1X AMPure XP beads. Note that the AMPure method cannot recover planarian DNA out of crude lysates, possibly due to resorptive competition of the abundant contaminants.

S2.1 Large insert libraries
To prepare large insert libraries, AMPure purified genomic DNA was fragmented to sizes of 12-30 kbp with the Megaruptor device (Diagenode). For shearing, either 20 kbp shearing settings for medium insert SMRT bell libraries or 35 kbp shearing settings for large insert SMRT bell libraries were used. PacBio SMRT bell libraries were prepared as described in the company-supplied “Procedure and Checklist – 20 kbp Template Preparation Using BluePippin™ Size-Selection System”. SMRT bell libraries were size selected with the BluePippin™ system (Sage Science) with a minimum fragment length cutoff of 7.5 and 10 kbp for medium insert libraries and at 15 and 17.5 kbp for large insert libraries.

PacBio RSII SMRT cells were loaded with SMRT bell libraries after primer annealing and polymerase binding using MagBead loading. A total of 197 SMRT cells loaded with the medium insert libraries were sequenced on the PacBio RSII instrument making use of the P4 polymerase and the C2 sequencing chemistry (P4/C2). In addition, a total of 38 SMRT cells loaded with large insert libraries were sequenced on the RSII instrument with the P6 polymerase and C4 sequencing chemistry (P6/C4). Movie times for all P4/C2 SMRT cells were 240 minutes. Movie times for P6/C4 SMRT cells were 240 min (20 SMRT cells) and 360 minutes (23 SMRT cells).
S3: Assembly Pipeline

3.1 MARVEL assembler
At the time of the initial sequencing of Smed with the P4/C2 chemistry, the standard workflow for assembling noisy long-reads from PacBio was to clean the reads, either with Illumina sequencing data or by reducing the PacBio native error rate of 12-15% to under 1% by high coverage sequencing (PacBioToCA). After this correction step, the original Celera assembler\(^6\) was used to assemble the long-reads. However, initial cleaning of the reads can have one of two disadvantages, especially in a genome with a high repeat content such as Smed. Either the repeats within the reads are not cleaned and the assembly of such regions by the Celera assembler is hindered due to a an error rate higher than expected (< 1%) or the repeats are incorrectly cleaned due to the ambiguous character of overlaps between repeats. Both approaches lead to incorrect and fragmented assemblies with possible misjoins. Therefore, MARVEL was developed to deal with notoriously complex genomes that contain a high fraction of repetitive genomic information and are sequenced with noisy long-read technologies (Novojilov et al., in preparation). For comparisons with the Canu assembler\(^7\) (see Table 1, main text), we used Canu v1.3 with standard parameters and the same uncorrected PacBio reads as for the MARVEL assembly.

The MARVEL assembler currently consists of three major phases namely the setup phase, patch phase and the assembly phase. In the setup phase, depending on the sequencing technology, the reads are extracted from the raw sequencing data and saved in an internal database. The patch phase detects and corrects read artefacts including untrimmed adapters, polymerase strand jumps, and ligation chimers that are the primary impediments to long contiguous assemblies, which are then used for the final assembly phase. The assembly phase stitches short alignment artefacts resulting from bad sequencing segments within overlapping read pairs. The default parameter is generally 50 bp in length, but for Smed 300 bp were used. This step is followed by repeat annotation and the generation of the overlap graph, which is subsequently toured in order to generate the final contigs.
A genome with a skewed GC-content, such as the Smed genome, has a different k-mer distribution than those with a uniform distribution of the GC content. This can be used to weigh certain k-mers and their expected distribution within the genome. Therefore, the AT bias option (-b) was used in DALIGNER* (the noisy read aligner used in MARVEL to find the overlaps between reads) as well as in DBdust, to avoid masking AT-rich regions as low complexity and thereby excluding them from the k-mer seeding.

In addition, alignments were forced through variable AT microsatellites (< 300 bp) despite often significant length discrepancies and non-uniformity of their instances in the reads. This allowed for a more contiguous assembly. The criteria for repeat module identification in overlaps was relaxed, allowing for better recognition of repeat modules and therefore resulting in improved resolution of repetitive elements.

Since the Smed dataset combined sequencing data generated by two different chemistries (P4/C2 and P6/C4), each chemistry batch was patched individually prior to combining the data sets as input for the assembly phase. In the assembly phase, the overlap of the combined dataset was calculated and assembled. After all tours were calculated, the contigs were corrected and a consensus sequence was generated by remapping the patched reads.

3.2 MARVEL - Determining read quality
In order to determine the quality of a read, the consensus of the aligned supporting reads were used. MARVEL does not determine the quality of individual bases, but rather the quality of the subread within the trace points (usually consisting of 100 bp). The end result of the process is the determination of a consensus base supported by the aligned read overlaps between the trace points.

3.3 MARVEL - Annotation of repeat regions
The repeat regions were initially annotated by the masking server and later during the scrubbing phase by DBrepeat. While both use the alignments to determine repeat regions, the masking server annotates the repeat region dynamically and once the
repeat coverage threshold is exceeded (in most cases twice the sequencing coverage) the region is masked and not used for subsequent k-mer seeding.

DBRepeat runs sequentially over the reads to annotate the repeat regions. Once a base with a coverage that exceeds the repeat coverage is identified, a repeat region is opened. A repeat region is closed once a base at or below the non-repeat coverage is found. The minimum overlap length of 1 kbp results in complete lack of repeat annotation for the 1 kbp regions at the start and end of the reads. DBhomogenize can be used to transitively transfer an existing repeat annotation based on the computed alignments, thereby properly annotating the tips of the reads.

3.4 MARVEL - Read correction

Despite the noisy nature of PacBio reads, MARVEL does not clean the reads to an error-rate of less than 1% as is common practice of other assemblers when dealing with noisy long-reads. The reads are first patched after the initial overlap phase, removing the major sequencing related insertions/deletions (indels), chimeras and additional artefacts, which is necessary for quality overlaps that will be used for the assembly. After a second round of overlaps, MARVEL improves the alignments that are corrupt due to short sequencing errors. Finally, Quiver can be used after the assembly phase in order to generate a high-quality consensus sequence, but this often results in long runtimes, especially with large genomes. This can be improved by circumventing the use of BLASR by transforming the previously calculated read overlaps into the position of the reads within the contigs.

3.5 MARVEL - Computational requirements

For anything beyond bacterial samples, which can be assembled rather quickly on a standard laptop, a cluster environment with sufficient CPU and storage is recommended. We recommend having at least 8 GB of RAM per CPU core for the cluster nodes and a fast interconnect to the cluster-attached storage. RAM per core demand can be lessened by further increasing the partitioning of the data into blocks, at the price of increased IO demand. Example CPU times can be found in Supplementary Table 1 and have been measured on our Sandy-Bridge and Haswell architecture based cluster. Storage requirements are highly dependent on coverage and the repetitiveness of the genome. Also, storage requirements for a single phase
of MARVEL can be found in Supplementary Table 1. For peak storage requirements, this number needs to be multiplied by two.

**S4: Chicago/HiRise Scaffolding**

**S4.1 Chicago library preparation and sequencing**

A Chicago library was prepared as described previously\(^3\). Briefly, ~500 ng of HMW gDNA (>150 kbp mean fragment size) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was then digested with DpnII, the 5’ overhangs were filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was sheared to a mean fragment size of ~350 bp, and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were then isolated using streptavidin beads before PCR enrichment of the library. The libraries were sequenced on an Illumina HiSeq 2500 in rapid run mode to produce 286 million 2x100 bp read pairs, amounting to 93x physical coverage (1-50 kbp spanning pairs) of the *Smed* genome.

