BMC Genomics

RESEARCH

Open Access

The genome and transcriptome of the snail *Biomphalaria sudanica s.l.*: immune gene diversification and highly polymorphic genomic regions in an important African vector of *Schistosoma mansoni*

Tom Pennance[1*], Javier Calvelo[2], Jacob A. Tennessen[3], Ryan Burd[1], Jared Cayton[1], Stephanie R. Bollmann[4], Michael S. Blouin[4], Johannie M. Spaan[1], Federico G. Hoffmann[5,6], George Ogara[7], Fredrick Rawago[7], Kennedy Andiego[7], Boaz Mulonga[7], Meredith Odhiambo[7], Eric S. Loker[8], Martina R. Laidemitt[8], Lijun Lu[8], Andrés Iriarte[2], Maurice R. Odiere[7] and Michelle L. Steinauer[1*]

Abstract

Background Control and elimination of schistosomiasis is an arduous task, with current strategies proving inadequate to break transmission. Exploration of genetic approaches to interrupt *Schistosoma mansoni* transmission, the causative agent for human intestinal schistosomiasis in sub-Saharan Africa and South America, has led to genomic research of the snail vector hosts of the genus *Biomphalaria*. Few complete genomic resources exist, with African *Biomphalaria* species being particularly underrepresented despite this being where the majority of *S. mansoni* infections occur. Here we generate and annotate the first genome assembly of *Biomphalaria sudanica* sensu lato, a species responsible for *S. mansoni* transmission in lake and marsh habitats of the African Rift Valley. Supported by whole-genome diversity data among five inbred lines, we describe orthologs of immune-relevant gene regions in the South American vector *B. glabrata* and present a bioinformatic pipeline to identify candidate novel pathogen recognition receptors (PRRs).

Results De novo genome and transcriptome assembly of inbred *B. sudanica* originating from the shoreline of Lake Victoria (Kisumu, Kenya) resulted in a haploid genome size of ~944.2 Mb (6,728 fragments, N50 = 1.067 Mb), comprising 23,598 genes (BUSCO = 93.6% complete). The *B. sudanica* genome contains orthologues to all described immune genes/regions tied to protection against *S. mansoni* in *B. glabrata*, including the polymorphic transmembrane clusters (*PTC1* and *PTC2*), *RADres*, and other loci. The *B. sudanica PTC2* candidate immune genomic region contained many PRR-like genes across a much wider genomic region than has been shown in *B. glabrata*, as well as a large inversion between species. High levels of intra-species nucleotide diversity were seen in *PTC2*, as well as in regions linked to *PTC1* and *RADres* orthologues. Immune related and putative PRR gene families were significantly over-represented

*Correspondence:
Tom Pennance
tpennance@westernu.edu
Michelle L. Steinauer
msteinauer@westernu.edu
Full list of author information is available at the end of the article

in the sub-set of *B. sudanica* genes determined as hyperdiverse, including high extracellular diversity in transmembrane genes, which could be under pathogen-mediated balancing selection. However, no overall expansion in immunity related genes was seen in African compared to South American lineages.

Conclusions The *B. sudanica* genome and analyses presented here will facilitate future research in vector immune defense mechanisms against pathogens. This genomic/transcriptomic resource provides necessary data for the future development of molecular snail vector control/surveillance tools, facilitating schistosome transmission interruption mechanisms in Africa.

Keywords *Biomphalaria sudanica*, *Biomphalaria choanomphala*, Schistosomiasis, Snail vector, De novo genome assembly, Polymorphism, Immunogenetics, Gene family evolution, Balancing selection, Pathogen recognition

Background

Freshwater pulmonate snails of the genus *Biomphalaria* are intermediate hosts for a diversity of trematode parasites but are most notorious for their role in the transmission of *Schistosoma mansoni*, to humans. Although several species of *Schistosoma* can cause schistosomiasis, including those that are transmitted by *Bulinus* and *Oncomelania* snails, *S. mansoni* is among the most prevalent in human infections. Schistosomiasis is a chronic and inflammatory disease with devastating impacts to human health, which likely are underestimated due to schistosomiasis associated morbidities and mortalities being attributed to non-communicable diseases for which the symptoms are similar [1]. Despite the widely recognized toll of schistosomiasis on public health there are few effective and implementable options for controlling transmission of the parasite [2].

The need for novel interventions to interrupt *S. mansoni* transmission has spurred on investigatory genomic research of the *Biomphalaria* vector hosts [3]. Genomic resources for *Biomphalaria* will facilitate the discovery of genetic resistance mechanisms to schistosome infection, which could be manipulated to block transmission to snails, and thus humans [4]. Currently, three *Biomphalaria* species represent the extent of genome resources, two South American species: *Biomphalaria glabrata* [5, 6]; *Biomphalaria straminea* [7]; and one sub-Saharan African species *Biomphalaria pfeifferi* [8]. African *Biomphalaria* species are therefore underrepresented in terms of complete genomic information, even though African species contribute to the vast majority of global *S. mansoni* transmission since approximately 90% of human infections occur in Africa [9]. Thus, genome-wide analysis of African *Biomphalaria* species would facilitate the development of genetic based vector control in areas where it is highly relevant to transmission. As more *Biomphalaria* genomes become available, evolutionary analysis of immunity of these major vectors will be possible, whilst putting it into the context of species divergence across the African continent after being introduced from South America sometime between 1.8 and 5 MYA [8, 10–12].

The African species *Biomphalaria sudanica* was originally described in 1870 from Djur and Rek tributaries of the White Nile in the Bahr el Ghazal region of Southern Sudan [13]. It is distributed throughout the Nile Basin in marsh and lacustrine habitats in Uganda, Kenya, Sudan, Tanzania and Ethiopia [14–19]. *Biomphalaria sudanica* distribution in East Africa corresponds to the geographic region where the genetic diversity of *S. mansoni* is the greatest [20, 21]. Most research regarding *B. sudanica* has been focused on populations from Lake Victoria, where *S. mansoni* remains highly endemic even following repeated and widespread mass drug administration of schistosomiasis preventative chemotherapy [22–24]. Although *B. sudanica* inhabits the marshy fringes and nearshore shallow waters of Lake Victoria where human-freshwater contact takes place [25], another snail vector of *S. mansoni*, described as *Biomphalaria choanomphala* [26], occurs in deep water habitats of Lake Victoria [14, 27, 28]. The taxonomic status of these two species of Lake Victoria snails is in question, as DNA divergence of mitochondrial genes (and the few nuclear genes that have been sequenced) between these species suggests they may represent ecomorphs of a single species [29–31]. However, distinct morphologies, habitats, and schistosome susceptibility profiles [32] make the distinction of these two forms critical in the context of a genome report. Thus, we follow conventional use of the species name, or *Biomphalaria sudanica* sensu lato. Genomic data of these species will facilitate future population genomic analyses aimed at better understanding the relationship between these taxa. Snail based schistosomiasis control cannot even be imagined without understanding these basics.

Experimental infections have shown that *B. sudanica* displays the greatest natural resistance to schistosome infection relative to *B. choanomphala*, and another closely related species, *B. pfeifferi* [28, 32], and thus offers an excellent target for the discovery of immune relevant genes in an African *Biomphalaria* species. While some immune genes can be characterized by conserved domains as a result of positive selection [33], others such as those involved in host–pathogen interactions are rapidly evolving under balancing selection due to the simultaneous arms races occurring between the host and its pathogens [34]. Indeed, immune loci are among the most diverse in many genomes, including the classic example of the vertebrate major histocompatibility complex (MHC) [35, 36]; human innate immunity genes [37]; R genes in plants [38, 39] and more recently shown in immune genes of invertebrate organisms such as *Caenorhabditis elegans* [40]. Virtually nothing is known regarding the *B. sudanica* immune defense except for what can be inferred from orthologous gene searching strategies related to experimental work with the South American congener, *B. glabrata* [3]. The identification of *B. glabrata* loci associated with resistance to schistosomes is an active research field with approaches such as Quantitative Trait Locus (QTL) analysis providing valuable new insights [6].

In this paper, we present the first description of the genome of *B. sudanica*. Our novel annotated genome of *B. sudanica* 111 (Bs111), an inbred line maintained at Western University of Health Sciences which originates from the Kisumu region (Kenya) of Lake Victoria, comprises PacBio HiFi long-read DNA and RNA sequence data, as well as Illumina short-read RNA sequence data. This combination of genomic and transcriptomic data provides a confident annotation of functional gene boundaries, exon–intron structure, and isoforms for the representative genome of this species. Here, we focus on identifying and describing gene regions orthologous to those involved in immunity of the South American vector *B. glabrata* to *S. mansoni*, such as the Polymorphic Transmembrane Cluster (*PTC1* and *PTC2*) genomic regions [41, 42] and fibrinogen-related proteins (FREPs) [43, 44]. We also used a new analysis pipeline to find novel pathogen recognition receptors (PRRs). Following the hypothesis that PRRs are under balancing selection as with other highly polymorphic immune loci, we searched in the most hyperdiverse genome regions through the genomic comparison of five genetic lines of *B. sudanica* for signatures of candidate PRRs. This allows us to identify immune related genes that do not maintain detectable sequence similarity with known gene families and are not only specific to schistosome immunity. The description of key features of the *B. sudanica* genome provides multiple exciting avenues for future research into this important vector of *S. mansoni*.

Results

Biomphalaria sudanica genome assembly and nuclear genome annotation

The PacBio assembled *B. sudanica* (Bs111) haploid genome size is ~ 944.2 Mb, comprising 6,728 contigs and scaffolds with an N50 of 1.067 Mb, and a mean sequencing coverage of ~ 23x (Supplementary Table 1) (NCBI BioProject: PRJNA1041389). The estimated size of the *B. sudanica* genome is somewhat larger than those of *B. pfeifferi* (~ 771.8 Mb [8]) and *B. glabrata* (iM line: ~ 871.0 Mb [6]), but smaller than that of *B. straminea* (~ 1,004.7 Mb [7]).

PacBio and Illumina RNA sequence data were obtained from Bs111 snails to aid in annotation of the assembled *B. sudanica* genome. Pooled RNA was processed following a standard PacBio IsoSeq procedure, which yielded 335.0 Gbases (N50 of ~ 115.5 Kbases) of long-read transcript data. Of the original 6,945,781 circular consensus sequencing (ccs) reads, 3,798,283 (54.68%) passed the Q20 quality threshold determined in the longQC software [45], of which 1,708,667 (44.99%) were identified as potentially complete isoforms (i.e. they bear both 5' and 3' adapter sequences) using Lima (github.com/PacificBiosciences/barcoding) (Supplementary Table 1). To supplement the RNA transcript long-read data, Illumina paired-end 150 short-read sequence data yielded 45,125,478 paired reads of which 44,008,723 (97.53%) passed the trimming process conducted in Trimommatic [46] (Supplementary Table 1). Overall mapping rate of long-read transcripts using minimap2 [47, 48] and Illumina RNA short-reads using STAR [49] to the assembled genome was close to 100%. Transcript characterization using StringTie2 v2.2.1 [50] identified 25,847 individual genes, of which 23,598 in TransDecoder.Predict v5.5 [51] had an assigned open reading frame (ORF) (Supplementary File 1 and Supplementary File 2). InterProScan v5.56–89.0 [52] identified at least one protein domain signature on 19,945 genes (~ 84%) (Supplementary File 3). BUSCO [53] completeness analysis shows that the latter set (23,598 genes) represents a close to complete genome annotation relative to mollusca_odb10 (of 5295 BUSCO groups, 93.6% complete, 1.4% fragments and 5.0% missing).

The *B. sudanica* genome contains orthologues to at least 18 candidate immune loci of *B. glabrata* that function in protection against *S. mansoni* (Supplementary Table 2). These include key genes/gene clusters/loci coding for: FREP2 and FREP3; *B. glabrata toll-like receptor* (*BgTLR*); *B. glabrata* phagocyte oxidase (*Phox*); Guadeloupe Resistance Complex 1 genes (*GRC*) – referred to from here as *PTC1*; *PTC2*; *RADres*; heat shock protein 90 (*HSP*90); *Granulin* (*GRN*); *B. glabrata* thioester-containing protein (*BgTEP*); Catalase (*cat*);

Biomphalysin; *Glabralysin*; OPM-04 (Knight marker); superoxide dismutase 1 (*sod1*); *Peroxiredoxin 4* (*prx4*), qRS-5.1 and qRS-2.1 (Supplementary Table 2).

A total of 919 tRNA and 107 rRNA genes were predicted in the *B. sudanica* nuclear genome (Supplementary Table 3, Supplementary Table 4 and Supplementary File 1). The number of tRNA genes identified in the *B. sudanica* genome is comparably higher than that observed in the genomes of *B. pfeifferi* ($n=514$) [8] and *B. glabrata* ($n=510$) [6]. As is the case for *B. pfeifferi*, one selenocysteinyl tRNA (tRNA-SeC) is present in the genome of *B. sudanica* (Supplementary Table 3 and Supplementary File 1), meaning this species is capable of synthesizing selenocysteine containing polypeptides, or selenoproteins [54, 55]. The tRNA-Sec gene has not been identified in *B. glabrata* [8]. Overall, fewer rRNA genes were predicted in this genome assembly of *B. sudanica* compared to that of *B. pfeifferi* (107 and 757, respectively), which could be a result of some of the highly repetitive rRNA genes being misassembled in the more fragmented *B. sudanica* genome assembly.

As with other *Biomphalaria*, repetitive elements composed a large proportion of the *B. sudanica* genome (40.3%) (Supplementary Fig. 1). About 87% of protein coding genes overlap with at least one annotated repeated element in their gene model. The overlap is primarily within introns and untranslated regions (UTRs); however, 1576 genes have repeat elements within their predicted coding sequence (CDS) (Supplementary Table 5). The repeat regions largely comprise unknown repeat elements, in addition to an abundance of unclassified long interspersed nuclear elements (LINE), LINE/retrotransposable element Bovine B (RTE-BovB) and unclassified DNA transposons (Supplementary Fig. 1, Supplementary Table 6).

Mitochondrial genome annotation and trimming processes

The mitochondrial genome comprises the same gene content (13 genes, 2 rRNA, 22 tRNA), and synteny as its congeners [56] (Supplementary Table 7). The mitochondrial genomes of gastropods are unique in the fact that they have acquired transcriptional processes during their evolutionary history that are not often observed in vertebrates [57]. Therefore, while an annotated mitochondrial genome of *B. sudanica* has been previously published [56], our novel contribution here is to validate gene boundaries and explore the transcription processes using long-read transcriptomic data.

Raw PacBio IsoSeq reads (Q20 or higher) were mapped to the mitochondrial genome with minimap2 [47, 48]. To explore the trimming process of the primary mitochondrial transcript, the intermediary pre-mRNA mitochondrial transcripts, i.e., that cover multiple features within the mitochondrial genome, were recovered and counted (Supplementary Fig. 2). In line with the tRNA punctuation model [58], we established that pre-mRNAs of the *B. sudanica* mitochondrial genome are trimmed at the tRNA genes (with the potential exception of *atp6*/*atp8*), with the minus strand being processed 3'-to-5' while the plus strand shows an odd mixture of both directions.

In *B. sudanica*, there are three mitochondrial gene re-arrangements in comparison to a typical animal mitochondrial genome that affect the transcription processes [59]. First, it is typical for *atp6* and *atp8* to be adjacent and remain together in the mature mRNA; however, in *Biomphalaria*, including *B. sudanica*, these genes are separated by a tRNA gene (trnN, see Fragment 5; Supplementary Fig. 2 and [56]). In the case of *atp6*/*atp8*, the presence of trnN suggest both are translated to proteins as monocistronic transcripts. However, only a few monocistronic transcripts of each *atp6* ($n=16$) and *atp8* ($n=1$) were identified from the raw RNA reads, in contrast with the far more abundant untrimmed intermediaries (Supplementary Fig. 2). Considering that the ancestral condition is the translation of both proteins from a bicistronic mRNA, it is a tempting hypothesis that this is still the case in *Biomphalaria* despite the extra trimming points.

Second, it appears that, *nad4l* is either a non-functional pseudogene in the *B. sudanica* mitochondrial genome or is only expressed in very low levels. The adjacency and bicistronic transcription of *nad4*/*nad4l*, is well conserved in invertebrate and vertebrate mitochondrial genomes, yet in many molluscan lineages, including *B. sudanica*, these genes are nonadjacent (Supplementary Fig. 2) [57]. Additionally, the transcript data demonstrates that *nad4* terminates on an abbreviated stop codon (T–) as was experimentally supported in its South American sister species *B. glabrata* [60], and was abundantly represented in the transcriptome ($n=83$ monocistronic transcripts from Fragment 3; Supplementary Fig. 2). On the other hand, *nad4l* transcripts were rare, represented by only four intermediary reads that were attached to transcripts for neighboring gene Cytochrome B (*cob*) (Fragment 1, Supplementary Fig. 2).

Third, *nad6*/*nad5*/*nad1* are found adjacent to one another and seem to be translated as a polycistronic mRNA (Fragment 1, Supplementary Fig. 2), since their gene boundaries predicted by MITOS2 [61] overlap; there are no clear cuts in the read coverage; and pure monocistronic reads were only recovered in small numbers for *nad5* ($n=16$) and *nad1* ($n=2$) and none were found for *nad6*. Given the lack of a clear trimming point in this region, the observed monocistronic and bicistronic reads such as *nad6*/*nad5* ($n=80$) and *nad5*/*nad1*

($n=85$), or *nad5*/*nad1*/OH_1 ($n=1929$) are likely the product of partial RNA degradation that were spuriously retained in the final dataset. With this considered, our data suggests that the three genes are translated into proteins from the same mRNA molecule.

Lastly, MITOS2 initially reported the gene rrnL as split in two subunits (rrnL-a coordinates 12,182–12,287 and rrnL-b coordinates 12,483–13,095). As this is likely an artifact where partial hits to rrnL were identified as separate genes, we defined our predicted rrnL as the region between the tRNAs that bookend the gene, trnV and trnL1. This corresponds to the last pre-rRNA recovered for this locus and is consistent with the rrnS gene that is flanked by trnE and trnM.

Location signals and transmembrane domains

Location signals for exportation, i.e. the signal peptides or mitochondrial targeting peptides, were identified with SignalP v6.0 [62] and TargetP v2.0 [63] in 3,339 genes (5,016 isoforms) and 69 genes (111 isoforms), respectively (Supplementary Table 8). Transmembrane domains were predicted in 4,922 genes (8,728 isoforms), and the location signals analysis suggests that 835 of these were firmly anchored to the plasma membrane, organellar membranes or vesicles, and seven to the mitochondria (Supplementary File 4).

