First *de-novo* transcriptome assembly of a South American frog, *Oreobates cruralis*, enables population genomic studies of Neotropical amphibians

Santiago Montero-Mendieta Corresp., [1] , **Manfred Grabherr** [2] , **Henrik Lantz** [3] , **Ignacio De la Riva** [4] , **Jennifer A Leonard** [1] , **Matthew T Webster** [5] , **Carles Vilà** Corresp. [1]

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

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

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

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

[5] 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 is opening the door to novel insights into the population structure and evolutionary history of poorly known species. In organisms with large genomes, which includes most amphibians, whole-genome sequencing is 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 to facilitate assembly and the transcriptome sequence must be assembled *de-novo*. We used RNA-seq to obtain the transcriptome 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 workflow to assist with pre-processing, assembling, evaluating and functionally annotating a *de-novo* transcriptome from RNA-seq data of non-model organisms. Our workflow 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/PRACTICAL-GUIDE-TO-BUILD-DE-NOVO-TRANSCRIPTOME-ASSEMBLIES-FOR-NON-MODEL-ORGANISMS/wiki

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2980v1 | CC BY 4.0 Open Access | rec: 16 May 2017, publ: 16 May 2017

First *de-novo* transcriptome assembly of a South American frog, *Oreobates cruralis*, enables population genomic studies of Neotropical amphibians

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

[1] Conservation and Evolutionary Genetics Group, Department of Integrative Ecology, Doñana Biological Station (EBD-CSIC), Avd. Américo Vespucio N26, 41092 Seville, Spain

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

* Corresponding authors:

E-mail: santiago.montero@ebd.csic.es (SMM)

E-mail: carles.vila@ebd.csic.es (CV)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2980v1 | CC BY 4.0 Open Access | rec: 16 May 2017, publ: 16 May 2017

Abstract

Whole genome sequencing is opening the door to novel insights into the population structure and evolutionary history of poorly known species. In organisms with large genomes, which includes most amphibians, whole-genome sequencing is 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 to facilitate assembly and the transcriptome sequence must be assembled *de-novo*. We used RNA-seq to obtain the transcriptome 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 workflow to assist with pre-processing, assembling, evaluating and functionally annotating a *de-novo* transcriptome from RNA-seq data of non-model organisms. Our workflow 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/PRACTICAL-GUIDE-TO-BUILD-DE-NOVO-TRANSCRIPTOME-ASSEMBLIES-FOR-NON-MODEL-ORGANISMS/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 genome sizes. The average genome size of frogs is 5.0 gigabases (Gb), while the fire salamander (*Salamandra salamandra*) genome averages 34.5 Gb (Gregory *et al.* 2007). For this reason, few genomics studies on amphibians have been carried out so far. To date, only the genome of three frogs of reduced genome size, *Xenopus tropicalis* (1.5Gb; Hellsten *et al.* 2010), *Xenopus laevis* (2.7Gb; Session *et al.* 2016) and *Nanorana parkeri* (2.3Gb; Sun *et al.* 2015), have been sequenced and published, in contrast to the larger number of genomes of reptiles (10), birds (53) and mammals (43). Due to the difficulties to obtain reference genome sequences for species with large genome sizes, reduced representation approaches are a cost-effective way to obtain information on genome-wide variability. For non-model organisms in which whole genome sequencing (WGS) is not feasible, transcriptome (e.g. Geraldes *et al.* 2011; De Wit *et al.* 2015) or exome (Lamichhaney *et al.* 2012) sequencing are commonly used as a reduced representation of the genome.

In amphibians, 24 transcriptomes from 19 species are currently available on the Transcriptome Shotgun Assemblies (TSA) database (https://www.ncbi.nlm.nih.gov/genbank/tsa/, January 2017), highlighting the importance of RNA sequencing (RNA-seq) for genomic studies in this group. RNA-seq is more affordable than whole genome sequencing and has rapidly become the preferred method for cataloguing and quantifying the complete set of transcripts or messenger RNA for a specific tissue, developmental stage or physiological condition (Wang *et al.* 2009). Nowadays, RNA-seq has a wide variety of uses but the core analyses include transcriptome profiling, differential gene expression and functional profiling (Conesa *et al.* 2016). As transcriptome assembly becomes more common for non-model and poorly-known organisms, we expect it will become a more popular tool also in phylogenomics as well as in demographic and population structure inference. However, what kind of RNA-seq data analysis is to be performed depends on the species of interest and the research goals. For model organisms and their close relatives, RNA-seq data is analyzed by mapping reads to a reference genome. By contrast, most non-model organisms do not have a reference genome from a sufficiently closely related species, and the transcriptome must be assembled *de-novo* (Martin & Wang 2011). Many bioinformatics tools to build a *de-novo* transcriptome are now available, yet contrasting opinions about the steps to follow may be disorienting. Some extremely simple pipelines have been developed to automatize the process (e.g. TRUFA; Kornobis *et al.* 2015), but this may limit the flexibility of the different pieces of software that have been integrated.

Here, we present the transcriptome profile for *Oreobates cruralis*, a direct-developing frog species from the Amazonian regions of Bolivia and Peru. To date, this is the first transcriptome available for a South American amphibian. We also present a simple workflow for pre-processing, building and functionally annotating a *de-novo* transcriptome from RNA-seq data of non-model organisms using available software.

Methods

Study model and sample collection

The genus *Oreobates* Jiménez de la Espada, 1872 (Anura: Craugastoridae) is a poorly studied clade of New World direct-developing frogs (Terrarana) distributed from the lower slopes of the eastern Andes into the upper Amazon basin, encompassing from southern Colombia and western and central Brazil up to northern Argentina. More than half of the 24 identified species have been described in the last decade and the species diversity in this genus is likely to be underestimated (Köhler & Padial 2016). One of these species, *O. cruralis* (Boulenger, 1902) occurs in a wide range of elevations and habitats across Bolivia and Peru. Its distribution includes lowland Amazonian rainforests (approximate range, from 100 to 600 m.a.s.l.), Yungas-montane Amazonian rainforests (600–2500 m.a.s.l.), and inter-Andean dry valleys (1300–3000 m.a.s.l.) (De la Riva *et al.* 2000). However, little is known about its ecology and evolutionary history.

For this study we used tissue samples from a single individual of *O. cruralis,* sampled in Bolivia (Villa Tunari, Cochabamba, Bolivia; 345 m.a.s.l.; 16º59'01.4"S 65º24'30.16"W) on November 28th, 2013 and deposited at the tissue collection of the Museo Nacional de Ciencias Naturales (MNCN-CSIC) in Madrid, Spain (MNCN/ADN:65263; Colección Boliviana de Fauna, CBF 7268). Samples of five tissues (intestine, liver, spleen, heart and skin) were isolated and preserved in Nucleic Acid Preservation (NAP) buffer (Camacho-Sánchez *et al.* 2013) at -80ºC.

Transcriptome sequencing

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

RNA-seq data analysis

The overall workflow is summarized in Figure 1 and our practical guide is available at: https://github.com/biomendi/PRACTICAL-GUIDE-TO-BUILD-DE-NOVO-TRANSCRIPTOME-ASSEMBLIES-FOR-NON-MODEL-ORGANISMS/wiki. Briefly, the first step after obtaining the raw sequence data in FASTQ format was to perform a preliminary quality control analysis with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC delivers quality metrics that are useful to identify if the data requires initial pre-processing before the transcriptome assembly. The pre-processing stage included three steps: first, removal of possible ribosomal RNA (rRNA) contamination; second, trimming low quality bases and PCR adapters; third, normalization to remove large excess of reads corresponding to moderately and highly expressed transcripts. Pre-processing is not always needed but it is highly recommendable to improve assembly quality. Once the data was pre-processed, a quality control was performed again and then, clean reads were *de-novo* assembled in absence of a reference genome. Subsequent analyses depend on the study goals. In our case, transcripts were functionally annotated using various databases to obtain a transcriptome profile. These steps are described in further detail in the following paragraphs.

