INSIGHT INTO *ELEUTHERODACTYLUS COQUI* RESILIENCE AND ANURAN HYPOXIA TOLERANCE UTILIZING PINCHO

A dissertation submitted in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

to the faculty of the

DEPARTMENT OF BIOLOGICAL SCIENCES

of

ST. JOHN'S COLLEGE OF LIBERAL ARTS AND SCIENCES

at

ST. JOHN'S UNIVERSITY
New York
by
Randy Ortiz

Date Submitted _____ DateApproved_____

_____ _____

Randy Ortiz Dr. Juan C. Santos

ABSTRACT

INSIGHT INTO ELEUTHERODACTYLUS COQUI RESILIENCE AND ANURAN HYPOXIA TOLERANCE UTILIZING PINCHO

Randy Ortiz

The onset of high-throughput RNA-sequencing (RNA-seq) technology allowed research into transcriptomics to accelerate exponentially aided by the rapid advancement of bioinformatic pipelines. The *Pincho* workflow, my first research endeavor, is a transcriptomic workflow developed to create an avenue of high-quality reconstructions from RNA-seq data. High-quality reconstruction standards entail longer transcripts, more complete transcripts, and more raw data utilization. We have discovered an ideal trio of assemblers between transABySS, rnaSPAdes and TransLiG that would best reconstruct next-generation sequencing data according to these standards.

We utilized *Pincho* to drive two distinct experiments: (1) exploring the genetic basis for the successful invasion of *Eleutherodactylus coqui* (*E. coqui*) from Puerto Rico (PR) to mainland US and (2) exploring the genetic variation across anuran oxygen delivery and consumption systems. *E. coqui* is one of the top four invasive anurans in the US; described as a pest that has destabilized ecosystems and cost the taxpayers millions of dollars. Few researchers delve into the genetic explanations as to how and why *E. coqui* are so successful in colonizing locations outside PR. We discovered several differentially expressed defense response transcripts that differ between the two populations; with a focus on a novel cathelicidin sequence that is only expressed in native *E. coqui*. The absence of cathelicidin expression in invasive *E. coqui* leads us to attribute their successful invasion to entering a cleaner environment and subsequently having more energy to utilize on reproduction and expansion.

As we further studied *E. coqui* cathelicidin we questioned how variability in transcript structure might be more widespread in anurans, especially within oxygen delivery and conservation systems. Anurans are described as hypoxia/anoxia resilient in literature, thus we hypothesized these systems would be highly conserved. Our results revealed that hemoglobin was instead under significant episodic diversifying selection. Sites neighboring crucial heme and oxygen binding sites were also found to be under positive selection leading us to believe that these changes could alter overall oxygen affinity and lead to hematological consequences. We speculate that even if anurans are hypoxia/anoxia resilient, resilience levels can differ between species as shown in the sequence divergence in anuran protein alignments.

DEDICATION

I dedicate this dissertation to my family which have supported me in every way imaginable.

ACKNOWLEDGMENTS

I am forever grateful for the leadership, guidance, and support of St. John's University administration and staff as well as the support of my colleagues. I thank Juan C. Santos for his continued support throughout my research and for his mentorship that allowed me to secure full-time employment as a college instructor even before I officially graduated. I thank all my laboratory members: Victoria Akilov, Leeann C. Dabydeen, Carolyn Kosinski, Priyanka Gera, Ronit Eliav, Md. Abu B. Siddique, and Juan D. Carvajal. Each one of the laboratory members has helped me grow and continue to inspire me with their achievements. I thank all my collaborators: Greg Pauly, Hunter J. Howell and Emily Powell. I thank all those who provided the knowledge I needed in course curriculum and/or advisement so that I could succeed in my career in biology: Dr. Yong Yu, Dr. Ales Vancura, Dr. Dianella G. Howarth, Dr. Christopher Bazinet, Dr. Javier F. Juárez, Dr. Richard Stalter and Dr. Rachel Zufferey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

PINCHO: A MODULAR APPROACH TO HIGH QUALITY DE NOVO TRANSCRIPTOMICS

1.1 INTRODUCTION

Homemade de novo transcriptomic workflows tend to be idiosyncratic to specific study goals, unoptimizable to other studies and, in many cases, left unpublished or buried in supplementary materials. We could say Rnnotator [1] in 2010 was the first single-assembler transcriptomic pipeline to be publicly available, while the Oyster River Protocol (ORP; [2]) in 2018 was the first multi-assembler pipeline available. This presumed eight-year period between single- and multi-assembler approaches is odd considering multi-assembler methods have been shown to produce reconstructions with higher degrees of completeness [2]. Nevertheless, the combinations of assemblers that produce the best reconstructions in the multi-assembly approach are not well explored nor classified. Adding to the complexity of the situation, assemblers are routinely updated, and new assemblers are created in a timely fashion, making assembler comparisons both a necessity and routine process. The closest comparison to our workflow would be the ORP; however, it employs a rigid tri-assembly approach to produce high quality transcriptomes via rnaSPAdes (k55, k75; [3]), Trinity (k25; [4]) and Shannon (k75; [2,5]). In comparison, we developed an opensource workflow that broadens the k-mers used to up to five total k-mers per assembler. Our software, Pincho [6], allows the user to design and customize their own k-mer list and number of assemblers, among other parameters.

To characterize our management software, we present two major goals of this study that we sought to complete. The first was to construct a publicly available and customizable modular management toolkit that could simplify de novo transcriptomic work for data scientists. This simplification took place via the amalgamation of well-established and reviewed genomic and transcriptomic software centralized in one quick download and even faster user implementation options. We customize this workflow with the most common software used in de novo transcriptomics along with the modularity to allow simple incorporation of new software as future tools become available. Our second goal is to provide a comprehensive analysis on de novo transcriptome assembler performance individually and in combination. To our knowledge, this is the first publication on synergistic effects of single-, bi-, and tri-assembly combinations between nine distinct de novo and reference-guided assemblers aimed to elevate de novo transcriptome quality and completeness.

1.2 MATERIALS AND METHODS

1.2.1 Components of the Pincho Workflow

Our software supports various applications and automates their parameters, computer resources and output management via Python3 and Bash (Supplementary Figure S1). Pincho consists of twenty-five functions which fall under six modules: preprocessing (adaptor removal with Trimmomatic [7,8] and error correction via Rcorrector [9,10]); de novo assembly (ABySS [11,12], Tadpole [13,14], BinPacker [15,16], IDBA-tran [17,18], MEGAHIT [19,20], Oases/Velvet [21,22], rnaSPAdes [3,23], Shannon Cpp [5,24], SPAdes [25,26], Trans-ABySS [27,28], TransLig [29,30], and Trinity [4,31], Table 1; post-assembly (consensus assembly generation with TransRate [32,33], isolation of short transcripts under bp length threshold and redundancy reduction via CDHIT [34,35]); assembly assessment (alignments to reference transcriptomes or to the original raw reads via HISAT2 [36,37], BUSCO [38,39] and TransRate); annotation using a user reference (NCBI BLASTX, BLASTN, and BLASTP; [40–42]); and expression analysis (kallisto [43,44] and RSEM [45,46], Figure 1 and Supplementary Figure S1). Several important notes: Pincho can process Sequence Read Archive (SRA, [47]) data accession numbers via SRAtoolkit [48], Trinity can be run in genome guided mode instead of De novo with help from Samtools [49,50], and TransLig was modified to include assembly lengths via SeqKit [51,52].

1.2.2 Dataset Criteria and Selection

We analyzed eight distinct non-model datasets from the SRA ([53]; Table 2. We focused on hyloid anurans (frogs) that have complex and usually large genomes (e.g., ~6.76 Gb for *Dendrobates pumilio*, [54]). Data was chosen via the following criteria: (a) publicly

sourced RNA-seq data, (b) paired-end reads of various insert sizes (Table 2), (c) fastq format, (d) Illumina sequencing, (e) non-model organisms, (f) data containing a base count lower than 2Gb and (g) data that passed Pincho's rapid assessment with a complete BUSCO score greater than 50%. Rapid assessment is composed of fasterq-dump download of SRR raw reads, removal of Illumina adaptors, if necessary, from raw data via Trimmomatic, assembly of reads via succinct de Bruijn graphs with MEGAHIT and assessment via BUSCO scores. Chosen SRA files were analyzed with FastQC [55], revealing that all files were adapter free.

Our datasets are purposely under the standard yield of RNA-seq experiments (2GB –4GB), to highlight the potential of the selected assemblers on low yield, low coverage datasets. As higher levels of sequencing coverage lead to higher quality NGS data [56], we chose NGS data that are most likely to contain low sample coverage owing to low read counts [57]. We selected smaller sized files on average 6.88M reads, which is well beneath the recommended sequencing read number of 20M [56] to ensure an NGS scenario of low coverage. As a balance we made sure that all files were at least above 50% in complete BUSCO scores to avoid scenarios where read coverage was insufficient. Low coverage datasets are prone to many types of assembly errors (i.e., fragmentation and incompleteness [32]), which allows us to accurately test the various types of algorithms employed by the tested transcriptome assemblers and their abilities to work with problematic datasets. It is only under this scope that we can ideally view assembler performance and synergy without the reliance on synthetic data. We expect that if assemblers succeed at reconstructing more from smaller datasets, then they are sensitive enough to use on larger datasets as well.

1.2.3 Pincho Workflow Implementation

Raw data was analyzed with the Pincho pipeline with the following configurations: SRA accession numbers were used to download data from the SRA database via fasterqdump followed by whitespace removal and compression. Leading and lagging low quality base removal was performed via Trimmomatic, followed by error correction by Rcorrector. Transcriptomes were assembled via Trans-ABySS, BinPacker, IDBA-tran, Shannon Cpp, rnaSPAdes, TransLig, Trinity, MEGAHIT (positive control) and Tadpole (negative control) with adaptive k-mer control enabled. Adaptive k-mer control utilizes a minimum k-mer of k21 and four k-mers generated based on their respective maximum insert length and middle three quartiles between k21 and the maximum. Consensus assembly generation was conducted via TransRate. Read mapping was performed via HISAT2 aligner, presence of ancestral genes was identified by BUSCO and n50/n90 were calculated via TransRate. Assessment was conducted in combinations between the nine assemblers individually and in groups of two and three. Oases was not utilized in this study due to the frequent unresolved bugs associated with the software and its lack of maintenance (last major update 20 May 2013). SPAdes and ABySS de novo genome assemblers were not utilized in this study as we used their transcriptomic counterparts designed for transcriptome assembly. Both rnaSPAdes and ABySS were demonstrated to outperform SPAdes and ABySS, respectively [3,27].

1.2.4 K-mer Size Determination

K-mer sizes were left to their default values (Table 1) if the assembler only allowed one k-mer size as input and assembler runtime was extensive. Therefore, default k-mers were used for BinPacker, TransLig, Trinity and Shannon Cpp. Assemblers that allowed the

selection of multiple k-mer sizes and/or were time efficient were assigned a broad range of five k-mer sizes.

1.2.5 Assessment Validation

We utilized three metrics (TransRate, BUSCO, and HISAT2) that best represent the quality of a de novo transcriptome. TransRate provides the n50/n90 statistic, among others, which is the largest contig size where 50%/90% of bases are contained in transcripts of this length. These n50/n90 scores are often used to ascertain the quality of a reconstruction, with longer n50/n90 lengths correlating to a more complete assembly. Other assessment metrics include complete BUSCO scores representing percent ancestral transcripts present and HISAT20s overall alignment score which is the percentage of raw data utilized within reconstructions. For our workflow, we used BUSCO's Eukaryota dataset as a reference.

