Obtaining the genome sequence of the mollusc *Biomphalaria glabrata*: a major intermediate host for the parasite causing human schistosomiasis.

Matty Knight, Coen M. Adema*, Nithya Raghavan, Eric S. Loker*, Fred A Lewis and Hervé Tettelin[#]

Biomedical Research Institute, 12111 Parklawn Dr. Rockville MD 20852, *University of New Mexico, 167 Castetter Hall, Albuquerque, NM 87131, [#]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850.

Freshwater snails of the genus *Biomphalaria* are important intermediate snail hosts for the widespread transmission of schistosomiasis in humans. This chronic and debilitating disease remains one of the most intractable public health concerns in 74 developing countries, infecting more than 200 million people. Prevalence of schistosomiasis is difficult to estimate, but according to the World Health Organization more than 600 million people are currently at risk for infection with either one or more of the three medically important schistosome species, *Schistosoma mansoni*, *Schistosoma japonicum and Schistosoma haematobium (*WHO Expert Committee, 1993). Although control measures involving the combined use of molluscicides and mass chemotherapy have been effective in slowing the spread of the disease, long-term prevention of schistosomiasis has been difficult to achieve because of re-infection in the human population following chemotherapy. Ideally, a protective vaccine against the parasite will be the best method to combat the spread of this disease, but efforts to develop a vaccine have proven to be challenging (Bergquist, 1998). Thus, with neither a vaccine nor a thorough understanding of the parasite/host interaction at both the human and snail stages of the parasite's life cycle, we continue to make less than optimal progress in reducing transmission of schistosomiasis around the world. Added to this current situation is the decline in public health measures in several affected countries due to poverty, a rise in civil wars, and the construction of dams and new irrigation schemes in areas at most risk for schistosomiasis.

In the advent of modern genome biology, it may be possible to reverse this current trend by obtaining the genome sequences of the three organisms that are pertinent to transmission of schistosomiasis -the parasite, the intermediate snail host, and the human definitive host. Of the three, the human genome has been now sequenced, and efforts to obtain the complete sequence of, *S. mansoni* are underway (Schistosoma genome network website: www.nhm.ac.uk/hosted_sites/schisto). To date, at The Institute for Genomic Research (TIGR) and the Sanger Center a combined total of 1.8 million sequencing reads have been generated, achieving a 4.5x coverage of the parasite genome (El-Sayed *et al.*, submitted). Estimates indicate that less than 2% of the 270Mb genome is unsequenced, with approximately 20,000 gaps remaining. Additionally, considerable progress is also being made in learning more about the transcriptome profiles of various stages of *S. mansoni* (Verjovski-Almeida *et al*., 2003) and *S. japonicum* (Hu *et al*., 2003; Peng *et al*., 2003).

Of the schistosome snail host species, the most thoroughly studied is *Biomphalaria glabrata*. It is the one most closely associated with schistosomiasis in the Western Hemisphere and the easiest to maintain in the laboratory. Consequently, the volume of scientific literature on this species alone exceeds that of all the others combined. Since this snail species is linked to such an important human health problem, studies on its genetics have focused largely on its interaction with *S. mansoni* (Lewis *et al.*, 2001). These studies began in the mid-1950s, when it was found that susceptibility of *B. glabrata* to *S. mansoni* infection is a heritable trait (Newton, 1955; Richards and Shade, 1987). Since then, a great deal has been learnt about the genetics of parasite/snail compatibility. These studies culminated in the discovery that

resistance to parasite infection in the snail is mediated by an active defense response of the snail host against the parasite (Adema, and Loker, 1997). By identifying genes involved in the snail's internal defense system and other products that interfere with parasite survival, we may identify targets that can be exploited to develop alternative methods of controlling transmission of schistosomiasis. Information on the nature of those genes involved in the host /parasite relationship is still rudimentary. However, from these studies, the identity of some of the factors involved in the innate immunity of *B. glabrata* snails is gradually becoming available. Results such as the identification of fibrinogen-related proteins, (FREPs) have led to the conclusion that one ancestral function of fibrinogen (FBG) domains was in non-self recognition (Adema *et al.*, 1997). The presence of FBG domains in non-self recognition factors from other invertebrates and vertebrates (including the ficolins from humans) has confirmed that this function has been maintained to date. In addition to an FBG domain, FREPs also contain immunoglobulin superfamily (IgSF) domains. Remarkably, these snail-derived IgSF sequences most resemble so-called "V-type" Ig domains and may be informative for the evolutionary development of specific antigen-receptors of vertebrates such as antibodies and T-cell receptors (DuPasquier and Flajnik, 1999). Clearly, these results are relevant for our understanding of vertebrate (human) immunology.