We filtered raw FASTQ reads using SORTMERNA-v2.1 (Kopylova *et al.* 2012) against 8 default rRNA databases (SILVA 16S bacteria, SILVA 16S archaea, SILVA 18S eukarya, SILVA 23S bacteria, SILVA 23s archaea, SILVA 28S eukarya, Rfam 5S archaea/bacteria, Rfam 5.8S eukarya) to remove rRNA. Then, we used TRIMMOMATIC-v0.32 (Bolger *et al.* 2014) to trim adaptors and sequences with Phred quality score < 20. We normalized cleaned data of each tissue using the *in-silico* normalization utility included in the TRINITY-2.2.0 package (Grabherr *et al.* 2011). Normalization is useful for large RNA-seq data sets (>300 million paired-end reads) because it will delete over-expressed transcripts, thus lowering memory consumption and speeding up the assembly process (Haas *et al.* 2013). We merged the resulting data for the three tissues into a single dataset and normalized again prior to assembly. We used TRINITY (Grabherr *et al.* 2011) to *de-novo* assemble normalized reads into contigs. This resulted in a large number of transcripts, higher than the expected number of genes, likely because of alternative splicing. To avoid redundant transcripts, we kept the longest isoform for each "gene" identified by TRINITY (unigene) using the "get_longest_isoform_seq_per_trinity_gene.pl" utility in TRINITY. Thus, each unigene represented a collection of expressed sequences (i.e. transcripts) that apparently came from the same transcription locus, representing a putative gene. This set of unigenes was kept for downstream analyses.

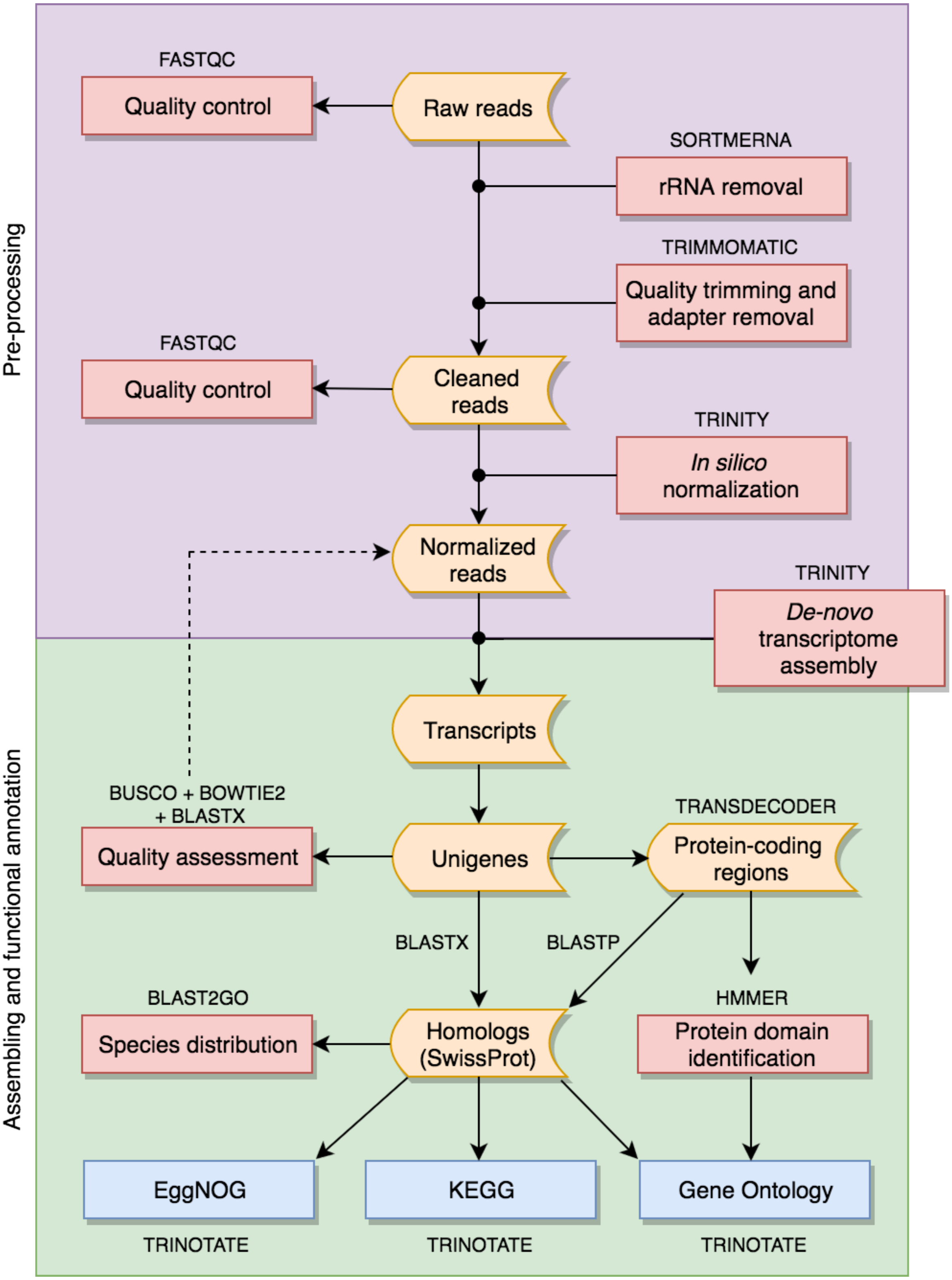
We evaluated the quality of the assembly and the transcript contiguity in terms of read representation by mapping normalized reads back to the set of unigenes using BOWTIE-1.1.2 (Langmead *et al.* 2009). We assessed the assembly completeness in terms of gene content using BUSCO-v1 (Simao *et al.* 2015) by searching the unigenes for the presence or absence of conserved orthologs in the tetrapoda-odb9 database (http://busco.ezlab.org/datasets/tetrapoda_odb9.tar.gz) that represents a collection of 3,950 single-copy tetrapoda orthologs. We also mapped with E-value ≤ 1E-20 the unigenes to the SwissProt database (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/) and to the Western clawed frog (*Xenopus tropicalis*) proteome (http://ftp.ensembl.org/pub/release81/fasta/xenopus_tropicalis/pep/Xenopus_tropicalis.JGI_4.2.pep.all.fa.gz) using BLASTX (searches within a protein database using a translated nucleotide query) included in the NCBI-BLAST-2.4.0+ package (Altschul *et al.* 1990). There is no perfect E-value cut-off in BLAST, but the smaller the most reliable the match. We used orthologous proteins found in SwissProt and *X. tropicalis* to assess completeness as described by Haas *et al.* (2013).

