Prediction and annotation of genomic repeat dynamics in the snail *Biomphalaria glabrata* using Hidden Markov Models

An essay submitted in partial fulfillment of

the requirements for graduation from the

Honors College at the College of Charleston

with an Artium Baccalaureatus in

Marine Biology

Kelsey L. Yetsko

May 2014

Advisor: Andrew M. Shedlock

Secondary Advisor: Paul E. Anderson

Prediction and annotation of genomic repeat dynamics in the snail *Biomphalaria glabrata* using Hidden Markov Models

Kelsey L. Yetsko[1], Paul E. Anderson[2], Andrew M. Shedlock[1]

Department of Biology[1], Department of Computer Science[2], College of Charleston

Abstract

Mobile elements cover an extensive amount of the genomes of both plants and animals. However, current homology, or similarity comparison, based search tools are optimized only for analyzing and annotating repeats in humans and well known experimental models. This skewed taxonomic distribution of reference data makes homology-based search tools less sensitive and less accurate, missing many targets in poorly examined genomically diverse lineages. With these limitations in mind, Hidden Markov Models (HMMs) were used for *de novo*, rather than homology-based, repeat annotation in the gastropod mollusk species *Biomphalaria glabrata*. Here we compare the HMM profile repeat annotation output to the commonly used homology-based search tool RepeatMasker in order to assess comparatively whether there is an advantage of using *de novo* model-driven repeat annotation methods over homology-based tools. Finally, we used PCR to amplify eight choice repeat segments across five cloned individual *B. glabrata* to verify experimentally the output received computationally.

Introduction

Since the Human Genome Project and the discovery that about half of the human genome is composed of not genes, but mobile elements, these repetitive segments of DNA have come under much scrutiny. They are prevalent in the genomes of most plants and animals [1, 2] and are important in building the structure of genomes [2]. There are two main

classes of repetitive elements, also known as transposable elements: class I elements, or retrotransposons, which move directly in the genome via a copy-and-paste mechanisms that requires an RNA intermediate; and class II elements, or DNA transposons, which move in the genome directly via a cut-and-paste mechanism [3]. Mobile elements are important because of their ability to shape genomes and also by providing the raw material for new gene functions. Mobile elements also aid in exaptation, a process in which a section of a mobile element consensus that has been under selection for transposition is placed under selection by the host for a new function [4]. So far, more than 10,000 functional elements in humans have been described as exapted from mobile elements that are identifiable in the modern human genome [4]. However, exaptation events older than the radiation of mammals are often difficult to detect when just analyzing the current human genome because most mobile elements that were active hundreds of millions of years ago have become inactivated somewhere in the lineage leading to humans [4]. Repeat consensus from species outside of mammals that still harbor ancient mobile elements in a near-ancestral state can be used to identify exaptations in the human lineage associated with now inactive transposons in humans [4]. Overall, genome-wide repetitive element analysis is important for several reasons. It aids our understanding about how genomes are built and how they change. The study of repeats throughout diverse lineages can also help to identify exaptations in our own genome, where repeats come under selection for new functions. Finally, through the current genome projects that have already been completed, we are presented with a great diversity in the percent genome composition of repeats and the dominant class of repeats throughout the animal kingdom [5, 6, 7, 8, 9].

There are several different tools available for repeat analysis but one of the most common repeat annotation tools is RepeatMasker [10]. RepeatMasker offers an online webserver that can accept DNA sequences in FASTA format as input and outputs a detailed

annotation of the repeats present in that DNA sequence [11]. RepeatMasker screens DNA sequences for repetitive elements by using sequence homology search tools such as Cross-match and Abblast against transposable element libraries in Repbase [12]. Most homology search tools search for pairwise similarity between a sequence of interest and the Repbase collection of transposable element sequences [12]. The sensitivity of transposable element detection using these methods depends both on the content of the database and the homology search method that is used [12]. While these tools have provided valuable information in the past, they do come with limitations. One of the main issues comes from the nature of the transposable elements themselves. Older transposable elements that have undergone extensive mutation or are heavily truncated or shortened due to reinsertion in the genome and may not be recognized when using homology search tools [12]. Also, 73% of the available downloadable transposable element libraries are from mammalian species, which limits the accuracy of homology-based repeat annotation for non-mammalian vertebrates and for all invertebrates.

In order to develop more comprehensive and accurate repeat annotators, hidden Markov models (HMMs) are being extensively used for *de novo* annotation. Markov models utilize simple probabilistic approaches to many different situations without becoming too computationally expensive. These HMMs are constructed using the concepts of conditional probability [13]. A Markov chain is a sequence of random variables whose probabilities depend on the transition probability, which is a conditional probability for the system to go to a particular state given the current state (Fig 1) [14].

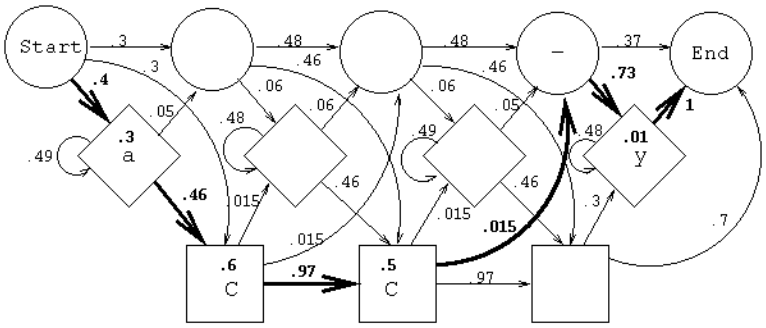

Figure 1. An example of a generic Hidden Markov Model (HMM), showing states (circles and squares), emission probabilities for those states, transitions between states (arrows), and transitional probabilities [15].

