A platform for efficient transgenesis in *Macrostomum lignano*, a flatworm model organism for stem cell research

Jakub Wudarski[1], Daniil Simanov[1,2], Kirill Ustyantsev[3], Katrien de Mulder[2,8], Margriet Grelling[1], Magda Grudniewska[1], Frank Beltman[1], Lisa Glazenburg[1], Turan Demircan[2,10], Julia Wunderer[4], Weihong Qi[5], Dita B. Vizoso[6], Philipp M. Weissert[1], Daniel Olivieri[1,9], Stijn Mouton[1], Victor Guryev[1], Aziz Aboobaker[7], Lukas Schärer[6], Peter Ladurner[4], Eugene Berezikov[1,2,3]

[1] European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Antonius Deusinglaan 1, 9713AV, Groningen, The Netherlands.

[2] Hubrecht Institute-KNAW and University Medical Centre Utrecht, Uppsalalaan 8, 3584CT, Utrecht, The Netherlands.

[3] Institute of Cytology and Genetics, Prospekt Lavrentyeva 10, 630090 Novosibirsk, Russia.

[4] Institute of Zoology and Center for Molecular Biosciences Innsbruck, University of Innsbruck , Technikerstr. 25, A-6020, Innsbruck, Austria.

[5] Functional Genomics Center Zurich, Winterthurerstrasse 190, Zurich, CH-8057, Switzerland.

[6] Evolutionary Biology, Zoological Institute, University of Basel, Vesalgasse 1, CH-4051, Basel, Switzerland.

[7] Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford, OX1 3PS, United Kingdom.

[8] Present address: Molecular laboratory, AZ St. Lucas Hospital, Gent, Belgium.

[9] Present address: Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, Basel, CH-4058, Switzerland.

[10] Present address: Department of Medical Biology, International School of Medicine, İstanbul Medipol University, Istanbul, Turkey.

Correspondence should be addressed to E.B. (e.berezikov@umcg.nl).

ABSTRACT

Regeneration-capable flatworms are informative research models to study the mechanisms of stem cell regulation, regeneration and tissue patterning. However, the lack of transgenesis methods significantly hampers their wider use. Here we report development of a transgenesis method for *Macrostomum lignano*, a basal flatworm with excellent regeneration capacity. We demonstrate that microinjection of DNA constructs into fertilized one-cell stage eggs, followed by a low dose of irradiation, frequently results in random integration of the transgene in the genome and its stable transmission through the germline. To facilitate selection of promoter regions for transgenic reporters, we assembled and annotated the *M. lignano* genome, including genome-wide mapping of transcription start regions, and showed its utility by generating multiple stable transgenic lines expressing fluorescent proteins under several tissue-specific promoters. The reported transgenesis method and annotated genome sequence will permit sophisticated genetic studies on stem cells and regeneration using *M. lignano* as a model organism.

INTRODUCTION

Animals that can regenerate missing body parts hold clues to advancing regenerative medicine and are attracting increased attention (Tanaka and Reddien, 2011). Significant biological insights on stem cell biology and body patterning were obtained using free-living regeneration-capable flatworms (Platyhelminthes) as models (Elliott and Sanchez Alvarado, 2013; Rink, 2013; Wagner et al., 2011). The most often studied representatives are the planarian species *Schmidtea mediterranea* (Elliott and Sanchez Alvarado, 2013) and *Dugesia japonica* (Umesono et al., 2011)*.* Many important molecular biology techniques and resources are established in planarians, including fluorescence-activated cell sorting, gene knockdown by RNA interference, *in situ* hybridization, and genome and transcriptome assemblies (Rink, 2013). One essential technique still lacking in planarians, however, is transgenesis, which is required for in-depth studies involving e.g. gene overexpression, dissection of gene regulatory elements, real-time imaging and lineage tracing. The reproductive properties of planarians, including asexual reproduction by fission and hard non-transparent cocoons containing multiple eggs in sexual strains, make development of transgenesis technically challenging in these animals.

More recently, a basal flatworm *Macrostomum lignano* (Macrostomorpha) emerged as a model organism that is complementary to planarians (Grudniewska et al., 2016; Mouton et al., 2009; Simanov et al., 2012; Wasik et al., 2015). The reproduction of *M. lignano*, a free-living marine flatworm, differs from planarians, as it reproduces by laying individual fertilized one-cell stage eggs. One animal lays approximately one egg per day when kept in standard laboratory conditions at 20ºC. The eggs are around 100 microns in diameter, and follow the archoophoran mode of development, having yolk-rich oocytes instead of supplying the yolk to a small oocyte via yolk cells (Morris et al., 2004). The laid eggs have relatively hard shells and can easily be separated from each other with the use of a fine plastic picker. These features make *M. lignano* eggs easily amenable to various manipulations, including microinjection (Sato et al., 2016). In addition, *M. lignano* has several convenient characteristics, such as ease of culture, transparency, small size, and a short generation

time of three weeks (Mouton et al., 2009; Simanov et al., 2012). It can regenerate all tissues posterior to the pharynx, and the rostrum (Egger et al., 2006). This regeneration ability is driven by stem cells, which in flatworms are called neoblasts (Wagner et al., 2011; Rink, 2013; Davies et al., 2017). Recent research in planarians has shown that the neoblasts population is heterogeneous and consists of progenitors and stem cells (Scimone et al., 2014; Van Wolfswinkel et al., 2014). The true pluripotent stem cell population is, however, not identified yet.

Here we present a method for transgenesis in *M. lignano* using microinjection of DNA into single-cell stage embryos and demonstrate its robustness by generating multiple transgenic tissue-specific reporter lines. We also present a significantly improved genome assembly of the *M. lignano* DV1 line and an accompanying transcriptome assembly and genome annotation. The developed transgenesis method, combined with the generated genomic resources, will enable new research avenues on stem cells and regeneration using *M. lignano* as a model organism, including in-depth studies of gene overexpression, dissection of gene regulatory elements, real-time imaging and lineage tracing.

RESULTS

Microinjection and random integration of transgenes

M. lignano is an obligatorily non-self-fertilizing simultaneous hermaphrodite (**Fig. 1a**) that produces substantial amounts of eggs (**Fig. 1b,c**). We reasoned that microinjection approaches used in other model organisms, such as *Drosophila*, zebrafish and mouse, should also work in *M. lignano* eggs (**Fig. 1d, Supplementary Video 1**). First, we tested how the egg handling and microinjection procedure itself impacts survival of the embryos (**Supplementary Table 1**). Separating the eggs laid in clumps and transferring them into new dishes resulted in a 17% drop in hatching rate, and microinjection of water decreased survival by a further 10%. Thus, in our hands more than 70% of the eggs can survive the microinjection procedure (**Supplementary Table 1**). When we injected fluorescent Alexa 555 dye, which can be used to track the injected material, about 50% of the eggs survived (**Supplementary Table 1**). For this reason, we avoided tracking dyes in subsequent experiments. Next, we injected *in vitro* synthesized mRNA encoding green fluorescent protein (GFP) and observed its expression in all successfully injected embryos (n > 100) within 3 hours after injection (**Fig. 1e**), with little to no autofluorescence detected in either embryos or adult animals (**Supplementary Fig. 1**). The microinjection technique can thus be used to deliver biologically relevant materials into single-cell stage eggs with a manageable impact on the survival of the embryos.

To investigate whether exogenous DNA constructs can be introduced and expressed in *M. lignano*, we cloned a 1.3 kb promoter region of the translation elongation factor 1 alpha (*EFA*) gene and made a transcriptional GFP fusion in the Minos transposon system (**Supplementary Fig. 2a**). Microinjection of the *Minos::pEFA::eGFP* plasmid with or without Minos transposase mRNA resulted in detectable expression of GFP in 5-10% of the injected embryos (**Supplementary Fig. 2c**). However, in most cases GFP expression was gradually lost as the animals grew (**Supplementary**

Fig. 2f), and only a few individuals transmitted the transgene to the next generation. From these experiments we established the HUB1 transgenic line with ubiquitous GFP expression (**Supplementary Fig. 2e**), for which stable transgene transmission has been observed for over 50 generations (Marie-Orleach et al., 2014, 2016).
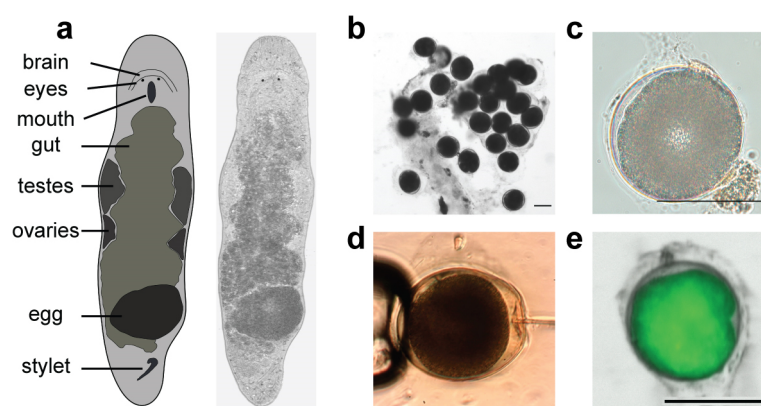

Figure 1. *Macrostomum lignano* embryos are amenable to microinjection. (**a**) Schematic morphology and a bright-field image of an adult *M. lignano* animal. (**b**) Clump of fertilized eggs. (**c**) DIC image of a one-cell stage embryo. (**d**) Microinjection into a one-cell stage embryo. (**e**) Expression of GFP in the early embryo 3 hours after injection with *in vitro* synthesized *GFP* mRNA. Scale bars are 100 µm.