We predicted protein coding regions in the unigenes based on the most likely longest-ORF using TransDecoder-v3 (Haas *et al.* 2013). We searched BLAST homologies for the predicted proteins using TRINOTATE-v.3 (https://trinotate.github.io/) with E-value ≤ 1E-5 via BLASTP (search protein database using a protein query) to the SwissProt database. Homologies were also searched with E-value ≤ 1E-5 using TRINOTATE via BLASTX to the *X. tropicalis* proteome and the SwissProt database. In both cases, BLASTP and BLASTX, we only kept top-hit matches. We used BLAST2GO (Conesa *et al.* 2005) to detect the species distribution of the top BLASTX results within the SwissProt database. We identified protein domains using TRINOTATE via HMMER-3.1b2 (Finn *et al.* 2011) to the Pfam-A database (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/). At the time of conducting this study, TRINOTATE was built around specific releases of SwissProt and Pfam databases (available at https://data.broadinstitute.org/Trinity/Trinotate_v3_RESOURCES/). Homologous proteins found in the SwissProt database were used to retrieve functional annotation comments from the GO (*Gene Ontology;* Ashburner *et al.* 2000), EggNOG (*Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups;* Powell *et al.* 2012) and KEGG (*Kyoto Encyclopedia of Genes and Genomes;* Kanehisa *et al.* 2012) databases via TRINOTATE. The software also searched GO terms in Pfam results ("GO-Pfam") and in the combined results of homology search via SwissProt and Pfam ("GO-SwissProt_Pfam"). We used BLAST2GO to categorize the annotated GO terms in the latter. EggNOG annotations were filtered to keep COGs (Clusters of Orthologous Groups) and those were categorized using the current version of the COG database (ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data). KEGG annotations were filtered to keep KOs (KEGG orthology) and those were categorized using the tool "Reconstruct Pathway" (http://www.kegg.jp/kegg/tool/map_pathway.html).

Data availability

Raw RNA-seq data in FASTQ format has been deposited at the NCBI Sequence Read Archive database (SRA) under the accession SRP106442. The transcriptome assembly in FASTA format has been deposited at DDBJ/EMBL/GenBank under the accession GFNJ00000000. The version described in this paper is the first version, GFNJ01000000. The quality of the assembly was examined through the NCBI contamination screen. The screen found 5 sequences to exclude, 105 sequences with locations to mask/trim and 6 potentially duplicated sequences (with 3 distinct checksums). As a result, the uploaded information contained 422,970 sequences (188,369,677 bp) rather than the initial 422,999 sequences (188,399,293 bp). All the data is available at NCBI BioProject under the accession PRJNA384528.

Figure 1: Overall workflow for the annotation of RNA-seq data.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2980v1 | CC BY 4.0 Open Access | rec: 16 May 2017, publ: 16 May 2017

Results and discussion

RNA sequencing and transcriptome assembly

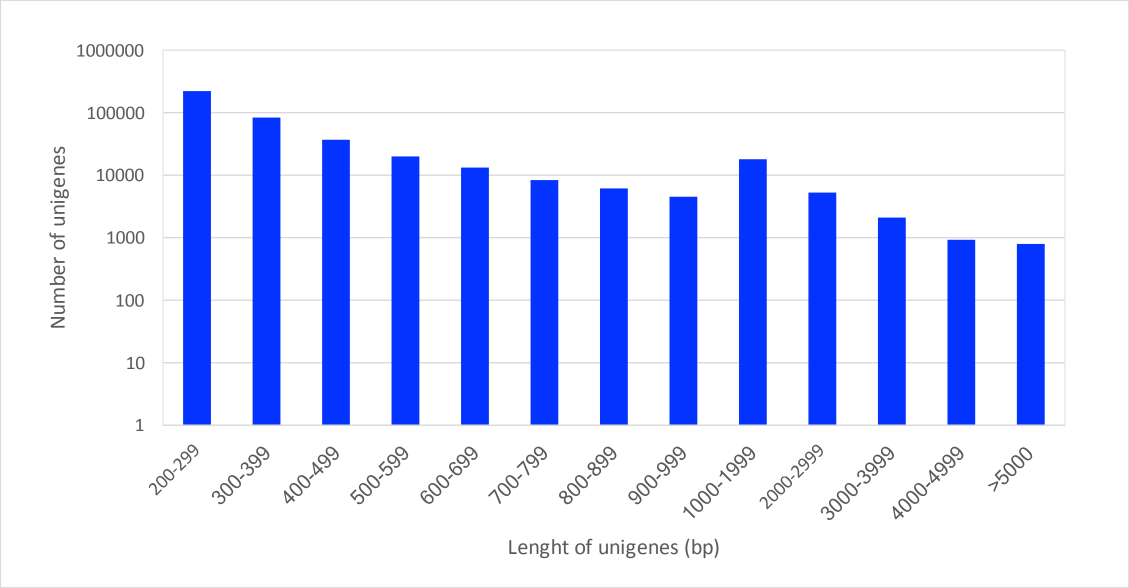
A summary of the RNA-seq data and transcriptome assembly is presented in Table 1. Illumina RNA sequencing for three tissues of *O. cruralis* in an Illumina HiSeq2500 instrument produced a total of 522,877,358 raw reads (intestine: 193,693,696; liver: 189,463,370; spleen: 139,720,292). Of those, 81.47% were kept after the pre-processing stage (426,003,462). The number of reads was further reduced to 6.97% after *in silico* normalization prior to assembly (36,428,858 reads). This highlights the importance of normalization to remove over-expressed transcripts in RNA-seq data. A total of 550,871 transcripts were obtained after *de-novo* transcriptome assembly. This large number of transcripts is not too surprising, both in terms of RNA-seq assembly as well as given the species and its likely large genome (see genome size for closely related genera at: http://www.genomesize.com/). First, transcriptome assemblies often include incompletely spliced introns, orphaned UTRs, read through off of the 3' ends, spuriously transcribed regions, active transposable elements, etc., so the number of assembled transcripts typically exceeds the expected number of protein coding genes by an order of magnitude. Second, large genomes tend to have large transcriptomes. In the axolotl (*Ambystoma mexicanum*) the transcriptome assembly had ~1,500,000 transcripts that clustered into ~1,300,000 putative genes (unigenes), and of those, 110,000 mapped to 30,000 SwissProt genes (Bryant *et al.* 2017). It is possible that these large genomes include a large number of repetitive sequences transcribed, which makes assembly more difficult and results in more fragmentation, especially when using diginorm (as in TRINITY) or any other in silico normalization. In *O. cruralis* the 550,871 transcripts clustered into 422,999 unigenes. This difference in number is likely because of alternatively spliced isoforms derived from paralogous genes (Wang *et al.* 2014). However, this will need to be confirmed with new amphibian genomes as they become available. Unigenes in the transcriptome of *O. cruralis* had an average GC content of 45.39%, which is very similar to other amphibians, such as the axolotl (*A. mexicanum* 45.56%; Hall *et al.* 2016), the green toad (*Bufotes viridis* 46.83%; Gerchen *et al.* 2016) or the common frog (*Rana temporaria* 44%; Price *et al.* 2015). The size of the unigenes in *O. cruralis* ranged from 201 to 16,804 bp with a mean length of 445 bp and a N50 length of 467 bp (Table 1; Figure 2). Here, the N50 value indicates that half of the transcriptome unigenes were at least 467 bp in length. The N50 length has been proposed as an estimator of genome assembly contiguity, since better assemblies will result in longer contigs (Li *et al.* 2014; Simpson 2014). However, in transcriptome data this measure can be highly misleading because it does not assess assembly completeness in terms of read representation or gene content (Simao *et al.* 2015).

