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

Kaja Wasik[a,1], James Gurtowski[a,1], Xin Zhou[a,b], Olivia Mendivil Ramos[a], M. Joaquina Delás[a,c], Giorgia Battistoni[a,c], Osama El Demerdash[a], Ilaria Falciatori[a,c], Dita B. Vizoso[d], Andrew D. Smith[e], Peter Ladurner[f], Lukas Schärer[d], W. Richard McCombie[a], Gregory J. Hannon[a,c,2], and Michael Schatz[a,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; [e]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089; and [f]Department of Evolutionary Biology, Institute of Zoology and Center for Molecular Biosciences Innsbruck, University of Innsbruck, A-6020 Innsbruck, Austria

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

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 sequences, 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 130× 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.

flatworm | regeneration | *Macrostomum* | neoblast | genome

Flatworms belong to the superphylum Lophotrochozoa, a vast assembly of protostome invertebrates (1, 2) (Fig. 1*A*). 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. 1*B*), a free-living, regenerating flatworm isolated from the coast of the Mediterranean Sea. *M. lignano* is an obligatorily cross-fertilizing simultaneous hermaphrodite (9) that belongs to Macrostomorpha, whereas the other often-studied free-living flatworms and human parasitic flatworms all belong to clades that are potentially more derived (less ancestral) in comparison with Macrostomorpha (2) (Fig. 1*C*).

Many flatworms can regenerate nearly their entire body or amputated organs. This regenerative capacity is thought to be attributable to the presence of somatic stem cells, termed neoblasts (10, 11). In *Schmidtea mediterranea* (planarian flatworm), even a single transplanted neoblast has the ability to rescue, regenerate, and change the genotype of a fatally irradiated worm (12). *M. lignano* can regenerate every tissue, with the exception of the head region containing the brain (13, 14).

Neoblasts in *M. lignano* (Fig. 1 *D* and *E*), in contrast to most vertebrate somatic stem cells, are plentiful, making up about ∼6.5% of all cells (15), and have a very high proliferation rate (16, 17). Of *M. lignano* neoblasts, 89% enter S-phase every 24 h (18). This high mitotic activity results in a continuous stream of progenitors, replacing tissues that are likely devoid of long-lasting, differentiated cell types (18). This makes *M. lignano* an ideal model to study tissue homeostasis because most other species have far fewer somatic stem cells, and these are usually more difficult to harvest.

Given its promise as a model for studying mechanisms governing pluripotency, a number of groups have worked to establish *M. lignano* as a model to study stem cell biology and regeneration (16, 19, 20), sexual selection and reproductive biology (21, 22), bioadhesion (23), and neurobiology (24). Efforts of the *M. lignano* community have resulted in the development of a number of tools that can be used to study *M. lignano* biology (15, 21, 25–27).

To facilitate use of *M. lignano* as a model organism more generally, we have produced genome and transcriptome assemblies. We found the *M. lignano* genome to be replete with dispersed tandem repeats of low-complexity sequences. To compensate for this complex genomic architecture, we generated over 100-times coverage of a PacBio long-read sequencing that gave rise to an assembly that is, on average, over 100-times more contiguous than the Illumina-only assembly.

Protein coding genes appear well assembled and ∼20,000 gene models are supported by our transcriptome libraries. *M. lignano*'s genome and transcriptome lack nearly all of the key mammalian pluripotency factors (i.e., *Oct4*/*Pou5f*, *Klf4*, and *c-Myc*). The availability of annotated genome and transcriptome assemblies enables comparison of gene-expression

Significance

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.

Author contributions: K.W., J.G., D.B.V., P.L., L.S., G.J.H., and M.S. designed research; K.W., J.G., X.Z., M.J.D., G.B., I.F., D.B.V., and P.L. performed research; K.W., J.G., O.M.R., O.E.D., A.D.S., P.L., L.S., W.R.M., G.J.H., and M.S. analyzed data; and K.W., J.G., O.M.R., G.J.H., and M.S. wrote the paper.

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

The authors declare no conflict of interest.

Freely 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. SRP059553).

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

[2]To 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.

