# A practical guide to build *de-novo* assemblies for single tissues of non-model organisms: The example of a Neotropical frog

**Santiago Montero-Mendieta** [Corresp., 1] , **Manfred Grabherr** [2] , **Henrik Lantz** [2] , **Ignacio De la Riva** [3] , **Jennifer A Leonard** [1] , **Matthew T Webster** [4] , **Carles Vilà** [Corresp. 1]

[1] Conservation and Evolutionary Genetics Group, Department of Integrative Ecology, Doñana Biological Station (EBD-CSIC), Consejo Superior de Investigaciones Científicas, Seville, Spain

[2] Department of Medical Biochemistry and Microbiology, National Bioinformatics Infrastructure Sweden (BILS), Uppsala Universitet, Uppsala, Sweden

[3] Department of Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Madrid, Spain

[4] Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala Universitet, Uppsala, Sweden

Corresponding Authors: Santiago Montero-Mendieta, Carles Vilà
Email address: santiago.montero@ebd.csic.es, carles.vila@ebd.csic.es

Whole genome sequencing (WGS) is a very valuable resource to understand the evolutionary history of poorly known species. However, in organisms with large genomes, as most amphibians, WGS is still excessively challenging and transcriptome sequencing (RNA-seq) represents a cost-effective tool to explore genome-wide variability. Non-model organisms do not usually have a reference genome and the transcriptome must be assembled *de-novo*. We used RNA-seq to obtain the transcriptomic profile for *Oreobates cruralis*, a poorly known South American direct-developing frog. In total, 550,871 transcripts were assembled, corresponding to 422,999 putative genes. Of those, we identified 23,500, 37,349, 38,120 and 45,885 genes present in the Pfam, EggNOG, KEGG and GO databases, respectively. Interestingly, our results suggested that genes related to immune system and defense mechanisms are abundant in the transcriptome of *O. cruralis*. We also present a pipeline to assist with pre-processing, assembling, evaluating and functionally annotating a *de-novo* transcriptome from RNA-seq data of non-model organisms. Our pipeline guides the inexperienced user in an intuitive way through all the necessary steps to build *de-novo* transcriptome assemblies using readily available software and is freely available at: https://github.com/biomendi/TRANSCRIPTOME-ASSEMBLY-PIPELINE/wiki

1  **A practical guide to build *de-novo* assemblies for single tissues of**
2  **non-model organisms: The example of a Neotropical frog**
3
4

6  **Santiago Montero-Mendieta[1*], Manfred Grabherr[2], Henrik Lantz[2], Ignacio De la Riva[3],**
7  **Jennifer A. Leonard[1], Matthew T. Webster[4] & Carles Vilà[1*]**
8
9
10  [1] Conservation and Evolutionary Genetics Group, Department of Integrative Ecology, Doñana
11  Biological Station (EBD-CSIC), Avd. Américo Vespucio N26, 41092 Seville, Spain
12
13  [2] Department of Medical Biochemistry and Microbiology, National Bioinformatics Infrastructure
14  Sweden (BILS), Uppsala University, Uppsala, Sweden
15
16  [3] Department of Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales,
17  Consejo Superior de Investigaciones Científicas, Madrid, Spain
18
19  [4] Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala
20  University, Uppsala, Sweden
21
22
23  * Corresponding authors:
24
25  E-mail: santiago.montero@ebd.csic.es (SMM)
26
27  E-mail: carles.vila@ebd.csic.es (CV)

28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

# Abstract

Whole genome sequencing (WGS) is a very valuable resource to understand the evolutionary history of poorly known species. However, in organisms with large genomes, as most amphibians, WGS is still excessively challenging and transcriptome sequencing (RNA-seq) represents a cost-effective tool to explore genome-wide variability. Non-model organisms do not usually have a reference genome and the transcriptome must be assembled *de-novo*. We used RNA-seq to obtain the transcriptomic profile for *Oreobates cruralis*, a poorly known South American direct-developing frog. In total, 550,871 transcripts were assembled, corresponding to 422,999 putative genes. Of those, we identified 23,500, 37,349, 38,120 and 45,885 genes present in the Pfam, EggNOG, KEGG and GO databases, respectively. Interestingly, our results suggested that genes related to immune system and defense mechanisms are abundant in the transcriptome of *O. cruralis*. We also present a pipeline to assist with pre-processing, assembling, evaluating and functionally annotating a *de-novo* transcriptome from RNA-seq data of non-model organisms. Our pipeline guides the inexperienced user in an intuitive way through all the necessary steps to build *de-novo* transcriptome assemblies using readily available software and is freely available at: https://github.com/biomendi/TRANSCRIPTOME-ASSEMBLY-PIPELINE/wiki

# Introduction

The word "genomics" refers to the study of the complete set of genes and gene products in an individual. With the ongoing reduction of costs, this is frequently achieved through the use of high-throughput sequencing technologies (Reuter *et al.* 2015). The "genomics era" formally started after the Human Genome Project (HGP) was first published in 2001 (Lander *et al.* 2001). Since then, genomics has drastically changed the way that we understand and study the genetic features of living organisms. Mainly due to novel gene discovery, genomics has proved useful in many fields, such as molecular medicine (Giallourakis *et al.* 2005), molecular anthropology (Destro-Bisol *et al.* 2010), social sciences (McBride *et al.* 2010), evolutionary biology (Wolfe 2006) and biological conservation (Mcmahon *et al.* 2014), among others. Nowadays, a main use of genomics is to profile genomes, transcriptomes, proteomes, and metabolomes (Schuster 2008). Genomics has also proved highly informative in elucidating evolutionary history of species and, for example, has enabled finding genes that could explain the variation in beak size within and among species of Darwin's finches, in addition to providing new insights into the evolutionary history of these birds (Lamichhaney *et al.* 2015, 2016).

At the time of writing this article (January 2017), 8951 genomes had been completely sequenced according to the Genomes OnLine Database (GOLD) (https://gold.jgi.doe.gov) (Mukherjee *et al.* 2017). These genomes include mainly unicellular organisms (4,958 bacteria; 240 archaea) and viruses (3,473) due to their small genome size. Eukaryote organisms usually have larger genomes and the sequencing effort to fully sequence them is much larger. Only 280 eukaryote genomes have been completed, most of them belonging to model organisms (i.e. species that have been widely studied because of particular experimental advantages or biomedical interest). However, the difficulties associated with the assembly of large genomes have resulted in very few of these being fully sequenced. Among terrestrial vertebrates, amphibians have the largest

94   genome sizes. The average genome size of frogs is 5.0 gigabases (Gb), while the fire salamander
95   (*Salamandra salamandra*) genome averages 34.5 Gb (Gregory *et al.* 2007). For this reason, few
96   genomics studies on amphibians have been carried out so far. To date, only the genome of three
97   frogs of reduced genome size, *Xenopus tropicalis* (1.5Gb; Hellsten *et al.* 2010), *Xenopus laevis*
98   (2.7Gb; Session *et al.* 2016) and *Nanorana parkeri* (2.3Gb; Sun *et al.* 2015), have been
99   sequenced and published, in contrast to the larger number of genomes of reptiles (10), birds (53)
100  and mammals (43). Due to the difficulties to obtain reference genome sequences for species with
101  large genome sizes, reduced representation approaches are a cost-effective way to obtain
102  information on genome-wide variability. For non-model organisms in which whole genome
103  sequencing (WGS) is not feasible, transcriptome (e.g. Geraldes *et al.* 2011; De Wit *et al.* 2015)
104  or exome (Lamichhaney *et al.* 2012) sequencing are commonly used as a reduced representation
105  of the genome.
106
107  In amphibians, 24 transcriptomes from 19 species are currently available in the Transcriptome
108  Shotgun Assemblies (TSA) database (https://www.ncbi.nlm.nih.gov/genbank/tsa/, January
109  2017), highlighting the importance of RNA sequencing (RNA-seq) for genomic studies in this
110  group. RNA-seq is more affordable than whole genome sequencing and has rapidly become the
111  preferred method for cataloguing and quantifying the complete set of transcripts or messenger
112  RNA for a specific tissue, developmental stage or physiological condition (Wang *et al.* 2009).
113  Nowadays, RNA-seq has a wide variety of uses but the core analyses include transcriptome
114  profiling, differential gene expression and functional profiling (Conesa *et al.* 2016). As
115  transcriptome assembly becomes more common for non-model and poorly known organisms, we
116  expect it will become a more popular tool also in phylogenomics as well as in demographic and
117  population structure inference. However, what kind of RNA-seq data analysis is to be performed
118  depends on the species of interest and the research goals. For model organisms and their close
119  relatives, RNA-seq data is analyzed by mapping reads to a reference genome. By contrast, most
120  non-model organisms do not have a reference genome from a sufficiently closely related species,
121  and the transcriptome must be assembled *de-novo* (Martin & Wang 2011). Many bioinformatics
122  tools to build a *de-novo* transcriptome are now available, yet contrasting opinions about the steps
123  to follow may be disorienting. Some extremely simple pipelines have been developed to
124  automatize the process (e.g. TRUFA; Kornobis *et al.* 2015), but this may limit the flexibility of
125  the different pieces of software that have been integrated.
126
127  Here, we present the transcriptome profile for *Oreobates cruralis*, a direct-developing frog
128  species from the Amazonian regions of Bolivia and Peru. To date, this is the first transcriptome
129  available for a South American amphibian. We also present a simple pipeline for pre-processing,
130  building and functionally annotating a *de-novo* transcriptome from RNA-seq data of non-model
131  organisms using available software.
132

## Methods

134

## Study model and sample collection

136
137      The genus *Oreobates* Jiménez de la Espada, 1872 (Anura: Craugastoridae) is a poorly
138  studied clade of New World direct-developing frogs (Terrarana) distributed from the lower
139  slopes of the eastern Andes into the upper Amazon basin, encompassing from southern Colombia

140 and western and central Brazil up to northern Argentina. More than half of the 24 identified
141 species have been described in the last decade and the species diversity in this genus is likely to
142 be underestimated (Köhler & Padial 2016). One of these species, *O. cruralis* (Boulenger, 1902)
143 occurs in a wide range of elevations and habitats across Bolivia and Peru. Its distribution
144 includes lowland Amazonian rainforests (approximate altitudinal range, from 100 to 600 meters
145 above sea level, m.a.s.l.), Yungas-montane Amazonian rainforests (600–2500 m.a.s.l.), and inter-
146 Andean dry valleys (1300–3000 m.a.s.l.) (De la Riva *et al.* 2000). However, little is known about
147 its ecology and evolutionary history.