**S4.2 HiRise scaffolding of PacBio-MARVEL assembly**

A draft genome assembly of *Smed* (PacBio-MARVEL, 782.1 Mbp with a contig N50 of 708.7 kbp), Illumina shotgun sequence data (Illumina HiSeq 2500 rapid run mode, 2x150 bp read pairs, SRR5408395), and Chicago library read pairs in FASTQ format were used as input data for HiRise, a software pipeline designed specifically for using Chicago data for error-correcting and scaffolding genome assemblies\(^3\). Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu/). The separations of Chicago read pairs mapped within draft scaffolds were analysed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative misjoins and score prospective joins. After scaffolding, shotgun sequences were used to close gaps between contigs.
**S4.3 Analysis of HiRise changes to PacBio-MARVEL assembly**

MARVEL raw contigs were mapped onto HiRise scaffolds with DALIGNER to find the exact break positions within the raw contigs. Then, each position was visually inspected using the layout of the toured overlap graph (similar to Figure 2e, left graph, main text), aided by the direct visualization of read overlaps with the MARVEL tool LAexplorer.

**S4.4 Scaffold gap closure using PBjelly**

Following HiRise scaffolding, we used PBjelly\(^\text{10}\) to fill or reduce as many scaffolding gaps as possible. All 632 HiRise scaffolds were used as input for PBjelly, as well as all patched P4/C2 and P6/C4 chemistry reads longer than 8 kbp. The 632 scaffolds contained 1,256 gaps of 100 bp (HiRise default gap size). PBjelly completely closed 482 gaps, 176 gaps were closed partially and overfilling 176 gaps, meaning the gap was actually larger than 100 bp. In addition to this, it was possible for PBjelly to join 32 scaffolds into bigger scaffolds, reducing the overall scaffold number to 600.

Overlap analysis between the scaffolds using DALIGNER (see below) showed that 119 PBjelly scaffolds were almost completely contained within other scaffolds and consisted of >95 % repeat bases. These were therefore moved into the alternative contig set (_alt, see below), leaving 481 scaffolds as the final haploid dd_Smed_g4 assembly.

**S5: Assembly Polishing**

**S5.1 Final Error Correction with Quiver**

The final consensus sequence was generated first by a round of the PacBio tool Quiver. Due to the runtime of BLASR\(^\text{9}\), which is used by Quiver, only reads that were part of the overlap graph of each contig (i.e. the true non-repeat induced overlaps) were used to polish the contig sequences. This step had mostly to do with the AT-bias and high repeat content of the Smed genome. In order to enrich the coverage and improve the final polished sequence, additional subreads were extracted from the Zero Mode Waveguides and used for polishing (if available). This brought the coverage to approximately 40x.
**S5.2 Redundancy filtering**

DALIGNER was used to identify the haplotype content of the *Smed* PacBio-MARVEL assembly, by mapping of the contigs against each other. If at least 70% of the bases in the two contigs as well as 70% of the reads used to build the contigs were part of both contigs, they were considered alternate haplotypes and separated from the main assembly into a second set of alternative contigs, designated by a `_alt` suffix to the contig name. For contigs shorter than 100 kbp that terminated in long repetitive sequences, these thresholds were reduced to 50%. This removed 1,509 alternative contigs with a total length of 168.1 Mb (note that this step was performed prior to scaffolding).

**S5.3 Consensus polishing using Pilon**

For fine-polishing of residual PacBio base pair inaccuracies, we used Pilon\(^\text{11}\). This tool uses the evidence from short-read sequencing data in order to correct single base differences, small indels, block substitution events, and gaps. Illumina 2x150 bp shotgun HiSeq 2500 paired-end sequencing data (SRR5408395; see S4.2 “HiRise scaffolding of PacBio-MARVEL assembly”) was aligned with bowtie2\(^\text{12}\) against the scaffolds of dd_Smed_g4 and the corresponding alignments were provided as input to Pilon. Since Pilon was designed mainly for smaller prokaryotic genomes, we ran it over each of the assembled scaffolds separately. To assess the assembly improvements, variants were called from the scaffolding PacBio data provided by Dovetail Genomics with samtools/bcftools before and after Pilon correction. Correction was assessed by counting InDels and SNPs per kbp, and independent rates were calculated for exons, genic regions and polished (dd_Smed_g4) vs. raw assembly (Pacbio-MARVEL). After Pilon polishing there was a 2.24-fold reduction of indels in intergenic, a 3.25-fold reduction in exonic and a 1.94-fold reduction in genic regions. The changes in SNP counts were minor, indicating that the detected differences correspond mostly to true variants.
S6: RNAseq and Transcriptomes

S6.1 RNA extraction and sequencing
For total RNA isolation, 30 sexual animals (~1 cm, 2 weeks starved) of the inbred genome sequencing Smed strain were homogenized in 20 ml Trizol reagent (Invitrogen) on ice using a Miccra D-8 homogenizer (Miccra) and processed using Trizol reagent protocol. Following phase separation by chloroform addition and isopropanol precipitation, the RNA pellet was dissolved in 350 µl RLT buffer and purified according to the RNeasy Mini Kit, including on-column DNase I digestion. The RNA was eluted in nuclease-free water and stored at -80 °C. The total RNA was quantified using a NanoDrop 1000 spectrophotometer and RNA integrity was checked by capillary electrophoresis on a Bioanalyzer using the RNA 6000 Nano Kit. The sample was poly-A selected and sequenced on an Illumina NextSeq 500 sequencer.

S6.2 Smed_v6 transcriptome assembly: Raw data, assembly parameters, filtering, annotations
The transcriptome assembly of the asexual strain has been published before. Briefly, we used Trinity to assemble RNA-Seq data from a mixed set of publically available data sources (Dresden, Berlin, Muenster). Transcripts were filtered for contaminants, directionality-corrected and subsequently annotated and filtered by ORF-content, domains, and RefSeq-protein BLAST hits to define a likely Protein Coding Fraction (PCF). For many applications, we further sub-filter the longest isoform of a Trinity graph component. Such uses of the transcriptome are indicated by a “PCFL” suffix, e.g. “Smed_v6.PCFL”. Details about the dataflow and the companion website to access the data are documented in the PlanMine publication.

S6.3 Smes_v1 transcriptome assembly: Raw data, assembly parameters, filtering, annotations
The transcriptome assembly Smes_v1 of the sexual strain of Smed was assembled from SRR955511 using the same pipeline as for Smed_v6, except that the strand
correction step and ordering of transcript numbers according to expression level were omitted. ORF directionality is therefore randomized in the Smes_v1 version. The PCFL-suffix, e.g., “Smes_v1.PCFL” designates the longest graph component subset of the protein coding fraction, as defined above.

**S7: Assembly Quality Control**

**S7.1 Transcriptome back-mapping**

*De novo* assembled transcriptomes of the asexual (Smed_v6) and sexual strains (Smes_v1) available at the PlanMine\textsuperscript{13} resource were used to assess the completeness of the dd_Smed_g4 assembly. For this purpose, the longest isoform per Trinity graph component of the protein-coding fraction (PCFL) was selected and mapped to the dd_Smed_g4 g assembly using the GMAP aligner\textsuperscript{15}, with a minimum query coverage cut-off of 0.6 and a minimum identity cut-off of 0.6. The resulting mappings were then used to investigate uniquely mapping, multi-mapping, and non-mapping transcripts, along with putative chimeric transcripts.

