The free-living flatworm, *Macrostomum lignano* has an impressive regenerative capacity. Following injury, it can regenerate almost an entirely new organism because of the presence of an abundant somatic stem cell population, the neoblasts. This set of unique properties makes many flatworms attractive organisms for studying the evolution of pathways involved in tissue self-renewal, cell-fate specification, and regeneration. The use of these organisms as models, however, is hampered by the lack of a well-assembled and annotated genome sequence, fundamental to modern genetic and molecular studies. Here we report the genomic sequence of *M. lignano* and an accompanying characterization of its transcriptome. The genome structure of *M. lignano* is remarkably complex, with ~75% of its sequence being comprised of simple repeats and transposon sequences. This has made high-quality assembly from Illumina reads alone impossible (N50 = 222 bp). We therefore generated 130x coverage by long sequencing reads from the Pacific Biosciences platform to create a substantially improved assembly with an N50 of 64 Kbp. We complemented the reference genome with an assembled and annotated transcriptome, and used both of these datasets in combination to probe gene-expression patterns during regeneration, examining pathways important to stem cell function.

**Flatworms**

Flatworms belong to the superphylum Lophotrochozoa, a vast assembly of protostome invertebrates (1, 2) (Fig. 1A). The evolutionary relationships within this clade are poorly resolved and the specific position of flatworms is currently debated (3, 4). Flatworms have attracted scientific attention for centuries because of their astonishing regenerative capabilities (5, 6), as well as their ability to “degrow” in a controlled way when starved (7). As far back as the early 1900s, Thomas Morgan recognized the potential of flatworms and conducted a number of fascinating regeneration experiments on planarian flatworms before his focus shifted to *Drosophila* genetics (8).

*Macrostomum lignano* is (Fig. 1B), a free-living, regenerating flatworm isolated from the coast of the Mediterranean Sea. *M. lignano* is an obligatorily cross-fertilizing simultaneous hermaphrodite (9) that is remarkably complex, with nearly its entire body following injury. The resources we present here will promote not only the studies of mechanisms of stem cell self-renewal, but also of regeneration and differentiation.

**Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano***

Kaja Wasik\(^a,\) James Gurtowski\(^a,\)\(^b,\) Xin Zhou\(^b,\) Olivia Mendivil Ramos\(^b,\) Osama El Demerdash\(^b,\) Ilaria Falcieri\(^c,\) Dita B. Vizoso\(^d,\) Andrew D. Smith\(^d,\) Peter Ladurner\(^d,\) Lukas Schärer\(^d,\) W. Richard McCombie\(^a,\) Gregory J. Hannon\(^a,\)\(^c,\)\(^d,\) and Michael Schatz\(^a,\)\(^d,\)\(^2\)

\(^a\)Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; \(^b\)Molecular and Cellular Biology Graduate Program, Stony Brook University, NY 11794; \(^c\)Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, United Kingdom; \(^d\)Department of Evolutionary Biology, Zoological Institute, University of Basel, 4051 Basel, Switzerland; \(^2\)Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089; and \(^d\)Department of Evolutionary Biology, Institute of Zoology and Center for Molecular Biosciences Innsbruck, University of Innsbruck, A-6020 Innsbruck, Austria

Contribution by Gregory J. Hannon, August 23, 2015 (sent for review June 25, 2015; reviewed by Ian Korf and Robert E. Steele)

The availability of high-quality genome and transcriptome assemblies is critical for enabling full exploitation of any model organism. Here we present genome and transcriptome assemblies for *Macrostomum lignano*, a free-living flatworm that can regenerate nearly its entire body following injury. The resources we present here will promote not only the studies of mechanisms of stem cell self-renewal, but also of regeneration and differentiation.

**Significance**


Reviewers: I.K., University of California, Davis; and R.E.S., University of California, Irvine.

The authors declare no conflict of interest.

Frequently available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the NCBI Sequence Read Archive database (accession no. SRP095533).

1.K.W. and J.G. contributed equally to this work.

