Supplementary Information: Widespread Horizontal Transfer of Retrotransposons.

A.M. Walsh, R.D. Kortschak, M.G. Gardner, T. Bertozzi and D.L. Adelson

Contents

1 Methods

1.1 Software Used

For local alignments and database searches BLAST (Basic Alignment Search Tool) version 2.2.25 [1] and LASTZ (Local Alignment Search Tool, blastZ-like) version 1.02.00 [2] were used. NCBI bl2seq was used for local alignments of two sequences [3]. Global alignments were done with MUSCLE (Multiple Sequence Comparison by Log-Expectation) version 3.8.3 [5]. Global alignments were refined manually and using Gblocks version 0.91b [6]. RepeatMasker version open-3.2.6 [7] was used to find repetitive elements and to annotate sequences.

Clustering was done with UCLUST version 4.1.93 [8]. Consensus sequences were extracted with PILER version 1.0 [9], HIV sequence database Advanced Consensus Maker [10] and a Perl script shown in section 2.14. Scripts were written in Perl and made use of the BioPerl modules available (Perl version 5.10.0, 5.8.8 and 5.10.1 were used) [11]. BEDTools version 2.11.2 were used to manipulate genomic intervals.

For genomic survey sequence short read assembly, Phrap version 1.090518 [12] was used to built contigs. MEGA version 5 [13] was used to calculate overall mean distances. GENSCAN [14, 15] was used to translate a BovB sequence into protein.

For building phylogenetic trees, FastTree version 2.1.3 [16, 17]; RAxML (Randomized Axelerated Maximum Likelihood) version 7.0.4 [18]; and BEAST (Bayesian evolutionary analysis sampling trees) version 1.6.2 [19] were used. Programs in the BEAST software package were also used to construct and analyse the BEAST tree, including BEAUti (Bayesian Evolutionary Analysis Utility), TreeAnnotator and Tracer version 1.5 [20]. Model generator version 0.85 [21] was used to determine the best model for building the phylogenetic trees

and Sprit [22, 23] was used to compare phylogenetic trees.

1.2 Presence of BovB in Genbank data

From the NCBI (National Center for Biotechnology Information) taxonomy database [24], a list was compiled of genera, families, superfamilies and orders to be screened for BovB, in order to get an overall picture of the distribution of BovB across the tree of life. Due to the limited maximum number of BLAST hits returned, smaller groups, e.g. families or genera were tested where BovB was expected, such as in ruminants, and larger groups, e.g. orders, were tested where it was not expected, such as in primates.

A BioPerl module, RemoteBlast, was used (script supplied in Section 2.1) to BLAST a file containing eight improved BovB/RTE sequences against the NCBI remote BLAST Nucleotide database. The hits corresponding to the taxon name from the list were then selected out. The eight BovB/RTE sequences in the query file were the BovB sequences from the snake (*Vipera ammodytes*) (BovB VA), cow (improved consensus)(*Bos taurus*) (BovB), opossum (*Monodelphis domestica*) (BovB Opos) and platypus (*Ornithorhynchus anatinus*) (BovB Plat); the RTE2 sequences from opossum (RTE2 MD) and wallaby (*Macropus eugenii*) (RTE2 ME) and RTE1 sequences from platypus (Plat RTE1) and purple sea urchin (*Strongylocentrotus purpuratus*) (RTE1X SP).

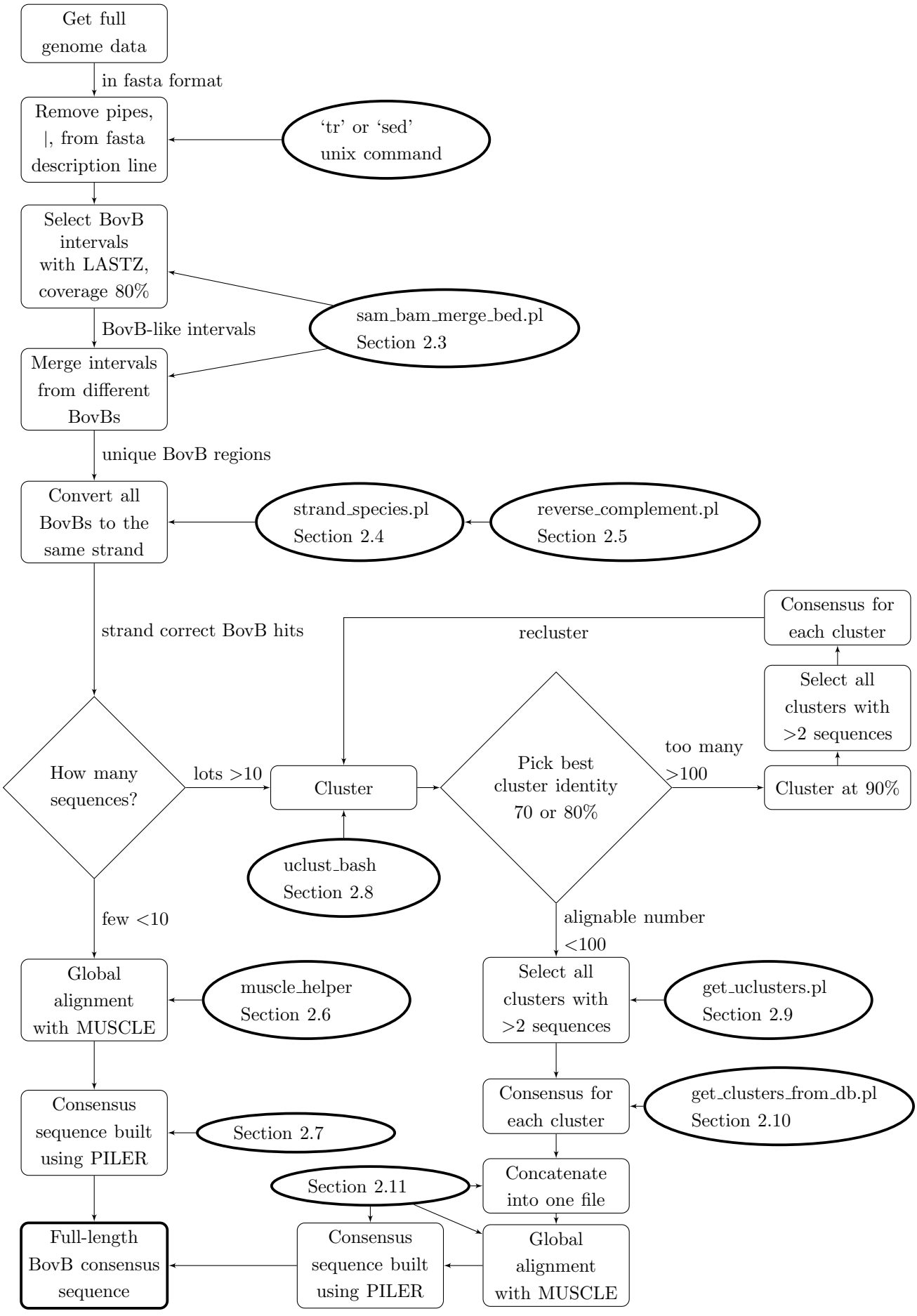
Two threshold e-values were used, e = 0 and e $\leq$ 1e-10 to identify significant hits.Significant BLAST hits were catalogued against the compiled list, seen in table 4.

In order to determine if sufficient sequence was available to infer the presence or absence of BovB in a group, the taxonomy database was queried for each of the groups and the number of available sequences ascertained.

1.3 Full Genomes search

Species where full genome data was available, shown in table 5, were searched for BovB. Scripts shown in Section 2.2 - 2.12 were used to generate full-length BovB consensus sequences for each species where BovB was found. A flow chart showing the pipeline for the analysis is shown in Fig. 1.

Figure 1: Pipeline to get nearly full-length BovBs from full genome data. Ellipses contain an indication of the command or script written to complete the task in the box. Scripts are shown in the appendices (2.2-2.12)



Script sam_bam_bed_merge_species_name, shown in section 2.3, was used to run LASTZ with 80% coverage.. BEDTools was used to process the LASTZ output and merge the intervals selected by LASTZ to get the unique fragments of the genome corresponding to BovB, as shown in Fig. 2.

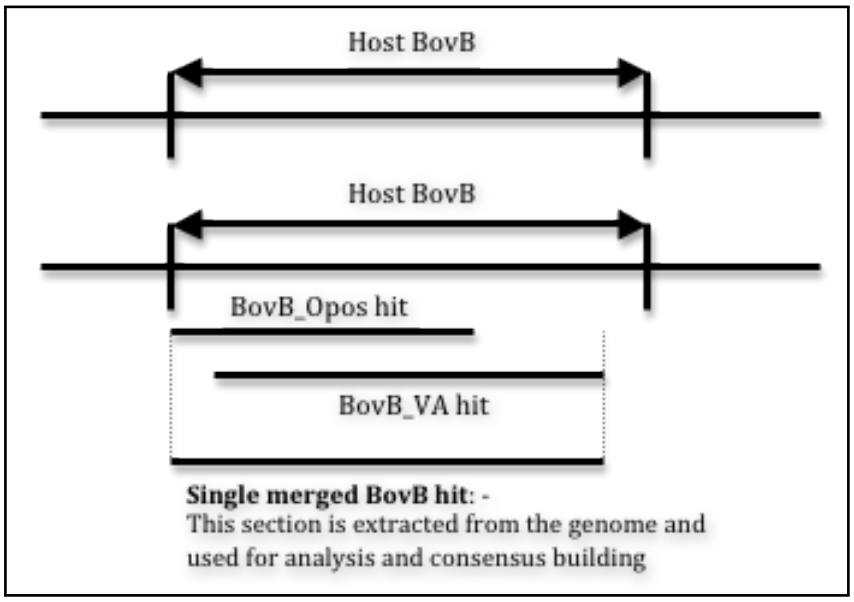


Figure 2: Illustration of potential BovB hits on a genome using LASTZ, this shows that the different BovB sequences may hit different parts of the host BovB and hence need to be merged, using BEDTools, before the host BovB can be extracted.

Script strand_species_name, in section 2.4, was then used to convert all sequences to the same strand.

Depending on the number of sequences extracted by LASTZ, the sequences were either clustered, using UCLUST, at 70 or 80% identity or directly globally aligned with MUSCLE, Section 2.6, PILER was then used to produce a consensus sequence from the alignment, Section 2.7.

If there were a large number of sequences, scripts uclust_bash (section 2.8), get_uclusters.pl (section 2.9) and get_clusters_from_db.pl (section 2.10) were used to cluster those sequences that were most similar and construct a consensus sequence for each cluster. These cluster

consensus sequence files were then concatenated together and an overall consensus sequence for the species was constructed, Section 2.11. If the initial clustering step produced very large clusters, e.g. >2000 sequences for the elephant and >600 for the cow, the sequences were clustered at 90% and consensus sequences for these clusters were constructed. These 90% cluster consensus sequences were then clustered at 80% to construct consensus sequences that were used to build the BovB for that species. The percentage used to cluster each species is shown in table 1.

Table 1: Clustering percentage for each of the species where BovB was found in the full genome sequence, scientific names can be found in table 5.

Cluster percentage	Species
No clustering	Platypus, Wallaby, Sea Urchin, Zebrafish, Silkworm
70%	Opossum, Tenrec
80%	Sheep, Anole, Horse, Rock Hyrax
90% then 80%	Cow, Elephant

Gblocks was used to refine the multiple alignments of the sequences used to build the consensus sequences and used to build the phylogenetic trees.

Once consensus sequences were built for the available full genome sequences , FastTree was used to build a phylogenetic tree using maximum likelihood methods in order to determine the relationships between the BovBs of the different species. Further information on the construction of the phylogenetic trees is in the tree method section below, section 1.8.

