SUPPLEMENTAL INFORMATION FOR

Discovery of a new TLR gene and gene expansion event through improved desert tortoise genome assembly with chromosome-scale scaffolds

Greer A. Dolby, Matheo Morales, Timothy H. Webster, Dale F. DeNardo, Melissa A. Wilson, Kenro Kusumi

Chicago library preparation and sequencing

Three Chicago libraries were prepared as described previously (Putnam et al, 2016). Briefly, for each library, ~500ng of HMW gDNA (mean fragment length =100 kb) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with *DpnII*, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq X platform. The number and length of read pairs produced for each library was: 100 million, 2x151 bp for library 1; 127 million, 2x151 bp for library 2, and 96 million, 2x151 bp for library 3. Together, these Chicago library reads provided 135.32X physical coverage of the genome (1–100 kb pairs).

Scaffolding the assembly with HiRise

The input *de novo* assembly, shotgun reads, and Chicago library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (Putnam et al, 2016). Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, shotgun sequences were used to close gaps between contigs.

Genome assembly results

The final assembly length for gopAga2.0 assembly is 2.34 Gb—slightly shorter than for gopAga1.0 (2,338,664,599 vs. 2,399,952,228 bp, respectively). The gopAga2.0 assembly dramatically improved scaffold contiguity. Its longest scaffold increased from 2 Mb to 106.5 Mb and the L50 decreased from 2,592 to 26 scaffolds, which means that roughly half the length of the genome is now in the first 26 scaffolds.

Genome annotation methods & results

To generate a *de novo* annotation for gopAga2.0 we made individual genome-guided transcriptome assemblies using original deep transcriptome data from male brain, lung, and skeletal muscle (accessions: SRX2342843–5). Raw paired-end reads were trimmed for quality using BBDuk v37.28 (ktrim=r, k=24, mink=11, hdist=1, tpe, tbo, qtrim=r, trimq=8) and trimmed reads were merged using BBMerge v37.28 (Bushnell 2014). On average, about 60% of reads were merged; both merged and unmerged reads were retained.

We mapped the trimmed merged and trimmed unmerged reads separately to the gopAga2.0 assembly using STAR v2.5.3a (Dobin et al. 2012) with the 2-pass approach where junctions identified during the first round of mapping were passed as input to the second phase using –sjdbFileChrStartEnd. The coordinate-sorted BAM files from mapping were passed into Trinity v2.5.1 to make tissue-specific genome-guided assemblies using a max intron limit of 100,000 (Grabherr et al. 2011). We modeled tortoise-specific repeat and low complexity regions with RepeatModeler v1.0.11 (Smit & Hubley 2008).

To annotate the gopAga2.0 assembly using MAKER v3 (Campbell et al. 2015) we performed one round of mapping evidence followed by three rounds of *ab initio* gene model training. For evidence, we provided the three genome-guided transcriptomes, predicted proteins from western painted turtle (NCBI BioProject PRJNA210179), and protein evidence from UniProtKB/Swiss-Prot database. We recorded the number of genes, average gene length, the Annotation Edit Distance (AED), and BUSCO results from MAKER-predicted transcripts for each round. To determine completeness, we ran BUSCO v3.0.2 (Waterhouse et al. 2017) on both the genome assembly (-m genome flag) as well as the MAKER-predicted transcripts (-m transcriptome flag) using the tetrapod gene set available via BUSCO at time of study (N = 3950 genes; Tables 1, S2).

The gopAga2.0 annotation has 25,469 genes, of which all but six have an AED < 1.0 and/or a PFAM domain (based on the quality_filter.pl script from Maker). This gene count is higher than the draft gopAga1.0 genome (20,172), which could result from higher confidence gene models as genes may have been previously split across scaffolds or because of the two-pass mapping and genome-guided transcriptomes (gopAga1.0 used *de novo* transcriptomes).

Evolution of *TLR8* in Testudines

Pseudogenization of the Testudines-specific TLR8C

Previous work (Liu et al 2019; Kahn et al 2019) showed a duplication of *TLR8* in turtles (forming TLR8B), as well as a second duplication in the Chinese softshell turtle (*Pelodiscus* sinensis) lineage, which has three copies of *TLR8*. Because this region appeared to be quite active, we did a fourth round of manual gene curation in the *TLR8* genomic region of all 22 species. In this fourth approach, we ignored all gene models and instead pulled the genomic sequence from the 3' end of TLR8-1 to the 5' end of TLR8-3 in all 22 species. For each species, we ran this region through the SMART motif finder to independently assess the number of supported *TLR8* genes. Doing this confirmed that:

(1) there are pseudogenized copies of *TLR8* (TLR8C) present in *Gopherus agassizii* and *Chelonia mydas* (the two non-freshwater testudines) based on the retention of the TIR domain, retention of some LRRs, and the presence of stop codons throughout.

(2) there is an intact TLR8B ortholog present in *Chrysemys picta*, which like *P. sinensis*, has 3 TLR8 paralogs. The TLR8B and TLR8C paralogs in the *C. picta* annotation are considered a single gene model—we propose instead they are separate genes.

(3) there is no evidence of a TLR8B gene in any of the other lineages examined (i.e. the TLR8B duplication event appears specific to Testudines).

Truncation of TLR8-1 *(TLR8B) in Gopherus agassizii*

A truncated TLR paralog has been observed in several fish species where there is both membrane-bound *TLR5M* gene and soluble *TLR5S* gene, which lacks a TIR and transmembrane domain but retains 21 leucine rich repeats (reviewed in Rebl et al. 2010). TLR5 proteins bind to flagellar antigens from pathogens, and the soluble TLR5S protein is thought to amplify TLR5M signaling through a positive feedback loop. Furthermore, TLR5M is ubiquitously expressed in fish while TLR5S exhibits tissue-specific expression, similar to the tissue-specific pattern found with *TLR8* in the Chinese softshell turtle, *P. sinensis* (Liu et al. 2019).

Within the *Gopherus agassizii* gopAga2.0 genome annotation, we identified a truncated gene model for *TLR8-1* (TLR8B) in the sequenced specimen due to the presence of a stop codon in the middle of the coding sequence. We performed additional analyses to determine whether **(1)** this stop codon is biologically replicated or the result of a technical artifact, and **(2)** whether a truncation effect similar to TLR5 in fishes may be occurring within TLR8-1 of the desert tortoise. To do so we mapped reads to the genome assembly from: **(A)** the deep transcriptome sequencing of skeletal muscle, lung, and brain originally used for the assembly and annotation of this individual; **(B)** reads from the blood transcriptomes sequenced from three additional unrelated *G. agassizii* individuals as well as three unrelated individuals each from sister species *G. morafkai* and *G. evgoodei* (Edwards et al 2016; SRX1004698, SRX1004679, SRX1004665, SRX1004662, SRX1004661, SRX1004618, SRX1004258, SRX1004169, SRX1002875), **(C)** low coverage whole-genome resequencing data from two *G. agassizii* individuals from western Arizona (raw data is unpublished but aligned .bam files for this region are available via archived data for this paper on Harvard Dataverse).

