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

Wasik K.A.*[1], Gurtowski J.*[1], Zhou X.[1,2], Ramos O.M[1], Delas M.J.[1,3], Battistoni G.[1,3], El Demerdash O.[1], Falciatori I.[1,3], Vizoso D.B.[4], Smith A.D.[5], Ladurner P.[6], Scharer L.[4], McCombie W.R.[1], Hannon G.J.[1,3] and Schatz M.[1]

Cold Spring Harbor Laboratory

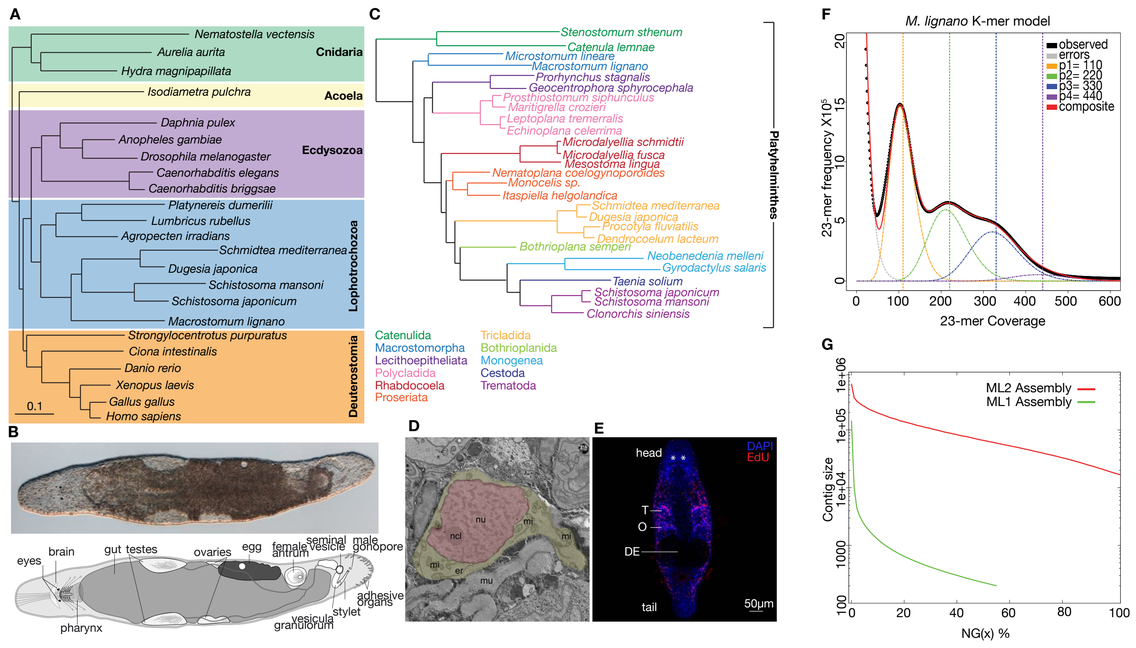
1 Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, New York 11724, USA; 2 Molecular and Cellular Biology Graduate Program, Stony Brook University, NY 11794; 3 Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge CB2 0RE, United Kingdom; 4 Department of Evolutionary Biology, Zoological Institute, University of Basel, 4051 Basel, Switzerland; 5 Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA; 6 Department of Evolutionary Biology, Institute of Zoology and Center for Molecular Biosciences Innsbruck, University of Innsbruck, A-6020 Innsbruck, Austria

The free-living flatworm, *Macrostomum lignano*, much like its better known planarian relative, *Schmidtea mediterranea*, has a nearly unlimited regenerative capacity. Following injury, this species has the ability to regenerate almost an entirely new organism. This is attributable to the presence of an abundant somatic stem cell population, the neoblasts. These cells are also essential for the ongoing maintenance of most tissues, as their loss leads to the rapid and irreversible degeneration of the animal. This set of unique properties makes flatworms an attractive species for studying the evolution of pathways involved in self-renewal, fate specification, and regeneration. The use of *Macrostomum lignano,* or other flatworms, 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 *Macrostomum lignano* and an accompanying characterization of its transcriptome. The genome structure of *Macrostomum 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=414bp). We therefore obtained 130X coverage by long sequencing reads from the PacBio platform and combined this with more than 250X Illumina coverage to create a mixed assembly with a significantly improved N50 of 64 kb.