1.4 Annotation of BovBs

RepeatMasker was used to determine the composition of BovB VA after it was noted that the ends, <600 and >4000, were overrepresented in BLAST and RepeatMasker output when searching for BovB, particularly in birds.

RepeatMasker was also used to analyse the composition of the horse BovB full-length sequences and to test the whole horse genome to determine if contamination was likely.

1.5 Genome coverage of BovB

Species where BovB was present in the full genome data used in section 1.3 were masked using RepeatMasker to determine the amount of the genome covered by BovB.

1.6 Substitution rates and percentage identity

Overall mean distances were computed for the nearly full-length BovBs using MEGA. The Jukes-Cantor model was used with gamma distribution and 90% partial deletion of missing data. Partial deletion of missing data was used because some species, such as the elephant, had so many BovB elements that global alignments produced no common sites among all sequences.

1.7 Low coverage genomic survey sequence BovB construction

Taxa: For the 65 taxa where low coverage genomic survey sequence data were available, see section 3.3, BLAST searches, using the BovB consensus sequences as the queries, were performed to identify reads that contained BovB. The species where BLAST provided sufficient hits to attempt to build a BovB were the reptile tick (*Bothriocroton hydrosauri*), reptile tick legs (*Amblyomma limbatum*), mardo (*Antechinus flavipes*), bilby (*Macrotis lagotis*), southern bandicoot (*Isoodon obesulus*), wallaroo (*Macropus antilopinus*), central pygmy possum (*Burramys parvus*(central ESU)), northern pygmy possum

(*Burramys parvus*(northern ESU)), eastern bandicoot (*Perameles gunni*), sugar glider (*Petaurus breviceps*), sea snake (*Hydrophis spiralis*), tree dragon (*Amphibolurus norrisi*), LHI skink (*Oligosoma lichenigerum*), *Leposoma scincoides*, Ca skink (*Ctenotus atlas*), Er skink (*Eremiascincus richardsonii*), Gd skink (*Glaphyromorphus douglasi*), G laz gecko (*Gehyra lazelli*), G var gecko (*Gehyra variegata*), and Howe Island gecko (*Christinus guentheri*). Due to the low coverage data and limited number of taxa, particularly reptiles, available, three additional species were sequenced. One gigabase of clean sequence data was sequenced by BGI (Beijing Genomics Institute) for the sleepy lizard (*Tiliqua rugosa*), which is the host of the two tick species being examined, Stoke's skink (*Egernia stokesii*) and the echidna (*Tachyglossus aculeatus*). One gigabase of sequence data represents a substantial fraction of the genomes in question, specifically about 1/3 genome coverage for the Echidna. Sequence reads in this data collection were each 100bp long, hence the initial BLAST step, used for the other data was skipped. Where possible, full-length BovBs were assembled from these sequences.

RepeatMasker: First RepeatMasker was run on all the data with the compiled BovB library including the four BovB sequences from the improved BovB file and the full-length BovBs built using the full genome search method described in section 1.3. For reptiles the library was modified to be free of the CR1 repeats that are incorporated onto the end of BovB VA. This was done by removing the first 650bp and the last 550bp of the BovB VA sequence. This was necessary because of the difficulty in assembling BovBs when a significant proportion of the reads used for assembly belong to a different repeat sequence.

Quality Control: Once the reads had been masked using RepeatMasker, the script RM_QC_for_phrap.pl, in section 2.13, was used to select out reads that masked as BovB over a percentage of their lengths. Initially 60% coverage was used as the cut off for BovB

masking but it was increased to 80% for those sequences that required more stringent conditions to build a BovB of sufficient quality for phylogenetic analysis. Species where 80% coverage was used were the tree dragon, mardo, bilby, southern bandicoot, and the three taxa sequenced by BGI because the reads were 100bp long.

Phrap Contigs: Phrap, a program for assembling shotgun DNA sequence data, was then used to build contigs from the reads. For most of the species the default parameters for Phrap were used, but for a few, more stringent parameters were needed to built a BovB of sufficient quality for phylogenetic analysis. A BovB of sufficient quality was defined as a BovB sequence that produced a good global alignment and a robust tree position when introduced to the BovB library or tree; it was of a similar length to the other sequences and so did not increase the total length of the alignment by more than 500bp. The more stringent parameters used for Phrap were penalty -15, shatter_greedy, bandwidth 30 and minscore 100. The more stringent Phrap conditions were used in the construction of the mardo, bilby and southern bandicoot BovBs. BovBs of sufficient quality could not be constructed for the LHI skink and *Leposoma scincoides*.

Quality Control 2: Once contigs had been built they were masked using RepeatMasker and the RM_QC_for_phrap.pl script was run to determine if they were masking as BovB over the percentage of their length that their reads were required to, for example 60% for the sea snake and 80% for the bilby.

Clustering: If Phrap built many contigs that masked as BovB over a high percentage of their length, they were clustered using UCLUST. The percentage identity with which the contigs were clustered varied between species, according to what percentage identity was

needed to produce clusters with >1 sequence, how many gaps the BovB produced by a cluster introduced to the global alignment and how many sequences were present in clusters that had to be manually curated. The wallaroo, central pygmy possum, northern pygmy possum, eastern bandicoot, sugar glider, sea snake and G var gecko were clustered at 70% identity. Ca skink, Gd skink, Er skink and G laz gecko were clustered at 80% identity and the tree dragon was clustered at 90% identity. The Howe Island gecko, mardo, bilby, southern bandicoot and two ticks were not clustered due to the small number of contigs built.

The BGI data were not clustered. However the number of contigs used to build the consensus sequence was reduced by selecting only the long contigs. For the two skinks this was contigs >500bp long and for the echidna this was contigs >1kb long.

Alignment and Scaffolding: As the contigs often masked as different regions of BovB, for example contig 1 might mask as the first 1kb of BovB Opos and contig 2 might mask as the last 1kb of Sheep_BovB, each sequence was aligned using MUSCLE with its corresponding BovB as a scaffold. Its corresponding BovB being the BovB used to mask it. These pairwise alignments were then aligned. The scaffolds were removed while manually curating the alignment. Alignments were also manually curated to remove short insertions present in one sequence that were absent in several others and to fix errors in the MUSCLE output, such as the one shown in Fig. 3. Insertions and deletions were an issue when consensus building particularly because we were dealing with repeats. The consensus being built was not from many sequencing runs of the same gene/region,but rather from distinct regions that have been evolving and mutating independently for some time. This process was an attempt to assemble and align them at the same time. Manual checking also ensured that the scaffold aligning process placed the contig approximately where Re-

peatMasker predicted it should be, for example if it masked as a 5′ part of BovB Opos it was not placed at the 3′ end of the alignment.

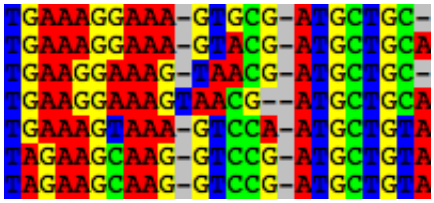


Figure 3: **MUSCLE error example:** One example of a misalignment found in the global alignment of sequences during consensus and tree building. These misalignments were corrected by manual curation.

Consensus Construction: Once the alignment was manually curated, consensus sequences were built in one of two ways. The first was using the HIV sequence database Advanced Consensus Maker. The other was the Perl script, cons.pl, shown in section 2.14. The HIV consensus builder was used before cons.pl was written. They use the same principle to build consensus sequences and differ only in their assignment of ambiguous bases, hence the consensus sequences built with HIV consensus builder were not rebuilt after cons.pl was written. Cons.pl was written so that it could be included in an automated pipeline.

1.7.1 Tenrec and Rock Hyrax

The tenrec and rock hyrax BovB sequences from the full genome method in section 1.3 were improved by taking the BovB sequences produced by LASTZ, RepeatMasking them, then running RM_QC_for_phrap.pl with 80% cutoff. This was done because these genomes were only partially assembled and hence produced shorter strand corrected BovB sequences than the other genomes. The sequences were then run through Phrap with the stringent conditions above. The tenrec sequences produced four contigs that masked as Elephant_BovB, these were each aligned with MUSCLE against Elephant_BovB as a scaffold, then all of

them were aligned together with MUSCLE and manually curated. From this cons.pl was used to extract a better consensus sequence than was previously constructed by the full genome BovB building method, in section 1.3. The rock hyrax sequences produced no contigs when Phrap was run, but after using RepeatMasker to filter out the poor quality sequences they were clustered at 80% identity and aligned. This alignment was manually curated and the HIV Advanced consensus builder was used to extract a consensus sequence that introduced fewer gaps into the multiple alignment used for tree building, compared to the previous rock hyrax BovB consensus sequence.

1.7.2 Control

Due to concern that the process, particularly the profile aligning of sequences, could cause a BovB to be built for a species that did not have BovB. The rat and brown toadlet BLAST hits were tested to determine if a BovB could be built. The methods described above did not produce BovB consensus sequences for these species, supporting the validity of our methodology.

1.7.3 PCR Verification of critical sequences

DNA was extracted from frozen or ethanol preserved tissue using a Puregene DNA isolation kit (Gentra Systems, Minneapolis, MN) following the manufacturers protocol for DNA purification from solid tissue. PCR was used to amplify single reads from the 5' and 3' ends of a contig consensus from each of the individuals in Table 2 using primers outlined in Table 3, which were developed using Primer3 [4]. Each PCR was carried out in a volume of 25 μl with a final concentration of 1X GeneAmp PCR Gold buffer, 2 mM MgCl2, 200 μM of each dNTP, 0.2 μM of each primer and 0.5 U of AmpliTaq Gold DNA polymerase

(Applied Biosystems, Foster City, CA). Amplifications consisted of an initial denaturation step of 94 °C for 9 min, followed by 34 cycles of PCR with the following temperature profile: denaturation at 94 °C for 45 s, annealing at 55-60 °C for 45 s, and extension at 72 °C for 1 min, with an additional final extension at 72 °C for 6 min. The double-stranded amplification products were visualised on 1.5% agarose gels and purified using Multiscreen PCR clean-up pates (Millipore Corporation, MA) before cycle-sequencing in both directions using the BigDye Terminator v3.1 cycle-sequencing kit (Applied Biosystems).The cycling protocol consisted of 25 cycles of denaturation at 96 °C for 30 s, annealing at 50 °C for 15 s, and extension at 60 °C for 4 min. All samples were sequenced on an Applied Biosystems 3730xl DNA sequencer.

All primer combinations produced a single amplicon of the expected size.

Specimen no	Taxon	Tissue	GenBank
ABTC123569	Equus caballus	blood	pending
AMSR90203	Christinus guentheri	liver	pending
ABTC111481	Amblyomma limbatum	legs	pending
ABTC123615	Bothriocroton hydrosauri	legs	pending
ABTC82613	Gehyra variegata	liver	pending

Table 2: Species used for PCR verification, AMS is an Australian Museum label and ABTC is a Australian Biological Tissue Collection, South Australian Museum label.

