Title:

De novo transcriptome assembly and polymorphism detection in ecologically important widely distributed Neotropical toads from the *Rhinella marina* species complex (Anura: Bufonidade)

Authors:

Coralie Nourisson[a*], Miguel Carneiro[a], Marcelo Vallinoto[ab], Fernando Sequeira[a]

Affiliations:

[a]CIBIO-InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal

[b]Institute of Coastal Studies (IECOS), Universidade Federal do Pará, Bragança Campus, 68600-000 Bragança - PA, Brasil

[*] Corresponding author: coralie.nourisson@gmail.com

Introduction:

The toads *Rhinella marina* and *R. schneideri* are large terrestrial true toads widely distributed in the Neotropical region, including most of South America, and in the case of the former also ranges up to the south of Texas, in North America (Frost 2014). These two species are morphologically similar but diagnosable by the presence of a tibial gland in *R. schneideri*. They have a broad parapatric distribution, but occur in sympatry in areas of transition between the Amazon rainforest (typically inhabited by *R. marina*) and Cerrado (typically inhabited by *R. schneideri*). Recent studies have reported instances of hybridization between *R. schneideri* and *R. marina* in the south part of the Amazon forest in “islands of Cerrado” (Vallinoto et al. 2010; Sequeira et al. 2011) that are likely remnants of past Amazon forest retraction (Pennington et al. 2000). Considering that ecological settings in this area of overlap are heterogeneous, it is possible that persistence of both toad species is determined by environmental selection (Harrison 1986).

Furthermore, an extensive mtDNA unidirectional introgression from *R. schneideri* into *R. marina* has been reported, which likely derived by the retreat of *R. schneideri* populations from the Amazon and subsequent southward expansion into the present-day Cerrado, associated with environmental changes during the Pleistocene/Holocene (Sequeira et al. 2011). However, laboratory crosses between *R. marina* females and males of two other *Rhinella* species resulted in hybrid offspring constituted only of females (Blair 1972; Malone and Fontenot 2008), suggesting that the hypothesis that asymmetric reproductive isolation may explain the direction of introgression. Despite these studies, key evolutionary questions associated with the impact of hybridization and/or local adaptation of these species are still poorly understood.

These and other related questions can be approached by investigations of hybrid zones. The study of hybrid zones and local adaptations has recently seen a deep conceptual transition from qualitative analysis based on a limited set of markers to genome-wide approaches. These have the potential to develop thousands of SNPs and the identification of candidate genes related to the processes of the acquisition of reproductive isolation and adaptive divergence.

Here, we report *de novo* transcriptome characterization and polymorphism detection of *R. marina* and *R. schneideri*, which will contribute to deepen the knowledge of the genetic architecture of reproductive isolation, and spatio-temporal dynamics of the hybrid zones between these toad species. Three additional important reasons make these toad species promising models for genomic-scale studies. First, they are associated with a suite of well described life-history and ecological traits (Zug and Zug 1979; Malone and Fontenot 2008). Second, genetics of speciation in amphibians has received little attention compared to other taxonomic groups, and therefore the genomic data of these species can be used for various ecological and evolutionary studies of closely related species. Finally, the development of these genomic data will be of special interest for *R. marina* since it is one of most successful invader of the world, in particular in Australia, where its rapid spread has been devastating for native biodiversity (Phillips and Shine 2004; Tingley et al. 2014). With these data we further hope to contribute to a more thorough understanding of the success of the global invasion of this toad.

Data Access:

- NGS sequence data: Sequence files can be found on NCBI Sequence Read Archive under project number: PRJNA255079 (accession number SRP044269)
  - Sequences of the non-redundant assembly transcripts (.bam file) and SNP data (.vcf file) can be found in DRYAD. doi:10.5061/dryad.3jm3n

Meta Information:

- Sequencing center – Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain
- Platform and model – Illumina HiSeq 2000
- Design Description- the goals of our study were to generate a transcriptome assembly for two species of *Rhinella*, and to identify SNPs that will allow us to examine patterns of introgression between the two species.
- Analysis type – mRNA
- Run date – samples loaded in three different flow cells : 2013-07-10, 2013-05-27, 2013-05-10

Library:

- Strategy – mRNA-Seq
- Taxon – *Rhinella marina* and *R. Schneideri*
- Sex – unknown
- Tissue – liver
- Location – *Rhinella marina* and *R. schneideri* were collected in the Brazilian states of Amapá (Macapá; - 0.021245°; -51.074525°) and Goiás (Goiânia; -16.65185°; -49.22994°), respectively.
- Sample handling – Liver tissue was freshly excised at the laboratory and placed immediately into RNA- later. Samples were first stored at room temperature for three days, and then placed at 4ºC until RNA processing for approximately one week.
  - Additional sample information –Total RNA was isolated from five individuals of each species, and then equimolar amounts of each individual RNA extracted sample were pooled together for each species. The TruSeq RNA Sample Preparation Kit was used to generate mRNA-focused libraries from total RNA through a polyA selection. The mRNA was not normalized.
- Selection –mRNA
- Layout – paired end fragments 2x76 bp, >240 M reads
- Library Construction Protocol- TruSeq RNA sample preparation kit (Illumina Inc)
  - Nominal Sizes were estimated directly from the assembly: 163 (stdev46) for *Rhinella Marina* and 164 (stdev46) for *Rhinella schnedeiri*

Processing:

Raw sequence processing and de novo assembly:

The quality of the reads generated by Illumina sequencing was assessed with the FastQC software Version 0.10.1. Based on a visual inspection of all sequenced lanes, raw reads were cleaned using Trimmomatic-0.30 (Lohse et al 2012) which 1) removed adaptors and other Illumina-specific sequences (given by the laboratory), 2) removed bases off the start and the end of a read if below quality 3, 3) we scanned the read with a 4-base wide sliding window and cut when the average quality per base dropped below 15, and 4) eliminated reads below 36 bases long. The quality of the reads was re-checked with FastQC after this step. In Table 1 we described the number of raw reads, number of reads after cleaning, and total number of aligned reads.

The reads were then concatenated and de novo assembled by means of the Trinity software (Grabherr et al. 2011) using default parameters following the protocol from Haas and collaborators (2013). We only used reads for which both pairs remained after the quality control step. Trinity stat was used to report the number of transcripts, number of components, and the transcripts contig N50 value. The largest and smallest transcripts, as well as the total, median and average sizes were calculated (Table 1).

Two quality control steps were carried out. First, software Bowtie was used as described in Haas and collaborators (2013) to map reads on the original assembly and mapping consistency was assessed by counting how many times paired reads mapped to the same transcript (Table 1). Second, the integrity of the transcripts was assessed. Predicted ORFs were defined from *Rhinella marina* transcripts and were blasted against the western clawed frog *Xenopus (Silurana) tropicalis* protein coding genes downloaded from ensembl.org (Flicek et al. 2013).

SNP calling:

SNP calling was performed by mapping reads from both species onto the R. marina transcriptome. Prior to SNP calling, transcripts were cleaned for non-redundancy, to have unique set of genes with no duplication of isoform. Duplicate reads were removed using PICARD and mapping was performed using BWA-MEM (Li H. and Durbin R., 2009). SNP calling was carried out using SAMtools (Li et al. 2009) with the following quality criteria: a minimum depth coverage of 10X, a mapping quality of 20, at least 10 bp from indels and a SNP quality of 30. FST for each SNP was calculated using allele counts according to Karlsson and collaborators (2007) only for SNPs with at least 20X coverage in both species (i.e a subset of the total SNPs called).

- Runs: Twelve files were submitted to NCBI SRA and divided in two experiments corresponding to each species. In each experiment, six files were submitted corresponding to the three differents flow cells and the two directions (paired ended, 1.fasq.gz and 2.fastq.gz).
- Run data file type : fastq.gz
- File Name : Rhinella_marina_a_1.fastq.gz, Rhinella_marina_a_2.fastq.gz, Rhinella_marina_b_1.fastq.gz, Rhinella_marina_b_2.fastq.gz, Rhinella_marina_c_1.fastq.gz, Rhinella_marina_c_2.fastq.gz, Rhinella_schneideri_a_1.fastq.gz, Rhinella_schneideri_a_2.fastq.gz, Rhinella_schneideri_b_1.fastq.gz, Rhinella_schneideri_b_2.fastq.gz, Rhinella_schneideri_c_1.fastq.gz, Rhinella_schneideri_c_2.fastq.gz

Results:

In total, 554,474,410 transcriptome sequencing reads were obtained for *R. marina* and 554,276,922 for *R. schneideri*. After removing the reads with adaptors and reads with low qualities, 531,369,210 reads (95.8%) for *R. marina* and 532,324,046 reads (96.0%) for *R. schneideri* (Table 1) remained.