The expected result for transposon-mediated transgenesis is genomic integration of the fragment flanked by transposon inverted terminal repeats. However, plasmid sequences outside the terminal repeats, including the ampicillin resistance gene, were detected in the HUB1 line, suggesting that the integration was not mediated by Minos transposase. Furthermore, Southern blot analysis revealed that HUB1 contains multiple transgene copies (**Supplementary Fig. 2g).** We next tried a different transgenesis strategy using meganuclease *I-SceI* (Thermes et al., 2002) to improve transgenesis efficiency (**Supplementary Fig. 2b**). We observed a similar 3-10% frequency of initial transgene expression, and only two instances of germline transmission, one of which resulted from the negative control experiment without co-injected meganuclease protein (**Supplementary Fig. 2c**). These results suggest that *I-SceI* meganuclease does not increase efficiency of transgenesis in *M. lignano*, but instead that exogenous DNA can be integrated in the genome by non-homologous recombination using the endogenous DNA repair machinery.

Improvement of integration efficiency

The frequency of germline transgene transmission in the initial experiments was less than 0.5% of the injected eggs, while transient transgene expression was observed in up to 10% of the cases (**Supplementary Fig. 2c,f**). We hypothesized that mosaic integration or mechanisms similar to extrachromosomal array formation in *C. elegans* (Mello and Fire, 1995) might be at play in cases of transient gene expression in *M. lignano.* We next tested two approaches used in *C. elegans* to increase the efficiency of transgenesis: removal of vector backbone and injection of linear DNA

fragments (Etchberger and Hobert, 2008a), and transgene integration by irradiation (Mello and Fire, 1995). Injection of PCR-amplified vector-free transgenes resulted in the germline transmission in 5 cases out of 269 injected eggs, or 1.86% (**Table 1**), and the stable transgenic line NL1 was obtained during these experiments (**Fig. 2a**). In this line, the *GFP* coding sequence was optimized for *M. lignano* codon usage. While we did not observe obvious differences in expression levels between codon-optimized and non-optimized *GFP* sequences, we decided to use codon-optimized versions in all subsequent experiments.
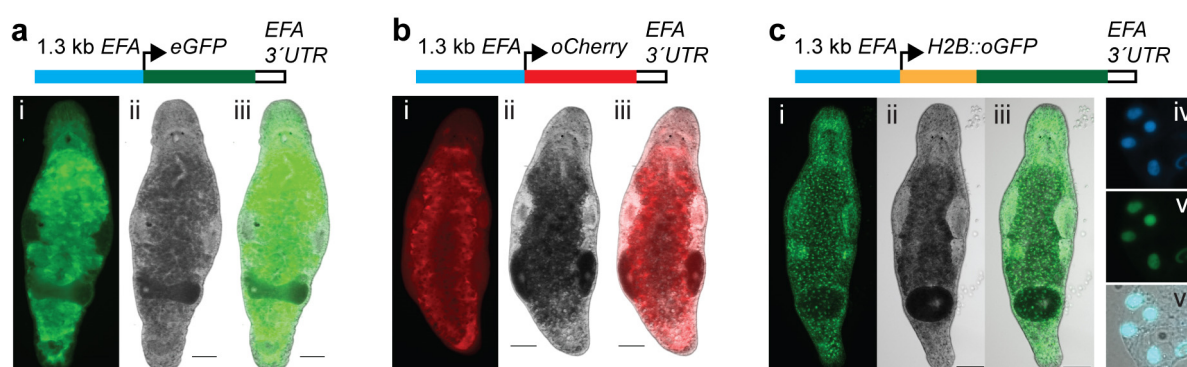

Figure 2. Ubiquitously expressed elongation factor 1 alpha promoter transgenic lines. (**a**) NL1 line expressing enchanced GFP (eGFP). (i) FITC channel; (ii) bright-field; (iii) merged. (**b**) NL3 line expressing codon-optimized Cherry (oCherry). (i) DsRed channel; (ii) bright-field; (iii) merged. (**c**) NL20 line expressing codon-optimized nuclear localized H2B::oGFP fusion. (i) FITC channel; (ii) bright-field; (iii) merged; (iv-vi) single cells from a macerated animal showing nuclear localization of GFP; (iv) DNA staining by Hoechst; (v) FITC channel; (iv) merge of Hoechst, FITC and bright-field. Scale bars are 100 μm.

M. lignano is remarkably resistant to ionizing radiation, and a dose as high as 210 Gy is required to eliminate all stem cells in an adult animal (De Mulder et al., 2010; Grudniewska et al., 2016). We reasoned that irradiation of embryos immediately after transgene injection might stimulate non-homologous recombination and increase integration rates. Irradiation dose titration revealed that *M. lignano* embryos are less resistant to radiation than adults and that a 10 Gy dose results in hatching of only 10% of the eggs, whereas more than 90% of eggs survive a still substantial dose of 2.5 Gy (**Supplementary Table 2**). Irradiating injected embryos with 2.5 Gy resulted in 1-8% germline transmission rate for various *EFA* promoter constructs in both plasmid and vector-free forms (**Table 1**). The stable transgenic line NL3 expressing codon-optimized red fluorescent protein Cherry was obtained in this way (**Fig. 2b**), demonstrating that ubiquitous expression of fluorescent proteins other than GFP is also possible in *M. lignano*. Finally, to test nuclear localization of the reporter protein, we fused GFP with a partial coding sequence of the histone 2B (H2B) gene as described previously(Kanda et al., 1998). The injection of the transgene fragment followed by irradiation demonstrated 5% transgenesis efficiency (**Table 1**), and the stable NL20 transgenic line with nuclear GFP localization was established (**Fig. 2c**).

Table 1. Efficiency of transgenesis with different reporter constructs and treatments

Reporter	Injected line	Injected DNA	Irradiation treatment	Injected eggs	Positive hatchlings (%)	Germline transmission (%)	Established lines
EFA::eGFP	DV1	PCR	-	269	39 (14.50)	5 (1.86)	NL1
EFA::oGFP	DV1	plasmid	-	114	28 (24.56)	0	-
EFA::oGFP	DV1	plasmid	2.5 Gy	42	13 (30.95)	2 (4.76)	-
EFA::oGFP	DV1	fragment	2.5 Gy	102	4 (3.92)	2 (1.96)	NL7
EFA::oCherry	DV1	plasmid	2.5 Gy	80	4 (5.00)	1 (1.25)	NL3
EFA::oCherry	DV1	fragment	2.5 Gy	36	6 (16.67)	3 (8.33)	NL4, NL5, NL6
EFA::H2B::oGFP	DV1	fragment	2.5 Gy	38	10 (26.32)	2 (5.26)	NL20
ELAV4::oGFP	DV1	fragment	2.5 Gy	56	29 (51.79)	2 (3.57)	NL21
MYH6::oGFP	DV1	fragment	2.5 Gy	103	13 (12.62)	1 (0.97)	NL9
APOB::oGFP	DV1	fragment	2.5 Gy	65	2 (3.08)	1 (1.54)	NL22
CABP7::oGFP	DV1	plasmid	-	20	2 (10.00)	1 (5.00)	NL23
CABP7::oNeon Green; ELAV4:: oScarlet-I	NL10	plasmid	-	137	3 (2.19)	2 (1.46)	NL24

Genome assembly and annotation

To extend the developed transgenesis approach to promoters of other genes, an annotated genome assembly of *M. lignano* was required. Towards this, we have generated and sequenced 29 paired-end and mate-pair genomic libraries of the DV1 line using 454 and Illumina technologies (**Supplementary Table 3**). Assembling these data using the MaSuRCA genome assembler (Zimin et al., 2013) resulted in a 795 Mb assembly with N50 scaffold size of 11.9 kb (data not shown). While this assembly was useful for selecting several novel promoter regions, it suffered from fragmentation. In a parallel effort, a PacBio-based assembly of the DV1 line, termed ML2, was recently published (Wasik et al., 2015). The ML2 assembly is 1,040 Mb large and has N50 contig size of 36.7 kb and NG50 contig size of 64.5 kb when adjusted to the 700 Mb genome size estimated from k-mer frequencies (Wasik et al., 2015). We performed fluorescence-based genome size measurements and estimated that the haploid genome size of the DV1 line is 742 Mb (**Supplementary Fig. 3**). It was recently demonstrated that *M. lignano* can have a polymorphic karyotype, where in addition to the basal 2n=8 karyotype, also animals with aneuploidy for the large chromosome, with 2n=9 and 2n=10 exist(Zadesenets et al., 2016). We confirmed that our laboratory culture of the DV1 line has predominantly 2n=10 and 2n=9 karyotypes (**Supplementary Fig. 3a,b**) and estimated that the size of the large chromosome is 240 Mb (**Supplementary Fig. 3f**). In contrast, an independently established *M. lignano* wild-type line NL10 has the basal karyotype 2n=8 and does not show detectable variation in chromosome number (**Supplementary Fig. 3**). This line, however, was established only recently and was not a part of the genome sequencing effort.