Primer	Primer sequence	Species	Annealing °C
G2250F	TGTGGGACGCCTGCCAAAGC	Equus caballus	60
G2251R	GTGTGGCACGCCGTGGGAC	Equus caballus	60
G2252F	GGCACATTGCGAGAAGGCAGGAC	Equus caballus	60
G2253R	AAAGCCATCACCCTTGACAGAGCCAG	Equus caballus	60
G2254F	CGCGAGACCATCCTCTCACAC	Amblyomma limbatum	55
G2255R	GGCAGAGACGCTGGAGTGAGT	Amblyomma limbatum	55
G2256F	GATAGATGGTGGAGGACAGGAAGG	Amblyomma limbatum	55
G2257R	GCATGAGGCGAAACAATGAGAA	Amblyomma limbatum	55
G2258F	CTCTCATCCTGCCCACTGACTC	Bothriocroton hydrosauri	55
G2259R	CCCCAGTAGCATAGTGGACACCTT	Bothriocroton hydrosauri	55
G2260F	AACGCCAGATTTCAAGACTGAACA	Bothriocroton hydrosauri	55
G2261R	TGGGGCGTAGGCTTGGACT	Bothriocroton hydrosauri	55
G2262F	AGCCACAGCCCTTAGTCTGC	Christinus guentheri	55
G2263R	GCTCCTCCTATTTGCCCATCTAT	Christinus guentheri	55
G2277F	AAAGGTCAGTTTACATCCCAATC	Gehyra variegata	55
G2278R	TCTCTTGAAGGACTTGCCATAG	Gehyra variegata	55

Table 3: Primers used for amplifying the BovB sequences from the species named.

1.8 Trees

1.8.1 Tree using BovB sequences

Trees were initially built with FastTree using defaults, or the general time reversible (GTR) model with gamma approximation on substitution rates. For the final tree the FastTree output was compared with the output produced from RAxML and BEAST. The trees were built from multiple alignments done by MUSCLE using the default parameters and from a version of this alignment that had been refined using Gblocks. For the final tree, FastTree was run with the GTR model using gamma approximation for substitution rates, so that all the trees could be compared using the same model. RAxML was run with 500 bootstraps using the substitution model GTRGAMMA. Model generator was used to determine that the GTR model with gamma rates was the best fit for the data when four rate categories were used. BEAUti was used to set up the BEAST MCMC (Markov chain Monte Carlo) run with the Tree Prior set to 'Speciation: Yule process'. For BEAST MCMC a chain of length 100,000,000 was used, sampling every 10,000 to produce 10,000 trees of which the

first 1,000, or 10%, were ignored (burnin value) when using TreeAnnotator to generate the best tree. This burnin value was verified using the program Tracer that showed that 10% burnin was sufficient to allow convergence.

1.8.2 Tree using orthologous sequences

For use as a control, a phylogenetic tree built from orthologous sequences was required. This was obtained from "OrthoDB: Database of Orthologous Groups" [25], supplied to us by Dr Evgeny Zdobnov. This tree contained only one non-avian reptile, the green anole lizard, so the breakdown of reptiles was determined using the TimeTree of Life publication [26, 27]. This publication provided the currently accepted breakdown of reptiles, which was used to replace the anole in the control tree built from orthologous sequences, to allow for analysis of the number of horizontal transfers.

1.8.3 BovB vs control tree comparison

Sprit was used to compare the control tree built from the orthologous sequences and the tree built from BovB sequences by estimating the number of horizontal transfers required to get the observed topology. Sprit calculated the minimum subtree prune and regraft (SPR) distance between phylogenies.

1.9 Exaptation

The protein sequence for BovB VA was found using GENSCAN. This sequence was used to determine if any part of the BovB repeat had been exapted into a gene in order to contribute to the protein coding content of the species. This was done by using the BLAST function on

UniProt to BLAST the BovB VA protein sequence against the SwissProt/UniProt protein sequence database [28, 29] in search of expressed BovB-like protein sequences.

2 Scripts

2.1 blastNCBI.pl

This script automates the identification of BovB sequences using megaBLAST. MegaBLAST requires a query to BLAST against the sequences in the subject. This script has the query set to the eight sequences in the improved BovB file and the sequences in the Nucleotide database that match the supplied taxon name as the subject. This program is currently set with its cutoff value at e=1e-10. If there are BLAST hits with e-values $\leq$1e-10 all BovB blast hits for that query and taxon will be written to an output file.

```
#!/usr/bin/perl -w
use Bio::Tools::Run::RemoteBlast;
use strict;
die "Useage: $0 <taxon><wordsize>\n" unless @ARGV>0;
my ($taxon, $wordsize) = @ARGV;
if ($wordsize !~ '\d+'){
        $wordsize = 16;
}
my $prog = 'blastn';
my $service = 'megablast';
my $db = 'nr';
my $e_val = '1';
my $penalty = '-1';
my $reward = '1';
my $other = '-G 5 -E 2';
my $query = '/Users/labadmin/Databases/BovB_improved.mfa';
my $entrez = '"'.$taxon.'"[Organism]';
print STDOUT "\nentrez query = ".$entrez."\n".$taxon."\n";
my @params = ( '-prog' => $prog,
                '-data' => $db,
                '-expect' => $e_val,
                '-service' => $service,
                '-word_size' => $wordsize,
                '-other_advanced' => $other,
                '-nucl_penalty' => $penalty,
                '-nucl_reward' => $reward,
                '-entrez_query' => $entrez );
my $fac = Bio::Tools::Run::RemoteBlast->new(@params);
my $v = 1;
my $r = $fac->submit_blast($query);

#code modified from http://doc.bioperl.org/releases/bioperl-1.6.1/
```

```
my $top_dir = "1_4_11";
mkdir $top_dir;
print STDERR "waiting..." if( $v > 0 );
my $dirname = $top_dir."/".$taxon;
print $dirname."\n";
while ( my @rids = $fac->each_rid ) {
        foreach my $rid ( @rids ) {
                my $rc = $fac->retrieve_blast($rid);
                if( !ref($rc) ) {
                        if( $rc < 0 ) {
                                $fac->remove_rid($rid);
                        }
                        print STDERR "." if ( $v > 0 );
                        sleep 5;
                } else {
                        my $result = $rc->next_result();
                        my $good_hit = 0;
                        my $e_cutoff = 1e-10;
                        print "\nQuery Name: ", $result->query_name(), "\n";
                        while ( my $hit = $result->next_hit ) {
                                next unless ( $v > 0);
                                while( my $hsp = $hit->next_hsp ) {
                                        if($hsp->evalue <= $e_cutoff){
                                                print "\thit name is ",$hit->name,"\n";
                                                print "\t\tscore is ",$hsp->score,"\n";
                                                $good_hit = 1;
                                        }
                                }
                        }
                        my $filename =
   $dirname."/".$result->query_name()."_ce".$e_cutoff."_w".$wordsize."\.blast";
                        if($good_hit){
                                mkdir $dirname;
                                $fac->save_output($filename);
                        }
                        $fac->remove_rid($rid);
                }
        }
}
```

2.2 count_any_program.pl

This script was used to run any of the scripts below, where the input parameter, $@, needs to be a range of numbers. For example when a program needs to be run on all 21 chromosomes or all 200 scaffolds.

```
#!/usr/bin/perl -w
die "Useage <start_value><end_value><program>" unless @ARGV>2;
my ($start_val, $end_val, $program) = @ARGV;
while($start_val<=$end_val){
        system("./".$program." ".$start_val);
        $start_val++;
}
```

2.3 sam_bam_bed_merge

This script uses LASTZ to identify the BovB interval locations in the genome, then merges the locations using BEDTools and selects out the unique coordinates in the genome that correspond to the BovB hits. This is the opossum version of the sam_bam_bed_merge script. This script was run on all chromosomes, e.g. use count_any_program.pl to run it from 1 - 8 then run it on the x chromosome and the chromosome unknown file.

```
lastz /export/genome/data/opossum/chr$@.fa[unmask] ../BovB/BovB_only.fasta[unmask]
        --chain --gapped --coverage=80 --format=sam >Opos_BovB_chr$@_80.sam

samtools view -b -o Opos_BovB_chr$@_80.bam -S Opos_BovB_chr$@_80.sam
bamToBed -i Opos_BovB_chr$@_80.bam >Opos_BovB_chr$@_80.bed
mergeBed -s -nms -i Opos_BovB_chr$@_80.bed >Opos_BovB_chr$@_80.merged.bed
fastaFromBed -fi /export/genome/data/opossum/chr$@.fa
        -bed Opos_BovB_chr$@_80.merged.bed -fo Opossum/BovB_chr$@.fasta
```

2.4 strand_anole

This script selects the LASTZ hits that are on the minus strand that need to be reverse complemented and then runs the reverse complement perl script, shown below, section 2.5. This is the anole version of the program, must be run on all chromosomes or scaffolds, e.g. `count_any_program.pl 1 6 strand_anole`.

```
grep -h -w '+' Anole_hits/Anole_BovB_scaf$@_80.merged.bed |fastaFromBed
        -fi ~/anole/scaf_$@.fasta -bed stdin -fo Anole/BovB_plus_scaf$@.fasta
```

```
grep -h -w '-' Anole_hits/Anole_BovB_scaf$@_80.merged.bed |fastaFromBed
        -fi ~/anole/scaf_$@.fasta -bed stdin -fo Anole/BovB_minus_scaf$@.fasta

perl reverse_comp.pl Anole/BovB_minus_scaf$@.fasta Anole/BovB_REVCOMP_scaf$@.fasta

cat Anole/BovB_REVCOMP_scaf$@.fasta Anole/BovB_plus_scaf$@.fasta
        >Anole/strand_correct_BovB_scaf$@.fasta
```

2.5 reverse_comp.pl

This script calculates the reverse complement of a DNA strand that is passed to it as input.

```
#!/usr/bin/perl -w
use strict;
use Bio::Seq;
use Bio::SeqIO;
die "Useage: $0 <input fasta file><output>" unless @ARGV>1;
my ($in, $out) = @ARGV;
unlink $out;
my $seqin = Bio::SeqIO->new( -format => 'Fasta' ,-file => $in);
my $seqout= Bio::SeqIO->new( -format => 'Fasta', -file => '>>'.$out);

while((my $seqobj = $seqin->next_seq())) {
        if( $seqobj->alphabet eq 'dna') {
                my $rev = $seqobj->revcom;
                my $id  = $seqobj->display_id();
                $id  = "$id.rev";
                $rev->display_id($id);
                $seqout->write_seq($rev);
        }
}
```

2.6 muscle_helper

This script performs the initial MUSCLE alignment on the output from above when no clustering is required.

```
cat strand_correct_BovB_chr* >all_sc_BovB_$@.fasta
muscle -in all_sc_BovB_$@.fasta -out $@_BovB_aligned_sc.fasta &
muscle -in all_sc_BovB_$@.fasta -out $@_BovB_aligned_sc.clw -clw &
```

2.7 PILER

From the MUSCLE output above a consensus sequence can be generated using PILER as shown below.

```
piler -cons $@_BovB_alinged_sc.fasta -out $@_consensus.fasta -label $@_cons
```

2.8 uclust_bash

For species where there are large numbers of hits this script performs clustering on the BovB hits at 70 and 80%.

```
usearch --sort all_sc_BovB_$@.fasta --output sorted.fasta
usearch --cluster sorted.fasta --id 0.8 --seedsout seeds_8_sorted.fasta
        --uc results_8_sorted.fasta
usearch --cluster sorted.fasta --id 0.7 --seedsout seeds_7_sorted.fasta
        --uc results_7_sorted.fasta
```

2.9 get_uclusters.pl

This script selects out the ids for all the BovBs that formed clusters with more than 2 elements (it can be set to more than 1 as well) when results_#_sorted.fasta is fed to it. Normally 80% clusters were used, #=8, but sometimes other percentage identities were used. This script saves a list of ids into a folder, called cluster_No, where No is the cluster number produced by uclust, so that the sequences can be extracted by the next script, get_clusters_from_db.pl, Section 2.10.

```
#!/usr/bin/perl -w
while (<>) {
   /^(H|S|C)\t(\d+)\t(\d+)+\t[^\t]+\t[^\t]+\t[^\t]+\t[^\t]+\t[^\t]+\t([^\t]+)\t.*$/;
   if(defined($1)&&defined($2)&&defined($3)&&defined($4)&&($1 eq 'H' || $1 eq 'S')){
      my $filename = 'cluster_'.$2;
      open(FILE,  ">>$filename");
      print FILE "$4\n";
      close(FILE);
   }
   if(defined($1) && defined($2) && defined($3) && defined($4) && ($1 eq 'C')){
      if($3<=2){
         system("rm cluster_$2");
         open(LEFTOVER, ">>unclustered");
         print LEFTOVER "$4\n";
```

```
        close(LEFTOVER);
      }
   }
}
```