An additional 82 proteins (146 isoforms) that showed no location signals in SignalP and TargetP were identified by SecretomeP [64] as potentially secreted through an alternative pathway. However, since 34 of these had at least one transmembrane domain, we suspect many of these are false negatives for either the signal or the mitochondrial targeting peptide. Furthermore four genes, which do not contain transmembrane domains and are potentially secreted, have signal peptides predicted in some but not all their isoforms (genes BSUD.7093, BSUD.10729, BSUD.12693 and BSUD.24440). This might represent cases of functional isoforms generated through alternative splicing that have different locations in the cell, as observed in other species [65].

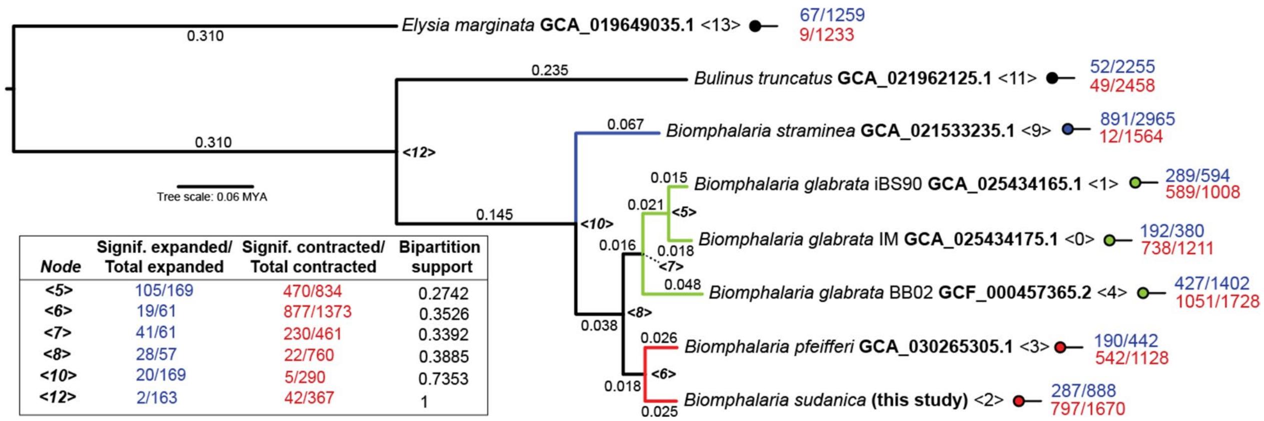
Two identical secreted proteins are worth pointing out: BSUD.4529 (contig 217) and BSUD.14556 (contig 559). These were identified as orthologs to the precursor protein of peptide P12 in *B. glabrata* (BGLB027975 [5]), which has been shown to trigger behavior modifications in *S. mansoni* miracidia, and thus is potentially an attractant [66, 67]. Compared to the *B. glabrata* P12, the orthologous 13 aa region in *B. sudanica*, and *B. pfeifferi* (protein NCBI accession: KAK0045317 [8]), contains a non-synonymous change, changing the 5th amino acid from Glycine to Valine (DITSVLDPEVADD), whereas the same 13 aa orthologous region in *B. straminea* (protein 046859-T1 [7]) contains 6 non-synonymous amino acid changes (EVASVLDPDVADN).

Gene family evolution in *Biomphalaria* species

The evolutionary dynamics of *Biomphalaria* protein families among *B. sudanica, B. glabrata, B. pfeifferi, B. straminea*, and two species used as outgroups, *Bulinus truncatus* and *Elysia marginata*, were estimated by first identifying orthology in Phylogenetic Hierarchical Orthogroups (HOG) with Orthofinder v2.5.4 [68]. Orthofinder clustered ~ 87% of identified genes into 29,664 orthogroups with representatives in at least two species. Orthogroups were distributed across 31,723 HOGs (Supplementary Table 9). HOGs were taken as an approximate estimation of protein families found on these genomes, and significant changes in their size estimated with CAFE 5 [69]. The phylogeny utilized for the estimation was the species tree generated by Orthofinder for the HOG definition, including 3,541 single copy orthogroups with representatives in all the genomes, and the root calibrated to be 20 Million Years based on the appearance of *Bulinus* in the fossil record 19–20 MYA [70] (Fig. 1). Gene Ontology (GO) terms were assigned to each HOG with eggNOG-mapper v2.0 [71, 72] assuming that a GO assigned to one member applied to the whole HOG (Supplementary Table 10), significantly enriched GO terms among the expanded and contracted HOGs on each node were identified with topGO [73] (Supplementary Table 11) and then summarized with REVIGO [74] to allow further interpretation (Supplementary Table 12 and Supplementary File 5).

Significant gene family expansions/contractions were identified throughout the analyzed phylogeny. Lineage-specific trends were identified within the *Biomphalaria* genus, in comparison to outgroups, between South American and African *Biomphalaria* species, within the African lineages (*B. sudanica* and *B. pfeifferi*), and between lines of *B. glabrata* (Fig. 1, Supplementary Table 12 and Supplementary File 5). Expansions in immune related gene families were detected in the common ancestor of all *Biomphalaria* species relative to the previous node representing the split from *Bulinus truncatus*, including those involved in acute inflammatory response, glomerular filtration and regulation of cell adhesion (see node 10, Fig. 1 and Supplementary File 5 and Supplementary Table 12).

In the branch leading to the common ancestor of the African species, *B. sudanica* and *B. pfeifferi*, we identified substantially more gene family contractions than expansions (see node 6, Fig. 1). Within this lineage we found the expansion or contraction of gene families associated with the circulatory system (e.g. GO:0001525, GO:0001987, GO:0007512 and GO:0061337), compound


0.310
Elysia marginata **GCA_019649035.1** <13>
67/1259
9/1233
0.235
Bulinus truncatus **GCA_021962125.1** <11>
52/2255
49/2458
0.067
Biomphalaria straminea **GCA_021533235.1** <9>
891/2965
12/1564
0.310
<12>
0.015
Biomphalaria glabrata iBS90 **GCA_025434165.1** <1>
289/594
589/1008
0.021
<5>
Tree scale: 0.06 MYA
0.145
<10>
0.016
0.018
<7>
Biomphalaria glabrata IM **GCA_025434175.1** <0>
192/380
738/1211
0.048
Biomphalaria glabrata BB02 **GCF_000457365.2** <4>
427/1402
1051/1728
<8>
0.038
0.026
Biomphalaria pfeifferi **GCA_030265305.1** <3>
190/442
542/1128
<6>
0.018
0.025
Biomphalaria sudanica **(this study)** <2>
287/888
797/1670

Node	Signif. expanded/ Total expanded	Signif. contracted/ Total contracted	Bipartition support
<5>	105/169	470/834	0.2742
<6>	19/61	877/1373	0.3526
<7>	41/61	230/461	0.3392
<8>	28/57	22/760	0.3885
<10>	20/169	5/290	0.7353
<12>	2/163	42/367	1

Fig. 1 Species tree generated in Orthofinder using the Species Tree of All Genes (STAG) algorithm [68, 75]. Root is time calibrated to be 20 Million Years Ago based on the appearance of *Bulinus* in the fossil record [70]. Node support values represent the bipartition proportions in each of the individual species tree estimates. Branch lengths represent the average number of substitutions per site across all the individual trees inferred from each gene family. The number of (significant/total) gene families expanded (blue) and contracted (red) in the ancestral populations of the *Biomphalaria* species, and outgroups *Elysia marginata* and *Bulinus truncatus* as determined in CAFE 5 [69] are shown for each internal and terminal node (< 0 > to < 13 >)

transport (e.g. GO:0098660, GO:0006811, GO:0035459 and GO:0006855), protein maturation (e.g. GO:0006508, GO:0036211 and GO:0016579), metabolic pathways and regulation (e.g. GO:1,901,293, GO:0006210, GO:0006212 and GO:0006164), development and growth (e.g. GO:0050767, GO:0001558 and GO:0036342), and protection towards environmental stressors including chemical stressors (e.g. GO:0009620, GO:0009636, GO:0010033 and GO:0010035) (Supplementary Table 12 for full details). Several genes associated with chemotaxis (e.g. GO:0006935, GO:0009410 and GO:0071310), which are often associated with immunity, were contracted in the African *Biomphalaria* (Supplementary Table 12). Regarding the *B. sudanica* lineage, gene families belonging to 287 GO terms were significantly expanded, including genes associated with the regulation of hippo signaling (see Node 2, Group 4 in Supplementary File 5 and Supplementary Table 12) that account for multiple immune response processes. However, many GO terms associated with the immune response were also identified in contracted gene families in *B. sudanica*, including defense responses and regulation of immune system processes suggesting that a complex evolutionary trend took place in this species.

Identification and phylogeny of variable immunoglobulin and lectin domain-containing molecules (VIgLs): FREPs and CREPs

Genes putatively belonging to FREP, C-type lectin-related protein (CREP) or galectin-related protein (GREP) families were predicted based on the presence of a secretion signal and in conjunction with either a fibrinogen (FBD), C-type lectin, or Galectin domain as predicted by InterProScan v5.56–89.0 [52] (Supplementary File 3) and hmmsearch v3.3.2 [76] searches with custom IgSF profiles (as described in [44]).

Following this selection pipeline, 246 genes distributed among 140 HOGs were determined to have key domains that made them FREP, CREP or GREP candidates (Supplementary Table 13). Upon cross-checking our candidate protein domains, none of the seven initial GREP candidates (i.e. Galectin domain-containing proteins) were identified to contain putative IgSF or other immunoglobin domains (Supplementary Table 13).

Unlike the FBD-containing proteins (i.e. candidate FREPs), proteins bearing C-type lectin domains showed a considerable structural diversity (i.e. displaying several domains non-related with the CREP family). Some of these genes may participate in the snail immune response, as they often had signatures compatible with IgSF domains that were identified by InterPro (accessions: IPR003598, IPR003599, IPR007110, IPR013098, IPR013783 and IPR036179), but were not determined as members of the CREP family following previously described criteria [77]. Considering the hallmarks of a CREP or FREP gene (signal peptide present plus one or two IgSF domains), a total of 10 potential CREP and 57 FREP (Supplementary Table 14) genes were identified (Table 1 and Supplementary Table 13).

The selected full CREP and FREP genes were aligned with a set of reference sequences from *B. glabrata* (Supplementary Table 15), their phylogeny estimated by maximum likelihood and annotated based on their position in the phylogeny (Fig. 2 and Fig. 3). Thirty-nine of these

Table 1 Summary of variable immunoglobulin and lectin domain-containing molecules (VIgLs), fibrinogen-related proteins (FREPs), C-type lectin-related protein (CREPs) and galectin-related proteins (GREPs) identified in *Biomphalaria sudanica* (this study), *B. glabrata* and *B. pfeifferi* [8]. No GREPs were identified in either of the African *Biomphalaria* species *B. sudanica* or *B. pfeiferi*. *One FREP in *B. sudanica* contained only a partial fibrinogen domain and is considered truncated (see BSUD.19120.1). Summaries of *B. sudanica* FREP and CREP gene compositions can be found in Supplementary Table 14

Complete VIgLs	*B. sudanica*	*B. glabrata*	*B. pfeifferi*
Fibrinogen-related proteins (FREPs)	57*	39	55
C-type lectin-related proteins (CREPs)	10	4	11
Galectin-related proteins (GREPs)	0	1	0

reference sequences were reported and curated previously [44], while 17 were annotated as members of these gene families (FREP/CREP) and available on the National Center for Biotechnology Information (NCBI) (Supplementary Table 15). The CREP family is divided into two main monophyletic groups, one of which includes all the reference sequences (Fig. 2). In this group containing reference sequences, the HOGs associated with CREPs identified for *B. sudanica* were N0.HOG0002779 and N0.HOG0016618, which are closely related to the references CREP1 and CREP3, respectively. The other monophyletic group, which comprises five HOGs, is formed exclusively by divergent *B. sudanica* CREPs, possibly representing new subtypes within the CREP family.

FREP genes were grouped in thirteen HOGs, eleven of which are closely related to classified reference sequences (Fig. 3). Thus, most FREPs in *B. sudanica* could be classified as a known subtype within the family, except for FREP genes belonging to N0.HOG0000107, N0.HOG0000652, N0.HOG0021173 and N0.HOG0003967. Genes from the HOG N0.HOG0000107 is by far the largest FREP subgroup in *B. sudanica* identified in this work, and it is subdivided into three monophyletic lineages, the largest lineage comprising 25 genes is associated with several reference sequences identified as FREP12. The sister clade to the FREP12 group is the smaller monophyletic FREPJ5 group and basal to these are a group of three genes from *B. sudanica* and two unclassified *B. glabrata* reference sequences, representing an undescribed FREP lineage. Another HOG containing FREPJ2 and FREP7 (N0.HOG0000652) is remarkable since it is paraphyletic

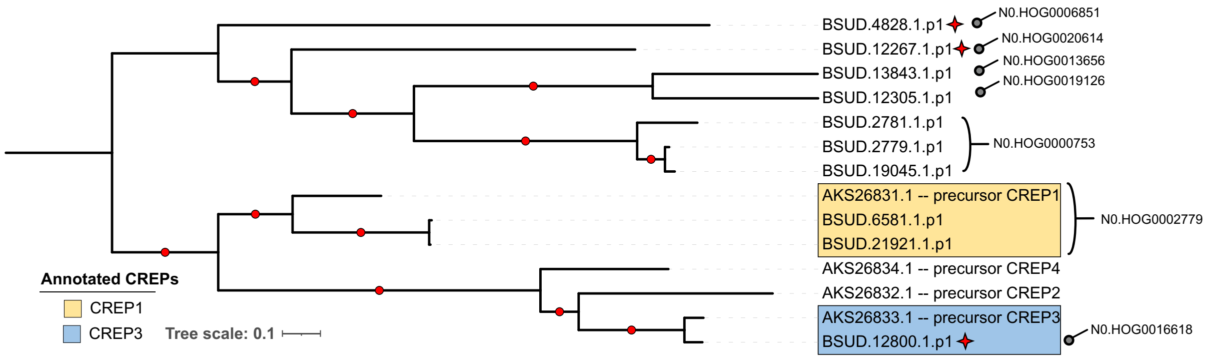


Fig. 2 Maximum likelihood tree of C-type lectin-related proteins (CREPs) identified from *Biomphalaria sudanica* in the current study (see Supplementary Table 14) and four CREPs identified previously from *B. glabrata* (see Supplementary Table 15) organized within hierarchical orthogroups (HOGs) as determined by Orthofinder [68]. Branch lengths represent the number of substitutions per site. Nodes with bootstrap values > 75 (estimated with 1000 replicates of non-parametric bootstrap) are signified by a red dot on the branch before bipartition. Red stars indicate three CREPs with unusual features, namely that they include weak hits for secondary immunoglobulin domains, which may overlap with C-lectin domains as well as containing a large interdomain region (see Supplementary Table 14)

(See figure on next page.)

Fig. 3 Maximum likelihood tree of the 57 fibrinogen-related proteins (FREPs) identified from *Biomphalaria sudanica* (bold) in the current study (see Supplementary Table 14) amongst reference sequences for FREPs identified previously from *B. glabrata* (see Supplementary Table 15) organized within hierarchical orthogroups (HOGs) as determined by Orthofinder [68]. Branch lengths represent the number of substitutions per site. Nodes with bootstrap values > 75 (estimated with 1000 replicates of non-parametric bootstrap) are signified by a red dot on the branch before bipartition. Red stars indicate FREPs with unusual features according to our annotation summarized in Supplementary Table 14, including those containing weak hits for additional immunoglobulin (IgSF) domains (e.g. BSUD.16968), IgSF rearrangements (e.g. BSUD.21927) or containing partial fibrinogen domains (e.g. BSUD.19120)

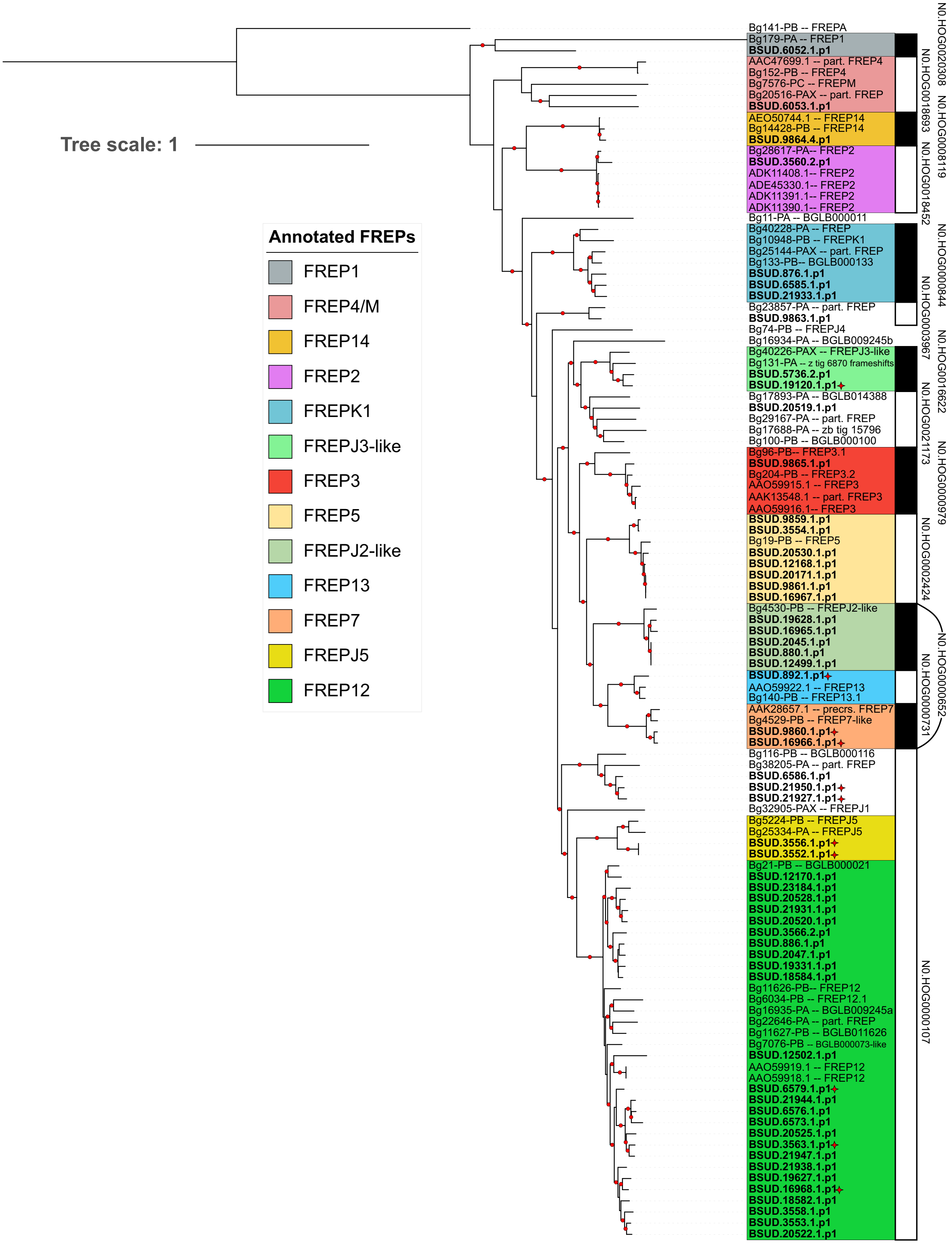


Fig. 3 (See legend on previous page.)