We re-assembled the DV1 genome from the generated Illumina and 454 data and the published PacBio data (Wasik et al., 2015) using the Canu assembler (Koren et al., 2017) and SSPACE scaffolder (Boetzer et al., 2011). The resulting Mlig_3_7 assembly is 764 Mb large with N50 contig and scaffold sizes of 215.2 Kb and 245.9 Kb respectively (**Supplementary Table 4**), which is greater than 3-fold continuity improvement over the ML2 assembly. To compare the quality of the ML2 and

6

Mlig_3_7 assemblies, we used the genome assembly evaluation tool REAPR, which identifies assembly errors without the need for a reference genome (Hunt et al., 2013). According to the REAPR analysis, the Mlig_3_7 assembly has 63.95% of error-free bases compared to 31.92% for the ML2 assembly and 872 fragment coverage distribution (FCD) errors within contigs compared to 1,871 in the ML2 assembly (**Supplementary Fig. 4a**). Another genome assembly evaluation tool, FRCbam, which calculates feature response curves for several assembly parameters (Vezzi et al., 2012), also shows better overall quality of the Mlig_3_7 assembly (**Supplementary Fig. 4b**). Finally, 96.9% of transcripts from the *de novo* transcriptome assembly MLRNA150904 (Grudniewska et al., 2016) can be mapped on Mlig_3_7 (>80% identity, >95% transcript length coverage), compared to 94.88% of transcripts mapped on the ML2 genome assembly, and among the mapped transcripts more have intact open reading frames in the Mlig_3_7 assembly than in ML2 (**Supplementary Fig. 4c**). Based on these comparisons, the Mlig_3_7 genome assembly represents a substantial improvement in both continuity and base accuracy over the ML2 assembly.

More than half of the genome is repetitive, with LTR retrotransposons and simple and tandem repeats accounting for 21% and 15% of the genome respectively (**Supplementary Table 5**). As expected from the karyotype of the DV1 line, which has additional large chromosomes, the Mlig_3_7 assembly has substantial redundancy, with 180 Mb in duplicated non-repetitive blocks that are longer than 500 bp and at least 95% identical. When repeat-annotated regions are included in the analysis, the duplicated fraction of the genome rises to 312 Mb.

Since genome-guided transcriptome assemblies are generally more accurate than *de novo* transcriptome assemblies, we generated a new transcriptome assembly based on the Mlig_3_7 genome assembly using a combination of the StringTie (Pertea et al., 2015) and TACO (Niknafs et al., 2016) transcriptome assemblers, a newly developed TBONE gene boundary annotation pipeline, previously published RNA-seq datasets (Cannon et al., 2016; Grudniewska et al., 2016) and the *de novo* transcriptome assembly MLRNA150904 (Grudniewska et al., 2016). Since many *M. lignano* transcripts are trans-spliced (Wasik et al., 2015; Grudniewska et al., 2016), we extracted reads containing trans-splicer leader sequences from raw RNA-seq data and mapped them to the Mlig_3_7 genome assembly after trimming the trans-splicing parts. This revealed that many more transcripts in *M. lignano* are trans-spliced than was previously appreciated from *de novo* transcriptome assemblies (6,167 transcripts in Grudniewska et al., 2016; 7,500 transcripts in Wasik et al., 2015; 28,273 in this study, **Supplementary Table 6**). We also found that almost 7% of the assembled transcripts are in fact precursor mRNAs, i.e. they have several trans-splicing sites and encode two or more proteins (**Supplementary Table 6, Supplementary Fig. 5a**). Therefore, in the transcriptome assembly we distinguish between transcriptional units and genes transcribed within these transcriptional units. For this, we developed computational pipeline TBONE (Transcript Boundaries based ON experimental Evidence), which relies on experimental data, such as trans-splicing and polyadenylation signals derived from RNA-seq data, to 'cut' transcriptional units and establish boundaries of mature mRNAs (**Supplementary Fig. 5a**). The new genome-guided transcriptome assembly, Mlig_RNA_3_7_DV1, has 66,777 transcriptional units, including duplicated copies and alternative forms, which can be collapsed to 33,715 non-redundant transcripts when clustered by 95% global sequence identity

(**Supplementary Table 6**). These transcriptional units transcribe 72,846 genes, of which 44,328 are non-redundant, 38.8% are trans-spliced and 79.98% have an experimentally defined poly(A) site (**Supplementary Table 6).** The non-redundant transcriptome has TransRate scores of 0.4360 and 0.4797 for transcriptional units and gene sequences respectively, positioning it among the highest quality transcriptome assemblies (Smith-Unna et al., 2016). The transcriptome is 98.1% complete according to the Benchmarking Universal Single-Copy Orthologs(Simao et al., 2015), with only 3 missing and 3 fragmented genes (**Supplementary Table 6**).

The Mlig_RNA_3_7_DV1 transcriptome assembly, which incorporates experimental evidence for gene boundaries, greatly facilitates selection of promoter regions for transgenesis. Furthermore, we previously generated 5´-enriched RNA-seq libraries from mixed stage populations of animals (Grudniewska et al., 2016) using RAMPAGE (Batut et al., 2013). In our hands, the RAMPAGE signal is not sufficiently localized around transcription start sites to be used directly by the TBONE pipeline, but it can be very useful for determining transcription starts during manual selection of promoter regions for transgenesis (**Supplementary Fig. 5b,c**). We used the UCSC genome browser software (Kent et al., 2002) to visualize genome structure and facilitate design of new constructs for transgenesis (**Supplementary Fig. 5**). The *M. lignano* genome browser, which integrates genome assembly, annotation and RNA-seq data, is publicly accessible at http://gb.macgenome.org.