Clean reads were assembled into a total of 199,799 transcripts with an average length of 949 bp and a N50 length of 3106 bp for *R. marina* and 172,671 transcripts with average length of 964 bp and a N50 length of 3106 bp for *R. schneideri* (Table 1). Statistics on the assembly, numbers of both reads mapping to the assembly, as well as

number of mapping inconsistencies, smallest and largest transcript, total, median and average sizes before and after redundancy step, are presented in Table 1. A total of 1,184,765 SNPs were called.

- Quality scoring system: phred+33
- Mean / Median coverage per contig : Table 1
- Polymorphism rate: Of the total called SNPs, 709,193 SNPs were identified for both species with a coverage >20x. The genome wide differentiation between *R. marina* and *R. schneideri* show shared, fixed and polymorphic SNPs over those SNPs (Figure 1 and 2).
- 3,386 predicted ORFs from *R. marina* blast again *Xenopus (Silurana) tropicalis*, and the great majority (60%) aligned to 80% or more of the full length of the *Xenopus* gene, indicating that the assembly shows good contiguity. Note that we could only establish orthology for a small fraction of the contigs due to high divergence between *Xenopus* and *Rhinella*.

Table 1: Results of the transcriptome read and assembly for *R. marina* and *R. schneideri.*

	Rhinella marina	*Rhinella schneideri*
total number of reads	554,474,410	554,276,922
number of reads after cleaning	531,369,210	532,324,046
Number of reads aligned	493,944,277 (92.96%)	501,923,873 (94.29%)
Mapping both	476,867,188 (96.54%)	486,792,608 (96.99%)
Mapping inconsistencies	5,699,018 (1.15%)	4,750,228 (0.95%)
N50	1,936	2,007
Total Trinity transcripts	199,799	172,671
Total Trinity components	131,020	117,518
Transcripts size		
Total	189,691,296	166,546,570
Largest	20,495	22,312
Smallest	201	201
Median	430	426
Average	949	964
Transcripts size after redundancy step		
Total	80,251,892	75,188,658
Largest	17,329	22,276
Smallest	201	201
Median size	331	336
Average size	612	640

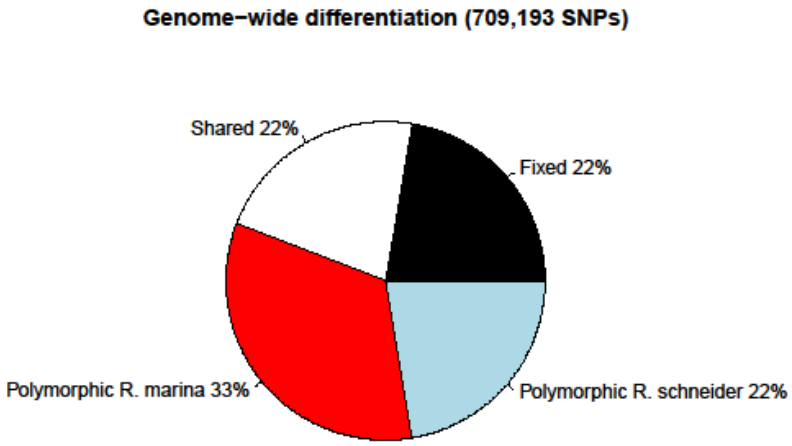

Figure 1: Relative proportion of fixed, shared, and exclusive polymorphisms between *R. marina* and *R. schneideri*