148
149 For this study we used tissue samples from a single individual of *O. cruralis,* sampled in Bolivia
150 (Villa Tunari, Cochabamba, Bolivia; 345 m.a.s.l.; 16º59'01.4"S 65º24'30.16"W) on November
151 28th, 2013 and deposited at the tissue collection of the Museo Nacional de Ciencias Naturales
152 (MNCN-CSIC) in Madrid, Spain (MNCN/ADN:65263; Colección Boliviana de Fauna, CBF
153 7268) for which some tissue samples were available. Unfortunately, the specimen was unsexed.
154 Samples of five tissues (intestine, liver, spleen, heart and skin) were isolated and preserved in
155 Nucleic Acid Preservation (NAP) buffer (Camacho-Sánchez *et al.* 2013) at the time of sampling
156 and were later kept at -80ºC. Unfortunately, no other tissues were preserved under the same
157 conditions.

158
## Ethics statement
160
161 Field surveys that led to sampling the studied specimen were approved by the Dirección General
162 de la Biodiversidad, Ministerio de Medio Ambiente y Agua, La Paz, Bolivia (Approval number:
163 MMAyA-VMA-DGBAP Nº 1592/12), and were supported and approved by the Spanish
164 Government (Ministerio de Economía y Competitividad, Project numbers: CGL2011-30393 and
165 CGL2013-47547-P, awarded to IDlR).

166
## Transcriptome sequencing
168
169 We extracted whole RNA for each tissue using the RNeasy Protect Mini Kit (Qiagen).
170 RNA quality was evaluated with RNA ScreenTape on TapeStation by Agilent. Due to poor RNA
171 quality, two tissues were discarded (skin and heart), thus only RNA extracts from intestine, liver
172 and spleen were used (RIN, RNA integrity number, scores of 6.2, 7.3 and 7.1, respectively).
173 Sequencing libraries were prepared and sequenced by the SNP&SEQ Technology Platform
174 (Uppsala University) from 1µg total RNA using the TruSeq stranded mRNA library preparation
175 kit (Illumina Inc.) and including poly-A selection. The library preparation was performed
176 according to the manufacturers' protocol. The quality of the libraries was evaluated using the
177 Agilent Technologies TapeStation and a DNA 1000-kit Screen Tape. The adapter-ligated
178 fragments were quantified by qPCR using the Library quantification kit for Illumina (KAPA
179 Biosystems) on a StepOnePlus instrument (Applied Biosystems/Life technologies) prior to
180 cluster generation and sequencing. A 14 pM solution of RNA was subjected to cluster generation
181 and paired-end sequencing with 125 bp (base pair) read length on a HiSeq2500 instrument
182 (Illumina Inc.) using the v4 chemistry according to the manufacturer's protocols.

183
## RNA-seq data analysis
185

186     The overall pipeline is summarized in Figure 1 and our practical guide is available at:
187     https://github.com/biomendi/TRANSCRIPTOME-ASSEMBLY-PIPELINE/wiki.  Briefly,  the
188     first step after obtaining the raw sequence data in FASTQ format was to perform a preliminary
189     quality            control            analysis            with            FastQC
190     (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC delivers quality metrics
191     that are useful to identify if the data requires initial pre-processing before the transcriptome
192     assembly. The pre-processing stage included three steps: first, removal of possible ribosomal
193     RNA (rRNA) contamination; second, trimming low quality bases and PCR adapters; third,
194     normalization to remove large excess of reads corresponding to moderately and highly expressed
195     transcripts. Pre-processing is not always needed but it is highly recommendable to improve
196     assembly quality. Once the data was pre-processed, a quality control was performed again and
197     then, clean normalized reads were *de-novo* assembled in absence of a reference genome.
198     Subsequent analyses depend on the study goals. In our case, transcripts were functionally
199     annotated using various databases to obtain a transcriptome profile. All steps are described in
200     further detail in the following paragraphs.
201
202     We filtered raw FASTQ reads using SORTMERNA-v2.1 (Kopylova *et al.* 2012) against 8
203     default rRNA databases (SILVA 16S bacteria, SILVA 16S archaea, SILVA 18S eukarya, SILVA
204     23S bacteria, SILVA 23s archaea, SILVA 28S eukarya, Rfam 5S archaea/bacteria, Rfam 5.8S
205     eukarya) to remove rRNA. Then, we used TRIMMOMATIC-v0.32 (Bolger *et al.* 2014) to trim
206     adaptors and sequences with Phred quality score < 20. We normalized cleaned data of each tissue
207     using the *in-silico* normalization utility included in the TRINITY-2.2.0 package (Grabherr *et al.*
208     2011). Normalization is useful for large RNA-seq data sets (>300 million paired-end reads)
209     because it will remove over-expressed transcripts, thus lowering computing memory
210     consumption and speeding up the assembly process (Haas *et al.* 2013). We merged the resulting
211     data for the three tissues into a single dataset and normalized again to remove redundant
212     sequences that could have been obtained from several tissues prior to assembly. We used
213     TRINITY (Grabherr *et al.* 2011) to *de-novo* assemble normalized reads into contigs. This
214     resulted in a large number of transcripts, much higher than the expected number of genes, likely
215     because of alternative splicing. To avoid redundant transcripts, we kept the longest isoform for
216     each    "gene"    identified    by    TRINITY    (unigene)    using    the
217     "get_longest_isoform_seq_per_trinity_gene.pl" utility in TRINITY. Thus, each unigene
218     represented a collection of expressed sequences (i.e. transcripts) that apparently came from the
219     same transcription locus, representing a putative gene. This set of unigenes was kept for
220     downstream analyses.
221
222     We evaluated the quality of the assembly and the transcript contiguity in terms of read
223     representation by mapping normalized reads back to the set of unigenes using BOWTIE-1.1.2
224     (Langmead *et al.* 2009). We assessed the assembly completeness in terms of gene content using
225     BUSCO-v1 (Simao *et al.* 2015) by searching the unigenes for the presence or absence of
226     conserved         orthologs         in         the         tetrapoda-odb9         database
227     (http://busco.ezlab.org/datasets/tetrapoda_odb9.tar.gz) that represents a collection of 3,950
228     single-copy tetrapoda orthologs. We also mapped with E-value ≤ 1E-20 the unigenes to the
229     SwissProt database (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/) and to the Western clawed frog
230     (*Xenopus                          tropicalis*)                          proteome
231     (http://ftp.ensembl.org/pub/release81/fasta/xenopus_tropicalis/pep/Xenopus_tropicalis.JGI_4.2.p

