MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2017) **17**, 546–549 doi: 10.1111/1755-0998.12563

GENOMIC RESOURCES NOTE

De novo transcriptome assembly and polymorphism detection in two highly divergent evolutionary units of Bosca's newt (*Lissotriton boscai*) endemic to the Iberian Peninsula

CORALIE NOURISSON,[1] ANTONIO MUÑOZ-MERIDA,[1] MIGUEL CARNEIRO and FERNANDO SEQUEIRA

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

Abstract

This article reports the de novo transcriptome assemblies of two highly divergent evolutionary units of the Iberian endemic Bosca's newt, *Lissotriton boscai*. These two units are distributed mostly allopatrically but overlap in the central–southwestern coastal region of Portugal. The resources we provide include the raw sequence reads, the assembled transcripts, the annotation and SNPs called for both lineages.

Keywords: amphibians, evolutionary theory, hybridization, speciation, transcriptomics

Received 13 October 2015; revision received 18 November 2015; accepted 19 November 2015

Introduction

The Bosca's newt, *Lissotriton boscai*, is an Iberian endemic species distributed throughout the western part of the Iberian Peninsula. This morphologically uniform species exhibits deep levels of genetic structure (Martínez-Solano *et al.* 2006; Teixeira *et al.* 2015) with two divergent evolutionary units that have been identified based on the analysis of nuclear and mtDNA data sets. Lineage A is distributed throughout most of the species distribution range, whereas lineage B is restricted to the central–southwest part of the Iberian Peninsula. Following Teixeira *et al.* (2015), the initial split between the ancestors of lineages A and B is thought to have occurred in the Miocene, approximately 9 million years ago (Myr). Both lineages were subsequently fragmented into multiple sublineages inferred to have diverged during the Pleistocene. Interestingly, the Teixeira *et al.* (2015) study based on cytonuclear patterns of gene flow and admixture suggested the occurrence of admixture likely to have occurred in the areas of secondary contact between the divergent lineages. Although the study of Teixeira *et al.* (2015) lacked detailed analysis of those contact areas, their results seem to match broad patterns of secondary contact and admixture found in several Iberian vertebrate species (e.g. Sequeira *et al.* 2005; Godinho *et al.* 2006).

A large body of work on speciation research has been focused on the genetics of reproductive isolation. In this context, amphibians have long been viewed as important model organisms for studying the architecture of species boundaries and the speciation process, because many taxa that have evolved in isolation for millions of years still exchange genes in areas of secondary contact (e.g. Mallet *et al.* 2007). However, most studies in nonmodel organisms still rely on cline model measurements of a limited set of markers to infer dynamics of secondary contact between divergent evolutionary units, such as the role of natural selection and dispersal rates on the outcome of genetic admixture. Recent developments of next-generation sequencing (NGS) offer the opportunity to generate genomewide sequence data sets in nonmodel organisms and to detect genomic regions associated with species-specific adaptations or reduced hybrid fitness, which is a crucial first step for identifying specific genes or mutations implicated in reproductive isolation and local adaptation.

Here, we report a transcriptome characterization and polymorphism detection in two highly divergent lineages of *L. boscai*. Considering that *L. boscai* includes several population groups with varying degrees of

Correspondence: Coralie Nourisson, Fax: +351 252 661 780; E-mail: coralie.nourisson@gmail.com

[1] The first two authors contributed equally to this work.

divergence, the development of these genomic data will be of special importance to investigate different levels of reproductive isolation. Finally, these data will also be valuable resource for related species as few-omics data are currently available for amphibians.