These models were used to build the transposable element library available in Dfam [12]. Dfam is a database of repetitive DNA elements put together by the HHMI Janelia Farm Research Campus [16], and it contains entries corresponding to all Repbase transposable entries from the human genome [12]. Each Dfam entry is represented by a profile HMM which is built from sequence alignments generated using other online transposable element libraries accessioned in RepeatMasker and Repbase [11, 17]. When used with the HMM search tool nhmmer [18], Dfam has been shown to increase the accuracy of human genome repeat annotation by 2.9%. However, Dfam is currently a human genome database only, which limits its effectiveness when working with other species. Our work in building profile HMMs for different mobile elements in other species is anticipated to recover a far greater relative proportion of unknown repeats from non-model host genomes not present in the heavily mammalian-biased reference database. As such it will provide a valuable resource for the genome and transposable element research community.

The main species on which we have focused our study is the freshwater snail *Biomphalaria glabrata*. The genome project for this species is currently underway. Dr. Shedlock has been asked to serve on the Steering Committee of the NIH-NHGRI funded International *Biomphalaria* Genome Sequencing Consortium and is leading the team of experts responsible for analyzing the repetitive genomic landscape for this gastropod species.

There is a large amount of data already available online as well as a large body of proprietary data accessible through the Shedlock Laboratory. This species was chosen for several reasons. One, it has a relatively small genome compared to other gastropods and mollusks (0.95 Gb), making its genome an ideal representative for this diverse group of organisms [19]. It is also one of the main species of snail responsible for a majority of the transmissions of the parasite *Schistosoma mansoni* in the Western hemisphere [19].Schistosomiasis is a debilitating disease that has infected more than 200 million people and remains a chief public health concern in 74 developing countries [20]. It is estimated that 600 million people are currently at risk for infection with either one or more of the three schistosome species responsible for causing schistosomiasis [20]. Its prevalence is further amplified by a decline in public health measures due to poverty, civil wars, and new irrigation schemes [20]. Efforts to develop a vaccine against schistosomiasis via molecular-based methods have proven to be challenging, which has prompted the effort to examine the species at a genomic level. By studying the genomes of the three organisms important to the life cycle and transmission of the parasite (*S. mansoni*, *B. glabrata*, and humans), we may be able to see evidence of horizontal transfer of transposable elements between the parasite and the snail or the human host [20].

With this study, we are proposing that the use of HMMs will provide another, more accurate, analysis of transposable element content in the *B. glabrata* genome. Our hypothesis was that *de novo* repeat annotation using HMMs would identify a larger percent genomic coverage of repetitive elements, as well as a larger percent coverage of unknown elements, than the commonly used homology-based search tool RepeatMasker. Finally, we verified the presence of repetitive elements identified in the *B. glabrata* genome by the generated HMM profiles *in vitro* by using PCR amplification and gel electrophoresis of eight randomly

selected loci across four diverse retrotransposon repeat families from isolated and purified *B. glabrata* DNA.

Methods

Hidden Markov Model comparison to RepeatMasker

The current available *B. glabrata* genome assembly was downloaded from VectorBase [21] as the BB02 strain genomic contig sequences file of the BglaB1 genome assembly and uploaded to the Charleston Computational Genomics Group (C2G2) cluster. Contig refers to overlapping DNA segments generated during genome sequencing that represent a consensus region of DNA. The current version of HMMER (HMMER 3.1b1) [22], which was made available by Janelia Farms of the Howard Hughes Medical Institute (HHMI), was also uploaded to the C2G2 cluster and used to generate multiple HMM profiles that were then used to search for different families of retrotransposons within the *B. glabrata* genome file. The HMMER 3.1b1 download included two main functions that were essential to this project: hmmbuild, which uses a multiple sequence alignment file in Stockholm format of a sequences of a specific repeat family as input to generate an HMM profile for that repeat family, and hmmsearch, which uses the probabilities in the HMM profile to identify repetitive sequences in any DNA sequence file in FASTA format that is used as input. In order to generate the multiple sequence alignment files that were required as input for hmmbuild, the *B. glabrata* genome file was run through a *de novo* repeat family identification and modeling package called RepeatModeler [23]. The RepeatModeler package includes two *de novo* repeat identification tools, RECON and RepeatScout, which employ l-mer and k-mer algorithms [24], a different method than HMMER, to identify repetitive elements. RepeatModeler then uses these outputs to generate consensus models of repeat families in the

genome [25]. The multiple sequence alignments used to generate the consensus for the RepeatModeler repeat library were stored in 892 html files (Fig 2).

Figure 2. A small portion of the multiple sequence alignment html file used to generate the consensus sequence for the *B. glabrata* repeat family 1014 from round 2 of the RepeatModeler output. This repeat family is identified as Unknown by RepeatModeler.

Using these html files, I wrote a program using the Python programming language (script codes are available upon request from the author) that would convert the last iteration of the multiple sequence alignment, which had fewer low quality scores (Fig 2), into a multiple sequence alignment text file in FASTA format (Fig 3) for each of the Class I element repeat families. These FASTA alignment files were then uploaded onto the online Format Converter (v2.3.5) tool provided by the HIV Sequence Database [26], which converted the FASTA formatted alignment files into Stockholm format (Fig 4). Finally, the Stockholm formatted multiple sequence alignment files were used as input for the hmmbuild function included in the HMMER 3.1b1 package to produce a profile HMM for each individual Class I element repeat family, and these profile HMMs were used with hmmsearch to identify *de novo* repeat sequences in the *B. glabrata* genomic sequence file.