Tissue-specific transgenic lines

Equipped with the annotated *M. lignano* genome and the developed transgenesis approach, we next set to establish transgenic lines expressing tissue-specific reporters. For this, we selected homologs of the *MYH6, APOB, ELAV4* and *CABP7* genes, for which tissue specificity in other model organisms is known and upstream promoter regions can be recognized based on genome annotation and gene boundaries (**Supplementary Fig. 5**). Similar to the *EFA* promoter, in all cases the transgenesis efficiency was in the range of 1-5% of the injected eggs (**Table 1**) and stable transgenic lines were obtained (**Fig. 3**). Expression patterns were as expected from prior knowledge and corroborated by the whole mount *in situ* hybridization results: the *MYH6::GFP* is expressed in muscle cells, including muscles within the stylet (**Fig. 3a, Supplementary Video 2**); *APOB::GFP* is gut-specific (**Fig. 3b**); *ELAV4::GFP* is testis-specific, including the sperm, which is accumulated in the seminal vesicle (**Fig. 3c**); and *CABP7::GFP* is ovary-specific and is also expressed in developing eggs (**Fig. 3d**). Finally, we made a double-reporter construct containing *ELAV4::oNeonGreen* and *CABP7::oScarlet-I* in a single plasmid (**Fig. 3e**). mNeonGreen (Shaner et al., 2013) and mScarlet (Bindels et al., 2016) are monomeric yellow-green and red fluorescent proteins, respectively, with the highest reported brightness among existing fluorescent proteins. The transgenesis efficiency with the double-reporter construct was comparable to other experiments (**Table 1**), and transgenic line NL24 expressing codon-optimized mNeonGreen (oNeonGreen) in testes and codon-optimized mScarlet-I (oScarlet) in ovaries was established (**Fig. 3e**), demonstrating the feasibility of multi-color reporters in *M. lignano*. The successful generation of stable transgenic reporter lines for multiple tissue-specific promoters validates the robustness of the developed transgenesis method and demonstrates the value of the generated genomic resource.
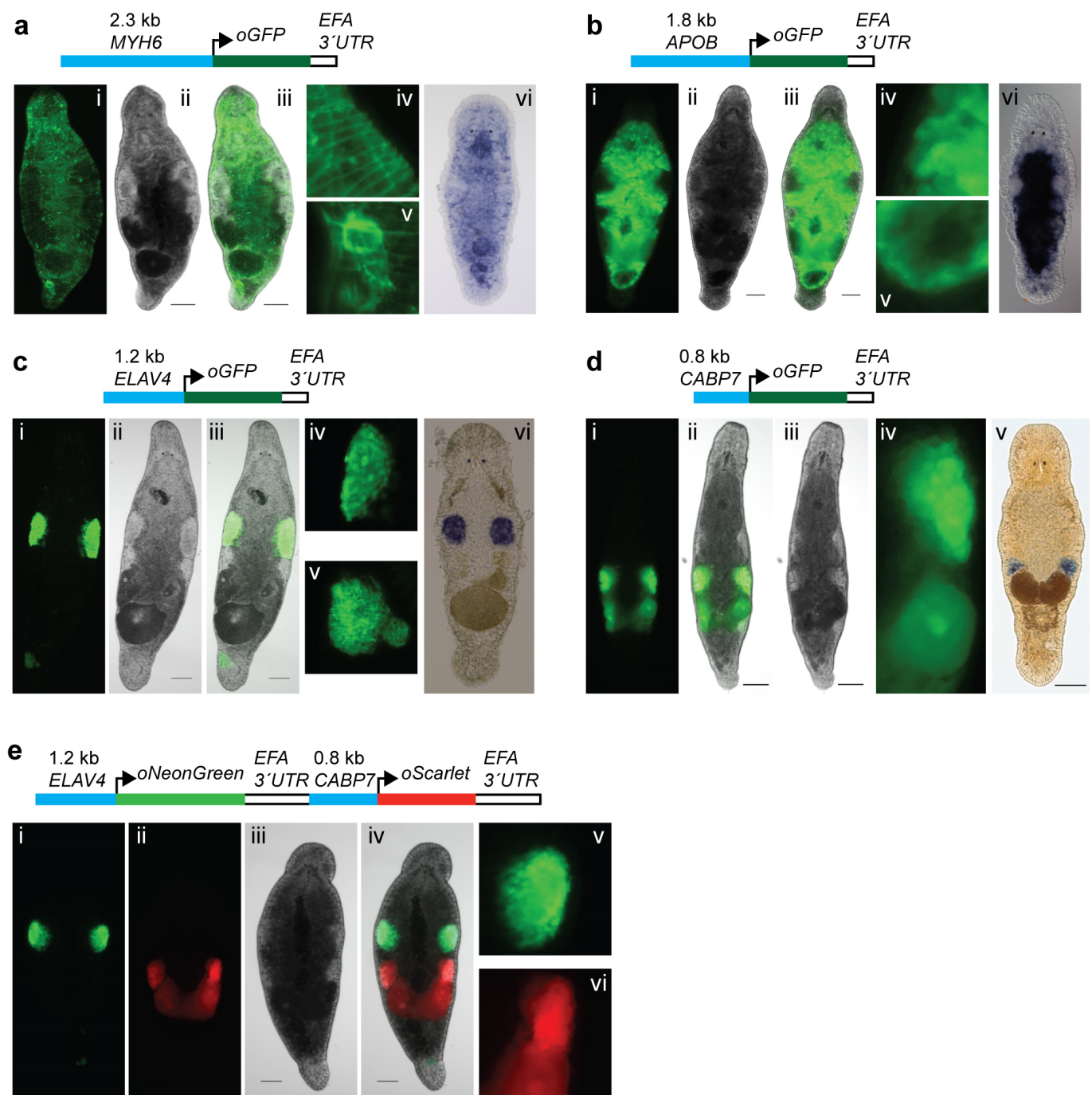
Figure 3. Tissue-specific promoter transgenic lines. (**a**) NL9 line expressing GFP under the muscle-specific promoter of the *MYH6* gene. (i) FITC channel; (ii) bright-field; (iii) merged; (iv,v) detailed images of the body wall and stylet respectively; (vi) whole-mount *in situ* hybridization expression pattern of *MYH6* transcript. (**b**) NL22 line expressing GFP under the gut-specific promoter of the *APOB* gene. (i) FITC channel; (ii) bright-field; (iii) merged; (iv,v) detailed images of the gut side and distal tip respectively; (vi) whole-mount *in situ* hybridization expression pattern of the *APOB* transcript. (**c**) NL21 line expressing GFP under the testis-specific promoter of the *ELAV4* gene. (i) FITC channel; (ii) bright-field; (iii) merged; (iv,v) detailed images of the testis and seminal vesicle, respectively; (vi) whole-mount *in situ* hybridization expression pattern of the *ELAV4* transcript. (**d**) NL23 line expressing GFP under the ovary-specific promoter of the *CABP7* gene. (i) FITC channel; (ii) bright-field; (iii) merged; (iv) detailed image of the ovary and developing egg; (v) whole-mount *in situ* hybridization expression pattern of the *CABP7* transcript. (**e**) NL24 line expressing in a single construct NeonGreen under the testis-specific promoter of the *ELAV4* gene and Scarlet-I under the ovary-specific promoter of the *CABP7* gene. (i) FITC channel; (ii) DsRed channel; (iii) bright-field; (iv) merged; (v,vi) detailed images of the testis and ovary regions, respectively. Scale bars are 100 μm.

9

Identification of transgene integration sites

To directly demonstrate that transgenes integrate into the *M. lignano* genome and to establish genomic locations of the integration sites, we initially attempted to identify genomic junctions by inverse PCR with outward-oriented transgene-specific primers (**Supplementary Fig. 6a**) in the NL7 and NL21 transgenic lines. However, we found that in both cases short products of ~200 nt are preferentially and specifically amplified from genomic DNA of the transgenic lines (**Supplementary Fig. 6b,c**). The size of the PCR products can be explained by formation of tandem transgenes (**Supplementary Fig. 6a**), and sequencing confirmed that this is indeed the case (**Supplementary Fig. 6d**). Next, we used the Genome Walker approach, in which genomic DNA is digested with a set of restriction enzymes, specific adapters are ligated and regions of interest are amplified with transgene-specific and adapter-specific primers. Similarly, many of the resulting PCR products turned out to be transgene tandems. But in the case of the NL21 line we managed to establish the integration site on one side of the transgene (**Supplementary Fig. 6e**), namely at position 45,440 in scaf3369 (Mlig_3_7 assembly) in the body of a 2-kb long LTR retrotransposon, 10.5 kb downstream from the end of the Mlig003479.g3 gene and 2.5 kb upstream from the start of the Mlig028829.g3 gene.

Transgene expression in regenerating animals

Our main rationale for developing *M. lignano* as a new model organism is based on its experimental potential to study the biology of regenerative processes *in vivo* in a genetically tractable organism. Therefore, it is essential to know whether regeneration could affect transgene stability and behaviour. Towards this, we monitored transgene expression during regeneration in the testis- and ovary-specific transgenic lines NL21 and NL23, respectively (**Supplementary Fig. 7**). Adult animals were amputated anterior of the gonads and monitored for 10 days. In both transgenic lines regeneration proceeded normally and no GFP expression was observed in the first days of regeneration (**Supplementary Fig. 7**). Expression in ovaries was first detected at day 8 after amputation, and in testes at day 10 after amputation (**Supplementary Fig. 7**). Thus, tissue-specific transgene expression is restored during regeneration, as expected for a regular genomic locus.

DISCUSSION

Free-living regeneration-capable flatworms are powerful model organisms to study mechanisms of regeneration and stem cell regulation (Elliott and Sanchez Alvarado, 2013; Rink, 2013). Currently, the most popular flatworms among researchers are the planarian species *S. mediterranea* and *D. japonica* (Rink, 2013). A method for generating transgenic animals in the planarian *Girardia tigrina* was reported in 2003 (González-Estévez et al., 2003), but despite substantial ongoing efforts by the planarian research community it has thus far not been reproduced in either *S. mediterranea* or *D. japonica*. The lack of transgenesis represents a significant experimental limitation of the planarian model systems. Primarily for this reason we focused on developing an alternative, non-planarian flatworm model, *Macrostomum lignano*. We reasoned that the fertilized one-cell stage eggs, which are

readily available in this species, will facilitate development of the transgenesis method, leveraging the accumulated experience on transgenesis in other model organisms.

In this study, we demonstrate a reproducible transgenesis approach in *M. lignano* by microinjection and random integration of DNA constructs. Microinjection is the method of choice for creating transgenic animals in many species and allows delivery of the desired material into the egg, whether it is RNA, DNA or protein (Sato et al., 2016). Initially, we tried transposon- and meganuclease-mediated approaches for integration of foreign DNA in the genome, but found in the course of the experiments that instead, random integration is a more efficient way for DNA incorporation in *M. lignano*. Random integration utilizes the molecular machinery of the host, integrating the provided DNA without the need for any additional components (Yan et al., 2013). The method has its limitations, since the location and the number of integrated transgene copies cannot be controlled, and integration in a functional site can cause unpredictable disturbances and variation in transgene expression (Yan et al., 2013). Indeed, we observed differences in the expression levels between independent transgenic lines for the *EFA* transgene reporter (**Supplementary Fig. 8**).Transgene silencing might occur in a copy-dependent manner, as is the case in the germline of *C. elegans* (Kelly et al., 1997). However, the fact that we readily obtained transgenic lines with germline-specific expression (**Fig. 3c-e**) indicates that germline transgene silencing is not a major issue in *M. lignano*.