(1) The CGA → TGA stop codon was confirmed in two independent datasets: the lung deep transcriptome from the gopAga2.0 type specimen, as well as a blood transcriptome from an unrelated conspecific. Mapping of the unpublished low coverage whole genome data from two unrelated *G. agassizii* individuals shows two other results. The first individual has the CGA codon present at this position where it occurs in other Testudines. The second individual shows to be heterozygous for the CGA/TGA arginine/stop codon polymorphism (Figure S9). These two individuals are from the same population in northwestern Arizona. Most of the transcriptomic sequences examined did not have read coverage over the position of this polymorphism.

(2) To assess whether transcripts in this type specimen were full-length or truncated as suggested by the gene model in the gopAga2.0 annotation, we assess read depth in this region from data sources A–C above. Within the gopAga2.0 type specimen, one tissue (lung) showed reads from mRNA produced from the full transcript, whereas brain and skeletal muscle only showed reads from mRNA transcribed to the 3' end of the stop codon. Within unrelated conspecifics for *G. agassizii* (individuals 1–3, Figure S9), sequenced reads from two individuals cover the length of the transcript whereas in another individual only covers the 3' end of the transcript, with lack of coverage around the stop codon polymorphism. In all six heterospecific individuals (*G. morafkai* and *G. evgoodei*) there are reads from the full length of the transcript.

In summary, we identified a stop-codon polymorphism within TLR8B orthologue that is exclusive to *G. agassizii*. We identified transcribed sequences 5' to the stop codon in full length transcripts. We observed transcripts that constituted only sequences 3' to the stop codon, which might be incompletely represented transcript sequences and reflect bias in RNA conversion to cDNA.

References:

Bushnell B. 2014. BBMap: A fast, accurate, splice-aware aligner. 2. doi: 10.1186/1471-2105-13-238.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Chapman JA, et al. 2011. Meraculous: de novo genome assembly with short paired-end reads. PLoS One. 6:e23501.

Dobin A et al. 2012. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15–21. doi: 10.1093/bioinformatics/bts635.

Edwards, T., Tollis, M., Hsieh, P., Gutenkunst, R., Liu, Z., Kusumi, K., Culver, M., Murphy, R. 2016. Assessing models of speciation under different biogeographic scenarios; an empirical study using multi-locus and RNA-seq analyses. Ecology and Evolution 6:379-396.

Smit AF, Hubley R. 2008. RepeatModeler Open-1.0. Available from http://www. repeatmasker. org.

Lieberman-Aiden E. et al. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 326:289–293.

Putnam NH et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Research. 26:342–350. doi: 10.1101/gr.193474.115.

Table S1 Species studied in TLR gene family analyses, the genome build versions used and their original sources. Study references are available in Appendix 1 or the main text.

Species	Common Name	Data source	Reference
Xenopus tropicalis	western clawed frog	Xenopus_tropicalis_v9.1	(Hellsten et al. 2010)
Homo sapiens	human	GRCh38.p12	(Collins et al. 2004)
Mus musculus	house mouse	GRCm38.p6	(Church et al. 2009)
Rattus norvegicus	brown Norway rat	Rnor_6.0	(Metzker et al. 2004)
Ornithorhynchus anatinus	platypus	Ornithorhynchus_anatinus-5.01	(Warren et al. 2008)
Anolis carolinensis	green anole	AnoCar2.0	(Alföldi et al. 2011)
Gekko japonicus	Schlegel’s Japanese gecko	Gekko_japonicus_V1.1	(Liu et al. 2015)
Pogona vitticeps	central bearded dragon	Pvi1.1	(Georges et al. 2015)
Python bivittatus	Burmese python	Python_molurus_bivittatus-5.0.2	(Castoe et al. 2013)
Thamnophis sirtalis	garter snake	Thamnophis_sirtalis-6.0	(Perry et al. 2018)
Chelonia mydas	green sea turtle	CheMyd_1.0	(Wang 2013)
Pelodiscus sinensis	Chinese softshell turtle	PelSin_1.0	(Wang 2013)
Chrysemys picta	western painted turtle	Chrysemys_picta_bellii-3.0.3	(Shaffer et al. 2013)
Gopherus agassizii	Mojave Desert tortoise	*this study*	this study
Chelonoidis abingdonii	giant Galápagos tortoise	ASM359739v1	(Quesada et al. 2018)
Alligator mississippiensis	American alligator	ASM28112v4	(St John et al. 2012)
Alligator sinensis	Chinese alligator	ASM45574v1	(Wan et al. 2013)
Crocodylus porosus	saltwater crocodile	CroPor_comp1	(St John et al. 2012)
Gavialis gangeticus	gharial	GavGan_comp1	(St John et al. 2012)
Gallus gallus domesticus	chicken	GRCg6a	(Hillier et al. 2004)
Haliaeetus leucocephalus	bald eagle	Haliaeetus_leucocephalus-4.0	(Zhang et al. 2014)
Columba livia domestica	domestic pigeon	Cliv_1.0	(Shapiro et al. 2013)

Table S2 Benchmarking Universal Single Copy Orthologs (BUSCO) results for the predicted transcripts of genome annotations for gopAga1.0 and gopAga2.0 based on the Tetrapoda conserved gene set (3950 BUSCO genes). Duplicated [d] BUSCOs are shown in parentheses. Analyses were run using the -m transcriptome flag.

	Complete [d]	**Fragmented**	**Missing**
gopAga1.0	73.1% [16.7%]	18.4%	8.5%
gopAga2.0	75.7% [1.3%]	16.7%	7.6%

Table S3 Scaffolds used in Figure 1, note that these are the 26 most gene-rich scaffolds, not necessarily the longest 26 scaffolds.

Figure 1 ID	Number of genes	Scaffold length (bp)	Scaffold ID in gopAga2.0
A	715	106,572,802	scaffold_0
B	861	90,694,790	scaffold_1
C	508	71,105,540	scaffold_2
D	455	69,455,760	scaffold_3
E	429	67,472,536	scaffold_4
F	376	55,848,362	scaffold_5
G	355	49,996,614	scaffold_6
H	446	41,678,395	scaffold_9
I	415	44,170,057	scaffold_7
J	503	40,959,907	scaffold_10
K	420	38,080,404	scaffold_12
L	528	36,123,519	scaffold_13
M	307	34,933,802	scaffold_16
N	580	34,090,330	scaffold_17
O	450	33,505,802	scaffold_19
P	348	33,617,390	scaffold_18
Q	364	33,293,575	scaffold_20
R	327	26,497,676	scaffold_29
S	395	26,164,675	scaffold_30
T	454	25,492,271	scaffold_31
U	718	23,685,959	scaffold_34
V	406	23,469,553	scaffold_35
W	401	20,976,089	scaffold_41
X	310	19,751,100	scaffold_44
Y	337	13,146,356	scaffold_50
Z	312	13,031,155	scaffold_51

Table S4 List of the reference sequences used in the BLAST analysis to find homologous sequences in the other tetrapods studied. Human was the first choice for a query sequence, when it was absent in human, we used mouse, when absent in mouse we used chicken.