(Fig. 3). HOGs N0.HOG0021173 and N0.HOG0003967, each comprise a single FREP gene of *B. sudanica*, and were not associated with FREP genes previously assigned to a subgroup within the family. Lastly, HOG N0.HOG0018693 possesses reference genes annotated for FREP4 and FREPM, yet unlike N0.HOG0000107 or N0.HOG0000652, they do not form clear monophyletic groups with strong bootstrap support.

When comparing the numbers of CREP and FREP genes assigned to HOGs across the eight analyzed genomes (*Biomphalaria* spp., *Bulinus truncatus*, *Elysia marginata*), it is evident that FREPs are not only more numerous, but also their HOGs, at least for *B. sudanica*, contain more genes that did not meet our selection criteria (i.e. containing well supported functional domains, see Methods) to be considered for phylogenetic comparison (Supplementary Table 16). For example, 12 of the 42 candidate genes initially identified in N0.HOG0000107 (FREPJ5/FREP12) were rejected because of not containing complete FREP characters, suggesting these are not typical FREP genes. Secondly, the number of FREP genes assigned to each HOG varies considerably across *B. glabrata* and the African *Biomphalaria* lineages. For instance, N0.HOG0000107 is relatively abundant in the African species *B. sudanica* ($n = 30$) and *B. pfeifferi* ($n = 30$), but this number varies considerably between *B. glabrata* lines iM ($n = 6$), BB02 ($n = 0$), and iBS90 ($n = 19$) (Supplementary Table 16). In addition, only a single copy of a FREP13 classified gene (N0.HOG0000731) is present in the *B. sudanica*, *B. pfeifferi*, *B. straminea* and *B. glabrata* iM line genome as opposed to *B. glabrata* iBS90 where FREP13 is absent, and in *B. glabrata* BB02 where it comprises 27 different genes. A full study of both CREP and FREP diversity and evolution is beyond the scope of this study, but these differences suggest that the FREP genes are under a complex dynamic of duplication and replacement, while the CREPs are more stable, at least among the laboratory strains for which there are genomes available.

Intraspecific diversity of the *Biomphalaria sudanica* genome: comparing four inbred lines

In addition to the Bs111 reference genome, we conducted Illumina whole genome sequencing of four other inbred lines of *B. sudanica* and aligned them to the reference (Table 2). The line Bs163 was the most divergent, with an average genomic divergence value of 0.44% from the four other inbred lines (Bs111, Bs110, BsKEMRI and Bs5-2). The Bs163 line is also the most resistant to *S. mansoni* infection in the laboratory [78]. Among the remaining four, pairwise divergences were similar and ranged between 0.28% and 0.41% with one exception: Bs110 and Bs111 were only 0.16% divergent showing high similarity. Among all five lines, median nucleotide diversity was 0.32% (95% CI = 0.06–0.83%). The heterozygosity per inbred line ranges from 0.11% to 0.21% (Table 2).

Characterization and categorization of highly diverse genes and genomic regions in *Biomphalaria sudanica* genome

We followed a novel bioinformatic pipeline (see Methods section: *Assessment of highly diverse genes and genome regions for novel pathogen recognition receptors*) to identify putative immune-related genes (not just those involved in schistosome immunity) in the *B. sudanica* genome that may be under balancing selection, further narrowing these down to genes coding for membrane-associated proteins that could represent PRRs. We first identified genes in genomic regions that showed high divergence between the inbred lines of *B. sudanica*, hyperdiverse gene selection being based on the highest nucleotide diversity across both the entire coding and/or noncoding gene regions, and genes that were contained within the top 0.1–1% of the most nucleotide diverse sliding windows between 10–100 kb. We found multiple gene clusters that were of notably high diversity (Fig. 4 and Fig. 5). Following the bioinformatic pipeline, 1,047 hyperdiverse genes (4.4% of all genes), from 184 different contigs/scaffolds were selected for further analysis (Supplementary Table 17).

Table 2 Genome statistics from Illumina sequencing (paired end 150 bp) for four *Biomphalaria sudanica* inbred lines (Bs110, Bs163, Bs5-2 and BsKEMRI) aligned to the ~ 944.2 Mb *B. sudanica* 111 reference genome (NCBI BioProject: PRJNA1041389)

Biomphalaria sudanica inbred line	NCBI SRA Accessions	Raw total sequences	Paired reads mapped	Bases mapped *CIGAR*	Coverage	Heterozygous Sites (percentage of genome)	Number of missing sites (percentage of genome)
Bs110	SRR26849483	174,760,210	86,796,593	23,247,544,501	24.6	0.13%	4.75%
Bs163	SRR26849482	207,521,490	102,626,199	27,823,100,537	29.4	0.11%	6.16%
Bs5-2	SRR26849484	548,701,568	272,013,259	73,112,262,010	77.2	0.17%	3.83%
BsKEMRI	SRR26849481	169,400,086	83,924,684	22,815,557,212	24.1	0.21%	5.31%

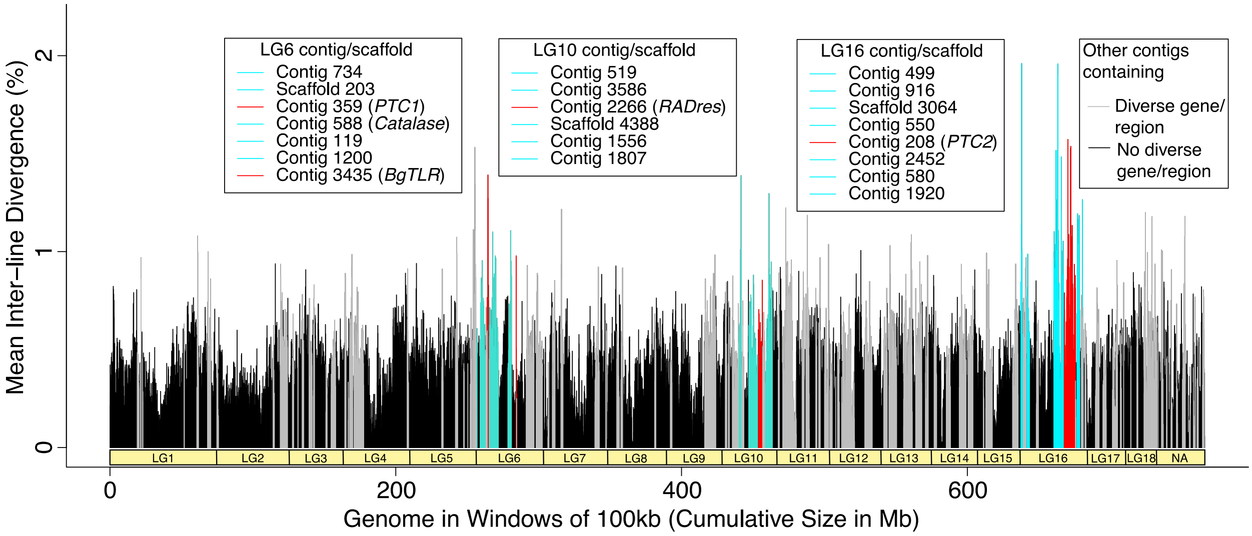


Fig. 4 Mean nucleotide diversity of five *Biomphalaria sudanica* inbred line genomes in windows of 100 kb (0 kb stagger). Linkage groups (LG1-LG18) for *B. sudanica* are inferred from *B. glabrata* [6], showing the hypothesized chromosomal position of contigs in *B. sudanica*. Notable clusters of highly diverse genomic regions and genes were seen in *B. sudanica* linkage groups LG6, LG10 and LG16 (highlighted in turquoise and red peaks). Peaks in red represent regions containing candidate immune loci that are orthologous to some of those previously associated with *Schistosoma mansoni* resistance in *B. glabrata* (*PTC1*, *PTC2*, *Catalase*, *BgTLR*, *RADres*, see Supplementary Table 2)

Characterized genes in diverse regions were categorized based on their annotation descriptions (BLAST results, InterPro, Pfam, DeepTMHMM) as to what predicted role in immune function they may have. We determined 182 (17.4% of shortlist) proteins had no role in immune function (group 1) and 245 (23.4%) proteins had a role (group 2), or 138 (13.2%) a potential role (group 3), in immune function. Of the other shortlisted genes, 81 (7.7%) had an unknown function but contained transmembrane domains (TMDs) (group 4), while the remaining 401 had an unknown function (could not be interpreted due to an absence of information) without containing a TMD (group 5). In performing comparative gene ontology analysis in topGO [73], several GO terms including immune-relevant terms concerning molecular binding and receptor activity classes, CD40 receptor complexes, and defense/inflammatory/immune responses as well as regulation of innate immune responses, were significantly enriched ($p<0.05$ Fisher's test) in this subset of 1047 highly diverse genes compared to the full *B. sudanica* genome (see Supplementary Table 18).

Although our novel pipeline was aimed at identifying immune genes under balancing selection in *B. sudanica* that could be due to interactions with any snail pathogen, some of those shortlisted were of relevance to candidate *S. mansoni* immune genes. Twelve of the 245 genes in diverse regions that were suspected to have immune function (group 2) in *B. sudanica* were orthologues to candidates for innate immune genes associated with *S. mansoni* resistance and likely act as PRRs in *B. glabrata*, namely *PTC1* ($n=5$), *PTC2* ($n=6$), and *BgTLR* ($n=1$) (Supplementary Table 17). In addition, four identified VIgLs were identified in highly diverse gene regions (Supplementary Table 17). These were FREP12 (BSUD.12502), and two neighboring FREPs on contig 777, BSUD.20519 (determined as an unknown FREP in N0.HOG0021173, Fig. 3) and BSUD.20520 (grouped within FREP12 N0.HOG0000107, Fig. 3). In addition, one full CREP (BSUD.13843, Fig. 2 and Supplementary Table 14), was also identified within the highly diverse genes/genomic regions analysis (Supplementary Table 17). Several other immune suspected (group 2) proteins contained functional domains of interest, such as Fibrinogen-related domains (FREDs), Fibronectin, C-type lectin (incomplete CREPs) and IgSF domains.

Pathogen recognition receptor candidates from the highly diverse genes and genomic regions in *Biomphalaria sudanica* genome

From the 1,047 hyperdiverse genes identified (see above), a list of 20 proteins that contained at least one transmembrane domain with the highest coding nucleotide diversity between the five *B. sudanica* inbred line genomes were identified as potential PRR candidates (Table 3).

The majority ($n=14$) of the most diverse immune/PRR genes (Table 3) were clustered on three linkage groups: 6, 10 or 16, which were also the regions of highest diversity

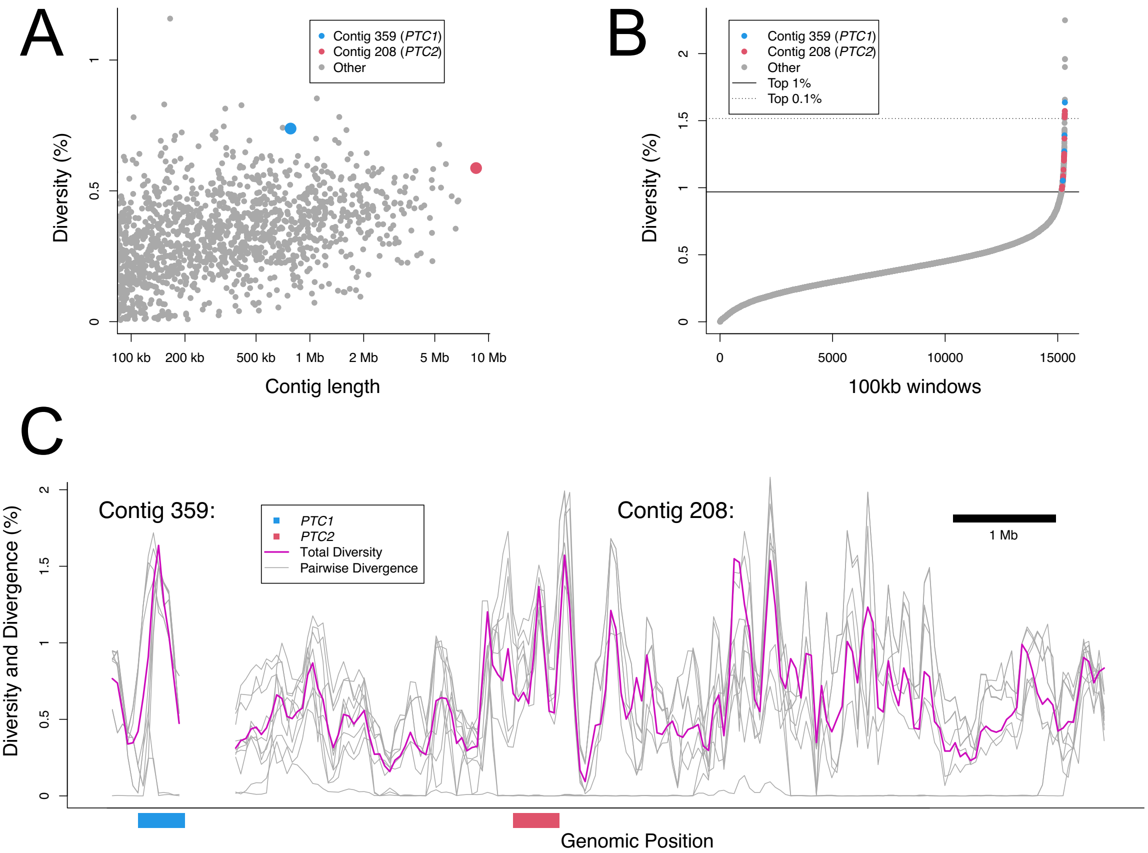


Fig. 5 **A** Patterns of nucleotide diversity across genomes of five inbred lines of *Biomphalaria sudanica*, highlighting two polymorphic transmembrane gene clusters that are orthologous to those previously associated with *Schistosoma mansoni* resistance in *B. glabrata* (*PTC1* [41] and *PTC2* [42]). **B** Genome-wide nucleotide diversity across overlapping 100-kb genomic windows (starting at 0-kb and 50-kb intervals) with windows on contigs 359 (*PTC1*) and 208 (*PTC2*) that occur in the top 1% of genome-wide nucleotide diversity between inbred lines being colored blue and red, respectively. **C** Genome-wide nucleotide diversity (purple line) and pairwise divergence for each haplotype pair (grey lines) across contigs 359 (*PTC1*) and 208 (*PTC2*). *PTC1* and *PTC2* regions are indicated by the blue and red bars, respectively. Similarly diverse regions span across several megabases of contig 208 in *B. sudanica*. Even in diverse regions, pairwise divergence can be near zero, indicating shared haplotypes

overall (Fig. 4 and Fig. 6). These linkage groups also contain genes/clusters of genes that have previously been inferred to have associations with schistosome resistance mechanisms in *B. glabrata* (*PTC1*, Catalase (*cat*), *BgTLR*, *sod*1, *prx*4, in LG6, *RADres* in LG10 and *PTC2* in LG16, see Fig. 6). Direct orthologs to the candidate PRR genes *PTC2* gene 9 (BSUD.3984) and *PTC1* gene 2 (BSUD.8884) were identified in the topmost diverse PRR-like genes in the *B. sudanica* genome (Table 3).

Pathogen recognition receptor candidates within Linkage Group 6

The order of genes in LG6 was highly conserved (Supplementary Fig. 3A). Within LG6, the hyperdiverse protein, BSUD.8884, was identified as a PRR candidate (Table 3), and is an ortholog of the *GRC*/*PTC1* gene 2 that is tied to schistosome resistance in *B. glabrata* [41]. This gene shows a tandem duplication in African *Biomphalaria* species (tandem neighboring gene in *B. sudanica* is BSUD.8885), and this duplication event has occurred independently of a similar duplication seen in some *B. glabrata* haplotypes (Fig. 7). A second shortlisted hyperdiverse protein, BSUD.20268 in contig 7608 (Table 3), could not be well described upon comparisons to other domains or proteins, perhaps due to its small (75 aa) size.

The ortholog to the *BgTLR* gene, BSUD.8256, is also contained within LG6 in a highly diverse region of the *B. sudanica* genome (Fig. 6), although the gene itself is relatively conserved within *B. sudanica* (0.1% coding nucleotide diversity). Additional immune-related genes on this linkage group are listed in Supplementary Table 17.