2To whom correspondence may be addressed. Email: greg.hannon@cruk.cam.ac.uk or mschatz@cshl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516718112/-/DCSupplemental.
patterns by RNA-Seq under different physiological conditions or in different cell types. We have demonstrated this by profiling gene-expression patterns in worms following posthead establishment of the inbred DV1 line used for the genome assembly. Nevertheless, 92% of ESTs could be aligned to the genome with an average identity of 94.6%. Of reads derived from 1,248 BACs, 91% pooled and sequenced using Illumina mapped to the ML2 assembly, with identity over 99.5%. Additionally, a sample of 10% of contigs from the ML1 (Illumina only) assembly aligned to the ML2 (PacBio only) assembly, with 99.56% identity. Considered together, these results indicated a reliable local assembly, and our annotation results (see below) show a high-quality representation of overall genic composition.

Even with the long PacBio reads, there remained a possibility that the unusually high rate of low-complexity sequences still had a profound impact on our ability to assemble the *M. lignano* genome. To examine this possibility, we analyzed contig ends and found that 55% of them had more than 50% of their bases masked by Tandem Repeat Finder, suggesting a possible cause for the contigs to terminate. Because we had obtained 21x coverage by reads larger than 10 Kb, this indicated the presence of repeat tracts of at least this length. The low-complexity repeats showed sequence biases, with complexity sequences were present in libraries prepared from both whole worms and sorted proliferating S-phase cells (*SI Appendix, Fig. S1B*), and were enriched in the nontranscribed fraction of the genome (*SI Appendix, Fig. S1A*). Roughly 25% of the *Macrostomum* genome was comprised of simple repeats, far greater than the fraction observed in *Caenorhabditis elegans*, *Drosophila melanogaster*, *Schistosoma mansoni*, or the human genome (*SI Appendix, Fig. S1C*). This percentage seemed high enough to contribute to the poor quality of our initial assembly. To overcome this problem, we sequenced the *M. lignano* genome using the SMRT sequencing from Pacific Biosciences (PacBio; 130x genomic coverage). This technology can generate reads long enough to span many more repeat elements than can short reads, leading to reports of greatly improved assemblies of several species. After error correction we had 21x coverage of reads greater than 10 Kb in length; these reads were used in the final assembly (ML2). Use of the PacBio reads significantly improved the genome assembly compared with Illumina only (Fig. 2B), including improving the contig N50 size from 222 bp to 64 Kbp and the largest assembled contig from 144 Kbp to 627 Kbp.

To assess the quality and coverage of the ML2 assembly, *M. lignano* expressed sequence tags (ESTs) from public databases (25) and sequences derived from an arrayed bacterial artificial chromosome (BAC) library of the *M. lignano* genome were aligned to the assembly. The *M. lignano* ESTs were generated before establishment of the inbred DV1 line used for the genome assembly. Nevertheless, 92% of ESTs could be aligned to the genome with an average identity of 94.6%. Of reads derived from 1,248 BACs, 91% pooled and sequenced using Illumina mapped to the ML2 assembly, with identity over 99.5%. Additionally, a sample of 10% of contigs from the ML1 (Illumina only) assembly aligned to the ML2 (PacBio only) assembly, with 99.56% identity. Considered together, these results indicated a reliable local assembly, and our annotation results (see below) show a high-quality representation of overall genic composition.

