Whole genome analysis of a schistosomiasis-transmitting freshwater snail

Adema, Coen M.; Hillier LaDeana W.; Jones, Catherine S.; Loker, Eric S.; Knight, Matty; Minx, Patrick; Oliveira, Guilherme; Raghavan, Nithya; Shedlock, Andrew; Amaral, Laurence Rodrigues do; Arican-Goktas, Halime D.; Assis, Juliana; Baba, Elio Hideo; Baron, Olga Lucia; Bayne, Christopher J.; Bickham-Wright, Utibe; Biggar, Kyle K.; Blouin, Michael; Bonning, Bryony C.; Botka, Chris; Bridger, Joanna M.; Buckley, Katherine M.; Buddenborg, Sarah K.; Caldeira, Roberta Lima; Carleton, Julia; Carvalho, Omar S.; Castillo, Maria G.; Chalmers, Iain W.; Christensens, Mikkel; Clifton, Sandy; Cosseau, Celine; Coustau, Christine; Cripps, Richard M.; Cuesta-Astroz, Yesid; Cummins, Scott; di Stephano, Leon; Dinguirard, Nathalie; Duval, David; Emrich, Scott; Feschotte, Cédric; Feyereisen, Rene; FitzGerald, Peter; Fronick, Catrina; Fulton, Lucinda; Galinier, Richard; Geusz, Michael; Geyer, Kathrin K.; Giraldo-Calderon, Gloria; Gomes, Matheus de Souza; Gordy, Michelle A.; Gourbal, Benjamin; Grossi, Sandra; Grunau, Christoph; Hanington, Patrick C.; Hoffmann, Karl F.; Hughes, Daniel; Humphries, Judith; Izinara, Rosse; Jackson, Daniel J.; Jannotti-Passos, Liana K.; Jeremias, Wander de Jesus; Jobling, Susan; Kamel, Bishoy; Kapusta, Aurélie; Kaur, Satwant; Koene, Joris M.; Kohn, Andrea B; Lawson, Dan; Lawton, Scott P.; Liang, Di; Limpanont, Yanin; Lockyer, Anne E.; Lovato, TyAnna L.; Liu, Sijun; Magrini, Vince; McManus, Donald P.; Medina, Monica; Misra, Milind; Mitta, Guillaume; Mkoji, Gerald M.; Montague, Michael J.; Montelongo, Cesar; Moroz, Leonid L; Munoz-Torres, Monica C.; Niazi, Umar; Noble, Leslie R.; Pais, Fabiano; Papenfuss, Anthony T.; Peace, Rob; Pena, Janeth J: Pila, Emmanuel A.; Quelais, T; Raney, Brian J.; Rast, Jonathan P.; Ribeiro, Fernanda; Rollinson, David; Rotgans, Bronwyn; Routledge, Edwin J.; Ryan, Kathryn M.; Scholte, Larissa; Silva, Francislon; Storey, Kenneth B; Swain, Martin; Tennessen, Jacob A.; Tomlinson, Chad; Trujillo, Damian L.; Volpi, Emanuela V.; Walker, Anthony J.; Wang, Tianfang; Wannaporn, Ittiprasert; Warren, Wesley C.;

Wu, Xiao-Jun; Yoshino, Timothy P.; Yusuf, Mohammed; Zhang, Si-Ming; Zhao, Min; Wilson, Richard K.

Summary

Biomphalaria snails are instrumental in transmission of the human blood fluke *Schistosoma mansoni*. With the World Health Organization's goal to eliminate schistosomiasis as a global health problem by 2025, there is now renewed emphasis on snail control. Our characterization of the genome of *Biomphalaria glabrata*, a lophotrochozoan protostome, provides timely and important information on snail biology. We describe phero-perception, stress resistance, immune function and multi-level regulation of gene expression that support the persistence of *B. glabrata* in aquatic habitats and define this species as a suitable snail host for *S. mansoni*. We identify several potential targets for developing novel control measures aimed at reducing snail-mediated transmission of schistosomiasis.

The fresh water snail *Biomphalaria glabrata* (Lophotrochozoa, Mollusca) is of medical relevance as this Neotropical gastropod contributes as obligate intermediate host of *Schistosoma mansoni* (Lophotrochozoa, Platyhelminthes) to transmission of the neglected tropical disease (NTD) human intestinal schistosomiasis[1]. An *S. mansoni* miracidium initiates a chronic parasite-snail infection that alters *B. glabrata* immunity and metabolism, causing parasitic castration such that the snail does not reproduce but instead supports generation of cercariae, the human-infective stage of *S. mansoni*. Patently infected snails continuously release free-swimming cercariae that penetrate the skin of humans that they encounter in their aquatic environment. Inside the human definitive host, *S. mansoni* matures to adult worms that reproduce sexually in the venous system surrounding the intestines, releasing eggs, many of which pass through the intestinal wall and are deposited in water with the feces. Miracidia hatch from the eggs and complete the life cycle by infecting another *B. glabrata*. Related *Biomphalaria* species transmit *S. mansoni* in Africa. Schistosomiasis is chronically debilitating; estimates of disease burden indicate that disability-adjusted life years (DALYs) lost due to morbidity rank schistosomiasis second only to malaria among parasitic diseases in impact on global human health[2].

In the absence of a vaccine, current control measures emphasize mass drug administration (MDA) of praziquantel (PZQ), the only drug available for large-scale treatment of schistosomiasis[3]. Schistosomes, however, may develop resistance and reduce the effectiveness of PZQ[4]. Importantly, PZQ treatment does not protect humans against rapid re-infection following exposure to water-borne cercariae released from infected snails. Snail-mediated parasite transmission must be interrupted to achieve long-term sustainable control of human schistosomiasis[5]. The World Health Organization has set a strategy for the global elimination of schistosomiasis as a public health threat by the year 2025 that recognizes both MDA approaches and targeting of the snail intermediate host as crucial

toward achieving this goal[6]. This significant undertaking provides added impetus for detailed study of *B. glabrata*.

Genome analyses of *B. glabrata* also provide evolutionary insights by increasing the relatively few taxa of the Lophotrochozoa that have been characterized. To date a majority of lophotrochozoan genomes have been derived from parasitic, sessile, or marine animals (i.e. platyhelminths, leech, bivalve, cephalopod, polychaete)[7-11]. Thus, *B. glabrata* likely possesses novel genomic features that characterize a free-living lophotrochozoan from a freshwater environment. Here we characterize the *B. glabrata* genome and describe biological properties that enable the snail's persistence in aquatic environments and define this snail as a suitable host for *S. mansoni*, including aspects of immunity and gene regulation. These efforts, we anticipate, will foster developments to interrupt snail-mediated parasite transmission in support of schistosomiasis elimination.

The *B. glabrata* genome comprises eighteen chromosomes (Supplementary Text 1 and Supplementary Fig. 1-3) and has an estimated size of 916Mb[12]. We assembled the genome of BB02 strain *B. glabrata*[13] from short reads, Sanger sequences, and 454 sequence data (~27.5X coverage) and used a linkage map to assign genomic contigs to linkage groups (Supplementary Text 2 and Supplementary Table 1). We captured transcriptomes from twelve different tissues (Illumina RNAseq) and mapped these reads to the assembly to aid annotation (Supplementary Text 3, Supplementary Figs. 4-8 and Supplementary Tables 2-7).

Olfaction in an aquatic environment

Aquatic molluscs employ proteins for communication; e.g. *Aplysia* attracts conspecifics using water-soluble peptide pheromones[14]. We collected *B. glabrata* proteins from snail conditioned water (SCW) and following electrostimulation (ES), which induces rapid release of proteins. NanoHPLC-MS/MS identified 177 secreted proteins shared among 533 proteins from SCW and 232 proteins

from ES (Supplementary Table 8). Detection of an ortholog of temptin, which plays a key role in pheromone attraction of *Aplysia*[15], suggests an operational pheromone sensory system in *B. glabrata*. To gain insight into the mechanisms for pheromone perception, we analyzed the *B. glabrata* genome for olfactory G-protein coupled receptor (GPCR)-like supergene families and identified 242 seven transmembrane domain *GPCR*-like genes belonging to fourteen subfamilies that are clustered in the genome (Supplementary Text 4, Supplementary Figs. 9,10 and Supplementary Table 9). RT-PCR and *in situ* hybridization confirmed expression of a *GPCR*-like gene within the mollusc's tentacles, known to be involved in chemosensation (Figure 1). This communication system allows *B. glabrata* to detect conspecifics in aquatic environments but may have a tradeoff effect by potentially exposing the snail as a suitable target for pathogens.