Pathogen recognition receptor candidates within Linkage Group 10

Synteny between *B. glabrata* and *B. sudanica* across LG10 was highly conserved (Supplementary Fig. 3B). Two proteins coded by neighboring genes located less than 1 Mb from the orthologous site of schistosome resistance

Table 3 Top 20 diverse proteins (as determined by coding nucleotide diversity) of *Biomphalaria sudanica* that contain at least one transmembrane domain, and thus represent putative pathogen recognition receptors. * Non-truncated BSUD.4885 protein is 405 aa in length as determined by manual alignment

Gene (contig/scaffold, *Linkage Group*)	Nucleotide Diversity % (coding)	Protein length (aa)	Inferred protein-receptor function / functional domains / protein family
BSUD.20937 (80, *LG1*)	5.7	453	Tumor Necrosis Factor (*TNF*)
BSUD.4884 (2266, *LG10*)	5.3	86	Unknown, linked (same contig) to *RADres* [79]. Bacteriocin domain
BSUD.4885 (2266, *LG10*)	4.8	354*	Unknown, linked (same contig) to RADres [79]. Domains: TMEM154 protein family, Rifin
BSUD.3983 (208, *LG16)*	4.0	817	Unknown, linked (same contig) to *PTC*2 [42]. Domains: SHP2-interacting transmembrane adaptor protein, SIT
BSUD.4096 (208, *LG16)*	3.7	270	Unknown, linked (same contig) to *PTC*2 [42]. Domains: Viral glycoprotein L
BSUD.20268 (7608, *LG6)*	3.3	75	Unknown. Domains of unknown function (DUF6768)
BSUD.9255 (379, *LG11)*	3.3	453	Unknown Domains of unknown function (DUF4781), Death domain
BSUD.15077 (580, *LG16)*	3.1	318	Cell adhesion molecule. Wnt and fibroblast growth factor (FGF) inhibitory regulator and Protocadherin domain.
BSUD.4112 (208, *LG16)*	2.9	341	Cell adhesion molecule linked (same contig) to *PTC*2 [42]. Protocadherin domain.
BSUD.18104 (69, *LG15)*	2.8	1154	G-protein coupled receptor. 7 transmembrane receptor (rhodopsin family), Leucine-rich repeat
BSUD.3984 (208, *LG16)*	2.7	324	Polymorphic transmembrane cluster 2 (*PTC*2) gene 9 [42].
BSUD.14211 (540, *LG5)*	2.7	516	Cell adhesion molecule. F5/8 type C domain.
BSUD.17807 (6862, *LG13*)	2.7	504	Cell adhesion molecule. Ephrin type-A receptor 2 transmembrane and immunoglobulin domain.
BSUD.14415 (550, *LG16)*	2.6	483	Unknown. Wolframin EF-hand domain
BSUD.9838 (3979, *LG4)*	2.6	203	Transient receptor potential cation channel subfamily M member 2-like. Ion transport domain
BSUD.12903 (499, *LG16)*	2.6	647	Receptor with C-type lectin and immunoglobulin domain
BSUD.23928 (4388, *LG10*)	2.5	944	G-protein coupled receptor. 7 transmembrane receptor (rhodopsin family), Low-density lipoprotein receptor domain class A, Leucine-rich repeat
BSUD.8884 (359, *LG6)*	2.4	565	Guadeloupe Resistance Complex (*GRC/PTC*1) gene 2 [41, 80]. Fibronectin type III domain, TMEM154 protein family.
BSUD.4003 (208, *LG16)*	2.4	473	Unknown, linked (same contig) to *PTC*2 [42]. Prodomain subtilisin 2, TMEM154 protein family.
BSUD.4089 (208, *LG16)*	2.3	331	Receptor with C-type lectin domain.

marker *RADres1* [79], BSUD.4884 and BSUD.4885 (contig 2266), showed extremely high (~5%) nucleotide diversity between the *B. sudanica* inbred lines (Table 3). The predicted function of BSUD.4884, an 86 aa protein partially matching a hypothetical protein recorded in *B. glabrata* (GenBank accession KAI8743980), could not be confidently characterized due to an absence of significant matches to orthologous proteins or domains (Supplementary Table 19). Comparisons with the annotated *B. glabrata* genome also revealed that the neighboring protein, BSUD.4885, had been erroneously truncated on the extracellular portion; the manually annotated gene codes for a 405 aa protein. The complete BSUD.4885 protein was predicted to contain both TMEM154 and Rifin domains (see [81]), features shared with some *PTC1* and/or *PTC2* genes (see BSUD.3980, BSUD.8873, BSUD.8874, BSUD.8876 and BSUD.8884, Table 3 and Supplementary Table 17). Remarkably, haplotype lineages of BSUD.4885 in *B. sudanica* also occur in *B. glabrata* (Fig. 8), suggesting that either this is a shared polymorphism within the genus that has survived the colonization of Africa, or that there were originally two divergent tandem paralogs but one or the other is independently deleted in every genome since no genome appears to have two copies of this gene.

Pathogen recognition receptor candidates within Linkage Group 16

Linkage group 16 contains many of the most diverse genomic regions (Fig. 9), such as contig 208, contig 499, contig 550, contig 580, and scaffold 3064, which include highly diverse individual genes (Supplementary Table 17). Six of the PRR candidates within LG16 originate from contig 208 (BSUD.3983, BSUD.4096, BSUD.4112, BSUD.3984, BSUD.4003, BSUD.4089, Table 3), the orthologous region to the *PTC*2 region in *B. glabrata* [42]. *PTC2* genes known to show high diversity in *B. glabrata* show comparable divergence among alleles in *B. sudanica* (e.g. BSUD.3979, Fig. 7B). As the orthologous *PTC2* genes are adjacent to these and many

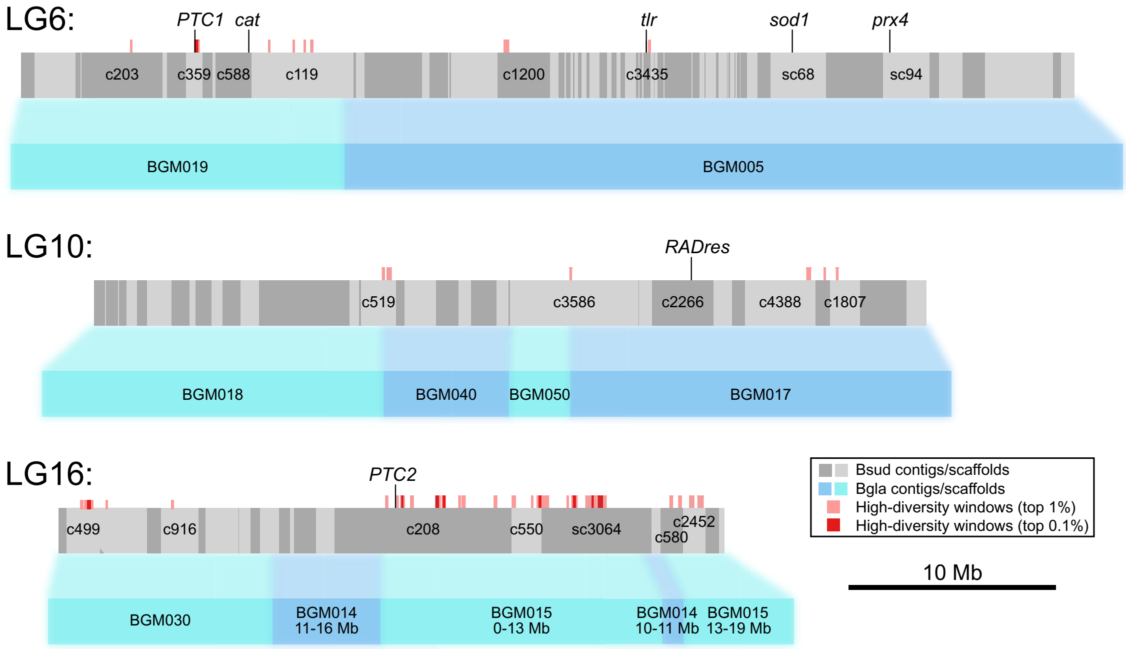


Fig. 6 Orthology of *Biomphalaria sudanica* contigs/scaffolds (grey boxes, with different shades of grey representing alternating contigs/scaffolds) to *B. glabrata* scaffolds (blue boxes, [6]) pertaining to linkage groups (LG) inferred from three *B. glabrata* linkage groups, highlighted here because they are notably enriched for both diverse regions/genes and orthologous candidate immune genes from *B. glabrata* (*PTC1*, *cat*, *BgTLR*, *sod1*, *prx4*, *RADres* and *PTC2*), positions of which are shown. Contigs/scaffolds with candidate genes and/or diverse regions (100 kb windows in top 1%, light red boxes, or top 0.1%, dark red boxes) are labeled. Synteny is relatively high between species, except for a large rearrangement on LG16 (see Supplementary Fig. 3C)

other diverse genes, several of which have single transmembrane domains and/or homology to *PTC2* genes (e.g. BSUD.4003), the resistance-associated *PTC2* region originally described in *B. glabrata* appears to be part of a much wider region of contig 208 (Fig. 5C; Fig. 6) with distinctive evolutionary and structural features. Furthermore, contigs neighboring contig 208 according to our inferred linkage map also contain single transmembrane domain proteins with homology to *PTC2* genes (BSUD.14425, BSUD.15084 and BSUD.23257 on contig 550, contig 580 and scaffold 3064, respectively).

PRR candidates BSUD.4112 (contig 208) and BSUD.15077 (contig 580) were predicted to be cell adhesion molecules since both contained a domain matching that of a protocadherin (Table 3; Supplementary Table 19). Two other PRR candidates, BSUD.12903 (contig 499) and BSUD.4089 (contig 208) (Table 3), both contain an extracellular C-type Lectin domain known to function in immune responses to pathogens. For BSUD.12903 at least, nonsynonymous diversity substantially exceeds synonymous diversity in the extracellular region of the protein where these predicted functional domains (C-type lectin and IgSF) were present (Fig. 10). BSUD.12903 occurs within a cluster of proteins representing one of the most diverse regions of the genome, many of which contain combinations of C-type lectin, IgSF and FN3 domains (Supplementary Table 17). BSUD.4089 encodes a longer isoform (331 aa) with a single transmembrane domain and an extracellular C-type lectin domain, and a shorter isoform (67 aa) with just the secreted peptide signal (confirmed through InterProScan, see Supplementary File 3) and a partial match to the C-type lectin domain, which would result in a truncated domain.

A large inversion (~15 Mb) between *B. sudanica* and *B. glabrata* iM line on LG16 appears to reverse parts of scaffold 3064, contig 208 and contig 550 relative to *B. glabrata* scaffolds BGM014 and BGM015 (Fig. 6; Supplementary Fig. 3C). BGM014 itself may be incorrectly assembled since it is split across *B. glabrata* linkage groups LG3 and LG16.

Discussion

A new genomic resource for vector biology

Long-read PacBio HiFi DNA and RNA sequencing, supplemented by Illumina short-read sequencing of RNA, were performed on an inbred line of *B. sudanica* sensu lato, originally collected from shoreline habitats of the Kisumu region of Lake Victoria. This sequencing approach and bioinformatic pipeline resulted in a

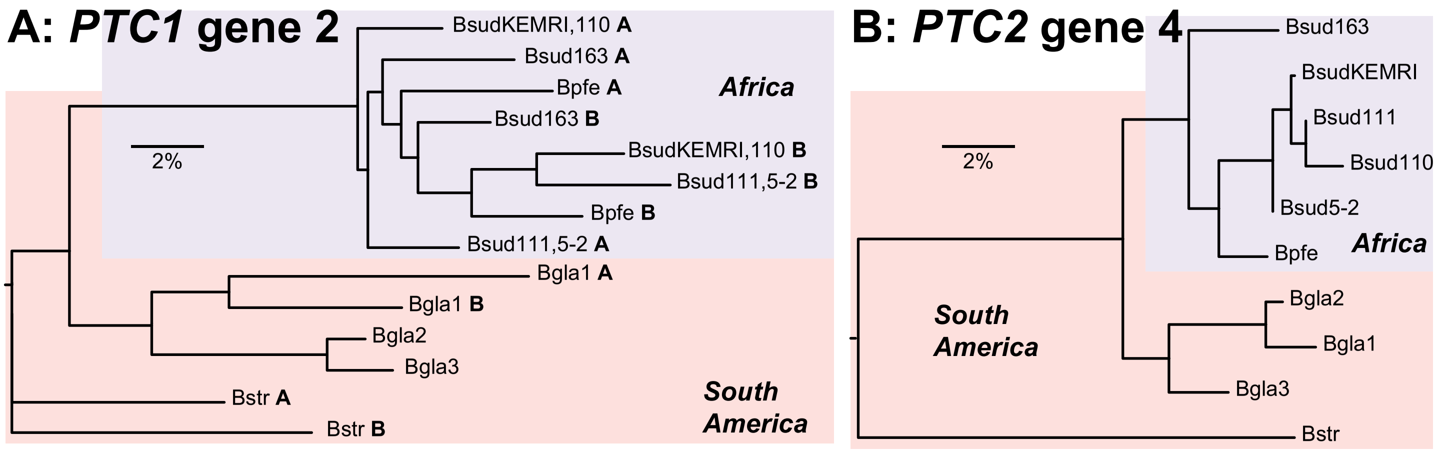


Fig. 7 Allelic phylogenies of *Biomphalaria sudanica*, *B. pfeifferi* (Bpfe), and *B. glabrata* (Bgla), rooted with *B. straminea* (Bstr). (A) Polymorphic transmembrane cluster 1 (*PTC1*) gene 2 (i.e. *grctm2* [41]) shows a tandem duplication (see A and B duplicates for relevant taxa indicated in bold) in the African species (A = BSUD.8884 and B = BSUD.8885 for *B. sudanica*) that is clearly independent of similar duplications seen in *B. glabrata* haplotype 1 (Bgla1) and *B. straminea*. Bgla1, Bgla2 and Bgla3 refer to *PTC1* haplotypes sequenced from *B. glabrata* [41]. (B) *PTC2* gene 4 (BSUD.3979 in *B. sudanica*) shows distinct haplotypes in each *Biomphalaria* species. Bgla1, Bgla2 and Bgla3 refer to *PTC2* haplotypes sequenced from *B. glabrata* [42]

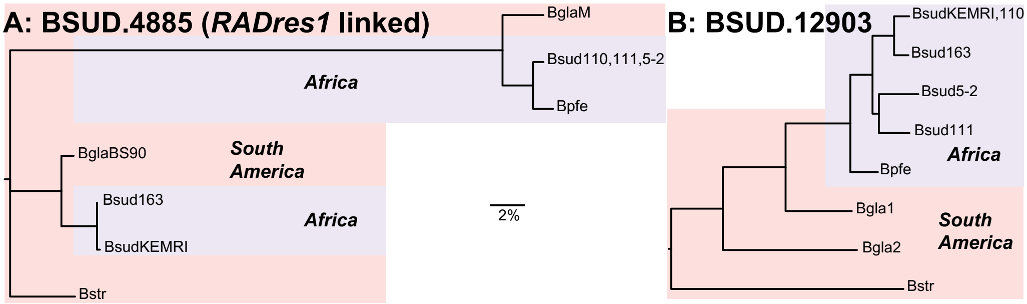


Fig. 8 Phylogenetic trees generated using RAxML [82] of exemplar genes showing unusually high diversity in *Biomphalaria sudanica*, alongside orthologous alleles in *B. pfeifferi* (Bpfe) and *B. glabrata* (Bgla), and rooted with *B. straminea* (Bstr). Bgla1 and Bgla2 refer to alleles sequenced from *B. glabrata* inbred lines [42], whilst BglaBS90 and BglaM represent alleles from two other *B. glabrata* inbred lines [6]. (A) BSUD.4885 (contig 2266, linkage group (LG) 10) is a gene with exceptionally high diversity, has a protein structure that is similar to genes previously inferred as pathogen recognition receptors, and shows an apparent trans-species polymorphism of divergent haplotypes in African and South American snails. (B) BSUD.12903 (contig 499, LG16) is another pathogen recognition receptor candidate, in one of the most polymorphic contigs in the *B. sudanica* genome, predicted to contain C-type lectin, immunoglobulin, TMEM154 and alternatively expressed fibronectin III domains (see Fig. 10)

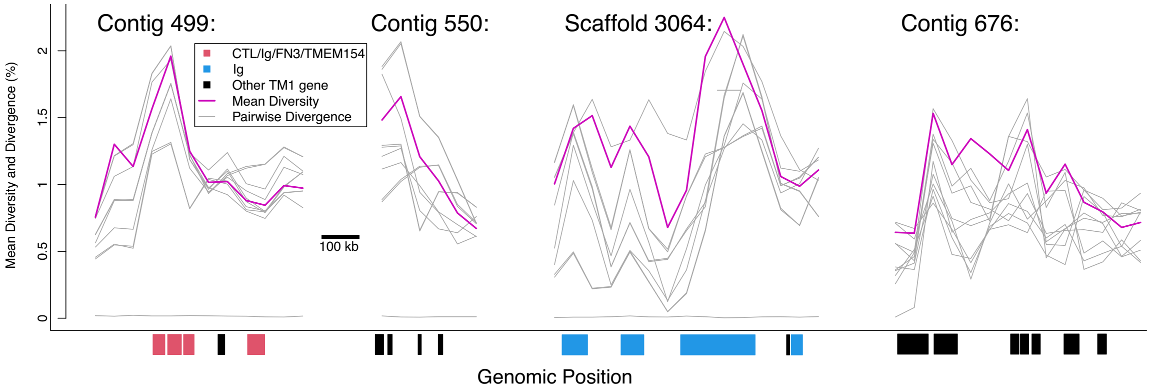


Fig. 9 Mean nucleotide diversity (purple line) and pairwise divergence for each haplotype pair (grey lines) across portions of four contigs in LG16 (contig 499 300-850 kb, contig 550 0–250 kb, scaffold 3064 2300–3000 kb) and LG5 (contig 676 150–850 kb) showing high diversity and clusters of transmembrane genes. All plotted genes (shown in red, blue and black) encode single-pass transmembrane proteins (TM1); other genes in these regions are not shown. Key functional domains potentially involved in pathogen recognition include C-type lectin (CTL), fibronectin type III (FN3) and immunoglobulin (Ig), and TMEM154, which is a membrane-spanning domain also found in several polymorphic transmembrane cluster 1 (*PTC1*) genes [41]. Genes shown in red, including BSUD.12903 (Fig. 10) all have at least three of these four functional domains, while genes shown in blue have only Ig and genes shown in black represent other genes encoding TM1 and various other protein domains

well-supported genome annotation, which adds to the growing repository of *Biomphalaria* species genomes. *Biomphalaria sudanica* remains a "neglected vector" despite its established role in transmission of *S. mansoni* in lake and marsh habitats in the African Rift Valley, where schistosomiasis transmission is among the highest in the world. Our analyses focused on characterizing immune genes using a comparative framework with that of the better-studied South American vector of schistosomiasis, *B. glabrata*, and another African vector, *B. pfeifferi*, for which genomic resources have recently been developed [8]. We also developed a novel bioinformatic pipeline to characterize regions of marked intraspecific diversity as a mechanism to identify novel genes that may be involved in pathogen recognition and are under balancing selection. These analyses and resources will facilitate future work to explore further and uncover additional vector immune defense mechanisms against pathogens such as schistosomes, with direct relevance to public health in Africa where the majority of *S. mansoni* infections occur. The hope is that the scientific community will use these resources to support the development of novel tools for schistosome control, potentially including gene-drive manipulation of the snail vector [83], and development of surveillance tools, as is described as a priority in World Health Organization guidelines [84].