Respective assessment scores were judged per assembler as greater than MEGAHIT's assessment scores or less than MEGAHIT's assessment scores. Assessment scores (AS) greater than MEGAHIT were subjected to the following formula:

$$\frac{AS_x}{AS_{max}} * 0.5$$

AS less than MEGAHIT were processed under a different formula to calculate underperformance:

$$\frac{AS_x}{-AS_{min}} * 0.5$$

while scores equal to MEGAHIT were counted as 0. Average assessment scores (AAS) were calculated as the average of HISAT20s overall alignment, complete BUSCO score, and TransRate's n50/n90 scores in a 1:1:1 ratio, so n50 and n90 scores were averaged

together before averaging with the other two assessment scores. Finally, the AAS were normalized between the numbers of 0.5 as overperforming versus MEGAHIT and −0.5 as underperforming.

1.3 RESULTS

1.3.1 Workflow Installation, System Build, and Performance

Pincho is packaged with an installer script written in Python3 and Bash which will install and configure required dependencies in Linux Ubuntu systems. Our workflow requires a minimum of 24 threads and 128GB of memory to run efficiently and is largely GPU independent. It is recommended to scale performance parameters evenly if higher performance is desired (i.e., 24:128 ratio). Our study was conducted on two new workstations including: AMD Ryzen 9 3900X 3.8GHz processor, G.Skill 128GB 4 × 32 D4 3200 memory modules, and an ASUS TUF GAMING X570-PLUS motherboard. An alternative replica build would be to purchase a PowerSpec G464 and upgrade the memory modules to a total of 128GB (net price 2200 USD). Our test data ranged in both number of bases and file size (Table 1) to provide an accurate depiction of the capacities of our workflow performance. We encountered no errors conducting the study with the parameters stated above. Methods can be easily replicated via Pincho's completely modular user interface.

1.3.2 Average Assessment Score Generation

We utilize three distinct assessment software—-HISAT2, TransRate, and BUSCO—-to derive raw scores for each single-, bi-, and tri- assembly run (see assessment validation in methods) and mark their over/underperformance in regard to a MEGAHIT single assembly run. Individual metric scores are normalized to a scale between −0.5 and 0.5, where 0 is equal to a MEGAHIT single assembly run assessment score. Negative integers denote underperformance and positive integers denote overperformance when compared to MEGAHIT genome assembler. Individual assessment scores are then

averaged together respectively to provide an AAS per assembler or assembler group. The following assemblers were utilized in this study: Trans-ABySS, BinPacker, IDBA-tran, Shannon Cpp, rnaSPAdes, TransLig, Trinity, MEGAHIT, and Tadpole.

1.3.3 Single-Assembly

According to our combination of assessment software criteria, rnaSPAdes outperformed all other assemblers with an AAS of 0.23, followed by Trans-ABySS (AAS: 0.18), TransLig (AAS: 0.17), IDBA-tran (AAS: 0.02), BinPacker (AAS: 0.02; Figure 2), and the MEGAHIT single-assembly baseline (AAS: 0). Shannon Cpp (AAS: −0.03), Trinity (AAS: −0.24), and Tadpole (AAS: −0.50) underperformed relative to the baseline (Figure 2). Runtime analysis highlights no correlation between total time consumption and performance, as assemblers that required the most time did not produce the best assemblies nor vice versa (Supplementary Figure S2). Assessment of raw data from our assessment software reveals rnaSPAdes and Trans-ABySS obtained the highest HISAT2 scores (>92%), rnaSPAdes and IDBA-tran scored the highest complete BUSCO scores (>199 complete eukaryotic ancestral transcripts), and TransLig and BinPacker contained the longest n50/n90 lengths (>1766 bp/>499 bp; Figure 3). Alternatively, IDBA-tran and BinPacker obtained the lowest HISAT2 scores (<85%), Trinity and Tadpole scored the lowest complete BUSCO scores (<169 complete transcripts) and also the shortest n50/n90 lengths (<1021 bp/<286 bp; Figure 3).

1.3.4 Bi-Assembly

The pairing of assemblers often increased the AAS; however, our negative control Tadpole caused a decrease in metric scores of our previous top three single-assemblers: rnaSPAdes (Net ΔAAS: −0.06), Trans-ABySS (Net ΔAAS: −0.13), and TransLig (Net

ΔAAS: −0.18; Figure 4). The combination of TransLig and rnaSPAdes outperformed all other single- and bi-assembly combinations achieving an AAS of 0.45 (Figures 2 and 4). Pairings between Trans-ABySS and rnaSPAdes achieved the second highest AAS of 0.42 (Figure 4). Bi-assemblies involving combinations between Tadpole and either Trinity, MEGAHIT, Shannon Cpp, Binpacker, or TransLig all underperformed when compared to a MEGAHIT single-assembly run (Figure 4).

1.3.5 Tri-Assembly

We observed the highest possible AAS of 0.50 in a tri-assembly approach containing Trans-ABySS, rnaSPAdes and TransLig (Figure 5). The higher AAS values are primarily located in the highest performing assembler groups: Trans-ABySS, rnaSPAdes, and TransLig (Figure 5). The lower AAS values are found not only in the negative control Tadpole, but in Trinity and Shannon Cpp as well. The rnaSPAdes bracket performed the best, yielding the highest AAS, while the Tadpole bracket performed the lowest, yielding the lowest AAS on average (Figure 5). The rnaSPAdes bracket also exhibited a smaller distribution of AAS, spanning 0.22 to 0.50, with a higher frequency of high AAS than other assembler groups (Figure 6). When tri-assembly runs are sorted from lowest AAS to highest, the rnaSPAdes group continues to lead the other tri-assembly groups at every datapoint (Supplementary Figure S3). Signs of over/underperformance amongst tri-assembly runs were observed, with Tadpole, Trinity, and Shannon Cpp tri-assembly approach underperforming by scoring equal to the MEGAHIT baseline previously set at 0 (Figure 5).

1.4 DISCUSSION

1.4.1 Single-Assembly Mode

Single-assembler comparisons yielded interesting results regarding the efficiency of de novo transcriptome assemblers compared to their genomic counterpart. Genome assemblers are known for their conservative style for reconstructions, whereas transcriptome assemblers take risks to assemble every transcript isoform identified. It is largely due to this deviation between the two software that we see gains or losses in average assessment scores. To further elaborate, isoforms are more common among longer transcripts as there is more genomic material, increasing the probability of the accumulation of mutations and change over time. As longer transcript isoforms are added to the assembly, the n50/n90 lengths increase. It also helps generate more ancestral transcripts, as each variant has a chance to align to the reference eukaryotic database of ancestral transcripts. Finally, more raw data containing variant fragments will be incorporated into the product, resulting in a higher HISAT2 alignment score. In summary, de novo transcriptome assemblers should be more than capable of outperforming de novo genome assemblers in part due to the identification and reconstruction of isoforms, which is why it is so bizarre to observe some transcriptome assemblers unable to outperform genome assemblers (i.e., MEGAHIT).

1.4.2 Trinity

Perhaps the most perplexing of all our results was the tendency for Trinity to underperform, as it has long been described in literature to be quite robust at de novo transcriptome assembly. Trinity incorporated roughly 89% of raw data into its assembly, which is average among assemblers tested (Figure 3). Trinity's n50/n90 scores, however,

were roughly half of what TransLig, a non-adaptive k-mer assembler, produced. The short n50/n90 lead us to believe that Trinity may be unable to bridge fragmented transcripts as well as other assemblers. Upon examination of the fragmented BUSCO scores, we observe that Trinity in fact did fragment more transcripts than other assemblers.

1.4.3 Shannon Cpp

Reconstructions produced by Shannon Cpp were fairly close to MEGAHIT's AAS. Shannon Cpp exhibited a higher n50 score than MEGAHIT, but a lower n90 score. Shannon Cpp tended to fragment more ancestral genes than MEGAHIT and that higher rate of fragmentation may account for the lower n90 score. Shannon Cpp had a higher complete BUSCO score than MEGAHIT; however, Shannon Cpp utilized roughly 1% less of the raw NGS datasets than MEGAHIT did, leading to a lower average assessment score total. Shannon Cpp utilizes a type of information theory algorithm built on a de Bruijn graph and we speculate that this algorithm is more conservative than MEGAHITs more normalized de Bruijn graph method, leading to more fragments and less raw NGS integration.

1.4.4 Tadpole and MEGAHIT

Tadpole, as a basic assembly tool, is not as complex as the other assemblers and tends to create many problematic reconstructions (i.e., chimeras, nonsense repeat sequences, etc.). This is evident as Tadpole scored the lowest in every assessment metric, except for raw NGS data incorporation. Further exploration of Tadpole assemblies reveals obvious misassemblies. This was known before the study and is why we chose Tadpole as a negative control: a metric to use as an indication of poor assembly methodology. In addition, we have included what we perceive to be the best genome assembler, MEGAHIT

(as a single assembler), to act as a baseline for transcriptome assemblies. Genome assemblers tend to be more conservative with their reconstructions and therefore will score moderately well according to assessment software metrics; however, they do not account for isoforms which account for large portions of transcriptomes. This allows for transcriptome assemblers to elevate themselves from the MEGAHIT baseline by providing isoform assemblies that are of high quality to increase their metric scores higher than that of a genome assembler.

1.4.5 BinPacker

With the second highest n50/n90 scores and decent complete BUSCOs, BinPacker ranks among the top assemblers, but on average, BinPacker's performance is only slightly better than MEGAHIT's. BinPacker was poor at integrating the raw NGS data into the completed reconstruction, scoring among the bottom two in the HISAT2 assessment bracket. Data integration depends on the quality of the raw reads, but also whether the algorithm designed for the assembler was able to incorporate that read within the assembly. BinPacker underperformed in raw NGS data incorporation; however, TransLig, the sequel to BinPacker, improves on this flaw.

1.4.6 IDBA-Tran

Suffering from the same issue as BinPacker, IDBA-tran's low raw NGS data utilization rate detracts from its impressive complete BUSCO score and decent n50/n90 metrics. Fortunately, its two strengths can carry IDBA-tran over the MEGAHIT baseline, providing evidence that IDBA-tran provides reconstructions of better quality than a genome assembler. An oddity is IDBA-tran's tendency to duplicate BUSCOs, which may

be caused by the addition of five k-mer sizes and the inability for IDBA-tran to reduce redundancy among the assembly.

1.4.7 TranLig, Trans-ABySS, and rnaSPAdes

The top three de novo assemblers are rnaSPAdes, Trans-ABySS, and TransLig. In single assembly comparisons these de novo transcriptome assemblers were able to largely outperform the other assemblers in various assessment metrics. Complete BUSCOs and raw data utilization rates for rnaSPAdes were both part of the top two metric scores, so it is no surprise rnaSPAdes scored the best among the three. rnaSPAdes was also able to produce one of the least redundant assemblies. Trans-ABySS incorporated the most NGS data into its assembly procedure but was not able to reconstruct as many transcripts as rnaSPAdes nor TransLig. TransLig outperformed all assemblers in n50/n90 scores, however its raw NGS data utilization was lacking. It is clear from the investigated assessment metrics that each of these assemblers excel in one area or another, which is precisely why multi-assembly provides higher quality transcriptomes.