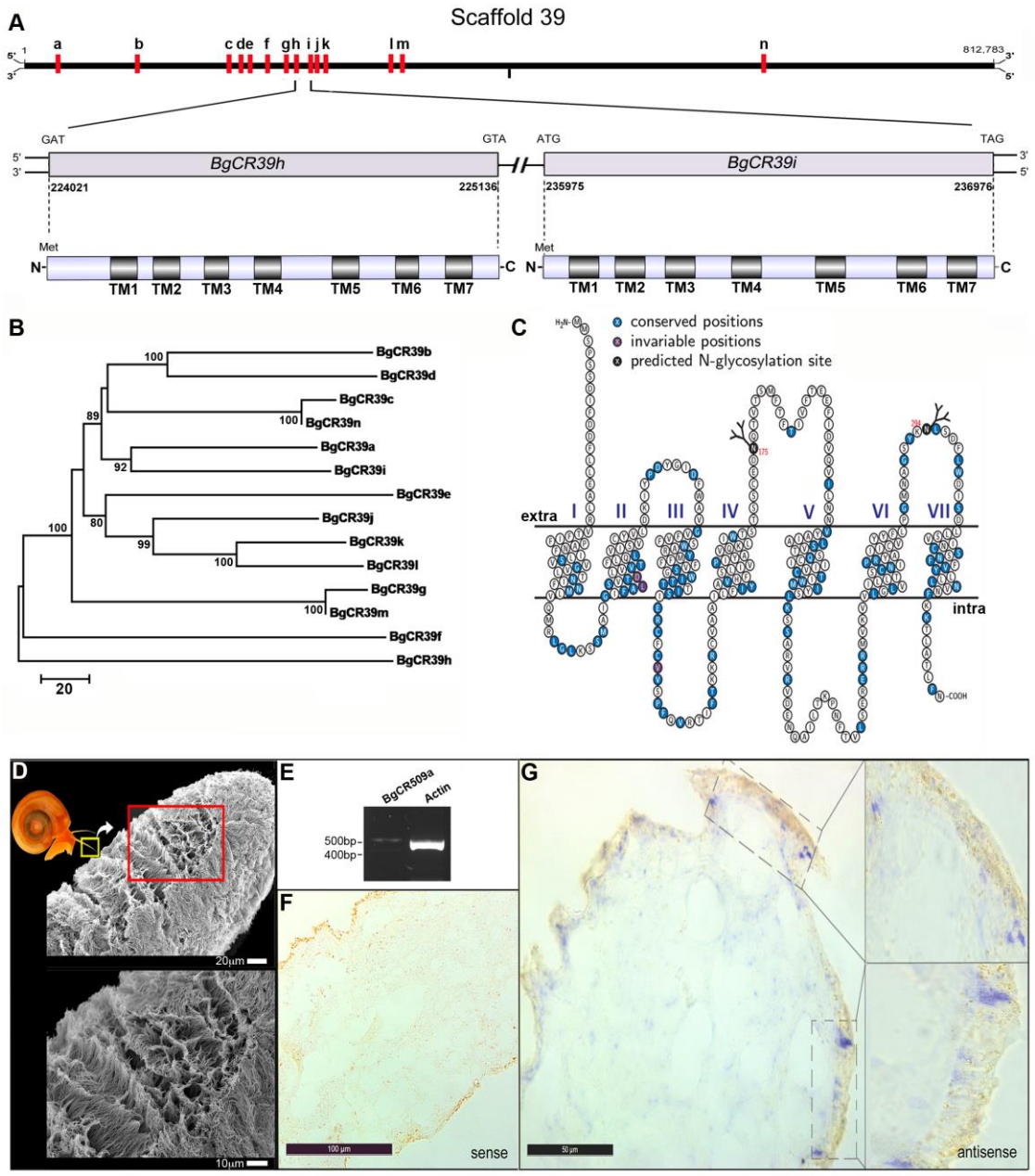


Fig 1. Candidate chemosensory receptors of *B. glabrata*. (A) Scaffold 39 contains fourteen candidate chemosensory receptor genes (BgCRa-n). Most encode 7-transmembrane domain proteins, BgCRm and BgCRn are truncated to six-transmembrane domains. Genes occur in both directions. (B) Phylogenetic analysis of scaffold 39 chemosensory receptors. (C) Schematic of receptor showing conserved and invariable amino acids, transmembrane domains I-VII; and location of glycosylation sites. (D) Scanning electron micrograph showing anterior tentacle, with cilia covering the surface. (B) RT-PCR gel showing amplicon for BgCR509a and actin from *B. glabrata* tentacle. (F, G) *In situ* hybridization showing sense (negative control) and antisense localization of BgCR509a in anterior tentacle section (purple).

Stress responses and immunity

Biomphalaria glabrata endures environmental insults, including biotic and abiotic stress from heat, cold, xenobiotics, pollutants, injury and infection. Additional to previous reports of *Capsaspora*[16] as single-cell eukaryote endosymbiont, we noted from the sequenced material an unclassified mycoplasma or related mollicute bacteria and viruses (Supplementary Text 5,6, Supplementary Figs. 11,12 and Supplementary Table 10). For anti-stress responses, *B. glabrata* genome has five families of heat shock proteins (HSP), including Hsp20, Hsp40, Hsp60, and Hsp 90. The Hsp70 gene family is the largest with six multi-exon genes, five single exon genes, and over ten pseudogenes (Supplementary Text 7, Supplementary Figs. 13-16, Supplementary Table 11). These genes are retained in *B. glabrata* embryonic (*Bge*) cells, the only available molluscan cell line[17], such that anti-stress function in *B. glabrata* can be investigated *in vitro* using cell cultures (Supplementary Text 8, Supplementary Figs. 17-21, and Supplementary Table 12,13).

The *B. glabrata* genome contains about 99 genes encoding heme-thiolate enzymes (CYP superfamily) for detoxifying xenobiotics. All major animal cytochrome P450 clans are represented, including ~33 CYP2 genes, ~35 CYP3 genes, and ~7 CYP4 genes. The presence of 18 genes of the mitochondrial clan suggests that molluscs, like arthropods, but unlike vertebrates, utilize mitochondrial P450s for detoxification[18]. Tissue-specific expression (e.g. four uniquely in ovotestis) suggests that fifteen P450 genes serve specific biological processes. Cataloguing the CYPome sets the groundwork for rational design of selective molluscicides, e.g. by inhibiting unique P450s or using *B. glabrata*-specific P450s for bioactivation of pro-molluscicides (Supplementary Text 9).

Biomphalaria glabrata possesses a diversity of pattern recognition receptors (PRRs) for immune recognition of pathogens[19]. This includes 56 Toll-like receptor (TLR) genes, of which 27 encode complete TLRs (Fig 2, Supplementary Text 10 and Supplementary Table 14), associated with a TLR

signaling network for transcriptional regulation through NF-κB transcription factors (Supplementary Text 11, Supplementary Fig. 22 and Supplementary Table 16). Like other lophotrochozoans, *B. glabrata* shows a moderate expansion of TLR genes relative to mammals and insects which have ~10 TLRs[20]. Other PRRs include eight peptidoglycan recognition-binding proteins (PGRPs), a single Gram-negative binding protein (GNBP) and seventeen genes encoding for single scavenger receptor cysteine-rich (SRCR) domains (Supplementary Table 15). A prominent category of *B. glabrata* PRRs consists of fibrinogen-related proteins (FREPs), plasma lectins that are somatically mutated to yield uniquely diversified FREP repertoires in individual snails[21]. The *B. glabrata* germline includes four FREP genes comprising one immunoglobulin (IgSF) domain and one C-terminal fibrinogen (FBG)-like domain, including one gene with an N-terminal PAN_AP domain, and twenty FREP genes with two upstream IgSF domains preceding an FBG domain. FREP genes cluster in the genome, often accompanied by partial FREP-like sequences (Supplementary Text 12 and Supplementary Figs. 23-26). A proteomics level study of anti-parasite immunity showed that some FREPs exhibiting binding reactivity to *S. mansoni* derive from different gene families between susceptible and resistant *B. glabrata* strains. Moreover, a FREP-like lectin that contains a galectin instead of a C-terminal fibrinogen domain (designated galectin-related protein; GREP[22]) is uniquely associated with the resistant snail phenotype (Supplementary Text 8).