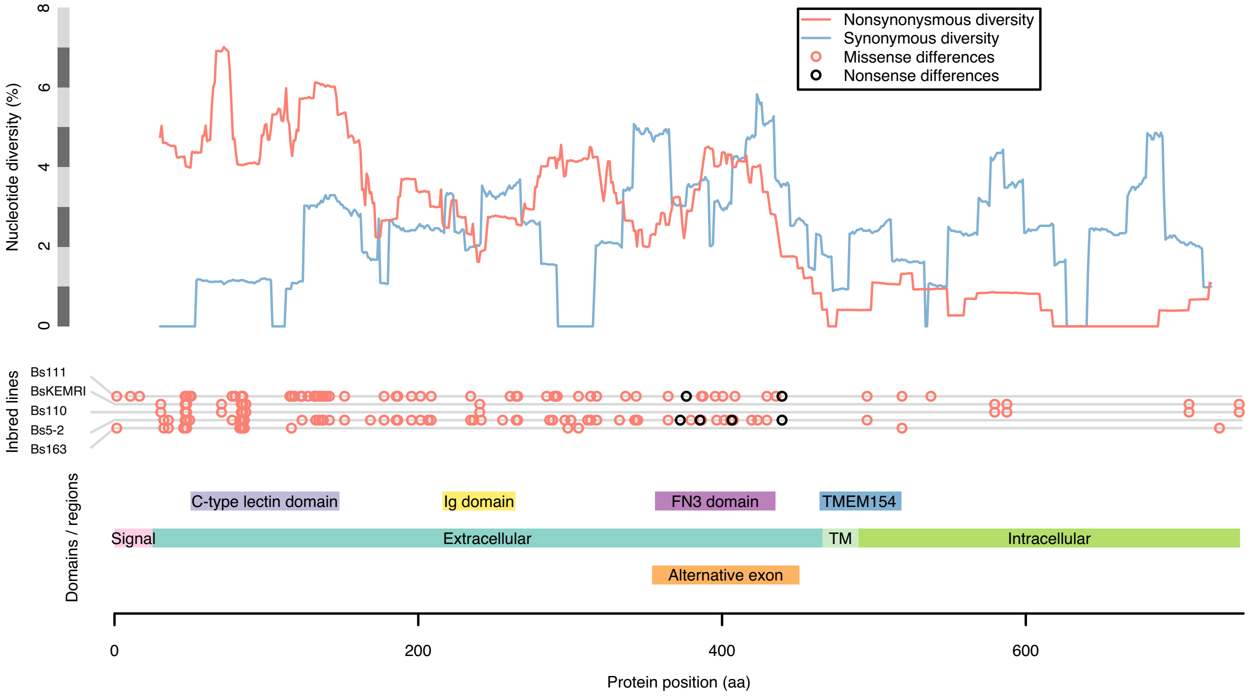


Fig. 10 Detailed view of protein BSUD.12903 containing the alternatively expressed exon of 741 aa, which demonstrates that in the extracellular region nonsynonymous diversity greatly exceeds synonymous diversity (shown here in 50 aa sliding windows) in regions where functional domains potentially involved in pathogen recognition are present (C-type lectin, immunoglobulin, and fibronectin III (FN3) domains). Two inbred lines of *Biomphalaria sudanica* (Bs111 and Bs5-2) have multiple nonsense variants (stop codon or frameshift) in a single exon containing a FN3 domain, which is expressed in the *B. pfeifferi* ortholog (KAK0057508 [8]), suggesting that FN3 may not be expressed in all *B. sudanica* lines. The FN3 exon also occurs in the *B. glabrata* ortholog BGLB024560, where it is also variably either expressed (NCBI Accession XM_056010427) or excluded (NCBI Accession SRX8534561). Both FN3 and/or TMEM154 domains are present in *B. sudanica* genes BSUD.8884, BSUD.8874 and BSUD.8876 that are orthologous to *B. glabrata PTC1* region genes associated with schistosome resistance: *grctm2*, *grctm3* and *grctm4* [41]

Comparative analysis of immune genes among species

Based on the average number of substitutions per site, the genomes of *B. sudanica* and *B. pfeifferi* are approximately 5.1% divergent from each other, which is comparable to that of the isolates of *B. glabrata* compared here (between 3.3–8.7%) (see Fig. 1). We found that *B. sudanica* possesses orthologous genes to all those previously identified from *B. glabrata* as having a potential function in immunity against schistosomes. However, the functions of these orthologs have yet to be verified for *B. sudanica*.

The number of complete variable IgSF and lectin domain-containing molecules (FREPs and CREPs) in *B. sudanica* was almost identical to that observed in *B. pfeifferi*, with both species having a higher number than observed in *B. glabrata* (Table 1). If the expansion of the FREP/CREP gene family in the African *Biomphalaria* is biologically real, and not a consequence of differing genome assembly methods compared to *B. glabrata* BB02, one hypothesis for this gene family expansion may be that the trans-Atlantic colonization of the *Biomphalaria* species may have necessitated an expansion in defense molecules such as FREPs whilst encountering a diverse array of new pathogens. The largest and most divergent FREP family in *B. sudanica* is FREP12, two genes of which were not identified by the VIgL annotation pipeline but instead were discovered with the pipeline to identify highly diverse immune-related proteins and PRRs. Within the African species, some FREP-like gene families were enriched in *B. sudanica* and underrepresented in *B. pfeifferi.* This association is interesting given the observation that *B. pfeifferi* is more susceptible to schistosome infections compared to *B. sudanica* [85, 86]; however the bulk of enriched FREPs in *B. sudanica* (FREPK1, FREPJ3, FREP5, FREPJ5, FREP12) are not known to play a role in schistosome resistance as based on studies of *B. glabrata* with comparable studies for African species yet to be undertaken. It is also apparent that FREP genes duplicate more readily in *B. sudanica*, resulting in more truncated FREPs than CREPs.

It should be acknowledged that the annotation and nomenclature of FREPs is not trivial. Diversification of FREPs by gene conversion, duplications, gene loss, exon shuffling, and other structural rearrangements, rather than the slow accumulation of mutations, complicates phylogenetic analysis and therefore classification and nomenclature of these genes [44]. Due to the complex molecular evolution of FREPs, the *B. sudanica* genome contains a diverse and potentially variable suite of FREP sequences as seen in *B. glabrata* [87], independent of nucleotide diversity at any particular gene. Further investigation into these hypervariable genes is necessary to fully describe the molecular mechanisms that drive their diversity and interactions with pathogens.

Elevated diversity at candidate innate immune gene regions, *PTC1* and *PTC2*, in both *B. sudanica* and *B. glabrata* appears to be independently generated, since the genes remain reciprocally monophyletic (Fig. 7), which suggests ongoing selection favoring novel diversity at these loci. Occasional transspecies polymorphisms, including one apparent at BSUD.4885 (Fig. 8), are suggestive of a long-term balancing selection, strong enough to overcome what was likely a narrow population bottleneck during the colonization of Africa.

Intraspecific genomic diversity

Hyperdiverse protein-coding genes in *B. sudanica* that were shortlisted as potential PRRs under balancing selection were enriched in molecules that have been associated with various aspects of immune functions in other organisms. The hyperdiverse transmembrane proteins included G-protein coupled receptors (GPCRs), TLRs, cluster of differentiation (CD) and cell adhesion molecules as well as proteins containing functional domains such as C-type lectins, FN3, IgSF, transient receptor potential (TRP) channels and many potentially immune related domains that await further immunological characterization. These often show high extracellular nucleotide non-synonymous diversity which overlaps with functional binding domains (see Fig. 10). While we have focused on candidate PPRs with high nucleotide diversity, we do not mean to discount the importance of conserved genes, or the between-paralog diversity found in multi-gene families. In particular, genes with well-studied immune roles in *B. glabrata*, such as FREPs [43] and components of the oxidative burst pathway [88], undoubtably also play essential roles in *B. sudanica* immunity regardless of their allelic diversity.

The most diverse candidate protein (Table 3; BSUD.20937) has a single transmembrane domain and a predicted tumor necrosis factor (TNF) domain. TNFs in invertebrates are far less characterized than those in mammals; however, recent work in mollusks and crustaceans indicates that just like in vertebrates, they have multiple roles in innate immunity and function in activation of antimicrobial peptides, apoptosis and phagocytosis of hemocytes, and activation of immune related enzymes such as lysozyme and phenoloxidase [89–91]. Recent work has also demonstrated the evolution of TNF signaling in parasitic platyhelminths (such as *Schistosoma* sp.) in response to TNF ligands of their mammalian hosts immune signaling [92]. Whether this kind of crosstalk between parasites and *B. sudanica* occurs cannot be determined from our results, but provides an avenue for future investigation. We could not detect this gene in two inbred lines (Bs163 and Bs5-2), presumably due to either deletion or extreme sequence divergence, and thus our diversity measurement may even be an underestimate. The high diversity of BSUD.20937 between *B. sudanica* inbred lines, and that of the other hyperdiverse transmembrane genes selected in our bioinformatic pipeline, suggests that these may well be PRR genes under balancing selection driven by variability in ligand binding sites that interact with pathogens.

Regarding genome-wide patterns of diversity, the identified high diversity windows were clustered on three linkage groups (6, 10, and 16), and near candidate immune genes. Linkage group 6 contains a cluster of highly diverse windows as well as several candidate immune genes including *PTC1*, *cat*, *tlr*, *sod1*, and *prx4*. Linkage group 10 also harbors diverse windows as well as resistance region *RADres*. The highest diversity genomic windows occur on linkage group 16, site of the highly diverse gene cluster *PTC*2 which we show is imbedded within a larger region of diversity (Fig. 4 and 6) that shows chromosomal rearrangement between *B. sudanica* and *B. glabrata* (Supplementary Fig. 3C). Such structural rearrangements may be enriched along with single-nucleotide polymorphisms in hyperdiverse regions like *PTC2* and may even help to maintain diversity if they are segregating within species; however, we also note that the 8.5 Mb contig 208 is the largest contig in our Bs111 assembly and thus it afforded the greatest power to observe large rearrangements that may be undetectable elsewhere in the genome. Immune-relevant genes may often be linked within chromosomal regions with distinct evolutionary dynamics, including perhaps the maintenance of elevated adaptive diversity, as also noted for *B. glabrata* [93, 94].

Although an effective method for PRR discovery as demonstrated by our results here, our genome wide nucleotide diversity estimates will inherently underestimate the true genome wide diversity. This is in part because the most highly divergent reads from *B. sudanica* inbred lines will not have aligned successfully to our Bs111 reference genome assembly, and also because divergent alleles may have been falsely assembled as distinct contigs in this Bs111 reference genome. Therefore, we demonstrate here the capacity to find diverse regions that are at least as diverse as they appear but emphasize that there may be additional diverse regions missed that may contain other PRR genes.

Molecular evolutionary legacy of colonization and adaptation in Africa

Biomphalaria originated in South America, but at least one lineage crossed the Atlantic based on previous estimates approximately 1.8 and 5 MYA [8, 10–12] and diversified into the contemporary African species. Such a transcontinental colonization, whenever it did take place, is striking for a freshwater gastropod of limited vagility.

Since the transcontinental colonization of *Biomphalaria* is predicted as sufficiently evolutionarily recent, its history can be explored more easily with the growing collection of *Biomphalaria* genomes. Despite the divergence between *B. glabrata* and the African species being between 10.5–10.8% (Fig. 1), thousands of orthologs could be aligned easily across these species and without saturation of neutral sites. These highly orthologous genomes allowed us to examine the expansion and contraction of gene families, with particular interest being paid to the genomic consequences of migration and adaptation of *Biomphalaria* to Africa. Originally, we hypothesized that immune genes of this African *Biomphalaria* ancestor would be substantially expanded as it encountered several new pathogens as it colonized Africa from South America; however, the analysis revealed no significantly expanded immunity-related genes. In fact, the contraction of genes outweighed expansions in the common ancestor of *B. sudanica* and *B. pfeifferi* compared to the South American congeners, suggesting an overall simplification of the genome in *B. sudanica* and *B. pfeifferi*. Perhaps this gene family contraction in African *Biomphalaria* was a maladaptive result of a founder effect, given that the ancestor of African *Biomphalaria* must have descended from a small number of colonizers, and/or perhaps many genes were freely lost without fitness consequences in the absence of the neotropical pathogens to which they were adapted. Despite this proteome simplification, the estimated genome size of *B. sudanica* was ~ 73 Mb larger than the genome of outcrossing *B. glabrata*, and also ~ 173 Mb larger than that of *B. pfeifferi*, which is hypothesized to have lost genes associated with mating given its reliance on selfing [8]. This size discrepancy could be a technical artifact from divergent allelic *B. sudanica* haplotypes getting assembled as separate contigs (see above), perhaps due to the higher heterozygosity in our Bs111 after only three generations of selfing, compared to naturally inbred *B. pfeifferi*, which is a preferentially selfing snail [95, 96], and the long-term inbred lines of *B. glabrata*. Alternatively, if the increased genome size of *B. sudanica* is biologically real, it could represent gene duplications and insertions not picked up in our current analysis, which may have led to novel gene functions and expression mechanisms favored in its African habitat. The mechanism could also be selectively neutral: the high repeat content of all *Biomphalaria* genomes indicates that total genome size can vary considerably with little direct effect on the content of coding genes.

As confirmed through our annotation of RNAs, a common feature in the two African species is the presence of tRNA-SeC, demonstrating the ability of *B. sudanica*, like its sister species *B. pfeifferi* [8], to synthesize polypeptides containing selenocysteine [54, 55]. Although selenoproteins are present in many gastropod lineages [97], tRNA-SeC has not been identified in *B. glabrata* [8], suggesting that the capacity to produce selenocysteinyl has been gained in the African *Biomphalaria* species, or lost across South American lineages of *B. glabrata*. With the growing repository of genome information for species basal to *Biomphalaria*, it will be possible to investigate this further. Selenocysteine-containing proteins, i.e. selenoproteins, are proposed to be involved in a wide range of bioactive processes in other invertebrates [98].

Interestingly, regarding rRNA genes, far fewer were predicted in *B. sudanica* compared to *B. pfeifferi*, although this is likely due to misassembly of the highly repetitive tandem repeats that rRNA genes are generally present in, causing near identical rRNA copies to be collapsed into single copies, rather than biological differences.

Transcriptomic analysis illuminates unusual mitochondrial transcription processes

Molluscan mitochondrial genomes are unusual in terms of their genomic features and transcriptomic processes [57, 59], and the latter has not been explored for African *Biomphalaria*. In *B. sudanica*, mitochondrial genes that were not separated by tRNA genes, namely *nad6*, *nad5*, and *nad1*, were highly represented by polycistronic mRNA transcripts with no clear trimming points. Arrangement and transcription processes of these genes therefore appear similar in other invertebrate and molluscan species [59, 99, 100]. In addition, *nad4l* lacked read coverage, with the few transcripts that were observed being polycistronic with *cob*. Lack of expression in key genes of the respiratory chain, such as the NADH dehydrogenase, is not unheard of within mollusks [101], but regardless the sharp increase on read coverage at the start of *cob*, suggests this is also processed by endonucleases.

Our transcript for *B. sudanica* across mitochondrial genes *atp6* and *atp8*, which in other species form a single mRNA [99], are punctuated by trnN in *B. sudanica* as in *B. glabrata* [60]. However, the bulk of the recovered reads belong to the pre-mRNAs. Considering their polycistronic status in other organisms, it is possible that translation could happen before the pre-mRNA is fully resolved, which would add another layer of complexity to molluscan mitochondrial genomes in that these genes are likely separated downstream into proteins through initiation on the ribosome (as reviewed in [59]). Additionally, given that *atp8* has been found under relaxed selective pressures and a high degree of variation in both *Biomphalaria* and *Bulinus* species [56, 102], it may play a less important role in the mitochondrial function within these planorbids.

Conclusions

The species *B. sudanica* is remarkable for two principal reasons: its dynamic evolutionary history involving recent colonization and diversification in Africa, and its tragic impact on public health as a schistosome vector. The genomic data presented here illuminate both aspects. Our observations include expansions and contractions of gene families in African snails, as well as numerous genomic regions of high diversity that contain genes that may play a role in host–pathogen coevolution and their vectorial capacity for parasites such as schistosomes. In combination with increasing genomic resources from other *Biomphalaria* isolates, this work will facilitate an enhanced understanding of the biology of these snails and future mechanisms for curbing transmission of schistosomiasis.

Methods

PacBio whole genome sequencing and assembly of *Biomphalaria sudanica*

High molecular weight DNA was isolated from the head-foot tissue of a single adult (>8 mm in shell diameter) *Biomphalaria sudanica* snail from the inbred line "111" (Bs111) following the Qiagen Blood & Tissue kit (Qiagen, MD, USA), the only modification to standard kit protocol being the overnight lysis of tissue at 37 °C. The inbred line, Bs111, was developed from snails originally collected from the Kisumu region of Lake Victoria, Kenya in 2012 and bred via selfing for three generations before the line was expanded through sibling mating and gDNA extracted. The library was generated using the SMRT-bell Express template Prep Kit 2.0 (Pacific Biosciences, CA, USA). PacBio high-fidelity (HiFi) circular consensus long-read sequencing [103] was carried out using two SMRT cells of a PacBio Sequel II at the University of Oregon. PacBio ccs reads (raw reads available in NCBI BioProject: PRJNA1041389) were assembled using Flye [104], settings: flye –pacbio-hifi Bsud111_ccs.fasta -g 1 g -I 0 -t 16 -o flye/Bsud111. Basic statistics of the genome assembly were calculated in seqkits v0.16.1[105].

Illumina whole genome sequencing, alignment and genotype calling of four genetic lines of *Biomphalaria sudanica*

Short-read Illumina data for single adults from four additional *B. sudanica* genetic lines Bs5-2, Bs110, Bs163 and BsKEMRI, all originally collected between 2010–2012 from the Kisumu region of Lake Victoria, Kenya, were also produced. Inbred lines were generated via selfing in isolation for 3 generations, after which, the lines were expanded through sibling mating. The BsKEMRI line snail was not subject to intentional inbreeding but was maintained in the laboratory for ~ 10 years (2010–2020) before gDNA was extracted for sequencing. High molecular weight DNA was extracted from single snails using the modified Qiagen Blood & Tissue kit (Qiagen, MD, USA) described above. The PrepX Robotic DNA Library Prep (Takara Bio, CA, USA) kits were used for library preparation. Illumina pair-ended reads (150 bp) were generated using the HiSeq 3000 at the Center for Quantitative Life Sciences at Oregon State University to obtain 15–20 × coverage for each sample. The FASTQ files had adapters removed with Cutadapt v1.15 [106] and then reads were trimmed with Trimmomatic v0.30 [46] with options: LEADING:20, TRAILING:20, SLIDINGWINDOW:5:20, MINLEN:50, and aligned using BWA version 0.7.12 [107] using command: bwa mem -P -M -t 4, with the PacBio assembly of *Bs*111 (see above). Aligned genome data and raw reads are available in archive NCBI Sequence Read Archive (NCBI BioProject: PRJNA1041389).

Genotypes for each snail genome were called against the Bs111 reference genome in variant call format (vcf) produced using BCFtools v1.9 [108]. Basic statistics of the alignments to the genome were calculated using vcftools v0.1.17 [109].