**S7.2 Genome completeness analysis**

The two back-mapped transcriptomes were used to estimate dd_Smed_g4 assembly completeness. The fact that 98.7 % of transcripts of the sexual (Smes_v1) and 97.3 % of the asexual (Smed_v6) transcriptome mapped to the dd_Smed_g4 assembly under our mapping criteria demonstrates that the genome is largely complete. The remaining non-mapping transcripts (538 or 667 in the Smes_v1 and Smed_v6 transcriptomes, respectively) could either represent true positives, i.e. genes genuinely missing from the dd_Smed_g4 assembly. Alternatively, they could represent transcriptome contaminants (e.g., from commensal organisms or other sources of sequencing noise) and thus false positives. To distinguish between these two possibilities, we BLAST searched (blastp) all non-mapping Smes_v1 and Smed_v6 transcript ORFs against NCBI RefSeq and recorded the species from which the best blastp hit of each transcript originated. We then checked if any species was specifically over-represented for non-mapping transcripts using the enricher function from the ClusterProfiler R package\textsuperscript{16}. In this regard we found overrepresentation of
the protist species *Tetrahymena thermophila, Ichthyophthirius multifilis,* and *Paramecium tetraurelia* in Smes_v1, whilst *Caenorhabditis elegans* and *Bos taurus* were over-represented in Smed_v6. After the annotation of these transcripts as likely contaminants, we further classified the remaining genes as likely missing genes if they uniquely mapped to the SmedSxl v4.0\(^1\) assembly and were annotated in PlanMine\(^1\) as having orthologues in at least 5 other planarian species. The remaining transcripts that did not fit the contaminant or likely missing gene criteria were marked as “unknown” and are likely enriched for assembly noise. Altogether, our analysis identified 46 out of a total of 31,966 *Smed* transcripts as genuinely missing from the dd_Smed_g4 assembly, amounting to a completeness of 99.9985%. Further, all missing genes could be identified on small contigs that had been removed by our initial assembly filtering. We therefore conclude that based on the mapping of previous transcriptome assemblies, the genome is of high quality. In contrast, if we consider Smes_v1 transcripts that map to our assembly (dd_Smed_g4), but do not map to the published Sanger genome assembly (SmedSxl v4.0)\(^1\) then we identify 2,193 such transcripts of which 1,229 also were annotated in PlanMine as having homologues in at least 5 other planarian species. Overall, this analysis therefore demonstrates conclusively that the dd_Smed_g4 assembly is the most complete *Smed* genome assembly to date.

**S7.3 Transcript-based genome assembly quality control**

We also used transcriptome backmapping to further probe the quality of the genome assembly. Since an entire *de novo* assembled transcriptome cannot provide an error-free “ground truth” data set, we first sub-selected a high confidence set of *Smed* transcripts from our Smes_v1 *de novo* transcriptome assembly on PlanMine\(^1\). Specifically, we aligned primary ORFs (i.e. the longest non-overlapping open reading frame of each transcript) against the 7 other in-house assembled planarian proteomes currently available in PlanMine. The search was performed with blastp using the BLOSUM45 similarity matrix and an e-value cut-off of 1E-3. Blastp results were filtered for subject and query coverage of at least 90% in at least 5 other transcriptomes. Altogether, this analysis identified 1,509 *Smed* ORFs that had clear
full-length homologues in multiple other planarian species and therefore represented high confidence cDNAs.

To stringently probe the quality of the dd_Smed_g4 assembly using the high confidence cDNA (HC-cDNA) subset, we further sub-filtered the previous GMAP transcriptome mapping results (see section S7.1) for a minimum query coverage of 0.9 and a minimum identity cut-off of 0.9 and categorized the HC-cDNA as uniquely mapping, non-mapping, multi-mapping. As shown in Extended Data Fig. 2a, 1,432 out of the 1,509 (94.90 %) transcripts mapped uniquely to the dd_Smed_g4 assembly, despite the high stringency and coverage filter criteria. Therefore, this provides a strong indication that the underlying gene sequences were correctly assembled. Of the 10 non-mapping transcripts (0.66 %), only 1 represented a genuine assembly mistake located at a contig end, whereby 4 of the 5'-exons were inverted and appended to the 3'-end of the gene (Extended Data Fig. 2 b,c). Also, 2 of the non-mapping transcripts represented GMAP false positives (both were BLAT-mappable with >0.9 query coverage and identity) and the other 7 were transcripts already identified as “missing gene (4)”, “unknown (2)” or “putative contaminant (1)”. Together, the high confidence gene analysis therefore confirms both the high assembly quality and completeness of the dd_Smed_g4 assembly.

Furthermore, 67 out of the 1,509 HC-cDNA (4.44 %) were classified as multi-mapping. Manual inspection demonstrated that all of these genes indeed had multiple (usually 2) “perfect” map locations in the dd_Smed_g4 assembly. Some, e.g. dd_Smes_v1_25458 (Extended Data Fig. 2d) mapped on opposite strands within non-repetitive and high-quality regions of the genome, thus representing likely biological gene duplications. However, other multi-mapping locations like that of dd_Smes_v1_28816, were situated in low-confidence regions close to assembly gaps (identified by bedtools closest\textsuperscript{18}), indicated by lack of a respective Quiver track in that region (Extended Data Fig. 2e). Indeed, we found that the 67 multi-mapping transcripts mapped close to contig ends in comparison with uniquely mapping high confidence genes (Extended Data Fig. 2 f,g) and the size of the duplicated genomic region was below 5 kbp in all cases. The size of the duplicated regions was determined by mapping transcripts containing intronic sequences to the dd_Smed_g4 assembly using GMAP with parameters --chimera-margin=200, --min-
trimmed-coverage=0.9 and --min-identity=0.9. If the transcript with intronic sequences still mapped to more than one locus in the genome, a flanking window of 2 kbp (1 kbp up- and 1 kbp downstream) was added to the transcript sequence and GMAP was run again. This process was repeated until only one location was found for a given transcript.

Altogether, this indicates that the majority of multi-mapping transcripts represent over-assembly artefacts in the dd_Smed_g4 assembly that are too small to be detected by the Chicago/HiRise scaffolding method. Likely, these contig end mistakes arise as a consequence of duplicated assembly of haplotypes and represent a known problem in assembling highly heterozygous genomes\textsuperscript{19}.

However, their low overall proportion minimally impact the quality of the dd_Smed_g4 assembly and the various quality control tracks provided in the PlanMine genome browser (e.g., gap locations, mappability, quivered intervals) provide tools for the interpretation of dubious cases.

Inter and intra-contig similarity in Extended Data Figure 2c-e was visualized using mirropeats\textsuperscript{20} with the described thresholds.

### S7.4 Contamination check

Although all animals used for genome sequencing were exposed to a stringent antibiotic treatment prior to DNA extraction, the sequencing of whole animal extracts always comes at the risk of co-sequencing parasitic or commensal organisms. To assay for the putative presence of non-Smed contigs in our dd_Smed_g4 assembly, we screened all contigs for GC-content and coverage (this method was recently used to identify contamination in the original \textit{Hypsibius dujardini} Tardigrade genome assembly)\textsuperscript{21}. Coverage of dd_Smed_g4 contigs was computed by first aligning reads from SRR954965 and SRR849810 to dd_Smed_g4 using bowtie2\textsuperscript{12}, and estimating coverage per scaffold with genomeCoverageBed\textsuperscript{18}. GC content was determined using infoseq from the EMBOSS tools suite\textsuperscript{22}. We could not detect major contig clusters that deviated from the Smed GC-content or haploid/diploid read coverage, thus indicating lack of bacterial or other contamination in the assembly (not shown).
S8: UCSC Genome Browser

To facilitate access to the dd_Smed_g4 assembly, we built a custom version of the “UCSC Genome Browser”. Further, we integrated the browser into PlanMine, the planarian transcriptome platform maintained by our group\textsuperscript{13}. The genome enhancement thus further improves the utility of PlanMine towards its mission objective as community platform for the planarian model system. The UCSC Genome Browser\textsuperscript{23} source code was downloaded from GitHub (https://github.com/ucscGenomeBrowser/kent) and several tracks were made available, which are briefly described in the following.