TLR Homolog	Query sequence	taxon	descriptor
TLR7	AAZ99026.1	*Homo sapiens*	TLR7
TLR8	AAZ95441.1	*Homo sapiens*	TLR8
TLR9	NP_059138.1	*Homo sapiens*	toll-like receptor 9 precursor
TLR11	NP_991388.2	*Mus musculus*	toll-like receptor 11
TLR12	EDL30230.1	*Mus musculus*	toll-like receptor 12
TLR13	EDL14060.1	*Mus musculus*	toll-like receptor 13
TLR21	NP_001025729.1	*Gallus gallus*	Toll-like receptor 21

Table S5 Bayesian Information Criterion (BIC) scores for different protein sequence evolution models as evaluated in ProtTest for TLR7 subfamily and TLR11 subfamily. The chosen models (highlighted in yellow) were the best-scoring models among those that can be implemented in MrBayes.

TLR7 subfamily results				**TLR11 subfamily results**		
model	***ΔBIC***	***BIC***		***model***	***ΔBIC***	***BIC***
JTT	0.00	96051.93		JTT	0.00	94817.92
WAG	576.12	96628.05		VT	561.50	95379.42
VT	652.06	96703.99		WAG	761.12	95579.04
CpREV	1100.78	97152.71		HIVb	1402.65	96220.57
HIVb	1170.21	97222.14		CpREV	1466.87	96284.79
FLU	1306.12	97358.05		LG	1506.18	96324.10
LG	1385.97	97437.90		FLU	1645.03	96462.95
Blosum62	2040.54	98092.47		Blosum62	1873.76	96691.68
DCMut	2219.89	98271.82		DCMut	2005.43	96823.35
Dayhoff	2225.21	98277.14		Dayhoff	2010.60	96828.52
RtREV	2624.89	98676.82		RtREV	2395.42	97213.34
HIVw	2893.86	98945.79		HIVw	3483.12	98301.04
MtREV	6462.45	102514.38		MtREV	6503.46	101321.38
MtMam	9814.05	105865.98		MtMam	9425.66	104243.58
MtArt	12126.36	108178.29		MtArt	11789.24	106607.16

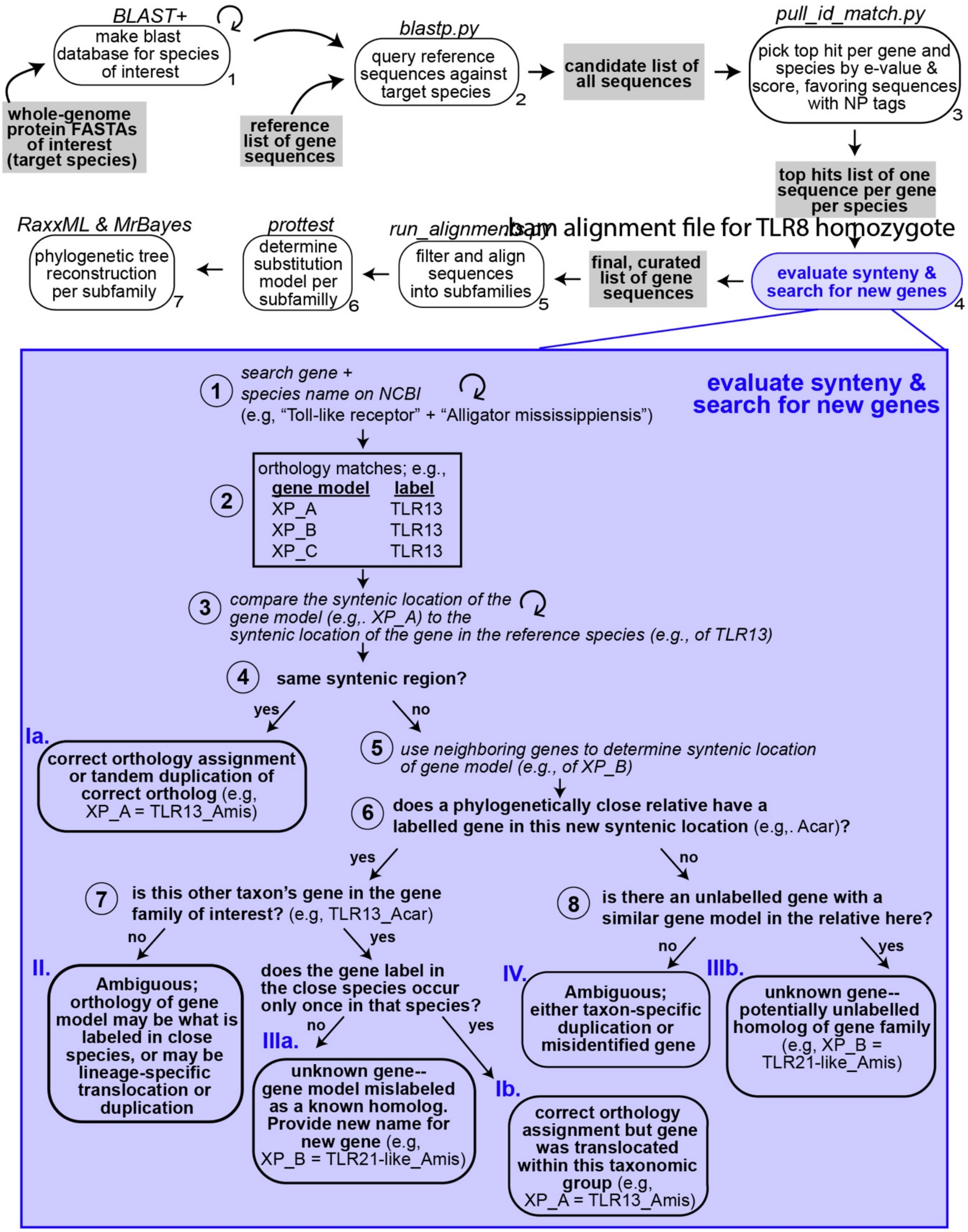
BLAST+
make blast database for species of interest
1
whole-genome protein FASTAs of interest (target species)
blastp.py
query reference sequences against target species
2
reference list of gene sequences
candidate list of all sequences
pull_id_match.py
pick top hit per gene and species by e-value & score, favoring sequences with NP tags
3
top hits list of one sequence per gene per species
evaluate synteny & search for new genes
4
final, curated list of gene sequences
run_alignments.py
filter and align sequences into subfamilies
5
prottest
determine substitution model per subfamilly
6
RaxxML & MrBayes
phylogenetic tree reconstruction per subfamily
7
bam alignment file for TLR8 homozygote
evaluate synteny & search for new genes
1 search gene + species name on NCBI (e.g, "Toll-like receptor" + "Alligator mississippiensis")
2 orthology matches; e.g.,
gene model label
XP_A TLR13
XP_B TLR13
XP_C TLR13
3 compare the syntenic location of the gene model (e.g,. XP_A) to the syntenic location of the gene in the reference species (e.g., of TLR13)
4 same syntenic region?
yes
no
Ia.
correct orthology assignment or tandem duplication of correct ortholog (e.g, XP_A = TLR13_Amis)
5 use neighboring genes to determine syntenic location of gene model (e.g., of XP_B)
6 does a phylogenetically close relative have a labelled gene in this new syntenic location (e.g,. Acar)?
yes
no
7 is this other taxon's gene in the gene family of interest? (e.g, TLR13_Acar)
8 is there an unlabelled gene with a similar gene model in the relative here?
no
yes
II.
Ambiguous; orthology of gene model may be what is labeled in close species, or may be lineage-specific translocation or duplication
does the gene label in the close species occur only once in that species?
no
yes
IIIa.
unknown gene-- gene model mislabeled as a known homolog. Provide new name for new gene (e.g, XP_B = TLR21-like_Amis)
Ib.
correct orthology assignment but gene was translocated within this taxonomic group (e.g, XP_A = TLR13_Amis)
no
yes
IV.
Ambiguous; either taxon-specific duplication or misidentified gene
IIIb.
unknown gene-- potentially unlabelled homolog of gene family (e.g, XP_B = TLR21-like_Amis)