In the field of neuroendocrinology, snails such as *Aplysia* and *Lymnaea* have been model organisms. Parasite infection has been shown to cause castration in infected snails due to the induction (by the parasite) of certain neuropeptides in the infected snail (de Jong-Brink, 1995; Sorensen *et al.*, 2001). Additionally, parasites down regulate or interfere with the internal defenses of the snail host (Lie, 1982). Therefore, it is reasonable to suggest that sequence information of the genomes of all three organisms relevant to the transmission of schistosomiasis (snail, parasite and human) will facilitate the identification of orthologs that may have similar effect on human infections. Efforts to identify genetic loci using a marker driven positional mapping approach is now under way, but this method is technically challenging because little in the way of either genetic/physical or RFLP maps exist for the *B. glabrata* genome. Obtaining the genome sequence of this snail host should help to identify molecular landmarks (microsatellites and SNPs) that will provide convenient predictors for epidemiological studies on parasite transmission in the field. Developing molecular markers should also improve field studies assessing the genetic make-up of snail populations, and benefit selection for genetic manipulation of parasite resistance genes in snails. These field applications developed from the *B. glabrata* genome sequence will also provide a model for many other snail borne infections, including emerging diseases such as cercarial dermatitis (Horak *et al*., 2002) and fascioliasis (Arjona *et al*., 1995).

B. glabrata is a highly derived member of the class gastropoda within the molluscan phylum. Its secondary adaptation to aquatic life suggests extensive selection during the course of evolution from old ancestry, that potentially preceded the Cambrian radiation. Compared to the genome size of *S. mansoni* (270 Mb) the genome of *B. glabrata* is considerably larger (estimated to be around 950 Mb, Gregory, 2003). The chromosomes (haploid number = 18) are small, relatively monomorphic, and have been organized into groups according to size and shape (Goldman *et al.*, 1984). To better understand the molecular make-up of *B. glabrata*, various gene libraries (cDNA, genomic, cosmid, BAC) have been constructed, and several laboratories are actively engaged in gene discovery efforts (ESTs). The full-length mitochondrial genome sequence (13670 nt) has also been obtained (DeJong *et al*., in preparation). The gene order of the mitochondrial genome sequence of *B. glabrata* is identical (with the exception of the order of two rRNA genes) to that of other gastropods. This further supports monophyly of the class gastropods (pulmonates, prosobranchs) within mollusca. The nuclear genome sequence may thus likely be informative for all gastropoda. The genome size estimate shows that *B. glabrata* has a small genome size among gastropods and even mollusca; thus it is economical to focus on this species

as a representative of the phylum. A recent search of the public nucleotide database GenBank (Release 137) revealed that a total number of 25,379 nucleotide sequences for molluscs (all species) have been deposited. On the basis of this information the number of nucleotide sequences for gastropods is shown in the table below.

Gastropod	**GenBank Entries**
Biomphalaria sp.	1964
B. glabrata	1709
Aplysia sp.	1175
Oncomelania sp.	159
Lymnaea sp.	169
Bulinus sp.	119

From this information, it is clear that the entire phylum Mollusca is vastly under-represented in proportion to its numbers and importance. For *B. glabrata*, several genes have been sequenced and characterized, and there currently are 1467 ESTs deposited in GenBank.