We identified several cytokines, including twelve homologs of IL-17, four MIF homologs, and eleven TNF molecules (Supplementary Text 10, Supplementary Table 15). Orthologs of complement factors indicate that *B. glabrata* can mount complement-like responses to pathogens (Supplementary Text 13, Supplementary Figs. 27,28 and Supplementary Tables 17-28). We discovered an

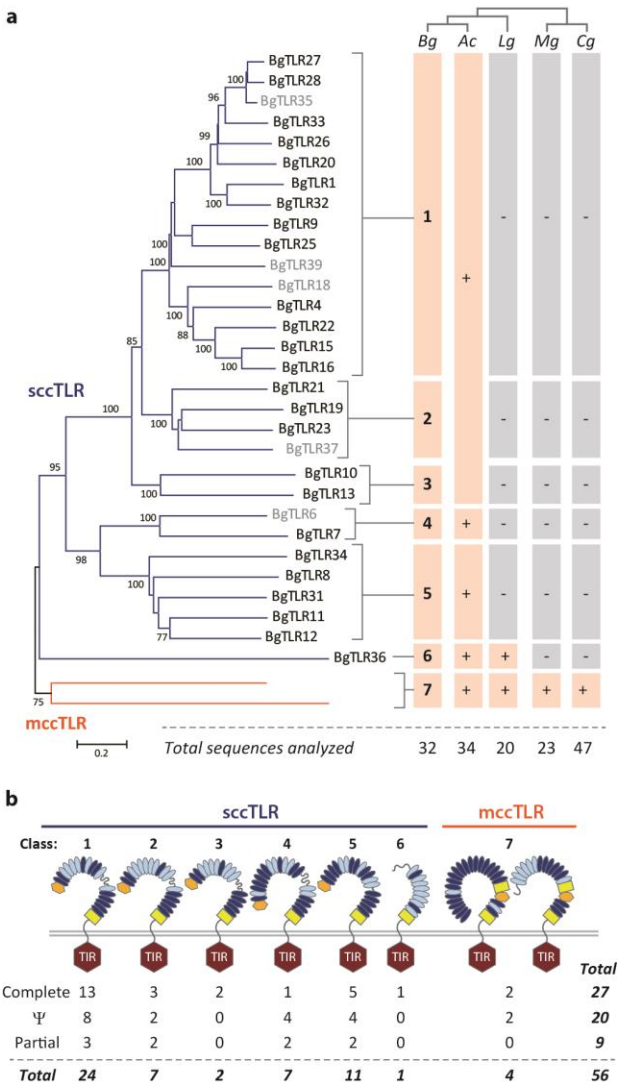


Fig 2. TLR genes in *B. glabrata*. (A) Analysis of the (only complete) TIR domains from BgTLRs identified seven classes (Neighbor-joining tree shown). Bootstrap values shown for 1,000 replicates. Comparisons included TLRs from *A. californica* (Ac), *L. gigantea* (Lg), *Mytilus galloprovincialis* (Mg), and *C. gigas* (Cg). The presence or absence of orthologs of each class in each molluscan species is indicated. A representative of the *B. glabrata* class 1/2/3 clade is present within *A. californica*, but is independent of the *B. glabrata* TLR classes (indicated by the large pink box). Grey font indicates pseudogenes or partial genes. (B) *B. glabrata* has both single cysteine cluster (scc; blue line)- and multiple cysteine cluster (mcc; orange line) TLRs. Domain structures are shown for each of BgTLR class. BgTLRs consist of an LRRNT (orange hexagon), a series of LRRs (ovals), a variable region (curvy line), LRRCT (yellow box), and transmembrane domain, and an intracellular TIR domain (red hexagon). The dark blue ovals indicate well defined LRRs (predicted by LRRfinder[23]); light blue ovals are less confident predictions. Each of the two class 7 BgTLRs has a distinct ectodomain structure. The numbers of complete, pseudogenes (Ψ) and partial and genes are indicated for each class.

extensive toolkit for apoptosis, a response that can regulate invertebrate immune defense[24]. This includes at least 50 genes encoding for Baculovirus IAP Repeat (BIR) domain-containing caspase inhibitors, localized among 26 genomic scaffolds. The expansion of this gene family in molluscs (17 genes in *Lottia gigantean*, 48 in *Crassostrea gigas*), relative to other animal clades, suggests important regulatory roles in apoptosis and innate immune responses[25] (Supplementary text 14, Supplementary Figs. 29-31 and Supplementary Table 29). Finally, *B. glabrata* possesses a gene complement to metabolize reactive oxygen species (ROS) and nitric oxide (NO), that are generated by hemocytes to exert cell-mediated cytotoxicity toward pathogens, including schistosomes (Supplementary Text 15, Supplementary Fig. 32 and Supplementary Table 30).

Searches for antimicrobial peptide (AMP) sequences of *B. glabrata* indicated only a single macin-type gene family, comprising six biomphamacin genes. While the AMP arsenal is reduced compared to other invertebrate species (e.g., bivalve molluscs have multiple AMP gene families[26]), *B. glabrata* possesses several multigenic families of antibacterial proteins including two achacins, five LBP/BPIs, and 21 biomphalysins (Supplementary Text 16,17, Supplementary Figs. 33,34 and Supplementary Tables 31,32). Gaps in functional annotation limit our interpretation of immune function in *B. glabrata* (Supplementary Text 18 and Supplementary Table 33,34). Computational analysis indicated that 15% of the predicted proteome consists of (583) secreted proteins. Immune-relevant tissues showed high expression of 100 secreted proteins, including FREP3, other lectins and novel proteins that potentially represent candidate immune factors for future study. (Supplementary text 3, Supplementary Fig. 7, Supplementary Tables 5-6).

Regulation of biological processes

Epigenetic regulation allows dynamic use of genes contained in the *B. glabrata* genome[27-29]. The chromatin-modifying enzymes include class I and II histone methyltransferases, LSD-class and

Jumonji-class histone demethylases, class I – IV histone deacetylases, and GNAT, Myst and CBP superfamilies of histone acetyltransferases. We identified homologs of DNA (cytosine-5)-methyltransferases 1 and 2 (no homolog of DNMT3), as well as putative methyl-CpG binding domain proteins 2/3. *In silico* analyses predicted a mosaic type of DNA methylation, as is typical for invertebrates (Supplementary Text 19, Supplementary Figs. 35-39, Supplementary Table 35). The potential role of DNA methylation in *B. glabrata* reproduction and *S. mansoni* interactions is reported elsewhere (K.G.G., U.H.N., D.D., Ce.C., C.T., I.W.C., M.T.S., U.B.-W., Sabrina E. Munshi, C.G., T.P.Y. and K.F.H., in preparation).

The *B. glabrata* genome also provides the protein machinery for biogenesis of microRNA (miRNAs) to regulate gene expression. Moreover, two computational methods independently predicted the same 95 pre-miRNAs, encoding 102 mature miRNAs. Of these, 36 predicted miRNAs were observed within our transcriptome data, while 53 displayed ≥ 90% nucleotide identity (with 25 at 100%) to *L. gigantea* miRNAs. One bioinformatics pipeline predicted 107 additional pre-miRNAs unique to *B. glabrata*. Several *B. glabrata* miRNAs are predicted to regulate transcripts for processes unique to snail biology, including secretory mucosal proteins and shell formation that may present possible targets for control of *B. glabrata* (Supplementary Text 20,21, Supplementary Figs. 40-67, Supplementary Tables 36-45).

Aspects of *B. glabrata* biology are periodic[30], likely due to control of circadian timing mechanisms. We identified seven candidate clock genes *in silico*, including a gene with strong similarity to the *period* gene of *A. californica.* Modification of expression of these putative clock genes may interrupt circadian rhythms and affect feeding and egg-laying of *B. glabrata* (Supplementary Text 22).

Neuropeptides expressed within the nervous system coordinate the complex physiology of *B. glabrata*, including regulation of male and female functions in this simultaneous hermaphrodite

snail. *In silico* searches identified 43 neuropeptide precursors in *B. glabrata* that were predicted to yield over 250 mature signaling products. Neuropeptide transcripts occurred in multiple tissues, yet some were most prominent within terminal genitalia (49%) and the CNS (56%), or even specific to the CNS, including gonadotropin-releasing hormone (GnRH) and insulin-like peptides 2 and 3 (Supplementary Text 23, Supplementary Fig. 68, Supplementary Tables 46-48).