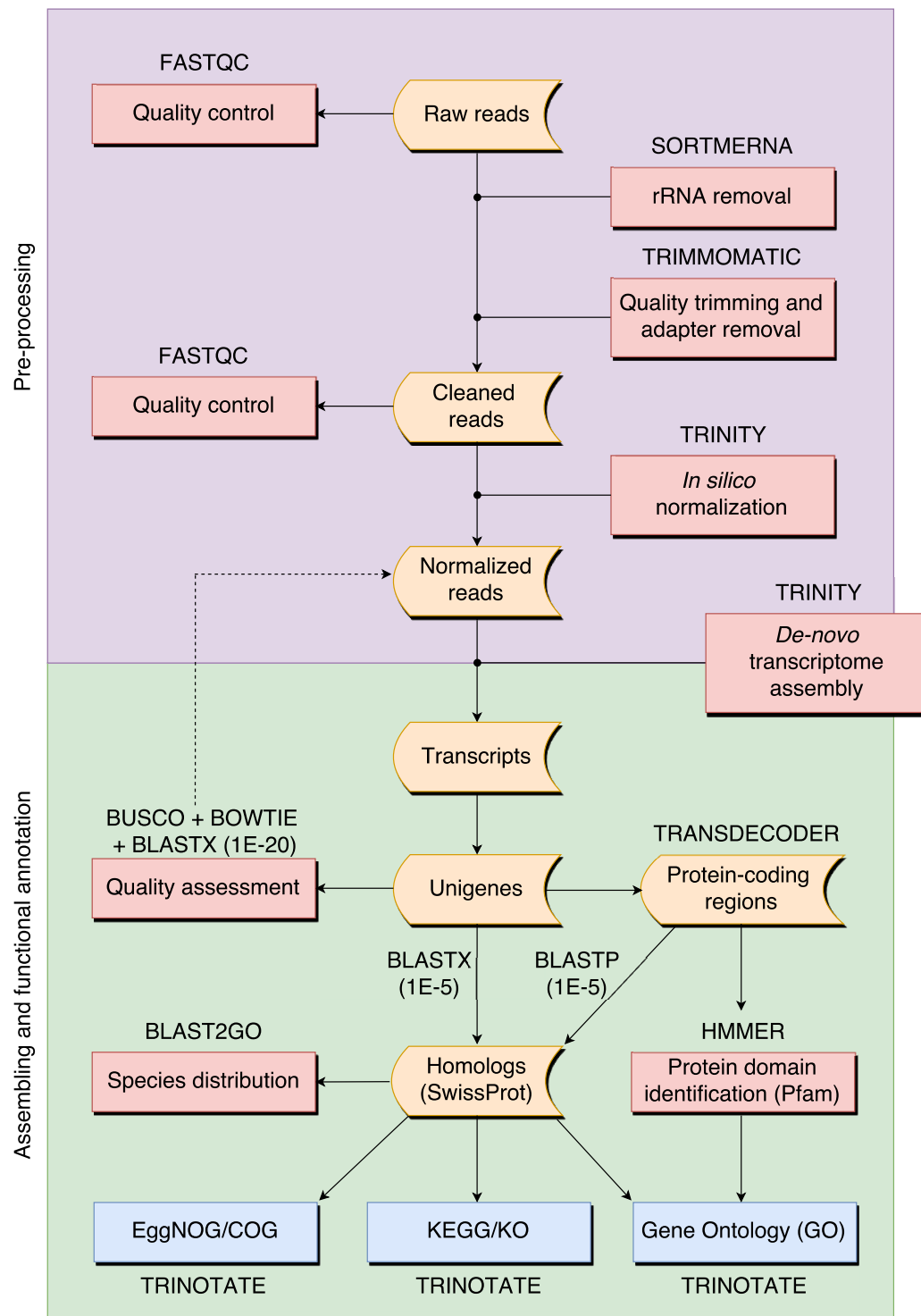
232 ep.all.fa.gz) using BLASTX (searches within a protein database using a translated nucleotide
233 query) included in the NCBI-BLAST-2.4.0+ package (Altschul *et al.* 1990). The SwissProt is a
234 curated protein sequence database aimed to provide a high level of annotation (e.g. the
235 description of the function of a protein), a minimal level of redundancy and high level of
236 integration with other databases (Bairoch & Apweiler, 200). There is no perfect E-value cut-off
237 in BLAST, but the smaller the most reliable the match. We used orthologous proteins found in
238 SwissProt and *X. tropicalis* to assess completeness as described by Haas *et al.* (2013).
239
240 We predicted protein-coding regions in the unigenes based on the most likely longest-ORF using
241 TransDecoder-v3 (Haas *et al.* 2013). In order to annotate the sequences we compared them to
242 public databases compiled for different purposes. We searched homolog sequences for the
243 predicted proteins using BLASTP (search protein database using a protein query) with E-value ≤
244 1E-5 to the SwissProt database. We also used BLASTX with E-value ≤ 1E-5 to search homolog
245 sequences for the unigenes compared to the SwissProt database. In both cases, BLASTP and
246 BLASTX, we only kept top-hit matches. We used BLAST2GO (Conesa *et al.* 2005) to detect the
247 species distribution of the top BLASTX results within the SwissProt database. We identified
248 protein domains using HMMER-3.1b2 (Finn *et al.* 2011) to the Pfam-A database
249 (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/). Homologous proteins found in the SwissProt database
250 were used to retrieve functional annotation comments from the GO (*Gene Ontology;* Ashburner
251 *et al.* 2000), EggNOG (*Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups;*
252 Powell *et al.* 2012) and KEGG (*Kyoto Encyclopedia of Genes and Genomes;* Kanehisa *et al.*
253 2012) databases using TRINOTATE-v.3 (https://trinotate.github.io). The software also searched
254 GO terms in Pfam results and in the combined results of homology search via SwissProt and
255 Pfam. At the time of conducting this study, TRINOTATE was built around specific releases of
256 SwissProt and Pfam databases (available at
257 https://data.broadinstitute.org/Trinity/Trinotate_v3_RESOURCES/). We used BLAST2GO to
258 categorize the annotated GO terms in the combined results of SwissProt and Pfam searches.
259 EggNOG annotations were filtered to keep COGs (Clusters of Orthologous Groups) and those
260 were categorized using the current version of the COG database
261 (ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data). KEGG annotations were filtered to keep KOs
262 (KEGG orthology) and those were categorized using the tool "Reconstruct Pathway"
263 (http://www.kegg.jp/kegg/tool/map_pathway.html).
264

## Data availability

266
267      Raw RNA-seq data in FASTQ format has been deposited at the NCBI Sequence Read
268 Archive database (SRA) under the accession SRP106442. The transcriptome assembly in
269 FASTA format has been deposited at DDBJ/EMBL/GenBank under the accession
270 GFNJ00000000. The quality of the assembly was examined through the NCBI contamination
271 screen. The screen found 5 sequences to exclude, 105 sequences with locations to mask/trim and
272 6 potentially duplicated sequences (with 3 distinct checksums). As a result, the uploaded
273 information contained 422,970 sequences (188,369,677 bp) rather than the initial 422,999
274 sequences (188,399,293 bp). All the data is available at NCBI BioProject under the accession
275 PRJNA384528.
276
277
278

279
280  **Figure 1**: Overall pipeline for the annotation of RNA-seq data. Boxes with curved sides represent sequence datasets. Red boxes
281  represent analyses, and the software used for those analyses is indicated outside the box. Reference databases are indicated as
282  blue boxes.
283

284

285

# Results and discussion

286

287

## RNA sequencing and transcriptome assembly

288

289

290        A summary of the RNA-seq data and transcriptome assembly is presented in Table 1.
291 Illumina RNA sequencing for three tissues of *O. cruralis* in an Illumina HiSeq2500 instrument
292 produced a total of almost $523 \times 10^6$ raw reads (intestine: $194 \times 10^6$; liver: $189 \times 10^6$; spleen: 140
293 $\times 10^6$). Of those, 81.47% were kept after the pre-processing stage ($426 \times 10^6$). The number of
294 reads was further reduced to 6.97% after *in silico* normalization prior to assembly ($36 \times 10^6$). This
295 highlights the importance of normalization to remove over-expressed transcripts in RNA-seq
296 data. A total of 550,871 transcripts were obtained after *de-novo* transcriptome assembly. This
297 large number of transcripts is not too surprising, both in terms of RNA-seq assembly as well as
298 given the species and its likely large genome (see genome size for closely related genera at:
299 http://www.genomesize.com/). First, transcriptome assemblies often include incompletely
300 spliced introns, orphaned UTRs, read through off of the 3' ends, spuriously transcribed regions,
301 active transposable elements, etc., so the number of assembled transcripts typically exceeds the
302 expected number of protein coding genes by an order of magnitude. Second, large genomes tend
303 to have large transcriptomes. In the axolotl (*Ambystoma mexicanum*) the transcriptome assembly
304 had $\sim 1.5 \times 10^6$ transcripts that clustered into $\sim 1.3 \times 10^6$ putative genes (unigenes), and of those,
305 110,000 mapped to 30,000 SwissProt genes (Bryant *et al.* 2017). It is possible that these large
306 genomes include a large number of repetitive sequences transcribed, which makes assembly
307 more difficult and results in more fragmentation, especially when using diginorm (as in
308 TRINITY) or any other in silico normalization. In *O. cruralis* the 550,871 transcripts clustered
309 into 422,999 unigenes. This difference in number is likely because of alternatively spliced
310 isoforms derived from paralogous genes (Wang *et al.* 2014). However, this will need to be
311 confirmed with new amphibian genomes as they become available. Unigenes in the
312 transcriptome of *O. cruralis* had an average GC content of 45.39%, which is very similar to other
313 amphibians, such as the axolotl (*A. mexicanum* 45.56%; Hall *et al.* 2016), the green toad (*Bufotes*
314 *viridis* 46.83%; Gerchen *et al.* 2016) or the common frog (*Rana temporaria* 44%; Price *et al.*
315 2015). The size of the unigenes in *O. cruralis* ranged from 201 to 16,804 bp with a mean length
316 of 445 bp and a N50 length of 467 bp (Table 1; Figure 2). The N50 value indicates that half of
317 the transcriptome unigenes were at least 467 bp in length. The N50 length has been proposed as
318 an estimator of genome assembly contiguity, since better assemblies will result in longer contigs
319 (Li *et al.* 2014; Simpson 2014). However, in transcriptome data this measure can be highly
320 misleading because it does not assess assembly completeness in terms of read representation or
321 gene content (Simao *et al.* 2015).

322

323

324

325

326

327

328

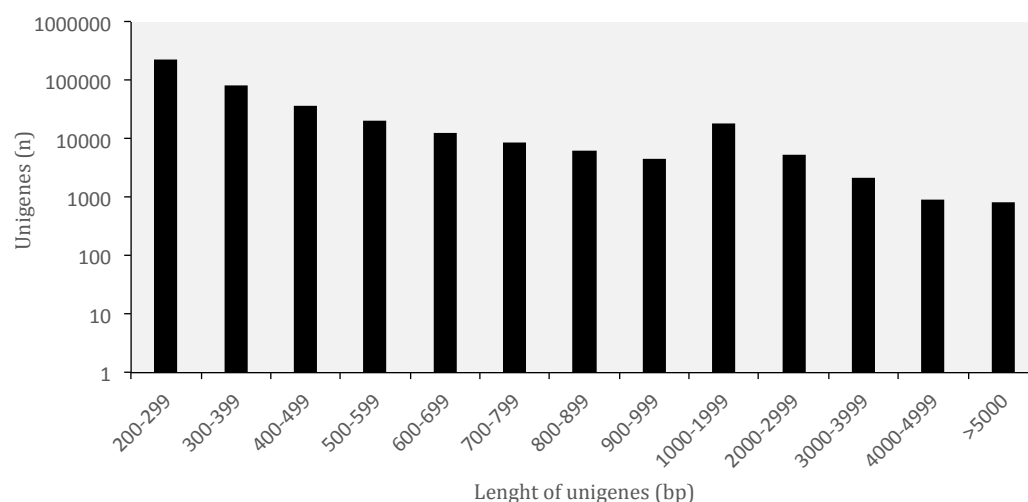
329

330

331
332
333  **Table 1**: Summary of the transcriptome data assembly for *Oreobates cruralis*.
334