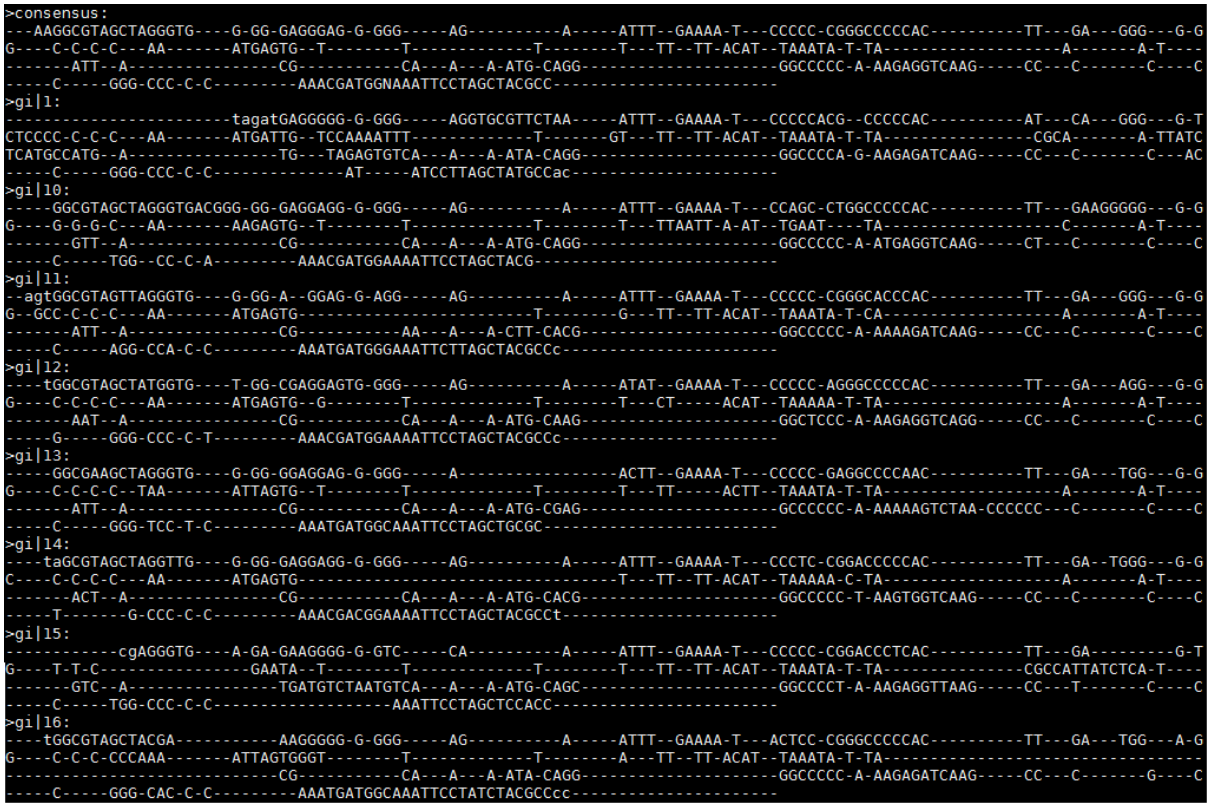

Figure 3. Example of a small subset of a FASTA format multiple sequence alignment file of the *B. glabrata* round 2 family 1014 repeat family, identified by RepeatModeler as Unknown.

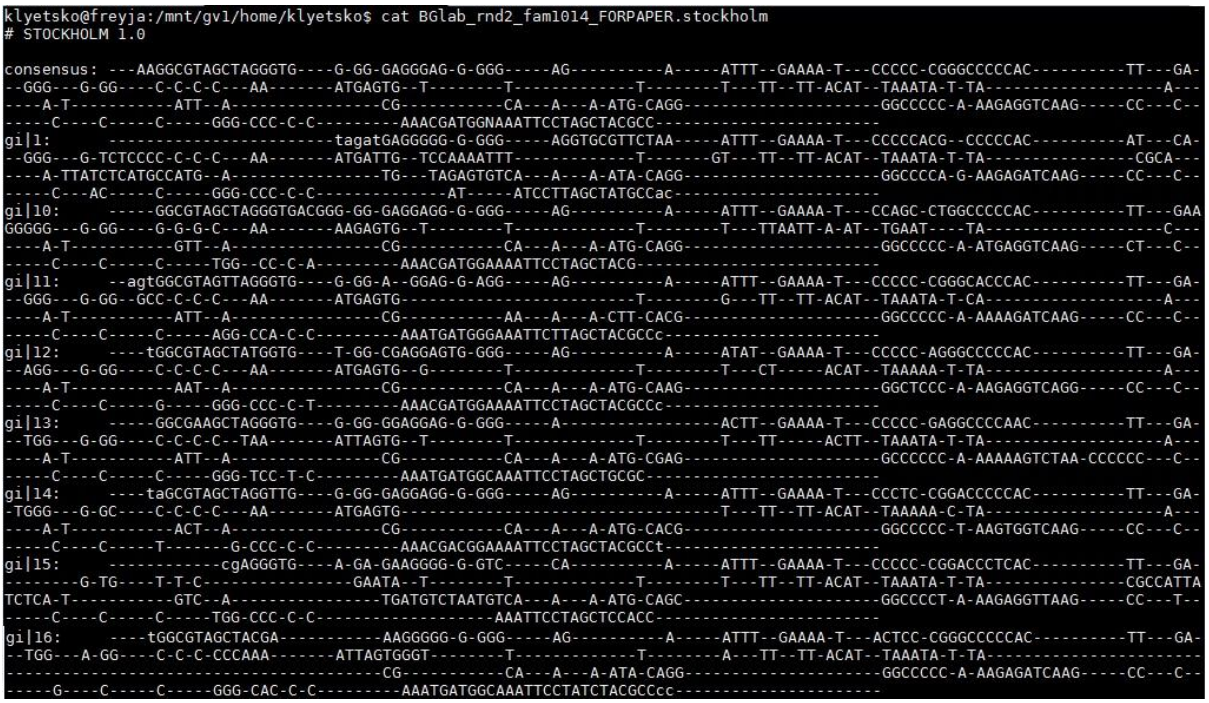

Figure 4. Example of a small subset of a Stockholm format multiple sequence alignment file of the *B. glabrata* round 2 family 1014 repeat family, identified by RepeatModeler as Unknown.