S8.1 GEM - Mappability track

Overall genome mappability was determined by the GEM mappability program\textsuperscript{24} (GEM-binaries-Linux-x86_64-core_i3-20130406-045632). Briefly, using the fast mapping-based algorithm, it produces mappability scores for different lengths of NGS reads on the genome. A higher GEM-score at a given position indicates a higher chance of uniquely mapping a read of the respective length to that position of the genome. Regions with lower score are less unique and will likely produce ambiguous mappings for the given read length. Generally, duplicated and repetitive regions generate lower mappability scores. One exception are repetitive regions in the assembly that were not error-corrected by the PacBio Quiver tool (see S8.5). These regions appear more unique, due to the higher error rate and tend to have high mappability scores, but can easily be spotted by the absence of a Quiver track.

To generate the mappability track, standard parameters along with a read length of 200 were applied. After that, the output files were converted to bigwig using standard parameters. The “mappability_200” track is available under “Mapping and Sequencing Tracks” at the custom UCSC Genome Browser.

S8.2 RepeatMasker – Repbase track

In order to provide information about known repeats (i.e. RepBase) a RepeatMasker run using parameters: -species “schmidtea Mediterranea” and -s (for sensitive) was performed on the dd_Smed_g4 assembly. Also, RepeatMasker\textsuperscript{25} was configured using rmblastn (v2.2.28+) and TRF (v4.09)\textsuperscript{26}. The output information is available at
the custom UCSC Genome Browser under “Variation and Repeats” section and track named “RepeatMasker Repbase”.

S8.3 RepeatMasker – RepeatModeler track

In order to also provide de novo repeat prediction, RepeatModeler (v1.0.8)\textsuperscript{27} was run on the dd_Smed_g4 assembly. Standard parameters were used and the tool was configured with RepeatMasker patched version of RECON (v1.08)\textsuperscript{28}, RepeatScout (v1.0.5)\textsuperscript{29}, TRF (v4.09)\textsuperscript{26}, rmblastn (v2.2.28+). The resulting track is available at “Variation and Repeats” section under “RepeatMasker - denovo”.

S8.4 Alternative Scaffolds track

The alternative scaffolds track highlights genomic regions for which an alternative haplotype exists. Each alternative haplotype (indicated with “_alt” suffix) is available in the browser (described in detail in section S12.3).

S8.5 Quiver track

Quiver status, described in details in section S5.1, is represented genome wide on this track. Lack of Quiver coverage indicates that a region was not error corrected and that the underlying sequence is to be interpreted with caution.

S8.6 PacBio coverage track

PacBio raw data (P4/C2 and P6/C4) were mapped to dd_Smed_g4 using bwa-mem\textsuperscript{30} with -x pacbio option. Using samtools\textsuperscript{31} and BEDTools – genomeCoverage\textsuperscript{18}, coverage information was generated and displayed on this track.

S8.7 Transcripts track

PlanMine\textsuperscript{13} Smed assembled transcripts (Smes_v1 and Smes_v6) were mapped to dd_Smed_g4 assembly using GMAP (described in section S7.1).

Coverage information was fetched from mapping raw data used for transcriptomes assemblies from PlanMine to dd_Smed_g4 using STAR\textsuperscript{32} with default parameters, samtools\textsuperscript{31} and BEDTools GenomeCoverage\textsuperscript{18} were used to build the track.
S8.8 Links to PlanMine

The custom UCSC Genome Browser version is embedded in PlanMine available at http://planmine.mpi-cbg.de. Also, each Smed transcript is linked between PlanMine and the UCSC Genome Browser.

S9: Repetitive Element Prediction

S9.1 De novo prediction of repetitive elements

An initial prediction of repetitive element classes was performed using RepeatModeler (v1.0.8)\(^{27}\) and the resulting library was classified using RepeatMasker (v4.0.6)\(^{25}\) using libraries from RepBase (20160829 snapshot)\(^{33}\). The RepeatModeler library contained large numbers of models that could not be classified (classification ‘Unknown’) and which, together, masked 25.29 % of the assembly. Upon manual inspection, high proportions of the models were chimeric or contained elements on varying strands. Removal of these models from the library increased masking ascribed to other repeat classes, also suggesting conflict between the models produced by RepeatModeler.

A library was manually created to ameliorate these issues by compiling repeat element classes from different sources, or from RepBase where no differences were noted between masking with RepBase entries in comparison to generated models. Simple repeats and DNA elements were taken directly from RepBase using the RepeatMasker “queryRepeatsDatabase.pl” script. Masking of LINE elements with the RepBase entries was poor in comparison to using models produced by RepeatModeler (4.39 % versus 10.98 % genome masking, respectively) and the latter were included in the library. LTR retroelement models were produced through a separate pipeline (see S10.1 “LTR element consensus construction”).

A hard-masked assembly was generated using this library and this assembly was again used as input into RepeatModeler. Masking from models classified as ‘Unknown’ was reduced to 14.39 %. Again, manual inspection revealed these contained high proportions of fragmentary and chimeric models (1,502 of 1,768 models) and these were excluded from downstream analysis.
For all masking, RepeatMasker was configured with rmblastn (v2.2.27+) and TRF (v4.09)\textsuperscript{26} and was run in sensitive (-s) mode.

**S10: LTR Analysis**

**S10.1 LTR consensus construction**

LTR elements were identified through an approach combining outputs from 4 prediction programs: LTR\_FINDER (v1.0.6)\textsuperscript{34} (run with and without inbuilt masking of highly repetitive sequences), LTRharvest (GenomeTools v1.5.9)\textsuperscript{35}, MGEScan-LTR\textsuperscript{36}, and Red\textsuperscript{37}. Where programs expose the options, maximum element size constraints were set to 50 kbp and minimum LTR sizes to 150 bp. As this was program-dependent, however, all identified features were further filtered on the presence of >150 bp LTRs and to be between 2 and 50 kbp in length. Passing features overlapping >=50 % were merged using BEDOPS (v2.4.20)\textsuperscript{38} to give the locations of final predictions.

Predicted elements were translated in all frames and scanned using hmmsearch (HMMER v3.1b2)\textsuperscript{39} for the 'rve' (PF00665.21), 'RVP' (PF00077.15), and 'RVT_1' (PF00078.22) Pfam HMMs\textsuperscript{40}, representing integrase (IN), protease (PR), and reverse transcriptase (RT), respectively. Matching regions were extracted for all elements and separate alignments built for IN, PR, and RT using MUSCLE (v3.8.31)\textsuperscript{41}. For elements containing all three HMM features, alignments were concatenated and a neighbour-joining tree built with MUSCLE. Clusters were generated using tree2clusters (USEARCH v9.0.2132)\textsuperscript{42} using a minimum identity threshold of 80 %. Not all elements contained all three HMM features and were represented in the tree, thus clusters were manually refined based on the individual trees to maximise inclusion and to exclude elements with sizes >=2 SDs from the mean.

Nucleotide sequences for all elements of each cluster were extracted from the assembly and aligned with MAFFT (v7.271)\textsuperscript{43}. Neighbour-joining trees were built using MUSCLE and any sub-clusters identified using USEARCH were separated. Levitsky consensus sequences were calculated and annotated using an automated pipeline within UGENE (v1.24.2)\textsuperscript{44}. Concatenated IN, PR, and RT sequences from the consensuses were added to the previous alignment and the tree inspected to ensure
representation of all elements. Finally, this was confirmed by using the consensus sequences as a custom library to mask the assembly using RepeatMasker and by comparison to the initial predictions.

To streamline further analysis, the LTRs were replaced with consensus sequences formed for each of the (sub-) clusters individually. To match the methods and nomenclature used in RepBase, the LTR and internal sequences were separated for the RepeatMasker library.

### S10.2 Gypsy LTR element annotation

Consensus elements were annotated with the locations associated with Pfam models used in their discovery (‘rve’, ‘RVP’, and ‘RVT_1’) and additionally with ‘RNase_H’ (PF00075.19). tRNA models predicted by tRNAscan-SE (v1.4) were scanned against the consensus elements using blastn (v2.2.30+) and the best matching region annotated. HMMs compiled from the GyDB 2.0 collection were also tested, alongside the entire PfamA and PfamB HMM sets. CCHC-type Zinc finger annotations, known to be associated with retroviral gag nucleocapsid (NC) proteins could be identified, however further hits were of low confidence and were omitted. Additional putative domains were predicted using ScanProsite and CD-search.