Compared to other invertebrate vectors of parasitic diseases (most notably several mosquito species), molecular studies of molluscs are considerably less advanced. The genomes of one of the most significant human malarial parasite, *Plasmodium falciparum* (Gardner *et al*., 2002) and its mosquito vector, *Anopheles gambiae* (Holt *et al.*, 2002) have now been sequenced. It is hoped that in the not too distant future, this wealth of molecular information will provide novel methods towards the eradication of malaria in the world. As a tropical disease, schistosomiasis is ranked as being only second to malaria in terms of its prevalence and ability to cause chronic malaise and morbidity in the human population. Similar to the *P. falciparum/A. gambiae* genome projects, the notion that learning more about genes that affect the outcome of the parasite/host relationship through a genome sequencing strategy is an ambitious undertaking. The snail host /schistosome relationship is known to be highly complex. Furthermore, the long term culture of these parasites (using the snail cell-line Bge) is not practical due to slow parasite growth and low yield of parasite material. It is obvious that since schistosomes alternate between a vertebrate and invertebrate host, ongoing sequencing efforts for *S. mansoni* will produce some significant gaps in our knowledge without having comparable sequence information for its snail host. This means that without a comparable genome sequence for the snail, progress towards unraveling what genes are associated with the *B. glabrata/S. mansoni* relationship may be slow to realize. As a part of the human genome sequencing effort, in addition to understanding the sequences present in the human genome, is a need for a broader overview of the human evolutionary ancestry. Current information is limited to deuterostome organisms (vertebrates and *Ciona*) and only to one class of protostome invertebrates (the ecdysozoa) as represented by *Anopheles, C. elegans* and *D. melanogaster*. No comprehensive information is currently available for higher developed representatives other major lineage of protostomes, the lophotrochozoa, of which molluscs, and gastropods are some of the most successful (in terms of number of species).

With the success of the human genome project (Lander *et al.*, 2001; Venter *et al.*, 2001), and the sequencing efforts of several other complex genomes, obtaining the genome sequence of the snail will be the most efficient way to advance our molecular knowledge of such an understudied complex organism. It is also clear that in recent years the technology and costs for generating genome sequences have significantly improved. The snail community is interactive and already pursues proteomics study of *Lymnaea stagnalis* at the VU Amsterdam, The Netherlands. In addition, expression studies, and

investigations towards developing micro-arrays are being pursued (Wellcome Trust-funded project in UK). International gene discovery efforts are expected to yield at least 12,500 EST entries within the next 3 years. With the realization that sequence data are essential for many relevant modern research methods, enthusiastic support for making a representative BAC library was gained from 5 continents (see below the list of investigators who provided letters of support). A website detailing efforts of the *B. glabrata* genome initiative is available (http://biology.unm.edu/biomphalaria-genome/index.html). Linkage maps are being composed using microsatellite markers, in anticipation of having actual genes and BACs toward physical mapping. The mitochondrial genome was sequenced by some international members of the *B. glabrata* genome initiative (UNM, USA and NHM, UK) without direct need or intent other than just to get *B. glabrata* on the list of organisms for which genomes are available, and because of the realization that the mitochondrial genome sequence (~13.7kb) would be too small to be recovered from a BAC library with standard (larger sized) inserts. The sequence (13,670 nt determined from two separate strains; M-line and isolate 1742) has been submitted to GenBank for release in the near future and a manuscript is in preparation. An individual initiative led to determination, and subsequent publication of the genome size of *B. glabrata* (Gregory, 2003). One modest BAC library (BS90 strain of *B. glabrata*, not currently proposed for genome sequencing) is available due to an in house effort (BRI). Several cDNA and genomic libraries have been produced and these have been shared among international labs (US, UK, Germany and others). Also, EST data and other experimental issues are frequently resolved informally between labs with the majority of ESTs generated by an individual initiative in collaboration with TIGR (Knight *et al.*, 1998, Raghavan *et al.*, 2003). Parasites and snails are exchanged upon request when possible. Significantly, in light of generating an NHGRI-sponsored BAC library and in anticipation of a genome initiative being undertaken, an international effort resulted in the collection of a new field isolate from Brazil (strain BB02). This snail was tested for parasite susceptibility (Carvalho, Brazil), and is maintained by selfing (to minimize haplotype diversity) in the US, South America and UK. Methods have been developed for the isolation of high molecular weight (HMW) DNA (Adema, Knight, combined with Wing, AGI) for the production of a representative BAC library (at least 100kb average insert size and 5x coverage), from BB02 *B. glabrata*, sponsored by NHGRI. This BAC library is currently being constructed (see below).