Table 1: Summary of transcriptome data assembly for *Oreobates cruralis*.

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

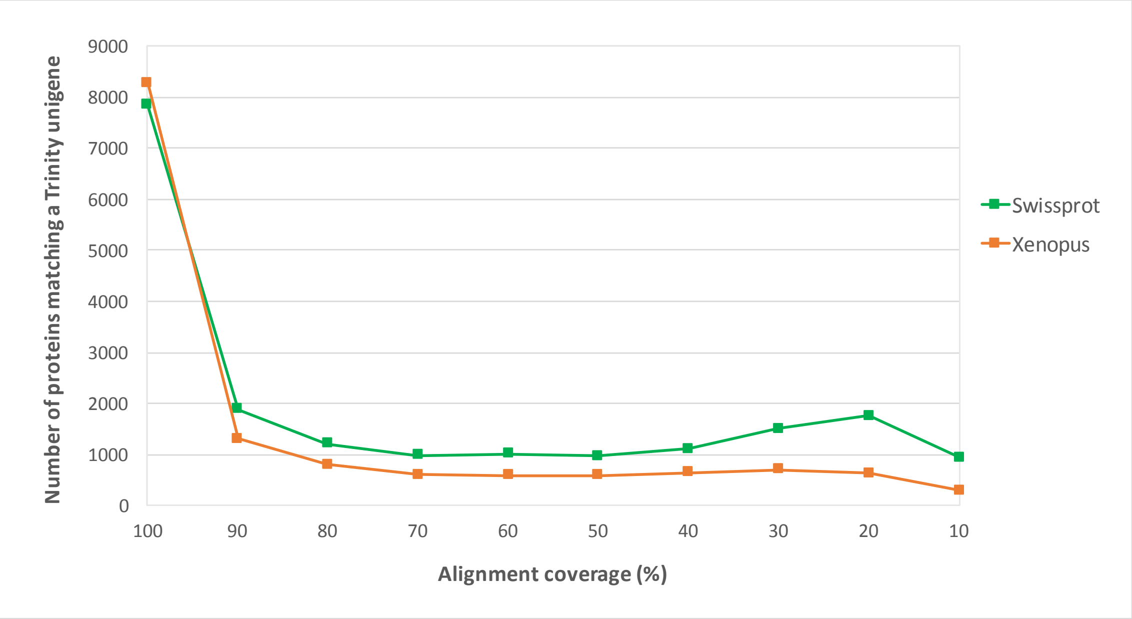
Figure 2: Length distribution of unigenes from *Oreobates cruralis*.



Transcriptome quality assessment

The set of assembled unigenes might not always perfectly correspond to all properly paired reads, as some unigenes might be built from just a portion of reads coming from the same transcription locus. When we evaluated assembly quality in terms of read representation, we found a high rate of reads that mapped back to unigenes (75.40%), thus confirming the presence of most of the initial reads in our final set of unigenes. When we evaluated the assembly completeness in terms of gene content, we found 2,830 complete orthologous genes (71.65%) out of the 3,950 genes available in the tetrapoda database (complete BUSCO hits). Of those, 2,501 were single-copy genes and 329 were duplicated genes. Only 462 (11.70%) of the genes in the database appeared fragmented and 658 (16.65%) were missing. We also obtained a high number of orthologous proteins in both the SwissProt and the *X. tropicalis* databases that fully matched (100% alignment coverage) or nearly fully corresponded (>80% alignment coverage) to unigenes in *O. cruralis* (Figure 3). Altogether, the high number of complete (or nearly complete) orthologous matches across the different databases provides a valuable validation of the depth and completeness of the assembly process.

Figure 3: Distribution of BLASTX hit coverage against SwissProt and *Xenopus* databases.



Functional annotation of unigenes

Gene annotation consists in adding relevant biological information to coding regions of the genome and it was arguably the most relevant section of our workflow, since it allows for describing and classifying the content of the *O. cruralis* transcriptome. Functional annotation was based on BLAST searches to find homologous proteins against a reference database (e.g. SwissProt) and then collect biological information from various sources (e.g. GO, KEGG, EggNOG or PFAM). We predicted a total of 45,466 protein-coding genes within the 422,999 unigenes using TansDecoder. After homology search using BLASTP, we found that 26,418 protein-coding genes in *O. cruralis* mapped to proteins in the SwissProt database. Search using BLASTX revealed a total of 54,425 unigenes that mapped to proteins in the *X. tropicalis* proteome and 47,349 unigenes that mapped to proteins in the SwissProt database. The relative low number of homologous proteins shared between *O. cruralis* and *X. tropicalis* is likely because of the very ancient divergence time between both species (estimated to be around 204 million years ago; http://www.timetree.org/). However, the observation of a number of matches (54,425) larger than the total number of proteins in *X. tropicalis* (N=22,718) may suggest the presence of a high number of duplicates or unresolved splice variants among the unigenes of *O. cruralis*. The version of the SwissProt database used included a selection of 553,231 protein sequences from 13,379 species, and the top-hit species distribution showed that 32% (13,099) of the *O. cruralis* unigenes were homologs to human (*Homo sapiens*) proteins and 19% (7661) to house mouse (*Mus musculus*) proteins (Figure 4). The larger number of hits to mammals than to other amphibians is likely due to the uneven distribution of species in the SwissProt database, in which the top twenty species accumulate 21.5% of the entries. Still, amphibian species were highly represented in the assembly with 10% (3909) of the *O. cruralis* unigenes having a highest match to *X. laevis* and 5% (1893) to *X. tropicalis* proteins. When we retrieved the functional comments for the homologous proteins found in the SwissProt database, the number of annotated unigenes varied depending on the source that was used (Figure 5).

Figure 4: Top-hit species distribution in "SwissProt-BLASTX" for unigenes from the transcriptome of *O. cruralis*.