### S10.3 The Burro family

We detected 14 classes of gypsy LTR elements in the *Smed* genome, all of which share the central protease (PR), reverse transcriptase (RT), ribonuclease H (RNAse) and integrase (INT) module of Ty3-gypsy-like retroelements. However, a phylogenetic analysis of the conserved PR/RT/Int/RNAse module could not assign any of the 14 classes to known retroelement classes. The most striking structural feature of the *Smed* retroelements is the massive sequence expansion of groups 4 (Burro-1), 5 (Burro-2) and 13 (Burro-3), which results in LTR elements of 25 kbp flanked by 5 kbp LTRs (compared to the 5-7 kbp of typical vertebrate LTR elements). The Ogre-elements found in plants so far provide the only example of similarly-sized LTRs, but priming by tRNA_{PRO} instead of tRNA_{ARG} in OGREs and the lack of an open reading frame upstream of the likely nucleocapsid protein (zinc knuckle domains) indicate that the *Smed* elements are likely not related to OGRE elements. Further, the presence of putative nuclease and BIR-domains that are not part of the typical
domain complement of Metaviridae further indicate that the giant Smed gypsy LTR elements represent likely novel branches of the Metaviridae, the unusual complexity of which harbours the potential for complex life cycles or interactions with the host transposon defence machinery.

S10.4 LTR element expression analysis

LTR element expression was estimated by TETranscripts\textsuperscript{51}. Briefly, TETranscripts estimates both, the expression of transposable elements (TE) and gene transcripts taking as inputs TE and gene annotations and alignment files (i.e. bam files). RNA-Seq reads (SRR5408394) were mapped to the genome using STAR (v2.5.2a_modified)\textsuperscript{32} with --winAnchorMultimapNmax 100, --outFilterMultimapNmax 100. A custom GTF containing coordinates of Smed_v6 transcripts that uniquely mapped to the genome by GMAP\textsuperscript{15} (--chimera-margin=200, --min-trimmed-coverage=0.6 and --min-identity=0.6) in addition to the genome were used as input to build STAR index. After the read alignment, a table with raw counts was generated using TETranscripts (default parameters).

S10.5 LTR element divergence analysis

Kimura substitution levels of masked regions against their corresponding model are automatically calculated by RepeatMasker and output by default. These were compiled for each (sub-) cluster and plotted as a means of assessing their divergence. LTR element age was calculated using mutation rates estimated for C. elegans\textsuperscript{52} (per generation: $2.70 \times 10^{-9}$ / per year $2.46375 \times 10^{-7}$). Assuming C. elegans mutation rates, LTR element invasion was calculated to be ~10 to 60 Mio years old (Supplementary Table 1).

S10.6 Solo/paired LTR element abundance analysis

We have estimated the proportion of solo LTRs and complete or nearly complete LTR/Gypsy. Solo LTRs were defined as events that lack any trace of internal region and have a length of at least 60 % of the specific consensus sequence. Complete or nearly complete repeat elements were defined as events that have LTRs flanking an internal region and an overall length (LTRs plus internal region) of at least 60 % of the consensus sequence.
**S10.7 Scaffold end analysis**

To examine the effect of different repeat classes on the existing assembly, their occurrence-frequency at scaffold ends was compared to the expected chance occurrence of the class. For each scaffold in the assembly, the terminal 1 kbp ends were examined for overlap with repeat annotations using BEDTools\textsuperscript{18} and scored as repetitive if more than 600 bp were covered by repeat annotations. For calculating the sum of repetitive sequence length, only the actual bases overlapping the repeat annotation were counted. Chance occurrence was determined by examining a set of 1,000 random genomic regions (1 kbp each) for repeat overlap annotation as above. Scaffold termini were excluded from the random genomic regions set.

**S11: AAT-Repeat Analysis**

**S11.1 AAT-microsatellite prediction and read alignment break analysis**

The analysis of microsatellites was based on a tandem repeat finder annotation of the dd_Smed_g4 assembly. All variations of AAT* patterns were converted into a Marvel interval track and incorporated into a Marvel dd_Smed_g4 contig database. Patched P6C4 reads (coverage 20x) were mapped with DALIGNER on the dd_Smed_g4 assembly. For each of the AAT* regions, the alignments of P6/C4 reads were analyzed and the fraction breaking within the AAT* region versus all reads spanning 500 bases left and right of the AAT* region was determined. Breaking and spanning reads were binned (bin size: 180 bp) according to length of the AAT* repeat within the interval [100, 1000]. Bin1: 100-280 bp, avg 166 bp; bin2: 280-460 bp, avg 351 bp; bin3: 460-640 bp, avg 536 bp; bin4: 640-820 bp, avg 719 bp; bin5: 820-1000 bp, avg 900 bp.

**S11.2 Microsatellite repeat length estimation by Circular Consensus Sequencing (CCS)**

In order to determine the origin of the length variations causing the read alignment breaks over AT-regions, circular consensus sequencing (CCS)\textsuperscript{53} reads with a minimal
genome coverage (< 1x) were generated. For 2 kbp CCS library, 2 µg of AMPure XP bead purified high molecular weight genomic DNA was sheared to 2 kbp fragments using clear miniTUBEs on the Covaris S2 instrument according to the manufacturers conditions. The majority of fragments were about 3.9 kbp in size. Of this 750 ng of sheared genomic DNA went into PacBio SMRT bell library preparation as described in “Procedure and Checklist – 2 kbp template preparation and sequencing” with the exception that the blunt sequencing adapter was used in a final concentration of 5 µM. Primer dimers and short fragments were removed after library preparation by 2x 0.6X AMPure XP bead washing steps. SMRT bell libraries were loaded on to PacBio RSII SMRT cells after primer annealing and polymerase binding, applying the MagBead loading protocol. Sequencing was done on the RSII instrument with P6 polymerase and C4 sequencing chemistry and 360 min movies to generate polymerase reads that lead to circular consensus sequences (CCS). The sequences were submitted to the SRA under accession SRR5408393.

S11.3 Overlap length variation in CCS and P6/C4 reads

CCS raw reads and patched P6/C4 reads were both mapped to the dd_Smed_g4 assembly with DALIGNER. The CCS consensus reads were generated by pbccs with the use of the default parameters and were only used to detect unique mapping locations for CCS reads, while ambiguously mapped P6/C4 reads were excluded from the analysis.

From this, AAT repeats were predicted as follows:

1. AAT-regions were selected, by choosing regions that can be spanned (~100 bp anchor right and left of AAT region) by uniquely mapping (see S11.3) circular consensus sequences (CCS). This resulted in 1,141 unique regions in the dd_Smed_g4 assembly, with length bins ranging from 300 – 1000 bp.

2. To eliminate sequencing coverage differences between different regions of the genome, 5 random CCS and 5 random P6/C4 reads were selected for each of those 1,141 dd_Smed_g4 regions.

3. Box plots for the fraction of CCS overlap length reads vs overlap length in CCS consensus read and fraction of P6/C4 overlap length vs dd_Smed_g4 overlap length were created.
4. The plot includes only length bins in the range of 400 - 800, because
   a. There are only a small number of events in bins of 900 - 1,000
   b. For bin 300 the difference between unique and AAT regions is very small as AAT regions can have up to 200 anchor bases.

The variance between the CCS reads and unique segments of the genome compared to the variance between standard P6/C4 reads and unique segments of the genome were under the error-rate of standard P6/C4 PacBio reads. The same was the case for the variance between the CCS reads and ATT* regions compared to the variance of standard P6/C4 and AAT* regions. However, if the AAT* loci within the dd_Smed_g4 assembly did vary, it would be expected that this variance was outside the expected error-rate of PacBio reads. Single regions were found that did have a high variance. However, at this point we cannot exclude that these were due to differences between haplotype differences being observed. Therefore, a biological source for the variance within AAT* loci could not be found and it can only be concluded that these are of a technical origin due resulting from pairwise alignment of repeat regions.