1.4.8 Bi-Assembly Mode

Bi-assembly methods, including Tadpole, led to lower overall assessment scores when compared to pairings without Tadpole, cementing the validity of our negative control. On average, all pairings excluding Tadpole achieved scores greater than the MEGAHIT baseline and greater AAS when compared to the single assembly approach. Bi-assembly increased n50/90 scores, complete BUSCOs and overall alignment scores from their single assembly counterparts on average. All increases are expected as we nearly double the coverage, while including several transcripts that were missed in the single assembler methodology. We note a significant increase in AAS from single- to bi-assembly

approaches across all assemblers. Lastly, we note rnaSPAdes produced the top three bi-assembly reconstructions, providing further evidence of the positive synergistic effects of our top single-assemblers.

1.4.9 Multi-Assembly Mode

We have demonstrated the potential of utilizing the multi-assembler approach to elevate the overall quality of reconstructions across three distinct assessment criteria. We highly recommend the usage of Trans-ABySS, rnaSPAdes, and TransLig in combination for de novo transcriptome assembly as they provided the highest metric scores. We observe the highest single-assembly AAS at 0.23, highest bi-assembly at 0.45 and the highest tri-assembly at the maximum 0.50 demonstrating that assemblers can not only synergize well together, but also that bi-assembly increased the quality of single-assembly by a large margin. We observe a significant increase in scores from bi- to tri-assembly as well. Although rnaSPAdes observed no significant change in average scores, there was still an increase in the number of novel transcripts recorded (via BUSCO) and this metric alone is worth the addition of a third assembler. We advocate for the usage of multi-assembler workflows as they provide the best chances of complete assemblies for non-model organisms.

1.5 CONCLUSION AND FUTURE DIRECTIONS

Over the past ten years, researchers have provided us with an extensive coverage of the strengths and weaknesses of the various de novo transcriptome assemblers in single-assembly approaches. However, there have been scarce publications to date offering a comprehensive comparison between multi-assembly approaches. We offer a broad comparative review of seven well-maintained de novo transcriptome assemblers and two de novo genome assemblers scored via three distinct assessment criteria. All our work was completed via a modular pipeline, Pincho, which we contribute to the scientific community as a modern modular de novo transcriptomic workflow written in Python3 for Ubuntu 20.04 Focal Fossa LTS on our GitHub page (https://github.com/RandyOrtiz/Pincho).

As assemblers continue to improve, we intend to continue updating our workflow with the latest versions released. As new assemblers arise, they can be seamlessly integrated into the Pincho workflow. We also seek to create several versions of our software that is cross-platform and low resource editions. It is our vision to have one version of Pincho that can work on any computer or laptop and perhaps eventually completely online without the use of local computer resources.

1.6 TABLES AND FIGURES

1.6.1 TABLES

Table 1. De novo Assemblers Utilized in Pincho.

Assembler	Genome or Transcriptome	K-mer Used	K-mer Default	Algorithm	Version	Version Release	Software Release	Cited by [2]	Datasets Explored
ABySS	Genome	Adaptive	32	*de Bruijn* Graph	v2.2.4	1/30/2020	11/26/2008	3481	Human
BinPacker	Transcriptome	25	25	Splice Graph	v1.0	10/17/2019	3/19/2015	95	Human, Mouse, Dog
IDBA-tran	Transcriptome	Adaptive	20, 30, 40, 50	*de Bruijn* Graph	v1.1.3	6/11/2016	6/19/2013	155	*Oryza sativa*
MEGAHIT	Genome	Adaptive	21, 41, 61, 81, 99	*de Bruijn* Graph	v1.2.9	10/14/2019	9/25/2014	1738	Soil
Oases/Velvet	Transcriptome	Adaptive	19, 21, 27, 31, 35	*de Bruijn* Graph	v0.2.08/ v1.2.10	05-20-2013/10-17-2013	12-11-2011/11-16-2007	1437	Human, Mouse
rnaSPAdes	Transcriptome	Adaptive	Automated k-mers	*de Bruijn* Graph	v3.14.1	5/2/2020	11/16/2018	122	Humans, Mouse, Corn, *Arabidopsis*
Shannon Cpp	Transcriptome	25	25	*de Bruijn* Graph	v0.4.0	12/19/2019	2/9/2016	27	Human
SPAdes	Genome	Adaptive	21, 33, 55	*de Bruijn* Graph	v3.14.1	5/2/2020	5/7/2012	12635	*Escherichia coli, Deltaproteobacteria*
Tadpole	Genome	Adaptive	31	Simple Kmer Code	v38.86	6/13/2020	1/9/2012	437	Fungus, Bacteria, Plant, Soil
Trans-ABySS	Transcriptome	Adaptive	32	*de Bruijn* Graph	v2.0.1	2/19/2018	6/18/2010	467	Human
TransLig	Transcriptome	31	31	Line Graph Iterations	v1.3	10/26/2019	11/23/2018	7	Human, Mouse
Trinity [1]	Transcriptome	25	25	*de Bruijn* Graph	v2.11.0	6/30/2020	12/3/2010	1175	*Drosophila melanogaster*

[1] genome guided mode available. [2] cited by column updated on 15 June 2021.

Table 2. Test NGS Dataset from NCBI SRA database.

Species	Accession	BUSCOs (%) [1]	Reads (M)	Bases (G)	Read Length (bp)	File Size (Mb)	Tissue
Allobates femoralis	SRR8288062	62.4	3.5	0.8	120	504.4	Skin
Amazophrynella minuta	SRR8288029	70.6	4.4	1.1	120	641.6	Skin
Dendrobates auratus	ERR3155280	91.0	3.3	1.9	294	1000.0	Skin
Dendrobates imitator	ERR3169394	66.3	16.3	1.6	50	782.5	Skin
Dendrobates sirensis	SRR8288043	72.2	4.9	1.2	120	710.7	Skin
Lithobates catesbeianus	SRR4048903	77.6	6.8	1.3	99	558.0	OB [2]
Pyxicephalus adspersus	SRR6890710	87.8	10.0	1.5	75	538.8	Testis
Scinax ruber	SRR8288044	73.7	5.8	1.4	120	840.1	Skin

[1] Complete BUSCO using Pincho's rapid assessment at default settings [2] Olfactory Bulb.

1.6.2 FIGURES AND FIGURE LEGENDS

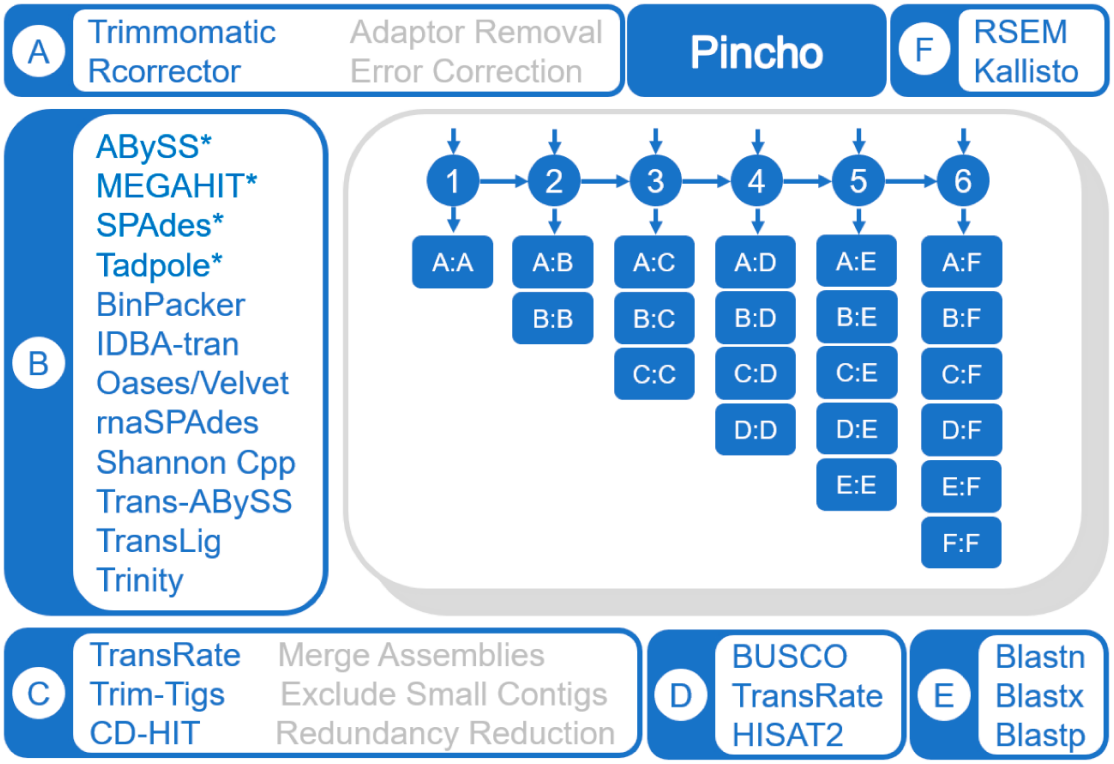

Figure 1. Pincho Management Workflow.

Software installed in the Pincho workflow v0.1, including (**A**) pre-processing, (**B**) transcriptome and * genome assemblers, (**C**) post-processing, (**D**) assessment software, (**E**) annotation software, and (**F**) expression analysis software. Modules may begin at any position (**A**–**F**) but must then process sequentially (i.e., **B**, **C**, **D**…). Possible avenues depicted in shorthand, where A:D represents steps **A**, **B**, **C** and **D**. Any number of items may be called from each module (i.e., module B: IDBA-tran, Trans-ABySS, Trinity = 3 items called from module B).

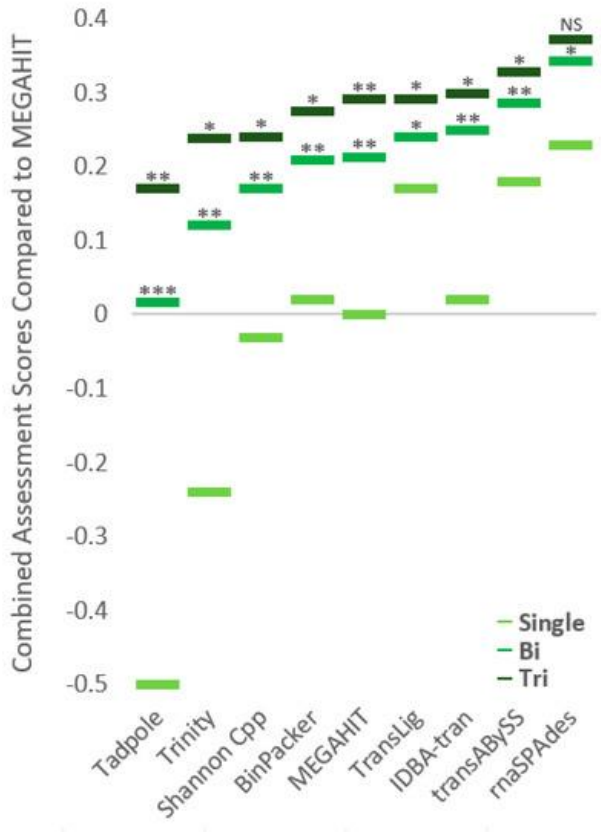

Figure 2. Single-, bi-, and tri-assembly assessment score averages.