2.10 get_clusters_from_db.pl

For this script the fasta file containing the sequence data must be formatted. The $@ parameter here is the database name in the next program, e.g. elephant or anole.

```
formatdb -p F -o T -i all_sc_BovB_$@.fasta -n $@
```

This script takes the clusters of BovB ids produced by get_uclusters.pl, Section 2.9 and selects the sequences out of the database, formed from all_sc_BovB_dbname.fasta above, and builds a consensus sequence for each cluster. This program needs to be run over all clusters, e.g. count_any_program.pl 0 biggest_cluster_number get_clusters_from_db.pl.

```
#!/usr/bin/perl -w
use strict;
die "Useage <start cluster><end cluster><database>\n"unless @ARGV>1;
my($start_val, $end_val, $database) = @ARGV;

while($start_val<=$end_val){
system("fastacmd -d ".$database." -p F -i cluster_".$start_val."
        -o cluster_".$start_val.".fasta");

system("muscle -in cluster_".$start_val.".fasta
        -out cluster_".$start_val."_mult_aligned.clw -clw");
system("muscle -in cluster_".$start_val.".fasta
        -out cluster_".$start_val."_mult_aligned.fasta");
system("piler -cons cluster_".$start_val."_mult_aligned.fasta
        -out ".$database."_cluster_".$start_val."_consensus.fasta
                -label ".$database."_cluster_".$start_val."_cons");
        $start_val++;
}
```

2.11 Concatenate, Alignment and Consensus

Next all of the consensus sequences for the clusters had to be concatenated into one file. The sequences were multiple aligned using MUSCLE and PILER was used to get a consensus sequence for the species.

```
cat database_name_cluster_*_consensus.fasta >species_all_cluster_cons.fasta
muscle -in species_all_cluster_cons.fasta
        -out species_all_cluster_cons_mult_aligned.fasta
piler -cons species_all_cluster_cons_mult_aligned.fasta -out species_consensus.fasta
        -label species_cons
```

2.12 Gblocks

2.12.1 Gblocks consensus

Gblocks was used on the cluster multiple alignments or on the cluster consensus sequence multiple alignment to get better consensus sequences.

Then the script below was used to get the consensus sequences from the Gblocks output.

```
piler -cons $@_mult_aligned.fasta-gb -out $@_gblocksHalf_consensus.fasta
        -label $@_gblocksHalf_cons
```

2.12.2 Gblocks tree

Gblocks was also used on final tree alignments.

Concatenate all the BovB sequences into one file.

```
cat *_consensus.fasta >tree1.fasta
```

Multiple aligning them with MUSCLE.

```
muscle -in tree1.fasta -out tree1_mult_aligned.fasta
```

Run Gblocks on tree1_mult_aligned.fasta to get tree1_mult_aligned.fasta-gb. Then build a tree using FastTree.

```
FastTree -nt tree1_mult_aligned.fasta-gb >tree1_gblocks.tree
```

This produced a tree where only the parts of the multiple alignment that were shared by most species were considered by the maximum likelihood tree building method.

2.13 RM_QC_for_phrap.pl

Script for extracting sequences that masked as BovB over a percentage of their length. Currently designed to select out reads that mask as something from the compiled BovB library over 60% of their length.

```
#!/usr/bin/perl -w
use strict;
use Bio::SimpleAlign;
use Bio::SeqIO;
use Bio::Seq;
use Bio::LocatableSeq;
use Bio::DB::Fasta;

die "Useage $0 <reads_file> <RepeatMasker_file> <output_file> \n" unless @ARGV >2;
my($reads_file, $RM_file, $out_file) = @ARGV;

my $out = Bio::SeqIO->new(-file => ">".$out_file, -format =>'fasta');
tie(my %sequences,'Bio::DB::Fasta',$reads_file);

#read in RM file
open (RepeatMasker_file, $RM_file) || die("Couldn't open RepeatMasker file\n");
my @RMfile = <RepeatMasker_file>;
my ($id,$start,$end,$left,$comp,$repName,$repClass,$repstart,$repend,$repleft)=0.0;
my $line;
foreach $line(@RMfile){
   if(($id,$start,$end,$left,$comp,$repName,$repClass,$repstart,$repend,$repleft)
=($line=~ m!^\s*\S+\s+\S+\s+\S+\s+\S+\s+(\S+)\s+(\d+)\s+(\d+)\s+\((\d+)\)\s+(\S)\s+
(\S+)\s+(\S+)\s+\(?(\d+)\)?\s+(\d+)\s+\(?(\d+)\)?.*$!)){

      if((($end-$start)/($end+$left) > 0.6) && $repClass eq "Unknown"){
         if($comp eq '+'){
            my $seq_read = Bio::Seq->new( -seq => $sequences{$id}, -id =>$id);
            $out->write_seq($seq_read);
         }else{
            my $seq_read = Bio::Seq->new( -seq => $sequences{$id}, -id =>$id);
            my $rev = $seq_read->revcom();
            $rev->display_id($id.".rev");
            $out->write_seq($rev);
         }
      }
   }
}
```

2.14 cons.pl

Perl script to build a consensus sequence that ignores gaps when choosing the best base for a position.

```
#!/usr/bin/perl -w
use Bio::SimpleAlign;
use Bio::AlignIO;
die "Useage <alignment_file>" unless @ARGV >0;
my ($infile, $name) = @ARGV;

my $in = Bio::AlignIO->new(-format => 'fasta', -file => $infile);
my $aln = $in->next_aln();
print ">".$name."\n".$aln->consensus_string()."\n";
```

3 Supplementary Material

3.1 BovB presence across the tree of life

Table 4: **Presence of BovB across the tree of life:** This table shows the presence of BovB in taxa throughout the tree of life as determined by BLAST searching the data available on NCBI, approximately 430,000 taxa. "**" indicates presence of BovB with e-value of 0.0; "*" indicates presence of BovB with e-value ≤1e-10; "?" indicates that BovB was expected in this taxa, from the literature, but not found; and "-" indicates BovB was not expected and not found. The e-value columns show which BovB/RTE sequence produced blast hits for that species at that cut off value. O = BovB Opos; V = BovB VA; C = BovB; PB = BovB Plat; O2 = RTE2 MD; W2 = RTE2 ME; PR = Plat RTE1; XSp = RTE1X SP; "All 8" indicates that all 8 sequences above were present; "4RTEs" indicates that O2,W2,PR and XSp were present and "4BovBs" indicates that O, V, C and PB were present. No. Nuc. seqs column is the number of nucleotide sequences available in the NCBI Nucleotide database for that group. Notes column provides additional information about the BLAST hit observed, such as which species within a large group has the BovB hit or if the hit is small or low complexity such as microsatellite DNA.

	e-value=0.0	e-value≤1e-10	No. Nuc. seqs	notes
Prototheria/Monotremes				
**Ornithorhynchidae	PB	4BovBs, PR	267,336	
*Tachyglossidae	-	PB, PR	313	
Metatheria/Marsupials				
*Dasyuromorphia	-	V, O, O2, W2	34,005	
**Didelphimorphia	O, O2	All 8	30,851	
**Diprotodontia	V, O, C, O2	All 8	132,851	
*Microbiotheria	-	V, O, O2, W2	192	
*Notoryctemorphia	-	V, O, O2, W2	63	
?Paucituberculata	-	-	127	
*Peramelemorphia	-	O2	349	
Eutheria				
Laurasiatheria				
-Insectivora	-	-	10,892	
*Perissodactyla	-	O, C, PB	70,093	horse
-Pholidota	-	-	287	
-Chiroptera	-	-	76,630	
-Carnivora	-	-	665,638	
Cetartiodactyla				
-Suina	-	-	493,784	
-Hippoptamidae	-	-	398	

Continued. . .

	e-value=0.0	e-value≤1e-10	No. Nuc. seqs	notes
-Cetacea	-	-	49,430	
-Camelidae	-	-	61,033	
Ruminantia				
**Tragulina	V, O, C	4BovBs	147	
?Moschidae	-	-	463	
**Giraffidae	V, O, C	4BovBs	318	
*Antilocapridae	-	O	125	
**Cervidae	V, O, C	4BovBs	7,665	
*Aepycerotinae	-	V	288	
*Alcelaphinae	-	V	528	
*Antilopinae	-	O, C	1,794	
*Cephalophinae	-	V, O	437	
?Hippotraginae	-	-	637	
?Peleinae	-	-	12	
*Reduncinae	-	C	287	
Bovinae				
*Bison	-	V, O, C	850	
**Bos	4BovBs	4BovBs	187,616	
**Bubalus	V, O, C	4BovBs	3,730	
*Tragelaphus	-	4BovBs	1,185	
?Other Bovinae	-	-	652	
Caprinae				
*Budorcas	-	V, O	52	microsatellite hits
**Capra	V, O, C	4BovBs	7,864	
*Ovibos	-	V, C	342	microsatellite hits
**Ovis	4BovBs	4BovBs	13,663	
?Other Caprinae	-	-	1494	
Afrotheria				
*Tenrecidae	-	4BovBs	777	

Continued...

	e-value=0.0	e-value≤1e-10	No. Nuc. seqs	notes
**Proboscidea	V, PB	4BovBs	28,317	
?Chrysochloridae	-	-	140	
*Sirenia	-	V, PB	658	
*Hyracoidea	-	4BovBs	146,835	
?Tubulidentata	-	-	104	
?Macroscelidea	-	-	759	
Euarchontoglires				
-Dermoptera	-	-	357	
-Scandentia	-	-	1,298	
**Haplorrhini	C	O, C	10,862,839	human construct, marmoset predicted gene
-Strepsirrhini	-	-	282,724	
-Lagomorpha	-	-	153,739	
Rodentia				
-Hystricognathi	-	-	31,659	
**Sciurognathi	V, O, C	V, O, C	1,777,441	brown rat construct, springhare SINEs
-Xenarthra	-	-	423,202	
Sauropsida/Reptiles				
Squamata/Snakes and Lizards				
**Iguania	4BovBs	4BovBs	50,738	
Scleroglossa/Lizards				
*Anguimorpha	-	V	1,720	
?Amphisbaenia	-	-	684	
Scincomorpha/Skinks				
**Lacertoidea	V, O, C	PB	9,709	
*Scincoidea	-	V, O	18,928	
?Teiioidea	-	-	1,512	
Gekkota/Geckos				
*Gekkonidae	-	V, O, C	12,172	
?Dibamidae	-	-	143	

Continued...