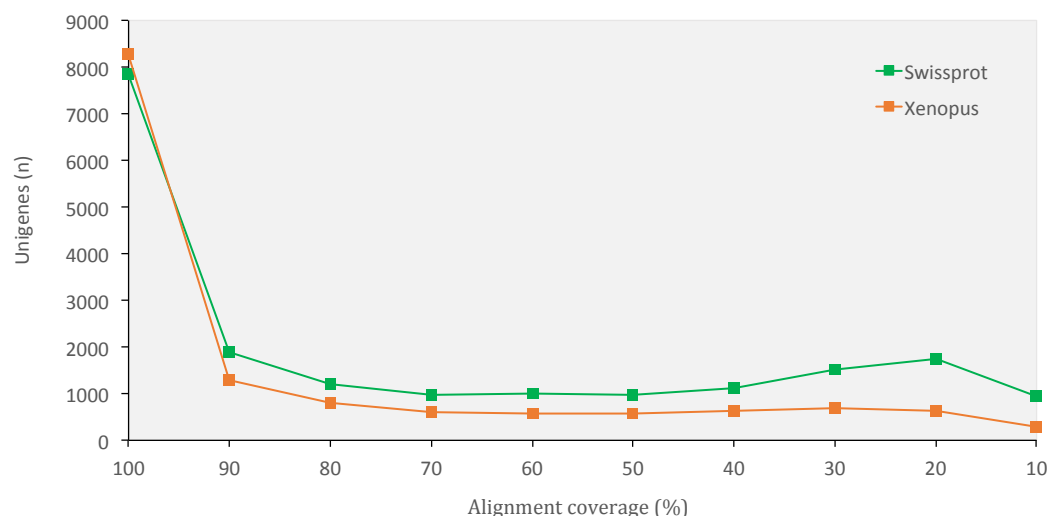
| PRIOR TO *DE-NOVO* TRANSCRIPTOME ASSEMBLY | |
|---|---|
| Length of raw reads (bp) | 125 |
| Total number of raw reads | 522,877,358 |
| Total number of clean reads | 426,003,462 |
| Total number of normalized reads | 36,428,858 |
| **AFTER *DE-NOVO* TRANSCRIPTOME ASSEMBLY** | |
| Total number of all transcripts / unigenes | 550,871 / 422,999 |
| GC-content of all transcripts / unigenes (%) | 45.88 / 45.39 |
| Total length of all transcripts / unigenes (bp) | 299,133,111 / 188,399,293 |
| N50 length of all transcripts / unigenes (bp) | 731 / 467 |
| Mean length of all transcripts / unigenes (bp) | 543 / 445 |
| Median length of all transcripts / unigenes (bp) | 309 / 290 |

335
336  **Figure 2:** Length distribution of unigenes from *Oreobates cruralis*.
337

338
339

# Transcriptome quality assessment

341
342    The set of assembled unigenes might not always perfectly correspond to all properly
343  paired reads, as some unigenes might be built from just a portion of reads coming from the same
344  transcription locus. When we evaluated assembly quality in terms of read representation, we
345  found a high rate of reads that mapped back to unigenes (75.40%), thus confirming the presence
346  of most of the initial reads in our final set of unigenes. When we evaluated the assembly
347  completeness in terms of gene content, we found 2,830 complete orthologous genes (71.65%)
348  out of the 3,950 genes available in the tetrapoda database (complete BUSCO hits). Of those,

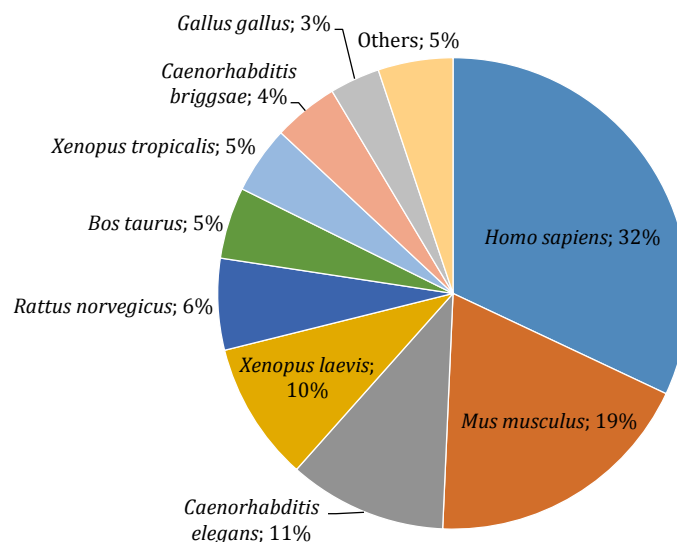
349   2,501 were single-copy genes and 329 were duplicated genes. Only 462 (11.70%) of the genes in
350   the database appeared fragmented and 658 (16.65%) were missing. We also obtained a high
351   number of orthologous proteins in both the SwissProt and the *X. tropicalis* databases that fully
352   matched (100% alignment coverage) or nearly fully corresponded (>80% alignment coverage) to
353   unigenes in *O. cruralis* (Figure 3). Altogether, the high number of complete (or nearly complete)
354   orthologous matches across the different databases provides a valuable validation of the depth
355   and completeness of the assembly process.
356
357   **Figure 3**: Distribution of BLASTX alignment coverage for *O. cruralis* unigenes against SwissProt and *Xenopus* databases. A
358   high number of orthologous proteins in the databases fully or nearly fully corresponded (>80% coverage) to unigenes in *O.*
359   *cruralis*
360

361
362

## Functional annotation of unigenes

364
365   Gene annotation consists of adding relevant biological information to coding regions of the
366   genome and it was arguably the most relevant section of our pipeline, since it allowed us to
367   describe and classify the content of the *O. cruralis* transcriptome. Functional annotation was
368   based on BLAST searches to find homologous proteins within a reference database (e.g.
369   SwissProt) and the collection of biological information from various sources (e.g. GO, KEGG,
370   EggNOG or Pfam). We predicted a total of 45,466 protein-coding genes within the 422,999
371   unigenes using TansDecoder. After homology search using BLASTP, we found that 26,418
372   protein-coding genes in *O. cruralis* mapped to proteins in the SwissProt database. Search using
373   BLASTX revealed a total of 54,425 unigenes that mapped to proteins in the *X. tropicalis*
374   proteome and 47,349 unigenes that mapped to proteins in the SwissProt database. The relative
375   low number of homologous proteins shared between *O. cruralis* and *X. tropicalis,* just 12.8% of
376   all unigenes identified in *O. cruralis,* is likely consequence of the very ancient divergence time
377   between both species (estimated to be around 204 million years ago; http://www.timetree.org/).
378   This ancient divergence implies long time for the accumulation of mutations. However, the
379   observation of a number of matches (54,425) that is larger than the total number of proteins in *X.*
380   *tropicalis* (22,718) may suggest that many of them might be duplicates or unresolved splice
381   variants among the unigenes of *O. cruralis*. The version of the SwissProt database used included

382    a selection of 553,231 protein sequences from 13,379 species, and the top-hit species distribution
383    showed that 32% (13,099) of the *O. cruralis* unigenes were homologs to human (*Homo sapiens*)
384    proteins and 19% (7661) to house mouse (*Mus musculus*) proteins (Figure 4). The larger number
385    of hits to mammals than to other amphibians is likely due to the uneven distribution of species in
386    the SwissProt database, in which the top twenty species accumulate 21.5% of the entries. Still,
387    amphibian species were highly represented in the assembly with 10% (3909) of the *O. cruralis*
388    unigenes having a highest match to *X. laevis* and 5% (1893) to *X. tropicalis* proteins. When we
389    retrieved the functional comments for the homologous proteins found in the SwissProt database,
390    the number of annotated unigenes varied depending on the source that was used: a total of
391    45,885, 38,120, 37,349 and 23,500 unigenes were annotated for GO, KEGG, EggNOG and Pfam
392    databases, respectively.

393
394    **Figure 4**: Top-hit species distribution for unigenes from the transcriptome of *O. cruralis* in the SwissProt database.
395