Nuclear genome annotation of *Biomphalaria sudanica*

The bioinformatic pipeline for the annotation of *B. sudanica* was performed using a custom script (see: github.com/J-Calvelo/Annotation-Biomphalaria-sudanica). To obtain the best possible genome annotation for *B. sudanica*, combined long- and short-read RNA-seq data was used (NCBI BioProject: PRJNA1041389). RNA for long- and short-read sequencing was obtained from multiple samples of pooled Bs111 snails during two developmental stages (juvenile and adult) or under different stressors (heat stress or exposure to *S. mansoni* NMRI strain from the NIH-NIAID Schistosomiasis Resource Centre [110]). Full details of samples used, and RNA extraction protocol and pooling are contained in the Supplementary Material (Supplementary File 6). The equal mass pool of eight samples was processed for sequencing using the standard

IsoSeq protocol (PacBio protocol: PN 101–763-800 v2). The NEBNext Single Cell/Low Input cDNA Synthesis and Amplification Module kit (New England BioLabs Inc. MA, USA) was used following manufacturers protocols for cDNA synthesis and amplification from 300 ng of RNA, and the library was generated using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, CA, USA). A Pronex bead purification (Promega, WI, USA) was performed on the amplified cDNA product following a size selection of 2 kb transcripts. The final cDNA library was sequenced on one SMRT cell on the PacBio Sequel II at GC3F, University of Oregon.

Short-read Illumina sequences of RNA were obtained from two samples, one of a single unchallenged Bs111 and the second a pool of three Bs111 adult snails that had been challenged to *S. mansoni* miracidia 24 h prior (Supplementary File 6). Library preparation and directional mRNA sequencing were performed at Novogene Corporation Inc. (CA, USA). Library preparation was performed using the NEBNext Ultra II Directional RNA Library Prep Kit (New England BioLabs Inc. MA, USA) for strand-specific Illumina libraries. Libraries were then sequenced on the NovaSeq 6000 platform (Illumina, CA, USA) and 150 bp pair-ended reads were generated to provide ~ 12 Gb of sequence data.

Bioinformatic analysis of long- and short-read RNA-seq data

Quality of the long-read data was first assessed with longQC software [45]. Then, reads were processed with Lima (github.com/PacificBiosciences/barcoding), with the options –-css and–-dump-clips”, and isoseq3 refine tool (github.com/PacificBiosciences/IsoSeq), with the option –-require-polya”, to recover all complete sequenced transcripts (both” and” adapters and poly-A tail should be present to consider a transcript as complete). Complete transcripts were then mapped to the assembled genome with minimap2 [47, 48] with options “-ax splice:hq -uf”. In parallel, short-read quality was verified using FASTQC software [111] and Illumina’s adapter sequences were removed along with low-quality bases using Trimommatic [46], with options “"ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:5:20 MINLEN:2″". The cleaned reads were then mapped to the genome using STAR [49] with the suggested options for genome assembly –-genomeChrBinNbits 13–-genomeSAindexNbases 13″. Alignment quality was measured with Bamtools v2.5.2 [112]. Transcripts were then characterized using both types of reads with StringTie2 v2.2.1 in –-mix” mode [50] plus the –-conservative” option. Transcriptome completeness was evaluated with BUSCO [53] using Mollusca as the reference. Nuclear coding sequences longer than 150 bases (50 amino acids) were predicted with the utility tool TransDecoder.Predict v5.5 of the Trinity platform [51], with the option–-single_best_only, making use of homology information from UniProt’s Swiss-Prot [113] and Pfam v35 protein domains [114]. Similarity searches were done with BLASTp, with the “-evalue 1e-5″ parameter [115] and hmmscan program with default parameters [116].

Functional annotation

Predicted proteins determined by nuclear coding sequences of TransDecoder.Predict v5.5 [51] were functionally annotated using G-mapper v2.0 (database downloaded 7/22/2022 [71, 72]) for Gene Ontology (GO Terms) assignation based on curated orthogroups, and InterProScan v5.56–89.0 [52] for protein identification based on domain prediction.

Repeat identification and masking of nuclear genome

Repetitive elements were annotated and masked using Earl Grey v 1.2 [117] pipeline. In short, repeat elements were identified with RepeatMasker v4.1.2 [118] with the “-norna, -nolow, -s” option and using Dfam v3.6 [119] curated database for the group Eumetazoa as reference. Then, RepeatModeler v2.0.3 [120], RECON v1.08 [121] and RepeatScout v1.0.5 [122] generated a de novo repeat library that was refined by a “BLAST, Extract and Extend” process to combine fragmented detections into a single repeat candidate. These repeats were finally used by a second run of RepeatMasker v4.1.2 [118] to identify novel repeat elements. Summarizing plots were generated with ggplot2 [123].

Non-coding RNA annotation

Transfer RNAs (tRNAs) and ribosomal genes (rRNA) were predicted using tRNAscan-SE v2.0.9 [124] and Barrnap v0.9 (github.com/tseemann/barrnap), respectively, options for both configured for eukaryotes.

Mitochondrial Genome Annotation

The contig encoding the mitochondrial genome was identified by BLAST searches of the genome assembly (-perc_identity 80) [115] against the known mitochondrial protein-coding genes from *B. glabrata* (NC_005439.1 [60]). Once retrieved, the DNA sequence (contig 7934) was independently annotated with the MITOS2 [61] web server (mitos2.bioinf.uni-leipzig.de/index.py: accessed July 2022). Genes not automatically annotated by mitos2 (*nad4l* and trnK, see Results) were localized with BLASTn [115] searches with respective *B. glabrata* mitochondrial genes (BLASTn with options -task blastn and -task blastn-short for *nad4l* and trnK, respectively).

RNA PacBio reads were mapped to the mitochondrial genome following the same procedure as for the nuclear genome annotation (see above). Transcript alignments to each mitochondrial gene were inspected on Integrative Genomics Viewer (IGV) [125] to manually improve gene annotation making the assumptions that a) translation started on the first viable start codon and b) stop codons were complete unless read coverage suggested a premature RNA end with a truncated stop codon, similar to the assumptions described previously [59]. Reads belonging to intermediary pre-mRNA were counted with htseq-count [126] using the options –-nonunique = all–--samout” and custom scripts (see: github.com/J-Calvelo/Annotation-Biomphalaria-sudanica). Using the pre-mRNA read data, a schematic representation of the post-transcriptional trimming/modification processes of the primary transcript was then manually generated using Inkscape (inkscape.org).

In line with other already published mitochondrial genomes for *B. sudanica* (NCBI RefSeq: NC_038060.1 [56]) and other *Biomphalaria* species (NCBI RefSeq: NC_038061.1 and NC_038059.1, GenBank Accession: MG431965.1 [56]) the start coordinate for the mitochondrial genome was set to the first codon of the *nad5* gene (Supplementary Fig. 4). However, for manual inspection of read alignments on IGV, a custom origin set to position 7222 (between two tRNA molecules, tRNA-s1 and tRNA-s2, lacking both gene predictions and read coverage from the PacBio RNA sequence dataset) was used.

Gene family analysis and evolutionary position of *Biomphalaria sudanica*

A general study of the evolutionary dynamics (copy number and selection pressure) of *Biomphalaria* protein families was carried out based on the longest proteins predicted for each gene, in combination with three *B. glabrata* strains: BB02 (NCBI RefSeq: GCF_000457365.2 [5]) iBS90 and iM (GenBank Accession: GCA_025434165.1 and GCA_025434175.1 [6]), and the species *B. pfeifferi* (NCBI BioProject Accession: GCA_030265305.1 [8]), *B. straminea* (GenBank Accession: GCA_021533235.1 [7]), with the planorbid snail *Bulinus truncatus* (GenBank Accession: GCA_021962125.1 [127]), and the Plakobranchidae *Elysia marginata* (GenBank Accession: GCA_019649035.1 [128]) as an outgroup. With the goal of homogenizing criteria and avoiding discrepancies between the reported mRNA sequence and their proteins, protein sequences taken from other works were predicted from the reported cDNA using getorfs from the EMBOSS v6.6.0.0 package [129]. The longest ORF among all isoforms (min size 150 bases) of each gene were selected, ties were broken by picking the most upstream candidates for each gene. Orthology relationships were estimated in HOGs with Orthofinder v2.5.4 [68] and treated as putative protein families. A species tree was generated in Orthofinder using the Species Tree from All Genes (STAG) algorithm [68, 75], to confirm that the topology of the tree matches those previously reported for these *Biomphalaria* species [10, 130].

Significant expansions/contractions above the background were identified withIE 5 [69]. The Enriched GO terms among significantly expanded/contracted HOGs I CAFE 5 were determined using the topGO R package v2.48.0 [73], with the Weight01 algorithm, node size = 10 and statistical test of Fisher (significant p-value ≤ 0.05). To this end, GO terms were assigned to each species' genes as described above, and those annotations were assigned to each HOG. Summary plots of enriched GO term lists were produced in REVIGO to allow further interpretation [74] based on the enrichment GO term p-value, using default parameters, and a small (0.5) result list.

Cellular location of proteins: secreted, mitochondrial translocated and transmembrane proteins

Location signals for exportation or mitochondrial translocation of proteins were predicted with SignalP v6.0 [62] and TargetP v2.0 [63], respectively. Isoforms with negative results for both tools were additionally analyzed with SecretomeP v1.0 [64] to identify additional secreted proteins based on their biochemical characteristics, that is, proteins exported through an alternative route or with location signals missed due to annotation errors. For all three analyses only hits with a probability/score above 0.95 were considered significant. Lastly, transmembrane domains (TMDs) were predicted with DeepTMHMM v1.0.13 [131].

VIgL (FREP/CREP/GREP) identification and analysis

Based on the presence or absence of known domains and features of FREPs, CREPs and GREPs, as reviewed previously [44], these VIgLs were identified following these criteria: 1) Evidence of secretion (see section: *Cellular location of proteins*), 2) presence of IgSF domains, which were identified using hmmsearch v3.3.2 [76] using the domain profiles generated previously [44], and 3) evidence of their respective 3' terminal domain in the InterPro annotation: FBD (IPR002181, IPR014716, IPR020837, IPR036056), C-lectin (IDs: IPR001304, IPR016186 IPR016187, IPR018378) or Galectin (IPR001079, IPR015533, IPR044156, IPR000922, IPR043159). For a summary of the expected hits for each family, see Dheilly et al. [77]. Lastly, a BLAST search between *B. sudanica* transcriptome and the *B. glabrata* FREP, CREP, and GREP genes reported by Lu et al. [44] and an additional

17 reference sequences (13 FREP and 4 CREP from NCBI (Supplementary Table 15) were performed.

Full FREP, CREP, and GREP candidates were defined as genes with a signal peptide (SP), at least one IgSF, and their respective 3' terminal signature domain (FBD, C-lectin, or Galectin domain respectively). Classification into subfamilies was carried out according to their phylogenetic relationship with the reference sequences. For each candidate gene the longest isoform reported for each gene and reported as a full candidate for each family were aligned in MAFFT v7.310 [132] and positions with more than 20% gaps removed with trimal v1.4.rev22 [133] (options -gt 0.8 -st 0). Then their phylogenetic relationships were estimated by maximum likelihood with IQ-TREE v.2.2.0.3 [134]. The best substitution model was selected by ModelFinder [135] using the Bayesian Information Criterion, by setting "-m MFP" as part of IQ-TREE options. The DNA substitution models VT + I + G4 and JTT + F + R6 models were selected as the best fitting for the CREP and FREP sequence alignments, respectively. Node support was estimated with 1000 replicates of non-parametric bootstrap using IQ-TREE options "-b 1000". Trees were re-rooted at the midpoint and inspected in iTOL v6 [136].

Identification of previously identified schistosome resistance genes

Several candidate immune genes have been identified in *B. glabrata* that could be involved in the response of snails to *S. mansoni*, inferring resistance (Supplementary Table 2). To test whether polymorphisms were shared within *B. sudanica* and between species of *Biomphalaria* in some of these candidate PRRs and immune genes that show the strongest support (BSUD.12903, *PTC1*, *PTC2*, *RADres*), phylogenies were generated using RAxML with -m GTRCAT [82], incorporating published sequences from the other species [6–8, 42].

Assessment of highly diverse genes and genome regions for novel pathogen recognition receptors

Intraspecific genomic diversity between the five *B. sudanica* inbred lines was determined by interrogating polymorphisms between the five inbred lines vcf file (see above) using statistical measures in vcftools v0.1.17 [109]. High-diversity genes were determined by calculating nucleotide diversity across both the entire gene (including noncoding regions such as introns and UTRs) and in coding regions only, determined by the Bs111 nuclear genome annotation, of which the top 1% were selected for further analysis (Supplementary Fig. 5). In addition, we identified genes occurring in or near diverse windows, as follows. We calculated nucleotide diversity across the genome in sliding windows of 10 kb (starting every 2.5 kb), 30 kb, (starting every 7.5 kb) or 100 kb (starting every 25 kb), and identified the top 0.1% of 10 kb windows, top 0.3% of 30 kb windows, and top 1% of 100 kb windows. We then identified all genes that occur within (or overlap partially with) 100 kb of the midpoint of any of these diverse windows.

For all unique genes identified following these criteria, the largest protein-coding amino acid sequence for each gene was summarized from the annotation (Supplementary File 7). Where available (those with matches), transcripts of coding regions from each gene (transcript determined from Bs111 RNAseq data) were matched with their annotated UniProt description (uniprot.org/, database as of May 25th, 2022) and InterPro v5.56–89.0 description [52] (Supplementary Table 19 and Supplementary Table 20). All amino acid sequences were further characterized by searching for orthologous proteins using BLASTp [115] on the NCBI protein database (ncbi.nlm.nih.gov/protein, database as of October 1st, 2022) (Supplementary Table 21). An e-value cut-off of 1e-50 was used for amino acid sequence matches. The proteins producing the most significant alignments (based on the lowest e-value and highest percentage identity from BLASTp), and those predicted using UniProt and Pfam to the query amino acid sequence were recorded and used to determine each peptide's function/biological process and key protein domains if this information was available and informative.

Based on the annotated protein, functional domain presence and structure (i.e. presence of TMDs), each protein was then placed in immune and non-immune gene-related categories, such as in similar analyses [137, 138], and placed under the broader groups of: 1) non-immune function suspected; 2) immune-related function; 3) potentially immune-related function; 4) unknown protein function but containing TMD(s); 5) unknown protein function because sufficient information could not be obtained and without TMD. To create a shortlist ($n = 20$) of candidate PRRs under balancing selection, proteins with the highest nucleotide diversity, and categorized as either in group 2, 3 or 4 that also contained at least one transmembrane domain were shortlisted. The genes and in some cases the genomic regions surrounding these were assessed in more detail, including assessing positions of functional domains and intra/extracellular regions relative to nucleotide diversity (Supplementary Fig. 5).

To determine if any enriched GO terms, particularly those associated with immunity, were amongst the proteins identified as potential PRRs, topGO v2.48.0 [73] in R v4.3.1 [139] with the Weight01 algorithm, node size = 10 and statistical test of Fisher (significant p-value ≤ 0.05) was used. The test was performed by comparing the *B.*

sudanica whole genome gene family composition to three separate lists of the highly diverse genes identified: 1) all 1047 highly diverse genes identified in the *B. sudanica* genome (as listed Supplementary Table 17), 2) 245 genes classified as immune suspected genes from the highly diverse genes (group 2 in Supplementary Table 17), and 3) 242 genes shortlisted as the protein containing a TMD and categorized in group 2, 3 (potential role in innate immunity) or 4 (an unknown function but contained TMD(s)) (Supplementary Table 17).

Inferring linkage groups and synteny with *Biomphalaria glabrata* linkage groups

We compared synteny and orthology with *B. glabrata* using the iM assembly, which is highly contiguous with an N50 of 22.7 Mb [6]. We defined orthologous genes as reciprocal best BLASTp hits of protein sequences. We defined orthologous contigs/scaffolds as those sharing the most orthologous genes. For *B. sudanica* contigs/scaffolds that were orthologous to scaffolds mapped in the 18 *B. glabrata* iM linkage groups (LGs), we assigned them to 18 LGs corresponding to the LGs and ordered them to mirror the orthologous order. A single iM scaffold (BGM014) is split between linkage groups, so we considered the distinctly-mapped portions of it separately. For LGs of particular interest, we generated dot plots of sequence similarity based on our previous approach [140]. We used these to refine the order of contigs/scaffolds, to identify contigs/scaffolds with little sequence similarity despite possessing nominally orthologous genes, and to characterize inversions and other chromosomal rearrangements.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10103-w.

Additional file 1: Supplementary Figure 1. Repetitive elements in the *Biomphalaria sudanica* genome. ^ "Other" indicates a Simple Repeat, Microsatellite, RNA. **Supplementary Figure 2.** Schematic presentation of the mitochondrial gene trimming process for the transcripts, showing regions of primary transcription (Fragment 1-6) on the plus and minus strand, and the trimming processes of these primary transcripts into pre-mRNA. Numbers above transcripts represent the RNA sequence depth from aligned PacBio IsoSeq ccs data. **Supplementary Figure 3.** Dot plots of *Biomphalaria sudanica* linkage groups 6 (A), 10 (B) and 16 (C), composed of multiple scaffolds determined using the *B. glabrata* iM line linkage map (Bu et al., 2022). Dots represent 600bp segments; dark blue is ≥97.5% sequence similarity, light blue is ≥90% sequence similarity. **Supplementary Figure 4.** Mitochondrial genome of *Biomphalaria sudanica* with point of origin set to the start of the *nad5* gene. **Supplementary Figure 5.** Schematic overview of the methods employed to delimit pathogen recognition receptors (PRR) and other immune genes of *Biomphalaria sudanica* under balancing selection, determined through the analysis of high intraspecific genetic diversity regions, potentially relevant to the resistance and susceptibility of this species to *Schistosoma mansoni*.

Additional file 2: Supplementary Table 1. Summary metrics and NCBI identifiers for the generated genome (PacBio long-read) and transcriptome (PacBio long-read and Illumina short-read) sequences, and the complete annotation statistics. ^ tRNA pseudogenes are fragments of genes that were identified and not used for annotation. * Genes models that were predicted by but Stringtie no valid ORF was identified by Transdecoder and therefore not used in the annotation.

Additional file 3: Supplementary Table 2. *Biomphalaria sudanica* orthologues to genes / genomic markers of 18 candidate immune loci that have been identified for their involvement in *Schistosoma mansoni* immunity in *B. glabrata.**The Knight marker, qRS-5.1, and qRS-2.1, truly represent a marker and not a coding gene, hence putative function and mechanism are unknown.

Additional file 4: Supplementary Table 3. Summary table of the 919 tRNA genes that were predicted by tRNAscan-SE 2.0 v2.0.9 [124] representing 24 tRNA types annotated in the *Biomphalaria sudanica* genome.

Additional file 5: Supplementary Table 4. Complete list showing the genome positions and prediction scores of 919 tRNA genes predicted in the *Biomphalaria sudanica* genome by tRNAscan-SE 2.0 v2.0.9 [124].

Additional file 6: Supplementary Table 5. Summary table of repeated elements composition of the *Biomphalaria sudanica* genome.

Additional file 7: Supplementary Table 6. Predicted coding genes that overlap with repeated elements. For each one it is specified the contig/scaffold, gene and coordinates, and each overlapping repeated element and the features they compromise: Intron region, 5'UTR, 3'UTR and/or CDS.