Figure S1 Schematic of three-step orthology curation. Steps taken to identify and curate TLR genes, including detailed decision tree for the manual curation and search for novel homologs (blue boxes). Scripts and programs are italicized, grey boxes are input records or data, and the goal of each step is provided. Amis, *Alligator mississippiensis.*

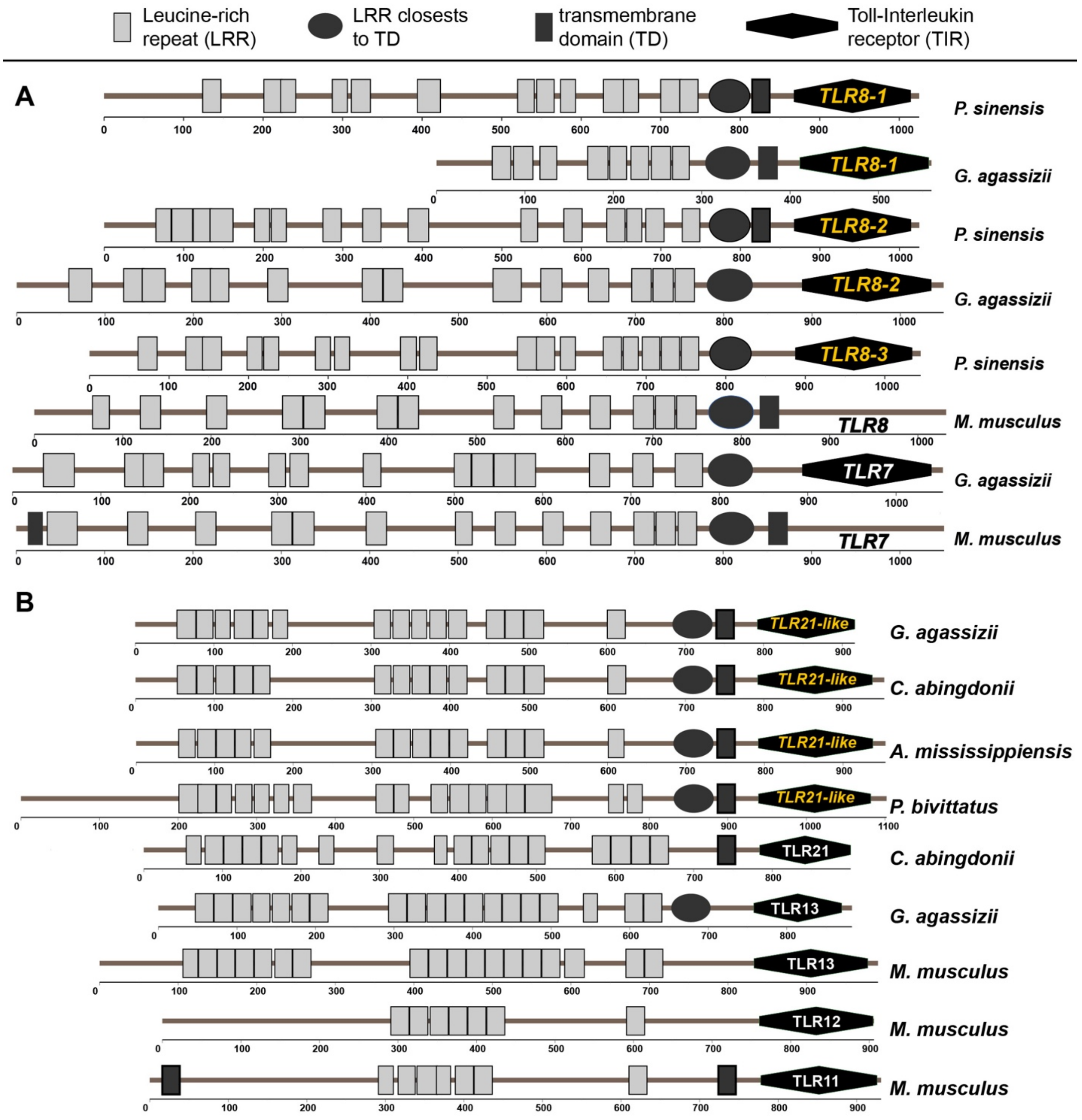

Figure S2 Motif evolution of Toll Like Receptor proteins. Based on motif detection of amino acid sequences from SMART: **A)** motif patterns of proteins from the TLR7 subfamily, and **B)** motif patterns from TLR21 and TLR21-like proteins. The number of leucine-rich-repeats (LRRs) vary in number and position (e.g., TLR7 of *G. agassizii* vs. *M. musculus*), and are largely responsible for the specificity and functionality of the protein. TLR11, 12, and 13 are shown for reference.