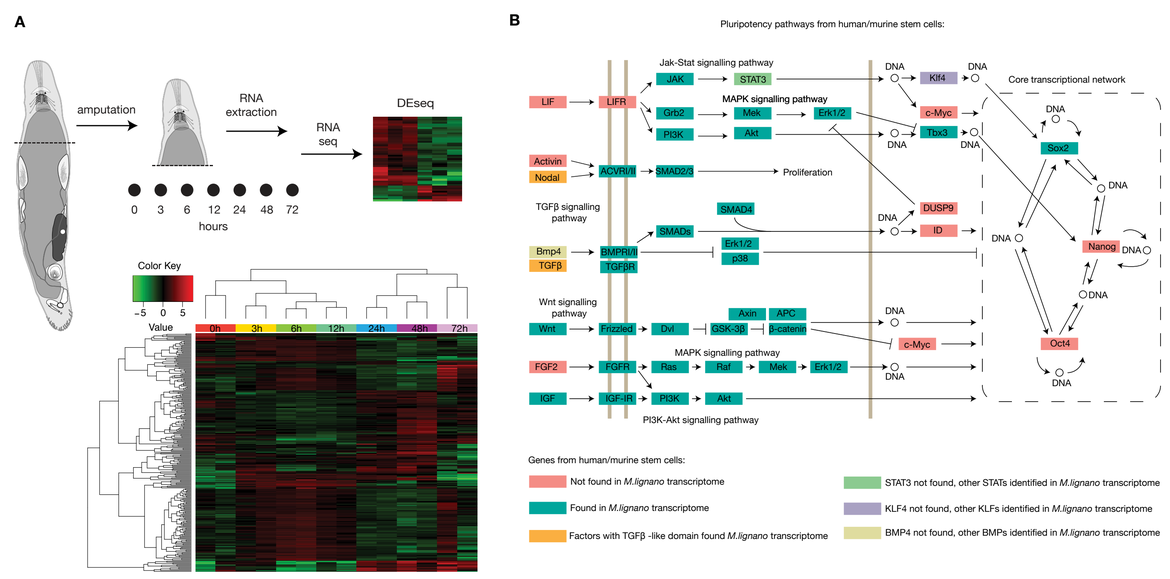
We complemented the reference genome with an assembled and annotated transcriptome, and used both of these datasets in combination to probe gene expression patterns during regeneration, examining pathways important to stem cell function. Aditionally we found evidence of low levels of CpG methylation in *Macrostomum lignano's* genome and evidence of trans-splicing in the worm's transcriptome. Interestingly we found that flatworms lack Myc - a very conserved pluripotency factor in Bilaterians and beyond (cnidarians, poriferans). As a whole, our data will provide a crucial resource for the community for the study not only of invertebrate evolution but also of regeneration and somatic pluripotency.

The Worm Genome and Transcriptome Assembly



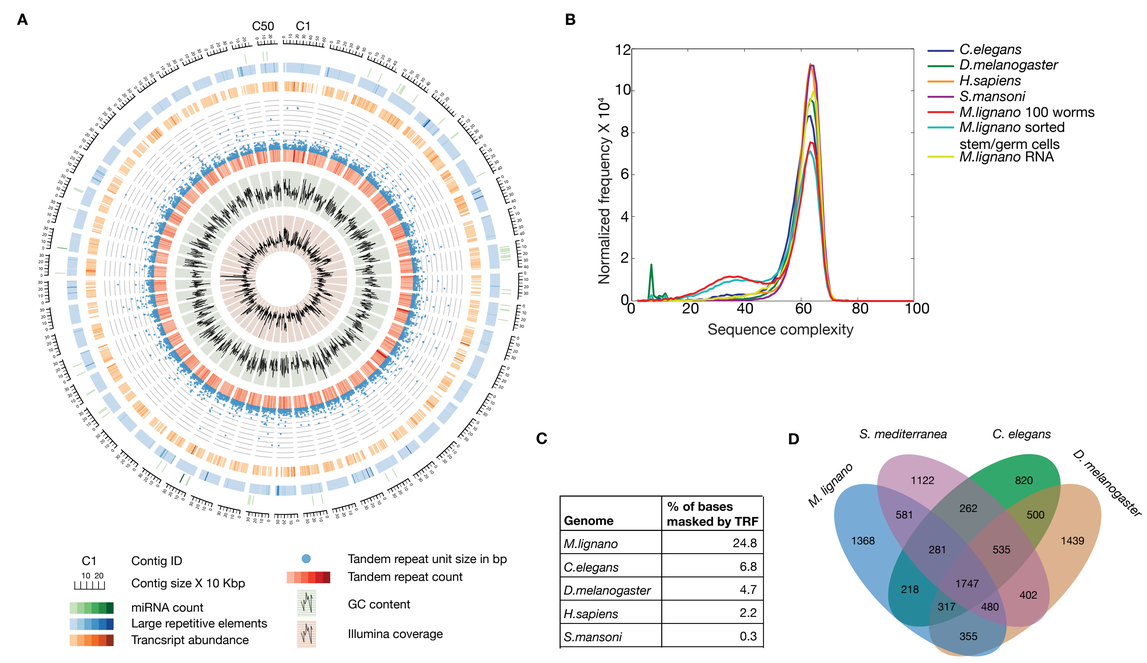
A. Phylogenetic analysis of 23 animal species using partial sequences of 43 genes. Fig. modified from Egger *et al.* (2015). **B**. Interference contrast image and a diagrammatic representation of an adult *Macrostomum lignano*. **C**. Phylogenomic analysis of 27 flatworm species (21 free-living and 6 neodermatan) using >100,000 aligned amino acids. Fig. modified from Egger *et al.* (2015). **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). T- testes, O - ovaries, DE – developing eggs, asterisks denote eyes. **F**. Representation of 23-mer frequency and coverage in the Illumina sequencing data generated from DNA extracted from a population of adult worms. *M. lignano* shows unusual 4-modal 23-mer distribution. **G**. Comparison of Illumina only (ML1) and Pacbio (ML2) assemblies. Contig length distribution (Log2 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.