The efficiency of integration and germline transmission varied between 1% and 8% of injected eggs in our experiments (**Table 1**), which is reasonable, given that a skilled person can inject up to 50 eggs in one hour. Although injection of a circular plasmid carrying a transgene can result in integration and germline transmission with acceptable efficiency (e.g. line NL23, **Table 1**), we found that injection of vector-free (Etchberger and Hobert, 2008b) transgenes followed by ionizing irradiation of injected embryos with a dose of 2.5 Gy gave more consistent results (**Table 1**). Irradiation is routinely used in *C. elegans* for integration of extrachromosomal arrays, presumably by creating DNA breaks and inducing non-homologous recombination (Mello and Fire, 1995). While irradiation can have deleterious consequences by inducing mutations, in our experiments we have not observed any obvious phenotypic deviations in the treated animals and their progeny. Nevertheless, for the downstream genetic analysis involving transgenic lines, several rounds of backcrossing to non-irradiated stock might be required to remove any introduced mutations, which is easily possible given that these worms are outcrossing and have a short generation time (Marie-Orleach et al., 2016, 2017). Despite the mentioned limitations, random integration of foreign DNA appears to be a straightforward and productive approach for generating transgenic lines in *M. lignano* and can be used as a basis for further development of more controlled transgenesis methods in this animal, including transposon-based (Ivics et al., 2009), integrase-based (Fogg et al., 2014), homology-based (Gerlai, 2016) or CRISPR/Cas9-based (Komor et al., 2016) approaches.

The draft genome assembly of the *M. lignano* DV1 line, which is also used in this study, was recently published (Wasik et al., 2015). The genome appeared to be difficult to assemble and even the 130x coverage of PacBio data resulted in the assembly with N50 of only 64 Kb (Wasik et al., 2015), while in other species N50 in the range of several megabases is usually achieved with such

PacBio data coverages (Chin et al., 2016). By adding Illumina and 454 data and using a different assembly algorithm, we have generated a substantially improved draft genome assembly, Mlig_3_7, with N50 scaffold size of 245.9 Kb (**Supplementary Table 4**). The difficulties with the genome assembly stem from the unusually high fraction of simple repeats and transposable elements in the genome of *M. lignano* (Wasik et al., 2015). Furthermore, it was shown that *M. lignano* has a polymorphic karyotype and the DV1 line used for genome sequencing has additional large chromosomes (Zadesenets et al., 2016, and **Supplementary Fig. 3**), which further complicates the assembly. The chromosome duplication also complicates genetic analysis and in particular gene knockout studies. To address these issues, we have established a different wild-type *M. lignano* line, NL10, from animals collected in the same geographical location as DV1 animals. The NL10 line appears to have no chromosomal duplications or they are present at a very low rate in the population, and its measured genome size is 500 Mb (**Supplementary Fig. 3**). While the majority of transgenic lines reported here are derived from the DV1 wild-type line, we observed similar transgenesis efficiency when using the NL10 line (**Table1,** line NL24). Therefore, we suggest that NL10 line is a preferred line for future transgenesis applications in *M. lignano*.

To facilitate the selection of promoter regions for transgenic reporter constructs, we have generated Mlig_RNA_3_7 transcriptome assembly, which incorporates information from 5´- and 3´-specific RNA-seq libraries, as well as trans-splicing signals, to accurately define gene boundaries. We integrated genome assembly, annotation and expression data using the UCSC genome browser software (**Supplementary Fig. 5**, http://gb.macgenome.org). For genes tested in this study, the regions up to 2 kb upstream of the transcription start sites are sufficient to faithfully reflect tissue-specific expression patterns of these genes (**Fig. 3**), suggesting the preferential proximal location of gene regulatory elements, which will simplify analysis of gene regulation in *M. lignano* in the future.

In conclusion, we demonstrate that transgenic *M. lignano* animals can be generated with a reasonable success rate under a broad range of conditions, from circular and linear DNA fragments, with and without irradiation, as single and double reporters, and for multiple promoters, suggesting that the technique is robust. Similar to transgenesis in *C. elegans, Drosophila* and mouse, microinjection is the most critical part of the technique and requires skill that can be developed with practice. The generated genomic resources and the developed transgenesis approach provide a technological platform for harvesting the power of *M. lignano* as an experimental model organism for research on stem cells and regeneration.

METHODS

M. lignano lines and cultures

The DV1 inbred *M. lignano* line used in this study was described previously (Janicke et al., 2013; Wasik et al., 2015; Zadesenets et al., 2016). The NL10 line was established from 5 animals collected near Lignano, Italy. Animals were cultured under laboratory conditions in plastic Petri dishes (Greiner), filled with nutrient enriched artificial sea water (Guillard's f/2 medium). Worms were fed ad libitum on the unicellular diatom *Nitzschia curvilineata* (Heterokontophyta, Bacillariophyceae) (SAG, Göttingen, Germany). Climate chamber conditions were set on 20°C with constant aeration, a 14/10h day/night cycle.

Cloning of the elongation factor 1 alpha promoter

The *M. lignano EFA* promoter sequence was obtained by inverse PCR. Genomic DNA was isolated using a standard phenol-chloroform protocol; fully digested by XhoI and subsequently self-ligated overnight (1ng/ul). Diluted self-ligated gDNA was used for inverse PCR using the *EFA* specific primers Efa_IvPCR_rv3 TCTCGAACTTCCACAGAGCA and Efa_IvPCR_fw3 CAAGAAGGAGGAGAC CACCA. Subsequently, nested PCR was performed using the second primer pair Efa_IvPCR_rv2 AAGCTCCTGTGCCTCCTTCT and Efa_IvPCR_fw2 AGGTCAAGTCCGTCGAAATG. The obtained fragment was cloned into p-GEM-T and sequenced. Later on, the obtained sequence was confirmed with the available genome data. Finally, the obtained promoter sequence was cloned into two different plasmids: the MINOS plasmid (using EcoRI / NcoI) and the I-SceI plasmid (using PacI / AscI).

Codon optimization

Highly-expressed transcripts were identified from RNA-seq data (Grudniewska et al., 2016) and codon weight matrices were calculated using the 100 most abundantly expressed non-redundant genes. *C. elegans* Codon Adapter code (Redemann et al., 2011) was adapted for *M. lignano* (http://www.macgenome.org/codons) and used to design codon-optimized coding sequences (**Supplementary Table 7)**. Gene fragments (IDT, USA) containing codon-optimized sequences, *EFA* 3'UTR and restriction cloning sites, were inserted into the pCS2+ vector to create optiMac plasmids used in the subsequent promoter cloning.

Cloning of tissue-specific promoters

Promoters were selected using Mlig_3_7 as well as several earlier *M. lignano* genome assemblies and MLRNA1509 transcriptome assembly (Grudniewska et al., 2016). RAMPAGE signal was used to identify the transcription start site and an upstream region of 1-2.5 kb was considered to contain the promoter sequence. An artificial ATG was introduced after the presumed transcription start site. This ATG was in-frame with the GFP of the target vector. The selected regions were cloned into optiMac vector using HindIII and BglII sites. Primers and cloned promoter sequences are provided in **Supplementary Table 7.**

Preparation and collection of eggs

Worms used for egg laying were kept in synchronized groups of roughly 500 per plate and transferred twice per week to prevent mixing with newly hatching offspring. The day before microinjections, around 1000 worms from 2 plates were combined (to increase number of eggs laid per plate) and transferred to plates with fresh f/2 medium and no food (to remove the leftover food from the digestive tracks of the animals as food debris can attach to the eggs and impair the microinjections by clogging needles and sticking to holders). On the day of the injections, worms were once again transferred to fresh f/2 without food to remove any debris and eggs laid overnight. Worms were kept in the dark for 3 hours and then transferred to light. After 30 minutes in the light, eggs were collected using plastic pickers made from microloader tips (Eppendorf, Germany), placed on a glass slide in a drop of f/2 and aligned in a line for easier handling.

Needle preparation

Needles used in the microinjection procedure were freshly pulled using either borosilicate glass capillaries with filament (BF100-50-10, Sutter Instrument, USA) or aluminosilicate glass capillaries with filament (AF100-64-10, Sutter Instrument, USA) on a Sutter P-1000 micropipette puller (Sutter Instrument, USA) with the following settings: Heat=ramp-34, Pull=50, Velocity=70, Time=200, Pressure=460 for borosilicate glass and Heat=ramp, Pull=60, Velocity=60, Time=250, Pressure=500 for aluminosilicate glass. The tips of the needles were afterwards broken and sharpened using a MF-900 microforge (Narishige, Japan). Needles were loaded using either capillary motion or microloader tips (Eppendorf, Germany). Embryos were kept in position using glass holders pulled from borosilicate glass capillaries without a filament (B100-50-10, Sutter Instrument, USA) using P-1000 puller with the following settings: Heat=ramp+18, Pull=0, Velocity=150, Time=115, Pressure=190. The holders where broken afterwards using a MF-900 microforge to create a tip of approximately 140 μm outer diameter and 50 μm inner diameter. Tips were heat-polished to create smooth edges and bent to a ~20° angle.