The output of the hmmsearch function was saved to a text file, and later analyzed in order to determine percent genomic coverage of the multiple Class I repeat families. The output

included the number of repeat sequences identified for each specific repeat family, as well as information on its exact location in the *B. glabrata* genome sequence (Fig 5). The e-value (Fig 5), also known as the expectation value, is the statistical significance score; repeat sequences with e-values close to zero are considered significant [27]. Using the contig header, the 'ali from' value, and the 'ali to' value, I wrote another Python program (script codes are available upon request from the author) to calculate percent genomic coverage of the multiple different Class I element repeat families. This was accomplished by determining the length of the sequence based on the positional information given for 'ali from' and 'ali to' for the specific contigs of the *B. glabrata* genomic sequence contigs file. The length of the identified repeat sequences was divided by the total base pairs in the genomic sequence contigs file, which was calculated by using another Python program I developed (script codes available upon request from the author), in order to generate a percent genomic coverage for each repeat sequence, and then group percent genomic coverage by major Class I repeat families.

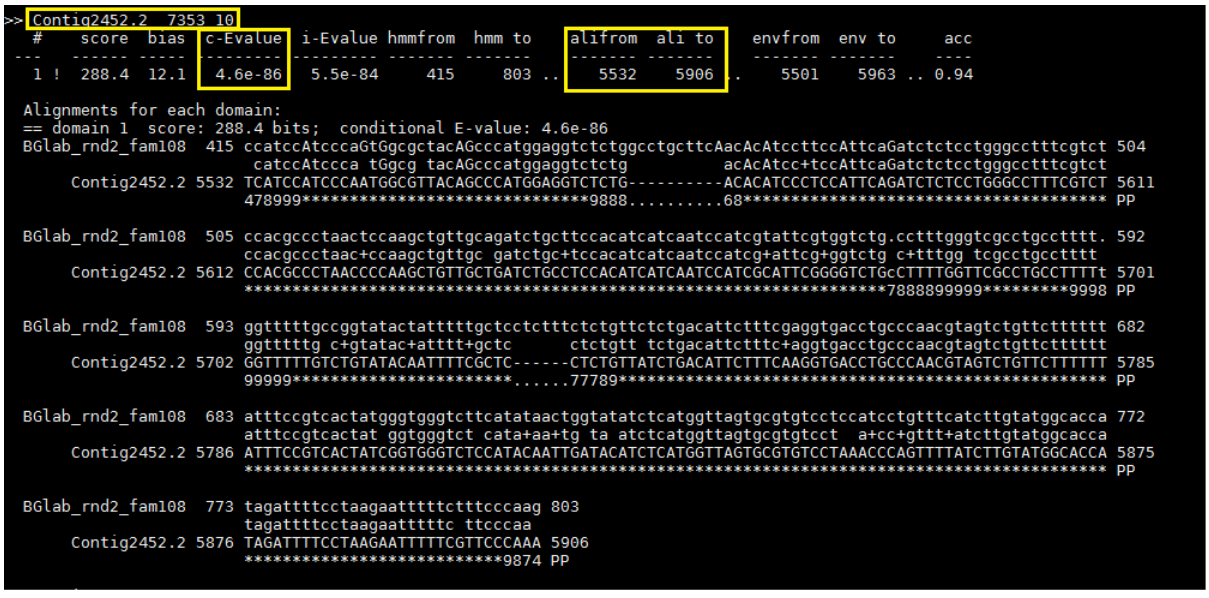

Figure 5. One repeat sequence identified from hmmsearch of the *B. glabrata* genomic sequence using a profile HMM of family 108 from round 2 of the RepeatModeler output, which was identified by RepeatModeler as a LINE/RTE-BovB element. Highlighted in the figure, from left to right: the contig header from the *B. glabrata* genome file, the e-value for the repeat sequence identified by the HMM, and the positional information of where the sequence is found within the *B. glabrata* contig.

The percent coverage values for major Class I element repeat families identified by HMMs were compared to the percent coverage values for sequences identified using the RepeatMasker repeat annotation package using the repeat family library file generated by RepeatModeler and the *B. glabrata* genome sequence file as input.