Average assessment scores from single-, bi-, and tri-assembly runs compared to MEGAHIT single-assembly as a baseline score (0). Scores lower than 0 underperformed when compared to MEGAHIT single-assembly, whereas, scores higher than 0 overperformed. Average assessment scores calculated by the average of HISAT2 overall alignment, BUSCO complete score, and TransRate n50 and n90 metrics averaged across all files processed. Assemblers utilized are included in the x-axis to denote both their average scores for single assembly and their average scores as part of a pair of two or three. Two tailed paired T-tests were conducted between single-assembly and bi-assembly, and between bi-assembly and tri-assembly. P-values are noted between single- and bi-assembly combinations and between bi- and tri-assembly combinations. All comparisons conform to $p < 0.05$ except for no-significance noted between bi- and tri-assembly associated with rnaSPAdes. *** is $p < 0.00001$, ** is $p < 0.001$, * is $p < 0.05$, and NS (No Significance) is $p > 0.5$. *p*-values are under FDR (False Discovery Rate) correction.

Single Assembly Raw Assessment Scores

Assembly	rnaSPAdes	trans-ABySS	TransLig	IDBA-tran	BinPacker	Shannon	Trinity	MEGAHIT	Tadpole
n90	317.8	309.9	624.0	433.8	499.5	312.6	285.5	363.9	239.5
n50	1200.8	1410.9	2043.8	1465.4	1766.4	1323.8	1020.1	1152.6	708.4
Complete BUSCOs	199.3	191.5	195.1	201.1	195.8	196.5	168.1	193.9	141.5
Duplicated BUSCOs	10.4	51.5	44.1	199.0	44.0	79.1	19.3	4.5	102.0
Fragmented BUSCOs	35.9	38.9	28.8	33.3	34.0	38.4	54.0	34.5	74.4
Missing BUSCOs	19.9	24.6	31.1	20.6	25.3	20.1	32.9	26.6	39.1
Overall Alignment Rate	0.93	0.94	0.87	0.82	0.84	0.89	0.89	0.90	0.88

Top 2 Metric Scores | Bottom 2 Metric Scores | Metrics Utilized in Study

Figure 3. Single-assembly raw average assessment scores.

Assessment metrics used in study: n50/n90 (via TransRate), complete BUSCOs (via BUSCO), and overall alignment rate (via HISAT2) are boxed in. Top two metric scores per assessment criteria are highlighted in green. Bottom two metric scores are highlighted in pink. Metrics not boxed in were provided to aid discussion but not for the generation of the average assessment scores.

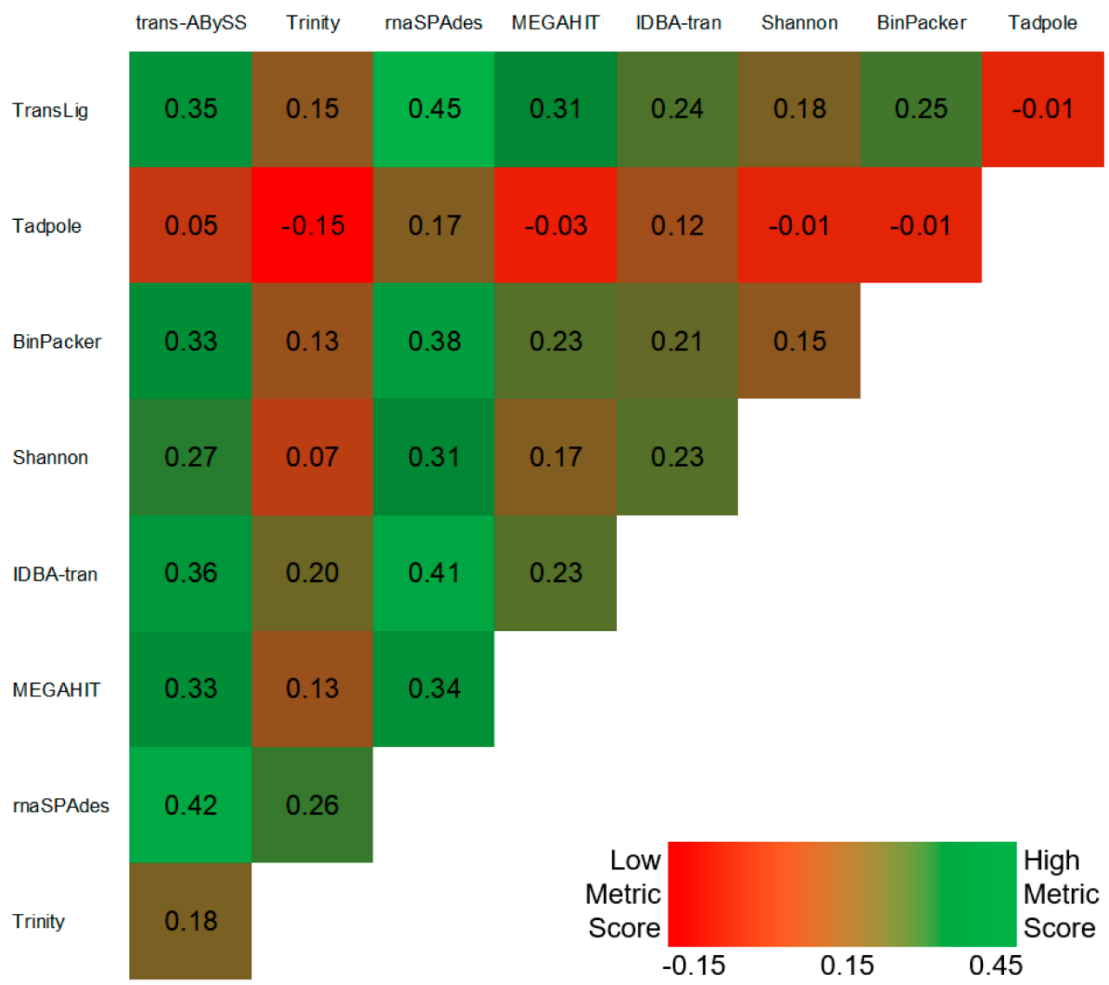

Figure 4. Bi-assembly assessment scores.

Heatmap of bi-assembly assessment scores from 36 combinations of 9 assemblers compared to MEGAHIT single-assembly as a baseline score (0). Scores lower than 0 underperformed when compared to MEGAHIT single-assembly, whereas, scores higher than 0 overperformed. Green denotes a higher assessment score and red denotes a lower assessment score among the 36 bi-assembly groups. Shannon denotes Shannon Cpp version.

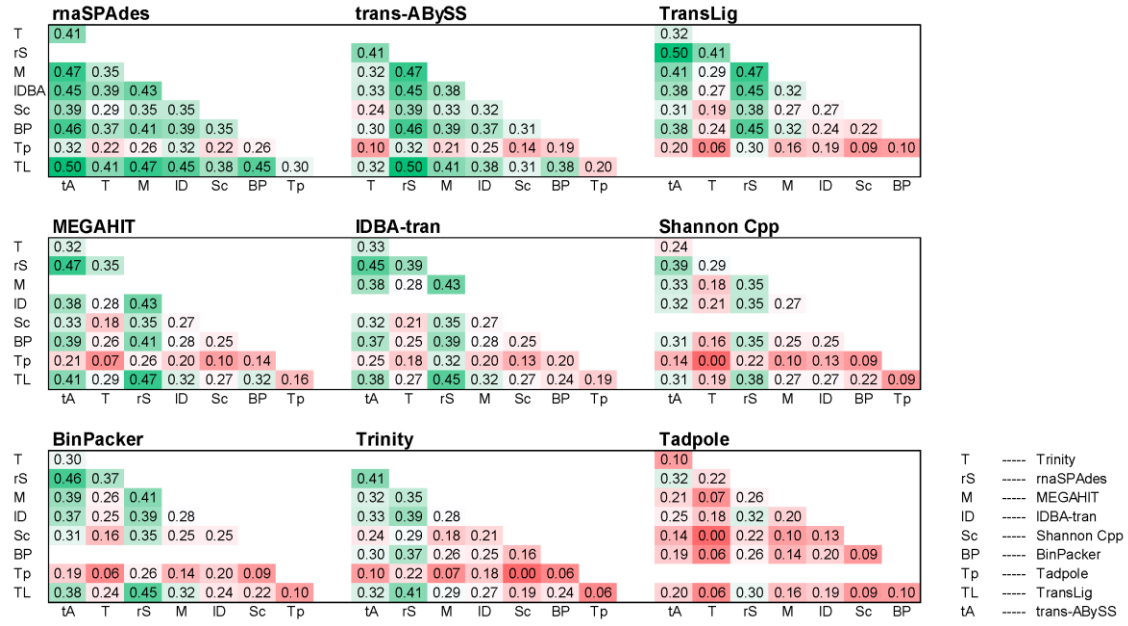

Figure 5. Tri-assembly Scores.

Tri-assembly assessment score results from 84 combinations of 9 assemblers, respectively. All assembler metrics are compared to over/underperformance to the average MEGAHIT single-assembly score. Highlighted values range from high average assembly scores up to 0.5 (green) to low average assessment scores down to 0.0 (red). Metric scores are consistent, with previous figures allowing for cross comparison.

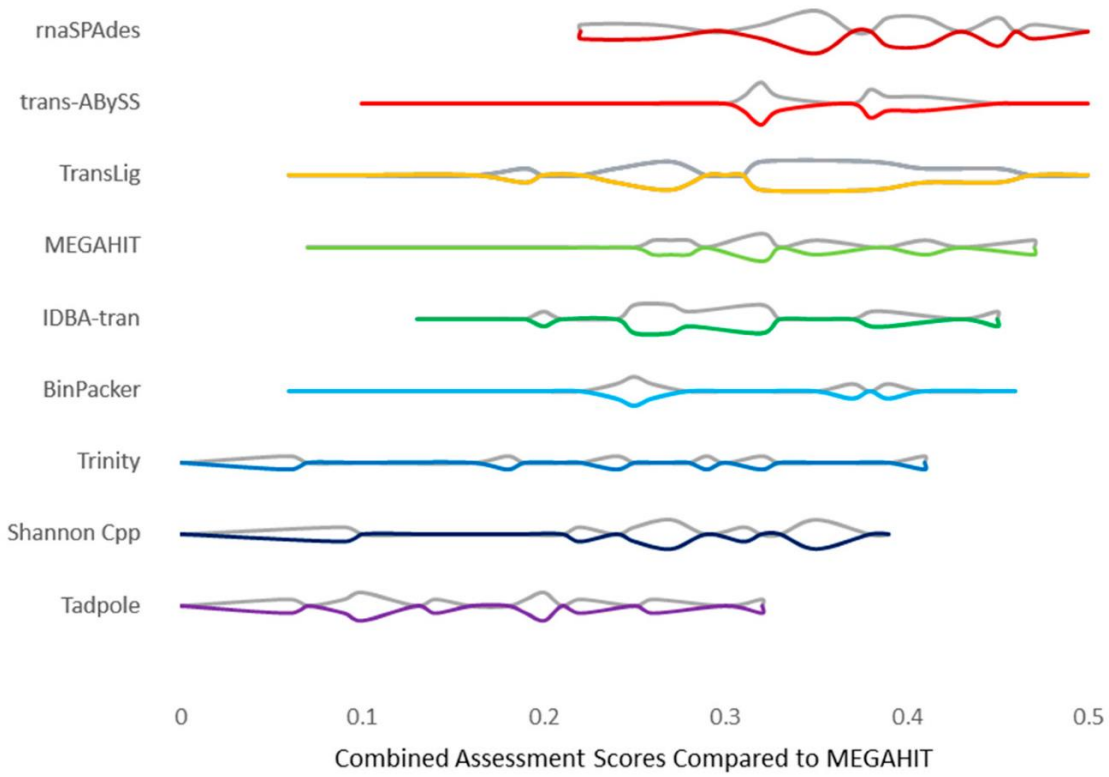

Figure 6. Tri-assembly score distributions.