S11.4 Genomic context of AT-rich regions
To characterise the genomic location of AT-rich microsatellite repeats, genomic features were defined on basis of the mapping results of the Smes_v1 transcriptome against the dd_Smed_g4 assembly. Then, locations of the previously defined AT-rich microsatellite repeats were extracted using IntersectBed. An overlap of at least 60 % of the query (>100 bp events) to a given genomic feature was required to be scored as either intergenic, intronic or exonic.

S12: Assembly Heterozygosity
S12.1 Drosophila assembly
For the comparison between the dd_Smed_g4 assembly and a D. melanogaster assembly, we used 50x coverage of the publicly available SRA dataset (SRX4999318) with a read length cut-off of 14 kbp. The resulting standard MARVEL assembly had a
contig N50 of 20 Mb and the longest contig was 28 Mb. The assembly displayed almost no alternative paths (bubbles) in the overlap graph, indicating a high level of homozygosity within the dataset. A representative 1.7 Mb portion of the longest contig was used to illustrate the differences between D. melanogaster and Smed in Fig. 2e.

**S12.2 Dot plot analysis**
The analysis and corresponding dot plot in Fig. 2f was generated by Mummer (v3.2.3)\textsuperscript{54}.

**S12.3 Coverage analysis alt vs. main contigs**
To quantify genome-wide coverage differences between alternative and main contigs, we only considered indels as regions of unambiguous difference between alt and main contigs. For each indel region, the fraction of breaking read alignments versus the coverage of spanning reads was quantified. Only positions of non AT-rich indels of length > 99 bases were considered. The coverage of breaking and spanning alignments was limited to [5, 25] and indel lengths had to be consistent at each position. The analysis was based on MARVEL assembled contigs (no correction, no scaffolding), and on the trimmed P6/C4 reads. 1,509 alternative contigs (bubbles) that mapped to 1,495 unique locations on 695 contigs in dd_Smed_g4 were considered (14 alternative regions (~1%) represented non-uniform structures such as stacked alternative contigs). Alternative contigs had an average of 13 break events (including AAT* patterns and inserts of any size) with a median spacing of 6,676 bp.

**S13: Gene Annotation**

**S13.1 Gene annotation and gene number estimate**
The transcriptome of the sexual strain (Smes_v1, PCFL) was mapped to the dd_Smed_g4 assembly using GMAP\textsuperscript{15} with a minimum query coverage of 0.6 and minimum identity of 0.6. Our transcriptome contains ~31 000 non-redundant and likely protein-coding transcripts. This number corresponds well to a previous estimate of gene numbers in the planarian genome\textsuperscript{54}. However, the actual number
of Smed genes is likely substantially lower due to fragmented transcripts or transcribed repeat elements.

**S13.2 Gene structure**

Gene structure statistics were calculated for a set of high-confidence transcripts, identified on basis of i) unique mapping location in the dd_Smed_g4 assembly and ii) having a homologue in at least 4 other species. In total, 12,413 transcripts fulfilled all criteria. The median/mean transcript size was 1,545 / 1,875 bp and mapped length 1,534 / 1,866 bp, respectively. Median/mean gene size was 7,473 / 15,395 bp, number of 4 / 5.76 introns, and intron length of 1,329 / 2,694, respectively.

**S13.3 Orthology analysis**

To establish an orthology model of Smed genes, we performed a pair-wise BLAST of (a) 20 selected Ensembl reference proteomes and (b) the proteome of Smed. To define the latter, we used the Smes_v1.PCFL transcriptome of the sexual strain and extracted just the primary Open Reading Frame (ORF) (i.e. longest non-overlapping ORF) of each transcript. BLAST search was performed using blastp using a BLOSUM45 similarity matrix and e-value cutoff of 1E-3. BLAST results were pre-filtered for subject and query coverage of at least 10 % and post-filtered for reciprocal best hits. Resulting data were clustered into orthology groups (i.e. graph components) using the igraph package in R56. Orthology groups were assessed with respect to presence/absence of a given species and/or combination of species.

**S14: Divergence Analysis**

**S14.1 Divergence analysis**

Orthologue groups of 52 highly conserved proteins across 22 species were generated using the orthoMCL pipeline (for details see S16.1) (Supplementary Table 2). Briefly, orthogroups that contained all species and where only one or two species had more than 1 orthologue were selected. Ambiguous amino acid residues (letter “X”) were first removed from each of the sequences. Each orthologue set was then aligned using PRANK57 (v10603, parameters: 10 iterations, +F) and trimmed using trimAl58 (v1.4rev15, parameters: gappyout). The orthologue sets were then concatenated to
form a supermatrix using a perl script. Maximum likelihood analysis was performed using RaxML (v8.2.9) with the PROTCATLG model. The topology of non-Platyhelminthes species was constrained using a guide tree (based on the phylogeny by Cannon et al. 2016).

S15: Synteny Analysis

S15.1 Synteny comparisons
We used the dd_Smed_g4 assembly as the reference genome and computed pairwise genome alignments to other Platyhelminthes (query) genomes using lastz (v1.03.54; parameters K=2400 L=3000 Y=3400 H=2000, HoxD55 scoring matrix) and the chain/net pipeline (parameters chainMinScore 1000, chainLinearGap loose). In addition, we used highly-sensitive local alignments with lastz parameters K=1500 L=2500 and W=5 to find co-linear alignments in the un-aligning regions that are spanned by local alignments (gaps in the chains). The same procedure was applied to align the M. lignano genome to other Platyhelminthes including Smed. For comparison, we also aligned the human genome against frog (xenTro7 assembly), zebrafish (danRer10) and lamprey (petMar2).

S16: Planarian Specific Gene Sequences

S16.1 Identification of planarian-specific genes
To search for putative novel or planarian-specific genes in Smed, we made use of OrthoMCL, which groups putative orthologues and paralogs using a Markov Chain algorithm. Here, we applied OrthoMCL (v2.0.9) to the analysis of a set of 27 species (Supplementary Table 3) representing the taxonomic groups Annelida (n=1), Arthropoda (n=4), Chordata (n=3), Cnidaria (n=1), Ctenophora (n=1), Mollusca (n=2), Nematoda (n=2) and Platyhelminthes (n=13, which included the transcriptomes of n=6 planarian species). To generate the input similarity matrix for OrthoMCL, we carried out an all-against-all reciprocal blastp analysis amongst the predicted protein sets of the 27 species with -matrix “BLOSUM45” and -value 1e-5 parameters. OrthoMCL was run with a parameter of “-l 1.5”. Only OrthoMCL
orthologue/paralogue groups comprised exclusively of flatworm entries were considered for further analysis.

Next, the Smed sequences (derived from Smes_v1.PCFL ORFs) were extracted from flatworm-specific orthogroups and further sub-filtered according to the following criteria:

1. ORF length equal or greater than 160 amino acids.
2. Similarity (blastp -value lower or equal to 1e-15) to at least one other distant planarian species (Dlac, Pten, Pnig or Ptor).
3. No Refseq BLAST hit (-value lower or equal to 1e-3) in non-Platyhelminthes species (RefSeq BLAST database snapshot from May 24th 2016).
4. Mapping location in the dd_Smed_g4 assembly has low overlap with repeats (less than 60 % of ORF length).
5. Mapping location in the dd_Smed_g4 assembly has no other transcripts mapped within 50 bp upstream and 150 bp downstream in order to exclude fragmented transcripts. Manual inspection confirmed that the latter two filtering criteria greatly reduced the presence of transcript fragments representing poorly conserved protein regions, yet the presence of some such sequences in the final data set is possible.