B. glabrata isolates have been maintained for decades in different laboratories across the world. Several strains are available with artificial selection for traits that determine compatibility with *S. mansoni*, and with *Echinostoma caproni* (France). We anticipate that the *B. glabrata* BB02 snails will become a new model organism for molluscs like *C. elegans* for nematodes and *D. melanogaster* for insects. Generation time of the snails is about 2 months and several generations can be obtained in a year. The DNA has been proven to be amenable to all standard molecular techniques, including high molecular weight DNA isolation. As stated above, snails with distinct phenotypes are available due to selection of compatibility qualities for human parasite *S. mansoni.* These snails were generated from complex crosses between geographically distinct isolates (Newton 1955; Richards and Shade 1987). To maintain these distinct snail isolates for the long term, careful selection is needed due to genetic diversity within strains. Several field-collected strains are also available. For example, the BS90 resistant strain has been maintained for decades without selection, and continues to be of the resistant phenotype since its first discovery in Brazil by Paraense and Correa in 1963. Also, selection experiments in France have yielded *B. glabrata* strains that are resistant for *Echinostoma caproni* (Langand *et al.*, 1998).

The availability of genome sequence information will enable the development and utilization of modern methods, such as RNAi to target (knock down as an alternative to knock-out) genes specifically associated with the schistosome snail-host relationship. These studies can be used to determine the regulation of gene expression and to facilitate the identification and manipulation of promoters; in particular, to locate and characterize pathogen-response factors. In addition, we will be in a better

position to study the extent and arrangement of gene families and understand complexities of transposable elements associated with the *B. glabrata* genome. Evidence of horizontal transfer of transposable elements between the parasite and the snail/human host will become obvious from sequence information generated from these organisms. Because of its smaller than anticipated genome size of *B. glabrata* to that of other molluscs, the possible study of occurrence of overlapping genes or alternative splicing will all be made easier once a complete genome sequence is available. Furthermore the sequence information obtained will complement the genomics research ongoing in several laboratories. Current research efforts to cultivate parasites *in vitro* using the lab maintained *B. glabrata* cell-line (Bge) should benefit significantly once sequence information becomes available, since this should provide a useful resource for the selection and design of promoters for gene expression/regulation and gene transfer studies. These studies (with the snail as a new model) will open up the field and attract new investigators with interest in comparative genomics involving another phylum.

We propose, that for a meaningful start of a genome project for this organism, our collective efforts could be organized into 2 major phases. These phases were first discussed at a snail genome sequencing initiative meeting at the 2001 American Society of Parasitology (ASP) in Albuquerque NM. The meeting was attended by several investigators, national and international (see list of current members of consortium) with research interests in the *B. glabrata* snail host and its interaction with trematodes.

The current snail genome consortium:

Coen Adema* - USA
Gennady Ataev- Russia
Chris Bayne* - USA
David Blair - Australia
Paul Brindley* - USA
Omar Carvalho - Brazil
Christine Coustau* - France
Jason Curtis* - USA
Lawrence A. Curtis - USA
Marijke deJong-Brink - Holland
Colette Dissous - France
Georges Dussart* - UK
Sharon File-Emperador – USA
Petr Horak - Czech
Renzo Nino Incani* - Venezuela
David Johnston - UK
Catherine Jones* - Scotland
José Jourdane* - France
Bernd Kalinna - Germany
Matty Knight* - USA
Thomas K. Kristensen* - Denmark
Hammou Laamrani* - Morocco
Fred Lewis***** - USA
Eric Loker* - USA
Nicholas Lwambo* - Tanzania
John Malone - USA
Don McManus* - Australia
Dennis Minchella* - USA
Gerald Mkoji* - Kenya
Helene Mone - France
Les Noble* - UK
Guri Roesijadi* - USA
David Rollinson* - UK
John Sullivan* - USA
Herve Tettelin - USA
Andre Théron - France
Jackie Trigwell* - UK
Mingyi Xia* - China
Tim Yoshino* - USA
Ulrike Zelck* - Germany

*Explicit indication of willingness to contribute experimental effort.