Violin plots representing assessment score frequency and distribution among tri-assembly runs. All assembler metrics are compared to over/underperformance to the average MEGAHIT single-assembly score (0). All tri-assembly scores performed equal to or greater than the baseline. Higher quality assembler combinations are represented via higher numerical scores up to a maximum of 0.5.

CHAPTER 2

Comparative skin transcriptomics of the invasive versus native *Eleutherodactylus coqui*: Stress genes, parasite signatures, and a novel antimicrobial peptide

2.1 INTRODUCTION

Invasive species contribute to declines in biodiversity worldwide, especially in habitats already under stress by climate change and human intervention (e.g., deforestation, agriculture, and urbanization). Most amphibians are impacted by such stresses, yet few thrive and become invasive, including a few Caribbean frogs of the genus *Eleutherodactylus*. The most prominent *Eleutherodactylus* is the *E. coqui* or common *coqui*, which is an iconic, cultural, and beloved symbol of Puerto Rico (PR). *E. coqui* are recognized by its characteristic mating call "CO-KEE." Given the prolific nature of common *coquis*, high population densities are commonplace, including a record for highest terrestrial vertebrate population density of ~20,000 individuals ha^{-1} in its native region [58]. Such large densities are also reported where common *coquis* are invasive, and their chorus call is considered a nuisance.

The history of the common *coqui* as an invasive species can be traced to the unregulated greenhouse-plant trade from PR, mostly during the 1980's. Since then, *E. coqui* has expanded its distribution beyond its natural range as an invasive species first as stowaways and then forming established populations into the southeastern US, throughout the Caribbean, Hawaii, and many Pacific islands. To date, most introductions

of *E. coqui* are unintentional, but few were the result of deliberate release of adult individuals such is the case of Costa Rica in 1998 [59]. Because of common *coqui* seemingly ubiquitous nature as well as its remarkable adaptability and resilience, this species has been the focus of radical and expensive programs of eradication where it is not native. Likewise, this unfavorable attention has resulted in significant basic research on this species in contrast to the other 202 congenerics within the Eleutherodactylidae family. Such studies cover many aspects of its ecology, behavior, development, physiology, and taxonomy; yet over the last 20 years this species is a new focus of molecular systematics, population genetics, and genomics research. We want to contribute to this effort by focusing on comparative transcriptomics and how it could help us understand how *E. coqui* differs between natural and invasive populations; and how genomic/transcriptomic tools can bring light into this remarkable species' resilience against pathogens.

At the generic level, *Eleutherodactylus* is the most numerous of the extensively diverse Eleutherodactylidae family of New World frogs with 203 species described as of 2022 [60]. In Puerto Rico, 16 species have been described, but only *E. coqui* is a widespread invasive outside this territory [61]. Currently, of the other PR *coqui* species, three are thought to be extinct, three are critically endangered, five are endangered, one is vulnerable, and the remaining three are considered least concern [62]. As other members of this genus, *E. coqui* is a direct-developer and its females deposit their eggs farther from larger bodies of water, in locations that have little moisture, like leaf litter [63-64]. This reproductive strategy has been hypothesized to help dispersal to new habitats and provide resiliency against deadly pathogens such as *Batrachochytrium dendrobatidis*

(*Bd*) fungus, the known-causal agent of chytridiomycosis [65]. For invasiveness, the direct development of *E. coqui* might provide a direct advantage in spreading to new humid habitats away from bodies of water, which other species may not be privy to [66]. Likewise, as common *coqui* and other direct-developer species complete their growth in the egg, they do not need to enter an aquatic environment where numerous aquatic predators thrive and feast on eggs and tadpoles; this furthers *Eleutherodactylus* possibility of survival and establishment in new habitats [67-68]. Finally, the small size of the concealed egg, sheltered away from water, allows common *coqui* eggs to be hidden in agricultural and horticultural products that are shipped outside PR and provide a constant stream of recolonization to locations both within the US (e.g., Florida, California, and Hawaii) and outside the US (mainly Central America) [69].

Ecological and invasion biology studies have shown that common *coquis* are extremely tolerant to habitat modification as well as prolific in pristine and humid primary forest (e.g., found in the Yunque national park in PR). This resilience has made the common *coqui* a habitat generalist, but it needs sites to retreat to in times of unfavorable weather as evidenced in invasive individuals in Hawaii [70]. Sites of retreat include areas where the *coqui* can conceal themselves from the sun, prevent desiccation, and to find warm areas in times of cold which may be more common in forests than in highly urbanized zones [71]. Such shelter requirements somehow have prevented the common *coquis* widespread dispersal to most US states with well-defined four seasons and freezing conditions during winter.

Natural history accounts are ample for the common *coqui* if compared to other *Eleutherodactylus*. *E. coqui* has a generalist diet and primarily feeds on diverse leaf litter

invertebrates. In their native habitat of Puerto Rico, these frogs feed on ants (Hymenoptera), crickets (Orthoptera), and weevils (Coleoptera) that account for >60% of their volumetric diet [72]. This remarkable diverse dietary pattern is also similar to those where *coqui* is invasive, such as in Hawaii, where these frogs feed on ants and crustaceans (e.g., amphipods and isopods) that account for ~60% of their diet [73]. Given the generalist diet, *E. coqui* seems to adapt easily to other habitats well outside their natural range and they may have many devastating effects when these frogs have achieved high densities. For example, invasive *E. coqui* in some locations of Hawaii have reached 2-3 times higher densities than those recorded in their natural habitat of PR, aided by the absence of native anurans in Hawaii [73]. In Hawaii, the impact of the introduced *coquis* is just beginning to be determined and these frogs seem to affect the nutrient dynamics, local arthropod diversity, and might cause an increase nutrient availability on the forest floor; which further influences floral and faunal compositions to favor the establishment of other non-native plants and invertebrate invaders [69; 74]. Therefore, the ecological impact of *E. coqui* also extends across trophic chains into the alteration of the naturally occurring nutrient cycle dynamics [59].

Most species have natural enemies including predators, parasites, competitors, and pathogens that keep their population numbers under control in their native environments. Few natural enemies of *E. coqui* have been reported in its native habitat but include the Puerto Rican racer snake (*Alsophis portoricensis*) and three avian species: the Puerto Rican screech owl (*Gymnasio nudipes*), the pearly-eyed thrasher (*Margarops fuscatus*), and the red legged thrush (*Turdus plumbeus*) [75]. These enemies are absent for invasive *coqui* in Hawaii and its only known vertebrate predator is another invasive

species, the Javan mongoose (*Herpestes javanicus*) [76]. Some arthropods are also predators and include giant crab spiders *Olios* (Sparassidae) and other huntsman spiders in both invasive and native habitats [77].

For parasite microorganisms, *E. coqui* are known to carry *Bd* fungus, as these frogs are resilient to it, which also means that they are a vector for spreading it to susceptible amphibian species [78]. Prior studies credit the success of *E. coqui* invasive expansion to the scarcity of parasites in these new regions that these frogs are vulnerable to [79]. Furthermore, introduction of these small frogs to Hawaii has been directly correlated with greater abundances of non-native avian species, as the frogs outcompete the native avian species Apapane (*Himatione sanguinea*), Hawaii ‘Amakihi (*Chlorodrepanis virens*), and Hawaii ‘Elepaio (*Chasiempis sandwichensis*) for invertebrate prey and serve as a food source for different birds that consume small vertebrates [80].

Given the ubiquitous and significant economic impact of common *coqui* frogs, molecular studies on this species are extensive if compared to other *Eleutherodactylus* and includes population genetics, phylogenetic relationships, microbiome, proteomics, and a draft genome. Phylogenetic studies on *E. coqui* from derived mitochondrial DNA (mtDNA) and nuclear loci revealed a significant genetic divergence among allopatric populations in Puerto Rico likely during late Pliocene at ~2.4 MYA associated development of xeric environments across the Caribbean [81]. This molecular systematics information helped to reveal the relatively low genetic diversity of Hawaii’s invasive populations, which were later complemented with population genetics based on microsatellites [82]. Further studies confirmed this low genetic diversity of Hawaiian

invasive *E. coqui* by analyzing 15 invasive and 13 native PR populations with 6-9 polymorphic microsatellite loci. Even though significantly lower genetic diversity was found in Hawaiian invasives, two separate introductions of common *coqui* are supported (i.e., on the islands of Hawaii and another on Maui) and both from source populations around San Juan city in PR where ornamental plants are grown. Interpopulation comparisons also revealed extensive mixing among Hawaiian populations likely facilitated by humans. The native sources of common *coqui* invaders to other US states (e.g., Los Angeles-California, LA-CA; Texas, and Florida) are yet to be determined. In this study, we explore the origin of the LA-CA invasion, and we further confirm that common *coquis* are successful invaders despite the loss of genetic variation.

During the last decade, next-generation sequencing technologies increased our understanding of the molecular basis of invasiveness and other associated adaptive traits such as the presence of antimicrobial peptides (AMPs). Many research articles that utilize skin swabs on *E. coqui* have explored either their health, microbiome, slime, or skin cell content, but there are almost no research articles that go further than the skin's surface. These samples have been used to assess antimicrobial defenses such as the research utilizing matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry on native *E. coqui*. Using skin swab samples, the authors were unable to detect any AMPs and concluded that the common *coqui* lacks these powerful innate defenses such those against *Bd* [78]. However, *E. coqui* is resilient against *Bd* and other mechanisms might be in action.

The most important molecular work in *E. coqui* is probably the public draft genome of *E. coqui* spanning all 13 chromosomes and a total of 2789.35 Mb (NCBI

UCB_Ecoq_1.0 accession number GCA_019857665.1). This was obtained from an invasive adult male (HN-11) collected in Hawaii and derived from kidney and liver tissues. The chromosomes of *E. coqui* were found to be highly conserved with other frogs including *Engystomops pustulosus*, *Hymenochirus boettgeri*, *Leptobrachium ailaonicum*, *Pyxicephalus adspersus*, *Xenopus laevis*, and *X. tropicalis* [85]. However, *E. coqui* gene sequences showed signs of divergence despite limited karyotype variation [85]. As most first drafts, *E.* coqui genome has 8% missing nucleotide sequences (as Ns) which account for over 230+ Mb. Some of such sequences might be in the 105,220 unplaced scaffolds that account for ~1007 Mb.

While a genome provides us with the identity and sequence structure of the genes present in an organism, it provides very little information about gene expression. In contrast, transcriptomes provide us with gene expression data that can help provide molecular answers to our ecological questions; however, transcriptomes are each a glimpse into a specific moment in time for an organism and not all genes are always produced. Transcriptomes also provide data on ncRNA sequences and transcript isoforms. In 2019, a PhD dissertation utilizing a whole-body transcriptome was published on the endocrine-mediated development of *E. coqui*; however, it does not provide us with an adequate transcriptome assembly for comparisons or any invasive-favoring genes such as antimicrobial peptides [83]. In this work and subsequent publications, *E. coqui* was highlighted as a great model for studying developmental biology (i.e., direct development), neuroethology (mating behavior), and invasion biology [84]. Therefore, more transcriptome datasets are needed to provide us with a better understanding of gene expression under the specific conditions such as invasive versus native populations. By

including transcriptomic work in conjunction with a whole genome, we must expect a better understating of how the common *coqui* has become such a successful invader and help to develop better tools to control this species spread using haplotype-specific tools rather that current "blunt" eradication approaches such as citric acid sprays.