Altogether, 1,165 transcripts passed the filtering and we therefore refer to this data set as flatworm-specific genes (Supplementary Table 4). Given the focus on transcribed sequences and the additional application of ORF length criteria, these largely represent actively transcribed, protein-coding genes that have no detectable sequence similarity outside the phylum by standard BLAST criteria (step 3). Our analysis underestimates the content of “new” genes in the Smed genome, since recently repeat-derived sequences are excluded by filtering (step 4) and the requirement for similarity in distant planarian species (step 2) removes recently generated or rapidly evolving gene sequences. Given the significant sequence divergence between planarian species at the transcriptome level (not shown), it is in fact very likely that many such genes exist and that planarians therefore constitute an interesting model system for studying the evolution of new genes.
**S16.2 Domain predictions**

Domains of flatworm-specific genes were predicted using SMART\(^66\) (“Normal mode” with the checked options “Outlier homologues and homologues of known structure”, “PFAM domains”, “signal peptides” and “internal repeats”) and InterProScan\(^67\) with SUPERFAMILY and Pfam applications.

**S16.3 Expression analysis**

To check the expression of flatworm-specific genes in *Smed*, transcript expression data sets for embryogenesis (SRA accession No.: SRR3629913 - SRR3629944), regeneration (SRR2051616 - SRR2051633) and stem cells (SRR2051616 - SRR2051633) were fetched from the NCBI – SRA database.

All data were mapped to the dd_Smed_g4 assembly using STAR\(^32\) (standard parameters) and a custom GTF file with genomic coordinates of mapped ORFs. Differential expression analysis was carried out using the DESeq2\(^68\) package for datasets with replicates (stem cells and embryogenesis). Log fold change greater than 0.5 or lower that -0.5 were used as cut-off for the regeneration dataset.

**S17: Gene Loss in Planarians**

**S17.1 Lost genes identification and verification**

The detection of high confidence gene loss events in *Smed* was performed using an in-house custom pipeline. Briefly, we performed a pair-wise reciprocal BLAST search against (a) Ensembl reference proteomes (see Supplementary Table 3 for database versions) and (b) 7 proteomes predicted from assembled planarian transcriptomes (including both Smed_v6 and Smes_v1). The inclusion of other planarian transcriptomes provided an important guard against false positive loss events due to accidental absence from one transcriptome. Orthology group clusters lacking any planarian species were then further filtered and validated using the dd_Smed_g4 genome assembly, as described in detail below:

**Step 1 – Pairwise reciprocal BLAST and coverage filtering**
For the pairwise reciprocal BLAST search, just the primary (i.e. longest non-overlapping open reading frame) of the longest transcript of each Trinity graph component was used as input. BLAST searches were performed using blastp with a BLOSUM45 substitution matrix and an e-value cutoff of 1e-3.

Step 2 - Orthology group construction
BLAST results were pre-filtered for subject and query coverage of at least 10 % and post-filtered for reciprocal best hits. Resulting data were clustered into orthology groups (i.e. graph components) using the igraph package in R<sup>56</sup>. Orthology groups were assessed with respect to species presence/absence, and groups containing at least 9 non-flatworm species, no Smed and no other planarian species were considered to be putatively lost in Smed.

Step 3 - Genomic loss verification using exonerate
To independently verify the loss of such genes in the Smed genome, all other protein members of planarian-deficient orthology groups were mapped onto the dd_Smed_g4 assembly using exonerate (--showvulgar no --showalignment no --model protein2genome -s 40 --showtargetgff yes --percent 2). Genes were considered as genuinely lost if no point in the genome was covered by overlapping exonerate alignments of more than 2 different non-planarian species. This threshold was necessary in order to filter for spurious alignments and residual 2-hit positions invariably had neither RNAseq coverage nor coding potential, as evidenced by a lack of blastx hits against NCBI nr database of those intervals.
Overall, our pipeline bases the definition of gene loss on the absence of detectable sequence similarity both in 6 independently assembled planarian transcriptomes and in the Smed genome. Further, the limitation to orthogroups with > 9 non-flatworm members generally restricts the analysis to genes that are highly conserved at the sequence level. Although false positives are consequently highly unlikely, we cannot rule out the presence of homologues that, only in planarians, have diverged beyond the point of detectable homology by conventional sequence comparisons.

Step 4 - Validation of lost gene candidates
The protein translations of putatively lost genes were additionally manually searched against Annelid, Molluscan and Platyhelminth transcriptomes using tblastn (parameters: -word size 2 -matrix BLOSUM45 -soft_masking TRUE -evalue 100). Hits were inspected manually to verify absence in the respective species (Supplementary Table 5). The fact that we did not recover the previously reported loss of the centrosome proteins\textsuperscript{69} indicates that our focus on deeply conserved genes with a high degree of sequence conservation likely underestimates the extent of gene loss in \textit{Smed}.

**S17.2 Essentiality assessment of lost genes**

The human and/or mouse homologues of genes lost in \textit{Smed} were used to query the Database of Essential Genes (DEG v13.3)\textsuperscript{70}. Positive hits were annotated as “essential” in Supplementary Table 5.

**S17.3 Mad1/Mad2 multiple sequence alignments**

Multiple sequence alignments for Mad1 and Mad2 were generated using the Constraint-based Multiple Alignment Tool (COBALT)\textsuperscript{71} (standard parameters) and resulting alignments were rendered using the BoxShade website with standard parameters (http://www.ch.embnet.org/software/BOX_form.html). Furthermore, the COBALT alignments were imported into Geneious\textsuperscript{72} (v9.1.5 Build 2016-06-11 21:21, http://www.geneious.com) and sequence similarity matrices between species were computed based on BLOSUM62 similarity matrix (threshold 0). The similarity matrices were rendered using matrix2png\textsuperscript{73}.

**S18: Planarian SAC Function**

**S18.1 Quantification of mitotic index in dissociated cell preparations**

**RNAi and nocodazole treatment**

RNA interference was performed as previously described\textsuperscript{74} with some minor modifications. \textit{In vitro} synthesized dsRNA was diluted 1:20 in H\textsubscript{2}O and quantified against a column purified standard eGFP control dsRNA (1 µg/µl) on a 0.7 % TAE agarose gel using ImageStudio Lite (v5.2.5, LI-COR Biosciences). The quantified
dsRNA was then added to liver paste to a final concentration of 1 µg/µl and frozen at -80 °C in small aliquots. The addition of red food dye as tracer was omitted, because it generated strong fluorescent background in the rhodamine channel (see “Imaging and data analysis”).

For each experimental replicate, 20 animals (~0.5 mm, asexual CIW4 strain) were fed 3x with dsRNA food every third day. On the eighth day, worms were split into 2 batches of 10 animals and either transferred to dishes with 50 ng/ml nocodazole + 1 % DMSO (v/v) in planarian water (+gentamycin) or 1 % DMSO (v/v) in planarian water (+gentamycin) as a vehicle control. The worms were treated for 24 hours and then macerated.

**Animal maceration, proteinase K treatment and formaldehyde fixation**

Animals were passively macerated in 4 ml of maceration solution (glycerol:glacial acetic acid:H₂O in a ratio of 1:1:13 (w/v/v)) for 10 min on the bench at RT, followed by 10 min on a vertical rotator at RT until the tissue was completely dissociated. The lysate was filtered through CellTrics® 50 µm filters (Sysmex, Cat. No.: 04-0042-2317). Depending on the size of the animals, the cell suspension was diluted 1:4 (small, 7-8 mm) to 1:10 (large, 16-18 mm) using maceration solution and 120 µl of the diluted cell suspension was transferred into each well of a black 96-well glass bottom plate (Greiner Bio-One, Cat. No.: 655090). For each condition, five or six separate wells were used as technical replicates. The cells were pelleted for 5 min at 2,000 x g at RT. Then 44 µl of 16 % formaldehyde (PFA) in 1X PBS were added and the cells were fixed for 20-25 min at RT and rinsed once with 200 µl 1X PBS, followed by a wash in 1X PBS for 5 min. After fixation, the cells were treated with 2 µg/ml Proteinase K in 1X PBS for 10 min at RT, rinsed once with 1X PBS and post-fixed with 4 % PFA in 1X PBS for 10 min at RT. Then, the cells were rinsed once in 1X PBS and washed in PBSTx0.1 (1X PBS; 0.1 % (v/v) Triton X-100) for 5 min.