Transcripts Involved in Regeneration in *M. lignano*



A. Schematic representation of the experimental design: 200 worms (per replicate) underwent amputation at a level between the brain and the gonads. The heads were allowed to regenerate, and regenerating animals were collected at different timepoints post amputation (0, 3, 6, 12, 24, 48, 72 hours). RNA-Seq libraries from each timepoint were analyzed for differentially expressed genes. Below - a heat map of differentially expressed genes at different regeneration timepoints. Each replicate is plotted separately. Downregulated and upregulated transcripts are labeled in green and red, respectively. Scale covers Log2 values. The samples are grouped with complete-linkage clustering using Euclidean distance. **B**.Known pluripotency pathways from *H. sapiens* and *M. musculus* were adapted from the Kyoto Encyclopedia of Genes and Genomes. Factors that had potential homologues in *M. lignano* are labeled.

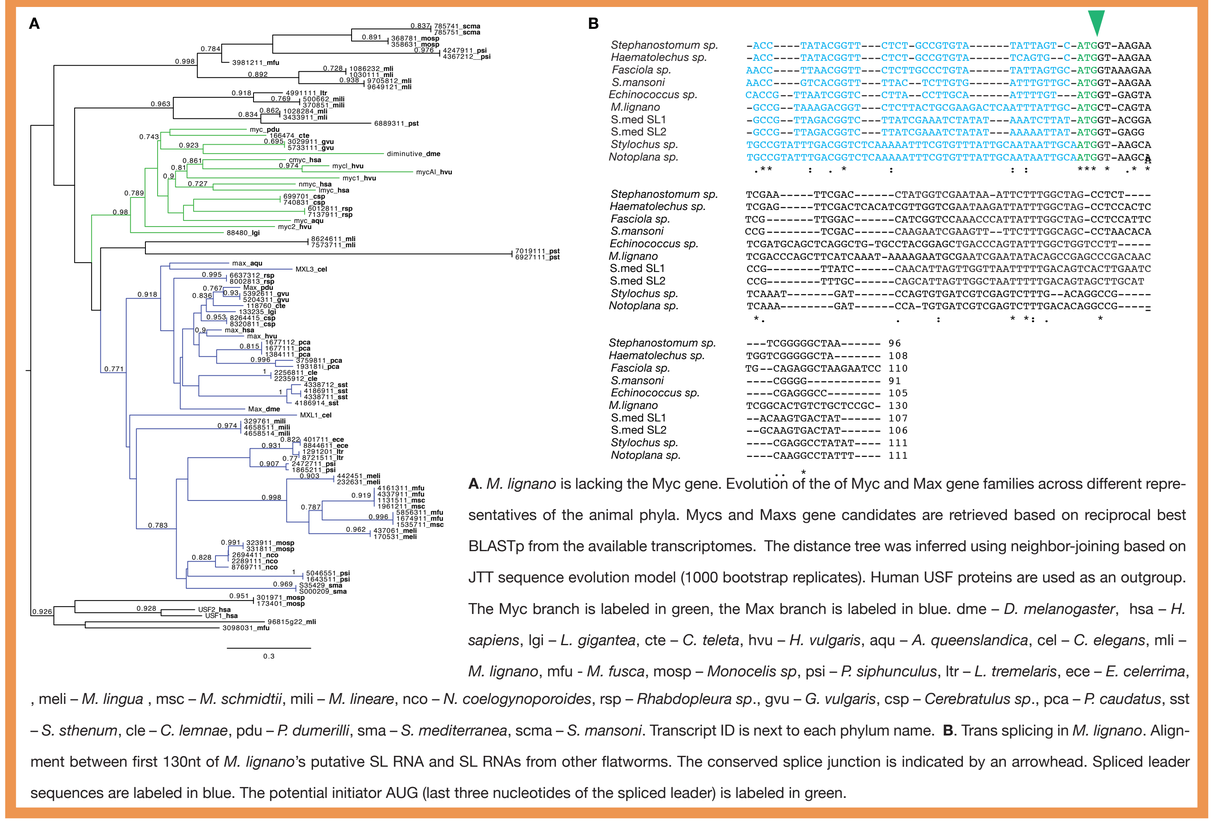
Genome and Transcriptome Annotation



Genome	% of bases masked by TRF
M.lignano	24.8
C.elegans	6.8
D.melanogaster	4.7
H.sapiens	2.2
S.mansoni	0.3

A.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 X 10 Kbp; miRNA count (1-54 mapped miRNAs); large repetitive elements (RepeatScout) (1-4476 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 (4-160X). The color gradients correspond to the range of values for each track (lower values are lighter, higher values are darker). **B**. Sequence complexity comparison across five organisms. *Drosophila melanogaster* has an abundance of very low complexity sequence, not found in the other species. *Macrostomum lignano* has a sizable amount of moderately complex sequences that are not found in other species and that do not appear to be expressed. **C**. Tandem Repeat Finder was run on five species to assess their tandem and low complexity sequence composition. *Macrostomum lignano* had far more bases masked by Tandem Repeat Finder than the other organisms in the test set. **D**.The number of reciprocal blast hits against the *Homo sapiens* transcriptome for four different species: *Macrostomum lignano, Schmidtea mediterranea, Caenorhabditis elegans*, and *Drosophila melanogaster*. Only the number of hits passing the E-value cutoff of ≤1e-10 is shown.