Microinjections

All microinjections were carried out on fresh one-cell stage *M. lignano* embryos. An AxioVert A1 inverted microscope (Carl Zeiss, Germany) equipped with a PatchMan NP2 for the holder and a TransferMan NK2 for the needle (Eppendorf, Germany) was used to perform all of the micromanipulations. A FemtoJet express (Eppendorf, Germany), with settings adjusted manually based on the amount of mucous and debris surrounding the embryos, was used as the pressure source for microinjections. A PiezoXpert (Eppendorf, Germany) was used to facilitate the penetration of the eggshell and the cell membrane of the embryo.

Irradiation

Irradiation was carried out using a IBL637 Caesium-137 source (CISbio International, France). Embryos were exposed to 2.5 Gy of γ-radiation within 1 hour post injection.

Establishing transgenic lines

Positive hatchlings (P_0) were selected based on the presence of fluorescence and transferred into single wells of a 24 well-plate. They were then crossed with single wild-type worms that were raised in the same conditions. The pairs were transferred to fresh food every 2 weeks. Positive F_1 animals from the same P_0 cross were put together on fresh food and allowed to generate F_2 progeny. After the population of positive F_2 progeny grew to over 200 hatchlings, transgenic worms were singled out and moved to a 24 well plate. The selected worms were then individually back-crossed with wild type worms to distinguish F_2 animals homozygous and heterozygous for the transgene. The transgenic F_2 worms that gave only positive progeny in the back-cross (at least 10 progeny observed) were assumed to be homozygous, singled out, moved to fresh food and allowed to lay eggs for another month to purge whatever remaining wild-type sperm from the back-cross. After the homozygous F_2 animals stopped producing new offspring, they were crossed to each other to establish a new transgenic line. The lines were named according to guidelines established at http://www.macgenome.org/nomenclature.html.

Microscopy

Images were taken using a Zeiss Axio Zoom V16 microscope with an HRm digital camera and Zeiss filter sets 38HE (FITC) and 43HE (dsRed), an Axio Scope A1 with a MRc5 digital camera or an Axio Imager M2 with an MRm digital camera.

Southern blot analysis

Southern blots were done using the DIG-System (Roche), according to the manufacturer's manual with the following parameters: vacuum transfer at 5 Hg onto positively charged nylon membrane for 2 h, UV cross-linking 0.14 J/cm^2, overnight hybridization at 68ºC.

Identification of transgene integration sites

The Universal GenomeWalker 2.0 Kit (Clontech Laboratories, USA) with restriction enzymes StuI and BamHI was used according to the manufacturer's protocol. Sanger sequencing of PCR products was performed by GATC Biotech (Germany).

Whole mount *in situ* hybridization

cDNA synthesis was carried out using the SuperScript III First-Strand Synthesis System (Life Technologies, USA), following the protocol supplied by the manufacturer. 2µg of total RNA were used as a template for both reactions: one with oligo(dT) primers and one with hexamer random primers. Amplification of selected DNA templates for ISH probes was performed by standard PCR with GoTaq Flexi DNA Polymerase (Promega, USA). Amplified fragments were cloned into pGEM-T vector system (Promega, USA) and validated by Sanger sequencing. Primers used for amplification are listed in **Supplementary Table 7.** Templates for riboprobes were amplified from sequenced plasmids using High Fidelity Pfu polymerase (Thermo Scientific, USA). pGEM-T backbone binding primers: forward (5'-CGGCCGCCATGGCCGCGGGA-3') and reversed (5'TGCAGGCGGCCGCACTAGTG-3') and

versions of the same primers with an upstream T7 promoter sequence (5'-GGATCCTAATACGACTCACTATAGG-3'. Based on the orientation of the insert in the vector either forward primer with T7 promoter and reverse without or vice versa, were used to amplify ISH probe templates. Digoxigenin (DIG) labelled RNA probe synthesis was performed using the DIG RNA labelling Mix (Roche, Switzerland) and T7 RNA polymerase (Promega, USA) following the manufacturer protocol. The concentration of all probes was assessed with the Qubit RNA BR assay (Invitrogen). Probes were then diluted in Hybridization Mix (Pfister et al., 2007) (20 ng/µl), and stored at −80°C. The final concentration of the probe and optimal hybridization temperature were optimized for every probe separately. Whole mount in situ hybridization was performed following a published protocol(Pfister et al., 2007). Pictures were taken using a standard light microscope with DIC optics and an AxioCam HRC (Zeiss, Germany) digital camera.

Karyotyping

DV1 and NL10 worms were cut above the testes and left to regenerate for 48 hours to increase the amount of dividing cells. Head fragments were collected and treated with 0.2% colchicine in f/2 (Sigma, C9754-100mg) for 4 hours at 20°C to arrest cells in mitotic phase. Head fragments were then collected and treated with 0.2% KCl as hypotonic treatment for 1 hour at room temperature. Fragments were then put on SuperfrostPlus slides (Fisher, 10149870) and macerated using glass pipettes while being in Fix 1 solution (H_2O : EtOH : glacial acetic acid 4:3:3). The cells were then fixed by treatment with Fix 2 solution (EtOH : glacial acetic acid 1:1) followed by Fix 3 solution (100% glacial acetic acid), before mounting by using Vectashield with Dapi (Vectorlabs, H-1200). At least three karyotypes were observed per worm and 20 worms were analyzed per line.

Genome size measurements

Genome size of the DV1 and NL10 lines was determined using flow cytometry approach(Hare and Johnston, 2011). In order eliminate the residual diatoms present in the gut, animals were starved for 24h. For each sample 100 worms were collected in an Eppendorf tube. Excess f/2 was aspirated and worms were macerated in 200µl 1x Accutase (Sigma, A6964-100ML) at room temperature for 30 minutes, followed by tissue homogenization through pipetting. 800µl f/2 was added to the suspension and cells were pelleted by centrifugation at 4°C, 1000 rpm, 5 min. The supernatant was aspirated and the cell pellet was resuspended in the nuclei isolation buffer (100 mM Tris-HCl pH 7.4, 154 mM NaCl, 1 mM $CaCl_2$, 0.5 mM $MgCl_2$, 0.2% BSA, 0.1% NP-40 in MilliQ water). The cell suspension was passed through a 35 µm pore size filter (Corning, 352235) and treated with RNase A and 10 mg/ml PI for 15 minutes prior to measurement. *Drosophila* S2 cells (gift from O. Sibon lab) and chicken erythrocyte nuclei (CEN, BioSure, 1006, genome size 2.5 pg) were included as references. The S2 cells were treated in the same way as *Macrostomum* cells. The CEN were resuspended in PI staining buffer (50 mg/ml PI, 0.6% NP-40 in Calcium and Magnesium free Dulbecco's PBS Life Technologies, 14190136). Fluorescence was measured on a BD FacsCanto II Cell Analyzer first separately for all samples and then samples were combined based on the amount of cells to obtain an even distribution

of different species. The combined samples were re-measured and genome sizes calculated using CEN as a reference and S2 as positive controls (**Supplementary Fig. 2**).

Preparation of genomic libraries

One week prior to DNA isolation animals were kept on antibiotic-containing medium. Medium was changed every day with 50 µg/ml streptomycin or ampicillin added in alternating fashion. Worms were starved 24 hours prior to extraction, and then rinsed in fresh medium. Genomic DNA was extracted using the USB PrepEase Genomic DNA Isolation kit (USB-Affymetrix, Cat. No. 78855) according to manufacturer's instructions. For the lysis step worms were kept in the supplied lysis buffer (with Proteinase K added) at 55 °C for 30-40 minutes and mixed by inverting the tube every 5 minutes. DNA was ethanol-precipitated once following the extraction and resuspended in TE buffer (for making 454 libraries Qiagen EB buffer was used instead). Concentration of DNA samples was measured with the Qubit dsDNA BR assay kit (Life Technologies, Cat. No. Q32850).

454 shotgun DNA libraries were made with the GS FLX Titanium General Library Preparation Kit (Roche, Cat. No. 05233747001), and for paired-end libraries the set of GS FLX Titanium Library Paired End Adaptors (Roche, Cat. No. 05463343001) was used additionally. All the libraries were made following the manufacturer's protocol and sequenced on 454 FLX and Titanium systems.

Illumina paired-end genomic libraries were made with the TruSeq DNA PCR-free Library Preparation Kit (Ilumina, Cat. No. FC-121-3001) following the manufacturer's protocol. Long-range mate-pair libraries were prepared with the Nextera Mate Pair Sample Preparation Kit (Illumina, Cat. No. FC-132-1001) according to manufacturer's protocol. Libraries were sequenced on the Illumina HiSeq 2500 system.