	e-value=0.0	e-value≤1e-10	No. Nuc. seqs	notes
?Phyllodactylidae	-	-	1,638	
?Pygopodidae	-	-	230	
Serpentes/Snakes				
?Acrochordoidea	-	-	52	
?Typhlopoidea	-	-	1153	
Henophidia				
**Boidae	4BovBs	4BovBs	1,090	
**Pythonidae	V, O, C	V, O, C	727	
?Other Henophidia	-	-	523	
Colubroidea				
**Viperidae	4BovBs	4BovBs	22,571	
*Hydrophiidae	-	V	751	
**Elapidae	V, O, C	V, O, C	3,328	
**Colubridae	V, O, C	V, O, C	11,754	
?Atractaspididae	-	-	158	
Sphenodontia/Beaked Reptiles				
?Sphenodontidae	-	-	442	
Archosauria				
-Crocodylidae	-	-	2,735	
*Dinosauria	-	V	410,239	birds, chicken repeat region hits
Testudines/Turtles				
**Cryptodira	O2	V, 4RTEs	12,117	
*Pleurodira	-	V	1,584	
Amphibia				
*Anura	-	O	189,450	Rana nigromaculata microsatellites
-Caudata	-	-	25,723	
-Gymnophiona	-	-	1,184	
Other Eukaryotes				
*Coelacanthimorpha	-	PR, W2	250,864	Latimeria menadoensis

Continued...

	e-value=0.0	e-value≤1e-10	No. Nuc. seqs	notes
-Dipnoi	-	-	458	
*Actinopterygii	-	V, O, PB, PR, O2, W2	1,062,267	Zebrafish
-Chondrichthyes	-	-	49,460	
-Hyperoartia	-	-	4,660	
*Hyperotreti	-	V, XSp	1,021	Inshore Hagfish, short gappy hit
-Cephalochordata	-	-	31,376	
-Tunicata	-	-	45,784	
-Chaetognatha	-	-	707	
*Echinodermata	-	V, PB, XSp	421,219	Purple Sea Urchin and Slate Pencil Urchin
-Hemichordata	-	-	75,696	
-Xenoturbellida	-	-	60	
*Protostomia	-	All 8	4,710,139	includes silkworm and other insects
-Acoelomata	-	-	197,740	
-Pseudocoelomata	-	-	452,369	
-Bilateria incertae sedis	-	-	75	
-Cnidaria	-	-	322,198	
-Ctenophora	-	-	5,423	
-Porifera	-	-	32,173	
-Placozoa	-	-	14,849	
-Mesozoa	-	-	112	
-Fungi	-	-	2,801,399	
-Choanoflagellida	-	-	10,267	
-Nucleariidae+Fonticula	-	-	34	
-Fungi/Metazoa incertae sedis	-	-	922	
-Alveolata	-	-	421,968	
-Amoebozoa	-	-	87,348	
-Apusozoa	-	-	263	
-Centroheliozoa	-	-	178	
-Cryptophyta	-	-	3,602	
-Euglenozoa	-	-	122,883	

Continued...

	e-value=0.0	e-value≤1e-10	No. Nuc. seqs	notes
-Fornicata	-	-	11,880	
-Glaucocystophyceae	-	-	195	
-Haptophyceae	-	-	3,442	
-Heterolobosea	-	-	18,044	
-Jakobida	-	-	78	
-Katablepharidophyta	-	-	124	
-Malawimonadidae	-	-	43	
-Oxymonadida	-	-	452	
-Parabasalia	-	-	191,299	
-Rhizaria	-	-	10,372	
-Rhodophyta	-	-	28,440	
-stramenopiles	-	-	147,405	
*Viridiplantae	-	O2	4,452,785	Asian Rice, very short hit
-Bacteria	-	-	4,954,898	
-Archaea	-	-	245,802	

3.2 Full Genome BovB results

Table 5 shows which species were tested for full-length BovBs using the method described in section 1.3. The table shows in which species BovB was identified and gives an indication of how abundant it is in the genome.

Table 5: **Presence of BovB in full genomes studied:** Y means BovB is found in the genome, N means BovB is not found. HA means highly abundant (>10% of the genome is covered by BovB), A means abundant (<10% and >5% of the genome is covered by BovB), P means present (<5% and >1% of the genome is covered by BovB), and R means rare (<1% of the genome is covered by BovB).

Common Name	**Species Name**	**BovB present**	
Cow	*Bos taurus*	Y	HA
Elephant	*Loxodonta africana*	Y	HA
Sheep	*Ovis aries*	Y	HA
Rock Hyrax	*Procavia capensis*	Y	A
Tenrec	*Echinops telfairi*	Y	A
Anole	*Anolis carolinensis*	Y	P
Opossum	*Monodelphis domestica*	Y	P
Platypus	*Ornithorhynchus anatinus*	Y	P
Wallaby	*Macropus eugenii*	Y	P
Horse	*Equus caballus*	Y	R
Sea Urchin	*Strongylocentrotus purpuratus*	Y	R
Silkworm	*Bombyx mori*	Y	R
Zebrafish	*Danio rerio*	Y	R
Common shrew	*Sorex araneus*	N	
Dog	*Canis familiaris*	N	
European Hedgehog	*Erinaceus europaens*	N	
Guinea Pig	*Cavia porcellus*	N	
Honey Bee	*Apis mellifera*	N	
Mosquito	*Aedes aegypti*	N	
Mouse	*Mus musculus*	N	
Nine-banded Armadillo	*Dasypus novemcinctus*	N	
Pig	*Sus scrofa*	N	
Rat	*Rattus norvegicus*	N	
Tree shrew	*Tupaia belangeri*	N	
Wasp	*Nasonia vitripennis*	N	
Zebrafinch	*Taeniopygia guttata*	N	

3.3 Taxa with low coverage genomic survey sequence

Table 6 shows the species that had low coverage genomic survey sequence available that were tested for BovB and the number of BLAST hits returned. Although the birds had significant numbers of BLAST hits, once they were masked using RepeatMasker with the BovB library that was free of CR1 repeats, no bird had more than three hits and none of the RepeatMasker hits were more than 72bp long. From all the marsupials, two of the four ticks and all but two, *Oligosoma lichenigerum* and *Leposoma scincoides*, of the reptiles sufficient sequence was available to reconstruct a BovB sequence long enough for phylogenetic analysis.

Table 6: This table shows the species names and common names of those taxa where genomic survey sequence was available that were tested for BovB and the number of BovB BLAST hits returned when using the improved BovB file containing the four BovBs, from cow, opossum, platypus and viper. The blue names indicate the birds, all of which show good numbers of hits. The red names indicate those species where BovB is abundant and therefore sufficient information was available in the low coverage data to confirm BovBs presence in that species and in most cases construct a nearly full length BovB sequence.

ID	Species	Taxa group	Family	Common name	BovB BLAST hits
AF34	*Nyctophilus gouldi*	Bat	Vespertilionidae	Gould's long-eared bat	12
AF35	*Nyctophilus geoffroyi*	Bat	Vespertilionidae	lesser long-eared bat	15
AF113	*Vanellus miles*	Bird	Charadriidae	Masked Lapwing (bird)	508
AF134	*Acridotheres tristis*	Bird		Indian Mynas	98
AF139	*Podargus strigoides*	Bird		tawny frogmouth	1155
AF18	*Pelecanus conspicillatus*	Bird	Pelecanidae	The Australian Pelican	418
AF23	*Leipoa ocellata*	Bird	Megapodidae	mallee fowl	1378
AF4	*Drymodes brunneopygia*	Bird	Petroicidae	southern scrub-robin	266
AF44	*Phalacrocorax fuscescens*	Bird	Phalacrocoracidae	cormorant (female)	460
AF47	*Gallinula mortierii*	Bird	Rallidae	Tasmanian native hen	434
AF50	*Aquila audax*	Bird	Accipitridae	wedgetail eagle	373
AF56	*Epthianura albifrons*	Bird	Meliiphagidae	white-fronted chat	445
AF80	*Neophema chrysogaster*	Bird	Psittacidae	orange bellied parrot	634
AF82	*Petroica phoenicea*	Bird	Petroicidae	Flame Robin	206
AF83	*Petroica goodenovii*	Bird	Petroicidae	red capped robin	183
AF84	*Petroica boodang*	Bird	Petroicidae	scarlet robin	145
AF85	*Eopsaltria australis*	Bird	Petroicidae	Eastern Yellow Robin	183
AF89	*Artamus personatus*	Bird	Corvidae	woodswallow	418
AF90	*Artamus superciliosus*	Bird	Corvidae	woodswallow	278
AF110	*Amphiprion mccullochi*	Fish	Pomacentridae	whitesnout anemone fish	73
AF111	*Chaetodon tricinctus*	Fish	Chaetodontidae	three-band butterflyfish	35
AF130	*Galaxias fuscus*	Fish		barred galaxid	25
AF147	*Pastinachus atius*	Fish		stingray	133
AF150	*Amphiprion sandaracinos*	Fish			70
AF45	*Cristiceps australis*	Fish	Clinidae	southern crested weedfish	12

Continued...

ID	Species	Taxa group	Family	Common name	BovB BLAST hits
AF61	*Argyrosomus japonicus*	Fish	Sciaenidae	Mulloway	21
AF67	*Henichorynchus siamensis*	Fish	Cyprinidae	siamese mud carp	40
AF68	*Henichorychus lobatus*	Fish	Cyprinidae		41
AF86	*Conorhynchos conirostris*	Fish	?	catfish	16
AF1	*Litoria aurea*	Frog	Hylidae	green & golden bell frog	75
AF102	*Litoria dentata*	Frog	Hylidae	bleating tree frog	34
AF104	*Mixophyes fleayi*	Frog	Myobatrachidae	Fleay's barred frog	53
AF105	*Philoria loveridgei*	Frog	Myobatrachidae	masked mountain frog	61
AF106	*Litoria booroolongensis*	Frog	Hylidae	Booroolong frog	24
AF117	*Litoria nannotis*	Frog	Hylidae	waterfall frogs	30
AF140	*Pseudophryne bibronii*	Frog		Brown toadlet	102
AF21	*Assa darlingtoni*	Frog	Myobatrachidae	pouched frog	29
AF108	*Rattus rattus* lll	Mammal	Muridae	rat lineage III	40
AF109	*Rattus rattus* l	Mammal	Muridae	rat lineage I	27
AF22	*Neophoca cinerea*	Mammal	Otariidae	sea lion	27
AF112	*Macropus antilopinus*	Marsupial	Macropodidae	Antilopine Wallaroo	9983
AF121	*Burramys parvus*(central ESU)	Marsupial	Burramyidae	mountain pygmy possum	1835
AF122	*Burramys parvus*(northern ESU)	Marsupial	Burramyidae	mountain pygmy possum	1668
AF129	*Perameles gunni*	Marsupial		eastern barred bandicoot	1423
AF14	*Antechinus flavipes*	Marsupial	Dasyuridae	Yellow-footed Antechinus(mardo)	1832
AF15	*Isoodon obesulus*	Marsupial	Peramelidae	southern brown bandicoot	1054
AF16	*Macrotis lagotis*	Marsupial	Peramelidae	Greater Bilby	1524
AF20	*Petaurus breviceps*	Marsupial	Petauridae	sugar glider	2982
AF107	*Halotydeus destructor*	Mites & ticks	Penthaleidae	redlegged earth mite	6
AF116	*Amblyomma limbatum*	Mites & ticks	Ixodidae	reptile tick	287
AF46	*Balaustium medicagoense*	Mites & ticks	Erythraeidae	a mite species	10
AF6	*Bothriocroton hydrosauri*	Mites & ticks	Ixodidae	reptile tick	59
AF24	*Amphibolurus norrisi*	Reptile	Agamidae	mallee tree dragon	6254
AF25	*Ctenotus atlas*	Reptile	Scincidae	skink	2019
AF29	*Gehyra variegata*	Reptile	Gekkonidae	gecko	1116

Continued...