Our goal is to present a comparative transcriptome analysis of total skin tissues of native and invasive *E. coqui* populations that highlights the trade-offs made by this species as it enters a new environment. Our transcriptome analyses try to bridge this gap using transcriptomic comparisons and try to identify differentially expressed genes between native *E. coqui* from Puerto Rico and an invasive population established in LA-CA (Los Angeles, California). We also elucidate some of the molecular mechanisms that contribute to this species' on-going successful expansion. Transcriptome data from both common *coqui* populations also helped us to describe how parasite genetic signatures were revealed in a native population, which are absent in the invasive one. These results suggest that native *E. coqui* populations have parasites that might be key in keeping natural population in check while invasives are present in a mostly innocuous habitats that might favor heightened reproduction and distribution. From native populations, we can report a new type of AMP present only on *E. coqui* that was confirmed with both genome and transcriptome data. Therefore, *E. coqui* might harbor more defenses at the molecular level than reported. Finally, we also summarized the life history of *E. coqui* as an invasive species and compare genetic differences using the mtDNA of *E. coqui* from Puerto Rico with the invasive populations established in California, and others from Hawaii.

2.2 MATERIALS AND METHODS

2.2.1 Sample Collection and Transcriptomic Sequencing

Three individuals of invasive *Eleutherodactylus coqui* were collected from Torrance California in the Los Angeles metropolitan area (Los Angeles County, California, United States). Four native *E. coqui* were collected from Río Blanco locality, Naguabo, Puerto Rico (PR, 18.261189, -65.796228). For comparison, we also included three individuals of *Pristimantis unistrigatus*, three individuals of *Eleutherodactylus johnstonei* from Cali, Colombia, two individuals of invasive *Eleutherodactylus planirostris* from Miami-Dade Florida (FL) and one individual of *Eleutherodactylus cochranae* were collected from the same locality in Río Blanco, PR.

All collected individuals were kept in containers with leaves and processed within 2-4 hours when they were humanely sacrificed through pithing and subsequent severance of the spinal cord. Tissues collections included skin and internal organs preserved in RNAlater at 4C for an initial 24 hours. For long term storage, samples were then kept in a -20C freezer. Collection permits were issued by Puerto Rican authorities: DRNA#: 2021-IC-011, O-VS-PVS15-SJ-01139-22072020 and we followed the humane euthanizing methods described in the IACUC protocol SJU-1965.

Total RNA was extracted from total skin tissue using the Trizol total RNA extraction protocol [86]. Total RNA was quantified with nanodrop, integrity determined with bioanalyzer, and kept at -80C until submission for sequencing. Only samples with RIN>7.0 were used for mRNA isolation, directional RNAseq library preparation, and Illumina 150 bp paired-end sequencing. Briefly, after the QC procedures using bioanalyzer, mRNA was isolated using oligo(dT) beads and rRNA was removed using

the Ribo-Zero kit. The isolated mRNA was used for cDNA synthesis and 150 bp pair-end sequenced directionally. The raw sequence data was submitted to the NCBI-SRA database under the BioProject PRJNA953648.

2.2.2 Skin Transcriptome Reconstruction, Annotation, and Expression Analysis

Raw RNA sequence reads (Table 3**)** were processed with Pincho v0.1 [6; 87]. This software provides a bioinformatics pipeline for high quality transcriptomics, which automates the following steps: Removal of Illumina adapter sequences with Trimmomatic [7-8]; insert error correction with Rcorrector [9-10]; three-assembler reconstructions using trans-ABySS [27-28] and rnaSPAdes [3; 23] with five computationally generated k-mer sizes derived from adapter-clean data inserts, and TransLig [29-30]; combined consensus transcriptome assembly derived from the three-assembler reconstructions using TransRate [32-33]; identical transcript redundancy reduction with CDHIT [34-35], consensus assembly quality assessment with BUSCO [38-39] for the eukaryota_odb10.2019-11-20 dataset; and transcript annotation against TrEMBL: amphibia with BLASTx [40-42], annotation against UniProt: Swiss-Prot with BLASTx with an e-value 1e-10, and annotation against KEGG: *Xenopus laevis* and *Nanorana parkeri* with BLASTn with an e-value 1e-10. For each transcript with a database match, only the first reference with the lowest e-value was used for annotation.

We further cleaned the annotated transcript by including a re-annotation step using a reference library *Xenopus tropicalis* (characterized UniProt) on the consensus assemblies using DIAMOND blast [88-90] to allow for isolation of single gene targets. Sequence alignment regions were used to trim reconstructions to portions that matched

the database only. Unmatched ends of transcripts were removed to ensure the removal of chimeric genes portions or erroneous connections before expression analysis. Duplicate sequences were removed with BBMap's dedupe protocol [13-14]. Weighted transcript/contig abundance was determined using the annotated transcriptome with Salmon [91-92] with default parameters. Data tables associated with these counts were then used for expression analyses with DESeq2 [93] in the DEBrowser webtool [94] pipeline. Read counts under ten were removed prior to analysis as a filtering option. We then used Panther17.0 [95] to determine the significant gene ontology groups from the overexpressed and underexpressed gene results from DESeq2.

2.2.3 Cathelicidin Antimicrobial Peptide Characterization

Cathelicidin (CATH) antimicrobial peptide sequences (potent against bacteria, viruses, and fungi) were discovered on our differential expression analysis for native and invasive *E. coqui*. We isolated CATH sequences from UniProt: Swiss-Prot files from *E. coqui, E. planirostris,* and *E. cochranae* annotated skin samples (**Supplementary Table S1, Table 4**). We used NCBI BLASTn to search our *E. coqui* CATH sequence against all *coqui* genomic datasets and we were able to retrieve a portion of a genomic scaffold matching our transcript within WNTK01001121. Known sequences of CATH AMP sequences in anurans were pulled from proteomics accounts in other frogs [96-99], from the NCBI database (XP_040211153.1, XP_018122516.1, ASU44943.1, and AVA30961.1), and from the *E. coqui* draft whole genome shotgun sequence library on the NCBI under accession number: WNTK01001121. CATH sequences were aligned in JalView utilizing cathelicidin-AL from *Amolops loloensis* as a reference for

Eleutherodactylus and trimmed accordingly so only the AMP portion of CATH remained (**Table 4**) [100].

2.2.4 Parasite Detection and Mitogenome

For parasite detection we downloaded a list of UniProt IDs for nematodes and platyhelminthes. Utilizing this list, we filtered out parasite genes from our *E. coqui* UniProt: Swiss-Prot annotations. Matching transcripts were filtered down to mitochondrial proteins and those proteins were validated via blastn against the complete NCBI nucleotide database at an evalue of 1e-10 yielding 18 distinct sequences (**Supplementary Table S3**). We identified the following roundworm mitochondrial sequences in PR *coqui*: *Ascaris suum, Anisakis simplex, Angiostrongylus cantonensis, Strongylus equinus, Oesophagostomum quadrispinulatum, Toxocara vitulorum, Oscheius tipulae, Pseudoterranova azarasi, Contracaecum osculatum B sensu Nascetti et al. (1993),* and *Contracaecum sp. ALS-2019* (**Supplementary Table S3**).

We also reconstructed the mitogenomes from RNAseq data, as these are common by-products of transcriptome assemblies. To isolate this genome, we filtered the annotated transcripts from the UniProt: Swiss-Prot consensus skin transcriptome assemblages for our study species: *E. coqui, E. planirostris, P. unistrigatus*, and *E. cochranae* (**Figure 4**, **Supplementary Figure S4**). If a mitogenome was already reconstructed in our UniProt: Swiss-Prot consensus skin transcriptome (i.e., PR *coqui*, validated against *Eleutherodactylus atkinsi* JX564864.1 reference partial mitogenome), then we used MitoZ's [101-102] *annotate* command along with the raw Illumina reads for the corresponding species. Some of these mt-genome transcripts cover the total length of the genome, and the best mitogenome reconstruction was used as a seed for further

validation with mt-genome assembler MitoZ under default parameters. MitoZ provides a GenBank-ready submission file with mitochondrial tRNA, rRNA, and protein coding genes delimitations; a stack-map of clean sequence reads that match the target mt-genomes; and a circular plot of each mt-genome of our five target species. If no mitogenomes were reconstructed completely we utilized each species unannotated transcriptomes along with MMseqs2 [103] under default parameters for nucleotide comparisons to locate the best transcript for mitogenome reconstruction using PR *coqui* mitogenome as our reference.

2.3 RESULTS

2.3.1 Draft Skin Transcriptomes:

All seven *E. coqui* assemblies achieved >96% BUSCO scores denoting high transcriptome completeness **(Table 3)**. Utilizing Uniprot: *Xenopus tropicalis*, UniProt: Swiss-Prot, TrEMBL: amphibia, and KEGG: *Xenopus laevis* and *Nanorana parkeri*, in total we annotated 62,512 and 76,241 unique alleles pertaining to LA-CA *E. coqui* and native *E. coqui* respectively and, collectively, 81,859 unique alleles were annotated between the two populations **(Table 5)**. Duplicate sequences were removed from final annotations, but we did not remove duplicate gene IDs. The UniProt: *Xenopus tropicalis* annotation was trimmed to matching coding sequences (CDS) to split chimeric/fused genes.

We used the protein class of each differentially expressed transcript among invasive and native *coqui* to create a general functional classification for overexpressed and underexpressed genes. We found chaperone, cytoskeletal, defense/immunity, extracellular matrix, and protein-binding activity modulator proteins all overexpressed in native *coqui* but not in the invasive population (**Fig. 11**). We found protein modifying enzyme and scaffold/adaptor protein to both be overexpressed in the invasive population but not in the native population (**Fig. 11**). Both populations showed overexpression of gene-specific transcriptional regulators, metabolite interconversion regulators, RNA metabolism proteins, translational proteins, and transporter proteins, but the invasive *coqui* exhibited a significant overexpression of metabolite interconversion regulator than native *coqui* (**Fig. 11**). Among the metabolite interconversion regulator proteins, seven were oxidoreductases and three were transferases which point towards increased

metabolic rate. Further comparison of *E. coqui* against invasive *E. johnstonei* reveals that *E. johnstonei* exhibit no overexpression of defense/immunity peptides when compared to native nor invasive coqui (**Fig. 12, 13**).

With the CDS aligned Uniprot: *Xenopus tropicalis* annotation, we conducted expression analysis by contrasting: LA-CA versus PR populations. By using PR are a reference, we found 102 differentially expressed genes that included 49 genes were underexpressed and 53 were overexpressed in the PR population **(Fig. 7; Supplementary Table S2)**. From these, Panther17.0 predicted three significant transcripts (A0A6I8Q9X7, A4QND9 and Q6PBF4) from the invasive population under the classification of mitochondrial ATP synthesis coupled electron transport related proteins. Panther17.0 could not determine any significant grouping for biological processes from the native population. Overall, these results suggest that these LA-CA individuals were more metabolically active, yet most differences between LA-CA and PR were not evident at the gene expression level. Therefore, we further explored unique genes being differentially expressed in both populations (i.e., present in one, but not in the other) that might contribute to organismal responses to pathogens and environmental conditions.