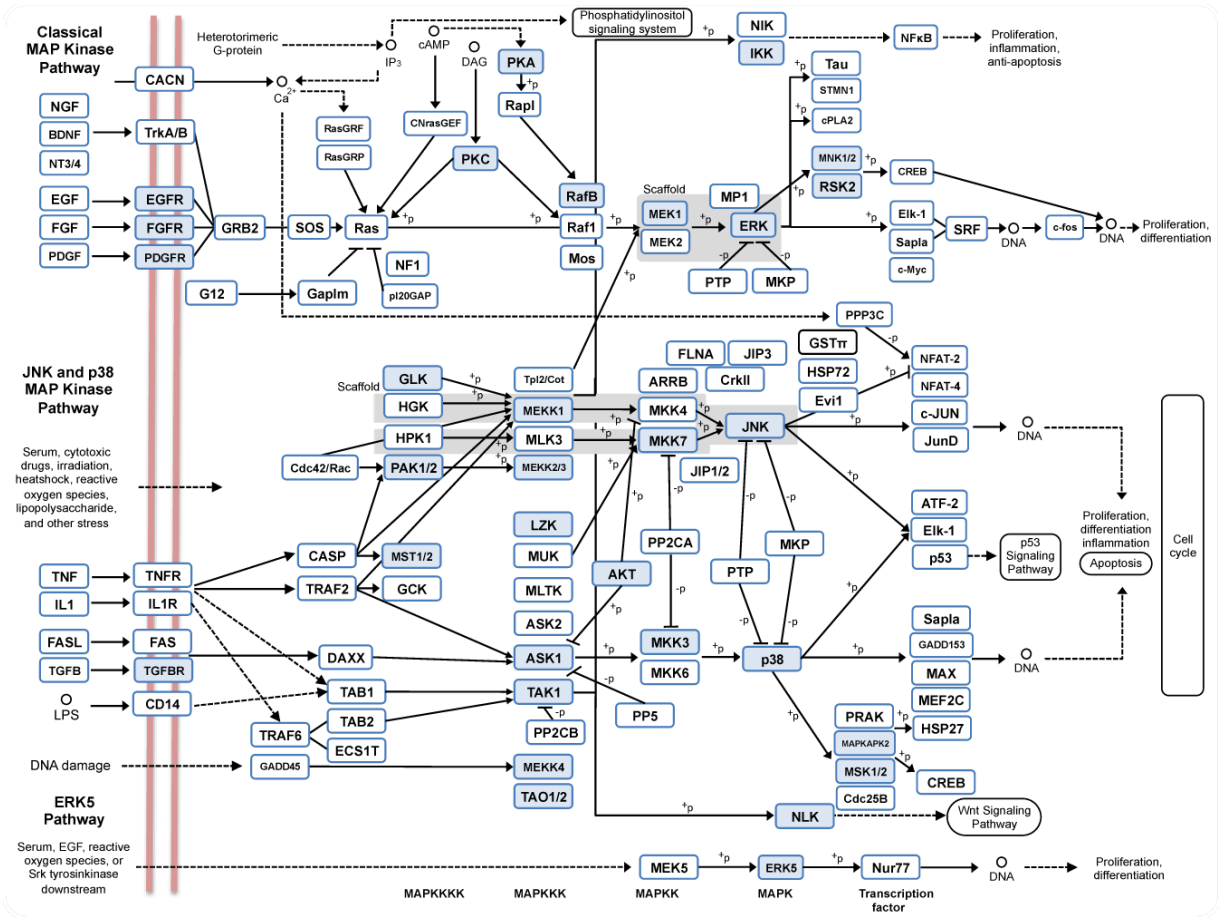
A role of steroid hormones in reproduction of hermaphrodite snails with male and female reproductive organs remains speculative. *Biomphalaria glabrata* has a CYP51 gene to biosynthesize sterols *de novo*, yet we found no orthologs of genes involved in either vertebrate steroid or arthropod ecdysteroid biosynthesis. The lack of CYP11A1 suggests that *B. glabrata* cannot process cholesterol to make vertebrate-like steroids. The absence of aromatase (CYP19), required for the formation of estrogens, is particularly enigmatic as molluscs possess homologs of mammalian estrogen receptors. Characterization of snail-specific aspects of steroidogenesis may identify targets to disrupt reproduction toward control of snails. (Supplementary Text 24, Supplementary Fig. 69, Supplementary Table 49).

The reproductive physiology of hermaphroditic snails is also modulated by male accessory gland proteins (MAGPs), which are delivered with spermatozoa to augment fertilization success[31]. The *B. glabrata* genome has sequences matching one such protein, Ovipostatin (LyAcp10), but none of the other MAGPs identified in *Lymnaea stagnalis*[32]. Perhaps MAGPs evolved rapidly and are putatively taxon-specific (Supplementary Text 25, Supplementary Fig. 70 and Supplementary Table 50).

KEGG Enrichment analyses[33] identified metabolic and biochemical pathways in *B. glabrata* and revealed distinct organ specific patterns of gene expression (Supplementary Text 3,26, Figs. 8, 71-80, Supplementary Tables 3,51). Half of 5,000 annotated sequences were assigned to 200 biochemical pathways (Supplementary Text 3 and Supplementary Tables 2,3). Mapping biological

processes onto the snail anatomy helps interpreting *B. glabrata*'s responses to environmental insults and pathogens.

For survival in *B. glabrata, S. mansoni* alters the snail host physiology, possibly by interfering with the extracellular signal-regulated kinase (ERK) pathway[34] and other signaling pathways. Eukaryotic protein kinases (ePKs) mediate signal transduction through phosphorylation in complex networks, and protein phosphatases counteract these effects towards effective signaling. Similarity searches for conserved domains identified 240 *B. glabrata* ePKs encompassing all main types of animal ePKs (Supplementary Text 3 and Supplementary Fig. 6). We also identified 60 putative protein phosphatases comprising ~36 protein Tyr phosphatases (PTPs) and ~24 protein Ser/Thr phosphatases (PSPs) (Supplementary Text 27, Supplementary Figs. 81-83). These sequences can be studied for understanding control of homeostasis, particularly in the face of environmental and pathogenic insults encountered by *B. glabrata* (Figure 3).



247

248 Fig 3. MAPK signaling pathway in *B. glabrata*. Blue shaded blocks correspond to proteins identified
249 in the *B. glabrata* predicted proteome, white blocks represent mammalian proteins without homologs
250 in the snail proteome. The +p signal represents phosphorylation and the –p signal represents protein
251 dephosphorylation.
252

Bilaterian evolution

We searched the *B. glabrata* genome for core cardiac-specification and -differentiation genes. A previously characterized short cDNA sequence from snail heart RNA led to identification of BGLB012592 as the *Biomphalaria* ortholog of *tin/Nkx2.5* [35]. Similarity searches with *Drosophila* orthologs identified most of the core cardiac regulatory factors and structural genes in the *B. glabrata* genome (Supplementary Text 28, Supplementary Table 52), with enriched expression of these genes in cardiac tissues (Figure 4). These results from a lophotrochozoan, in conjunction with ecdysozoans and deuterostomes, further support the hypothesis that a primitive heart-like structure, which developed through the actions of a core heart toolkit, was present in the urbilaterian ancestor.

Actins are conserved proteins that function in cell motility (cytoplasmic actins) and muscle contraction (sarcomeric actins)[36]. The clustering across seven scaffolds suggests that some of the ten *B. glabrata* actin genes arose through tandem duplication. Expression across all tissues indicates that four genes encode cytoplasmic actins (Figure 4). Protein sequence comparisons placed all *B. glabrata* actins as most closely related to mammalian cytoplasmic rather than sarcomeric actins (Supplementary Text 29 and Supplementary Table 53), a pattern also observed for all six actin genes of *D. melanogaster* [37]. The actin genes of *B. glabrata* and other molluscs were most similar to paralogs within their own genomes, rather than to other animal orthologs (Figure 4). One interpretation is that actin genes diverged independently multiple times in molluscs, similar to an earlier hypothesis for independent actin diversification in arthropods and chordates[38]. Alternatively, a stronger appearance of monophyly than really exists may result if selective pressures due to functional constraints keep actin sequences similar within a genome, for example if the encoded proteins have overlapping functions.