Additional file 8: Supplementary Table 7. Mitochondrial genome annotation of *Biomphalaria sudanica* and RNAseq read counts from the Illumina short-read datasets. For each feature, detail is given on the coordinates, encoding strand, source of the annotation and if their boundaries were manually adjusted based on the read mappings. The starting coordinate was selected to be the start of *nad5* in order to be consistent with the already published mitochondrial genomes of *Biomphalaria* [56].

Additional file 9: Supplementary Table 8. Summary table of the predicted location signals found in the *Biomphalaria sudanica* protein coding genes using the programs SignalP v6.0 [62] for the identification of Signal Peptides (SP), TargetP v2.0 [63] for the identification of mitochondrial transit peptide (mTP) and SecretomeP v1.0 [64] for proteins putatively secreted through other pathways. In addition to the total number of trans-membrane domains predicted by DeepTMHMM v1.0.13 [131] and the identified InterPro v5.56-89.0 [52] signals. Each isoform was putatively classified into 4 types: Secreted (SP identified or positive result with SecretomeP and no trans-membrane domains), Mitochondrial Cytoplasmatic (mTP identified and no trans-membrane domains), Membrane (SP identified or positive result with SecretomeP and at least one trans-membrane domain), Mitochondrial Membrane (mTP identified and at least one trans-membrane domain). When either an SP or a mTP was identified, its location and probability is included in the column Location Signal Details.

Additional file 10: Supplementary Table 9. Phylogenetic Hierarchical Orthogroups (HOG) estimated by Orthofinder [68]. For each HOG, detail is provided on their ID, the original Orthogroup ID and the node in the gene tree from which the HOG was determined together with the protein ID for the eight genomes analyzed: *B. glabrata* strain BB02 (GCF_000457365.2 [5]), *B. glabrata* strain iBS90 (GCA_025434165.1 [6]), *B. glabrata* strain iM (GCA_025434175.1 [6]), and the species *B. pfeifferi* (GCA_030265305.1 [8]), *B. straminea* (GCA_021533235.1 [7]), with the planorbid snail *Bulinus truncatus* (GCA_021962125.1 [127]) and the Plakobranchidae *Elysia marginata* (GCA_019649035.1 [128]).

Additional file 11: Supplementary Table 10. Gene Ontology (GO) terms assigned to each Phylogenetic Hierarchical Orthogroups (HOG) by eggNOG-mapper [71, 72].

Additional file 12: Supplementary Table 11. Enriched gene ontology (GO) terms found in the protein families that significantly expanded or contracted in the CAFE 5 [69] analysis. At each node it is specified the IDs and total count of the Phylogenetic Hierarchical Orthogroups (HOG) identified to be either expanded or contracted, followed by the enrichment results produced by v2.48.0 [73]: the GO terms enriched, the total number of annotated HOGs with each GO terms, how many GO terms were observed, and the total number of GO terms expected given the list size and the p-value.

Additional file 13: Supplementary Table 12. Summary of the enriched gene ontology (GO) terms found in the protein families that significantly expanded or contracted in the CAFE 5 [69] analysis. For each group in the REVIGO [74] TreeMap summary (see Supplementary File 5) it is specified their member GO terms ID description, and the semantic similarity measurements calculated by REVIGO: Frequency of the term in the UniProt database (uniport.org/, database as of May 25th 2022), its semantical Uniqueness in the whole list and their Dispensability if the term was nested with another GO term. In the latter case (dispensable terms), the GO term number it is nested with is reported, or displayed as "-" otherwise.

Additional file 14: Supplementary Table 13. *Biomphalaria sudanica* protein domain location and Signal Peptides (SP) found among the initial candidates identified for the protein families C-type lectin-related protein (CREP), Fibrinogen-related protein (FREP) or galectin-related protein (GREP), determined by the presence of C-type lectin, fibrinogen (FBD) and Galectin like domains, respectively. For each examined protein it is detailed to which of the three families it is considered a candidate for, and the details of each domain and SP, the analysis that identified it, the signature accession, score (e-value or probability depending on the analysis, domain coordinates, InterPro v5.56-89.0 [52] annotation and description (if applicable), and what type of signature it was classified as: C-type lectin like domain, FBD like domain, Galectin like domain, IgSF like domain, other immunoglobulin like domain or a secreted protein. If the protein was utilized in the phylogenetic analysis it is indicated with an "X" in the Final Selection column, and "" otherwise.

Additional file 15: Supplementary Table 14. Summary table of *Biomphalaria sudanica* C-type lectin-related proteins (CREP) and Fibrinogen-related proteins (FREP) annotated, providing protein ID, CREP/FREP subfamily, protein length, signal peptide, immunoglobulin domain (IgSF) and the c-lectin domain (CREP) or fibrinogen (FBD) domain (FREP) positions, notes on each genes features and protein sequence. Coordinates given in the notes column specify the overlap of any remarkable features with domains C-lectin, FBD or IgSF.

Additional file 16: Supplementary Table 15. Reference sequences of Fibrinogen-related protein (FREP) and C-type lectin-related protein (CREP) included in the phylogenetic analysis for both these protein families. Table details the sequence ID, source (reference), identifier provided within the family, length in amino acids and sequence. Sequences marked with an X at the end of their Sequence ID were used with the modifications detailed in Lu et al. (2020), rather than the original genome.

Additional file 17: Supplementary Table 16. Total number of genes included on each of the Fibrinogen-related protein (FREP) and C-type lectin-related protein (CREP) Phylogenetically Hierarchical Orthogroups (HOGs) across the analyzed *Biomphalaria* speciesgenomes and outgroups (*B. glabrata* strain BB02 (GCF_000457365.2 [5]), *B. glabrata* strain iBS90 (GCA_025434165.1 [6]), *B. glabrata* strain iM (GCA_025434175.1 [6]), and the species *B. pfeifferi* (GCA_030265305.1 [8]), *B. straminea* (GCA_021533235.1 [7]), *Bulinus truncatus* (GCA_021962125.1 [127]) and *Elysia marginata* (GCA_019649035.1 [128])). For *B. sudanica*, we specify both the raw counts (i.e. 'Raw', all the members of the HOG) and how many of these were selected for the phylogenetic analysis (i.e. 'Selected').

Additional file 18: Supplementary Table 17. List of 1047 genes that were present in the most diverse regions of the *Biomphalaria sudanica* genome, providing information on genomic position, BLASTp hits, InterPro v5.56-89.0 [52] and Pfam v35 [114] domains and inferred protein function categorized into immune groups (as detailed in Methods).

Additional file 19: Supplementary Table 18. Results of *Biomphalaria sudanica* gene family expansion and contraction analysis performed using topGO v2.48.0 [73] in R v4.3.1 [139] with the Weight01 algorithm, node size=10 and statistical test of Fisher (significant p-value ≤ 0.05). The test was performed by comparing the *B. sudanica* whole genome gene family composition to three separate lists (column=List) of the highly diverse genes identified; 1) 'All' 1047 highly diverse genes identified in the *B. sudanica* genome (as listed Supplementary Table 17), 2) 'Group-2' 245 genes classified as immune suspected genes from the highly diverse genes (group 2, Supplementary Table 17), and 3) 'Group-2-3-4' 247 genes shortlisted as the protein containing a transmembrane domain (TMD) and categorized in Immune Group 2, 3 (potential role in innate immunity) or 4 (an unknown function but contained TMD(s)) (Supplementary Table 17). Each group contains three tabs, showing gene family expansions and contractions in each set of genes in respects to their Molecular Function (MF), Cellular Component (CC) and Biological Process (BP).

Additional file 20: Supplementary Table 19. Pfam v35 [114] matches for function domains the 1047 most diverse genes identified in the *Biomphalaria sudanica* genome.

Additional file 21: Supplementary Table 20. UniProt (uniport.org/ database as of May 25th 2022) matches for the 1047 most diverse genes identified in the *Biomphalaria sudanica* genome.

Additional file 22: Supplementary Table 21. BLASTp matches for the 1047 most diverse genes identified in the *Biomphalaria sudanica* genome.

Additional file 23: Supplementary File 1. Complete functional annotation of *Biomphalaria sudanica*, including 23,598 genes assigned to have an open reading frame (ORF) and therefore predicted protein by TransDecoder.Predict v5.5.0 using the Trinity platform [51], the 2,248 raw RNA genes without predicted proteins predicted by StringTie2 v2.2.1 [50], 919 tRNA's predicted in tRNAscan-SE 2.0 v2.0.9 [124], and 107 rRNA's predicted in Barrnap v0.9 (github.com/tseemann/barrnap).

Additional file 24: Supplementary File 2. Summary table of the 23,598 genes assigned to have an open reading frame (ORF) giving ORF type, length (amino acid) and prediction score from TransDecoder.Predict v5.5.0 using the Trinity platform [51].

Additional file 25: Supplementary File 3. Summary InterProScan v5.56-89.0 [52] results for the predicted proteins. It includes the specific analysis, the signature accession and description identified, the start and stop location of the signature, its score, and associated InterPro Accession information.

Additional file 26: Supplementary File 4. Location of transmembrane domains identified on 4,922 genes (8,728 isoforms) determined through DeepTMHMM v1.0.13 [131].

Additional file 27: Supplementary File 5. File providing the REVIGO [74] TreeMap summaries of the GO terms enriched among the protein families, showing significant expansions and contractions for each node (excluding outgroups) of the species tree generated in Orthofinder using the Species Tree of All Genes (STAG) algorithm [68, 75].

Additional file 28: Supplementary File 6. Description of RNA extraction methods, quantification, and pooling for sequencing.

Additional file 29: Supplementary File 7. Amino acid sequences of the 1047 shortlisted most diverse genes in the *Biomphalaria sudanica* genome, used for searches for novel pathogen recognition receptors.

Acknowledgements

The NMRI schistosomes were provided by the NIAID Schistosomiasis Resource Center of the Biomedical Research Institute (Rockville, MD) through NIH-NIAID Contract HHSN272201700014I. We also acknowledge Hannah Tavalire for her work in developing the inbred lines of *B. sudanica*. We thank the time spent by three anonymous reviewers, whose insightful comments significantly improved this manuscript.

Authors' contributions

TP, JCal., JAT, SRB and MLS collected, analyzed and interpreted the majority of the data contributing to this manuscript. TP, JCal., JAT, AI and MLS were major contributors to writing the initial manuscript. JCal. developed the bioinformatic pipeline for the annotation of the genome. JAT developed and conceptualized the bioinformatic pipelines for the analysis and description of hyperdiverse genes under balancing selection. RB and JCay. collected and analyzed data regarding the hyperdiverse immune genes in the genome. SRB and MSB generated and assembled/aligned whole genome sequence data

from the inbred lines and assisted in generating transcriptome sequence data. MLS, TP, JMS, assisted in the maintenance and development of inbred snail lines. FGH, GO, FR, kA, BM, MOdh., and MOdi. assisted in manuscript preparation and editing. MOdi assisted in project planning. ESL, MRL and LL assisted in the comparative work with African *Biomphalaria* species, and LL assisted with the FREPs identification pipeline and analysis. MLS and MOdi. were responsible for funding acquisition to complete this project. All authors reviewed and approved the final manuscript.

Funding

Funding for this project was provided by National Institutes of Health, National Institute of Allergy and Infectious Disease R01AI141862, which supported TP, JAT, GO, FR, kA, BM, MOdh., MOdi. and MLS. ESL was funded by NIH grant R37AI101438.'

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the NCBI BioProject PRJNA1041389.

The bioinformatic pipeline for the genome annotation is available at github.com/J-Calvelo/Annotation-Biomphalaria-sudanica.

Declarations

Ethics approval and consent to participate

This project was undertaken following approval from the relevant bodies, including Kenya Medical Research Institute (KEMRI) Scientific Review Unit (Approval # KEMRI/RES/7/3/1 and KEMRI/SERU/CGHR/035/3864), Kenya's National Commission for Science, Technology, and Innovation (License # NACOSTI/P/15/9609/4270 and NACOSTI/P/22/14839), Kenya Wild- life Services (permit # 0004754 and # WRTI-0136–02-22), and National Environment, Management Authority (permit # NEMA/AGR/46/2014 – Registration # 0178 and NEMA/AGR/159/2022 – Registration # 201).

Consent for publication

Not applicable.

Competing interests

Not applicable.

Author details

[1]College of Osteopathic Medicine of the Pacific – Northwest, Western University of Health Sciences, Lebanon, OR, USA. [2]Laboratorio de Biología Computacional, Departamento de Desarrollo Biotecnológico, Facultad de Medicina, Instituto de Higiene, Universidad de La República, Montevideo 11600, Uruguay. [3]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [4]Oregon State University, Corvallis, OR, USA. [5]Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University, Starkville, MS, USA. [6]Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi State, MS, USA. [7]Centre for Global Health Research, Kenya Medical Research Institute (KEMRI), P. O. Box 1578-40100, Kisumu, Kenya. [8]Center for Evolutionary and Theoretical Immunology, Parasite Division Museum of Southwestern Biology, Department of Biology, University of New Mexico, Albuquerque, NM, USA.

Received: 7 November 2023 Accepted: 8 February 2024
Published online: 19 February 2024

References

1. Hotez PJ, Daar AS. The CNCDs and the NTDs: Blurring the Lines Dividing Noncommunicable and Communicable Chronic Diseases. PLoS Negl Trop Dis. 2008;2(10):e312.
2. WHO. Schistosomiasis: Key facts. 2023 [cited 2023 Apr 14]. Available from: https://www.who.int/news-room/fact-sheets/detail/schistosomiasis.
3. Castillo MG, Humphries JE, Mourão MM, Marquez J, Gonzalez A, Montelongo CE. *Biomphalaria glabrata* immunity: Post-genome advances. Dev Comp Immunol. 2020;104:103557.
4. Mitta G, Gourbal B, Grunau C, Knight M, Bridger JM, Théron A. The Compatibility Between *Biomphalaria glabrata* Snails and *Schistosoma mansoni*: An Increasingly Complex Puzzle. Adv Parasitol. 2017;97:111–45.
5. Adema CM, Hillier LW, Jones CS, Loker ES, Knight M, Minx P, et al. Whole genome analysis of a schistosomiasis-transmitting freshwater snail. Nat Commun. 2017;8:15451.
6. Bu L, Zhong D, Lu L, Loker ES, Yan G, Zhang S-M. Compatibility between snails and schistosomes: insights from new genetic resources, comparative genomics, and genetic mapping. Commun Biol. 2022;5(1):940.
7. Nong W, Yu Y, Aase-Remedios ME, Xie Y, So WL, Li Y, et al. Genome of the ramshorn snail *Biomphalaria straminea*—an obligate intermediate host of schistosomiasis. Gigascience. 2022;11:giac012.
8. Bu L, Lu L, Laidemitt MR, Zhang S-M, Mutuku M, Mkoji G, et al. A genome sequence for *Biomphalaria pfeifferi*, the major vector snail for the human-infecting parasite *Schistosoma mansoni*. PLoS Negl Trop Dis. 2023;17(3):e0011208.
9. Hotez PJ, Kamath A. Neglected Tropical Diseases in Sub-Saharan Africa: Review of Their Prevalence, Distribution, and Disease Burden. PLoS Negl Trop Dis. 2009;3(8):1–10.
10. DeJong RJ, Morgan JAT, Paraense WL, Pointier J-P, Amarista M, Ayeh-Kumi PFK, 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. 2001;18(12):2225–39.
11. Campbell G, Jones CS, Lockyer AE, Hughes S, Brown D, Noble LR, et al. Molecular evidence supports an African affinity of the Neotropical freshwater gastropod, *Biomphalaria glabrata*, Say 1818, an intermediate host for *Schistosoma mansoni*. Proc R Soc London Ser B Biol Sci. 2000;267(1460):2351–8.
12. Woodruff DS, Mulvey M. Neotropical schistosomiasis: African affinities of the host snail *Biomphalaria glabrata* (Gastropoda: Planorbidae). Biol J Linn Soc. 1997;60(4):505–16.
13. von Martens E. Conchylien aus dem obern Nilgebiet. Malakozool Blätter. 1870;17:32–6.
14. Brown DS. Freshwater Snails of Africa and their Medical Importance. 2nd ed. London, UK: Taylor & Francis; 1994.
15. Erko B, Balcha F, Kifle D. The ecology of *Biomphalaria sudanica* in Lake Ziway. Ethiopia Afr J Ecol. 2006;44(3):347–52.
16. Kazibwe F, Makanga B, Rubaire-Akiiki C, Ouma J, Kariuki C, Kabatereine NB, et al. Ecology of *Biomphalaria* (Gastropoda: Planorbidae) in Lake Albert, Western Uganda: snail distributions, infection with schistosomes and temporal associations with environmental dynamics. Hydrobiologia. 2006;568:433–44.
17. Williams SN, Hunter PJ. The distribution of *Bulinus* and *Biomphalaria* in Khartoum and Blue Nile Provinces, Sudan. Bull World Health Organ. 1968;39(6):949.
18. Loker ES, Moyo HG, Gardner SL. Trematode-gastropod associations in nine non-lacustrine habitats in the Mwanza region of Tanzania. Parasitology. 1981;83(2):381–99.
19. Magendantz M. The biology of Biomphalaria choanomphala and *B. sudanica* in relation to their role in the transmission of *Schistosoma mansoni* in Lake Victoria at Mwanza, Tanzania. Bull World Health Organ. 1972;47(3):331.
20. Platt RN, Le Clec'h W, Chevalier FD, McDew-White M, LoVerde PT, Ramiro de Assis R. Genomic analysis of a parasite invasion: Colonization of the Americas by the blood fluke *Schistosoma mansoni*. Mol Ecol. 2022;31(8):2242–63.
21. Morgan JAT, DeJong RJ, Adeoye GO, Ansa EDO, Barbosa CS, Brémond P, et al. Origin and diversification of the human parasite *Schistosoma mansoni*. Mol Ecol. 2005;14(12):3889–902.
22. King CH, Binder S, Shen Y, Whalen CC, Campbell CH, Wiegand RE, et al. SCORE Studies on the Impact of Drug Treatment on Morbidity due to *Schistosoma mansoni* and *Schistosoma haematobium* Infection. Am J Trop Med Hyg. 2020;103(1):30–5.
23. Secor WE, Wiegand RE, Montgomery SP, Karanja DMS, Odiere MR. Comparison of school-based and community-wide mass drug administration for schistosomiasis control in an area of western Kenya with high

initial *Schistosoma mansoni* infection prevalence: a cluster randomized trial. Am J Trop Med Hyg. 2020;102(2):318.

