Comparative genomic resources for spiny lizards (genus *Sceloporus*)

R. B. Harris*, B. L. Banbury and A. D. Leaché

Affiliation

Department of Biology & Burke Museum of Natural History and Culture, University of Washington, Seattle, WA, 98195-1800, USA
* Corresponding Author: rbharris@u.washington.edu

Keywords

Gene structure and function, reptiles, genomics/proteomics, comparative biology

Introduction

Spiny lizards (Genus *Sceloporus*) are a large (90+ species) and diverse clade of North American squamate reptiles (Bell et al., 2003) that has become a focal genus for integrative biological research: numerous studies have detailed the high degree of variation in morphology, behavior, life history, chromosome number, and sexual dimorphism. Research on *Sceloporus* has remained focused on ecological and evolutionary topics, with genetic analyses constrained to phylogenetics and systematics (Leaché, 2010; Leaché et al., 2013). No studies have yet attempted to annotate any of the *Sceloporus* genomes, yet next-generation sequencing are yielding large quantities of genome-scale data with the probability of capturing transcribed genes in the process.

Here, we provide annotations for 35 *Sceloporus* genomes to help expedite comparative genomic studies. Most of our sequencing effort is directed towards the Western Fence Lizard *S. occidentalis*. For this species, we used a whose genome shotgun approach to obtain large quantities of genome-scale data containing many transcribed genes. We obtain partial (∼2.7%) genomes of 34 other *Sceloporus* species using a reduced representation library.

Data Access

- *Sequence files* - All raw reads are freely available on the NCBI Sequence Read Archive under project accession number SRP041983
- *Assembly files* - All assemblies are available on Dryad.
- *Annotation files* - The functional annotations are available in Dryad (doi:10.5061/dryad.n2q7f)

Meta Information

- *Sequencing center* - Both the reduced-representation library (RRL) and whole genome shotgun (WGS) datasets were sequenced at the Vincent J. Coates Genomic Sequencing Laboratory at the University of California Berkeley (http://qb3.berkeley.edu/qb3/gsl/index.cfm).
- *Platform and model* - All individuals were run on an Illumina Hiseq 2000 with the exception of *S. cowlesi* and *S. tristicus* which were run on an Illumina Genome Analyzer IIx.
- *Design description* - We sampled 35 species of *Sceloporus* for comparative genome annotation (Table 1). The details of the RRL and WGS library preparation, sequencing, and de novo assembly are published in a recent study by Leaché et al. (2013). In this study, we conducted comparative population divergence analysis on eight species triplets, using a total of 22 species. The

RRL datasets for the species not used in this study (Table 1) were generated and assembled using the same methods. *S. occidentalis* was chosen for WGS as it is the most well-studied species in the *Sceloporus* genus and has a broad distribution throughout western North America. Genomic resources for this species will be useful for a maximal number of studies.

- *Run date* - All runs were completed between March 2010 and July 2012

Library

- *Strategy* - Whole-genome shotgun and reduced-representation library of whole-genomic DNA.
- *Taxon, Sex, and Location* - See Table 1.
- *Tissues* - Liver.
- *Sample handling* - All individuals used in this study are vouchered and deposited in museum collections as noted in Table 1.
- *Layout* - Paired end reads (2 x 100bp)
- *Library Construction Protocol* - The details of our library construction is published in Leaché *et al.* (2013). Briefly, we prepared the WGS using standard TruSeq protocol and conducted 100 bp, paired-end sequencing. For the RRL datasets, genomic DNA was sheared using StuI and fragments ranging in size from 1.5-2 kb were captured. These fragments were sheared into smaller fragments, libraries were prepared using standard TruSeq multiplexing protocols, and then paired-end sequenced in 100 bp reads.