In anticipation of a snail genome project becoming a reality we also sought and obtained the support of NHGRI for the construction of a representative *B. glabrata* BAC library. As mentioned above, a modest BAC library is available and another is currently under construction by AGI, a BAC-resource laboratory from NHGRI. This library will be completed later this year or early next year. The isolation of high molecular weight DNA from this organism has proven to be challenging (as seen for other mollusc systems). The BAC resource centers (Wing at AGI/de Jong at Children's Hospital Oakland Research Institute) working with NIH/NHGRI have experienced problems with obtaining HMW DNA from gastropod molluscs. In a joint effort researchers from UNM and AGI have exchanged laboratory visits to test, modify and apply protocols employed by AGI for obtaining HMW DNA. With the HMW DNA obtained from 40 selfed BB02 *B. glabrata* snails, AGI has initiated the production of a high quality BAC library. Recent analysis (October 2003) of transformed clones indicated an average insert size of

>140kb. Mass transformation is being performed to allow picking of clones. AGI intends for 10x coverage of the *B. glabrata* genome, additional ligations will be performed if need be. Consequently, a BAC library meeting or exceeding the Quality Assessment Standards predetermined by NHGRI for production of BAC libraries is imminently available.

The 2 phases for the sequencing efforts are:

Phase 1 - Presequencing approaches

Using a multigroup effort, the labor for this could be partitioned in the following way:

1. Construct BAC libraries with an average insert size of at least 100 Kb with 5-fold genome coverage (ca. 50000 clones). This is currently ongoing at UNM and AGI as stated above. Estimated time-line for completion is within this calender year (2003).
2. Construct cDNA libraries from specific regions or tissues of interest for EST-based sequencing. cDNA libraries are currently available for hemocytes, albumen glands, cerebral ganglia, hepatopancreas and whole snail and the Bge cell line. Several gene discovery projects or additional molecular characterization projects are in progress (US: Bayne, Loker, Adema, Minchella, Knight, Raghavan, Lewis; UK: Rollinson, Jones, Noble; Germany: Zelck; France: Capron, Dissous; Brazil: Carvalho and others). Mostly, these projects have generated ESTs or specific gene sequences, and have not yet moved into sequencing large genomic fragments.
3. Sequence full-length cDNAs for gene-prediction training.
4. Map new or already known genes to the BACs, so that contigs can be identified. Several BAC clones corresponding to a known retrotransposable element in the *B. glabrata* genome have been isolated and these BACs will be sequenced within the next 4 months.

Phase 1 will provide the starting material needed for Phase 2 and will allow us to proceed to the whole genome shotgun sequencing in a timely manner.

Phase 2 - Sequencing approaches

We will either employ whole genome shotgun sequencing, or shotgun sequencing of relevant chromosomes or BACs depending on the availability of funds. There are no alternate sources of funding for this project at this stage, but investigators in the UK are exploring possibilities with the Sanger Center (Wellcome Trust). The US members of the initiative have inquired with NIAID, and DOE (at the ASP meeting, 2001, considering *B. glabrata* as an environmental bioindicator species). WHO (Dr. Carlos Morel, Director TDR program, Dr. Ayo Oduola) has also expressed an interest. However, no commitments have been made by any of the above groups to fund a snail genome sequencing project.

Proposed steps towards sequencing are as follows:

1. Use the BAC libraries to generate BAC-end sequences that will provide paired sequences for scaffolding of sequence assemblies. These BAC-end sequences will also serve as markers along the snail chromosomes. To obtain one marker per 10 Kb, both ends of *ca*. 50,000 BAC clones will be sequenced.

2. Construct sheared genomic libraries with average insert sizes of 2 Kb, 10 Kb and 30-50 Kb in preparation of the shotgun sequencing phase.

3. Map BACs to chromosomes by FISH in a limited fashion; no more than 50-100, in order to be able to attribute contigs to specific chromosomes.