Genome assembly

PacBio data (acc. SRX1063031) were assembled with Canu (Koren et al., 2017) v. 1.4 with default parameters, except the errorRate was set to 0.04. The resulting assembly was polished with Pilon(Walker et al., 2014) v. 1.20 using Illumina shotgun data mapped by Bowtie (Langmead and Salzberg, 2012) v. 2.2.9 and RNA-seq data mapped by STAR (Dobin et al., 2013) v. 2.5.2b. Next, scaffolding was performed by SSPACE (Boetzer et al., 2011) v. 3.0 using paired-end and mate-pair Illumina and 454 data. Mitochondrial genome of *M. lignano* was assembled separately from raw Illumina reads using the MITObim software (Hahn et al., 2013) and the *Dugesia japonica* complete mitochondrial genome (acc. NC_016439.1) as a reference. The assembled mitochondrial genome differed from the recently published *M. lignano* mitochondrial genome(Egger et al., 2017) (acc. no. MF078637) in just 1 nucleotide in an intergenic spacer region. The genome assembly scaffolds containing mitochondrial sequences were filtered out and replaced with the separately assembled mitochondrial genome sequence. The final assembly was named Mlig_3_7. Genome assembly evaluation was performed with REAPR (Hunt et al., 2013) and FRCbam (Vezzi et al., 2012) software using HUB1_300 paired-end library and DV1-6kb-1, HUB1-3_6kb, HUB1-3_7kb, ML_8KB_1 and ML_8KB_2 mate-pair libraries (**Supplementary Table 2**).

Transcriptome assembly

Previously published *M. lignano* RNA-seq data (Cannon et al., 2016; Grudniewska et al., 2016) (SRP082513, SRR2682326) and the *de novo* transcriptome assembly MLRNA150904 (Grudniewska et al., 2016) were used to generate an improved genome-guided transcriptome assembly. First, trans-splicing and polyA-tail sequences were trimmed from MLRNA150904 and the trimmed transcriptome was mapped to the Mlig_3_7 genome assembly by BLAT (Kent, 2002) v. 36x2 and hits were filtered using the pslCDnaFilter tool with the parameters "-ignoreNs -minId=0.8 -globalNearBest=0.01 -minCover=0.95 –bestOverlap". Next, RNA-seq data to were mapped to genome by STAR (Dobin et al., 2013) v. 2.5.2b with parameters "--alignEndsType EndToEnd --twopassMode Basic --outFilterMultimapNmax 1000". The resulting bam files were provided to StringTie (Pertea et al., 2015) v. 1.3.3 with the parameter "--rf", and the output was filtered to exclude lowly expressed antisense transcripts by comparing transcripts originating from the opposite strands of the same genomic coordinates and discarding those from the lower-expressing strand (at least 5-fold read count difference). The filtered StringTie transcripts were merged with the MLRNA150904 transcriptome mappings using meta-assembler TACO (Niknafs et al., 2016) with parameters "--no-assemble-unstranded --gtf-expr-attr RPKM --filter-min-expr 0.01 --isoform-frac 0.75 --filter-min-length 100" and novel transcripts with RPKM less than 0.5 and not overlapping with MLRNA150904 mappings were discarded. The resulting assembled transcripts were termed 'Transcriptional Units' and the assembly named Mlig_RNA_3_7_DV1.TU. To reflect closely related transcripts in their names, sequences were clustered using cd-hit-est from the CD-HIT v. 4.6.1 package (Fu et al., 2012) with the parameters "-r 0 -c 0.95 -T 0 -M 0", and clustered transcripts were given the same prefix name. Close examination of the transcriptional units revealed that they often represented precursor mRNA for trans-splicing and contained several genes. Therefore, further processing of the transcriptional units to identified boundaries of the encoded genes was required. For this, we developed computational pipeline TBONE (Transcript Boundaries based ON experimental Evidence), which utilizes exclusively experimental data to determine precise 5' and 3' ends of trans-spliced mRNAs. Raw RNA-seq data were parsed to identify reads containing trans-splicing sequences, which were trimmed, and the trimmed reads were mapped to the genome assembly using STAR (Dobin et al., 2013). The resulting wiggle files were used to identify signal peaks corresponding to sites of trans-splicing. Similarly, for the identification of polyadenylation sites we used data generated previously (Grudniewska et al., 2016) with CEL-seq library construction protocol and T-fill sequencing method. All reads originating from such an approach correspond to sequences immediately upstream of poly(A) tails and provide exact information on 3'UTR ends of mRNAs. The generated trans-splicing and poly(A)_ signals were overlapped with genomic coordinates of transcriptional units by TBONE, 'cutting' transcriptional units into processed mRNAs with exact gene boundaries, where such experimental evidence was available. Finally, coding potential of the resulting genes was estimated by TransDecoder (Haas et al., 2013), and transcripts containing ORFs but missing a poly(A) signal and followed by transcripts without predicted ORF but with poly(A) signal were merged if the distance between the transcripts was not greater than 10kb and the spanning region was repetitive.

Annotation of transposable elements and genomic duplications

Two methods were applied to identify repetitive elements *de novo* both from the raw sequencing data and from the assembled scaffolds. Tedna software (Zytnicki et al., 2014) v. 1.2.1 was used to assemble transposable element models directly from the repeated fraction of raw Illumina paired-end sequencing reads with the parameters "-k 31 -i 300 -m 200 -t 37 --big-graph=1000". To mine repeat models directly from the genome assembly, RepeatModeler package (http://www.repeatmasker.org) was used with the default settings. Identified repeats from both libraries were automatically annotated using RepeatClassifier perl script from the RepeatModeler package against annotated repeats represented in the Repbase Update – RepeatMasker edition database (Jurka et al., 2005) v. 20170127. Short (< 200 bp) and unclassified elements were filtered out from both libraries. Additional specific *de novo* screening for full-length long terminal repeats (LTR) retrotransposons was performed using the LTRharvest tool (Ellinghaus et al., 2008) with settings "-seed 100 -minlenltr 100 -maxlenltr 3000 -motif tgca -mindistltr 1000 -maxdistltr 20000 -similar 85.0 -mintsd 5 -maxtsd 20 -motifmis 0 -overlaps all". Identified LTR retrotransposons were then classified using the RepeatClassifier perl script filtering unclassified elements. Generated repeat libraries were merged together with the RepeatMasker (Jurka et al., 2005) library v. 20170127. The resulted joint library was mapped on the genome assembly with RepeatMasker. Tandem repeats were annotated and masked with Tandem Repeat Finder (Benson, 1999) with default settings. Finally, to estimate overall repeat fraction of the assembly, the Red *de novo* repeat annotation tool (Girgis, 2015) with default settings was applied.

To identify duplicated non-repetitive fraction of the genome, repeat-masked genome assembly was aligned against itself using LAST software (Kiełbasa et al., 2011), and aligned non-self blocks longer than 500 nt and at least 95% identical were calculated.

Data availability

All raw data were submitted to NCBI Sequence Read Archive under accession numbers SRX2866466-SRX2866494. Annotated genome assembly has been deposited at DDBJ/ENA/GenBank under the accession NIVC00000000. The version described in this paper is version NIVC01000000.

AUTHOR CONTRIBUTIONS

E.B., P.L. and L.S. conceived the project. E.B. supervised the project and provided resources. J.Wudarski, K.M., T.D., D.S., P.W., M.Gre, K.U. made constructs and performed transgenesis. J.Wudarski optimized transgenesis efficiency. L.G., F.B., M.Gre., M.Gru and D.V. maintained *M. lignano* cultures. D.O. and L.G. established the NL10 line. D.S. and M.Gru generated genomic and RAMPAGE libraries. A.A., W.Q., L.S., E.B. contributed to sequencing genomic libraries. F.B. and S.M. performed genome size measurement and karyotyping. V.G. and E.B. performed genome and transcriptome assemblies and annotation. K.U. performed transposon annotation. J.Wudarski, D.S. and J.Wunderer performed *in situ* hybridizations. J.Wudarski and E.B. wrote the manuscript. All authors read the manuscript and provided edits.

REFERENCES

Batut, P., Dobin, A., Plessy, C., Carninci, P., and Gingeras, T.R. (2013). High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. *23*, 169–180.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Bindels, D.S., Haarbosch, L., van Weeren, L., Postma, M., Wiese, K.E., Mastop, M., Aumonier, S., Gotthard, G., Royant, A., Hink, M.A., et al. (2016). mScarlet: a bright monomeric red fluorescent protein for cellular imaging. Nat. Methods *14*, 53–56.

Boetzer, M., Henkel, C. V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

Cannon, J.T., Vellutini, B.C., Smith, J., Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. Nature *530*, 89–93.

Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050–1054.

Davies, E.L., Lei, K., Seidel, C.W., Kroesen, A.E., McKinney, S.A., Guo, L., Robb, S.M., Ross, E.J., Gotting, K., and Sánchez Alvarado, A. (2017). Embryonic origin of adult stem cells required for tissue homeostasis and regeneration. Elife *6*.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Egger, B., Ladurner, P., Nimeth, K., Gschwentner, R., and Rieger, R. (2006). The regeneration capacity of the flatworm Macrostomum lignano - On repeated regeneration, rejuvenation, and the minimal size needed for regeneration. Dev. Genes Evol. *216*, 565–577.

Egger, B., Bachmann, L., and Fromm, B. (2017). Atp8 is in the ground pattern of flatworm mitochondrial genomes. BMC Genomics *18*, 414.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.