ID	Species	Taxa group	Family	Common name	BovB BLAST hits
AF30	*Gehyra lazelli*	Reptile	Gekkonidae	gecko	3729
AF31	*Hydrophis spiralis*	Reptile	Hydrophiidae	sea snake	5989
AF54	*Eremiascincus richardsonii*	Reptile	Scincidae	skink	4602
AF55	*Glaphyromorphus douglasi*	Reptile	Scincidae	skink	1304
AF64	*Christinus guentheri*	Reptile	Gekkonidae	Howe Island gecko	1208
AF65	*Oligosoma lichenigerum*	Reptile	Scincidae	LHI skink	811
AF66	*Leposoma scincoides*	Reptile	Scincidae		616
AF27	*Mustelus antarcticus*	Shark	Triakidae	gummy shark	125
AF70	*Heterodontus potusjacksoni*	Shark	Heterodontidae	Port Jackson shark	30
AF48	*Euperipatoides rowelli*	Arthropod	Peripatopsidae	velvet worm	80

3.4 Annotation of BovB VA

RepeatMasker was used to determine the regions of BovB VA that masked as Chicken Repeat 1 (CR1), shown in Fig. 4. Table 7 shows the coordinates and orientation of the incorporated elements. These sections were removed when searching bird and reptile genomes for BovB to avoid detecting the abundant CR1 elements in sauropsids.

Figure 4: **BovB VA with annotations:** This image represents the RepeatMasker annotation of BovB VA when it is masked with the chicken repeat library and the BovB library containing BovB Opos, BovB and BovB Plat.

position in query				position in repeat					
query sequence	begin	end	(left)	C +	matching repeat	repeat class/family	(left) begin	end end	begin (left)
BovB VA	95	604	(4002)	+	CR1-E	LINE/CR1	3866	4378	(146)
BovB VA	4100	4584	(22)	C	CR1-Y2_Aves	LINE/CR1	(12)	3327	2821

Table 7: Shows the coordinates of the CR1 repeats that are incorporated onto the ends of the BovB VA according to RepeatMasker.

Note that the figure was generated several months before the table and in the interim the RepeatMasker database must have been updated, resulting in slightly different coordinates and annotation of CR1-Y2_Aves instead of the very similar CR1-Y4.

3.5 Chicken Repeats

Vipera ammodytes was the first squamate in which BovB was found and the BovB consensus, available from Repbase, for BovB VA is significantly longer than the other Repbase BovBs. Interestingly the BovB VA sequence has CR1 type elements on both ends of the

full-length BovB element. This means that at some point during its movement it has acquired the portions of the elements now present on both ends of the BovB for all of its future copy and paste movements around the genome. The presence of CR1 fragments at the ends of the BovB VA has made the construction of other squamate BovB consensus sequences more challenging.

The CR1 parts of BovB VA also mean that when searching bird genomes with BLAST or RepeatMasker huge numbers of hits appear. For example the low coverage genomic survey sequence from the mallee fowl had in excess of 1,000 BLAST hits to BovB VA. However, when the CR1 part of BovB VA was removed no hits were found. Indicating that CR1 like repeats are abundant in the mallee fowl, and all birds, as expected, but BovB is not present.

It is possible that other squamates have CR1 fragments on their BovB consensus sequences too. However due to the abundance of CR1 in the squamate genomes and the low coverage reads from which the squamate BovBs were built, all CR1 fragments had to be removed in order to reliably assemble a BovB consensus. Hence further work on full genome sequences or using PCR in a greater range of reptiles would be required to determine when CR1 ends were acquired by the squamate BovB lineage. Interestingly the BovB sequences for the python and the copperhead that were extracted from RepBase do not have the CR1 like ends that are present in BovB VA. This could be due to a different repeat building process used by Castoe *et al.*[30].

3.6 Divergence of BovB consensus sequences with respect to BovB VA

Consensus sequences for BovB were masked using RepeatMasker defaults with BovB VA. Divergence values from the RepeatMasker output were averaged if there was more than one value. For many there was only one section of the repeat masked.

BovB Consensus Sequence	Average Divergence from BovB VA
BovB *Amblyomma limbatum* (reptile tick)	15.3
BovB *Amphibolurus norrisi* (tree dragon)	17.3
BovB *Anolis carolinensis* (green anole)	24.7
BovB *Antechinus flavipes* (mardo)	25.7
BovB *Bombyx mori* (silkworm)	32.4
BovB *Bos taurus* (cow)	16.9
BovB *Bothriocroton hydrosauri* (reptile tick)	21.6
BovB *Burramys parvus* (central ESU pygmy possum)	23.7
BovB *Burramys parvus* (northern ESU pygmy possum)	21.2
BovB *Christinus guentheri* (Howe Island gecko)	29.3
BovB *Ctenotus atlas* (skink)	16.0
BovB *Danio rerio* (zebrafish)	34.2
BovB *Echinops telfairi* (tenrec)	31.5
BovB *Egernia stokesii* (stokes skink)	16.2
BovB *Equus caballus* (horse)	30.5
BovB *Eremiascincus rchardsonii* (skink)	19.1
BovB *Gehyra lazelli* (gecko)	16.6
BovB *Gehyra variegata* (gecko)	21.7
BovB *Glaphyromorphus douglasi* (skink)	16.1
BovB *Hydrophis spiralis* (seasnake)	7.0
BovB *Isoodon obesulus* (southern brown bandicoot)	27.6
BovB *Loxodonta africana* (elephant)	32.7
BovB *Macropus antilopinus* (antilopine wallaroo)	18.9
BovB *Macropus eugenii* (wallaby)	23.9
BovB *Macrotis logotis* (greater bilby)	30.4
BovB *Ovis aries* (sheep)	15.5
BovB *Perameles gunni* (eastern barred bandicoot)	24.0
BovB *Petaurus breviceps* (sugar glider)	20.8
BovB *Procavia capensis* (rock hyrax)	32.9
BovB *Strongylocentrotus purpuratus* (purple sea urchin)	32.9
BovB *Tachyglossu aculeatus* (echidna)	32.6
BovB *Tiliqua rugosa* (sleepy lizard)	17.0

Table 8: Percent Divergence of Consensus Sequences vs BovB VA.

3.7 Validation of BovB from low coverage species and ticks

Sequences amplified by PCR from independent biological samples were sequenced and aligned to the contigs from which sequencing primers were designed Fig. 8. All validation samples were aligned to our contigs and BLASTed against GenBank sequences. In this fashion, we confirmed the occurrence of BovB in our original sequence samples. We also annotated the sequences using RepeatMasker and those annotations are shown below.

3.7.1 Results From Validation Sequences

```
BLASTN 2.2.27+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: 5BKEG71W014


Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           16,502,370 sequences; 42,424,746,788 total letters
Query= Bothriocroton_5'_TG235_

Length=310


                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

gb|AF332697.1|AF332697  Vipera ammodytes Bov-B LINE, complete ...   233    4e-58

ALIGNMENTS
>gb|AF332697.1|AF332697 Vipera ammodytes Bov-B LINE, complete sequence
Length=4606

 Score =  233 bits (258),  Expect = 4e-58
 Identities = 191/233 (82%), Gaps = 3/233 (1%)
 Strand=Plus/Plus

Query  77   CATGRATCAGTGCCTTGTTGCGGCGAAGGTGCTTTAGTAGCTCAATGAAGCTATGAGTTA  136
            |||| || | |||||||| | |||||||| ||||   ||  ||||||||||||||| ||
Sbjct  702  CATGGATTACTGCCTTGTCGTGGCGAAGGGGCTTGCATAATTCAATGAAGCTATGAGCTA  761

Query  137  TGCCATCTAGGATTACCCAAGATGGACAGGTCATAGTAGAGAGTTGTGACTAAACGTGAT  196
            |||| |  |||   |||||||| ||| ||||||||| |||||||| |||| |||||||||
Sbjct  762  TGCCGTGCAGGGCCACCCAAGACGGAAAGGTCATAGCAGAGAGTTCTGACAAAACGTGAT  821

Query  197  CCGCTGGAGAAGGAAATGGCAATCCACTCCAGTAGTCCTGCCAAGAAAACCCGATGAATT  256
            || ||||||||||||||||| ||||||||||    ||||| |||||||| ||| |
Sbjct  822  CCACTGGAGAAGGAAATGGCAACCCACTCCAGTATCTTTGCCATGAAAACCCTATGGA--  879

Query  257  GCAGAACTAAAAGGCTAAACGATATGACACTGGAAKATGAGACCCTCAGGTCG  309
             ||| || ||||||| | |||||||||| |||||| ||||| ||||||||||
Sbjct  880  -CAGTACCAAAAGGCAATACGATATGACGCTGGAAGATGAGCCCCTCAGGTCG  931
```

Figure 5: **BLASTN Result for *Bothriocroton* 5' Sequence**

```
BLASTN 2.2.27+
Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and
Webb Miller (2000), "A greedy algorithm for aligning DNA
sequences", J Comput Biol 2000; 7(1-2):203-14.


RID: 5BKKMUCR016


Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           16,502,370 sequences; 42,424,746,788 total letters
Query= Amblyomma_5'_ABTC111481_

Length=168


                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

gb|AF332697.1|AF332697  Vipera ammodytes Bov-B LINE, complete ...   161    1e-36

ALIGNMENTS
>gb|AF332697.1|AF332697 Vipera ammodytes Bov-B LINE, complete sequence
Length=4606

 Score =  161 bits (87),  Expect = 1e-36
 Identities = 123/141 (87%), Gaps = 0/141 (0%)
 Strand=Plus/Plus

Query  28   CATAGACTACTGCCTTGTCGTGGCGAAGGGGCTTGCGTAGCTCACGGAAGCTATGAGTTA  87
            ||| || ||||||||||||||||||||||||||| ||  |||  |||||||||||| ||
Sbjct  702  CATGGATTACTGCCTTGTCGTGGCGAAGGGGCTTGCATAATTCAATGAAGCTATGAGCTA  761

Query  88   TGCCGTGCAGGGTCACCTAAGACGAGCAGCTCATAGCAGATAGCTCTAACAAAACGTGAT  147
            |||||||||||| |||| ||||||   || |||||||||| || ||| ||||||||||||
Sbjct  762  TGCCGTGCAGGGCCACCCAAGACGGAAAGGTCATAGCAGAGAGTTCTGACAAAACGTGAT  821

Query  148  TCACTGGAGAAGGAAATGGCA  168
             |||||||||||||||||||
Sbjct  822  CCACTGGAGAAGGAAATGGCA  842
```

Figure 6: **BLASTN Result for *Amblyomma* 5' Sequence**

```
BLASTN 2.2.27+
Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and
Webb Miller (2000), "A greedy algorithm for aligning DNA
sequences", J Comput Biol 2000; 7(1-2):203-14.


RID: 67SYU7EE016


Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           16,545,181 sequences; 42,537,579,184 total letters
Query= G.variegata ABTC82613 5'

Length=276


                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

gb|AF332666.1|AF332666  Boa constrictor clone BC Bov-B LINE, c...   302    1e-78

ALIGNMENTS
>gb|AF332666.1|AF332666 Boa constrictor clone BC Bov-B LINE, complete sequence
Length=1767

 Score =  302 bits (163),  Expect = 1e-78
 Identities = 220/250 (88%), Gaps = 0/250 (0%)
 Strand=Plus/Plus

Query  1     CCAAAGAAGGGCAACACCAAGGAATGYTCCAACTATCGCACAATTGCACTCATYTCACAC  60
             ||||||||||||||  |||| ||||| || ||||| || ||||||||||||||| |||||
Sbjct  1222  CCAAAGAAGGGCAATGCCAAAGAATGTTCTAACTACCGTACAATTGCACTCATTTCACAT  1281

Query  61    GCTAGCAAGGTCATGCTCAAGATCCTACAAGCTAGGCTTCAGCAGTATGTGGACAGAGAA  120
             |||||||||||| |||||||| |||||||||||||||||||||||||||||| ||  ||||
Sbjct  1282  GCTAGCAAGGTGATGCTCAAAATCCTACAAGCTAGGCTTCAGCAGTATGTGAACCAAGAA  1341

Query  121   TTGCCAGAAGTACAAGCTGGGTTTCGAAGAGGCAGAGGAACKAGAGACCAAATTGCCAAC  180
              | |||||||| ||||||||||||||||||||||||||||||  |||| || |||||||||
Sbjct  1342  CTACCAGAAGTGCAAGCTGGGTTTCGAAGAGGCAGAGGAACTCGAGATCAGATTGCCAAC  1401

Query  181   ATTCGCTGGATTATGGAGAAAGCAAGGGAGTACCAGAAAAACATCTACTTCTGGTTCTCT  240
              |||||||||| ||||||||||||| |||| ||||||||||||||||||||| |||  |
Sbjct  1402  CTTCGCTGGATCATGGAGAAAGCAAGAGAGTTCCAGAAAAACATCTACTTCTGCTTCATT  1461

Query  241   GCCTATGCTA  250
             | ||| ||||
Sbjct  1462  GACTACGCTA  1471
```