Processing

- *Pipeline* - The full details of data filtering and de novo assembly are given in Leaché *et al.* 2013. Briefly, we used CLC Genomics Workbench v6 to qual-

ity filter and de novo assemble both the WGS and RRl datasets. Following assembly, consensus sequences from each species with length >1,000 bp and coverage >8x were combined into single-species fasta files. The gene prediction and annotation pipeline MAKER version 2.31.3 (Holt and Yandell, 2011) (last accessed April 21, 2014) was used to annotate each species based on *Anolis carolinensis* (AnoCar2.0.74) (Eckalbar et al., 2013). Each dataset was run through the MAKER pipeline twice. In the first round, MAKER implements RepeatMasker version 4.0.5 (http://www.repeatmasker.org/, last accessed April 21, 2014) to identify repetitive regions using the *Anolis* repeat library. The repeat masked sequences are then aligned to *Anolis* cDNA sequences using BLAST and *Anolis* peptide sequences are used to polish the resulting BLAST hits using the program Exonerate version 2.2 (prot2genome = 1) (Slater and Birney, 2005). Upon completion of MAKER round one, a draft training set was generated for ab initio gene prediction using the gene finding program SNAP (Korf, 2004). The second round of MAKER entails optimizing SNAP using this training set (prot2genome = 0). MAKER was run in parallel using the mpi version of the program on the University of Washington's HYAK computing cluster using 128 processors.

Both ab initio and evidence based gene predictions (the first and second pass through MAKER, respectively) were analyzed using InterProScan version 5.47 (Quevillon et al., 2005). Only ab initio gene predictions with positive InterProScan results are included in the final annotations. We did not filter Interproscan results beyond this, as e-values are dependent on the member database method and researchers may be interested in different criteria. Gene ontology and domains are included in the final gff file output.

Finally, to detect orthologs and paralogs, we input predicted genes from all species into the program OrthoMCL version 2.0.9 (Li et al., 2003) . OrthoMCL clusters unusually similar sequences into groups of high similarity. We include the chicken and human protein sequences for more detailed annotation information. The final group file contained all 34 RRL datasets, the *S. occidentals*

WGS, and the reference *Anolis*, chicken, and human genomes.

- *Runs* - The filtered reads were uploaded to the NCBI Sequence Read Archive in fastq format and are accessible from accession SRP041983.

Results

While all data is freely available via NCBI and Dryad, we have also made an easily searchable database through R shiny available at: https://rstudio.stat.washington.edu/shiny/sceloporus. All information about read characteristics are shown in Table 1.

- *Quality Scoring System* - Phred+33
- *Quality Scoring ASCII character* - ! to J
- *Annotation and Gene Ontology* - Annotation and gene ontology results are included in the final gff3 files available on Dryad along with the orthologous groups predicted by OrthoMCL.

Acknowledgements

We thank Ben Rubin, Carson Holt, and the MAKER development-wiki group for technical help. We also thank A. Gottscho, J. Lemos-Espinal, A. Nieto Montes de Oca, M. McElroy, L. Gray, C. Linkem, and J. Grummer for help collecting specimens. We thank our lab managers D. Reid and T. Gill for their assistance with library preparations. This work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system, supported in part by the University of Washington eScience Institute. This work was supported by grants from the National Science Foundation (DEB-1144630) and the University of Washington Royalty Research Fund (A61649).

References

E. L. Bell, H. M. Smith, and D. Chiszar. An annotated list of the species-group names applied to the lizard genus sceloporus. *Acta Zoologica Mexicana*, 91:103–174, 2003.

W. L. Eckalbar, E. D. Hutchins, G. J. Markov, A. N. Allen, J. J. Corneveaux, K. Lindblad-Toh, F. Di Palma, J. Alfoldi, M. J. Huentelman, and K. Kusumi. Genome reannotation of the lizard *Anolis carolinensis* based on 14 adult and embryonic deep transcriptomes. *BMC Genomics*, 14:49, 2013.

C. Holt and M. Yandell. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12:491, 2011.

I. Korf. Gene finding in novel genomes. *BMC Bioinformatics*, 5:59, 2004.

A. D. Leaché. Species trees for spiny lizards (genus: Sceloporus): identifying points of concordance and conflict between nuclear and mitochondrial data. *Molecular Phylogenetics and Evolution*, 54:162–171, 2010.

A.D. Leaché, R. B. Harris, M. E. Maliska, and C. W. Linkem. Comparative species divergence across eight triplets of lizards (sceloporus) using genomic sequence data. *Genome Biology and Evolution*, 5:2410–2419, 2013.

L. Li, C. J. Stoeckert, and D. S. Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13, 2003.

E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. Interproscan: protein domains identifier. *Nucleic Acids Research*, 33: 116–200, 2005.