396
397

## Protein domain identification

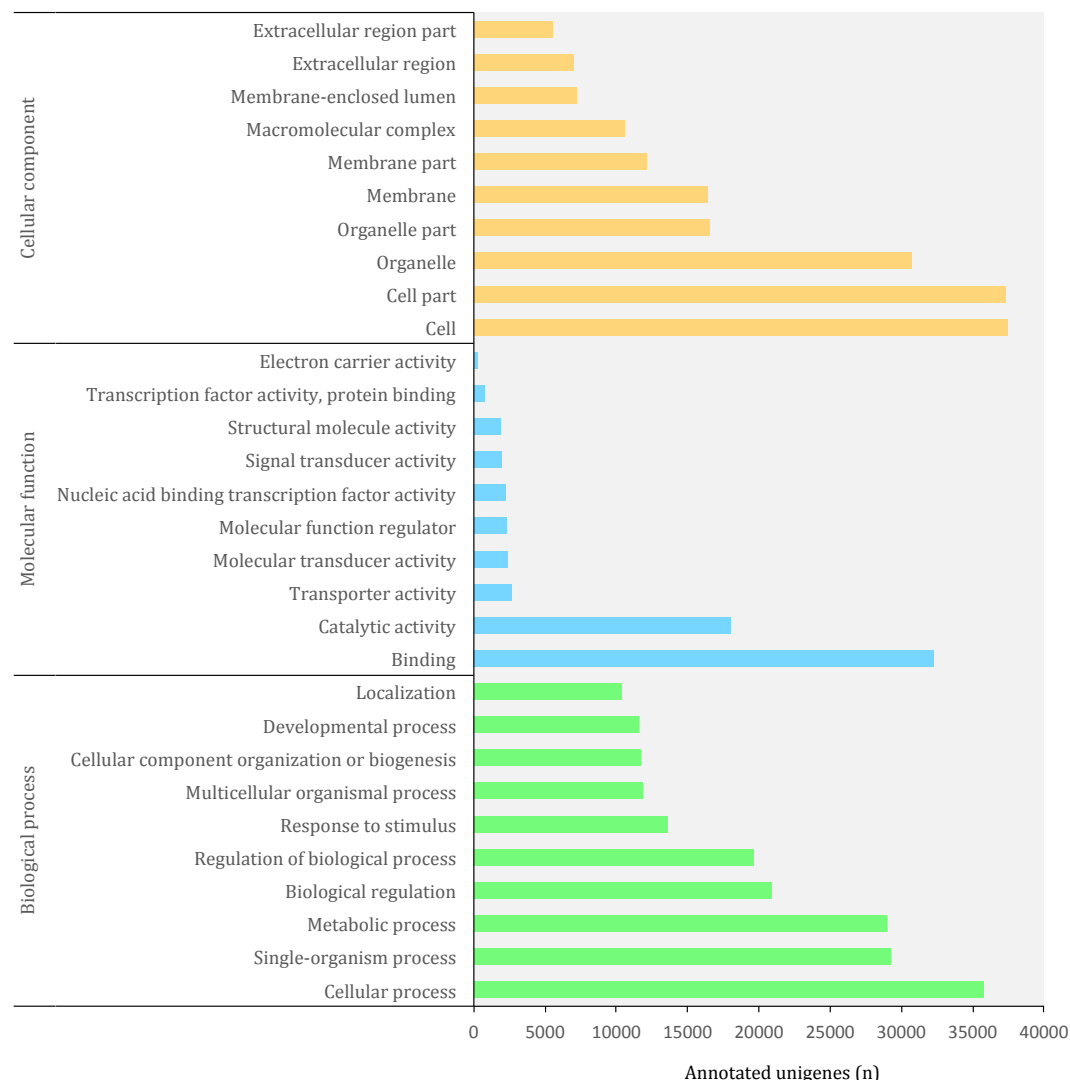
399
400         Protein domains are preserved portions of proteins with tertiary structure that can act,
401    evolve and exist independently of the rest of the protein chain (Jacob 1977). Prediction of protein
402    domains is an important step of transcriptome annotation since they provide insights in specific
403    cellular functions that assist comparative genomics of domain families across species (Ochoa *et*
404    *al.* 2011). The Pfam database is a large collection of protein families that currently contains
405    16,303 families (Pfam v30.0). From the predicted 45,466 protein-coding genes in the
406    transcriptome of *O. cruralis*, we identified 23,500 that are present in the Pfam-A database,
407    consisting of 5,686 protein domain families. We found that the most common Pfam domain in
408    the transcriptome of *O. cruralis* is the 'Zinc finger, C2H2 type' (961 hits; 4.09%). The C2H2
409    zinc finger proteins are very frequent in eukaryotic genomes (e.g. the human genome has 564
410    C2H2 zinc fingers; Tadepally *et al.* 2008), and their functions are extraordinarily diverse,
411    including DNA recognition, RNA packaging, transcriptional activation, regulation of apoptosis,
412    protein folding and assembly, and lipid binding (Laity *et al.* 2001). Interestingly, this protein
413    family was also reported as the most common for other amphibians, such as the green frog

414 (*Lithobates clamitans*) and the Pacific tree frog (*Pseudacris regilla*) (Robertson & Cornman
415 2014).
416
417 The 'WD domain, G-beta repeat' was the second most common Pfam domain in *O. cruralis*
418 transcriptome (840 hits; 3.57%). The G protein family is involved in signal transduction from
419 outside a cell to its interior (Umbarger *et al.* 1992), and in frog oocytes they are important
420 regulating the maturation process (Kalinowski *et al.* 2003). Another essential domain for frogs is
421 the 'Protein kinase domain' that we found as the third more abundant (643 hits; 2.74%). This
422 domain is supposed to play an important role in frogs in freezing tolerance during cold winters,
423 likely inducing the transcription of antioxidant response genes (Dieni & Storey 2014). Although
424 freezing winters are not common within the current range of *O. cruralis*, the relative abundance
425 of protein kinase domains could have been important in the evolutionary history of *Oreobates*, a
426 genus that may have originated at high altitude in the Andes (Padial *et al.* 2008). It is also
427 remarkably the high number of immunoglobulin-related domains found within the top 10 Pfam
428 domains in the transcriptome of *O. cruralis* (1,066 hits; 4.54%) (Table 2). Immunoglobulin
429 domains are involved in a wide range of functions, including cell-cell recognition, cell-surface
430 receptors, muscle structure and immune system function (Isenman *et al.* 1975). In frogs, as in the
431 Yunnan firebelly toad (*Bombina maxima*) (Zhao *et al.* 2014), these domains are essential for the
432 regulation of immune responses, allowing them to survive in harsh environmental conditions. It
433 is possible that tropical rainforests could host a large diversity of potential pathogens imposing a
434 positive selection on immunoglobulin-related domains in *Oreobates* frogs, but this hypothesis
435 remains to be tested.
436
437 **Table 2**: Top 10 Pfam domains identified in the transcriptome of *O. cruralis*.
438

| No | Pfam domain | Pfam ID | N-hits |
|----|-------------|---------|--------|
| 1 | Zinc finger, C2H2 type | PF00096.23 | 961 |
| 2 | WD domain, G-beta repeat | PF00400.29 | 840 |
| 3 | Protein kinase domain | PF00069.22 | 643 |
| 4 | Protein tyrosine kinase | PF07714.14 | 608 |
| 5 | C2H2-type zinc finger | PF13912.3 | 593 |
| 6 | C2H2-type zinc finger | PF13894.3 | 570 |
| 7 | Ankyrin repeat | PF00023.27 | 553 |
| 8 | Immunoglobulin I-set domain | PF07679.13 | 549 |
| 9 | Immunoglobulin domain | PF00047.22 | 517 |
| 10 | Leucine rich repeat | PF13855.3 | 482 |

439
440
441
442
443
444
445
446
447
448
449
450
451

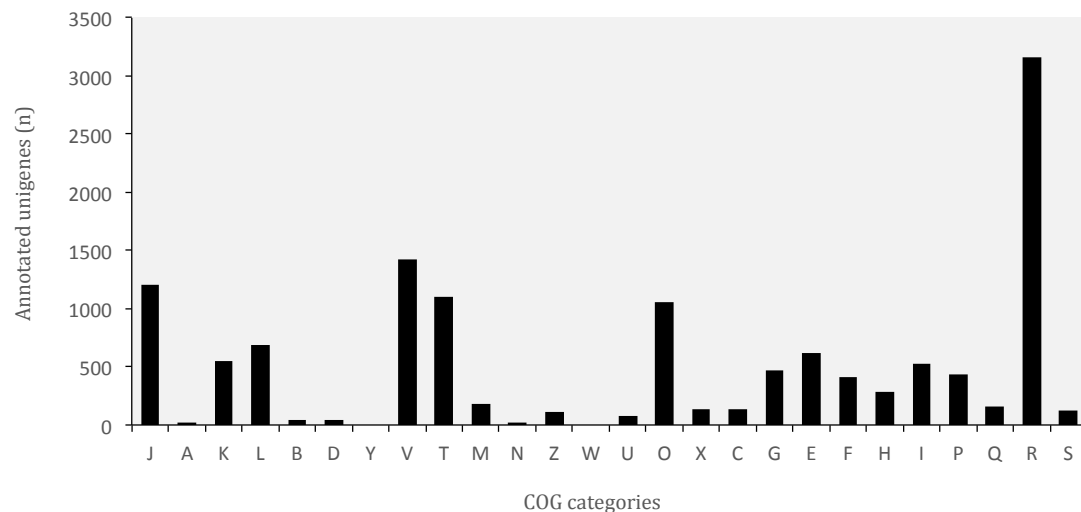
452 **Figure 5**: Distribution of top-10 gene ontology GO terms in the transcriptome of *O cruralis* identified by homology with the
453 databases via SwissProt and Pfam. Categories shown correspond to gene ontology level 2.
454