2.3.2 Defense and Stress Responses:

We revisited the gene expression results for gene ontology groupings that were missed by Panther17.0. We identified CATH and pentraxin 3 (PTX3) as two immune response genes among all DE genes **(Fig. 8)**. CATH and PTX3 were found to be overexpressed by PR coqui, with CATH being completely absent in LA-CA coqui **(Fig. 8; Supplementary Table S2)**. Reactive oxygen species modulator 1 (ROMO1), major histocompatibility complex, class I-related-like (MR1-LIKE), and mal, T cell

differentiation protein 2 (MAL2) were recognized as general defense response genes **(Fig. 8)**. ROMO1, MR1-LIKE and MAL2 were found to be overexpressed by LA-CA coqui, with MR1-LIKE and MAL2 being completely absent in PR coqui **(Fig. 8, Supplementary Table S2)**. PR coqui overexpressed a classic stress response protein DNAJ heat shock protein family HSP40 member C12 (DNAJC12), cell migration genes RAS homolog family member C (RHOC) and RAS homolog family member Q (RHOQ), and oxidative phosphorylation mitigation genes COX7A2 and CYTOCHROME B-C1 **(Fig. 8, Supplementary Table S2)**. LA-CA coqui overexpressed the antioxidants thioredoxin (TRX) and MGC76137, the wound healing genes epiplakin 1 (EPPK1) and collagenase 3 (MMP-13), and stress mitigators tyrosinase-like (TYR-LIKE) and NADPH dehydrogenase quinone 1 **(**NQO1; **Fig. 8)**. PR coqui did not express any MMP-13 and LA-CA coquis did not express any RHOC **(Supplementary Table S2)**. We then extracted all AMPs found in PR *E. coqui* along with their nucleotide and amino acid sequences and discovered DMBT1 and FAM3A to also be expressed by native coqui but not at significant levels (**Supplementary Table S1**).

2.3.3 Cathelicidin Characterization:

Isolated CATH AMP transcripts from native *E. coqui* transcriptomes and aligned to known anuran CATH AMP sequences. *E. coqui* CATH, *E. cochranae* CATH, and cathelicidin-AL sequences share high sequence similarity (**Fig. 9**). CATH AMP regions from *E. coqui*, *E. cochranae*, and cathelicidin-AL were 79AA, 74AA, and 48 AA respectively (**Table 4**). CATH AMP regions from *E. coqui* and E. *cochranae* are on average more electronegative, longer, and have a higher normalized hydrophobicity than other anuran species (**Fig. 9, Table 4**).

Both *E. coqui* and *E. cochranae* contain a repeat region, which only occurs in these two species, consisting of NGGRG that spans 10 times and 8 times respectively **(Table 4)**. Comparison with the published genome of *E. coqui* provided a sequence nearly identical to our transcriptomic sequence, except the genomic counterpart was 1 NGGRG repeat shorter than the transcriptomic version **(Table 4)**. This repeat sequence lengthens the *Eleutherodactylus* AMP to a length that is longer than all other anuran CATH AMP sequences **(Fig. 9, Table 4)**. *Eleutherodactylus* cathelcidin contains hydrophobic residues near the N-terminus of the peptide like other CATH AMPs; however, its C-terminal side is largely hydrophilic **(Fig. 9).** *E. coqui* CATH is also a peptide that is the highest positive charge seen to date in anurans due to the introduction of arginine in the repeat sequence.

2.3.4 Parasite signatures in native *E. coqui* transcriptomes:

We were able to identify ten distinct species of nematodes in our UniProt annotation in *E. coqui* from Puerto Rico (**Supplementary Table S3**). These include 16 sequences that match to roundworm mitogenomes; however, even the sequence with the best BLAST max score and the lowest evalue of 0 (contig:Graph_29462_0_3_6.31) had at least 60 distinct species that also yielded an evalue of exactly 0 as well. We present 10 distinct species of roundworms as determined by BLAST max score although we realize that these may not be the exact parasites that were within the PR *coqui*. We present this knowledge simply as evidence of distinct parasite signatures within PR *coqui* with no parasite signatures detected in CA *coqui*.

2.3.5 Mitogenomics:

E. coqui draft mitogenome is fairly complete, containing 35 of 37 total genes (**Fig. 10**). *trnL* and *trnF* are two genes which encode tRNAs which are missing from the depiction of the mitogenome (**Figure 4**); however, it may be likely these tRNAs are present in the mitogenome sequence but did not align well to the reference. Cytochrome c oxidase subunit 1 (COI) transcripts gathered from all three CA *coqui* replicates were 100% identical to each other. All three COI sequences from CA *coqui* mapped best to an *E. coqui* voucher for COI (KY033486) collected from Kukuihaele, Hawaii at 98.72% sequence identity and 100% query cover proving the strongest evidence of their relatedness.

2.4 DISCUSSION

2.4.1 Introduction

We find that by studying the transcriptome of Puerto Rican *E. coqui*, we were able to identify antimicrobial peptides in the species that were currently believed to not exist. Our paper is not the first to employ the usage of transcriptomics in *E. coqui*.

2.4.2 History of *E. coqui* invasion

The common coqui is a successful invader, which is considered a noisy nuisance, a pest, and a voracious predator with evident ecological impacts. Like other invasive species, *E. coqui* are a major concern for declines in local biodiversity, especially in vulnerable habitats with representative or endangered endemic flora and fauna [104-106]. In the U.S., the common coqui has been mainly reported from California, Florida, and Hawaii; and these are regions with the most active programs of eradication. In California, the species was introduced in 2005 with plants imported for nurseries [61]. Current eradication program includes the identification of infested colonies and then spraying citric acid for their immediate removal. In Florida, *E. coqui* were introduced to Miami-Dade County in the 1980s via the horticultural trade from Puerto Rico. Yet, most populations might not be viable in Florida due to the colder winter climates and thus depend on constant reintroduction [107]. Currently *E. coqui* are presumably eradicated from Florida [84].

Hawaii is the US state with the most active program of *E. coqui* eradication and this species has reached high population densities with up to 91,000 frogs per ha [61]. In places where this frog is established, a common complaint is the loudness of males' calls due to their high densities. For this reason, introduced coquis may be a causal factor in

deterring tourism in Hawaii [108] and infestations have been reported to lower property values by ~0.16% [109]. Moreover, Hawaii milder climatic condition has made *E. coqui* a formidable pest and noticeable threat to local economy such as the tropical horticulture industry where coqui-infested nurseries may require quarantine to reduce the risk of spread [110]. Such mitigating efforts usually result in harsh fines for importing coquis, economic losses due to quarantines, reputational damage, and might require costly programs of eradication [111].

The US is not the only country where *E. coqui* has been introduced. In 1998 six individuals of *E. coqui* were intentionally released in the city of Turrialba, Cartago Province, Costa Rica [59]. This population was revisited in 2010; however, only 100 individuals were observed, and revisited again in 2019 where a predicted population of 200 in the city of Turrialba, Cartago Province and sixty individuals in the Invu neighborhood in Juan Viñas [59]. In this scenario *E. coqui* have not been able to reach high densities. Additional locations where *E. coqui* are reported to be invasive in the Dominican Republic, St. Thomas, and St. Croix; although iNaturalist sightings report them to be expansive throughout the Caribbean [84; 112].

2.4.3 Approaches to *E. coqui* eradication

There are several forms of pest control for *E. coqui*: 16% citric acid treatment, 3% hydrated lime, 2% caffeine, Thionex, sodium bicarbonate and hot water. Citric acid treatment is currently used as treatment against *E. coqui* as it has 100% lethality after one hour of direct contact, but it has some phytotoxic effects and is very expensive compared to hydrated lime [113-114]. Regardless of the side-effects 16% citric acid is considered a low-risk pesticide with less restrictions than the other alternatives [115]. Utilization of

16% citric acid has proven to be effective at eradication of *E. coqui* from Wahiawā after consistent application over five years [115]. The application of 3% hydrated lime reports an 80% mortality rate after 48 hours of direct contact [116]. Hydrated lime is not phytotoxic; however, it does leave behind an unsightly white residue that detracts from the aesthetic of ornamental plants and can be corrosive at higher concentrations [114; 117]. 2% caffeine is specific to killing frogs, but there were concerns about potential human health effects of widespread applications of caffeine [114]. Thionex is an insecticide which is very effective against coqui, but its usage is also highly restricted due its effects on other species [116]. Sodium bicarbonate can also be applied as a fine powder and reach mortality rates greater than 80% twenty-four hours after contact [116]. Hot water treatment is described as a non-phytotoxic treatment for potted plants that involves application of 39C water for five minutes or 45C water for one minute for 100% mortality of adult coqui frogs and their eggs [118]. The biggest hurdles would be ensuring direct contact and accounting for the decrease of effectiveness of the treatment on days with high humidity [115]. In addition to these methods, there is also the option of capture and extermination if the species is of low abundance.

2.4.4 The discovery of AMPs in *E. coqui*

Utilizing *E. coqui* skin samples from Puerto Rico and California we have identified several overexpressed genes in both *E. coqui* populations that indicate high amounts of stress-mitigation. In PR coqui we notice high amounts of immune response and general stress response genes. Most notably we discover high amounts of CATH, a potent antimicrobial peptide (AMP), within the PR *E. coqui* population. In contrast, CA *E. coqui* exhibit no differential expression of CATH nor any other AMP.

AMPs are known to act both directly on microorganisms and indirectly via acting as cell-signaling molecules [119]. Previous research has been unsuccessful in the identification of any AMPs within *E. coqui*, noting that coqui must rely on other mechanisms for protection against pathogens such as *Batrachochytrium dendrobatidis* (*Bd*) [78]. We present the first skin draft transcriptome assembly from *E. coqui*, which highlights the presence of one potent AMP: CATH, and strong immune system mediator: PTX3. Upon further inspection of the microbes from native and invasive E. coqui, we confirm that PR coqui were infected with a larger and diverse number of roundworms, all of which were absent in CA coqui. We also note peculiar aspects of *Eleutherodactylus coqui* CATH.

CATH are known to defend skin tissue against a wide range of bacteria, viruses, and fungi, as it is highly expressed in epithelial cells, neutrophils, as well as, dendritic cells, lymphocytes, mast cells, macrophages, monocytes, and NK cells in vertebrates [119-124]. CATH are also known to be produced in dermal serous glands where they are stored within granules until released [125]. CATH can exert direct antimicrobial effects and can trigger specific defense responses within the host [70]. The cationic and amphipathic properties of CATH allow them to associate with the negatively charged phospholipids in bacterial membranes and penetrate the invading microbe's membrane leading to the fragmentation of the membrane and subsequent cell death [119-120]. The CATH gene in *Rana muscosa* and *Rana sierrae* was found to be one of two AMPs overexpressed in response to *Bd* infection [126]. We hypothesize that this newfound CATH may also play a significant role in *E. coqui*'s resilience to *Bd* infection as well although we did not detect the presence of *Bd* infection [78]. Instead, we detected the

mitochondrial signature of ten roundworms within PR coqui, any of which could have triggered CATH production.

CATH consist of a signal sequence, a well conserved pro-region and highly diverse active peptides that are classified under five different structural groups (Glycine and Serine-rich or Type-GS, Proline and Arginine-rich or Type-PR, Tryptophan and Arginine-rich or Type-WR, β-haripin or Type-β, and α-helix or Type-α) [121]. Anuran CATH active peptide regions conform to Type-GS; however, we note our newly discovered CATH's active peptide to be Glycine and Asparagine-rich **(Table 4, Supplementary Table S1)**. It has been noted that CATH produced by amphibians are species-specific and vary tremendously between species [127]. We have evidence which leads us to believe that *Eleutherodactylus* CATH are similar but also contain key differences as well in amino acid composition to each other; however, they are highly variable compared to other families leading us to hypothesize a drastic change in function **(Fig. 9, Table 4)**. This is of great interest to us considering CATHs' potential for wound healing, angiogenesis, antimicrobial, anti-inflammatory, and anti-cancer research, and application [127].