PCR verification of HMM identified repeat sequences

To develop the primers for PCR verification, eight repeat sequences in the *B. glabrata* genome identified by HMM profiles were randomly selected across four different Class I repeat families: the chicken repeat one (CR1) subfamily of the long interspersed nucleotide element (LINE) repeat family, short interspersed nucleotide elements (SINEs), long terminal repeats (LTRs), and the LINE subfamily Nimbus, which has also been identified in the genome of the parasite *S. mansoni*. Using a program I wrote in Python (script codes available upon request from the author), these eight sequences were extracted from the original *B. glabrata* genome sequence file and uploaded onto the NCBI Primer-BLAST tool [28] and the resulting forward and reverse primers were synthesized (Table 1). We received DNA from Dr. Coen Adema (Univ. NM Biology Dept., Albuquerque, NM USA) that was extracted using a CTAB extraction of whole body DNA from five individual BB02 strain *B. glabrata* snails. The DNA was sent as pellets under 8% EtOH and was resuspended and diluted 1:50 using MilliQ water. Reactions for two 96-well gradient PCR plates were set up, and each well included 25 µl of the following reaction mixture: 2.5 µl of buffer, 0.5 µl of a dNTP solution, 0.75 µl of one of the eight forward primers, 0.75 µl of the matching reverse primer, 0.13 µl of Taq polymerase to catalyze the reaction, 2.27 µl of *B. glabrata* DNA from one of the five cloned individuals, and 18.11 µl of MilliQ water. The plates were then placed in a thermocycler which was calibrated to run the following heating cycles: one two-minute cycle at 95 °C; 30 cycles of 95 °C for 30 seconds, a range of annealing temperatures between 49

and 61 °C for 30 seconds, and 68 °C for 60 seconds; and a final extension cycle at 68 °C for five minutes. Each row of the two gradient PCR plates had a different annealing temperature in the range of 49 to 61 °C (Table 2, 3) in order to determine the optimal annealing temperature for each of the eight sets of primers, and this was calibrated by the thermocycler. For the two PCR plates, only DNA from two of the BB02 *B. glabrata* individuals was used, BG1 and BG4. These samples were determined to have the highest concentration of DNA following DNA quantification with a spectrophotometer. Following PCR amplification of the eight loci, products for the annealing temperatures 51, 54, and 57 °C were analyzed and visualized using agarose gel electrophoresis. The 2% agarose gel was mixed with GelRed nucleic acid dye, which stains DNA products and can be visualized using a UV imager.

Name	Primer sequence (5' --> 3')	Expected Product Lengths
Fwd CR1 Contig899.8	AGTATGGAACGGGTTGCCTG	531
Rev CR1 Contig899.8	TGAACGAAGTGGTCGTCTCC	
Fwd CR1 Contig13668.1	ACTGCTGATGGTTGCTGTGT	509
Rev CR1 Contig13668.1	ATCCGTTCCACGCTCTGAAG	
Fwd SINE Contig246.29	CTTTGCTGGGACAAAGTCGC	770
Rev SINE Contig246.29	TCTATTGCTTGTTCTCCGTTTGA	
Fwd SINE Contig7790.2	TGCGGGTTCATTACGCCTAC	323
Rev SINE Contig7790.2	GTCGCGTCGTTGCTTACAAA	
Fwd LTR Contig7285.1	ACGCGCCCTTGTCAATAGAA	839
Rev LTR Contig7285.1	TGGAGAACTGCCTGAACTGG	
Fwd LTR Contig261.29	ACTTAAGGAGGACCCCACGA	974
Rev LTR Contig261.29	TCTCTTTTGGTGAGCGGGAC	
Fwd Nimbus Contig52312.1	CCAAAACGGCCTAGCAAACC	372
Rev Nimbus Contig52312.1	AAATAAGGCCGGGGGAGTTG	
Fwd Nimbus Contig6.87	AAGGGGCTAGAGTCGCCTAA	286
Rev Nimbus Contig6.87	GCTAAGGAAAGAGGAGCGCA	

Table 1. List of synthesized PCR primers for eight randomly selected loci identified in the *B. glabrata* genome by profile HMMs.

CR1 Contig899.8			Nimbus Contig6.87			LTR Contig261.29			SINE Contig246.29			Annealing temps
												61
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	60
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	57
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	54
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	51
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	49

Table 2. Table of a 96-well PCR plate for the primers for CR1 Contig899.8, Nimbus Contig6.87, LTR Contig261.29, and SINE Contig246.29. In the table, 'neg' stands for the negative control in which no DNA was added, and the range of annealing temperatures used for each row is shown on the right.

CR1 Contig13668.1			Nimbus Contig52312.1			LTR Contig7285.1			SINE Contig7790.2			Annealing temps
												61
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	60
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	57
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	54
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	51
BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	BG1	BG4	neg	49

Table 3. Table of a 96-well PCR plate for the primers for CR1 Contig13668.1, Nimbus Contig52312.1, LTR Contig7285.1, and SINE Contig7790.2. In the table, 'neg' stands for the negative control in which no DNA was added, and the range of annealing temperatures used for each row is shown on the right.

Results

The RepeatMasker output using the *B. glabrata* genome sequence and repeat family library file has a total percent coverage of 15.58% for Class I repeat families and 20.06% for Unclassified repeat sequences (Table 4). The RepeatMasker table output is tailored for the human genome, in that it divides the major repeat families into subfamilies, but only includes the major subfamilies found in humans, such as ALUs and LINE1s. According to

RepeatMasker, most repeat sequences in *B. glabrata* that are not Unclassified are LINEs, and the most abundant LINE subfamily shown is the L3/CR1 subfamily.

RepeatMasker: *Biomphalaria glabrata*		
	length occupied	percentage of sequence
SINEs (total):	18758200 bp	2.09%
ALUs	0 bp	0.00%
MIRs	0 bp	0.00%
LINEs (total):	111336209 bp	12.39%
LINE1	0 bp	0.00%
LINE2	2794355 bp	0.31%
L3/CR1	19716902 bp	2.19%
LTR elements (total):	9923293 bp	1.10%
ERVL	0 bp	0.00%
ERVL-MaLRs	0 bp	0.00%
ERV_classI	0 bp	0.00%
ERV_classII	0 bp	0.00%
Unclassified:	180350020 bp	20.06%

Table 4. RepeatMasker output table for *B. glabrata*. The RepeatMasker output is tailored for the human genome in that it only includes the major repeat subfamilies found in humans.