Figure 7: **BLASTN Result for *Gehyra* 5' Sequence**

SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin	end	(left)		matching repeat	repeat class/family	position in repeat begin	end	(left)	ID
432	11.5	0.0	0.0	Amblyomma_3'_ABTC111481_	2	62	(156)	+	BovB_ACo	LINE/RTE-BovB	3152	3212	(70)	1
1016	12.3	0.7	0.0	Amblyomma_5'_ABTC111481_	23	168	(0)	+	BovB_ACo	LINE/RTE-BovB	2	148	(3134)	2
1153	19.6	0.5	0.0	Bothriocroton_3'_TG235_	3	216	(0)	+	BovB_Mars	LINE/RTE-BovB	468	682	(2537)	3
1423	15.6	0.4	1.3	Bothriocroton_5'_TG235_	76	309	(1)	+	BovB_ACo	LINE/RTE-BovB	6	237	(3045)	4
421	21.5	2.1	1.1	Christinus__3'_ABTC6983_	1	94	(5)	+	RTE-1_EC	LINE/RTE-BovB	3108	3202	(6)	5
1391	10.1	1.0	0.0	Equus_3'_Chief	1	198	(153)	+	RTE-1_EC	LINE/RTE-BovB	2999	3198	(10)	6
769	15.7	0.0	0.7	Equus_3'_Chief	204	351	(0)	C	MER63B	DNA/hAT-Blackjack	(245)	191	45	7
622	16.7	0.9	0.9	Equus_5'_Chief	1	109	(205)	C	ERE1C	SINE/tRNA	(162)	109	1	8
1324	8.3	6.3	0.5	Equus_5'_Chief	110	314	(0)	+	RTE-1_EC	LINE/RTE-BovB	135	351	(2857)	9
1820	11.7	0.0	0.0	G.variegata_ABTC82613	1	256	(20)	+	BovB_PMo	LINE/RTE-BovB	1903	2158	(1130)	10

Figure 8: **RepeatMasker annotation of validation sequences:**This figure shows the RepeatMasker .out file for the validation sequences. Sequences of amplicons from *Equus caballus, Amblyomma limbatum, Bothriocroton hydrosauri, Christinus guentheri, Gehyra variegata.*

3.8 Phylogenetic tree of BovB and orthologues

3.8.1 Tree built from orthologous sequences

Fig. 9 shows a tree developed using the orthologous sequences present in OrthoDB and generously provided by Dr Evgeny Zdobnov. This shows the expected phylogenetic relationships between the species and acts as a control from which to determine what HTs have occurred.

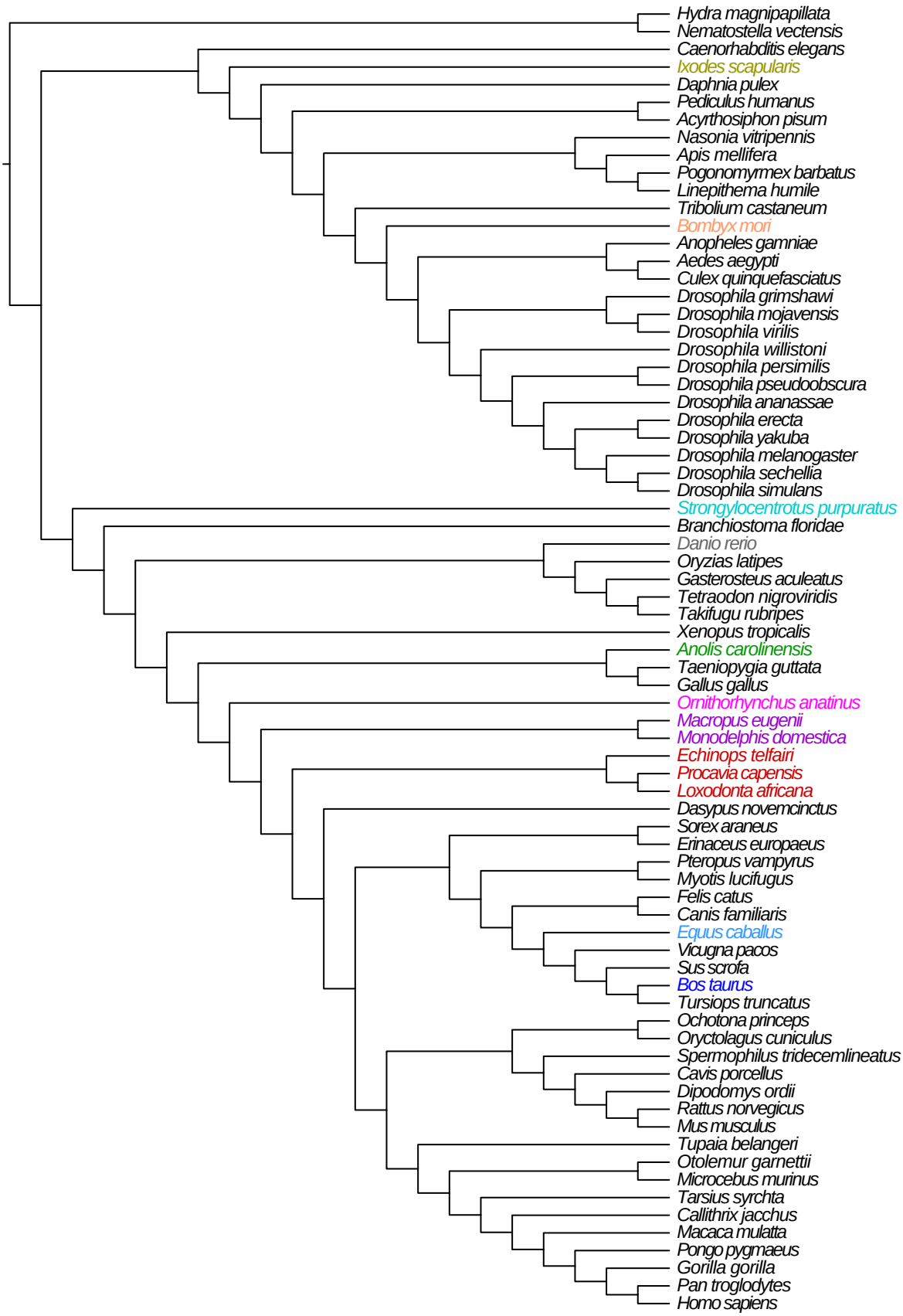


Figure 9: **Tree built from orthologues:** Tree provided by Dr Evgeny Zdobnov for comparison to the phylogenetic trees built from the BovB sequences. Colours indicate the taxonomic groups that have BovB.

3.8.2 Trees built from BovB sequences

RAxML tree, in Fig. 10 in section 3.8.3, shows a maximum likelihood tree built using 500 bootstraps to determine the bootstrap support for the nodes in the tree. The differences between the FastTree output, shown in the paper, and RAxML show that some of the nodes of the tree are not supported when different tree building parameters are used. For example the marsupial clade in the FastTree output is a sister group to the clade that contains reptiles, ticks and ruminants, however in the RAxML tree the marsupial clade is a sister group to the ruminant clade and together they group with the reptiles. The bootstrap support for the monophyly of marsupials is strong but the bootstrap support values within the marsupial clade are very low, as seen for the local support values in the FastTree output.

BEAST tree, in Fig. 11 in section 3.8.4, shows that the basic topology is robust, regardless of which tree building method is used. This allows conclusions about the origins of BovB elements to be inferred. There are however several differences between the BEAST tree and FastTree output. The position of the zebrafish BovB in the FastTree output and BEAST tree is not robust. In the FastTree output the zebrafish BovB has strong support for being basal to the Afrotherian/monotreme/horse clade, whereas in the BEAST tree it has strong support for being basal to the marsupial/reptile/ruminant group. The main snake clade is basal to the ruminants in the BEAST tree unlike in the FastTree output. The tree dragon BovB is also not robust across the two trees. In the FastTree output it is basal to the reptile/marsupial group but with BEAST it is sister to the skinks. Again the marsupial clade has strong support for monophyly but weak support for the resolution within the clade.

All three tree building methods group the ruminants and reptiles together, and the placement of the ticks is well supported in all trees. The marsupials and the reptiles form a clade that is robust to the tree building method, despite the weak support for some internal branches and nodes. The Afrotherian/monotreme/horse clade is well supported by all methods and shows concordance across maximum likelihood and Bayesian MCMC tree building methods.

3.8.3 RAxML

The parameters used to produce the RAxML tree in Fig. 10 are shown below.

RAxMLHPC -fa -N 500 -s tree_withRepBase_mult_aligned_gblocks.phylip
-n tree_withRepBase_faxgtrgamma -m GTRGAMMA -x 51011 -p 51011