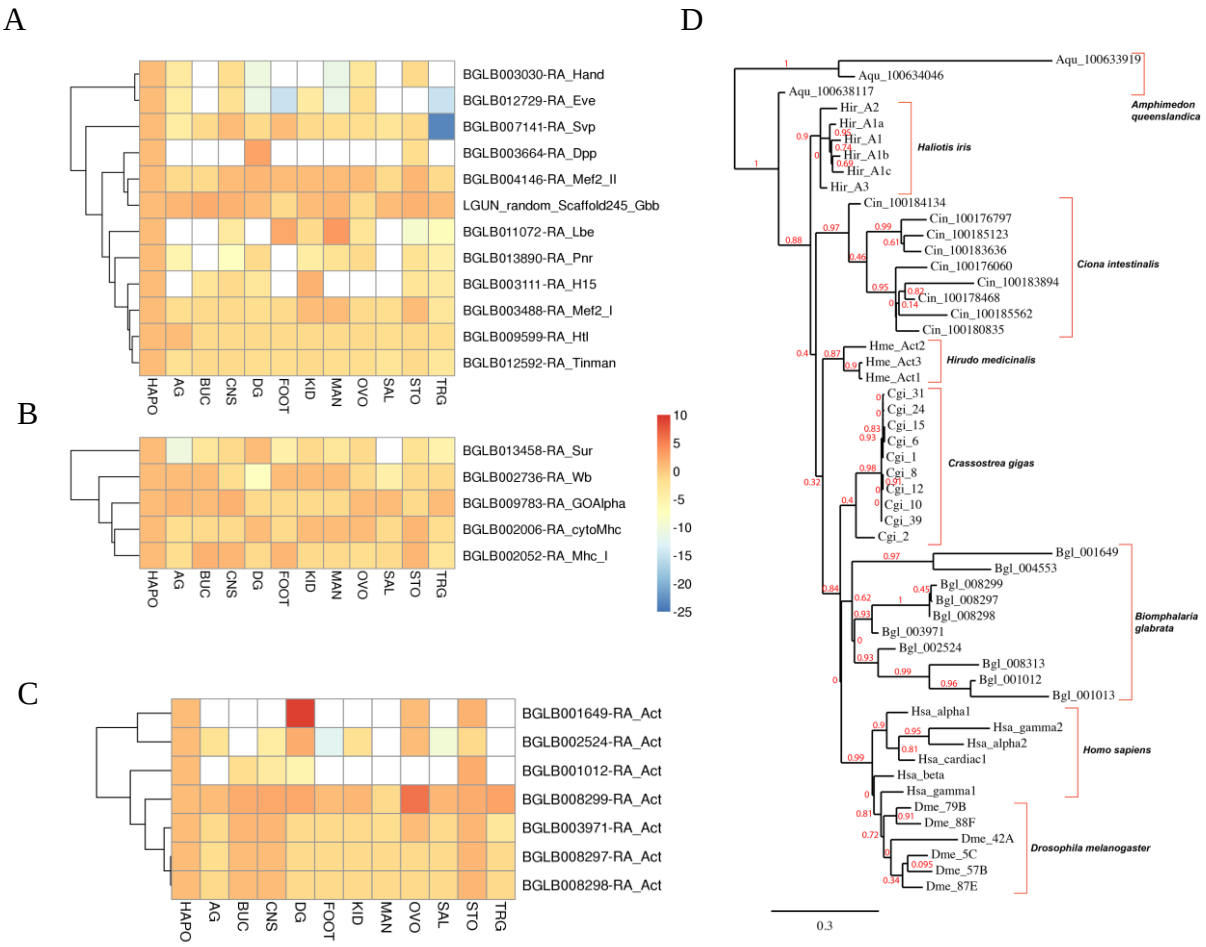


Fig 4. Expression of cardiac genes and actin genes in *B. glabrata* tissues determined through RNA-seq analysis. (A) cardiac regulatory genes. (B) cardiac structural genes. (C) Relative expression of actin genes in *B. glabrata* tissues. (D) Phylogenetic relationships of actin genes.

AG - Albumen gland; BUC - buccal mass; CNS - central nervous system; DG - digestive gland; FOOT – headfoot; HAPO - heart/APO; KID – kidney; MAN - mantle edge; OVO – ovotestes; SAL - salivary glands; STO - stomach; TRG - terminal genitalia *Amphimedon* – marine sponge; *Haliotis* – abalone; *Ciona* – sea squirt; *Hirudo* – leech; *Crassostrea* – oyster; *Biomphalaria* – pond snail; *Homo* – human; *Drosophila* – fruit fly.

We analyzed the transcriptomic data for *B. glabrata* genes involved in biomineralization. Of 1,211 transcripts that were >2-fold up-regulated in the mantle relative to other tissues, 34 shared similarity with molluscan sequences known to be involved in shell formation and biomineralization. Another 177 candidate sequences putatively involved in shell formation including eighteen genes (10.2%) with similarity to sequences of shell forming secretomes of other marine and terrestrial molluscs were identified from the entire mantle transcriptome (Figure 5). Like previous studies, our results show both diverse and conserved aspects of molluscan shell-forming strategies. Highly conserved components of the molluscan shell forming toolkit include carbonic anhydrases and tyrosinases[9]. (Supplementary Text 30, Supplementary Fig. 84, Supplementary Tables 54).

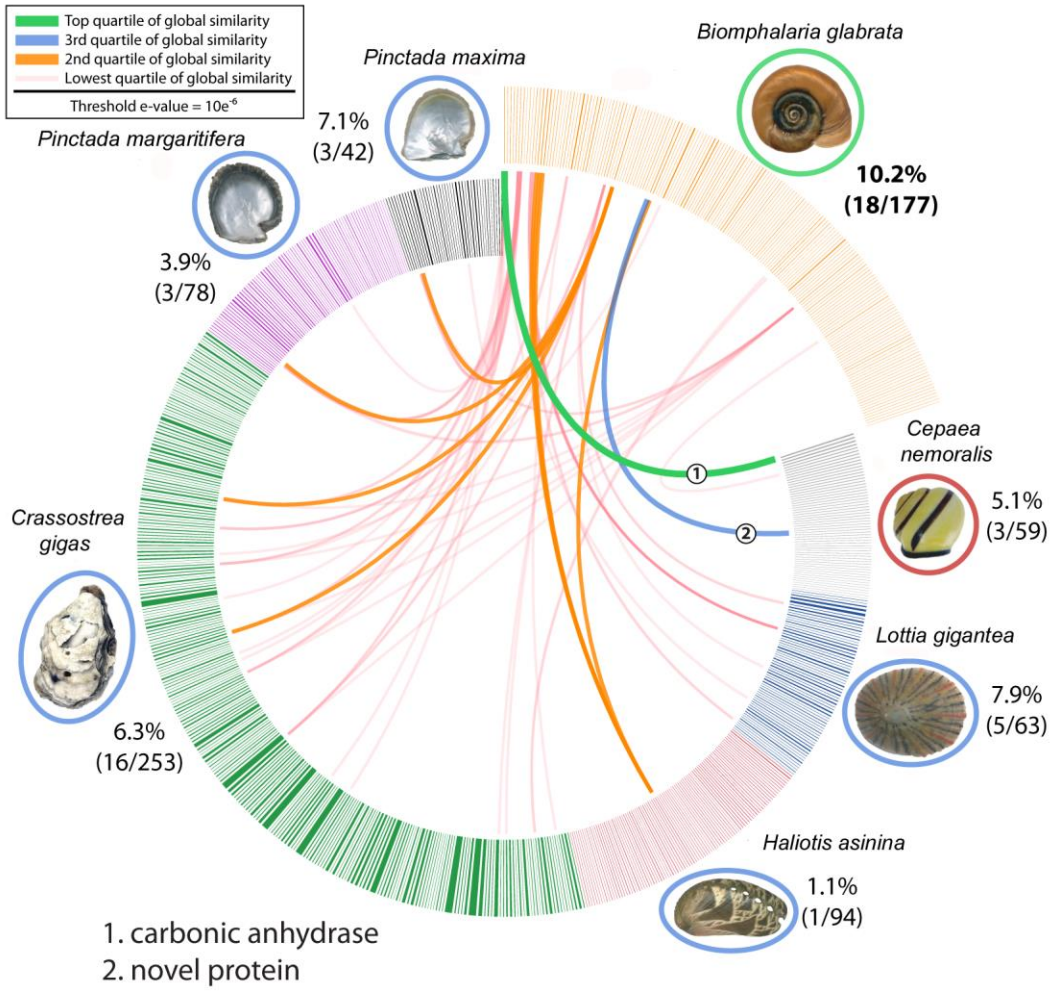


Fig 5. Circos diagram of 177 mantle specific, secreted *Biomphalaria* gene products compared against six other molluscan shell forming proteomes (BLASTp threshold $\leq 10e^{-6}$). Protein pairs that share sequence similarity in the top quartile are linked in green, the second quartile is linked in blue, third quartile has orange links and lowest quartile of similarity has red links. Marine species are circled in blue, freshwater species in green, terrestrial species in red. Percentages (proportions in brackets) indicate the number of proteins that shared similarity with a *Biomphalaria* shell forming candidate gene.