On the other hand, HMMs can be used to identify multiple subfamilies that RepeatMasker does not include in its output table file (Table 5). Subfamilies identified by RepeatModeler and included in the HMM output that are not specifically identified by RepeatMasker are the LINE/RTE-BovB, LINE/I-Nimb, LINE/Jockey, and LTR/Gypsy elements. Profile HMMs identified a higher percent coverage of total LINEs, and specifically L3/CR1s compared to RepeatMasker, and had a total percent coverage of 17.59% for Class I repeat families. The HMM pipeline also identified the same percent coverage of LINE2 elements as RepeatMasker. However, profile HMMs also identified 0.1% fewer total LTR elements and 0.77% fewer SINE elements, which is not what was originally predicted to occur.

HMM: *Biomphalaria glabrata*		
	length occupied	percentage of sequence
SINEs (total):	12100965 bp	1.32%
ALUs	0 bp	0.00%
MIRs	0 bp	0.00%
LINEs (total):	139516539 bp	15.26%
LINE1	0 bp	0.00%
LINE2	2878335 bp	0.31%
L3/CR1	39928337 bp	4.37%
LINE/RTE-BovB	59253791 bp	6.48%
LINE/I-Nimb	1291417 bp	0.14%
LINE/Jockey	29072630 bp	3.18%
LTR elements (total):	9243646 bp	1.01%
ERVL	0 bp	0.00%
ERVL-MaLRs	0 bp	0.00%
ERV_classI	0 bp	0.00%
ERV_classII	0 bp	0.00%
LTR/Gypsy	8463732 bp	0.93%

Table 5. Length occupied and percent coverage calculated for repeat sequences identified in the *B. glabrata* genome by profile HMMs.

All eight primers that were used to verify loci identified by the HMM pipeline amplified a specific segment of DNA and produced product (Fig 6, 7). The LINE/Nimbus Contig6.87 primers were the only pair of primers that produced nonspecific amplification, and instead amplified three different loci (Fig 6A). The banding pattern for all four gels is smeared, and appears to be an artifact of the agarose gel or the gel electrophoresis since the DNA ladder exhibits the same curved pattern. This has prevented us from determining the exact size of the bands by comparison to the ladder. However, all of the bands fall within a similar range of the expected product lengths for the eight primers (Table 1), except for the larger two DNA sequences that were amplified by the LINE/Nimbus Contig6.87 primers (Fig 6A). The two primers that amplified a target sequence closest to the expected product were

the smallest sequence band of the LINE/Nimbus Contig6.87 primer set (Fig 6A, Table1) and the SINE Contig7790.2 primer set (Fig 7B, Table1). For some of the primers, there appear to be different sized bands between the two *B. glabrata* individuals, but close in size (Fig 6B, 7B, see arrows). Both individuals are clones of the same BB02 strain, so they would be expected to share the same genome. A subset of the leftover PCR products will be sent for sequencing to confirm product lengths and other fragment anomalies identified by the gel analysis.

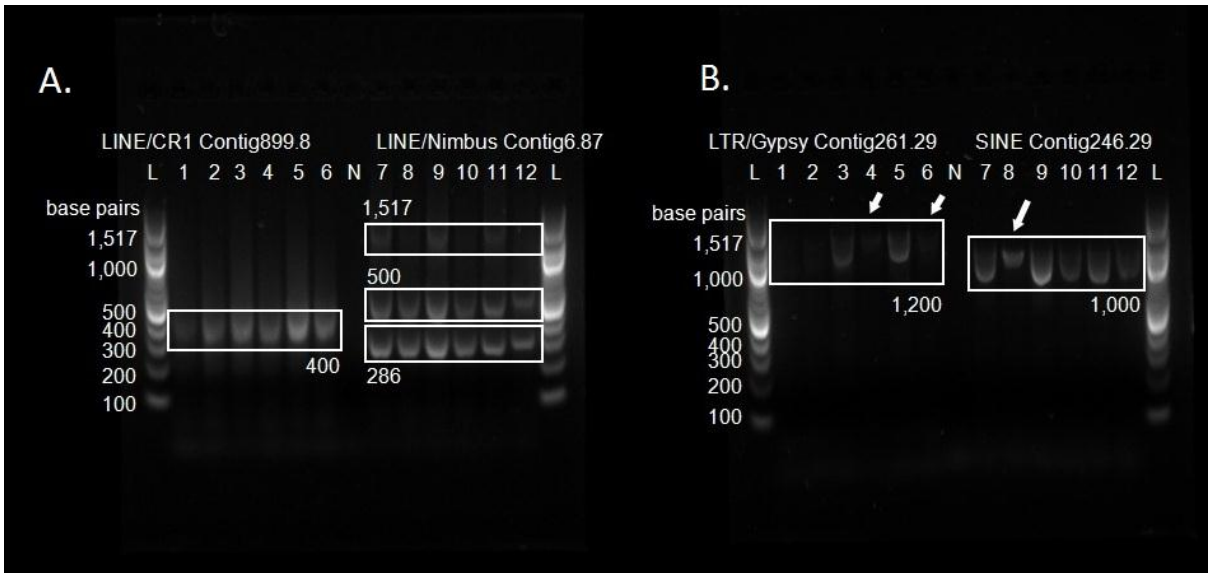

Figure 6. Gels with primers for LINE/CR1 Contig899.8 and LINE/Nimbus Contig6.87 (A), and primers LTR/Gypsy Contig261.29 and SINE Contig246.29 (B) (L = New England Biolabs 100 bp DNA ladder, N = negative control). Odd numbered lanes contained DNA from *B. glabrata* individual 1 and even numbered lanes contained DNA from *B. glabrata* individual 4. Lanes 1, 2, 7, and 8 had an annealing temperature of 51 °C; lanes 3, 4, 9, and 10 had an annealing temperature of 54 °C; and lanes 5, 6, 11, and 12 had an annealing temperature of 57 °C.