455
456

## Gene ontology

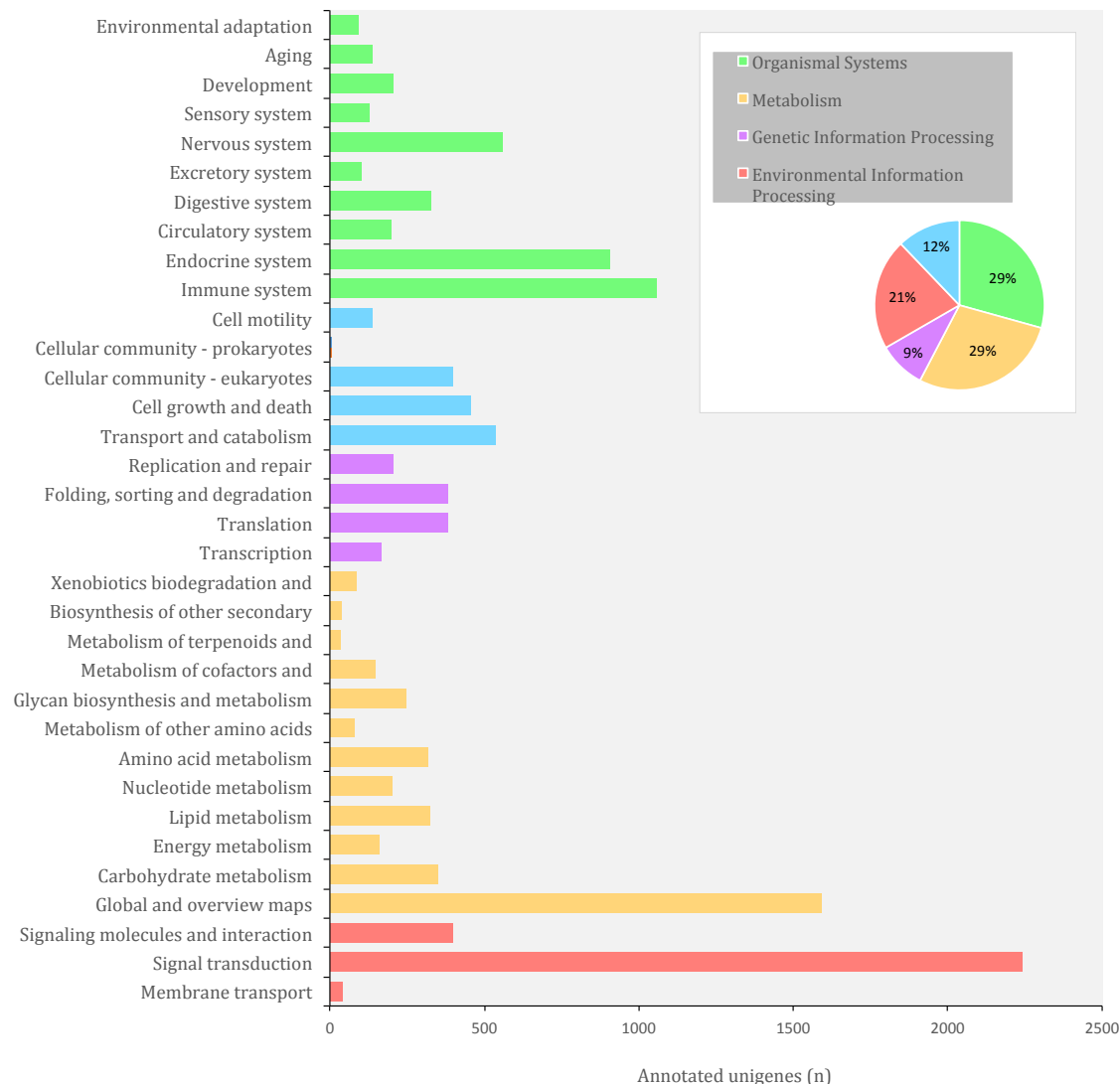
458
459 The Gene Ontology (GO) (http://geneontology.org/) is a standardized functional
460 classification system aimed to describe gene and gene product attributes across species, using a
461 controlled vocabulary (i.e. ontology terms). The GO classification comprises three domains:
462 cellular component, molecular function and biological process. These domains have a
463 hierarchical structure and a GO term can belong to different levels depending on the path
464 followed and the number of steps between the term and the root (Ashburner *et al*. 2000). Using
465 the combined results of a homology search via SwissProt and Pfam, we detected a total of
466 3,094,863 GO terms (19,407 unique) corresponding to 45,885 (10.85%) unigenes in the
467 transcriptome of *O. cruralis*. This contrasts previous studies that have reported that between 50
468 and 80% of the transcripts reconstructed from RNA-seq data can be annotated with GO terms

469  (Conesa *et al.* 2016). However, the relatively low percentage of annotation may reflect the
470  scarcity of amphibian sequences in the GO database, and therefore the presence of undetected
471  novel transcripts. Still, the GO database produced the highest number of annotated unigenes
472  compared to other sources, such as Pfam, KEGG or EggNOG (Figure 5). The largest number of
473  GO terms corresponded to the category of "Biological Process" (49%) followed by "Cellular
474  Component" (38%) and "Molecular Function" (13%). At ontology level-2, which represents the
475  second most general category in the GO database, there were 65 different GO terms (Figure 6).
476  Within the "Biological Process"  category, the most frequent GO terms were "cellular process"
477  (35,730) and "single-organism process" (29,237). Within the "Molecular Function" category,
478  unigenes were mainly associated to "binding" (32,275) and "catalytic activity" (18,023). Within
479  the "Cellular Component" category, unigenes were mostly associated with "cell" (37,424) and
480  "cell part" (37,293). These highly abundant GO terms are likely associated to genes involved in
481  essential cell functions and metabolism regulation, since they describe very general terms. A
482  similar distribution of GO terms was found in a comparative transcriptome study of seven anuran
483  species (Huang *et al.* 2016). We found 185 unigenes with antioxidant activity, most of them with
484  peroxidase activity (128). This number is relatively high compared to the 63 antioxidant genes
485  present in humans (Gelain *et al.* 2009) and it might be related to the high number of protein
486  kinase domains that we recorded earlier, as well as to the habitat of *O. cruralis*. Specimens are
487  usually encountered in tropical rainforest leaf litter, where amphibian pathogens are common
488  (Pounds *et al.* 2006). Antioxidant genes have previously been reported from the skin of
489  amphibians, contributing to resistance against microorganism infection or radiation injury (Yang
490  *et al.* 2009). However, since the transcriptome of *O. cruralis* was built from tissues of intestine,
491  liver and spleen, our results suggest that antioxidant genes in amphibians can also be expressed
492  in different tissues besides skin. Because *O. cruralis* is mainly a lowland Amazonian rainforests
493  frog, it would be interesting to compare this results with closely-related species living in higher
494  altitudes (e.g. *Oreobates ayacucho*), where temperature is lower and microbial activity too.
495  
496  **COG classification**
497  
498  The database of Clusters of Orthologous Groups (COGs) is another common tool for
499  functional annotation (Galperin *et al.* 2015). In this database, orthologous genes from 722
500  prokaryote genomes are grouped according to their biological function. The current version
501  consists of 4,632 COGs classified into 26 functional categories. The EggNOG database is based
502  on the original idea of COGs and expands it to non-supervised orthologous groups from
503  numerous organisms, including eukaryotes and viruses (Huerta-Cepas *et al.* 2016). We identified
504  a total of 37,349 (8.83%) unigenes that are present in the EggNOG database (Figure 5). Of these,
505  12,993 belonged to the COG database, corresponding to 24 functional categories (Figure 7). The
506  "general function" category (3,166; 24.37%) represented the largest group, followed by "defense
507  mechanisms" (1,421; 10.94%). Our results showed that genes related to defense functions may
508  be relatively abundant in the transcriptome of *O. cruralis*, particularly compared to the seven
509  anurans studied by Huang *et al.* (2016) and also to *A. mexicanum* (Wu *et al.* 2013). In both
510  studies, only about 2% of unigenes corresponded to defense mechanisms. Within the unigenes
511  involved in defense mechanisms, we identified 1,163 (81.84%) that are related to Cytochrome
512  P450 enzymes (CYPs), while only 57 of those genes have been found in humans (Zanger &
513  Schwab 2013). CYPs are a protein superfamily in charge of metabolizing potentially toxic
514  compounds, such as drugs or products of endogenous metabolism (Fujita *et al.* 2004). This large

515    difference in the number of genes in humans and *O. cruralis* may indicate the presence of
516    duplicates in our data, but it could also be associated with some degree of myrmecophagy
517    (feeding on ants) in this group of frogs. Because the eating habits of *Oreobates* frogs have not
518    been studied yet, protein data from strict myrmecophagous species (e.g. poison dart frogs in the
519    family Dendrobatidae) are needed to confirm these results.
520

521    **Figure 6**: Distribution of Clusters of Orthologous Groups (COG) categories in the transcriptome of *O cruralis*.
522

523
524    J: Translation, ribosomal structure and biogenesis. A: RNA processing and modification. K: Transcription. L: Replication,
525    recombination and repair. B: Chromatin structure and dynamics. D: Cell cycle control, cell division, chromosome partitioning. Y:
526    Nuclear structure. V: Defense mechanisms. T: Signal transduction mechanisms. M: Cell wall/membrane/envelope biogenesis. N:
527    Cell motility. Z: Cytoskeleton. W: Extracellular structures. U: Intracellular trafficking, secretion, and vesicular transport. O:
528    Posttranslational modification, protein turnover, chaperones. X: Mobilome: prophages, transposons. C: Energy production and
529    conversion. G: Carbohydrate transport and metabolism. E: Amino acid transport and metabolism. F: Nucleotide transport and
530    metabolism. H: Coenzyme transport and metabolism. I: Lipid transport and metabolism. P: Inorganic ion transport and
531    metabolism. Q: Secondary metabolites biosynthesis, transport and catabolism. R: General function prediction only. S: Function
532    unknown.

533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550

551 **Figure 7**: Distribution of KEGG Orthology (KO) categories in the transcriptome of *O cruralis*.
552

553
554
## KEGG pathways
556
557     In the KEGG (Kyoto Encyclopedia of Genes and Genomes) database, genes from
558     completely sequenced genomes are linked to higher-level systemic functions of the cell, the
559     organism and the ecosystem (Kanehisa & Goto 2000). Molecular-level functions are stored in the
560     KO (KEGG Orthology) database, where each KO is defined as a functional ortholog of genes
561     and gene products (Kanehisa *et al.* 2016). We identified a total of 38,120 (9.01%) unigenes from
562     *O. cruralis* in the KEGG database (Figure 5). Of these, 25,619 unigenes have orthologs in the
563     KO database. Many unigenes were classified under the category of organismal systems (3704;
564     29.32%), followed by metabolism (3580; 28.34%), environmental information processing (2678;
565     21.20%), cellular processes (1535; 12.15%) and genetic information processing (1135; 8.99%)
566     (Figure 7). We found the largest number of unigenes to be related with signal transduction
567     (2241) within the category of environmental information processing. Particularly, the PI3K-Akt

568  signaling pathway was the most frequent (184; 8.21%) among the signal transduction unigenes,
569  followed by the MAPK signaling pathway (152; 6.78%). Both the PI3K-Akt and the MAPK
570  signaling pathways play a major role in the development of immune cells (Liu *et al.* 2007;
571  Juntilla & Koretzky 2008). Interestingly, the immune system category was also highly enriched
572  (1057 unigenes) and within the immune category, the chemokine signaling pathway comprised
573  the highest number of unigenes (105; 9.93%). Chemokine receptors associate with G proteins to
574  promote signaling cascades, including MAPK pathways, that cause immune responses such as
575  degranulation, a cellular process that releases antimicrobial cytotoxic molecules to destroy
576  invading microorganisms (Murdoch & Finn 2000). This suggests that, compared to other genes,
577  those related to the immune system are relatively abundant in the transcriptome of *O. cruralis*.
578  We hypothesize that tropical conditions, in which high temperature and humidity are constant
579  throughout the year, impose a crucial challenge to amphibian fitness. Although based on a single
580  transcriptome our results lack of statistical power, this study provides a first view towards the
581  understanding of gene evolution in Neotropical amphibians.