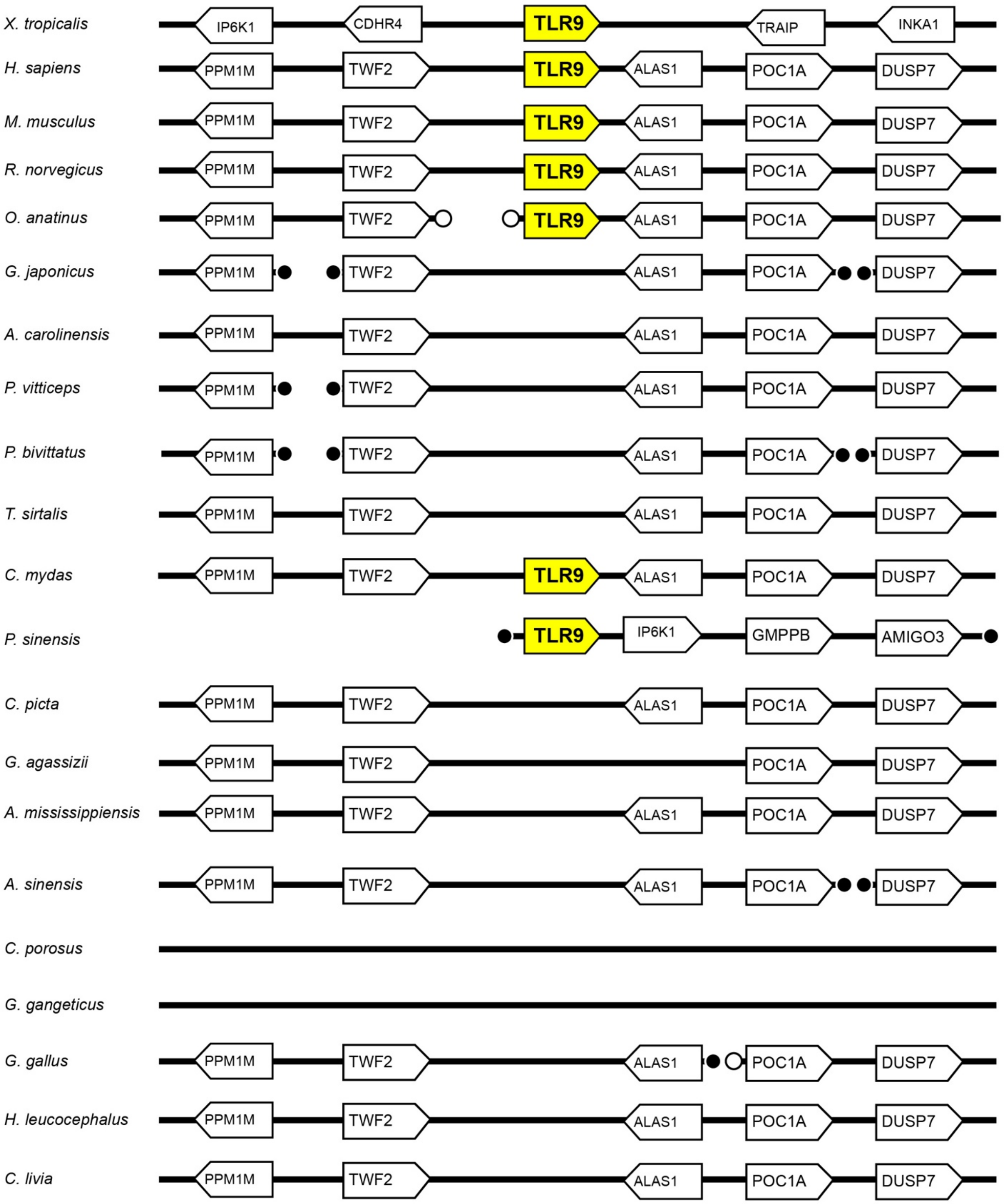

Figure S3 Syntenically conserved region for *TLR9*, which is a member of the TLR7 gene subfamily. *TLR9* has not been found in crocodilian and squamate genomes but is present in mammalian and some chelonian genomes. Filled circles represent the end of a scaffold and open circles indicate that the scaffold continues.