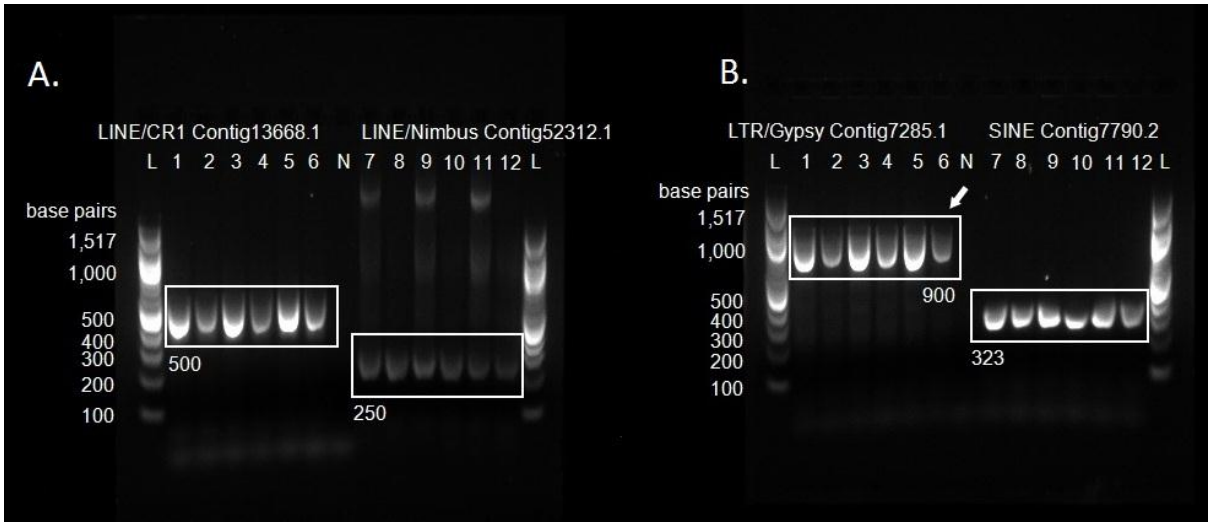

Figure 7. Gels with primers for LINE/CR1 Contig13668.1 and LINE/Nimbus Contig52312.1 (A), and primers LTR/Gypsy Contig7285.1 and SINE Contig7790.2 (B) (L = New England Biolabs 100 bp DNA ladder, N = negative control). Odd numbered lanes contained DNA from *B. glabrata* individual 1 and even numbered lanes contained DNA from *B. glabrata* individual 4. Lanes 1, 2, 7, and 8 had an annealing temperature of 51 °C; lanes 3, 4, 9, and 10 had an annealing temperature of 54 °C; and lanes 5, 6, 11, and 12 had an annealing temperature of 57 °C.

Discussion

The comparison of percent genomic coverage of Class I repeat families in the *B. glabrata* genome identified by RepeatMasker and by HMM profiles supports our initial hypothesis that *de novo* HMM profile repeat identification produces a higher percent coverage than sequence-homology based search tools such as RepeatMasker. The percent coverage of total LINEs, specifically L3/CR1s, significantly increased when using HMM profiles for repeat identification (Table 4, 5). The L3/CR1 family is incredibly diverse and not well characterized, especially in non-model organisms, so percent coverage of this repeat family is likely to increase when using *de novo* annotation. There was no difference in the percent coverage of total LINE2 elements between RepeatMasker and the HMM pipeline. However, there was a slight decrease in percent coverage of total LTR elements and a greater decline in percent coverage of total SINEs identified by the HMM pipeline compared to RepeatMasker (Table 4, 5). The HMM percent coverage of SINEs may in fact be more accurate if RepeatMasker is inflating total SINE content. This is likely due to the relatively

short length and lack of diagnostic structural features in SINEs, and the potential of processed retropseudogenes resembling SINEs that would generate false positives. This will require more testing in order to determine the most accurate method; however, the initial results of the HMM pipeline are promising, especially for annotating repeats in genomes of non-model species such as *B. glabrata*.

The PCR verification of the repeat sequences identified in the *B. glabrata* genome by HMMs provides experimental support that the *in silico* results are biologically informative. All eight of the primers amplified DNA sequences that were similar in size to the expected product (Fig 6, 7, Table 1). The smearing that is evident in the gels prevents more accurate identification of product sequence size (Fig 6, 7). In order to completely determine primer amplification accuracy, unused PCR products will need to be sequenced. This will also aid in determining whether there is a difference in sequence length between *B. glabrata* individuals 1 and 4. Based on the gel results, the DNA sequences amplified by primers LTR/Gypsy Contig7285.1, LTR/Gypsy Contig261.29, and SINE Contig246.29 appear to be of slightly different lengths between *B. glabrata* individuals 1 and 4 (Fig 6B, 7B). If the products are indeed different sequence lengths, this would provide evidence for insertions, deletions, or other mutations occurring within these specific repeat sequences between the two individuals. Finally, analyzing the PCR products using gel electrophoresis for the higher range of annealing temperatures may show higher specificity of amplification and produce fewer bands for the LINE/Nimbus Contig6.87 primer set since higher annealing temperatures tend to be more specific (Fig 6A). Otherwise, 54 °C and 57 °C appear to produce the clearest bands for all eight loci, and may be the more optimal annealing temperatures for these primers (Fig 6, 7).