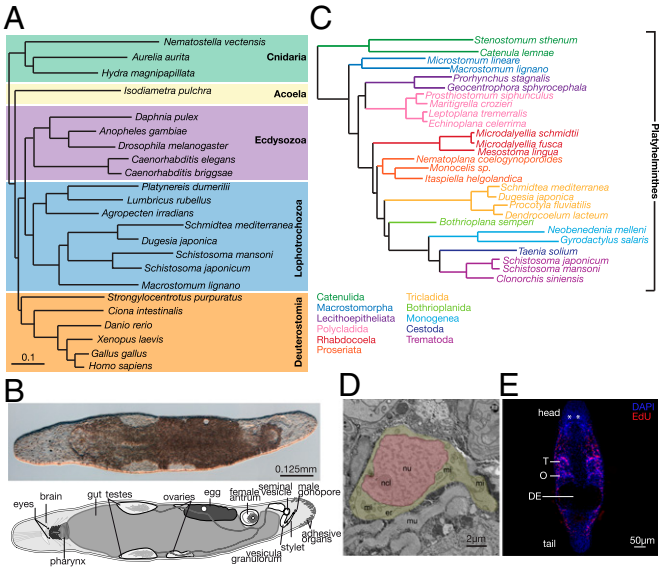

Fig. 1. (*A*) Phylogenetic analysis of 23 animal species using partial sequences of 43 genes. Figure modified from Egger et al. (80). (*B*) Interference contrast image and a diagrammatic representation of an adult *M. lignano*. (*C*) Phylogenomic analysis of 27 flatworm species (21 free-living and 6 neodermatan) using >100,000 aligned amino acids. Figure modified from Egger et al. (2). (*D*) Electron micrograph of a *M. lignano* neoblast. Note the small rim of cytoplasm (yellow) and the lack of cytoplasmic differentiation. Er, endoplasmic reticulum; mi, mitochondria; mu, muscle; ncl, nucleolus; nu, nucleus (red). (*E*) Immunofluorescence labeling of dividing neoblasts with EdU (red) in an adult worm. All cell nuclei are stained with DAPI (blue). DE, developing eggs; O – ovaries; T, testes. Asterisks denote eyes.

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 amputation. It is our hope that the assembled *M. lignano* genome and transcriptome will serve as a valuable reference for studies of evolutionary relationships, will shed light on the evolution and origins of Bilateria, and will comprise an important resource for regenerative biology.

Results

Genome Assembly. Sequencing efforts were focused on the DV1 line, which was generated through 35 generations of sibling crosses (28). Using a whole-genome shotgun approach, a total of a 170× genomic coverage of Illumina paired-end 101-bp reads were generated. Based on its K-mer (23-mer) distribution, the *M. lignano* genome size was estimated to be ∼700 Mbp, which is roughly 1.5× the estimated size of the *S. mediterranea* genome, the closest relative of *M. lignano* with genomic information available (29). The first assembly draft, the ML1 assembly, had a very unusual four-modal K-mer distribution (Fig. 2*A*), suggesting a high frequency of genomic duplications (peaks 3 and 4). Indeed *M. lignano* has four (2n = 8) sets of chromosomes in comparison with three (2n = 6) sets of chromosomes found in the majority of other *Macrostomum* species studied to date (30), suggesting a potential chromosomal duplication. The proportion of duplicated sequences was higher, however, than what one would expect based on the duplication of one small chromosomal pair. This finding suggested another layer of multiplication, potentially an ancestral whole-genome duplication or more recent large segmental duplications.

The ML1 assembly was highly fragmented, with an average contig size of only 532 bp, an N50 of 222 bp, and a maximum contig size of 144 Kbp (Fig. 2*B*). A potential explanation for such low values may be the observed prevalence of low-complexity sequences in the *M. lignano* genome (*SI Appendix*, Fig. S1*A*), which is higher than that seen in many other sequenced genomes. The low-complexity sequences were present in libraries prepared from both whole worms and sorted proliferating S-phase cells (*SI Appendix*, Fig. S1*B*), and were enriched in the nontranscribed fraction of the genome (*SI Appendix*, Fig. S1*A*). 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. S1*C*). 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; 130× 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 21× 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. 2*B*), 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 datasets (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 assemblies. 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 contings 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 21× coverage by reads larger than 10 Kbp, this indicated the presence of repeat tracts of at least this length. The low-complexity repeats showed sequence biases, with