Repetitive landscape

Repeat content analysis showed that 44.8% of the *B. glabrata* assembly consists of transposable elements (TEs; Fig 6, Supplementary Text 31, Supplementary Figs. 87-87 and Supplementary Table 55), higher than observed from other molluscs: Pacific oyster, *C. gigas* (36%)[9], owl limpet, *L. gigantea* (21%)[8], sea hare, *A. californica* (30%)[39], and comparable to *Octopus bimaculoides* (43%)[11]. The fraction of unclassified elements in *B. glabrata* was high (17.6%). Most abundant classified repeats were LINEs, including Nimbus[40] (27% of TEs, 12.1% of the genome), and DNA TEs (17.7% of TEs, 8% of the genome), long terminal repeats (LTRs) were a small percentage of the repetitive elements (6% of TEs, 1.7% of the genome). Non-mobile simple repeats comprised 2.6% of the genome with an abundance of short dinucleotide satellite motifs. Divergence analyses of element copy and consensus sequences indicated that DNA TEs were not recent invaders of the *B. glabrata* genome, no intact transposases were detected. A hAT DNA transposon of *B. glabrata* (~1000 copies) has significant identity with *SPACE INVADERS (SPIN)* which horizontally infiltrated a range of animal species, possibly through host-parasite interactions[41]. Overall, our results reinforce a model in which diverse repeats comprise a large fraction of molluscan genomes.

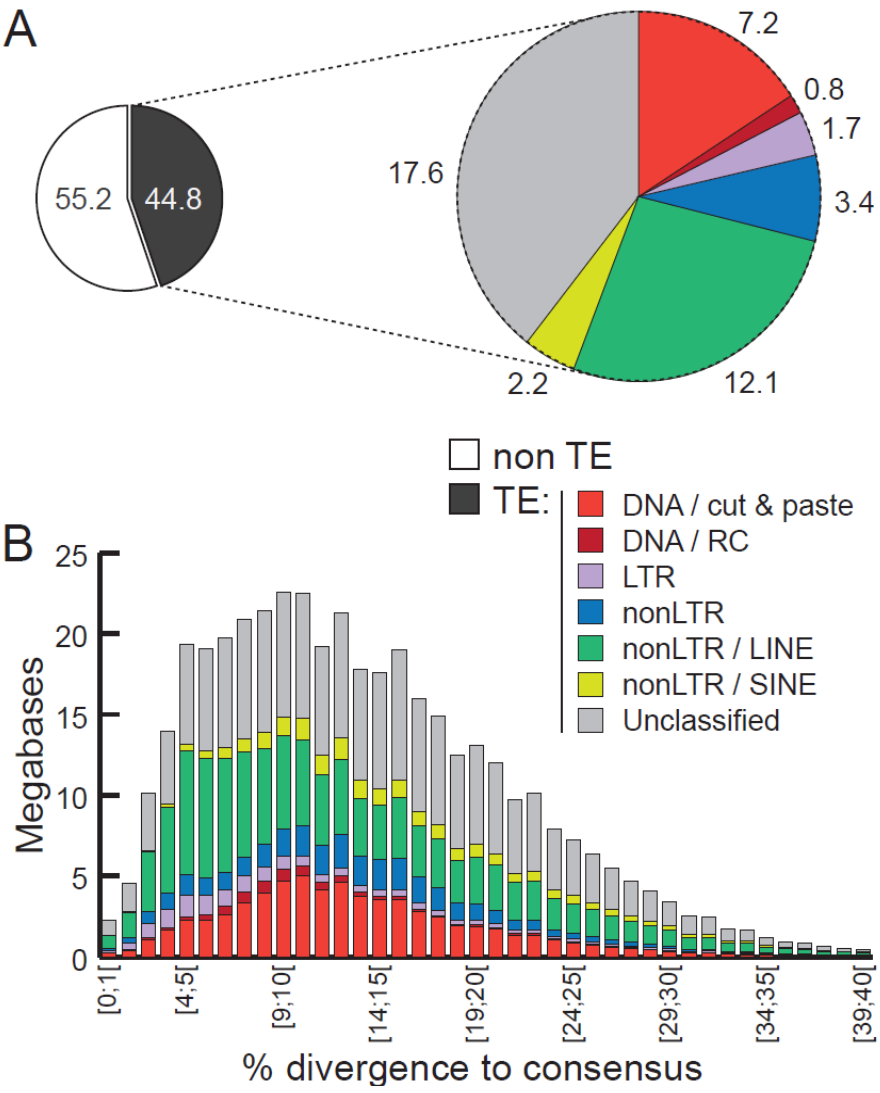


Fig 6. Transposable element (TE) landscape of *B. glabrata*. (A) Left: proportion of DNA annotated as TE (in black) or not annotated (white). Right: TE composition by class. Numbers represent the percentage of the genome corresponding to each class. **(**B) Evolutionary view of TE landscape. For each class, cumulative amounts of DNA (in Mb) are shown in function of the percentage of divergence to the consensus (by bins of 1%, first one being ≥0 and <1; see Methods). Percentage of divergence to the consensus is used as a proxy for age: the older the invasion of the TE is, the more copies will have accumulated mutations (higher percentage of divergence, right of the graph). Conversely, sequences corresponding to youngest elements show little divergence to consensus (left of the graph). RC: rolling circle.

Concluding remarks

The genome of the Neotropical freshwater snail *B. glabrata* expands insights into animal biology by further defining the lineage of the Lophotrochozoa relative to Ecdysozoa and Deuterostomia among higher animals. An important rationale for analysis of the genome of *B. glabrata* pertains to its role in transmission of *S. mansoni* in the New World. Moreover, most of the world's cases of *S. mansoni* infection occur in sub-Saharan Africa where other *Biomphalaria* species are responsible for transmission, most notably *Biomphalaria pfeifferi*. Due to a shared common ancestor, *B. glabrata* likely provides a good representation of the genomes of African *Biomphalaria* species[42,43] . This notion is supported by at least 90% sequence identity shared among 196 assembled transcripts collected from *B. pfeifferi* (Illumina RNAseq) with the transcriptome of *B. glabrata* (Supplementary Text 32 and Supplementary Tables 56-58). This report provides novel details on the biological properties of *B. glabrata* and points to potential strategies for more effective surveillance and control efforts against *Biomphalaria* to limit the transmission of schistosomiasis.

Methods

The genetic material used for sequencing the hermaphroditic freshwater snail *B. glabrata* was derived from multiple snails of the BB02 strain, established at the University of New Mexico, USA from a filed isolate collected from Minas Gerais, Brazil. Using a genome size estimate of 0.9-1Gb[12], we sequenced fragments (15X coverage), 3kb long inserts (10X), and 8kb long inserts (3X) with reads generated on Roche 454 instrumentation, plus 1X coverage from plasmids and 0.1X from bacterial artificial chromosome (BAC) ends[13] on the ABI3730xl. Reads were assembled using Newbler (v2.6)[44]. Additional sequencing reads were collected using Illumina instrumentation, including 200bp short inserts (45X), 3kb long inserts (15X), and 8kb long inserts (10X), and assembled de novo using SOAP de novo (v1.0.5)[45]. Finally, the Newbler assembly was merged with the SOAP assembly using GAA[46].

Redundant contigs in the merged assembly were collapsed and gaps between contigs were closed through iterative rounds of Illumina mate-pair read alignment and extension using custom scripts. We removed from the assembly all contaminating sequences, trimmed vectors (X), and ambiguous bases (N). Shorter contigs (≤200bp) were removed prior to public release.