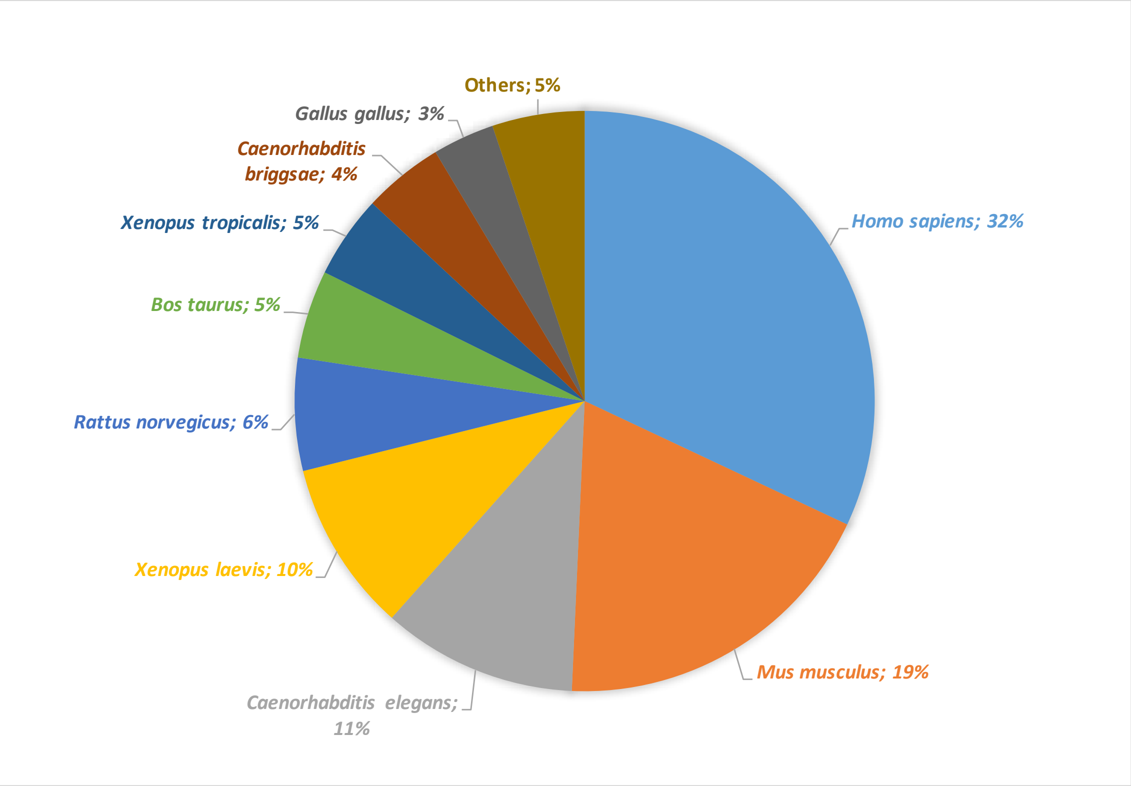
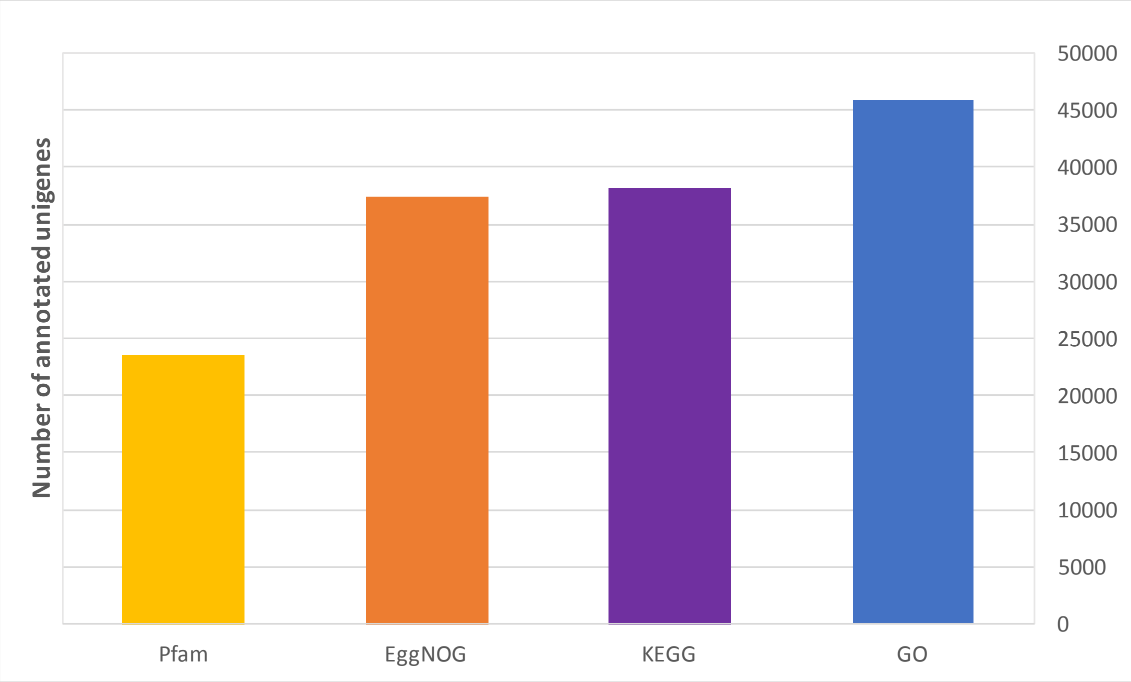


Figure 5: Number of annotated unigenes in the transcriptome of *O. cruralis* using various sources.



Protein domain identification

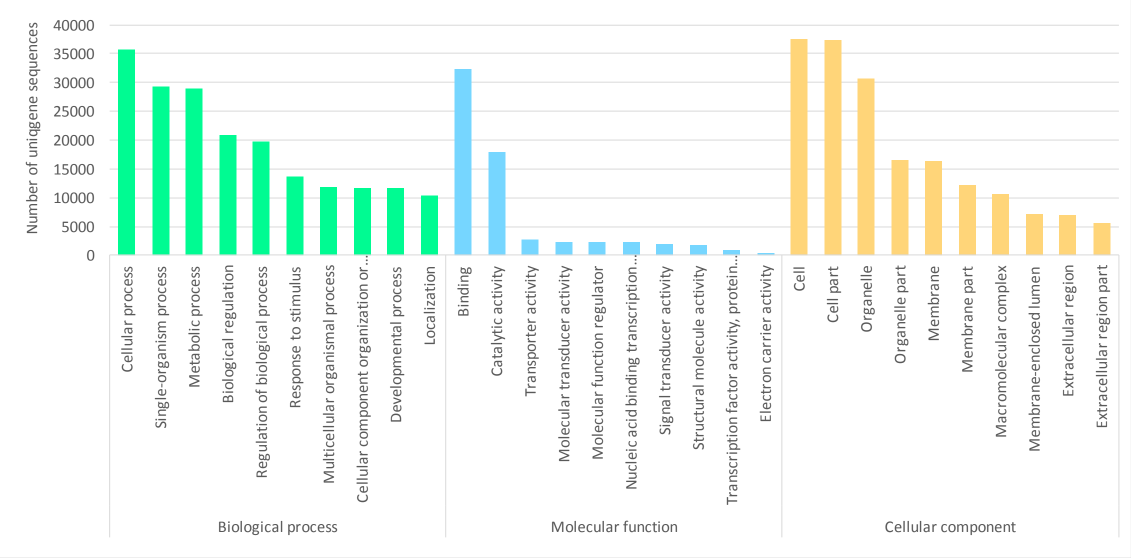
Protein domains are preserved portions of proteins with tertiary structure that can act, evolve and exist independently of the rest of the protein chain (Jacob 1977). Prediction of protein domains is an important step of transcriptome annotation since they provide insights in specific cellular functions that assist comparative genomics of domain families across species (Ochoa *et al.* 2011). The Pfam database is a large collection of protein families that currently contains 16,303 families (Pfam v30.0). From the predicted 45,466 protein-coding genes in the transcriptome of *O. cruralis*, we identified 23,500 that are present in the Pfam-A database, consisting of 5,686 protein domain families. We found that the most common Pfam domain in the transcriptome of *O. cruralis* is the 'Zinc finger, C2H2 type' (961 hits; 4.09%). The C2H2 zinc finger proteins are very frequent in eukaryotic genomes (e.g. the human genome has 564 C2H2 zinc fingers; Tadepally *et al.* 2008), and their functions are extraordinarily diverse, including DNA recognition, RNA packaging, transcriptional activation, regulation of apoptosis, protein folding and assembly, and lipid binding (Laity *et al.* 2001). Interestingly, this protein family was also reported as the most common for other amphibians, such as the green frog (*Lithobates clamitans*) and the Pacific tree frog (*Pseudacris regilla*) (Robertson & Cornman 2014).

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

Table 2: Summary of top 10 PFAM domains in the transcriptome of *O. cruralis.*

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

Figure 5: Distribution of top-10 GO terms for the transcriptome of *O cruralis* derived from "GO-SwissProt_Pfam" results. Categories shown correspond to gene ontology level 2.