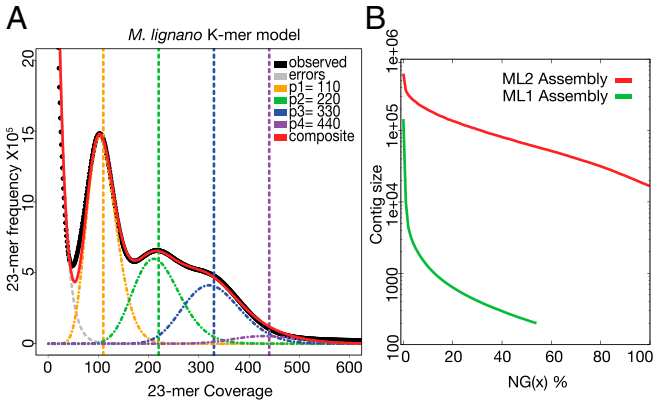

Fig. 2. (*A*) Representation of 23-mer frequency and coverage in the Illumina sequencing data generated from DNA extracted from a population of adult worms. Peak modeling was performed by fitting a mixture model of four Poisson distributions and calculating their composite distribution in R. (*B*) Contig length distribution ($\log_2$ scale) over the *M. lignano* genome in the ML1 (green) and ML2 (red) assemblies. Note that the ML1 assembly covers only about 55% of the genome.

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. S3*A*). 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. S3*A*). The 20–24 mers often occurred as tandem repeats and were evenly dispersed throughout the genome (Fig. 3 and *SI Appendix*, Fig. S3*B*). 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), indicating 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 domain 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).

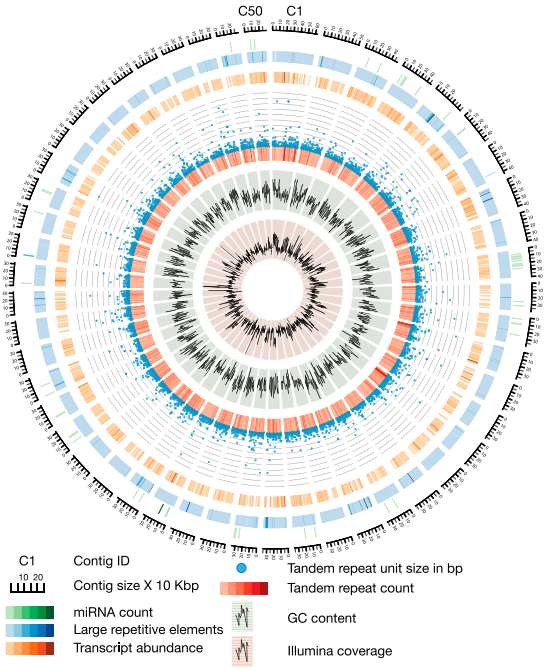

Fig. 3. Overview of the 50 largest contigs in the *M. lignano* genome, making up about 2.6% of the total assembly. Different tracks denote (moving inwards): contig size × 10 Kbp; miRNA count (1–54 mapped miRNAs); large repetitive elements (RepeatScout) (1–4,476 identified repeats); transcript count (1–43 mapped transcripts); Tandem repeat unit size in base pairs (1–500); Tandem repeat count (1–28); GC content (0–1); and Illumina coverage (4–160×). The color gradients correspond to the range of values for each track (lower values are lighter, higher values are darker).

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 assembled genome of *M. lignano* included ~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. S7*A*). 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. S7*B*). 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* transciptome assembly in the form of 7,500 transcripts with potential spliced leader (SL) sequence at their 5′ ends (*SI Appendix*, Table S2). Those transcripts encoded 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. S8*A*). 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. S9*A*). 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. S9*B* 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*A* and Dataset S4). Those genes belong to a variety of pathways (i.e., Jak-Stat, Pi3k-Akt, Egf, Igf, Vegf, Fgf, Pdgf,

Tgfβ/Bmp, Mapk, p53, Hedgehog, Notch) (*SI Appendix*, Fig. S10*B* and Dataset S4).

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 most well-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 metoazoan 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 PI3K-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), *Dbx* (60), and *Prrx* (61). We also observed that some families have undergone independent lineage duplications, leaving multiple copies of *Hox1*, *NK2.2*, *NK2.1*, *Cdx*, *Irx*, *Meis*, and *Pknox* (*SI Appendix*, Figs. S13 and S14). The genes of the homeobox superclass are often organized in clusters wherein 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 *Iroquois* 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 Tgf-β/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,