582

## Conclusions

584

585      Although large genome size renders complete genome sequencing practically unfeasible
586  in many species, such as most amphibians, transcriptome sequencing represents a cost-effective
587  alternative to obtain a large amount of genome-wide data. This can allow the study of selection
588  and adaptation in natural populations, but it will also lead to advances in the study of ecological
589  and evolutionary processes beyond the limits imposed by the use of small panels of markers. In
590  this study, we have provided and discussed a pipeline that covers the basic elements needed to
591  build a *de-novo* transcriptome from RNA-seq data of non-model organisms for which sequencing
592  and assembling a genome is not a practical option. We have successfully applied this pipeline to
593  obtain the transcriptome profile of *Oreobates cruralis*, a poorly known Neotropical frog. The
594  data obtained here has some limitations: the specimen was unsexed, and only three tissues and
595  one life stage were represented. Thus, the results should be taken with caution in the context of
596  sex-specific gene expression. Nevertheless, this is the first transcriptome data available for a
597  South American amphibian, and therefore, a stepping-stone towards the study of the
598  diversification patterns across this group of vertebrates using genomic approaches. Once a
599  reference transcriptome is available, capture-based approaches can help to obtain homologous
600  sequences for a large array of closely related species at a reduced cost. In this regard, this
601  transcriptome will serve as a valuable resource for the inference of orthologous sequences in
602  closely related species. This, for example, will allow solving phylogenomic relationships among
603  the species of the genus *Oreobates*, as well as studying population differentiation, demographic
604  history and gene evolution for the different species.

605

## Acknowledgements

607

608      The tissue samples used for this study were provided by the frozen tissue collection of the
609  Museo Nacional de Ciencias Naturales (MNCN-CSIC) in Madrid, Spain (MNCN/ADN
610  collection). Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala,
611  Sweden. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science
612  for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council
613  and the Knut and Alice Wallenberg Foundation. Computations were performed on resources

614 provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational
615 Science (UPPMAX) under Project b2015409. We thank all the members of the Conservation and
616 Evolutionary Genetics Group, as well as Dr. José Manuel Padial for constructive comments and
617 support in the study. We also thank Anna Olsson for laboratory support.
618

# References

620

621 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search
622      tool. *Journal of molecular biology*, **215**, 403–10.

623 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K,
624      Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese
625      JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the
626      unification of biology. *Nature genetics*, **25**, 25–29.

627 Bairoch, A., & Apweiler, R. (2000) The SWISS-PROT protein sequence database and its
628      supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48.

629 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence
630      data. *Bioinformatics*, **30**, 2114–2120.

631 Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, Lee TJ, Leigh
632      ND, Kuo TH, Davis FG, Bateman J, Bryant S, Guzikowski AR, Tsai SL, Coyne S, Ye WW,
633      Freeman RM Jr, Peshkin L, Tabin CJ, Regev A, Haas BJ, Whited JL (2017) A Tissue-
634      Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration
635      Factors. *Cell Reports*, **18**, 762–776.

636 Camacho-Sanchez M, Burraco P, Gomez-Mestre I, Leonard JA (2013) Preservation of RNA and
637      DNA from mammal samples under field conditions. *Molecular Ecology Resources*, **13**,
638      663–673.

639 Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: A
640      universal tool for annotation, visualization and analysis in functional genomics research.
641      *Bioinformatics*, **21**, 3674–3676.

642 Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak
643      MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A (2016) A survey of best practices for
644      RNA-seq data analysis. *Genome Biology*, **17**, 13.

645 Destro-Bisol G, Jobling MA, Rocha J, Novembre J, Richards MB, Mulligan C, Batini C, Manni
646      F (2010) Molecular Anthropology in the genomic era. In: *Journal of Anthropological*
647      *Sciences*, pp. 93–112.

648 Dieni CA, Storey KB (2014) Protein kinase C in the wood frog, *Rana sylvatica*: reassessing the
649      tissue-specific regulation of PKC isozymes during freezing. *PeerJ*, **2**, e558.

650 Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity
651      searching. *Nucleic Acids Research*, **39**, W29–W37.

652 Fujita Y, Ohi H, Murayama N, Saguchi K, Higuchi S (2004) Identification of multiple
653      cytochrome P450 genes belonging to the CYP4 family in *Xenopus laevis*: cDNA cloning of
654      CYP4F42 and CYP4V4. *Comparative Biochemistry and Physiology Part B: Biochemistry*
655      *and Molecular Biology*, **138**, 129–136.

656 Galperin MY, Makarova KS, Wolf YI, Koonin E V (2015) Expanded microbial genome
657      coverage and improved protein family annotation in the COG database. *Nucleic acids*
658      *research*, **43**, D261-9.

659 Gelain DP, Dalmolin RJ, Belau VL, Moreira JC, Klamt F, Castro MA (2009) A systematic

660   review of human antioxidant genes. *Frontiers in bioscience (Landmark edition)*, **14**, 4457–
661       63.
662   Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann M,
663       Birol I, Jones SJ, Cronk QC, Douglas CJ (2011) SNP discovery in black cottonwood
664       (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology*
665       *Resources*, **11**, 81–92.
666   Gerchen JF, Reichert SJ, Röhr JT, Dieterich C, Kloas W, Stöck M (2016) A single transcriptome
667       of a green toad (*Bufo viridis*) yields candidate genes for sex determination and -
668       differentiation and non-anonymous population genetic markers. *PLoS ONE*, **11**, 1–14.
669   Giallourakis C, Henson C, Reich M, Xie X, Mootha VK (2005) Disease gene discovery through
670       integrative genomics. *Annual review of genomics and human genetics*, **6**, 381–406.
671   Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
672       Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F,
673       Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length
674       transcriptome assembly from RNA-Seq data without a reference genome. *Nature*
675       *biotechnology*, **29**, 644–52.
676   Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF,
677       Greilhuber J, Bennett MD (2007) Eukaryotic genome size databases. *Nucleic Acids*
678       *Research*, **35**.
679   Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D,
680       Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman
681       R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A (2013) De novo
682       transcript sequence reconstruction from RNA-seq using the Trinity platform for reference
683       generation and analysis. *Nature Protocols*, **8**, 1494–1512.
684   Hall KW, Eisthen HL, Williams BL (2016) Proteinaceous pheromone homologs identified from
685       the cloacal gland transcriptome of a male axolotl, *Ambystoma mexicanum*. *PLoS ONE*, **11**,
686       1–18.
687   Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam
688       NH, Shu S, Taher L, Blitz IL, Blumberg B, Dichmann DS, Dubchak I, Amaya E, Detter JC,
689       Fletcher R, Gerhard DS, Goodstein D, Graves T, Grigoriev IV, Grimwood J, Kawashima T,
690       Lindquist E, Lucas SM, Mead PE, Mitros T, Ogino H, Ohta Y, Poliakov AV, Pollet N,
691       Robert J, Salamov A, Sater AK, Schmutz J, Terry A, Vize PD, Warren WC, Wells D, Wills
692       A, Wilson RK, Zimmerman LB, Zorn AM, Grainger R, Grammer T, Khokha MK,
693       Richardson PM, Rokhsar DS (2010) The genome of the Western clawed frog *Xenopus*
694       *tropicalis*. *Science*, **328**, 633–6.
695   Huang L, Li J, Anboukaria H, Luo Z, Zhao M, Wu H (2016) Comparative transcriptome
696       analyses of seven anurans reveal functions and adaptations of amphibian skin. *Scientific*
697       *Reports*, **6**, 24069.
698   Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR,
699       Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P (2016) eggNOG 4.5: a hierarchical
700       orthology framework with improved functional annotations for eukaryotic, prokaryotic and
701       viral sequences. *Nucleic Acids Research*, **44**, D286–D293.
702   Isenman DE, Painter RH, Dorrington KJ (1975) The structure and function of immunoglobulin
703       domains: studies with beta-2-microglobulin on the role of the intrachain disulfide bond.
704       *Proceedings of the National Academy of Sciences of the United States of America*, **72**, 548–
705       52.

706     Jacob F (1977) Evolution and tinkering. *Science*, **196**, 1161–1166.

707     Juntilla MM, Koretzky GA (2008) Critical roles of the PI3K/Akt signaling pathway in T cell
708          development. *Immunology letters*, **116**, 104–10.

709     Kalinowski RR, Jaffe LA, Foltz KR, Giusti AF (2003) A receptor linked to a Gi-family G-
710          protein functions in initiating oocyte maturation in starfish but not frogs. *Developmental*
711          *Biology*, **253**, 139–149.

712     Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2016) KEGG: new perspectives on
713          genomes, pathways, diseases and drugs. *Nucleic Acids Research*, **45**, D353–D361.

714     Kanehisa M, Goto S (2000) Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids*
715          *Research*, **28**, 27–30.

716     Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and
717          interpretation of large-scale molecular data sets. *Nucleic Acids Research*, **40**, D109–D114.

718     Köhler J, Padial JM (2016) Description and phylogenetic position of a new (singleton) species of
719          Oreobates Jiménez De La Espada, 1872 (Anura: Craugastoridae) from the yungas of
720          Cochabamba, Bolivia. *Annals of Carnegie Museum*, **84**, 23–38.

721     Kopylova E, Noé L, Touzet H (2012) SortMeRNA: Fast and accurate filtering of ribosomal
722          RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.

723     Kornobis E, Cabellos L, Aguilar F, Frías-López C, Rozas J, Marco J, Zardoya R (2015)
724          TRUFA : A User-Friendly Web Server for de novo RNA-seq Analysis Using Cluster
725          Computing. *Evolutionary bioinformatics online*, **11**, 97–104.