Even with the long PacBio reads, there remained a possibility that the unusually high rate of low-complexity sequences still had a profound impact on our ability to assemble the *M. lignano* genome. To examine this possibility, we analyzed contig ends and found that 55% of them had more than 50% of their bases masked by Tandem Repeat Finder, suggesting a possible cause for the contigs to terminate. Because we had obtained 21x coverage by reads larger than 10 Kb, this indicated the presence of repeat tracts of at least this length. The low-complexity repeats showed sequence biases, with complexity sequences were present in libraries prepared from both whole worms and sorted proliferating S-phase cells (*SI Appendix, Fig. S1B*), and were enriched in the nontranscribed fraction of the genome (*SI Appendix, Fig. S1A*). Roughly 25% of the *Macrostomum* genome was comprised of simple repeats, far greater than the fraction observed in *Caenorhabditis elegans*, *Drosophila melanogaster*, *Schistosoma mansoni*, or the human genome (*SI Appendix, Fig. S1C*). This percentage seemed high enough to contribute to the poor quality of our initial assembly. To overcome this problem, we sequenced the *M. lignano* genome using the SMRT sequencing from Pacific Biosciences (PacBio; 130x genomic coverage). This technology can generate reads long enough to span many more repeat elements than can short reads, leading to reports of greatly improved assemblies of several species. After error correction we had 21x coverage of reads greater than 10 Kb in length; these reads were used in the final assembly (ML2). Use of the PacBio reads significantly improved the genome assembly compared with Illumina only (Fig. 2B), including improving the contig N50 size from 222 bp to 64 Kbp and the largest assembled contig from 144 Kbp to 627 Kbp.

To assess the quality and coverage of the ML2 assembly, *M. lignano* expressed sequence tags (ESTs) from public databases (25) and sequences derived from an arrayed bacterial artificial chromosome (BAC) library of the *M. lignano* genome were aligned to the assembly. The *M. lignano* ESTs were generated before establishment of the inbred DV1 line used for the genome assembly. Nevertheless, 92% of ESTs could be aligned to the genome with an average identity of 94.6%. Of reads derived from 1,248 BACs, 91% pooled and sequenced using Illumina mapped to the ML2 assembly, with identity over 99.5%. Additionally, a sample of 10% of contigs from the ML1 (Illumina only) assembly aligned to the ML2 (PacBio only) assembly, with 99.56% identity. Considered together, these results indicated a reliable local assembly, and our annotation results (see below) show a high-quality representation of overall genic composition.

Even with the long PacBio reads, there remained a possibility that the unusually high rate of low-complexity sequences still had a profound impact on our ability to assemble the *M. lignano* genome. To examine this possibility, we analyzed contig ends and found that 55% of them had more than 50% of their bases masked by Tandem Repeat Finder, suggesting a possible cause for the contigs to terminate. Because we had obtained 21x coverage by reads larger than 10 Kb, this indicated the presence of repeat tracts of at least this length. The low-complexity repeats showed sequence biases, with complexity sequences were present in libraries prepared from both whole worms and sorted proliferating S-phase cells (*SI Appendix, Fig. S1B*), and were enriched in the nontranscribed fraction of the genome (*SI Appendix, Fig. S1A*). Roughly 25% of the *Macrostomum* genome was comprised of simple repeats, far greater than the fraction observed in *Caenorhabditis elegans*, *Drosophila melanogaster*, *Schistosoma mansoni*, or the human genome (*SI Appendix, Fig. S1C*). This percentage seemed high enough to contribute to the poor quality of our initial assembly. To overcome this problem, we sequenced the *M. lignano* genome using the SMRT sequencing from Pacific Biosciences (PacBio; 130x genomic coverage). This technology can generate reads long enough to span many more repeat elements than can short reads, leading to reports of greatly improved assemblies of several species. After error correction we had 21x coverage of reads greater than 10 Kb in length; these reads were used in the final assembly (ML2). Use of the PacBio reads significantly improved the genome assembly compared with Illumina only (Fig. 2B), including improving the contig N50 size from 222 bp to 64 Kbp and the largest assembled contig from 144 Kbp to 627 Kbp.