Data access

- NGS sequence data: Sequence files can be found on NCBI Sequence Read Archive under project number: PRJNA296559 (Accession no. SRP063964), Experiments SRX1271756 (boscai_portugal) and SRX1271510 (boscai_spain).
- Sequences of the nonredundant assembly transcripts (.fasta file), the annotation (text tabulated file) and SNP data (.vcf file) can be found in DRYAD: doi: 10.5061/dryad.gp5j0

Meta information

- Sequencing centre—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 lineages of *Bosca's newts* and to identify SNPs that will allow us to examine patterns of introgression between the two lineages.
- 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—*Lissotriton boscai*
- Sex—unknown
- Tissue—liver
- Location—*Lissotriton boscai* were collected in the Iberian Peninsula, in Sierra de Aracena, Spain (−6.547919, 37.880160) and Vila Nova de Milfontes, Portugal (−8.797032, 37.750631), respectively.
- Sample handling—liver tissue was freshly excised at the laboratory and placed immediately into RNA later.
  - ◦ Additional sample information—total RNA was isolated from five individuals of each evolutionary unit (*boscai_portugal* and *boscai_spain*), 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 2 × 76 bp, 239.8 M reads
- Library Construction Protocol—TruSeq RNA sample preparation kit (Illumina Inc)
  - ◦ Nominal sizes were estimated directly from the assembly: 176 (stdev185) for *boscai_spain* and 170 (stdev129) for *boscai_portugal*.
- Runs: twelve files were submitted to NCBI SRA and divided into two experiments corresponding to each evolutionary unit (*boscai_spain* and *boscai_portugal*). In each experiment, six files were submitted corresponding to the three different flow cells and the two directions (paired ended, 1.fasq.gz and 2.fastq.gz).
- Run data file type: fastq.gz
- File Name:

  Boscai_spain_a_1.fastq, Boscai_spain_a_2.fastq,
  Boscai_spain_b_1.fastq, Boscai_spain_b_2.fastq,
  Boscai_spain_c_1.fastq, Boscai_spain_c_2.fastq,
  Boscai_portugal_a_1.fastq, Boscai_portugal_a_2.fastq,
  Boscai_portugal_b_1.fastq, Boscai_portugal_b_2.fastq,
  Boscai_portugal_c_1.fastq, Boscai_portugal_c_2.fastq,

Processing

Raw sequence processing and de novo assembly

A first-quality assessment of the reads generated by the sequencer was performed by means of FASTQC software version 0.10.1 (Andrews 2010). After visualizing the quality of the reads, sequence trimming was made using TRIMMOMATIC-0.30 (Lohse *et al.* 2012). Several steps were performed: (i) the removal of adaptors and other Illumina-specific sequences (as provided by the sequencing centre), (ii) trimming of bases in the ends of reads with quality below 30, (iii) a read scan with a 4-base wide sliding window to remove those read fragments with an average quality per base below 15 and (iv) removal of reads below 36 bases long. After the cleaning step, the quality of the reads was rechecked with FASTQC. Table 1 shows a summary of the number of raw reads, number of reads after cleaning and total number of aligned reads.

The resulting reads were then used to perform a de novo assembly by means of the TRINITY software version 2.0.4 (Grabherr *et al.* 2011) following the protocol from Haas *et al.* (2013). Only reads with both pairs remaining after the trimming were selected for the assembly step. TRINITY stat was used to report several statistics summarizing the overall length of the resulting assemblies and inferred transcripts, such as number of transcripts, the contig N50 value, the largest and smallest transcripts, as well as the total, median and average sizes (Table 1).

Two further steps were carried out after the initial assembly. First, all contigs were clustered using CD-HIT (parameters: -c 0'9 -n 8) (Weizhong & Godzik 2006) to

Table 1 Results of the transcriptome reads and assembly for *boscai_spain* and *boscai_portugal*

	Boscai_spain	*Boscai_portugal*
Total number of reads	568 552 398	537 221 754
Number of reads after cleaning	545 925 346	518 758 606
Number of reads aligned	523 102 838 (95.81%)	487 846 523 (94.04%)
Mapping both paired ends	512 373 616 (97.95%)	471 196 924 (96.59%)
Mapping inconsistencies	2 990 586 (0.57%)	5 920 244 (1.21%)
N50	2805	2706
Total TRINITY transcripts	153 270	141 317
Total transcripts after redundancy	119 365	109 080
Transcript size		
Total	173 736 688	157 842 550
Largest	21 325	20 485
Smallest	224	224
Median	433	434
Average	1134	1117
Transcripts size after redundancy step		
Total	121 309 544	109 720 360
Largest	21 325	20 485
Smallest	224	224
Median size	407	410
Average size	1016	1006