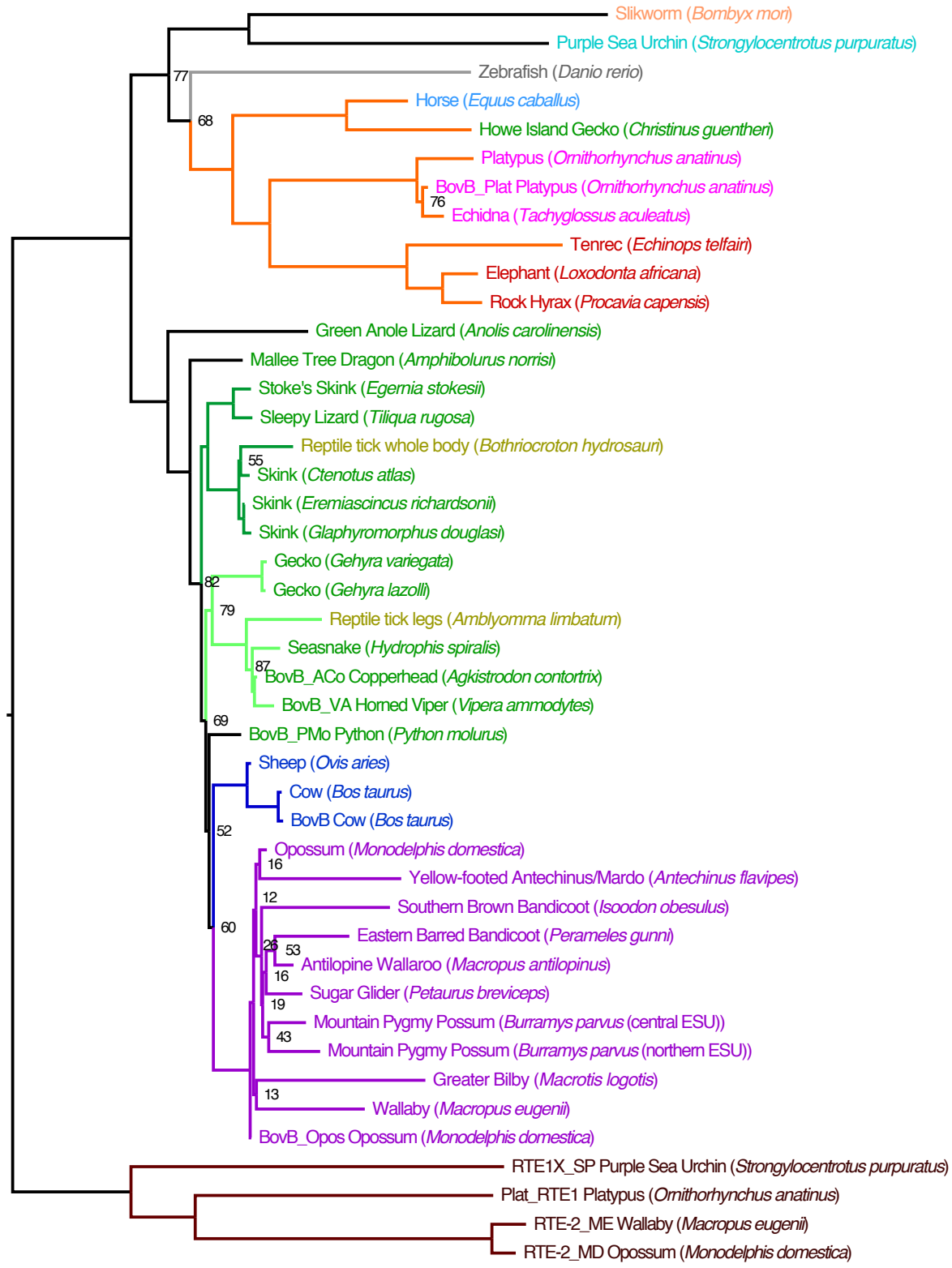


Figure 10: **RAxML tree:** RAxML maximum likelihood tree with only those bootstrap values below 90% shown (500 replicates). Tree built from the full-length BovB sequences extracted from full genome sequence and those constructed from low coverage reads. The sequences were aligned with MUSCLE and processed by Gblocks to limit the effect of indels, making an alignment that was 2858bp long. Branch colours indicate important BovB clades, marsupials in purple, skinks/tick in green, gecko/snake/tick in light green, ruminants in blue and monotremes/Afrotheria/horse in orange, and the RTE clade, in maroon, used to root the tree. Taxa showing BovB are coloured taxonomically, with marsupials in purple, reptiles in green, ruminants in dark blue, arthropods in yellow, Afrotheria in red, monotremes in pink, horse in blue, zebrafish in grey, sea urchin in light blue and silkworm in orange. The RTEs are in maroon.

3.8.4 BEAST

Fig. 11 is a tree built using the BEAST software [19] after the correct model was chosen using ModelGenerator [21]. We used the GTR with gamma model, because the Bayesian Information Criteria (BIC) ranked it best and it ranked second best for the AIC (Akaike information criterion) 1 and 2. Yule process was used for tree priors because each BovB comes from a different species and therefore each branch is a speciation event. This assumption breaks down for three of the branches, the BovB Plat vs Platypus, BovB vs Cow and BovB Opos vs Opossum but this was recognised and a test of the tree structure with yule priors and the duplicated species removed showed an almost identical topology as the tree with the duplicates included. The only difference was the position of the central pygmy possum sequence but given the low posterior support value for its placement in both the tree with and without duplicates the fact that the position of this sequence is not robust if sequences are removed is not surprising and does not provide sufficient evidence to invalidate the original tree.

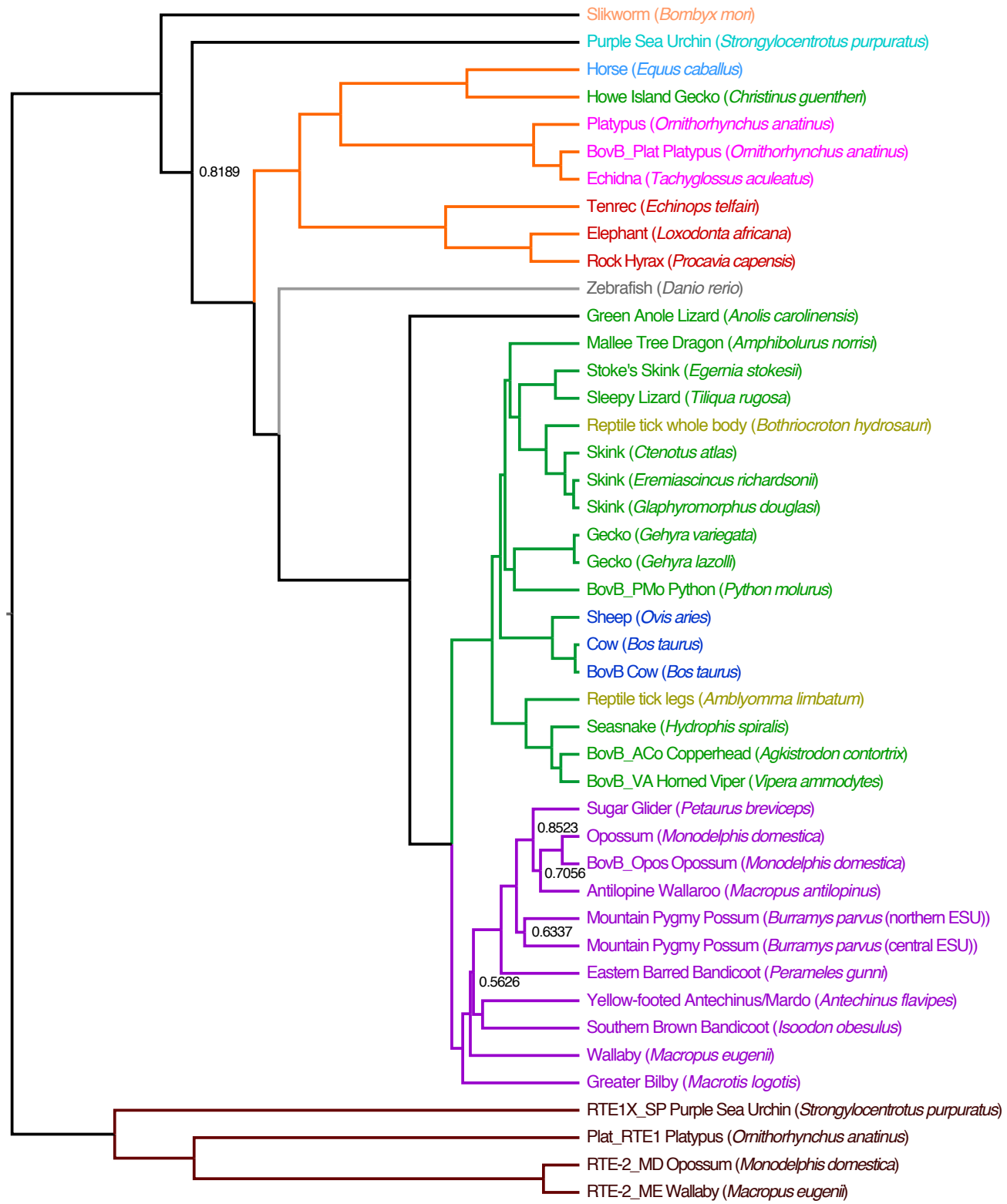


Figure 11: **BEAST tree:** Tree built by BEAST and TreeAnotator with only those posterior probabilities that are below 0.9 shown. MCMC chain length of 100,000,000 sampling every 10,000; burnin = 1000 trees. Tree built from the full-length BovB sequences extracted from full genome sequence and those constructed from low coverage reads. The sequences were aligned with MUSCLE and processed by Gblocks to limit the effect of indels, making an alignment that was 2858bp long. Branch colours indicate important BovB clades, marsupials in purple, reptiles/ruminants/ticks in green and monotremes/Afrotheria/horse in orange, and the RTE clade, in maroon, used to root the tree. Taxa showing BovB are coloured taxonomically, with marsupials in purple, reptiles in green, ruminants in dark blue, arthropods in yellow, Afrotheria in red, monotremes in pink, horse in blue, zebrafish in grey, sea urchin in light blue and silkworm in orange. The RTEs are in maroon.

References

[1] Altschul, S., Gish, W., Miller, W., Myers, E. & D.J., L. Basic local alignment search tool. *J Mol Biol* **215**, 403–10 (1990).

[2] Harris, R. *Improved pairwise alignment of genomic DNA*. Ph.D. thesis, The Pennsylvania State University (2007).

[3] NCBI. Blast home (2011). URL `http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome`.

[4] Rozen, S. & Skaletsky, HJ Primer3 (1998). URL `Http://Www-Genome.Wi.Mit.Edu/Genome_Software/Other/Primer3.Html`.

[5] Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004).

[6] Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**, 540–552 (2000).

[7] Smit, A., Hubley, R. & Green, P. Repeatmasker (1996-2011). URL `http://repeatmasker.org`.

[8] Edgar, R. C. Search and clustering orders of magnitude faster than blast. *Bioinformatics* (2010).

[9] Edgar, R. C. & Myers, E. W. Piler: identification and classification of genomic repeats. *Bioinformatics* **21**, i152–i158 (2005).

[10] Los Alamos National Laboratory, . Advanced consensus maker (2011). URL `http://www.hiv.lanl.gov/content/sequence/CONSENSUS/AdvCon.html`.

[11] Stajich, J. E. *et al.* The bioperl toolkit: Perl modules for the life sciences. *Genome Research* **12**, 1611–1618 (2002).

[12] Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequencefinishing. *Genome Research* **8**, 195–202 (1998).

[13] Tamura, K. *et al.* Mega5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* (2011).

[14] Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic dna. *Journal of Molecular Biology* **268**, 78–94 (1997).

[15] Burge, C. The genscan web server at mit (2011). URL `http://genes.mit.edu/GENSCAN.html`.

[16] Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26**, 1641–1650 (2009).

[17] Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

[18] Stamatakis, A. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).

[19] Drummond, A. & Rambaut, A. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214 (2007).

[20] Rambaut, A. & Drummond, A. Tracer v1.5 (2009). URL `http://beast.bio.ed.ac.uk/Tracer`.

[21] Keane, T., Creevey, C., Pentony, M., Naughton, T. & McLnerney, J. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology* **6**, 29 (2006).

[22] Hill, T. *et al.* Sprit: Identifying horizontal gene transfer in rooted phylogenetic trees. *BMC Evol Biol* **10**, 42 (2010).

[23] Linz, S. On hill et al's conjecture for calculating the subtree prune and regraft distance between phylogenies. *BMC Evolutionary Biology* **10**, 334 (2010).

[24] NCBI. Taxonomy browser (2011). URL `http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi`.

[25] Waterhouse, R. M., Zdobnov, E. M., Tegenfeldt, F., Li, J. & Kriventseva, E. V. Orthodb: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Research* **39**, D283–D288 (2011).

[26] Kumar, S. & Hedges, S. B. Timetree2: Species divergence times on the iphone. *Bioinformatics* (2011).

[27] Hedges, S. B., Dudley, J. & Kumar, S. Timetree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).

[28] The UniProt Consortium, . Ongoing and future developments at the universal protein resource. *Nucleic Acids Research* **39**, D214–D219 (2011).

[29] Jain, E. *et al.* Infrastructure for the life sciences: design and implementation of the uniprot website. *BMC Bioinformatics* **10**, 136 (2009).

[30] Castoe, T. A. *et al.* Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. *Genome Biology and Evolution* **3**, 641–653 (2011).