To assess the quality and coverage of the ML2 assembly, *M. lignano* expressed sequence tags (ESTs) from public databases (25) and sequences derived from an arrayed bacterial artificial chromosome (BAC) library of the *M. lignano* genome were aligned to the assembly. The *M. lignano* ESTs were generated before establishment of the inbred DV1 line used for the genome assembly. Nevertheless, 92% of ESTs could be aligned to the genome with an average identity of 94.6%. Of reads derived from 1,248 BACs, 91% pooled and sequenced using Illumina mapped to the ML2 assembly, with identity over 99.5%. Additionally, a sample of 10% of contigs from the ML1 (Illumina only) assembly aligned to the ML2 (PacBio only) assembly, with 99.56% identity. Considered together, these results indicated a reliable local assembly, and our annotation results (see below) show a high-quality representation of overall genic composition.

Even with the long PacBio reads, there remained a possibility that the unusually high rate of low-complexity sequences still had a profound impact on our ability to assemble the *M. lignano* genome. To examine this possibility, we analyzed contig ends and found that 55% of them had more than 50% of their bases masked by Tandem Repeat Finder, suggesting a possible cause for the contigs to terminate. Because we had obtained 21x coverage by reads larger than 10 Kb, this indicated the presence of repeat tracts of at least this length. The low-complexity repeats showed sequence biases, with complexity sequences were present in libraries prepared from both whole worms and sorted proliferating S-phase cells (*SI Appendix, Fig. S1B*), and were enriched in the nontranscribed fraction of the genome (*SI Appendix, Fig. S1A*). Roughly 25% of the *Macrostomum* genome was comprised of simple repeats, far greater than the fraction observed in *Caenorhabditis elegans*, *Drosophila melanogaster*, *Schistosoma mansoni*, or the human genome (*SI Appendix, Fig. S1C*). This percentage seemed high enough to contribute to the poor quality of our initial assembly. To overcome this problem, we sequenced the *M. lignano* genome using the SMRT sequencing from Pacific Biosciences (PacBio; 130x genomic coverage). This technology can generate reads long enough to span many more repeat elements than can short reads, leading to reports of greatly improved assemblies of several species. After error correction we had 21x coverage of reads greater than 10 Kb in length; these reads were used in the final assembly (ML2). Use of the PacBio reads significantly improved the genome assembly compared with Illumina only (Fig. 2B), including improving the contig N50 size from 222 bp to 64 Kbp and the largest assembled contig from 144 Kbp to 627 Kbp.
the most common repeats being GA-rich (SI Appendix, Fig. S2). Additionally, the low-complexity repeats had an unusual length distribution and frequency of occurrence (SI Appendix, Fig. S3A). The most commonly repeated sequences were 20–24 mers, but repeats were found as long as 100 bp. This high prevalence of long repeats was unique to M. lignano and was not observed in the other analyzed species (SI Appendix, Fig. S3A). The 20–24 mers often occurred as tandem repeats and were evenly dispersed throughout the genome (Fig. 3 and SI Appendix, Fig. S3B). Detailed analysis of the tandem repeats revealed that they could be composed of units even longer than 100 bp. For example, a 150 bp long tandem repeat was identified that comprises 1.6% of the entire M. lignano genome (SI Appendix, Table S1).

Tandem repeats are usually present as constitutive heterochromatic loci (31–33). This can be associated with the presence of CpG methylation, which can act as a repressive epigenetic mark in some contexts (34). Cytosine methylation makes the base susceptible to deamination, resulting in C-to-T transitions (35, 36). Thus, CpG frequency can be used as an indication of the extent of methylation. Quantifying dinucleotide occurrences in M. lignano revealed that CpG, which was present at only 71% of the anticipated frequency (SI Appendix, Fig. S4), indicated that CpG methylation might be present at low rates. In accord with this hypothesis, we detected putative homologs of MBD-1, -2, and -3 (Methyl-CpG binding proteins) in both the genome and transcriptome and of DNMT1 and DNMT3A (de novo methyltransferases) in the genome of M. lignano (Dataset S1). To test for the presence of modified cytosines directly, we sequenced the M. lignano genome after bisulfite conversion. This process revealed that low levels (~2.5% of CpGs, based upon a genome-wide average) of modification are present, although we cannot strictly distinguish based upon our analysis between cytosine methylation and other modifications, such as hydroxymethylation. Notably, DNA methylation was not detected in S. mediterranea (37) and there are conflicting reports for S. mansoni (38, 39).