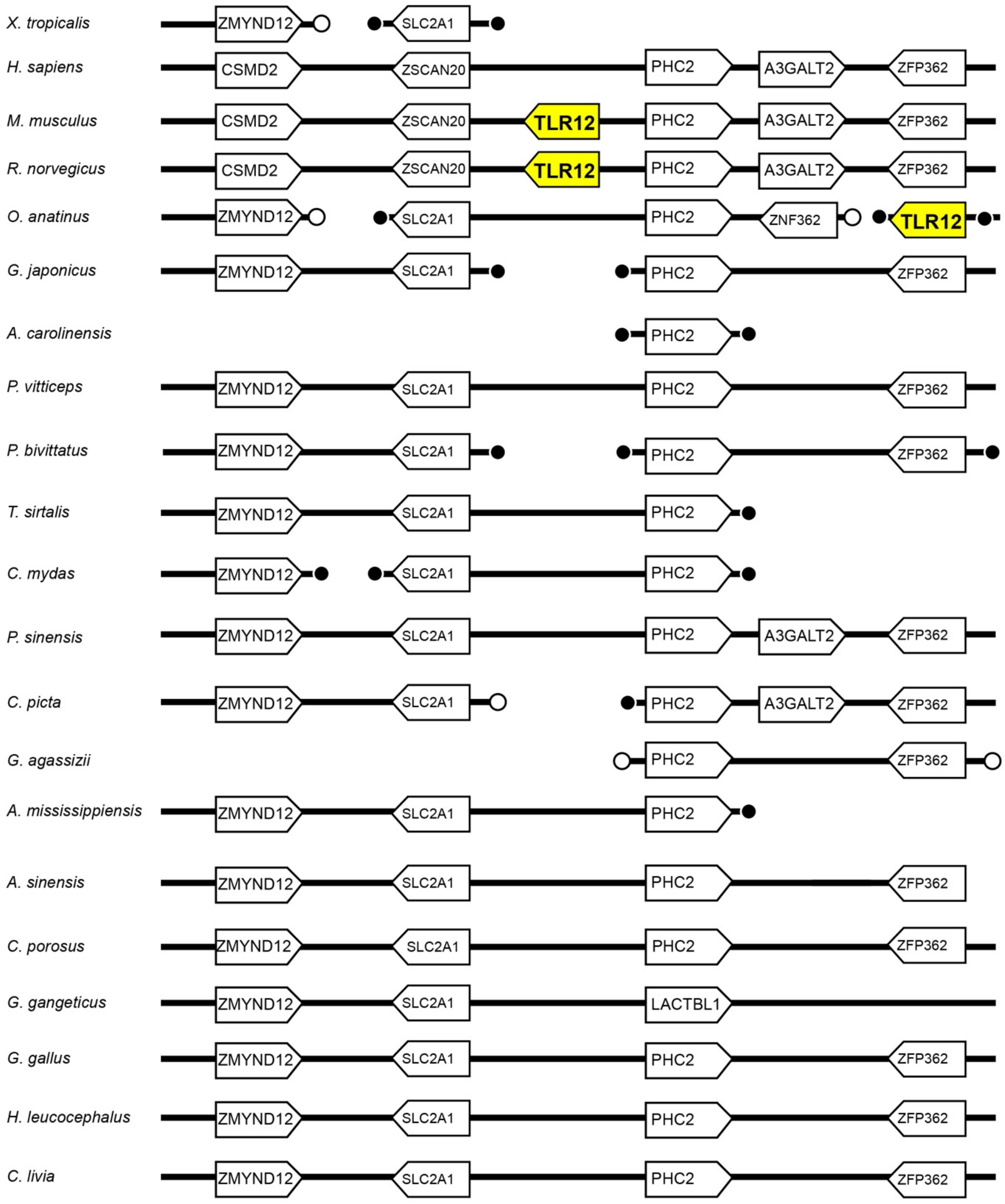

Figure S4 Syntenically conserved region for *TLR12*, which is a member of the TLR11 subfamily and has only been observed in mammals. Filled circles represent the end of a scaffold and open circles indicate that the scaffold continues.

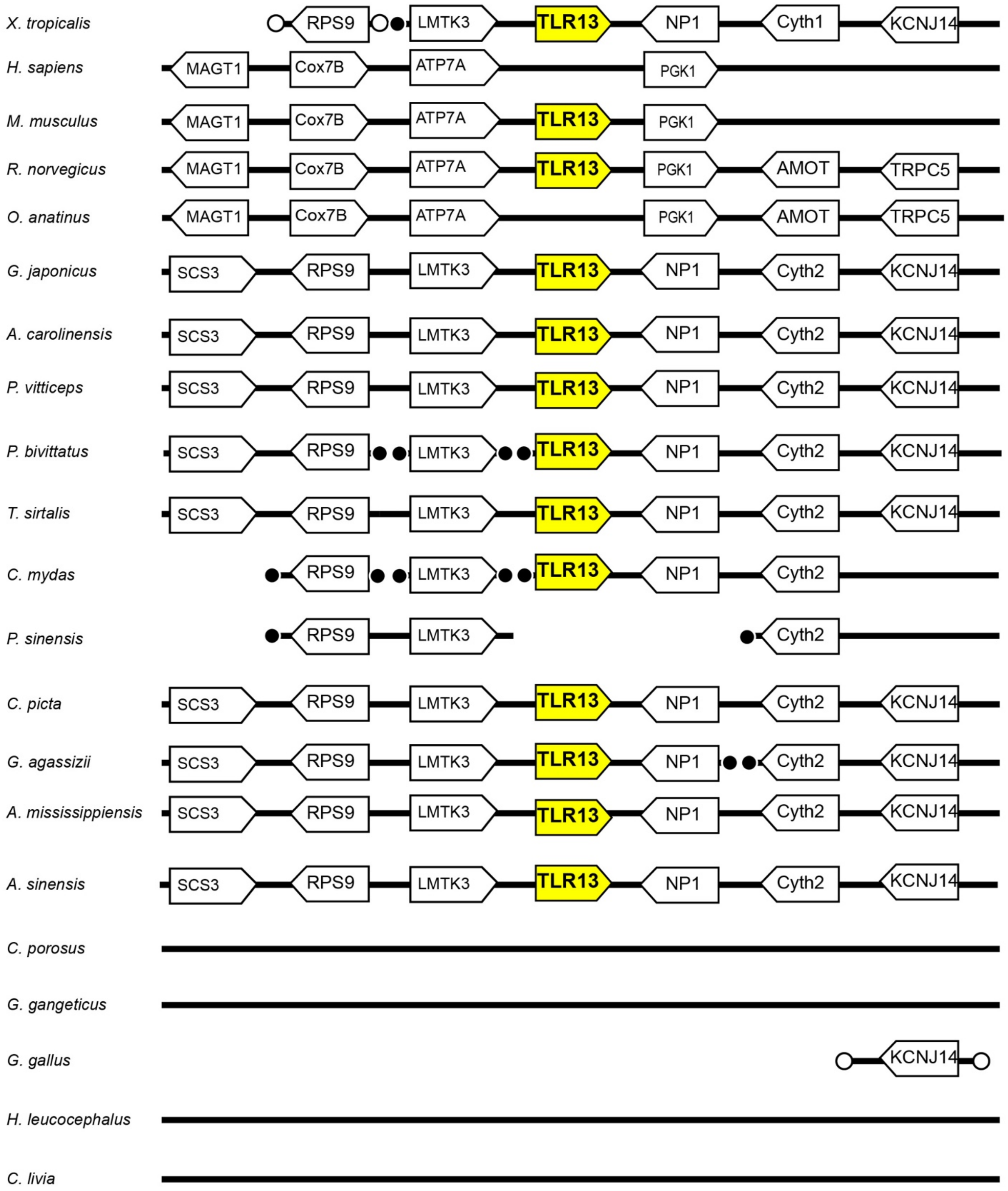

Figure S5 Syntenically conserved region for *TLR13* showing *TLR13* is present in most species with some variability among its neighboring genes. *TLR13* is part of the TLR11 subfamily. Species without a gene box do not have these genes present in the current genome annotation (e.g., *C. porosus*). Filled circles represent the end of a scaffold and open circles indicate that the scaffold continues.