remove redundancy (Table 1). Second, ORFs were defined from transcripts (ORF predictions were made using the online tool ORFPREDICTOR (http://proteomics.ysu.edu/tools/OrfPredictor.html) and those with no ORF in any frame were removed from the final contig set. The remaining set of contigs for each population was annotated by SMA3S software (parameters: -a 123 -d uniprot_vertebrates.dat -p F) (Muñoz-Mérida *et al.* 2013) using UNIPROT (only the vertebrates taxonomy) and compared between populations to detect the proportion of contigs in common.

SNP calling

SNP calling was performed by mapping reads from both species onto the reference transcriptomes created for *boscai_spain* and *boscai_portugal*. Duplicate reads were removed using PICARD with the option MarkDuplicates, and mapping was performed using BWA-MEM with default parameters (Li & Durbin 2009). SNP calling was carried out using SAMTOOLS (Li *et al.* 2009) with the following quality criteria: a minimum depth coverage of 10×, a mapping quality of 20 and a SNP quality of 30.

Results

In total, 568 552 398 transcriptome sequencing reads were obtained for *Boscai_spain* and 537 221 754 for *Boscai_portugal*. After removing the reads with adaptors and reads with low quality, 545 925 346 reads (96.02%) for *Boscai_spain* and 518 758 606 reads (96.56%) for *Boscai_portugal* remained (Table 1).

Clean reads were assembled into a total of 119 365 transcripts with an average length of 1134 bp and a N50 length of 2805 bp for *Boscai_spain* and 109 080 transcripts with an average length of 1117 bp and a N50 length of 2706 bp for *Boscai_portugal* (Table 1). Statistics of the assembly, number of reads mapping to the assembly, number of mapping inconsistencies, smallest and largest transcript, as well as total, median and average sizes before and after redundancy step, are presented in Table 1. A total of 1 104 799 SNPs were called.

- Quality scoring system: phred+33, quality scoring ASCII character range: 19–40
- Mean/Median coverage per contig: Table 1
- Polymorphism rate: 668 492 SNPs were identified when mapped to the *boscai_spain* transcriptome and

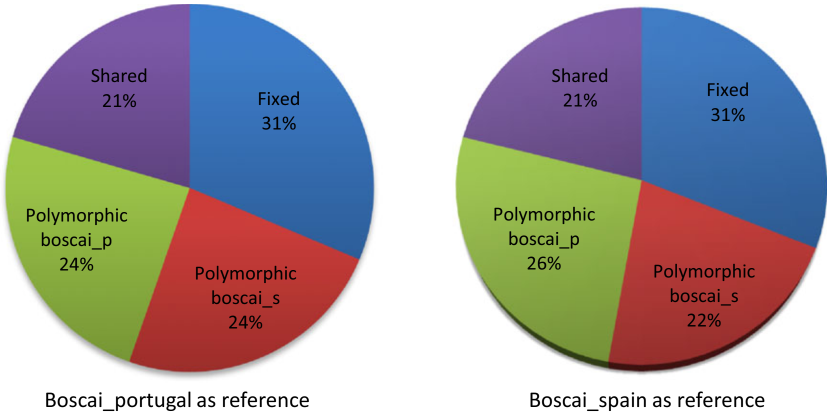

Fig. 1 Genomewide differentiation: 644 465 and 668 492 SNPs mapping to *boscai_portugal* and *boscai_spain* transcriptomes, respectively. Relative proportion of fixed between populations, shared and exclusive polymorphisms between the two lineages of *Lissotriton boscai* (Portugal and Spain). [Colour figure can be viewed at wileyonlinelibrary.com]

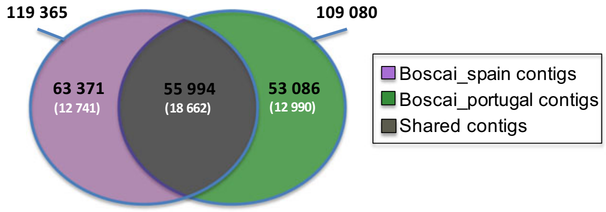

Fig. 2 Number of unique and shared contigs in the two *Lissotriton boscai* evolutionary units. The number of contigs annotated with Sma3s is shown in white. [Colour figure can be viewed at wileyonlinelibrary.com]