Elliott, S.A., and Sanchez Alvarado, A. (2013). The history and enduring contributions of planarians to the study of animal regeneration. Wiley Interdiscip. Rev. Dev. Biol. *2*, 301–326.

Etchberger, J.F., and Hobert, O. (2008). Vector-free DNA constructs improve transgene expression in C. elegans. Nat. Methods *5*, 3.

Fogg, P.C.M., Colloms, S., Rosser, S., Stark, M., and Smith, M.C.M. (2014). New applications for phage integrases. J. Mol. Biol. *426*, 2703–2716.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Gerlai, R. (2016). Gene targeting using homologous recombination in embryonic stem cells: The future for behavior genetics? Front. Genet. *7*, 1–10.

Girgis, H.Z. (2015). Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics *16*, 227.

González-Estévez, C., Momose, T., Gehring, W.J., and Saló, E. (2003). Transgenic planarian lines obtained by electroporation using transposon-derived vectors and an eye-specific GFP marker. Proc. Natl. Acad. Sci. U. S. A. *100*, 14046–14051.

Grudniewska, M., Mouton, S., Simanov, D., Beltman, F., Grelling, M., de Mulder, K., Arindrarto, W., Weissert, P.M., van der Elst, S., and Berezikov, E. (2016). Transcriptional signatures of somatic neoblasts and germline cells in *Macrostomum lignano*. Elife *5*, 1–23.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads--a baiting and iterative mapping approach. Nucleic Acids Res. *41*, e129.

Hare, E.E., and Johnston, J.S. (2011). Genome size determination using flow cytometry of propidium iodide-stained nuclei. Methods Mol. Biol. *772*, 3–12.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T.D. (2013). REAPR: a universal tool for genome assembly evaluation. Genome Biol. *14*, R47.

Ivics, Z., Li, M.A., Mates, L., Boeke, J.D., Nagy, A., Bradley, A., and Izsvak, Z. (2009). Transposon-mediated genome manipulation in vertebrates. Nat. Methods *6*, 415–422.

Janicke, T., Marie-Orleach, L., De Mulder, K., Berezikov, E., Ladurner, P., Vizoso, D.B., and Schärer, L. (2013). Sex allocation adjustment to mating group size in a simultaneous hermaphrodite. Evolution *67*, 3233–3242.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Kanda, T., Sullivan, K.F., and Wahl, G.M. (1998). Histone-GFP fusion protein enables sensitive analysis of chromosome dynamics in living mammalian cells. Curr. Biol. *8*, 377–385.

Kelly, W.G., Xu, S., Montgomery, M.K., and Fire, A. (1997). Distinct requirements for somatic and germline expression of a generally expressed Caenorhabditis elegans gene. Genetics *146*, 227–238.

Kent, W.J. (2002). BLAT — The BLAST -Like Alignment Tool. Genome Res. *12*, 656–664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Res. *21*, 487–493.

Komor, A.C., Badran, A.H., and Liu, D.R. (2016). CRISPR-Based Technologies for the Manipulation

of Eukaryotic Genomes. Cell *168*, 1–17.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. Genome Res. *27*, 722–736.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359.

Marie-Orleach, L., Janicke, T., Vizoso, D.B., Eichmann, M., and Schärer, L. (2014). Fluorescent sperm in a transparent worm: validation of a GFP marker to study sexual selection. BMC Evol. Biol. *14*, 148.

Marie-Orleach, L., Janicke, T., Vizoso, D.B., David, P., and Schärer, L. (2016). Quantifying episodes of sexual selection: Insights from a transparent worm with fluorescent sperm. Evolution (N. Y). *70*, 314–328.

Marie-Orleach, L., Vogt-Burri, N., Mouginot, P., Schlatter, A., Vizoso, D.B., Bailey, N.W., and Schärer, L. (2017). Indirect genetic effects and sexual conflicts: Partner genotype influences multiple morphological and behavioral reproductive traits in a flatworm. Evolution (N. Y).

Mello, C., and Fire, A. (1995). DNA transformation. Methods Cell Biol. *48*, 451–482.

Morris, J., Nallur, R., Ladurner, P., Egger, B., Rieger, R., and Hartenstein, V. (2004). The embryonic development of the flatworm Macrostomum sp. Dev. Genes Evol. *214*, 220–239.

Mouton, S., Willems, M., Braeckman, B.P., Egger, B., Ladurner, P., Schärer, L., and Borgonie, G. (2009). The free-living flatworm Macrostomum lignano: A new model organism for ageing research. Exp. Gerontol. *44*, 243–249.

De Mulder, K., Kuales, G., Pfister, D., Egger, B., Seppi, T., Eichberger, P., Borgonie, G., and Ladurner, P. (2010). Potential of Macrostomum lignano to recover from gamma-ray irradiation. Cell Tissue Res. *339*, 527–542.

Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M., and Iyer, M.K. (2016). TACO produces robust multisample transcriptome assemblies from RNA-seq. Nat. Methods *14*, 68–70.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Pfister, D., De Mulder, K., Philipp, I., Kuales, G., Hrouda, M., Eichberger, P., Borgonie, G., Hartenstein, V., and Ladurner, P. (2007). The exceptional stem cell system of Macrostomum lignano: screening for gene expression and studying cell proliferation by hydroxyurea treatment and irradiation. Front. Zool. *4*, 9.

Redemann, S., Schloissnig, S., Ernst, S., Pozniakowsky, A., Ayloo, S., Hyman, A.A., and Bringmann, H. (2011). Codon adaptation-based control of protein expression in C. elegans. Nat. Methods *8*, 250-U94.

Rink, J.C. (2013). Stem cell systems and regeneration in planaria. Dev. Genes Evol. *223*, 67–84.

Sato, M., Ohtsuka, M., Watanabe, S., and Gurumurthy, C.B. (2016). Nucleic acids delivery methods for genome editing in zygotes and embryos: the old, the new, and the old-new. Biol. Direct *11*, 16.

Scimone, M.L., Kravarik, K.M., Lapan, S.W., and Reddien, P.W. (2014). Neoblast specialization in regeneration of the planarian schmidtea mediterranea. Stem Cell Reports *3*, 339–352.

Shaner, N.C., Lambert, G.G., Chammas, A., Ni, Y., Cranfill, P.J., Baird, M.A., Sell, B.R., Allen, J.R., Day, R.N., Israelsson, M., et al. (2013). A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. Nat. Methods *10*, 407–409.

Simanov, D., Mellaart-Straver, I., Sormacheva, I., and Berezikov, E. (2012). The flatworm

macrostomum lignano is a powerful model organism for ion channel and stem cell research. Stem Cells Int. *2012*.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M., and Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *26*, 1134–1144.

Tanaka, E.M., and Reddien, P.W. (2011). The Cellular Basis for Animal Regeneration. Dev. Cell *21*, 172–185.

Thermes, V., Grabher, C., Ristoratore, F., Bourrat, F., Choulika, A., Wittbrodt, J., and Joly, J.-S. (2002). I-SceI meganuclease mediates highly efficient transgenesis in fish. Mech. Dev. *118*, 91–98.

Umesono, Y., Tasaki, J., Nishimura, K., Inoue, T., and Agata, K. (2011). Regeneration in an evolutionarily primitive brain - the planarian Dugesia japonica model. Eur. J. Neurosci. *34*, 863–869.

Vezzi, F., Narzisi, G., and Mishra, B. (2012). Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. PLoS One *7*.

Wagner, D.E., Wang, I.E., Reddien, P.W., Evans, M.J., Evans, M.J., Kaufman, M.H., Martin, G.R., Wagers, A.J., Sherwood, R.I., Christensen, J.L., et al. (2011). Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. Science *332*, 811–816.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*.

Wasik, K., Gurtowski, J., Zhou, X., Ramos, O.M., Delás, M.J., Battistoni, G., El Demerdash, O., Falciatori, I., Vizoso, D.B., Smith, A.D., et al. (2015). Genome and transcriptome of the regeneration-competent flatworm, Macrostomum lignano. Proc. Natl. Acad. Sci. *112*, 201516718.

Van Wolfswinkel, J.C., Wagner, D.E., and Reddien, P.W. (2014). Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. Cell Stem Cell *15*, 326–339.

Yan, B.W., Zhao, Y.F., Cao, W.G., Li, N., and Gou, K.M. (2013). Mechanism of random integration of foreign DNA in transgenic mice. Transgenic Res. *22*, 983–992.

Zadesenets, K.S., Vizoso, D.B., Schlatter, A., Konopatskaia, I.D., Berezikov, E., Sch??rer, L., and Rubtsov, N.B. (2016). Evidence for karyotype polymorphism in the free-living flatworm, macrostomum lignano, a model organism for evolutionary and developmental biology. PLoS One *11*, 1–24.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics *29*, 2669–2677.

Zytnicki, M., Akhunov, E., and Quesneville, H. (2014). Tedna: A transposable element de novo assembler. Bioinformatics *30*, 2656–2658.