726     De la Riva I, Köhler J, Lötters S, Reichle S (2000) Ten years of research on Bolivian
727          amphibians: updated checklist, distribution, taxonomic problems, literature and
728          iconography. *Revista Española de Herpetología*, **14**, 19–164.

729     Laity JH, Lee BM, Wright PE (2001) Zinc finger proteins: new insights into structural and
730          functional diversity. *Current opinion in structural biology*, **11**, 39–46.

731     Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A,
732          Promerová M, Rubin CJ, Wang C, Zamani N, Grant BR, Grant PR, Webster MT,
733          Andersson L (2015) Evolution of Darwin's finches and their beaks revealed by genome
734          sequencing. *Nature*, **518**, 371–375.

735     Lamichhaney S, Han F, Berglund J, Wang C, Almén MS, Webster MT, Grant BR, Grant PR,
736          Andersson L (2016) A beak size locus in Darwin´s finches facilitates character
737          displacement during a drought. *Science*, **352**, 470–474.

738     Lamichhaney S, Martinez Barrio A, Rafati N, Sundström G, Rubin CJ, Gilbert ER, Berglund J,
739          Wetterbom A, Laikre L, Webster MT, Grabherr M, Ryman N, Andersson L (2012)
740          Population-scale sequencing reveals genetic differentiation due to local adaptation in
741          Atlantic herring. *Proceedings of the National Academy of Sciences*, **109**, 19345–50.

742     Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle
743          M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J,
744          LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor
745          J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y,
746          Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S,
747          Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A,
748          Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A,
749          Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A,
750          Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW,
751          McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR,

752      Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL,
753      Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P,
754      Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C,
755      Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ,
756      Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL,
757      Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C,
758      Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F,
759      Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L,
760      Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G,
761      Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A,
762      Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson
763      M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K,
764      Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J,
765      Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H,
766      Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S,
767      Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley
768      RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C,
769      Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA,
770      Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D,
771      Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP,
772      Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D,
773      Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang
774      SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A,
775      Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ,
776      Szustakowki J; International Human Genome Sequencing Consortium (2001) Initial
777      sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
778 Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment
779      of short DNA sequences to the human genome. *Genome biology*, **10**, R25.
780 Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN (2014) Evaluation of de
781      novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, **15**, 553.
782 Liu Y, Shepherd EG, Nelin LD (2007) MAPK phosphatases — regulating the immune response.
783      *Nature Reviews Immunology*, **7**, 202–212.
784 Martin J a., Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*,
785      **12**, 671–682.
786 McBride CM, Bowen D, Brody LC, Condit CM, Croyle RT, Gwinn M, Khoury MJ, Koehly LM,
787      Korf BR, Marteau TM, McLeroy K, Patrick K, Valente TW (2010) Future Health
788      Applications of Genomics. Priorities for Communication, Behavioral, and Social Sciences
789      Research. *American Journal of Preventive Medicine*, **38**, 556–565.
790 Mcmahon BJ, Teeling EC, Höglund J (2014) How and why should we implement genomics into
791      conservation? *Evolutionary Applications*, **7**, 999–1007.
792 Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, Thomas AD,
793      Ali R, Sharma K, Kyrpides NC, Reddy TB (2017) Genomes OnLine Database (GOLD) v.6:
794      data updates and feature enhancements. *Nucleic acids research*, **45**, D446–D456.
795 Murdoch C, Finn A (2000) Chemokine receptors and their role in inflammation and infectious
796      diseases. *Blood*, **95**, 3032–43.
797 Ochoa A, Llinás M, Singh M (2011) Using context to improve protein domain identification.

798      *BMC Bioinformatics*, **12**, 90.
799   Padial JM, Chaparro JC, De la Riva I (2008) Systematics of Oreobates and the Eleutherodactylus
800      discoidalis species group (Amphibia, Anura), based on two mitochondrial DNA genes and
801      external morphology. *Zoological Journal of the Linnean Society*, **152**, 737–773.
802   Pounds JA, Bustamante MR, Coloma L, Consuegra JA, Fogden MP, Foster PN, La Marca E,
803      Masters KL, Merino-Viteri A, Puschendorf R, Ron SR, Sánchez-Azofeifa GA, Still CJ,
804      Young BE (2006) Widespread amphibian extinctions from epidemic disease driven by
805      global warming. *Nature*, **439**, 161–7.
806   Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I,
807      Doerks T, Jensen LJ, von Mering C, Bork P (2012) eggNOG v3.0: Orthologous groups
808      covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research*, **40**,
809      D284-9.
810   Price SJ, Garner TWJ, Balloux F, Ruis C, Paszkiewicz KH, Moore K, Griffiths AG (2015) A de
811      novo assembly of the common frog (*Rana temporaria*) transcriptome and comparison of
812      transcription following exposure to Ranavirus and Batrachochytrium dendrobatidis. *PLoS*
813      *ONE*, **10**, 1–23.
814   Reuter JA, Spacek D V., Snyder MP (2015) High-Throughput Sequencing Technologies.
815      *Molecular Cell*, **58**, 586–597.
816   Robertson LS, Cornman RS (2014) Transcriptome resources for the frogs *Lithobates clamitans*
817      and *Pseudacris regilla*, emphasizing antimicrobial peptides and conserved loci for
818      phylogenetics. *Molecular Ecology Resources*, **14**, 178–183.
819   Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nature Methods*, **5**,
820      16–18.
821   Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A,
822      Suzuki A, Kondo M, van Heeringen SJ, Quigley I, Heinz S, Ogino H, Ochi H, Hellsten U,
823      Lyons JB, Simakov O, Putnam N, Stites J, Kuroki Y, Tanaka T, Michiue T, Watanabe M,
824      Bogdanovic O, Lister R, Georgiou G, Paranjpe SS, van Kruijsbergen I, Shu S, Carlson J,
825      Kinoshita T, Ohta Y, Mawaribuchi S, Jenkins J, Grimwood J, Schmutz J, Mitros T,
826      Mozaffari SV, Suzuki Y, Haramoto Y, Yamamoto TS, Takagi C, Heald R, Miller K,
827      Haudenschild C, Kitzman J, Nakayama T, Izutsu Y, Robert J, Fortriede J, Burns K, Lotay
828      V, Karimi K, Yasuoka Y, Dichmann DS, Flajnik MF, Houston DW, Shendure J,
829      DuPasquier L, Vize PD, Zorn AM, Ito M, Marcotte EM, Wallingford JB, Ito Y, Asashima
830      M, Ueno N, Matsuda Y, Veenstra GJ, Fujiyama A, Harland RM, Taira M, Rokhsar DS
831      (2016) Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, **538**, 1–15.
832   Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM (2015) BUSCO:
833      Assessing genome assembly and annotation completeness with single-copy orthologs.
834      *Bioinformatics*, **31**, 3210–3212.
835   Simpson JT (2014) Exploring genome characteristics and sequence quality without a reference.
836      *Bioinformatics*, **30**, 1228–1235.
837   Sun Y-B, Xiong Z-J, Xiang X-Y, Liu SP, Zhou WW, Tu XL, Zhong L, Wang L, Wu DD, Zhang
838      BL, Zhu CL, Yang MM, Chen HM, Li F, Zhou L, Feng SH, Huang C, Zhang GJ, Irwin D,
839      Hillis DM, Murphy RW, Yang HM, Che J, Wang J, Zhang YP (2015) Whole-genome
840      sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod
841      genomes. *Proceedings of the National Academy of Sciences of the United States of America*,
842      **112**, E1257-62.
843   Tadepally HD, Burger G, Aubry M (2008) Evolution of C2H2-zinc finger genes and subfamilies

844    in mammals: Species-specific duplication and loss of clusters, genes and effector domains.
845    *BMC Evolutionary Biology*, **8**, 176.
846    Umbarger KO, Yamazaki M, Hutson LD, Hayashi F, Yamazaki A (1992) Heterogeneity of the
847    retinal G-protein transducin from frog rod photoreceptors: Biochemical identification and
848    characterization of new subunits. *Journal of Biological Chemistry*, **267**, 19494–19502.
849    Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics.
850    *Nature reviews. Genetics*, **10**, 57–63.
851    Wang W, Wang J, You F, Ma L, Yang X, Gao J, He Y, Qi J, Yu H, Wang Z, Wang X, Wu Z,
852    Zhang Q (2014) Detection of alternative splice and gene duplication by RNA sequencing in
853    Japanese flounder, *Paralichthys olivaceus*. *G3 (Bethesda, Md.)*, **4**, 2419–24.
854    De Wit P, Pespeni MH, Palumbi SR (2015) SNP genotyping and population genomics from
855    expressed sequences - Current advances and future possibilities. *Molecular Ecology*, **24**,
856    2310–2323.
857    Wolfe KH (2006) Comparative genomics and genome evolution in yeasts. *Philosophical
858    transactions of the Royal Society of London. Series B, Biological sciences*, **361**, 403–412.
859    Wu C-H, Tsai M-H, Ho C-C, Chen C-Y, Lee H-S (2013) De novo transcriptome sequencing of
860    axolotl blastema for identification of differentially expressed genes during limb
861    regeneration. *BMC genomics*, **14**, 434.
862    Yang H, Wang X, Liu X, Wu J, Liu C, Gong W, Zhao Z, Hong J, Lin D, Wang Y, Lai R (2009)
863    Antioxidant peptidomics reveals novel skin antioxidant system. *Molecular & cellular
864    proteomics : MCP*, **8**, 571–83.
865    Zanger UM, Schwab M (2013) Cytochrome P450 enzymes in drug metabolism: Regulation of
866    gene expression, enzyme activities, and impact of genetic variation. *Pharmacology &
867    Therapeutics*, **138**, 103–141.
868    Zhao F, Yan C, Wang X, Yang Y, Wang G, Lee W, Xiang Y, Zhang Y (2014) Comprehensive
869    transcriptome profiling and functional analysis of the frog (*Bombina maxima*) immune
870    system. *DNA Research*, **21**, 1–13.