**Riboprobe hybridization, washing**

The riboprobe hybridization buffers were prepared according to Pearson et al. The cells were incubated 5 min in 1:1 (v/v) PBSTx0.1:WashHyb and subsequently blocked in PreHyb for 45-60 min at 56 °C. PreHyb was removed and replaced with
pre-warmed Hyb containing DIG-labeled riboprobe against smedwi-1 and hybridized over 16 h at 56 °C. Hyb was removed and the cells were briefly rinsed with pre-warmed WashHyb. The cells were then successively washed once 5 min with pre-warmed 1:1 SSC/WashHyb, followed by 2x 15 min 2X SSC + 0.1 % (v/v) Triton X-100, 2x 15 min 0.2X SSC + 0.1 % (v/v) Triton X-100. Then, the cells were brought to RT and rinsed and washed once with PBSTx0.1 for 10 min. Endogenous peroxidase activity was inactivated with 100 mM sodium azide in PBSTx0.1 for 15 min at RT and rinsed once and washed twice with PBSTx0.1 for 10 min each.

**Tyramide signal amplification**

The cells were blocked 1 h with 5 % (v/v) horse serum + 0.5 % (v/v) Roche Western Blocking Reagent (RWBR; Roche, Cat No.: 11921673001) in PBSTx0.1 at RT and then incubated with anti-DIG-POD Fab fragments (Roche, Cat No.: 11207733910) in 1 % (v/v) horse serum/0.1 % (v/v) RWBR in PBSTx0.1 for 2 h at RT. After washing 3x 20 min with PBSTx0.1 the signal was developed with rhodamine-tyramide in TSA buffer (2 M NaCl, 0.1 M boric acid, pH 8.5) containing 0.006 % H2O2 for 10 min.

**Antibody staining**

Antibody staining was essentially performed as previously described with some modifications77. After removing the TSA buffer and washing 3x for 10 min with PBSTx0.1, primary antibody (1:1,000 anti-Histone H3 antibody (rabbit, phospho S10 + T11) [E173]; Abcam, Cat. No.: ab32107) in 1 % goat serum + 0.1 % RWBR was added and incubated 16 h at 4 °C. After washing 4x 10 min with PBSTx0.1, the solution was replaced with secondary antibody in 1% (v/v) goat serum + 0.1 % (v/v) RWBR (1:1,000 goat anti-rabbit IgG (H+L), Alexa Fluor 647; ThermoFisher Scientific, Cat. No.: A-21245) and DAPI (1 µg/ml) and incubated 2-3 h at RT in the dark. The secondary antibody was washed off for 3x 20 min with PBSTx0.1 and the plates were stored in 0.02 % (w/v) sodium azide in 1X PBS at 4 °C in the dark.

**Imaging and data analysis**
The plates were imaged on an Operetta High-Content Imaging System (Perkin Elmer) using a 20x lens and the following excitation/emission filters: DAPI (Ex. 360-400/Em. 410-480 nm), Rhodamine (520-550/560-630 nm) and far red (620-640/650-760 nm). For each well, five separate locations were imaged. Image analysis was performed using the integrated Harmony High Content Imaging and Analysis Software and FIJI.

**Statistical analysis**

Comparisons of group means (± S.D.) for the effect of RNAi interference of SAC components on the ability of the planarian neoblasts to elicit a SAC response upon nocodazole treatment, a two-way analysis of variance (ANOVA) was performed and in cases of significance, followed by Dunnett’s test. Two-way ANOVA revealed a significant effect of SAC component RNAi on the activation of the SAC by nocodazole treatment ($F(4, 24) = 124.7, P < 0.0001$). For the tested RNAi targets, *rod, zwilch, zw10* yielded significant reductions of anti-H3ser10P-positive cells (M-phase marker) upon nocodazole treatment ($P < 0.0001$), while *bub3* did not differ from control RNAi (egfp) ($P = 0.78$). RNAi itself did not significantly alter the baseline number of cells in M-phase in the DMSO controls. All statistical analyses were performed using GraphPad Prism version 7.0c for Mac OS X, GraphPad Software, La Jolla California USA, www.graphpad.com.

**S19: Miscellaneous Experimental Methods**

**S19.1 Animal husbandry**

Sexual (S2) and asexual (CIW4) strains of *Smed* were kept in plastic containers in 1X Montjuïc salt water with 25 mg/L gentamycin sulfate. The animals were fed homogenized organic calf liver paste.

**S19.2 Transcript cloning**

Target cDNA fragments were PCR-amplified from *Smed* cDNA using KAPA HiFi HotStart DNA Polymerase (Kapa Biosystems), TAE agarose gel purified and cloned into pPR-T4p. The cloned insert was used for preparation of riboprobe templates for in situ hybridization and dsRNA production for RNA interference. The PlanMine
transcript IDs and gene specific forward and reverse primers used for cloning are provided in the Supplementary Table 7.

**S19.3 Whole mount in situ hybridization**

Whole mount *in situ* hybridization (WISH) was essentially performed as previously described\(^76,80\).

**S19.4 Karyotyping**

For karyotyping, animals (~5 mm) of sexual *Smed* were fed three times every third day using calf liver paste. The third day after the last feed, the planarian water was supplemented with 50 ng/ml nocodazol (Sigma Cat # M1404) and 1 % (v/v) DMSO (Sigma Cat # 276855) for a final 24 h incubation period. Subsequently, 15 animals were macerated in 5 ml 6.6 % (v/v) glacial acetic acid supplemented with 5 µg/ml Hoechst 34580 (ThermoFisher Scientific Cat # H21486). Single cell dispersion was achieved by carefully rotating the maceration solution for the last 10 min of the 20 min total incubation period. Subsequently, 50 µl of the cell suspension was pipetted as a single drop on a 0.17 mm thick/No. 1.5 borosilicate coverslip (Menzel). To assist chromosome spreading, 100 µl of 0.05 % (v/v) Triton X-100 solution was added drop-wise to the slide. Chromosome spreads were aged at 37 °C for 60 min until the liquid completely evaporated. Coverslips were mounted in 80 % (v/v) glycerol prior to imaging. Chromosomes were imaged on an Andor Revolution WD Borealis confocal spinning disc system: The Olympus IX83 stand was equipped with an Andor iXon Ultra 888 EMCCD camera and an Olympus 150x U Apochromat NA 1.45 Oil immersion objective. Hoechst 34580 stained chromosomes were excited with a 405 nm laser diode and a 452/45 bandpass filter was used to detect the emission light. Image stacks were 3D deconvolved using ImageJ/Fiji plugins\(^78,81\) and finally maximum projected for better visualization.

**Supplementary Tables**

**Supplementary Table 1** - Calculated Kimura distances and age for Smed LTR elements

**Supplementary Table 2** - MARVEL computational requirements & assembly stats
**Supplementary Table 3** - Conserved genes used for divergence analysis

**Supplementary Table 4** - Species and database versions used for orthology analysis

**Supplementary Table 5** - Planarian specific genes

**Supplementary Table 6** - Lost gene genes & essentiality analysis

**Supplementary Table 7** - PlanMine identifiers for transcripts and primers used for cloning used in this study

**References**


40. Finn, R. D. *et al.* The Pfam protein families database: towards a more


54. Kurtz, S. *et al.* Versatile and open software for comparing large genomes.


68. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and


81. Kirshner, H., Aguet, F., Sage, D. & Unser, M. 3-D PSF fitting for fluorescence