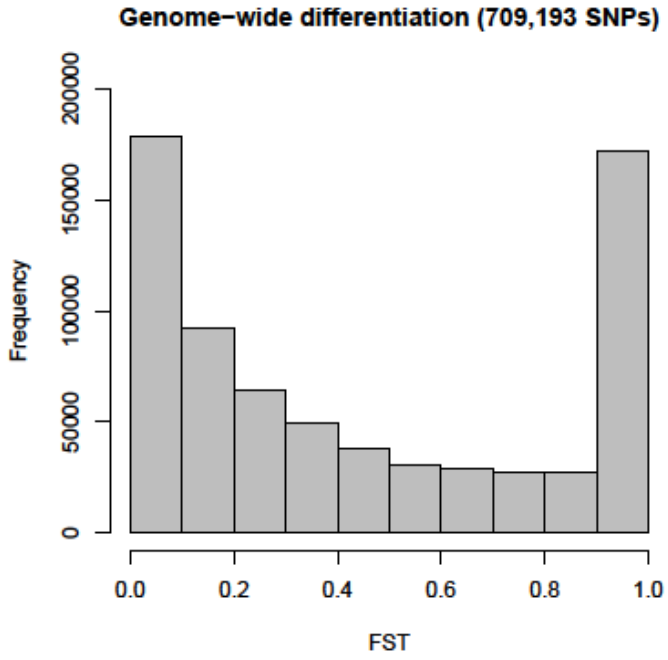

Figure 2: Histogram of the genome wide differentiation between *R. marina* and *R. schneideri* using the total number of SNPs.

Acknowledgments

This project was funded by Fundação para a Ciência e a Tecnologia (FCT) through the research project PTDC/BIA BEC/105093/2008 (funded by FEDER through the COMPETE program and Portuguese national funds). CN was funded by the project "Genomics and Evolutionary Biology" cofinanced by North Portugal Regional Operational Programme 2007/2013 (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF). FS and MC were supported by Postdoctoral grants (SFRH/BPD/87721/2012; SFRH/BPD/72343/2010, respectively) from FCT under the Programa Operacional Potencial Humano-Quadro de Referência Estratégico Nacional funds from the European Social Fund and Portuguese Ministério da Educação e Ciência. MV was supported by a Postdoctoral grant (CNPq/CSF 232916/2013-6), a researcher fellowship (CNPq/CSF 232916/2013-6) and was funded by the project MCT/CNPq 14/2013 (473313/2013-8) from Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil. We thank the Centro Nacional de Análisis Genómico (CNAG) for their valuable services. We also thank T. Brunes and Y. Oliveira for field work, and N. Lins and S. Afonso for valuable help at the Lab.

References

Blair WF (1972) Evidence from hybridization In Evolution in the Genus Bufo. Edited by: Blair WF. Austin: University of Texas Press:196-232.

Flicek P, I Ahmed, M. R Amode et al. (2013) Ensembl. Nucleic Acids Research 2013 41 Database issue:D48-D55

Frost, Darrel R. 2014. Amphibian Species of the World: an Online Reference. Version 6.0 (4 August 2014). Electronic Database accessible at http://research.amnh.org/herpetology/amphibia/index.html. American Museum of Natural History, New York, USA.

Grabherr MG, Haas BJ, Yassour M, et al. (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 15;29(7):644-52.

Haas BJ, Papanicolaou A, Yassour M, et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(8):1494-512.

Harrison, R. G. (1986) Pattern and process in a narrow hybrid zone. Heredity. 56: 337–349

Karlsson EK, Baranowska I, Wade CM, et al. (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet. 39(11):1321-8.

Li H. and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

Li H., Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

Lohse, M, Bolger AM, Nagel A, et al. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res (40(Web Server issue)), W622-627.

Malone J, Fontenot B (2008) Patterns of reproductive isolation in toads. PLoS one, 3(12):e3900.

Pennington R, Prado D, Pendry C (2000) Neotropical seasonally dry forests and Quaternary vegetation changes. J. Biogeogr., 27:261-273.

Phillips BL, Shine R (2004) Adapting to an invasive species: Toxic Cane Toads induce morphological change in Australian snakes. Proceedings of the National Academy of Sciences USA 101(49):17150-17155.

Sequeira F, Sodre D, Ferrand N, et al. (2011) Hybridization and massive mtDNA unidirectional introgression between the closely related Neotropical toads *Rhinella marina* and *R. schneideri* inferred from mtDNA and nuclear markers. BMC Evolutionary Biology 11(1): 264.

Tingley R, Vallinoto M, Sequeira F, Kearneyd MR (2014) Realized niche shift during a global biological invasion. Proc. Natl. Acad. Sci. U.S.A. (Early edition).

Vallinoto M., Sequeira F, Sodré D, Bernardi JAR, Sampaio I, Schneider H (2010) Phylogeny and biogeography of the Rhinella marina species complex (Amphibia, Bufonidae) revisited: implications for Neotropical diversification hypotheses. ZoologicaScripta, 39, 128–140.

Yang Y and Smith SA (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics, 14:328