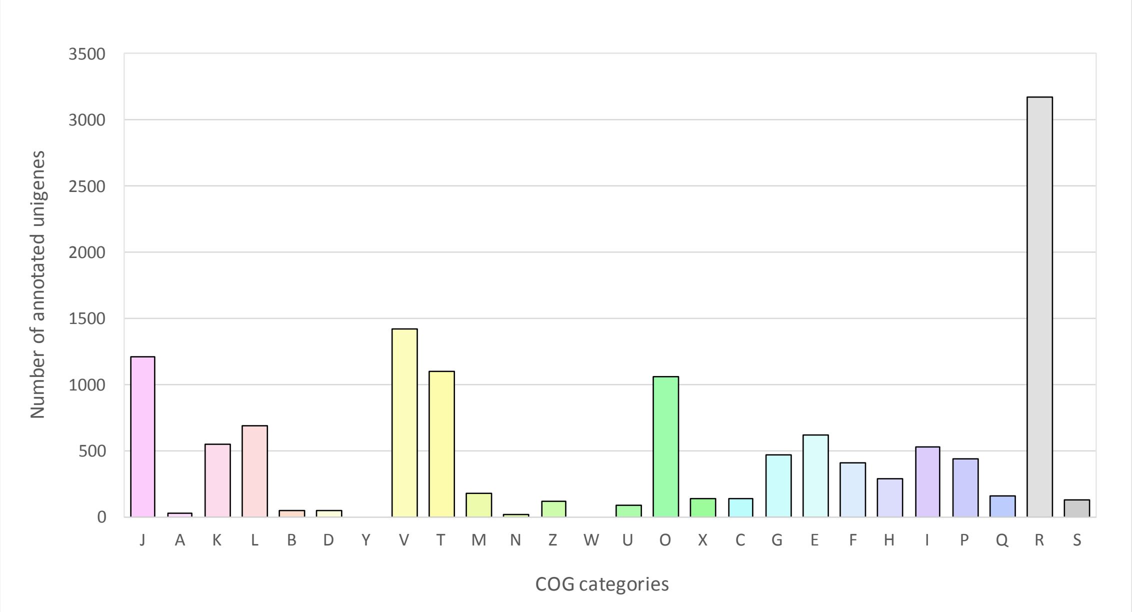
Gene ontology

The Gene Ontology (GO) (http://geneontology.org/) is a standardized functional classification system aimed to describe gene and gene product attributes across species, using a controlled vocabulary (i.e. ontology terms). The GO classification comprises three domains: cellular component, molecular function and biological process. These domains have a hierarchical structure and a GO term can belong to different levels depending on the path followed and the number of steps between the term and the root (Ashburner *et al.* 2000). Using the combined results of a homology search via SwissProt and Pfam, we detected a total of 3,094,863 GO terms (19,407 unique) corresponding to 45,885 (10.85%) unigenes in the transcriptome of *O. cruralis*. This contrasts previous studies that have reported that between 50 and 80% of the transcripts reconstructed from RNA-seq data can be annotated with GO terms (Conesa *et al.* 2016). However, the relatively low percentage of annotation may reflect the scarcity of amphibian sequences in the GO database, and therefore the presence of undetected novel transcripts. Still, the GO database produced the highest number of annotated unigenes compared to other sources, such as Pfam, KEGG or EggNOG (Figure 5). The largest number of GO terms corresponded to the category of "biological process" (BP; 49%) followed by "cellular component" (CC; 38%) and "molecular function" (MF; 13%). At ontology level-2, which represents the second most general category in the GO database, there were 65 different GO terms (Figure 6). Within the BP category, the most frequent GO terms were "cellular process" (35,730) and "single-organism process" (29,237). Within the MF category, unigenes were mainly linked to "binding" (32,275) and "catalytic activity" (18,023). Within the CC category, unigenes were mostly associated with "cell" (37,424) and "cell part" (37,293). These highly abundant GO terms are likely associated to genes involved in essential cell functions and metabolism regulation, since they describe very general terms. A similar distribution of GO terms was found in a comparative transcriptome study of seven anuran species (Huang *et al.* 2016). We found 185 unigenes with antioxidant activity, most of them with peroxidase activity (128). This number is relatively high compared to the 63 antioxidant genes present in humans (Gelain *et al.* 2009) and it might be related to the high number of protein kinase domains that we recorded earlier, as well as the habitat of *O. cruralis*. Specimens are usually encountered in tropical rainforest leaf litter, where amphibian pathogens are common (Pounds *et al.* 2006). Antioxidant genes have previously been reported from the skin of amphibians, contributing to resistance against microorganism infection or radiation injury (Yang *et al.* 2009). However, since the transcriptome of *O. cruralis* was built from tissues of intestine, liver and spleen, our results suggest that antioxidant genes in amphibians can also be expressed in different tissues besides skin. Because *O. cruralis* is mainly a lowland Amazonian rainforests frog, it would be interesting to compare this results with closely-related species living in higher altitudes (e.g. *Oreobates ayacucho*), where temperature is lower and microbial activity too.