644 465 when mapped to *boscai_portugal*. The genome-wide differentiation between *boscai_spain and boscai_portugal* was summarized in shared, fixed and polymorphic SNPs from the total number of SNPs (Fig. 1). In this study, we considered a fixed SNP when the two populations displayed alternative and diagnostic alleles. These SNPs provide good markers to characterize each population and study patterns of admixture in contact areas.

- After removing the redundancy from the data sets, contigs were annotated withSma3s (GO terms, domains from InterPro, pathway, keywords, interactions from IntAct) for 31 403 sequences in *boscai_spain* and 31 652 in *boscai_portugal* (Fig. 2).
- A comparison between the two populations was performed using BLAST taking as database *boscai_spain* sequences to find the corresponding contig in the *boscai_portugal* assembly. We considered two sequences to be homologous if they presented at least an e-value of 10^{-3} and a minimum of 70% of the length of the contig (Fig. 2).

Acknowledgements

This project was funded by Fundação para a Ciência e a Tecnologia (FCT) through the research project PTDC/BIA BEC/105083/2008 (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 was supported by Postdoctoral grant (SFRH/BPD/87721/2012) 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. The work was supported by POPH-QREN funds from the European Social Fund and Portuguese MCTES [FCT Investigator (IF/00283/2014) and postdoc grant to MC (SFRH/BPD/72343/2010)]. We thank the Centro Nacional de Análisis Genómico (CNAG) for their valuable services. We also thank Jesús Díaz Rodríguez for sampling and T. Brunes and S. Afonso for valuable help at the laboratory.

References

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Godinho R, Mendonça B, Crespo EG, Ferrand N (2006) Genealogy of the nuclear R-fibrinogen locus in a highly structured lizard species:comparison with mtDNA and evidence for intragenic recombination in the hybrid zone. *Heredity*, **96**, 454–463.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

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. *Nature Protocols*, **8**, 1494–1512.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.*, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lohse M, Bolger AM, Nagel A *et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**(Web Server issue), W622–W627.

Mallet J, Beltran M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, **7**, 28.

Martínez-Solano I, Teixeira J, Buckley D, García-París M (2006) Mt-DNA phylogeography of *Lissotriton boscai* (Caudata, Salamandridae): evidence for old, multiple refugia in an Iberian endemic. *Molecular Ecology*, **15**, 3375–3388.

Muñoz-Mérida A, Viguera E, Claros MG, Trelles O, Pérez-Pulido AJ (2013) Sma3s: a three-step modular annotator for large sequence datasets. *DNA Research*, **21**, 341–353.

Sequeira F, Alexandrino J, Rocha S, Arntzen JW, Ferrand N (2005) Genetic exchange across a hybrid zone within the Iberian endemic golden-striped salamander, *Chioglossa lusitanica*. *Molecular Ecology*, **14**, 245–254.

Teixeira J, Martínez-Solano I, Buckley D, Tarroso P, García-París M, Ferrand N (2015) Genealogy of the nuclear β-fibrinogen intron 7 in Lissotriton boscai (Caudata, Salamandridae): concordance with mtDNA and implications for phylogeography and speciation. *Contributions to Zoology*, **84**, 193–215.

Weizhong L, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

FS and MC conceived the study. CN, MC and AMM analyzed the data. CN and AMM wrote the first draft of the manuscript. All authors contributed to the writing of the manuscript.

Data accessibility

NGS sequence data: Sequence files can be found on NCBI Sequence Read Archive under project number: PRJNA296559 (Accession no. SRP063964), Experiments SRX1271756 (boscai_portugal) and SRX1271510 (boscai_spain).

Sequences of the nonredundant assembly transcripts (.fasta file), the annotation (text tabulated file) and SNP data (.vcf file) can be found in DRYAD: http://dx.doi.org/10.5061/dryad.gp5j0