In the creation of the linkage group AGP files, we identified all scaffolds (145Mb total) that were uniquely placed on a single linkage group (Supplementary Text 2 and Supplementary Table 1). Because of low marker density and scaffolds could not be ordered and oriented within linkage groups. The final draft assembly, (NCBI: ASM45736v1) is comprised of 331,400 scaffolds with an N50 scaffold length of 48kb and an N50 contig length of 7.3kb. The assembled coverage (Newbler) is 27.5X, and the assembly spans over 916Mb (with a coverage of 98%, 899Mb of sequence with ~17Mb of estimated gaps). The draft genome sequence of *Biomphalaria glabrata* was aligned with assemblies of *Lottia* and *Aplysia* (http://genome-test.cse.ucsc.edu/cgi-

bin/hgGateway?hgsid=389472876_vEhDXpybetKHBbwuzZFov0KE6Qhl&clade=other&org=Snail &db=0) and also deposited in the DDBJ/EMBL/GenBank database (Accession Number APKA00000000.1). It includes the genomes of an unclassified mollicute and viruses (Supplementary 5 and 6: Accession Numbers CP013128, KT728710-12). Total RNA was extracted from 12 different tissues dissected from multiple adult BB02 snails. Illumina RNAseq was used to generate tissue-specific transcriptomes for albumen gland (AG); buccal mass (BUC); central nervous system (CNS); digestive gland/hepatopancreas (DG/HP); muscular part of the headfoot (FOOT); heart including amebocyte producing organ (HAPO); kidney (KID); mantle edge (MAN); ovotestis (OVO); salivary gland (SAL); stomach (STO); terminal genitalia (TRG). The genome assembly was also deposited in Vectorbase[47], (https://www.vectorbase.org/organisms/biomphalaria-glabrata). Computational annotation using Maker2[48] yielded 14,423 predicted gene models, including 96.5% of the 458 sequences from the CEGMA core set of eukaryotic genes[49]. RNAseq data were mapped to the genome assembly and Web Apollo[50] was applied to aid manual annotation of genes of interest by contributors. Methods and results are described in the Supplementary Information.

References

1. Paraense, W. L. The schistosome vectors in the Americas. *Mem. Inst. Oswaldo Cruz* **96 Suppl**, 7-16 (2001).

2. King, C. H. Parasites and poverty: The case of schistosomiasis. *Acta Trop.* **113**, 95-104 (2010).

3. Doenhoff, M. J. *et al*. Praziquantel: Its use in control of schistosomiasis in sub-Saharan Africa and current research needs. *Parasitology* **136**, 1825-1835 (2009).

4. Melman, S. D. *et al*. Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.* **3**, e504-e504 (2009).

5. Rollinson, D. *et al*. Time to set the agenda for schistosomiasis elimination. *Acta Trop.* **128**, 423-440 (2013).

6. World Health Organization. *Accelerating work to overcome the global impact of neglected tropical diseases: A roadmap for implementation*, (Geneva: Department of Control of Neglected Tropical Diseases. Available: http://www.who.int/entity/neglected_diseases/NTD_RoadMap_2012_Fullversion.pdf, 2012).

7. Berriman, M. *et al*. The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352-358 (2009).

8. Simakov, O. *et al*. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526-531 (2013).

9. Zhang, G. *et al*. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49-54 (2012).

10. Tsai, I. J. *et al*. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**, 57-63 (2013).

11. Albertin, C. B. *et al*. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**, 220-224 (2015).

12. Gregory, T. R. Genome size estimates for two important freshwater molluscs, the zebra mussel (*Dreissena polymorpha*) and the schistosomiasis vector snail (*Biomphalaria glabrata*). *Genome* **46**, 841-844 (2003).

13. Adema, C. M. *et al*. A bacterial artificial chromosome library for *Biomphalaria glabrata*, intermediate snail host of *Schistosoma mansoni*. *Mem. Inst.Oswaldo Cruz* **101**, 167-177 (2006).

14. Cummins, S. F. & Degnan, B. Sensory sea slugs: Towards decoding the molecular toolkit required for a mollusc to smell. *Commun. Integr. Biol.* **3**, 423-426 (2010).

15. Cummins, S. F. *et al*. *Aplysia* temptin - the 'glue' in the water-borne attractin pheromone complex. *FEBS J.* **274**, 5425-5437 (2007).

16. Hertel, L. A., Bayne, C. J. & Loker, E. S. The symbiont *Capsaspora owczarzaki*, nov. gen. nov. sp., isolated from three strains of the pulmonate snail *Biomphalaria glabrata* is related to members of the Mesomycetozoea. *Int. J. Parasitol.* **32**, 1183-1191 (2002).

17. Odoemelam, E., Raghavan, N., Miller, A., Bridger, J. M. & Knight, M. Revised karyotyping and gene mapping of the *Biomphalaria glabrata* embryonic (Bge) cell line. *Int. J. Parasitol.* **39**, 675-681 (2009).

18. Feyereisen, R. Arthropod CYPomes illustrate the tempo and mode in P450 evolution. *Biochim Biophys Acta* **1814**, 19-28 (2011).

19. Janeway, C. A. & Medzhitov, R. Innate immune recognition. *Immunol.* **20**, 197-216 (2002).

20. Buckley, K. M. & Rast, J. P. Diversity of animal immune receptors and the origins of recognition complexity in the deuterostomes. *Dev. Comp. Immunol.* **49**, 179-189 (2015).

21. Zhang, S.-M., Adema, C. M., Kepler, T. B. & Loker, E. S. Diversification of Ig superfamily genes in an invertebrate. *Science* **305**, 251-254 (2004).

22. Dheilly, N. M. *et al*. A family of variable immunoglobulin and lectin domain containing molecules in the snail *Biomphalaria glabrata*. *Dev. Comp. Immunol.* **48**, 234-243 (2015).

23. Offord, V. & Werling, D. LRRfinder2.0: A webserver for the prediction of leucine-rich repeats. *Innate Immun.* **19**, 398-402 (2013).

24. Falschlehner, C. & Boutros, M. Innate immunity: Regulation of caspases by IAP-dependent ubiquitylation. *EMBO J.* **31**, 2750-2752 (2012).

25. Vandenabeele, P. & Bertrand, M. J. M. The role of the IAP E3 ubiquitin ligases in regulating pattern-recognition receptor signalling. *Nat. Rev. Immunol.* **12**, 833-844 (2012).

26. Mitta, G., Vandenbulcke, F. & Roch, P. Original involvement of antimicrobial peptides in mussel innate immunity. *FEBS L* **486**, 185-190 (2000).

27. Fneich, S. *et al*. 5-methyl-cytosine and 5-hydroxy-methyl-cytosine in the genome of *Biomphalaria glabrata*, a snail intermediate host of *Schistosoma mansoni*. *Parasit. Vectors* **6**, 167-167 (2013).

28. Geyer, K. K. *et al*. Cytosine methylation regulates oviposition in the pathogenic blood fluke *Schistosoma mansoni*. *Nat. Commun.* **2**, 424-424 (2011).

29. Knight, M. *et al*. Polyethyleneimine (PEI) mediated siRNA gene silencing in the *Schistosoma mansoni* snail host, *Biomphalaria glabrata*. *PLoS Negl. Trop. Dis.* **5**, e1212-e1212 (2011).

30. Rotenberg, L. *et al*. Oviposition rhythms in *Biomphalaria* spp. (Mollusca: Gastropoda). 1. Studies in *B. glabrata*, *B. straminea* and *B. tenagophila* under outdoor conditions. *Prog. Clin. Biol. Res.* **341B**, 763-761 (1990).

31. Koene, J. M. *et al*. Male accessory gland protein reduces egg laying in a simultaneous hermaphrodite. *PLoS ONE* **5**, e10117-e10117 (2010).

32. Nakadera, Y. *et al*. Receipt of seminal fluid proteins causes reduction of male investment in a simultaneous hermaphrodite. *Curr. Biol.* **24**, 859-862 (2014).

33. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109-D114 (2012).

34. Zahoor, Z., Davies, A. J., Kirk, R. S., Rollinson, D. & Walker, A. J. Nitric oxide production by *Biomphalaria glabrata* haemocytes: effects of *Schistosoma mansoni* ESPs and regulation through the extracellular signal-regulated kinase pathway. *Parasit. Vectors* **2**, 18-18 (2009).

35. Holland, N. D., Venkatesh, T. V., Holland, L. Z., Jacobs, D. K. & Bodmer, R. AmphiNk2-tin, an amphioxus homeobox gene expressed in myocardial progenitors: Insights into evolution of the vertebrate heart. *Dev. Biol.* **255**, 128-137 (2003).

36. Fyrberg, E. A., Mahaffey, J. W., Bond, B. J. & Davidson, N. Transcripts of the six *Drosophila* actin genes accumulate in a stage- and tissue-specific manner. *Cell* **33**, 115-123 (1983).

37. Fryberg, E. A., Bond, B. J., Hershey, N. D., Mixter, K. S. & Davidson, N. The actin genes of *Drosophila*: Protein coding regions are highly conserved but intron positions are not. *Cell* **24**, 107-116 (1981).

38. Mounier, N., Gouy, M., Mouchiroud, D. & Prudhomme, J. C. Insect muscle actins differ distinctly from invertebrate and vertebrate cytoplasmic actins. *J. Mol. Evol.* **34**, 406-415 (1992).

39. Panchin, Y. & Moroz, L. L. Molluscan mobile elements similar to the vertebrate Recombination-Activating Genes. *Biochem. Biophys. Res. Commun.* **369**, 818-823 (2008).