**Genome Annotation.** To identify and evaluate protein-coding genes within the assembly, we used a combination of CEGMA and the MAKER annotation system. Of the 248 conserved eukaryotic genes, 232 (93.55%) were complete and 246 (99.19%) were partial hits in the M. lignano genome. This finding indicated that the M. lignano gene space was well assembled, but that the assembly was fragmented in noncoding regions because of the high frequency of low-complexity and tandem repeats. As predicted using MAKER, the genome included sequences of M. lignano including ~61,000 gene models, constituting an estimated 10% of the genome (SI Appendix, Fig. S5). This was likely an overestimate of the true gene number because of overprediction, unrecognized transposable elements, pseudogenes, and gene fragmentation at contig or scaffold boundaries. Only 19,794 gene models had over 50% of their exons supported by RNA-Seq data (SI Appendix, Fig. S5).

RepeatMasker masked only 7.7% of the ML2 assembly and indicated that known retroelements and DNA transposons constitute only 0.06% and 0.11% of the genome, respectively. RepeatScout was used to more broadly detect repetitive substrings in the M. lignano genome (40) and detected 23,064 types of elements with an average length of 946 bp, the longest being 20 Kbp long. These elements make up ~51% of the genome (Fig. 3). Of the 23,000 elements detected, 1,693 were annotated (SI Appendix, Fig. S6). The low fraction of annotated transposons suggests that there may be novel classes of transposons in M. lignano.

**Transcriptome Assembly and Annotation.** To facilitate gene annotation, we generated RNA sequencing libraries from whole worms and assembled them using the Trinity assembler; this generated 149,647 putative transcripts totaling 77 million base pairs. The average assembled transcript length was 516 bp and the N50 of the transcriptome was 649 bp (SI Appendix, Fig. S7A). Transcripts were annotated using the Trinotate pipeline; 64,842 transcripts were annotated representing 43.3% of the assembled transcripts. Of the transcripts present in the transcriptome, 99.47% align to the genome at an average identity of 98.31%. The average transcript had an alignment covering 98.6% of its length. Gene Ontology analysis on the annotated transcriptome defined the most predominant classes of transcripts (SI Appendix, Fig. S7B). The annotation of transposons in the transcriptome assembly (5% of transcripts) suggests the presence of actively transposing families (particularly Mos1) (Dataset S2).

Some flatworm species have been shown to carry out trans-spliced leader addition (41, 42). We have found evidence of trans-splicing in the M. lignano transcriptome assembly in the form of 7,500 transcripts with potential spliced leader (SL) sequence at their 5′ ends (SI Appendix, Table S2). Those transcripts encode proteins from a range of protein families and had introns, suggesting that they undergo both trans- and cis-splicing. The longest version of the putative leader sequence was 45 nt long, but shorter versions were also observed. All versions were identical at the 3′ end and differed only by length of the 5′ end, and contained a potential initiator AUG, similarly to what was observed in other flatworms (41, 42) (SI Appendix, Fig. 84/8). We identified a longer transcript that potentially gives rise to the leader sequence. Is shares sequence similarity with other planarian SL RNAs—especially the splice site conservation—but it is ~two-times longer than planarian SL RNAs (SI Appendix, Fig. S8 and Table S2).