Additional Findings



A. *M. lignano* is lacking the Myc gene. Evolution of the of Myc and Max gene families across different representatives of the animal phyla. Mycs and Maxs gene candidates are retrieved based on reciprocal best BLASTp from the available transcriptomes. The distance tree was inferred using neighbor-joining based on JTT sequence evolution model (1000 bootstrap replicates). Human USF proteins are used as an outgroup. The Myc branch is labeled in green, the Max branch is labeled in blue. dme – *D. melanogaster*, hsa – *H. sapiens*, lgi – *L. gigantea*, cte – *C. teleta*, hvu – *H. vulgaris*, aqu – *A. queenslandica*, cel – *C. elegans*, mli – *M. lignano*, mfu - *M. fusca*, mosp – *Monocelis sp*, psi – *P. siphunculus*, ltr – *L. tremelaris*, ece – *E. celerrima*, , meli – *M. lingua* , msc – *M. schmidtii*, mili – *M. lineare*, nco – *N. coelogynoporoides*, rsp – *Rhabdopleura sp*., gvu – *G. vulgaris*, csp – *Cerebratulus sp*., pca – *P. caudatus*, sst – *S. sthenum*, cle – *C. lemnae*, pdu – *P. dumerilli*, sma – *S. mediterranea*, scma – *S. mansoni*. Transcript ID is next to each phylum name. **B**. Trans splicing in *M. lignano*. Alignment between first 130nt of *M. lignano*'s putative SL RNA and SL RNAs from other flatworms. The conserved splice junction is indicated by an arrowhead. Spliced leader sequences are labeled in blue. The potential initiator AUG (last three nucleotides of the spliced leader) is labeled in green.

Conclusions:

- We have assembled and annotated a highly repetitive genome using a mix of Pacbio and Illumina sequencing
- We have found that:
  - *M. lignano's* genome shows evidence of CpG methylation
  - It has retained a large number of homeoboxes as compared to other flatworms
  - The transcriptome shows evidence of trans-splicing
  - Flatworms lost the very conserved Myc gene
- We have characterized the gene expression patterns during regeneration in *M. lignano*
- Wasik *et al*. PNAS (2015); doi: 10.1073/pnas.1516718112