40 Raghavan, N., *et al. Nimbus (BgI):* An active non-LTR retrotransposon of the *Schistosoma mansoni* snail host *Biomphalaria glabrata*. *Int. J. Parasitol*. **37**,1307-1318 (2007).

41. Pace, J. K., Gilbert, C., Clark, M.S. & Feschotte, C. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proc. Nat. Acad. Sci. USA* **105**, 17023-17028 (2008).

42. DeJong, R. J. *et al.* Evolutionary relationships and biogeography of *Biomphalaria* (Gastropoda: Planorbidae) with implications regarding its role as host of the human bloodfluke, *Schistosoma mansoni*. Mol. Biol. Evol.**18**, 2225-2239 (2001).

43. Campbell, G. *et al.* Molecular evidence supports an african affinity of the Neotropical freshwater gastropod, *Biomphalaria glabrata*, Say 1818, an intermediate host for *Schistosoma mansoni*. *Proc. Biol. Sci.* **267**, 2351-2358 (2000).

44. Margulies, M. *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).

45. Li, R. *et al*. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265-272 (2010).

514 46. Yao, G. *et al*. Graph accordance of next-generation sequence assemblies. *Bioinformatics* **28**, 13-
515 16 (2012).

516 47. Giraldo-Calderón, G. I. *et al*. VectorBase: An updated bioinformatics resource for invertebrate
517 vectors and other organisms related with human diseases. *Nucleic Acids Res.* **43**, D707-13
518 (2015).

519 48. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management
520 tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).

521 49. Parra, G., Bradnam, K. & Korf, I. CEGMA: A pipeline to accurately annotate core genes in
522 eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).

523 50. Lee, E. *et al*. Web Apollo: A web-based genomic annotation editing platform. *Genome Biol.* **14**,
524 R93-R93 (2013).

525 **Supplementary Information** is linked to the online version of the paper at www.nature.com/nature

526 **Acknowledgments**

527 We thank S. Newfeld for discussion of actin evolution. Sequence characterization of the
528 *Biomphalaria glabrata* genome was funded by National Institutes of Health (NIH) grant HG003079
529 to R.K.W. and McDonnell Genome Institute, Washington University School of Medicine. Further
530 work was supported by FAPEMIG (RED-00014-14) and CNPq (304138/2014-2, 309312/2012-4) to
531 G.O.. K.F.H. and M.S. acknowledge the support of the UK BBSRC (BB/K005448/1) K.F.H. and
532 M.S. acknowledge the support of BBSRC (BB/K005448/1) and CNPq (503275/2011-5) to R.L.C.
533 C.B. acknowledges the support of NIH grant AI016137. J.M.K. acknowledges support from the
534 Research Council for Earth and Life Sciences (ALW; 819.01.007) with financial aid from the
535 Netherlands Organization for Scientific Research (NWO). S.E. acknowledges NIAID contract

536 HHSN272201400029C. C.M.A. was supported by NIH grant P30GM110907from the National
537 Institute of General Medical Sciences (NIGMS). R.M.C. acknowledges NIH grant GM061738 and
538 support from the American Heart Association, Southwest Affiliate (14GRNT20490250). D.T. was
539 supported by NIH grant R25 GM075149. M.K. acknowledges NIH-NIAID grant R01-AI063480.
540 C.F. acknowledges NIH grant R01-GM077582. D.J.J. acknowledges D.F.G grant JA2108/1-2.
541 E.S.L. acknowledges NIH grant P30GM110907 and ROI AI101438. BG acknowledges ANR JCJC
542 INVIMORY (ANR-13-JSV7-0009). K.M.B and J.P.R acknowledge Natural Sciences and
543 Engineering Research Council (NSERC 312221) and the Canadian Health Institutes for Research
544 (CIHR MOP74667). C.S.J, L.R.N, S.J, E.J.R, S.K and A.E.L acknowledge funding from NC3R's
545 Grant (ref GO900802/1).
546

547 **Author Contributions**

548 C.M.A., E.S.L., M.K., N.R. conceived the study, scientific objectives. C.M.A. led the project and
549 manuscript preparation with input from steering committee members M.K., C.S.J., G.O., P.M.,
550 L.W.H., A.S., E.S.L., assisted by S.E. O.C. provided field collected snails. C.M.A. and E.S.L
551 cultured snails and provided materials. P.M., L.W.H, S.C., L.F., W.C.W, R.K.W., V.M., C.T
552 developed the sequencing strategy, managed the project, conducted assembly and evaluation. B.R.
553 performed genome alignments. M.C., D.H., S.E., G.G-C. and D.L. perfomed the genebuild, managed
554 metadata, performed genome annotation and data analysis, and facilitated Community Annotation
555 with M.C.M-T. H.D. A-G., M.Y., E.V.V., M.K. and J.M.B. performed Karyotyping and FISH
556 analysis. J.A.T and M.B. performed linkage mapping. G.O., F.P., F.R., J.A., I.R., Y.C., S.G., F.S.
557 and L.S. performed computational analyses of genomic, proteomic, and transcriptomic data, SNP
558 content, secretome, metabolic pathways and annotation of eukaryote protein kinases (ePKs). S.F.C.,
559 L.Y., D.L., M.Z. and D.McM conducted pheroreception studies. M.T.S., K.K.G., U.N. and K.F.H

conducted bacterial symbiont analysis. S.L., S-M.Z., E.S.L. and B.C.B. performed virus analyses. M.K., P.F., W.I. and N.R. performed annotation of HSP. T.P.Y., X-J.W., U.B-W. and N.D. conducted proteogenomic studies of parasite-reactive snail host proteins and data analysis. R.F., A. E.L. and C.S.J. performed annotation of CYP. J.H. performed annotation of NFkB. K.M.B. and J.P.R. performed annotation of conserved immune factors. C.M.A. and J.J.P performed annotation of FREPs. M.C. and C.M. performed annotation of complement. D.D. performed annotation of apoptosis. B.G. and C.J.B. performed annotation of REDOX balance. O.L.B., D.D., R.G., Ch.C. and G.M. performed annotation of antibacterial defenses. L.d.S. and A.T.P performed search for antibacterial defense genes. P.C.H., M.A.G. and E.A.P. performed annotation of unknown novel sequences. K.K.G., I.W.C., U.N., K.F.H., Ce.C., T.Q. and C.G. performed annotation and analysis of epigenetic sequences. E.H.B., L.R. doA., M.deS.G., R.L.C, and W.deJ.J. performed annotation of miRNA (Brazil), K.K.B., R.P. and K.B.S. performed annotation of miRNA (Canada). M.G. performed annotation of periodicity. S.F.C., B.R., T.W., A.E.L and S.K.conducted neuropeptide studies and data analysis. A.E.L., R.F., S.K., E.J.R., S.J., D.R., C.S.J. and L.R.N.performed annotation neuroendocrinology (CYP). J.M.K., B.R. and S.F.C.performed annotation of ovipostatin. A.B.K and L.L.M performed tissue location of transcripts analysis. A.J.W. and S.P.L. performed annotation of phosphatases. T.L.L., K.M.R., M.Mi., and R.C. performed annotation of actins and annotation of cardiac transcriptional program with D.L.T. D.J.J., B.K. and M.Me. performed annotation of biomineralization genes. J.C., A.K. and C.F. performed annotation of DNA transposons, global analysis of transposable element landscape, and horizontal transfer events. A.S and C.B. performed repeat/TE analysis. E.S.L, S.M.Z., G.M.M and SKB conducted comparative *B. pfeifferi* transcriptome studies and data analysis. C.M.A, P.M., M.L.M did most of the writing with contributions from all authors.

Author Information

The *Biomphalaria glabrata* genome project has been deposited at DDBJ/EMBL/ GenBank under the accession number APKA00000000.1. All short-read data have been deposited into the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) as follows: Illumina HiSeq 2000 WGS reads:SRX648260-71; 454 GS FLX WGS: SRX005828, SRX008161,2; 454 GS FLX RNA reads: SRX014813, SRX014894-7

Genome, transcriptome and predicted proteome data are also available at Vectorbase (ref. 47).

The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.M.A. (coenadem@unm.edu).

List of Supplementary Files

PDF file

1. Supplementary information (SIZE)
   This file contains Supplementary Text and Data sections 1-32 (see Contents list for details), Supplementary Figures 1-87

Zip files

1. Supplementary Figure 8 (JPG, 8500 KB).

2 Supplementary Tables 1-56 (Excel 1919048 KB).