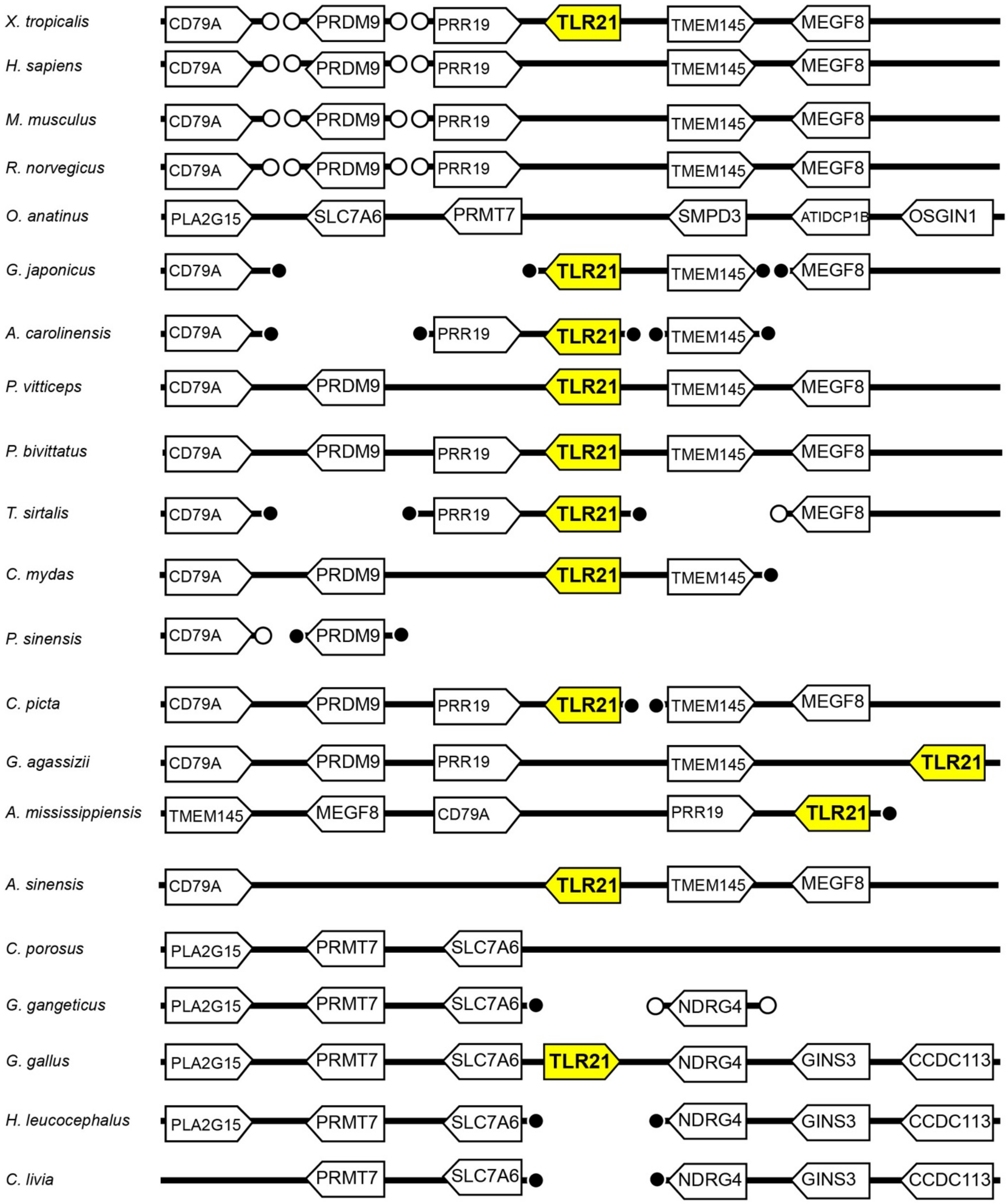

Figure S6 Syntenically conserved region for *TLR21*. The *TLR21* is not co-localized to the *TLR21-like* homolog within the genome, but their homology is based on phylogenetic analysis (see Figure 3). Filled circles represent the end of a scaffold and open circles indicate that the scaffold continues.

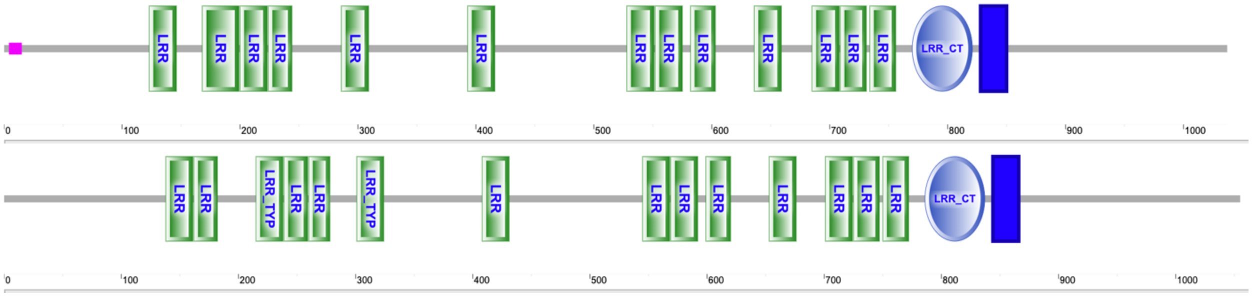

Figure S7 Comparison of protein motifs from *Xenopus tropicalis* for *TLR8-1* (top), *TLR8-2* (bottom). This duplication of *TLR8* may have been an independent event from the duplication observed in chelonians and crocodilians based on the fact that there is very high motif similarity between *TLR8-1* and *TLR8-2* of *X. tropicalis* (shown above) and they are phylogenetically sister to one another yet cluster separately from the *TLR8-1* and *TLR8-2* clades from other species (Figure 2B). There is not a clear parsimonious explanation based on *TLR8* losses and gains in tetrapods. If the ancestor of *X. tropicalis* and other species in this study had two *TLR8* homologues it would require five loss events of *TLR8* on three separate lineages, whereas an ancestral state of one *TLR8* homologue yields either three gain events and two loss events across five lineages, or two gain events and three loss events over four lineages (based on the phylogeny and presence/absence depicted in Figure 2A).

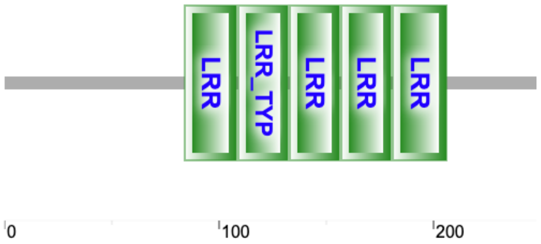

Figure S8 Motif analysis for Leucine rich repeat and Ig domain containing 1 (*LINGO1*) in *Anolis carolinensis*. *LINGO1* occurs in the syntenic region of *A. carolinensis* that is otherwise occupied by *TLR21-like* in snakes, turtles, and crocodilians (Figure 3A). It is unclear whether *LINGO1* translocated to precisely the syntenic location that *TLR21-like* occupies in other non-avian reptiles, or whether this ancestrally was the *TLR21-like* gene that experienced subsequent loss of the transmembrane domain and toll-interleukin receptor, leaving only the LRR repeats that make it identifiable as *LINGO1*.