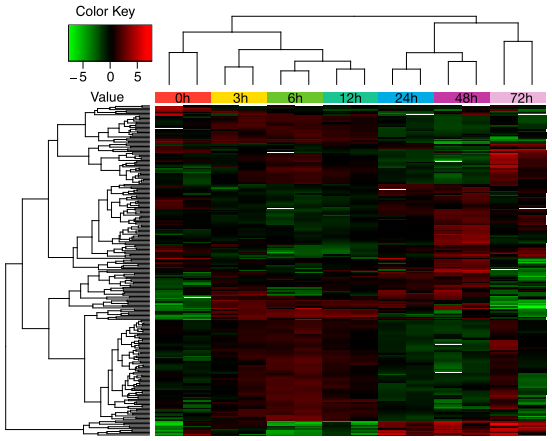

Fig. 4. Heat map of differentially expressed genes at different regeneration timepoints. Each replicate is plotted separately. Down-regulated and up-regulated transcripts are labeled in green and red, respectively. Scale covers $\log_2$ values. The samples are grouped with complete-linkage clustering using Euclidean distance.

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

GENETICS

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 curvilineata*). 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 Ovation Ultralow Library Systems (Nugen). For PacBio 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 GAII 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 115× 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," gypsy, 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.

Differential Expression. Reads were aligned to the transcriptome using RSEM (74). Differentially expressed genes (false-discovery rate $\leq$ 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 Warren et al. (76).

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

Tandem Repeat Finder Masking for Low Complexity. Tandem Repeat Finder (78) was run on each sample with the following parameters: 2 7 7 80 10 50 500 -f -d -m -ngs -h.

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 Berezikov 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 DBI-1350041 (to M.S.); 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 5P30CA045508.

1. Giribet G (2008) Assembling the lophotrochozoan (=spiralian) tree of life. *Philos Trans R Soc Lond B Biol Sci* 363(1496):1513–1522.
2. Egger B, et al. (2015) A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Curr Biol* 25(10):1347–1353.
3. Paps J, Baguñà J, Riutort M (2009) Lophotrochozoa internal phylogeny: New insights from an up-to-date analysis of nuclear ribosomal genes. *Proc Biol Sci* 276(1660):1245–1254.
4. Struck TH, et al. (2014) Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of spiralia. *Mol Biol Evol* 31(7):1833–1849.
5. Aboobaker AA (2011) Planarian stem cells: A simple paradigm for regeneration. *Trends Cell Biol* 21(5):304–311.
6. Reddien PW (2013) Specialized progenitors and regeneration. *Development* 140(5): 951–957.
7. González-Estévez C, Felix DA, Rodríguez-Esteban G, Aboobaker AA (2012) Decreased neoblast progeny and increased cell death during starvation-induced planarian degrowth. *Int J Dev Biol* 56(1-3):83–91.
8. Morgan TH (1901) Regeneration and liability to injury. *Science* 14(346):235–248.
9. Schärer L, Ladurner P (2003) Phenotypically plastic adjustment of sex allocation in a simultaneous hermaphrodite. *Proc Biol Sci* 270(1518):935–941.
10. Brøndsted H (1955) Planarian regeneration. *Biol Rev Camb Philos Soc* 30(1):65–125.
11. Baguñà J (2012) The planarian neoblast: The rambling history of its origin and some current black boxes. *Int J Dev Biol* 56(1-3):19–37.
12. Wagner DE, Wang IE, Reddien PW (2011) Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science* 332(6031):811–816.
13. Egger B, Ladurner P, Nimeth K, Gschwentner R, 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(10):565–577.
14. Nimeth KT, et al. (2007) Regeneration in *Macrostomum lignano* (Platyhelminthes): Cellular dynamics in the neoblast stem cell system. *Cell Tissue Res* 327(3):637–646.
15. Bode A, et al. (2006) Immunogold-labeled S-phase neoblasts, total neoblast number, their distribution, and evidence for arrested neoblasts in *Macrostomum lignano* (Platyhelminthes, Rhabditophora). *Cell Tissue Res* 325(3):577–587.
16. Ladurner P, Rieger R, Baguñà J (2000) Spatial distribution and differentiation potential of stem cells in hatchlings and adults in the marine platyhelminth *Macrostomum* sp.: A bromodeoxyuridine analysis. *Dev Biol* 226(2):231–241.
17. Nimeth KT, et al. (2004) Stem cell dynamics during growth, feeding, and starvation in the basal flatworm *Macrostomum* sp. (Platyhelminthes). *Dev Dyn* 230(1):91–99.
18. Ladurner P, Egger B, De Mulder K, Pfister D, Kuales G (2008) The stem cell system of the basal flatworm *Macrostomum lignano*. *Stem Cells*, ed Bosch TCG (Springer, Dordrecht, The Netherlands), pp 75–94.
19. Egger B, et al. (2009) The caudal regeneration blastema is an accumulation of rapidly proliferating stem cells in the flatworm *Macrostomum lignano*. *BMC Dev Biol* 9:41.
20. De Mulder K, et al. (2009) Stem cells are differentially regulated during development, regeneration and homeostasis in flatworms. *Dev Biol* 334(1):198–212.
21. Marie-Orleach L, Janicke T, Vizoso DB, Eichmann M, Schärer L (2014) Fluorescent sperm in a transparent worm: Validation of a GFP marker to study sexual selection. *BMC Evol Biol* 14:148.