COG classification

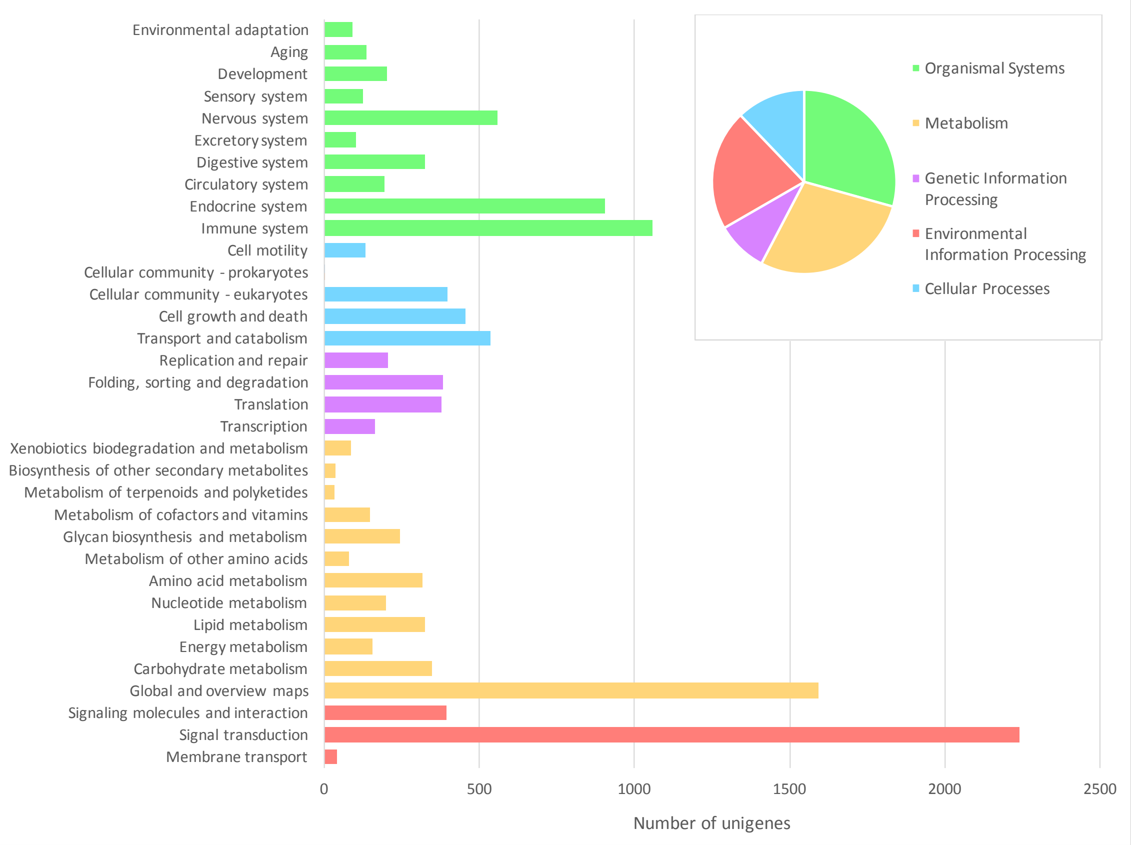
The database of Clusters of Orthologous Groups (COGs) is another common tool for functional annotation (Galperin *et al.* 2015). In this database, orthologous genes from 722 prokaryote genomes are grouped according to their biological function. The current version consists of 4,632 COGs classified into 26 functional categories. The EggNOG database is based on the original idea of COGs and expands it to non-supervised orthologous groups from numerous organisms, including eukaryotes and viruses (Huerta-Cepas *et al.* 2016). We identified a total of 37,349 (8.83%) unigenes that are present in the EggNOG database (Figure 5). Of these, 12,993 belonged to the COG database, corresponding to 24 functional categories (Figure 6). The "general function" category (3,166; 24.37%) represented the largest group, followed by "defense mechanisms" (1,421; 10.94%). Our results showed that genes related to defense functions may be relatively abundant in the transcriptome of *O. cruralis*, particularly compared to the seven anurans studied by Huang *et al.* (2016) and also to *A. mexicanum* (Wu *et al.* 2013). In both studies, only about 2% of unigenes corresponded to defense mechanisms. Within the unigenes involved in defense mechanisms, we identified 1,163 (81.84%) that are related to Cytochrome P450 enzymes (CYPs), while only 57 of those genes have been found in humans (Zanger & Schwab 2013). CYPs are a protein superfamily in charge of metabolizing potentially toxic compounds, such as drugs or products of endogenous metabolism (Fujita *et al.* 2004). This large difference in the number of genes in humans and *O. cruralis* may indicate the presence of duplicates in our data, but it could also be associated with some degree of myrmecophagy in this group of frogs. Because the eating habits of *Oreobates* frogs have not been studied yet, protein data from strict myrmecophagous species (e.g. poison dart frogs in the family Dendrobatidae) are needed to confirm these results.

Figure 6: Distribution of COG categories for the transcriptome of *O cruralis*.



Color-coded according to the functional categories adopted in the COG system. Abbreviations as follow; J: Translation, ribosomal structure and biogenesis. A: RNA processing and modification. K: Transcription. L: Replication, recombination and repair. B: Chromatin structure and dynamics. D: Cell cycle control, cell division, chromosome partitioning. Y: Nuclear structure. V: Defense mechanisms. T: Signal transduction mechanisms. M: Cell wall/membrane/envelope biogenesis. N: Cell motility. Z: Cytoskeleton. W: Extracellular structures. U: Intracellular trafficking, secretion, and vesicular transport. O: Posttranslational modification, protein turnover, chaperones. X: Mobilome: prophages, transposons. C: Energy production and conversion. G: Carbohydrate transport and metabolism. E: Amino acid transport and metabolism. F: Nucleotide transport and metabolism. H: Coenzyme transport and metabolism. I: Lipid transport and metabolism. P: Inorganic ion transport and metabolism. Q: Secondary metabolites biosynthesis, transport and catabolism. R: General function prediction only. S: Function unknown.

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



KEGG pathways

In the KEGG (Kyoto Encyclopedia of Genes and Genomes) database, genes from completely sequenced genomes are linked to higher-level systemic functions of the cell, the organism and the ecosystem (Kanehisa & Goto 2000). Molecular-level functions are stored in the KO (KEGG Orthology) database, where each KO is defined as a functional ortholog of genes and gene products (Kanehisa *et al.* 2016). We identified a total of 38,120 (9.01%) unigenes from *O. cruralis* in the KEGG database (Figure 5). Of these, 25,619 unigenes have orthologs in the KO database. Many unigenes were classified under the category of organismal systems (3704; 29.32%), followed by metabolism (3580; 28.34%), environmental information processing (2678; 21.20%), cellular processes (1535; 12.15%) and genetic information processing (1135; 8.99%) (Figure 7). We found the largest number of unigenes to be related with signal transduction (2241) within the category of environmental information processing. Particularly, the PI3K-Akt signaling pathway was the most frequent (184; 8.21%) among the signal transduction unigenes, followed by the MAPK signaling pathway (152; 6.78%). Both the PI3K-Akt and the MAPK signaling pathways play a major role in the development of immune cells (Liu *et al.* 2007; Juntilla &

Koretzky 2008). Interestingly, the immune system category was also highly enriched (1057 unigenes) and within the immune category, the chemokine signaling pathway comprised the highest number of unigenes (105; 9.93%). Chemokine receptors associate with G proteins to promote signaling cascades, including MAPK pathways, that cause immune responses such as degranulation, a cellular process that releases antimicrobial cytotoxic molecules to destroy invading microorganisms (Murdoch & Finn 2000). This suggests that, compared to other genes, those related to the immune system are relatively abundant in the transcriptome of *O. cruralis*. We hypothesize that tropical conditions, in which high temperature and humidity are constant throughout the year, impose a crucial challenge to amphibian fitness. Although based on a single transcriptome our results lack of statistical power, this study provides a first view towards the understanding of gene evolution in neotropical amphibians.

Conclusions

Although large genome size renders complete genome sequencing practically unfeasible in many species, such as most amphibians, transcriptome sequencing represents a cost-effective alternative to obtain a large amount of genome-wide data. This can allow advances in the study of ecological and evolutionary processes beyond the limits imposed by the use of small panels of markers. In this study, we have provided and discussed a workflow that covers the basic elements needed to build a *de-novo* transcriptome from RNA-seq data of non-model organisms for which sequencing and assembling a genome is not a practical option. We have successfully applied this workflow to obtain the transcriptome profile of *Oreobates cruralis*, a poorly known neotropical frog. To date, this is the first transcriptome available for a South American amphibian, and therefore, a stepping stone towards the study of the diversification patterns across neotropical amphibians using genomic approaches. Once a reference transcriptome is available, capture-based approaches can help to obtain homologous sequences for a large array of closely-related species at a reduced cost. In this regard, this transcriptome will serve as a valuable resource for the inference of orthologous sequences in closely-related species. This, for example, will allow solving phylogenomic relationships among the species of the genus *Oreobates*, as well as studying population differentiation, demographic history and gene evolution for the different species.

Acknowledgements

The tissue sample used for this study was provided by the frozen tissue collection of the Museo Nacional de Ciencias Naturales (MNCN-CSIC) in Madrid, Spain (MNCN/ADN collection). Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, Sweden. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. Computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project b2015409. We thank all the members of the Conservation and Evolutionary Genetics Group for constructive comments.

References

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

Ashburner M, Ball CA, Blake JA *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature genetics*, **25**, 25–29.

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

Bryant DM, Johnson K, DiTommaso T *et al.* (2017) A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*, **18**, 762–776.