As of right now, this research is an ongoing process. While some of the results support our initial hypothesis that *de novo* repeat annotation using profile HMMs will

identify a higher percent coverage of certain repeat families, we also hypothesize that profile HMMs will also identify a higher percent coverage of Unknown or Unclassified elements. However, the profile HMMs for Unknown repeat sequences still need to be generated. In this vein, we are in the process of running an internal control to verify the accuracy of the model. Finally, the program for calculating percent coverage of major repeat families will need to be modified in order to include a calculation of total number of repeats identified for each major repeat family. This research will be continued and integrated with the results of the International *Biomphalaria* Genome Sequencing Consortium.

Acknowledgements

We would like to thank the Undergraduate Research and Creative Activities (URCA) Program for providing the MAYS Grant (MA2014-009) that provided funding for ordering PCR primers and reaction solutions. We would also like to thank Coen Adema for providing the isolated snail DNA from five cloned *B. glabrata* individuals that were used in our PCR validation experiments. Finally, we would like to thank Elizabeth Cushman for providing laboratory training and optimization of PCR protocols and the Tanya Darden Lab in the SC Department of Natural Resources (DNR) Genetics Lab at Hollings Marine Laboratory for providing access to facilities for conducting this research.

References

1. Weiner AM: **SINEs and LINEs: the art of biting the hand that feeds you**. Current Opinion in Cell Biology 2002, **14**: 343-350.

2. Kazazian Jr. HH: **Mobile elements: drivers of genome evolution**. Science 2004, **303**: 1626-1632.

3. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, *et al*: **A unified classification system for eukaryotic transposable elements**. Nature Reviews 2007, **8**: 973 – 982.

4. Lowe CB, Bejerano G, Salama SR, Haussler D: **Endangered species hold clues to human evolution.** J Hered 2010, **101**(4): 437-447.

5. Jaillon O, Aury JM, Brunet F, Petit JL, Strange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A *et al*: **Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype**. Nature 2004, **431**: 946-957.

6. Shedlock AM: **Phylogenomic investigation of CR1 LINE diversity in reptiles**. Syst Biol 2006, **55**(6): 902-911.

7. Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L *et al*: **The genome of the western clawed frog *Xenopus tropicalis***. Science 2010, **328**: 633-636.

8. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S: **The genome of a songbird**. Nature 2010, **464**: 757-762.

9. Alfoldi J, Palma FD, Grabherr M, Williams C, Kong L, Mauceli E, Russell P, Lowe CB, Glor RE, Jaffe JD *et al*: **The genome of the green anole lizard and a comparative analysis with birds and mammals.** Nature 2011, **477**: 587-591.

10. Smit AFA, Hubley R, Green P, unpublished data: **Current Version: open-4.0.1 (RMLib: 20120418 & Dfam: 1.1)**. Web. < http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker>.

11. Smit A, Hubley R: **RepeatMasker.** Institute for Systems Biology 2003. Web. < http://www.repeatmasker.org/>.

12. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AFA, Finn RD: **Dfam: a database of repetitive DNA based on profile hidden Markov models.** Nucleic Acids Research 2013, **41**(D1): D70-D82.

13. Sharma KR: **Bioinformatics: Sequence Alignment and Markov Models**. New York: McGraw-Hill, 2009. Print.

14. Majoros WH: **Methods for Computational Gene Prediction**. New York: Cambridge University Press, 2007. Print.

15. Karchin R: **Hidden Markov Models and Protein Sequence Analysis**. Web. <http://compbio.soe.ucsc.edu/ismb99.handouts/KK185FP.html#hmm>.

16. Howard Hughes Medical Institute Janelia Farms Research Campus. <http://www.janelia.org/>.

17. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase update, a database of eukaryotic repetitive elements.** Cytogenet. Genome Res. 2005, **110**(1–4): 462–467.

18. Eddy SR: **Accelerated profile HMM searches**. PLoS Comput. Biol. 2011, **7**(10): e1002195.

19. Raghavan N, Knight K: **The snail (*Biomphalaria glabrata*) genome project**. Trends Parasitol 2006, **22**(4): 148-151.

20. Knight M, Adema CM, Raghavan N, Loker ES, Lewis FA, Tettelin H: **Obtaining the genome sequence of the mollusk *Biomphalaria glabrata*: a major intermediate host for the parasite causing human schistosomiasis.** National Human Genome Research Institute Sequence Proposals 2003. <http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/BiomphalariaSEQv.2.pdf>.

21. VectorBase. **Data files – *Biomphalaria glabrata***.
<https://www.vectorbase.org/downloads?field_organism_taxonomy_tid=1241&field_status_value=Current>.

22. **The current version of HMMER** <http://hmmer.janelia.org/software>.

23. **RepeatModeler** <http://www.repeatmasker.org/RepeatModeler.html>

24. Price AL, Jones NC, Pevzner PA: ***De novo* identification of repeat families in large genomes**. Bioinformatics 2005, **21**(1): i251-i258.

25. Smit AFA, Hubley R: **RepeatModeler Open-1.0**. 2008-2010.
<http://www.repeatmasker.org>.

26. **Format Converter v2.2.5**. HIV Sequence Database 2013.
<http://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html>.

27. Eddy SR: **HMMER User's Guide**. Howard Hughes Medical Institute, 2010. Web.
<ftp://selab.janelia.org/pub/software/hmmer/CURRENT/Userguide.pdf>.

28. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T: **Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction.** BMC Bioinformatics 2012, **13**:134.