22. Arbore R, et al. (2015) Positional RNA-Seq identifies candidate genes for phenotypic engineering of sexual traits in *Macrostomum lignano*. *Front Zool* 12:14.
23. Lengerer B, et al. (2014) Biological adhesion of the flatworm *Macrostomum lignano* relies on a duo-gland system and is mediated by a cell type-specific intermediate filament protein. *Front Zool* 11(1):12.
24. Morris J, Cardona A, De Miguel-Bonet MdelM, Hartenstein V (2007) Neurobiology of the basal platyhelminth *Macrostomum lignano*: Map and digital 3D model of the juvenile brain neuropile. *Dev Genes Evol* 217(8):569–584.
25. Morris J, et al. (2006) The *Macrostomum lignano* EST database as a molecular resource for studying platyhelminth development and phylogeny. *Dev Genes Evol* 216(11): 695–707.
26. Ladurner P, et al. (2005) Production and characterisation of cell- and tissue-specific monoclonal antibodies for the flatworm *Macrostomum* sp. *Histochem Cell Biol* 123(1): 89–104.
27. Pfister D, et al. (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.
28. Janicke T, et al. (2013) Sex allocation adjustment to mating group size in a simultaneous hermaphrodite. *Evolution* 67(11):3233–3242.
29. Robb SM, Ross E, Sánchez Alvarado A (2008) SmedGD: The *Schmidtea mediterranea* genome database. *Nucleic Acids Res* 36(Database issue):D599–D606.
30. Egger B, Ishida S (2005) Chromosome fission or duplication in *Macrostomum lignano* (Macrostomorpha, Plathelminthes)—Remarks on chromosome numbers in 'archoophoran turbellarians'. *J Zoological Syst Evol Res* 43(2):127–132.
31. McClintock B (1951) Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol* 16:13–47.
32. Melters DP, et al. (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* 14(1):R10.
33. Pardue ML, Gall JG (1970) Chromosomal localization of mouse satellite DNA. *Science* 168(3937):1356–1358.
34. Holoch D, Moazed D (2015) RNA-mediated epigenetic regulation of gene expression. *Nat Rev Genet* 16(2):71–84.
35. Frederico LA, Kunkel TA, Shaw BR (1990) A sensitive genetic assay for the detection of cytosine deamination: Determination of rate constants and the activation energy. *Biochemistry* 29(10):2532–2537.
36. Shen JC, Rideout WM, 3rd, Jones PA (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res* 22(6):972–976.
37. Jaber-Hijazi F, et al. (2013) Planarian MBD2/3 is required for adult stem cell pluripotency independently of DNA methylation. *Dev Biol* 384(1):141–153.
38. Geyer KK, et al. (2011) Cytosine methylation regulates oviposition in the pathogenic blood fluke *Schistosoma mansoni*. *Nat Commun* 2:424.
39. Raddatz G, et al. (2013) Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc Natl Acad Sci USA* 110(21):8627–8631.
40. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21(Suppl 1):i351–i358.
41. Zayas RM, Bold TD, Newmark PA (2005) Spliced-leader trans-splicing in freshwater planarians. *Mol Biol Evol* 22(10):2048–2054.
42. Rajkovic A, Davis RE, Simonsen JN, Rottman FM (1990) A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. *Proc Natl Acad Sci USA* 87(22):8879–8883.
43. Nichols J, Smith A (2012) Pluripotency in the embryo and in culture. *Cold Spring Harb Perspect Biol* 4(8):a008128.
44. Nichols J, et al. (1998) Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* 95(3):379–391.
45. Chambers I, et al. (2003) Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* 113(5):643–655.
46. Jiang J, et al. (2008) A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* 10(3):353–360.
47. Cartwright P, et al. (2005) LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. *Development* 132(5):885–896.
48. Masui S, et al. (2007) Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol* 9(6):625–635.
49. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126(4):663–676.