M. lignano shared 6,217 transcripts with S. mediterranea (SI Appendix, Fig. S9A). Furthermore, based on the transcriptome analysis M. lignano had a similar number of gene losses compared with humans as did S. mediterranea. C. elegans showed the highest number of gene losses relative to humans, and D. melanogaster the lowest (SI Appendix, Fig. S9B and Dataset S3). Interestingly, the M. lignano genome encodes ~2,000 genes that are present in humans but absent in both C. elegans and D. melanogaster (SI Appendix, Fig. S10/1 and Dataset S4). Those genes belong to a variety of pathways (i.e., Jak-Stat, PI3K-Akt, Egf, Igf, Vegf, Gf, Pdgf,
Putative Pluripotency Genes and Pathways. A number of transcription factors have been shown to play pivotal roles in stem cell maintenance and determination of pluripotency in mammals (43). The best-characterized set of mammalian pluripotency factors includes Oct4/Pou5f1 (44), Nanog (45), Klf4 (46), c-Myc (47), and Sox2 (48). These have been successfully used for the dedifferentiation of adult somatic cells into induced pluripotent stem cells (49–51). Of the five key mammalian pluripotency factors, only Sox2 was identified with high confidence in the M. lignano genome (SI Appendix, Fig. S11 and Dataset S5). Interestingly, Myc orthologs seem to be lost entirely in Platyhelminthes, making it the only animal group besides Nematodes (52) that does not encode this conserved protein (SI Appendix, Fig. S12). Hydra magnipapillata, another metazoan species with striking regenerative capacity, contains only homologs of Myc and Sox2 in its genome and is missing the remainder of the mammalian pluripotency factors (53). In S. mediterranea, only Sox2 and other Sox family transcription factors were present and these are expressed specifically in the neoblasts (54). Interestingly, even though Oct4/Pou5f1, Nanog, Klf4, and c-Myc could not be identified in the M. lignano genome/transcriptome, the main mammalian stem cell pluripotency maintenance pathways (i.e., Jak-Stat, Wnt, TGFβ, MAPK, and P3K-Akt) seem to be conserved (SI Appendix, Fig. S11 and Dataset S5) (55).

Homeobox Genes. The homeobox superclass of genes, in particular the Hox family, is responsible for patterning of the anterior–posterior axis in bilateral animals and is critical for organ regeneration in planarians (56, 57). We found that M. lignano has 49 homeobox-containing genes, represented across 11 classes of homeoboxes (SI Appendix, Figs. S13 and S14). We found interesting retention of homeobox families not seen in other Platyhelminthes sequenced so far (58) (SI Appendix, Figs. S13 and S14). The most prominent examples of those retained homeobox-gene families that could play a role in regeneration were Cdx (59), Dlx (60), and Prx (61). We also observed that some families have undergone independent lineage duplications, leaving multiple copies of Hox1, NK2.2, NK2.1, Cdx, Irf, Meis, and Pcnx (SI Appendix, Figs. S13 and S14). The genes of the homeobox superclass are often organized in clusters where the order in the cluster reflects positional or temporal expression patterns in the animal (62, 63). The clustering of homeobox genes in a genome is often of functional significance because it reflects coregulation as well as remnants of ancestral states (64). For the homeobox complement of M. lignano, we observed various instances of clustering, most likely because of independent lineage duplications, except for the case of Mnx-Barh (SI Appendix, Table S3). The most prominent examples were the TALE-class, in which a cluster of four Irogouis genes are found within the same scaffold and a scaffold containing three Meis paralogs.

A Transcriptional Profile of M. lignano Regeneration. To examine gene signatures associated with regeneration, we cut worms between the pharynx and the testes and let the head fragment regenerate for 3, 6, 12, 24, 48, and 72 h (SI Appendix, Fig. S15), and we searched for gene-expression changes across the time course (Fig. 4). We first focused on early response genes (i.e., those that are up-regulated within 3–12 h after amputation) (Fig. 4, SI Appendix, Fig. S15, and Dataset S6). Among those genes there were a number of growth factors (EGF-like growth factors and Von Willebrand growth factors). Those types of growth factors are known to participate in cell growth/division in response to stimuli (65). Interestingly, homologs of genes from the Tgfβ/Bmp pathway, one of the regulators of mammalian pluripotency, were also present among the early response genes. Additionally there were multiple up-regulated transcripts involved in cell signaling (kinase, ATPase, and GTPase domains). Finally, there were a number of up-regulated factors involved in cellular organization: cell adhesion, response to wounding, and cytoskeletal organization. This group is likely essential for wound closure and blastema formation (66) (SI Appendix, Fig. S15 and Dataset S6). We next analyzed transcripts that change expression levels at 24 or 48 h postamputation, because this time point exhibits the largest expansion of S-phase cells (putative dividing neoblasts) (19, 66). At this time point, there was an enrichment of transcription factors with zinc-finger domains, Klf transcription factors, and a TNF-like protein, a systemic signaling cytokine. Among the factors that are down-regulated 48-h postamputation, we identified a potential pluripotency determinant, a Smad4-like transcript, supporting the previous observations that the blastema at this stage enters a differentiation phase (66) (SI Appendix, Fig. S16 and Dataset S6). In summary we identified six different synexpression classes (SI Appendix, Fig. S16 and Dataset S6) of genes specifically up- or down-regulated at different time points postamputation. Even though the majority of transcripts measured were not yet annotated, these datasets can provide a valuable resource for future regeneration studies.