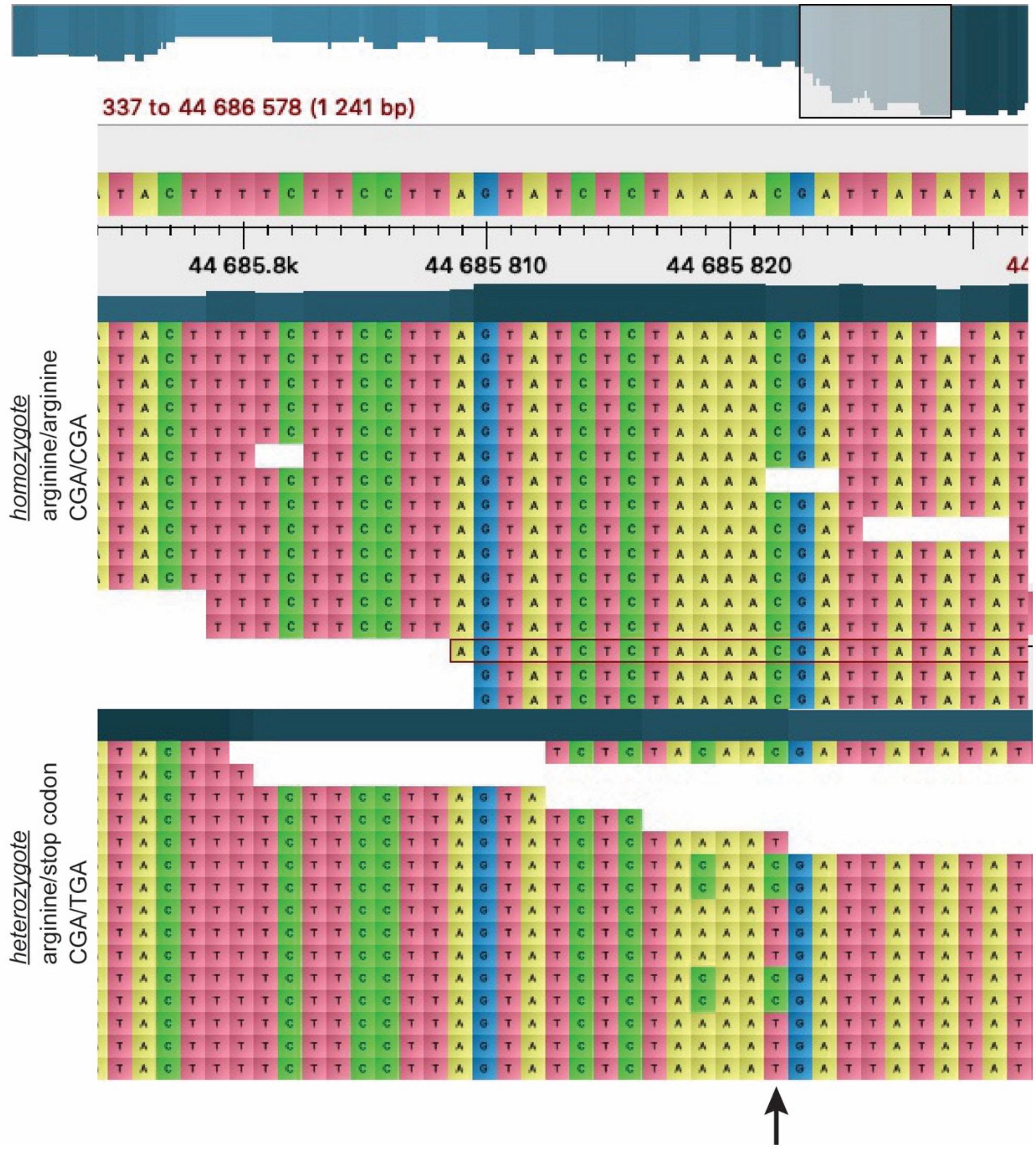
337 to 44 686 578 (1 241 bp)
T A C T T T T C T T C C T T A G T A T C T C T A A A A C G A T T A T A T A T
44 685.8k
44 685 810
44 685 820
44
homozygote
arginine/arginine
CGA/CGA
heterozygote
arginine/stop codon
CGA/TGA

Figure S9 TLR8B stop codon polymorphism. Mapping results of reads from resequencing data from two *G. agassizii* individuals: the top panel represents an individual with the arginine codon (CGA), while the bottom panel represents an individual (same population) that is heterozygous for the C/T allele that controls the CGA (arginine)/TGA (stop codon) polymorphism. The C/T polymorphism is at position 44,685,822 on Scaffold 3 in gopAga2.0.

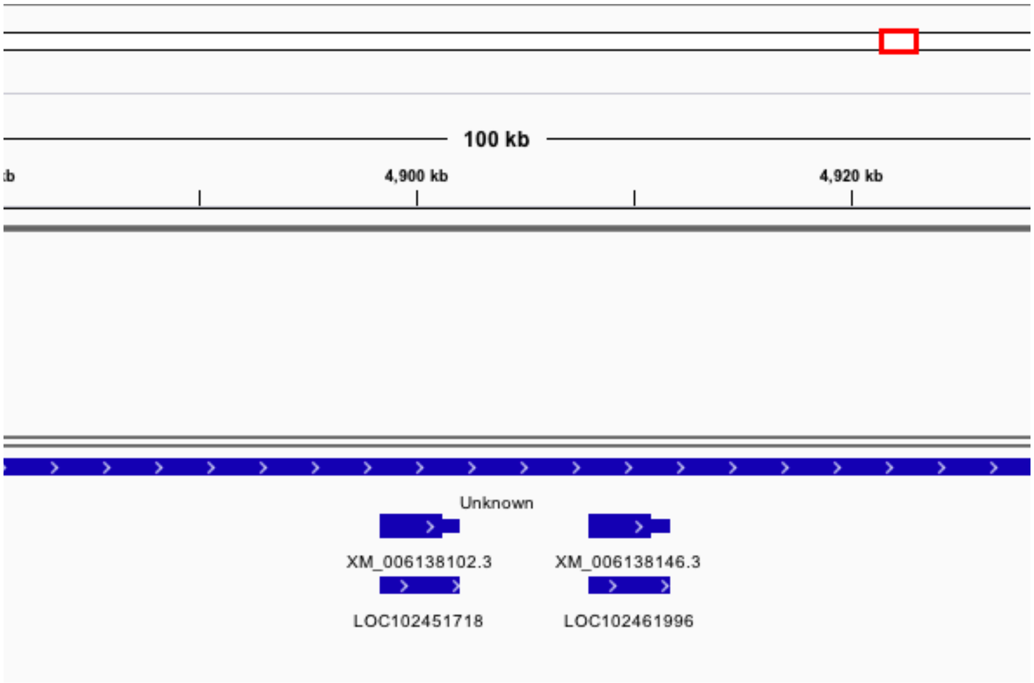

Figure S10 IGV map showing the co-localization of the *TLR21-like* 1 and *TLR21-like* 2 genes in the *Pelodiscus sinensis* genome, suggesting origins of these paralogues through a tandem duplication event. There is evidence of other TLR gene duplications in the *P. sinensis* genome (Figure 2).