50. Okita K, Ichisaka T, Yamanaka S (2007) Generation of germline-competent induced pluripotent stem cells. *Nature* 448(7151):313–317.
51. Wernig M, et al. (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* 448(7151):318–324.
52. McFerrin LG, Atchley WR (2011) Evolution of the Max and Mlx networks in animals. *Genome Biol Evol* 3:915–937.
53. Chapman JA, et al. (2010) The dynamic genome of Hydra. *Nature* 464(7288):592–596.
54. Onal P, et al. (2012) Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells. *EMBO J* 31(12):2755–2769.
55. Nakaya A, et al. (2013) KEGG OC: A large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res* 41(Database issue):D353–D357.
56. Roberts-Galbraith RH, Newmark PA (2015) On the organ trail: Insights into organ regeneration in the planarian. *Curr Opin Genet Dev* 32:37–46.
57. Hudry B, et al. (2014) Molecular insights into the origin of the Hox-TALE patterning system. *eLife* 3:e01939.
58. Tsai IJ, et al.; *Taenia solium* Genome Consortium (2013) The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* 496(7443):57–63.
59. Sanchez-Ferras O, et al. (2012) Caudal-related homeobox (Cdx) protein-dependent integration of canonical Wnt signaling on paired-box 3 (Pax3) neural crest enhancer. *J Biol Chem* 287(20):16623–16635.
60. Lacin H, Zhu Y, Wilson BA, Skeath JB (2009) dbx mediates neuronal specification and differentiation through cross-repressive, lineage-specific interactions with eve and hb9. *Development* 136(19):3257–3266.
61. Satoh A, makanae A, Hirata A, Satou Y (2011) Blastema induction in aneurogenic state and Prrx-1 regulation by MMPs and FGFs in *Ambystoma mexicanum* limb regeneration. *Dev Biol* 355(2):263–274.
62. Akam M (1989) Hox and HOM: Homologous gene clusters in insects and vertebrates. *Cell* 57(3):347–349.
63. Graham A, Papalopulu N, Krumlauf R (1989) The murine and *Drosophila* homeobox gene complexes have common features of organization and expression. *Cell* 57(3): 367–378.
64. Hui JH, et al. (2012) Extensive chordate and annelid macrosynteny reveals ancestral homeobox gene organization. *Mol Biol Evol* 29(1):157–165.
65. Yusuf D, et al. (2012) The transcription factor encyclopedia. *Genome Biol* 13(3):R24.
66. De Mulder K, et al. (2009) Characterization of the stem cell system of the acoel *Isodiametra pulchra*. *BMC Dev Biol* 9:69.
67. Vergnaud G, Denoeud F (2000) Minisatellites: Mutability and genome architecture. *Genome Res* 10(7):899–907.
68. Yi S (2012) Birds do it, bees do it, worms and ciliates do it too: DNA methylation from unexpected corners of the tree of life. *Genome Biol* 13(10):174.
69. Young SL, et al. (2011) Premetazoan ancestry of the Myc-Max network. *Mol Biol Evol* 28(10):2961–2971.
70. Zhou X, et al. (August 31, 2015) Dual functions of Macpiwi1 in transposon silencing and stem cell maintenance in the flatworm *Macrostomum lignano*. *RNA*, 10.1261/rna.052456.115.
71. Andersen RA, Berges JA, Harrison PJ, Watanabe MM (2005) *Recipes for Freshwater and Seawater Media* (Elsevier, Amsterdam).
72. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644–652.
73. Simakov O, et al. (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature* 493(7433):526–531.
74. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.
75. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
76. Warren IA, et al. (2014) Extensive local gene duplication and functional divergence among paralogs in Atlantic salmon. *Genome Biol Evol* 6(7):1790–1805.
77. Gabrielian A, Bolshoy A (1999) Sequence complexity and DNA curvature. *Comput Chem* 23(3-4):263–274.
78. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573–580.
79. Dos Santos CO, Dolzhenko E, Hodges E, Smith AD, Hannon GJ (2015) An epigenetic memory of pregnancy in the mouse mammary gland. *Cell Reports* 11(7):1102–1109.
80. Egger B, et al. (2009) To be or not to be a flatworm: The acoel controversy. *PLoS One* 4(5):e5502.