4. Based on the above map, enter the genome at a couple locations by choosing the appropriate BACs and generate large contigs (e.g. 2x1 Mb or 1x2 Mb which represents *ca*. 20 BACs) by sequencing partially overlapping BACs from those areas. The approach is to fully sequence each seed BAC by shotgun, then to identify the next partially overlapping BAC (overlap of *ca*. 15 kb) based on the alignment of the BAC-end sequences generated, sequence this next BAC and repeat the process. These large contigs will provide an idea of gene organization and repeat content in the snail genome. In addition to biological insights, this information will also be of great value in order to assess the possibility of using the whole genome shotgun approach.
5. Based on the success of the steps above and the available budget, perform whole genome shotgun sequencing of the snail genome to a certain sequence coverage (e.g. 3x).

The sequencing approaches described above have already been applied extensively to ongoing genomes at TIGR such those of *S. mansoni* and *Trypanosoma brucei.* Thus, TIGR and the Joint Technology Institute (JTC) has the necessary expertise to achieve these sequencing goals and properly manage the data generated.

In summary, schistosomiasis research has been supported by NIH-NIAID for roughly 50 years. As a consequence, great strides have been made in understanding the biology of the parasite and, most notably, deciphering immune components of the disease process in the mammalian host (Cheever, *et al*., 2000). We feel the present time is to sequence the genome of *B. glabrata*, thus giving more relevance to the information coming from the sequencing efforts for S. *mansoni*, and spearheading the NIH-supported sequencing movement into an entirely new phylum of medically and economically important organisms

Bibliography

1. Adema, C.M. and Loker, E. S. (1997) In: Advances in Trematode Biology, Fried, B. and Graczyk T.K. (Eds) CRC Press, Boca Raton and New York. pp. 229-263.

2. Adema, C. M. *et al.* (1997) Proc. Natl. Acad. Sci. USA. 94:8691-8696.

3. Arjona, R. J. *et al.* (1995) Medicine. 74:13-23.

4. Berquist, N. R. (1998) Tech. Rep. Ser. 830:1-86.

5. Cheever, A. W. *et al.* (2000) Immunology Today. 21:465-466.

6. de Jong-Brink, M. (1995) Advances In Parasitology. 35:177-256.

7. DuPasquier, L. and Flajnik, M. (1999) In Fundamental Immunology, 4th editiob, Ed Paul W. E,. Lippincott-Raven Publishers New York, pp. 605-650.

8. Gardner, M. J. *et al.* (2002) Nature. 419:498-511.

9. Goldman, M. A. *et al.* (1984) Malacologia. 25:427-446.

10. Gregory, T. R. (2003) Genome. 46:841-844.

11. Holt, R. A. *et al.* (2002) Science. 298:129-149.

12. Horak, P. *et al.* (2002) Advances In Parasitology. 52:155-233.

13. Hu, W. *et al.* (2003) Nat Genet. 35:139-147.

14. Knight, M. *et al.* (1998) Malacologia. 39:175-182.

15. Lander, E. S. *et al.* (2001) Nature. 409:928-933.
erratum: Nature (2001) 411:720
correction: Nature (2001) 412:565-566.

16. Langand, J (1998) Heredity. 80:320-325.

17. Lewis, F. A. *et al.* (2001) Parasitol. 123:S169-S179.

18. Lie, K. J. (1982) Tropical and Geographical Medicine. 34:11-122.

19. Newton, W. L. (1955) J. Parasitol. 41:526-528.

20. Peng, H. J. *et al.* (2003) Parasitol Res. 90:287-93.

21. Raghavan, N. *et al.* (2003) Mol. Biochem. Parasitol. 26:181-91.

22. Richards, C. S. and Shade, P. C. (1987) J. Parasitol. 73: 1146-1151.

23. Sorensen, R. E. *et al.* (2001) Parasitology. 123:Suppl:S3-18.

24. Venter J. C. *et al.* (2001) Science. 291:1304-1351.
Erratum : Science (2001) 292:1838.

25. Verjovski-Almeida, S. *et al.* (2003) Nat. Genet. 35:148-57.