Discussion
To serve as a resource for future studies of flatworm biology, we have sequenced, assembled, and annotated the genome and transcriptome of the free-living flatworm M. lignano. The genome of this animal is highly enriched in dispersed, low-complexity repeats, making de novo assembly exceptionally difficult. It is currently unclear why the M. lignano genome is so rich in low-complexity tandem repeats. There is evidence pointing to minisatellites as units that cause mutability and promote evolution because of their recombinogenic properties (67). If minisatellites are present in the M. lignano genome in the tens of thousands, as the data suggests, they could contribute to meiotic mutability and potentially cause genomic instability. The impact of this repeat burden is thus a clear area for further investigation and would benefit from comparisons with other closely related species.

The M. lignano genome showed an indication of low levels of DNA methylation (~2.5% CpG). Many commonly used nonmammalian model organisms (including yeast, C. elegans, and D. melanogaster) completely lack or have very low levels of genomic DNA methylation (68). As the M. lignano genome is indeed methylated, albeit to a low extent, this organism will provide an important invertebrate model for studying the evolution of methylation in metazoans.

Our initial analysis of the M. lignano genome and transcriptome has begun to reveal a range of interesting properties. The homeobox complement of M. lignano has retained distinct homeobox families in contrast to other Platyhelminthes analyzed (58). The M. lignano transcriptome shows evidence of trans-splicing. The evolutionary
history and significance of trans-splicing remain open questions. *M. lignano* and other flatworms lack Myc orthologs. This is an interesting observation because Myc is very conserved in Bilaterians and even beyond (cnidarians, poriferans), although it is also absent from Nematodes (52, 69). Because the Myc transcriptional network predates the origin of animals (69), Nematodes and Platyhelminthes must have independently lost the Myc genes, although other parts of the Myc transcriptional network (as suggested by the retention of Max) may be intact and remain to be investigated across Platyhelminthes. *M. lignano* also provides an interesting model for the study of germ cell biology, because neoblasts are able to differentiate into germ cells. As an example, we are already beginning to probe the roles of the piRNA pathway in transposon silencing and neoblast maintenance in *M. lignano* (70).

*M. lignano* has a number of properties that make it advantageous as a model for studying stem and germ cell biology, differentiation, regeneration, and perhaps also aspects of neuroscience. Moreover, viewed in comparison with those of other Platyhelminthes, the resource we provide might shed further light on the evolution of the molecular toolkit of regeneration and also on the evolution and conservation of genes and pathways in protostomes.

**Materials and Methods**

Detailed materials and methods are available in SI Appendix.

**Animal Culture and Regeneration.** *M. lignano* was kept in Petri dishes with nutrient-enriched f/2 medium (71) and fed ad libitum with diatom algae (Nitzschia curvilinea). For regeneration, worms were cut at the postpharyngeal level to completely remove the gonads.