G. S. C. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, 2005.

Table 1. Sampling, sequencing, and annotation information.

Species	Seq. Date	Voucher No.	Field No.	Sex	Collection Locality	Total Unfilt. Reads	Reads into Assembly	De novo contigs	N50	Mean coverage	Filtered contigs	MAKER evidence based predictions	MAKER ab initio based predictions	Ab initio predictions with positive IPR results
S. adleri	Jan 2012	UWBM 6608	ADL 4105	F	MEXICO; Guerrero, Asoleadero	61,396,744	31,351,090	368,090	474	13x	19,726	214	1,799	606
*S. angustus**	July 2012	LACM 13478	LA 457	?	MEXICO; Baja California Sur, Isla Santa Cruz	59,008,874	28,606,707	534,720	522	28x	21,129	156	1,485	426
S. bicanthalis	July 2012	UWBM 7307	ADL 4153	M	MEXICO; Districto Federal, 11 km W Rio Frio	50,963,292	31,215,601	247,932	495	16x	22,722	114	1,235	618
*S. carinatus**	July 2012	UWBM 6614	ADL 4050	M	MEXICO; Chiapas, Sierra Madre de Chiapas	79,553,536	46,700,905	894,635	474	8x	25,580	208	1,727	517
S. clarkii	Mar 2011	MVZ 245876	TJD 101	F	USA; AZ, Santa Cruz Co., Coronado N.F.	NA**	18,542,404	59,562	376	55x	1,183	17	67	29
S. cowlesi	Mar 2010	AMNH 154059	ADL 432	F	USA; AZ, Apache Co.	48,762,864	26,224,281	278,468	949	10x	45,510	483	5,490	396
S. edwardtaylori	Jan 2012	UWBM 6588	MTM 005	F	MEXICO; Oaxaca, Juchiten de Zargoza	45,692,228	23,424,584	272,080	495	12x	19,798	184	1,589	583
*S. exsul**	Jan 2013	UWBM 6590	ADL 4113	M	MEXICO; Queretaro, Pena Blanca	35,733,442	14,996,169	191,313	373	8x	440	249	4,548	4,258
S. formosus	Jan 2012	UWBM 6623	ADL 4088	F	MEXICO; Guerrero, Omeltemi	64,971,516	35,813,869	590,161	495	9x	25,571	251	2,297	834
S. gadoviae	July 2012	UWBM 7309	ADL 4163	F	MEXICO; Puebla, Zapotitlan Salinas	58,190,400	26,021,380	288,885	383	14x	15,018	160	1,589	493
*S. graciosus**	Mar 2011	MVZ 240898	ADL 876	F	USA; CA, Tuolumne Co., Yosemite N.P.	NA**	10,215,985	9,838	309	71x	25	0	0	0
S. grammicus	Jan 2012	UWBM 6585	ADL 4096	F	MEXICO; Guerrero, Asoleadero	47,583,134	25,273,167	258,309	539	13x	26,712	228	2,212	85
S. horridus	Jan 2013	UWBM 6632	POE 3887	F	MEXICO; Guerrero, Tierra Colorada	37,356,428	19,275,595	131,289	567	20x	15,518	30	1,029	922
S. hunsakeri	Jan 2013	SDSNH 76079	ADG 098	F	MEXICO; Baja California Sur	44,180,416	25,580,920	158,212	533	17x	16,292	340	5,178	4,686
S. jalapae	July 2012	UWBM 7318	ADL 4159	M	MEXICO; Puebla, San Luis Temalacayuca	69,585,852	38,721,933	741,561	467	8x	21,367	215	2,120	657
S. licki	Jan 2013	SDSNH 76080	ADG100	F	MEXICO; Baja California Sur	33,801,198	16,485,334	133,173	550	17x	14,702	271	5,065	4,647
S. magister	Jan 2013	UWBM 7395	ADL 4471	F	USA; Arizona, Coconino Co., Marble Canyon	34,953,494	17,964,775	103,055	650	19x	12,020	298	4,943	4,559
*S. malachiticus**	Mar 2011	MVZ 263420	SMR 450	F	HONDURAS Corts, Parque Nacional Cusuco	NA**	21,965,000	81,711	369	49x	1,702	36	86	55
*S. mucronatus**	Jan 2012	UWBM 6636	ADL 4092	F	MEXICO; Guerrero, Asoleadero	55,355,942	26,574,363	331,892	475	12x	19,885	166	1,047	418
S. occidentalis	**Mar 2011**	**MVZ**	**ADL 3279**	**F**	**USA; CA, Tuolumne Co., Yosemite N.P.**	**NA****	**40,849,442**	**955,511**	**2,967**	**29x**	**413,800**	**6,806**	**134,144**	**30,991**
S. ochoterenae	Jan 2012	UWBM 6641	ADL 4111	M	MEXICO; Guerrero, Omeltemi	66,333,598	31,248,947	292,345	533	15x	25,003	376	2,521	1,173
*S. olivaceus**	Jan 2013	UWBM 7968	JWS 631	?	USA; TX, Arlington	31,389,948	16,658,706	121,157	650	18x	16,213	236	4,382	3,882
S. orcutti	Jan 2013	UWBM 7654	ADG 102	M	USA; CA, Riverside Co.	38,845,798	23,213,887	154,480	514	15x	14,267	300	4,906	4,540
S. palaciosi	July 2012	UWBM 7313	ADL 4155	M	MEXICO; Districto Federal	65,853,622	32,395,045	163,616	605	22x	21,754	140	1,109	371
S. scalaris	Jan 2012	UWBM 6589	ADL 4126	F	MEXICO; Jalisco, Rancho las Papas	33,561,800	24,697,422	465,770	454	10x	15,411	229	1,720	766
*S. siniferus**	Jan 2012	UWBM 6653	ADL 4067	F	MEXICO; Oaxaca, Mixtequila	50,630,798	21,938,063	311,347	468	11x	13,866	121	1,124	299
*S. smithi**	Jan 2012	UWBM 6662	ADL 4071	F	MEXICO; Oaxaca, Mixtequila	47,525,652	25,097,617	279,889	493	12x	22,162	138	1,794	438
S. spinosus	Jan 2012	UWBM 6672	ADL 4124	M	MEXICO; Jalisco, Rancho las Papas	59,078,332	32,562,785	546,964	475	9x	21,779	155	1,361	431