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

Conesa A, Götz S, García-Gómez JM *et al.* (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Conesa A, Madrigal P, Tarazona S *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biology*, **17**, 13.

Destro-Bisol G, Jobling MA, Rocha J *et al.* (2010) Molecular Anthropology in the genomic era. In: *Journal of Anthropological Sciences*, pp. 93–112.

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

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

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

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

Gelain DP, Dalmolin RJS, Belau VL *et al.* (2009) A systematic review of human antioxidant genes. *Frontiers in bioscience (Landmark edition)*, **14**, 4457–63.

Geraldes A, Pang J, Thiessen N *et al.* (2011) SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources*, **11**, 81–92.

Gerchen JF, Reichert SJ, Röhr JT *et al.* (2016) A single transcriptome of a green toad (*Bufo viridis*) yields candidate genes for sex determination and -differentiation and non-anonymous population genetic markers. *PLoS ONE*, **11**, 1–14.

Giallourakis C, Henson C, Reich M, Xie X, Mootha VK (2005) Disease gene discovery through integrative genomics. *Annual review of genomics and human genetics*, **6**, 381–406.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, **29**, 644–52.

Gregory TR, Nicol JA, Tamm H *et al.* (2007) Eukaryotic genome size databases. *Nucleic Acids Research*, **35**.

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.

Hall KW, Eisthen HL, Williams BL (2016) Proteinaceous pheromone homologs identified from the cloacal gland transcriptome of a male axolotl, *Ambystoma mexicanum*. *PLoS ONE*, **11**, 1–18.

Hellsten U, Harland RM, Gilchrist MJ *et al.* (2010) The genome of the Western clawed frog *Xenopus tropicalis*. *Science*, **328**, 633–6.

Huang L, Li J, Anboukaria H *et al.* (2016) Comparative transcriptome analyses of seven anurans reveal functions and adaptations of amphibian skin. *Scientific Reports*, **6**, 24069.

Huerta-Cepas J, Szklarczyk D, Forslund K *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, **44**, D286–D293.

Isenman DE, Painter RH, Dorrington KJ (1975) The structure and function of immunoglobulin domains: studies with beta-2-microglobulin on the role of the intrachain disulfide bond. *Proceedings of the National Academy of Sciences of the United States of America*, **72**, 548–52.

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

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

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

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

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

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

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

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

Kornobis E, Cabellos L, Aguilar F *et al.* (2015) TRUFA : A User-Friendly Web Server for de novo RNA-seq Analysis Using Cluster Computing. *Evolutionary bioinformatics online*, **11**, 97–104.

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

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

Lamichhaney S, Berglund J, Almén MS *et al.* (2015) Evolution of Darwin’s finches and their beaks revealed by genome sequencing. *Nature*, **518**, 371–375.

Lamichhaney S, Han F, Berglund J *et al.* (2016) A beak size locus in Darwin´s finches facilitates character displacement during a drought. *Science*, **352**, 470–474.

Lamichhaney S, Martinez Barrio A, Rafati N *et al.* (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences*, **109**, 19345–50.

Lander ES, Linton LM, Birren B *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.

Li B, Fillmore N, Bai Y *et al.* (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, **15**, 553.

Liu Y, Shepherd EG, Nelin LD (2007) MAPK phosphatases — regulating the immune response. *Nature Reviews Immunology*, **7**, 202–212.

Martin J a., Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*, **12**, 671–682.

McBride CM, Bowen D, Brody LC *et al.* (2010) Future Health Applications of Genomics. Priorities for Communication, Behavioral, and Social Sciences Research. *American Journal of Preventive Medicine*, **38**, 556–565.

Mcmahon BJ, Teeling EC, Höglund J (2014) How and why should we implement genomics into conservation? *Evolutionary Applications*, **7**, 999–1007.

Mukherjee S, Stamatis D, Bertsch J *et al.* (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic acids research*, **45**, D446–D456.

Murdoch C, Finn A (2000) Chemokine receptors and their role in inflammation and infectious diseases. *Blood*, **95**, 3032–43.

Ochoa A, Llinás M, Singh M (2011) Using context to improve protein domain identification. *BMC Bioinformatics*, **12**, 90.

Padial JM, Chaparro JC, De la Riva I (2008) Systematics of Oreobates and the Eleutherodactylus discoidalis species group (Amphibia, Anura), based on two mitochondrial DNA genes and external morphology. *Zoological Journal of the Linnean Society*, **152**, 737–773.

Pounds JA, Bustamante MR, Coloma L a *et al.* (2006) Widespread amphibian extinctions from epidemic disease driven by global warming. *Nature*, **439**, 161–7.

Powell S, Szklarczyk D, Trachana K *et al.* (2012) eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research*, **40**, D284-9.

Price SJ, Garner TWJ, Balloux F *et al.* (2015) A de novo assembly of the common frog (*Rana temporaria*) transcriptome and comparison of transcription following exposure to Ranavirus and Batrachochytrium dendrobatidis. *PLoS ONE*, **10**, 1–23.

Reuter JA, Spacek D V., Snyder MP (2015) High-Throughput Sequencing Technologies. *Molecular Cell*, **58**, 586–597.

Robertson LS, Cornman RS (2014) Transcriptome resources for the frogs *Lithobates clamitans* and *Pseudacris regilla*, emphasizing antimicrobial peptides and conserved loci for phylogenetics. *Molecular Ecology Resources*, **14**, 178–183.

Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nature Methods*, **5**, 16–18.

Session AM, Uno Y, Kwon T *et al.* (2016) Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, **538**, 1–15.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Simpson JT (2014) Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, **30**, 1228–1235.

Sun Y-B, Xiong Z-J, Xiang X-Y *et al.* (2015) Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E1257-62.

Tadepally HD, Burger G, Aubry M (2008) Evolution of C2H2-zinc finger genes and subfamilies in mammals: Species-specific duplication and loss of clusters, genes and effector domains. *BMC Evolutionary Biology*, **8**, 176.

Umbarger KO, Yamazaki M, Hutson LD, Hayashi F, Yamazaki A (1992) Heterogeneity of the retinal G-protein transducin from frog rod photoreceptors: Biochemical identification and characterization of new subunits. *Journal of Biological Chemistry*, **267**, 19494–19502.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, **10**, 57–63.

Wang W, Wang J, You F *et al.* (2014) Detection of alternative splice and gene duplication by RNA sequencing in Japanese flounder, *Paralichthys olivaceus*. *G3 (Bethesda, Md.)*, **4**, 2419–24.

De Wit P, Pespeni MH, Palumbi SR (2015) SNP genotyping and population genomics from expressed sequences - Current advances and future possibilities. *Molecular Ecology*, **24**, 2310–2323.

Wolfe KH (2006) Comparative genomics and genome evolution in yeasts. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **361**, 403–412.

Wu C-H, Tsai M-H, Ho C-C, Chen C-Y, Lee H-S (2013) De novo transcriptome sequencing of axolotl blastema for identification of differentially expressed genes during limb regeneration. *BMC genomics*, **14**, 434.

Yang H, Wang X, Liu X *et al.* (2009) Antioxidant peptidomics reveals novel skin antioxidant system. *Molecular & cellular proteomics : MCP*, **8**, 571–83.

Zanger UM, Schwab M (2013) Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, **138**, 103–141.

Zhao F, Yan C, Wang X *et al.* (2014) Comprehensive transcriptome profiling and functional analysis of the frog (*Bombina maxima*) immune system. *DNA Research*, **21**, 1–13.