**Sequencing Library Preparation, DNA and RNA Isolation.** DNA-Seq libraries were prepared using the PacBio Ultralow Library Systems (Nugen). For PaclBio sequencing the libraries were prepared using the PacBio library preparation kit, RS II, according to the manufacturer’s instructions. The libraries were sequenced using either the p4c2 or p5c3 chemistry and standard run parameters. For transcriptome assembly, three Script Seg V2 (Epibio) libraries were constructed according to manufacturer’s specifications. RNA-Seq libraries for the regeneration studies were generated using the Encore Complete RNA-Seq DR Multiplex System according to manufacturer’s instructions. All Illumina samples were sequenced using Illumina GAIL or HiSeq. 2000 (PE100) platforms.

**Transcriptome Assembly and Annotation.** The transcriptome assembly was done using the Trinity package (Broad Institute). The transcriptome annotation was performed using Trinotate, the Trinity annotation pipeline (72).

**Genome Assembly and Annotation.** The Illumina Assembly (ML1) was built using SGA using 115x coverage of 101-bp paired-end Illumina HiSeq data. Pacbio data were self-corrected using HGAP. After correction, reads were assembled using the Celera Assembler v8.2beta generating the ML2 assembly. A sample of 81,665 contigs from the Illumina assembly (~10%) were aligned to all of the contigs in the Pacbio assembly using Mummer v3.23. Genome annotation was performed using MAKER v2.31.8 (December 2014).

**Transposon Analysis.** RepeatScout v1.0.5 was run on both the Illumina and Pacbio assemblies (40). Only repeats that occur at least 10 times in the genome were kept for further analysis. Repeats were annotated using a custom nonredundant library from National Center for Biotechnology Information (NCBI) entries (keywords: retrotransposon, transposase, “reverse transcriptase,” gympos, copia) obtained from O. Simakov et al. (73).

**K-mer Analysis and Peak Modeling.** K-mers were counted in the Illumina data using Jellyfish 1.1.10 with the -C parameter. Peak modeling was performed by fitting a mixture model composed of four Poisson distributions and calculating their composite in R (74).

**Differential Expression.** Reads were aligned to the transcriptome using RSEM (74). Differentially expressed genes (false-discovery rate ≤ 0.001, with a minimum fourfold change) were identified using DESeq (75).

**Analysis of the Transcript Conservation.** Control script (reciprocalblast_allsteps.py) for running reciprocal BLASTp search was obtained from Waren et al. (76).

**Sequence Complexity Analysis.** Sequence complexity was calculated on a per basis using a previously described algorithm (77).

**D tandem Repeat Finder Masking for Low Complexity.** Tandem Repeat Finder was used with the -n option to mask low complexity repeats in RNA-Seq data.

**Estimating CpG Content.** CpG histograms were built using a previously described method (73).

**Bisulfite Genomic DNA Sequencing and Analysis.** The DNA was bisulfite converted using Zymo EZ methylation gold kit following manufacturer’s instructions. Reads were aligned to the ML2 assembly and analyzed as previously described (79).

**Data Access.** The genome and transcriptome sequencing data are available in the NCBI Sequence Read Archive under accession no. SRP059553.

**ACKNOWLEDGMENTS.** We thank Dr. Eugene Berezhkov and Turan Demircan for helping us to establish *Macrostomum lignano* culture at Cold Spring Harbor Laboratory, and discussions in the early phases of the project; Willi Salvenmoser for help with electron microscopy; Dr. Dario Bressan for help with experimental design; and Emily Lee and the Genome Center of Cold Spring Harbor Laboratory for preparing and running the Pacific Biosciences and Illumina sequencing libraries. This work is supported by National Institutes of Health Grants R37 GM062534 (to G.J.H.) and R01-HG006677 (to M.S.); National Science Foundation Grant CNS-0844449 (to M.S.); National Science Foundation Grant 0819079 (to P.A.M.); and a Swiss National Science Foundation Grant 31003A-143732 (to L.S.). This work was performed with assistance from Cold Spring Harbor Laboratory Shared Resources, which are funded, in part, by Cancer Center Support Grant SP30CA045508.