24. Wiegand RE, Mwinzi PNM, Montgomery SP, Chan YL, Andiego K, Omedo M, et al. A persistent hotspot of *Schistosoma mansoni* infection in a five-year randomized trial of praziquantel preventative chemotherapy strategies. J Infect Dis. 2017;216(11):1425–33.
25. Gouvras AN, Allan F, Kinung'hi S, et al. Longitudinal survey on the distribution of *Biomphalaria sudanica* and *B. choanomophala* in Mwanza region, on the shores of Lake Victoria, Tanzania: implications for schistosomiasis transmission and control. Parasit Vectors. 2017;10:316. https://link.springer.com/article/10.1186/s13071-017-2252-z.
26. Martens E von. Recente Conchylien aus dem Victoria Nianza (Ukerewe). Sitzungsberichte der Gesellschaft Naturforschender Freunde zu Berlin. 1879:103–5. https://www.biodiversitylibrary.org/page/7795591#page/2/mode/1up. https://www.biodiversitylibrary.org/bibliography/7922.
27. Loker ES, Hofkin BV, Mkoji GM, Mungai B, Kihara JH, Koech DK. Distributions of freshwater snails in southern Kenya with implications for the biological control of schistosomiasis and other snail-mediated parasites. J Med Appl Malacol. 1993;5:1–20.
28. Mutuku MW, Laidemitt MR, Beechler BR, Mwangi IN, Otiato FO, Agola EL, et al. A Search for Snail-Related Answers to Explain Differences in Response of *Schistosoma mansoni* to Praziquantel Treatment among Responding and Persistent Hotspot Villages along the Kenyan Shore of Lake Victoria. Am J Trop Med Hyg. 2019;101(1):65–77.
29. Standley CJ, Wade CM, Stothard JR. A fresh insight into transmission of schistosomiasis: a misleading tale of *Biomphalaria* in Lake Victoria. PLoS ONE. 2011;6(10):e26563.
30. Andrus PS, Stothard JR, Kabatereine NB, Wade CM. Comparing shell size and shape with canonical variate analysis of sympatric *Biomphalaria* species within Lake Albert and Lake Victoria, Uganda. Zool J Linn Soc. 2023;199(3):713–22.
31. Standley CJ, Goodacre SL, Wade CM, Stothard JR. The population genetic structure of *Biomphalaria choanomphala* in Lake Victoria, East Africa: implications for schistosomiasis transmission. Parasit Vectors. 2014;7(1):524.
32. Mutuku MW, Laidemitt MR, Spaan JM, Mwangi IN, Ochanda H, Steinauer ML, et al. Comparative vectorial competence of *Biomphalaria sudanica* and *Biomphalaria choanomphala*, snail hosts of *Schistosoma mansoni*, from transmission hotspots In Lake Victoria. Western Kenya J Parasitol. 2021;107(2):349–57.
33. Shultz AJ, Sackton TB. Immune genes are hotspots of shared positive selection across birds and mammals. Elife. 2019;8:e41815.
34. Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, et al. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. Science. 2013;339(6127):1578–82.
35. Takahata N, Satta Y, Klein J. Polymorphism and balancing selection at major histocompatibility complex loci. Genetics. 1992;130(4):925–38.
36. Hughes AL, Yeager M. Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet. 1998;32(1):415–35.
37. Ferrer-Admetlla A, Bosch E, Sikora M, Marquès-Bonet T, Ramírez-Soriano A, Muntasell A, et al. Balancing Selection Is the Main Force Shaping the Evolution of Innate Immunity Genes. J Immunol. 2008;181(2):1315–22.
38. Chisholm ST, Coaker G, Day B, Staskawicz BJ. Host-Microbe Interactions: Shaping the Evolution of the Plant Immune Response. Cell. 2006;124(4):803–14.
39. Van de Weyer A-L, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, et al. A Species-Wide Inventory of NLR Genes and Alleles in *Arabidopsis thaliana*. Cell. 2019;178(5):1260-1272.e14.
40. Lee D, Zdraljevic S, Stevens L, Wang Y, Tanny RE, Crombie TA, et al. Balancing selection maintains hyper-divergent haplotypes in *Caenorhabditis elegans*. Nat Ecol Evol. 2021;5(6):794–807.
41. Tennessen JA, Théron A, Marine M, Yeh J-Y, Rognon A, Blouin MS. Hyperdiverse Gene Cluster in Snail Host Conveys Resistance to Human Schistosome Parasites. PLoS Genet. 2015;11(3):1–21.
42. Tennessen JA, Bollmann SR, Peremyslova E, Kronmiller BA, Sergi C, Hamali B, et al. Clusters of polymorphic transmembrane genes control resistance to schistosomes in snail vectors. Elife. 2020;9:e59395.
43. Adema CM, Hertel LA, Miller RD, Loker ES. A family of fibrinogen-related proteins that precipitates parasite-derived molecules is produced by an invertebrate after infection. Proc Natl Acad Sci. 1997;94(16):8691–6.
44. Lu L, Loker ES, Adema CM, Zhang S-M, Bu L. Genomic and transcriptional analysis of genes containing fibrinogen and IgSF domains in the schistosome vector *Biomphalaria glabrata*, with emphasis on the differential responses of snails susceptible or resistant to *Schistosoma mansoni*. PLoS Negl Trop Dis. 2020;14(10):1–35.
45 Fukasawa Y, Ermini L, Wang H, Carty K, Cheung M-S. Long QC: A Quality Control Tool for Third Generation Sequencing Long Read Data. G3. 2020;10(4):1193–6.
46. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
47. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
48. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37(23):4572–4.
49. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
50. Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLOS Comput Biol. 2022;18(6):1–18.
51. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494–512.
52. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res. 2023;51(D1):D418–27.
53. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: Assessing Genomic Data Quality and Beyond. Curr Protoc. 2021;1(12):e323.
54. Serrão VHB, Silva IR, da Silva MTA, Scortecci JF, de Freitas FA, Thiemann OH. The unique tRNASec and its role in selenocysteine biosynthesis. Amino Acids. 2018;50(9):1145–67.
55. Commans S, Böck A. Selenocysteine inserting tRNAs: an overview. FEMS Microbiol Rev. 1999;23(3):335–51.
56. Zhang S-M, Bu L, Laidemitt MR, Lu L, Mutuku MW, Mkoji GM, et al. Complete mitochondrial and rDNA complex sequences of important vector species of *Biomphalaria*, obligatory hosts of the human-infecting blood fluke. Schistosoma mansoni Sci Rep. 2018;8(1):1–10.
57. Grande C, Templado J, Zardoya R. Evolution of gastropod mitochondrial genome arrangements. BMC Evol Biol. 2008;8(1):61.
58. Ojala D, Montoya J, Attardi G. tRNA punctuation model of RNA processing in human mitochondria. Nature. 1981;290(5806):470–4.
59. Ghiselli F, Gomes-dos-Santos A, Adema CM, Lopes-Lima M, Sharbrough J, Boore JL. Molluscan mitochondrial genomes break the rules. Philos Trans R Soc B Biol Sci. 1825;2021(376):20200159.
60. DeJong RJ, Emery AM, Adema CM. The mitochondrial genome of *Biomphalaria glabrata* (Gastropoda: Basommatophora), intermediate host of *Schistosoma mansoni*. J Parasitol. 2004;90(5):991–8.
61. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: Improved *de novo* metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013;69(2):313–9.
62. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol. 2022;40(7):1023–5.
63. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance. 2019;2(5):e201900429.
64. Bendtsen JD, Jensen LJ, Blom N, von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel. 2004;17(4):349–56.
65. Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, et al. The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad Sci. 2007;104(13):5495–500.
66. Wang T, Wyeth RC, Liang D, Bose U, Ni G, McManus DP, et al. A *Biomphalaria glabrata* peptide that stimulates significant behaviour modifications in aquatic free-living *Schistosoma mansoni* miracidia. PLoS Negl Trop Dis. 2019;13(1):e0006948.
67. Fogarty CE, Phan P, Duke MG, McManus DP, Wyeth RC, Cummins SF, et al. Identification of *Schistosoma mansoni* miracidia attractant candidates in infected *Biomphalaria glabrata* using behaviour-guided comparative proteomics. Front Immunol. 2022;13:954282.

68. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
69. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 2020;36(22–23):5516–8.
70. Pickford M. Freshwater and terrestrial Mollusca from the Early Miocene deposits of the northern Sperrgebiet. Namibia Mem Geol Surv Namibia. 2008;20:65–74.
71. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38(12):5825–9.
72 Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):D309-14.
73. Alexa A, Rahnenfuhrer J. topGO: enrichment analysis for gene ontology. R package version 2.48.0. 2022. Available at: https://bioconductor.org/packages/topGO.
74. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS ONE. 2011;6(7):e21800.
75. Emms DM, Kelly S. STAG: Species Tree Inference from All Genes. bioRxiv. 2018;267914. https://doi.org/10.1101/267914.
76. Eddy SR. Accelerated Profile HMM Searches. PLOS Comput Biol. 2011;7(10):e1002195.
77. Dheilly NM, Duval D, Mouahid G, Emans R, Allienne J-F, Galinier R, et al. A family of variable immunoglobulin and lectin domain containing molecules in the snail *Biomphalaria glabrata*. Dev Comp Immunol. 2015;48(1):234–43.
78. Spaan JM, Pennance T, Laidemitt MR, Sims N, Roth J, Lam Y, et al. Multi-strain compatibility polymorphism between a parasite and its snail host, a neglected vector of schistosomiasis in Africa. Curr Res Parasitol Vector-Borne Dis. 2023;3:100120.
79. Tennessen JA, Bonner KM, Bollmann SR, Johnstun JA, Yeh J-Y, Marine M, et al. Genome-Wide Scan and Test of Candidate Genes in the Snail *Biomphalaria glabrata* Reveal New Locus Influencing Resistance to *Schistosoma mansoni*. PLoS Negl Trop Dis. 2015;9(9):1–19.
80. Allan ERO, Tennessen JA, Bollmann SR, Hanington PC, Bayne CJ, Blouin MS. Schistosome infectivity in the snail, *Biomphalaria glabrata* is partially dependent on the expression of Grctm6, a Guadeloupe Resistance Complex protein. PLoS Negl Trop Dis. 2017;11(2):1–15.
81. Goel S, Palmkvist M, Moll K, Joannin N, Lara P, Akhouri RR, et al. RIFINs are adhesins implicated in severe *Plasmodium falciparum* malaria. Nat Med. 2015;21(4):314–7.
82. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
83. Maier T, Wheeler NJ, Namigai EKO, Tycko J, Grewelle RE, Woldeamanuel Y, et al. Gene drives for schistosomiasis transmission control. PLoS Negl Trop Dis. 2019;13(12):1–21.
84. WHO. WHO guideline on control and elimination of human schistosomiasis. 2022.
85 Mutuku MW, Lu L, Otiato FO, Mwangi IN, Kinuthia JM, Maina GM, et al. A Comparison of Kenyan *Biomphalaria pfeifferi* and *B. sudanica* as Vectors for *Schistosoma mansoni*, Including a Discussion of the Need to Better Understand the Effects of Snail Breeding Systems on Transmission. J Parasitol. 2017;103(6):669–76.
86 Lu L, Zhang S-M, Mutuku MW, Mkoji GM, Loker ES. Relative compatibility of *Schistosoma mansoni* with *Biomphalaria sudanica* and *B. pfeifferi* from Kenya as assessed by PCR amplification of the *S. mansoni* ND5 gene in conjunction with traditional methods. Parasit Vectors. 2016;9(1):166.
87. Moné Y, Gourbal B, Duval D, Du Pasquier L, Kieffer-Jaquinod S, Mitta G. A Large Repertoire of Parasite Epitopes Matched by a Large Repertoire of Host Immune Receptors in an Invertebrate Host/Parasite Model. PLoS Negl Trop Dis. 2010;4(9): e813.
88. Larson MK, Bender RC, Bayne CJ. Resistance of *Biomphalaria glabrata* 13–16-R1 snails to *Schistosoma mansoni* PR1 is a function of haemocyte abundance and constitutive levels of specific transcripts in haemocytes. Int J Parasitol. 2014;44(6):343–53.
89. Sun Y, Zhou Z, Wang L, Yang C, Jianga S, Song L. The immunomodulation of a novel tumor necrosis factor (CgTNF-1) in oyster *Crassostrea gigas*. Dev Comp Immunol. 2014;45(2):291–9.
90. Zheng Y, Liu Z, Wang L, Li M, Zhang Y, Zong Y, et al. A novel tumor necrosis factor in the Pacific oyster *Crassostrea gigas* mediates the antibacterial response by triggering the synthesis of lysozyme and nitric oxide. Fish Shellfish Immunol. 2020;98:334–41.
91. Huang Y, Si Q, Du S, Du J, Ren Q. Molecular identification and functional analysis of a tumor necrosis factor superfamily gene from Chinese mitten crab (*Eriocheir sinensis*). Dev Comp Immunol. 2022;134: 104456.
92. Bertevello CR, Russo BRA, Tahira AC, Lopes-Junior EH, DeMarco R, Oliveira KC. The evolution of TNF signaling in platyhelminths suggests the cooptation of TNF receptor in the host-parasite interplay. Parasit Vectors. 2020;13(1):491.
93. Blouin MS, Bonner KM, Cooper B, Amarasinghe V, O'Donnell RP, Bayne CJ. Three genes involved in the oxidative burst are closely linked in the genome of the snail. Biomphalaria glabrata Int J Parasitol. 2013;43(1):51–5.
94 Tennessen JA, Bollmann SR, Blouin MS. A Targeted Capture Linkage Map Anchors the Genome of the Schistosomiasis Vector Snail. Biomphalaria glabrata G3. 2017;7(7):2353–61.
95. Mimpfound R, Greer GJ. Allozyme variation among populations of *Biomphalaria pfeifferi* (Krauss, 1848) (Gastropoda: Planorbidae) in Cameroon. J Molluscan Stud. 1990;56(4):461–7.
96. Tian-Bi Y-N, Jarne P, Konan J-N, Utzinger J, N'Goran EK. Contrasting the distribution of phenotypic and molecular variation in the freshwater snail *Biomphalaria pfeifferi,* the intermediate host of *Schistosoma mansoni*. Heredity (Edinb). 2013;110(5):466–74.
97. Baclaocos J, Santesmasses D, Mariotti M, Bierła K, Vetick MB, Lynch S, et al. Processive Recoding and Metazoan Evolution of Selenoprotein P: Up to 132 UGAs in Molluscs. J Mol Biol. 2019;431(22):4381–407.
98. Budachetri K, Karim S. An insight into the functional role of thioredoxin reductase, a selenoprotein, in maintaining normal native microbiota in the Gulf Coast tick (Amblyomma maculatum). Insect Mol Biol. 2015;24(5):570–81.
99. Gao S, Ren Y, Sun Y, Wu Z, Ruan J, He B, et al. PacBio full-length transcriptome profiling of insect mitochondrial gene expression. RNA Biol. 2016;13(9):820–5.
100. Fourdrilis S, de Frias Martins AM, Backeljau T. Relation between mitochondrial DNA hyperdiversity, mutation rate and mitochondrial genome evolution in *Melarhaphe neritoides* (Gastropoda: Littorinidae) and other Caenogastropoda. Sci Rep. 2018;8(1):17964.
101. Yévenes M, Núñez-Acuña G, Gallardo-Escárate C, Gajardo G. Adaptive mitochondrial genome functioning in ecologically different farm-impacted natural seedbeds of the endemic blue mussel *Mytilus chilensis*. Comp Biochem Physiol Part D Genomics Proteomics. 2022;42: 100955.
102. Zhang S-M, Bu L, Lu L, Babbitt C, Adema CM, Loker ES. Comparative mitogenomics of freshwater snails of the genus *Bulinus* obligatory vectors of *Schistosoma haematobium*, causative agent of human urogenital schistosomiasis. Sci Rep. 2022;12(1):5357.
103. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–62.
104. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
105. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS ONE. 2016;11(10): e0163962.
106. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
107. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
108. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
109. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
110. Lewis FA, Liang Y-S, Raghavan N, Knight M. The NIH-NIAID schistosomiasis resource center. PLoS Negl Trop Dis. 2008;2(7):e267–e267.

111. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Internet]. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
112. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691–2.
113. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2020;49(D1):D480–9.
114. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2020;49(D1):D412–9.
115. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
116. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013;29(19):2487–9.
117. Baril T, Imrie RM, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. bioRxiv. 2022;2022.06.30.498289. https://www.biorxiv.org/content/10.1101/2022.06.30.498289v2.
118. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013.
119. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12(1):2.
120. Smit A, Hubley R. RepeatModeler Open-1.0. 2008. Available at: https://www.repeatmasker.org/RepeatModeler/.
121. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12(8):1269–76.
122. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(suppl_1):i351-8.
123. Wickham H. ggplot2. Wiley Interdiscip Rev Comput Stat. 2011;3(2):180–5.
124. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49(16):9077–96.
125. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
126. Putri GH, Anders S, Pyl PT, Pimanda JE, Zanini F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics. 2022;38(10):2943–5.
127. Young ND, Stroehlein AJ, Wang T, Korhonen PK, Mentink-Kane M, Stothard JR, et al. Nuclear genome of *Bulinus truncatus*, an intermediate host of the carcinogenic human blood fluke *Schistosoma haematobium*. Nat Commun. 2022;13(1):977.
128. Maeda T, Takahashi S, Yoshida T, Shimamura S, Takaki Y, Nagai Y, et al. Chloroplast acquisition without the gene transfer in kleptoplastic sea slugs. Plakobranchus ocellatus Elife. 2021;10: e60176.
129. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276–7.
130. Laidemitt MR, Buddenborg SK, Lewis LL, Michael LE, Sanchez MJ, Hewitt R, et al. *Schistosoma mansoni* Vector Snails in Antigua and Montserrat, with Snail-Related Considerations Pertinent to a Declaration of Elimination of Human Schistosomiasis. Am J Trop Med Hyg. 2020;103(6):2268–77.
131. Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv. 2022;2022.04.08.487609. https://www.biorxiv.org/content/10.1101/2022.04.08.487609v1.
132. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
133. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
134. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 2015;32(1):268–74.
135. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9.
136. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256–9.
137. Dinguirard N, Cavalcanti MGS, Wu X-J, Bickham-Wright U, Sabat G, Yoshino TP. Proteomic Analysis of *Biomphalaria glabrata* Hemocytes During in vitro Encapsulation of *Schistosoma mansoni* Sporocysts. Front Immunol. 2018;9:2773.
138. Wu X-J, Dinguirard N, Sabat G, Lui H, Gonzalez L, Gehring M, et al. Proteomic analysis of *Biomphalaria glabrata* plasma proteins with binding affinity to those expressed by early developing larval *Schistosoma mansoni*. PLOS Pathog. 2017;13(5):1–30.
139. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018. Available online at https://www.R-project.org/.
140. Blouin MS, Bollmann SR, Tennessen JA. *PTC2* region genotypes counteract *Biomphalaria glabrata* population differences between M-line and BS90 in resistance to infection by *Schistosoma mansoni*. PeerJ. 2022;10:e13971.

Publisher's Note