Species	Seq. Date	Voucher No.	Field No.	Sex	Collection Locality	Total Unfilt. Reads	Reads into Assembly	De novo contigs	N50	Mean coverage	Filtered contigs	MAKER evidence based predictions	MAKER ab initio based predictions	Ab initio predictions with positive IPR results
S. taeniocnemis	Mar 2011	MVZ 264322	SMR 657	F	GUATEMALA; Departamento El Progreso	NA**	17,959,114	74,107	388	41x	2,136	29	78	28
*S. torquatus**	Jan 2012	UWMB 6600	ADL 4125	F	MEXICO; Jalisco, Rancho las Papas	67,811,820	33,838,916	296,861	522	17x	25,539	365	2,740	1,133
S. tristichus	Mar 2010	AMNH 153948	ADL 403	F	USA; AZ, Navajo County, Holbrook	53,101,800	31,013,091	311,638	937	10x	51,533	610	6,673	2,356
*S. utiformis**	Mar 2011	MVZ 236299	TJP 26512	M	MEXICO; Guerrero, 17 km E Bajos del Ejido	NA**	15,587,168	93,407	356	29x	1,653	32	217	105
*S. variabilis**	Jan 2012	UWBM 6678	MTM 002	M	MEXICO; Oaxaca, San Pedro Tapanatepec	75,896,002	44,504,186	752,328	505	9x	27,802	239	2,167	806
*S. woodi**	Jan 2013	UWBM 7265	RA X64	F	USA; FL, Marion County, Ocala N.F.	35,209,562	15,225,180	227,886	423	34x	9,957	280	4,893	4,534
S. zosteromus	Jan 2013	SDSNH 76081	ADG 074	M	MEXICO; Baja California Sur	23,051,026	9,746,243	88,389	628	16x	10,793	236	4,691	4,260

* denotes those species not used in the previous study by Leaché et al. (2013).

** Raw Illumina reads unavailable. Only CLC Genomics filtered data is available.