We also noted a significantly higher level of PTX3 in the native coqui population than in the invasive population. Long pentraxin PTX3 are produced by endothelial cells and macrophages [128]. PTX3 works within innate immunity, binding to both gram-negative bacteria and fungal spores in mouse models [129]. Observing significantly higher levels of PTX3 in the native population supports our hypothesis that the native coqui are in an environment more saturated in pathogens than in the environment in which they are invasive to.

Finding high amounts of CATH and PTX3 leads us to believe that perhaps the environment in PR is not as ideal for the species as the locations in which they have become invasive and expansive (i.e., highest population densities in Hawaii). Fighting infections may be one of many limiters in the population density of *E. coqui* in PR compared to CA and even Hawaii. It may also be the cause of the under expression of several oxidoreductases within the PR population which signals decreased metabolic activity [130]. We postulate the likely scenario that the native *E. coqui* population experienced a current or past exposure to a highly transmissible pathogen. We hypothesize on the pathogen's transmissibility because even though the frogs were kept isolated, they all showed similar symptoms of infection or post-infection. It is likely that CATH and PTX3 played a role in the defense of the organism via direct interaction with the pathogen or signaling local cells to initiate an immune response.

California coqui exhibited high amounts of wound healing genes, along with stress mitigation in the form of antioxidants. The cause of the wound healing response is unclear; however, the history of the area in which we collected sheds some clarity, although it may just be happenstance. At the time of collections and even currently in California, 16% citric acid treatments are used as a minimum risk pesticide to euthanize coqui [113]. This pesticide eliminates coqui in any stage of life upon direct contact with 100% effectiveness [113]. Citric acid is quickly absorbed by coqui skin and sends the frog into osmotic shock within 30-60 minutes; however, partial contact or indirect contact may not be enough to secure lethality [113]. The invasive coqui from California were collected from an area in Torrance that was under citric acid treatment. The collected coqui was kept overnight so they were therefore not killed by the citric acid treatment;

however, we cannot rule out that the coqui was not indirectly nor partially exposed to the pesticide application.

Upon further inspection utilizing *E. johnstonei* as a third group in our differential expression analysis against native and invasive coqui we note several key takeaways. The *E. johnstonei* population sampled, although an invasive population, does not show overexpression of any wound healing peptides. As this population was not under citric acid treatment nor did we notice any noticeable wounds, this further supports our hypothesis regarding citric acid treatment and wound healing response. We also note a lack of any overexpression of immune peptides, providing evidence that these anurans are not currently combating infection as our native coqui were. Upon closer examination of the *E. johnstonei* transcriptome, we were unable to locate any actively expressed cathelicidin sequences as we did for *E. coqui*, *E. cochranae* and *E. planirostris*.

2.5 CONCLUSION AND FUTURE DIRECTIONS

Draft transcriptomes of a non-model organism can serve as a reliable summary into the molecular processes within the organism at a given place and time. We offer several variations of the *Eleutherodactylus coqui* transcriptome publicly in the hopes that it leads to future discoveries concerning *E. coqui*'s resilience to dangerous infections, such as *Bd*, and general conservation studies. We provide the hypothesis, via differential expression analysis, that *E. coqui* may utilize a novel form of CATH and PTX3 as the main immune defense mechanism via direct and/or indirect interactions with pathogens entering the skin. We note high levels of different forms of stress between native and non-native populations. From this evidence we deduce that the localities that were sampled were both not completely ideal for the species, although it did not deter their expansion; however, we note that invasive coqui are able to escape an environment rich in pathogens, allowing the species to conserve energy and spend it on reproduction and subsequent expansion. We admit that our study is limited to only two localities and can be greatly expanded with sampling more populations from different locations; however, we provide sufficient information to suggest that *E. coqui* are under stress and have successful genetic toolsets to overcome this stress to become an expansive species in its native locations and even more so in its non-native settings. We also suggest additional sampling of *E. coqui* would greatly aid conservation efforts for other species of *Eleutherodactylus* that are severely endangered. All our data is publicly available on the NCBI database.

2.6 TABLES AND FIGURES

2.6.1 TABLES

Table 3: *Eleutherodactylus coqui* Skin Illumina NGS Data

ID[1]	Location	Number of Reads	Number of Bases (bp)	Insert Length (bp)	BUSCO Score
CC22	California	25,063,868	3,759,580,200	150	98.1%
CC41	California	25,216,519	3,782,477,850	150	96.5%
CC52	California	22,232,192	3,334,828,800	150	97.7%
PC1-1	Puerto Rico	41,194,569	6,179,185,350	150	98.1%
PC3-1	Puerto Rico	32,494,855	4,874,228,250	150	99.2%
PC5-1	Puerto Rico	25,381,971	3,807,295,650	150	97.3%
PC8-1	Puerto Rico	31,162,042	4,674,306,300	150	98.1%

[1] IDs of NGS data submitted to the SRA database. CC: California coqui, PC: Puerto Rico coqui

Table 4: Cathelicidin AMP Sequences in Anurans (DBAASPv3.0 MF)

Species	**Peptide**	**Length**	**Net Charge**	**Normalized Hydrophobicity**	**Sequence**	**Source**
A. loloensis	cathelicidin-AL	48	12	0.44	RRSRRGRGGGRR GGSGGRGGRGGG GRSGAGSSIAGVG SRGGGGGRHYA	Alford et al. 2020
E. coqui	EC-CATH	79	17	0.66	RRSRNGGRGNGG RGNGGRGNGGRG NGGRGNGGRGNG GRGNGGRGNGGR GNGGRGGRGGGR TGSGSFIAGGGSR GKGSYA	Río Blanco, Naguabo PR Collection
E. coqui	EC-CATH	74	16	0.64	RRSRNGGRGNGG RGNGGRGNGGRG NGGRGNGGRGNG GRGNGGRGNGGR GGRGGGRTGSGSF IAGGGSRGKGSYA	WNTK01001121*
E. cochranae	ECo-CATH	74	15	0.67	RRSRNGGRGSGG RGNGGRGNGGQG NGGRGNGGRGNG GRGNGGRGNGGR GGRGGGRTGSGSS IAGGGSRGKGSYA	Río Blanco, Naguabo PR Collection
E. cochranae	ECo-CATH	69	15	0.69	RRSRNGGRGNGG RGNGGRGNGGRG NGGRGNGGRGNG GRGNGGRGGRGG GRTGSGSSIAGGG SRGKSSYA	Río Blanco, Naguabo PR Collection
E. planirostris	EP-CATH	45	9	0.62	RRSRNGGRGSGG RGNGGRGNGGQG NGGRGNGGRGGR GGGRTGSGSSIAG GGSRGKGSYA	Miami FL Collection
B. gargarizans	cathelicidin-Bg	34	8	-0.42	RPCRGRSCSPWLR GAYTLIGRPAKNQ NRPKYMWV	Alford et al. 2020
B. gargarizans	BG-CATH37	37	10	-0.18	SSRRP**C**RGRS**C**GP RLRGGYTLIGRPV KNQNRPKYMWV	Alford et al. 2020
D. melanostictus	cathelicidin-DM	37	9	-0.24	SSRRKPCKGWLC KLKLRGGYTLIGS ATNLNRPTYVRA	Alford et al. 2020
H. rugulosus	HR-CATH	33	8	0.16	ASKKGKCNLLCK LKQKLRSVGAGT HIGSVVLKG	Chen et al. 2021
L. fragilis	Lf-CATH1	30	5	-0.83	PPCRGIFCRRVGSS SAIARPGKTLSTFI TV	Alford et al. 2020
L. fragilis	Lf-CATH2	30	5	-0.21	GKCNVLCQLKQK LRSIGSGSHIGSVV LPRG	Alford et al. 2020

N. ventripunctata	cathelicidin-NV	24	5	0.25	ARGKKECKDDRC RLLMKRGSFSYV	Alford et al. 2020
N. yunnanensis	cathelicidin-PY	29	6	0.04	RKCNFLCKLKEKL RTVITSHIDKVLRP QG	Alford et al. 2020
O. andersonii	cathelicidin-OA1	27	0	-0.37	IGRDPTWSHLAAS CLKCIFDDLPKTH N	Alford et al. 2020
O. livida	OL-CATH1	33	5	-0.45	KKCKGYRCRPVG FSSPISRRINDSENI YLPFGV	Alford et al. 2020
O. livida	OL-CATH2	33	8	0.17	RKCNFLCKVKNK LKSVGSKSLIGSAT HHGIYRV	Alford et al. 2020
P. nigromaculata	PN-CATH1	34	10	0.41	KKCNFFCKLKKK VKSVGSRNLIGSA THHHRIYRV	Wang et al. 2021
P. nigromaculata	PN-CATH2	29	6	-0.14	EGCNILCLLKRKV KAVKNVVKNVVK SVVG	Wang et al. 2021
P. puerensis	cathelicidin-PP1	32	6	-0.2	ASENGKCNLLCLV KKKLRAVGNVIKT VVGKIA	Alford et al. 2020
P. puerensis	cathelicidin-PP2	46	10	0.37	SRGGRGGRGGGG SRGGRGSSGRGRT GSGSFIAGGGNRG SRGGRQYA	Chen et al. 2021
R. catesbeiana	cathelicidin-RC1	28	9	-0.26	KKCKFFCKVKKKI KSIGFQIPIVSIPFK	Alford et al. 2020
R. catesbeiana	cathelicidin-RC2	33	8	0.23	KKCGFFCKLKNKL KSTGSRSNIAAGT HGGTFRV	Alford et al. 2020
R. temporaria	cathelicidin-like	32	7	-0.23	KCNFLCKVKNRL KSLSSTSVIAAGVP RGTYRG	XP_040211153.1*
T. verrucosus	cathelicidin-TV	17	7	1	PRQTRKCVRQNN KRVCK	Chen et al. 2021
X. laevis	cathelicidin	49	12	0.13	RRNGGRGRGGGR GGGIRGGGKSGFG SPIAGVKRNLINLP NAKIKRHFSA	XP_018122516.1*
X. tropicalis	cathelicidin	27	5	-0.18	KKCKTSGCRFTGA GSAIAGVKPLQSIG	Gao et al. 2016
Z. puerensis	cathelicidin-1 precursor	28	7	-0.25	GKCNLLCLVKKK LRAVGNVIKTVVG KIA	ASU44943.1*
Z. puerensis	cathelicidin precursor	48	12	0.45	RRSRGGRGGRGG GGSRGGRGSSGRG RTGSGSFIAGGGN RGSRGGRQYA	AVA30961.1*

*Sequences were aligned and trimmed to estimate AMP region

Table 5: Unique Alleles Annotated in Native and Invasive *Eleutherodactylus coqui*

ID[1]	UniProt: *Xenopus tropicalis*	UniProt: Swiss-Prot	TrEMBL: amphibia	KEGG: *Xenopus laevis* and *Nanorana parkeri*
CC	17,723	22,634	37,675	12,183
PC	18,144	26,937	50,949	16,695

[1] IDs of samples: CC: California coqui, PC: Puerto Rico coqui.

2.6.2 FIGURES AND FIGURE LEGENDS

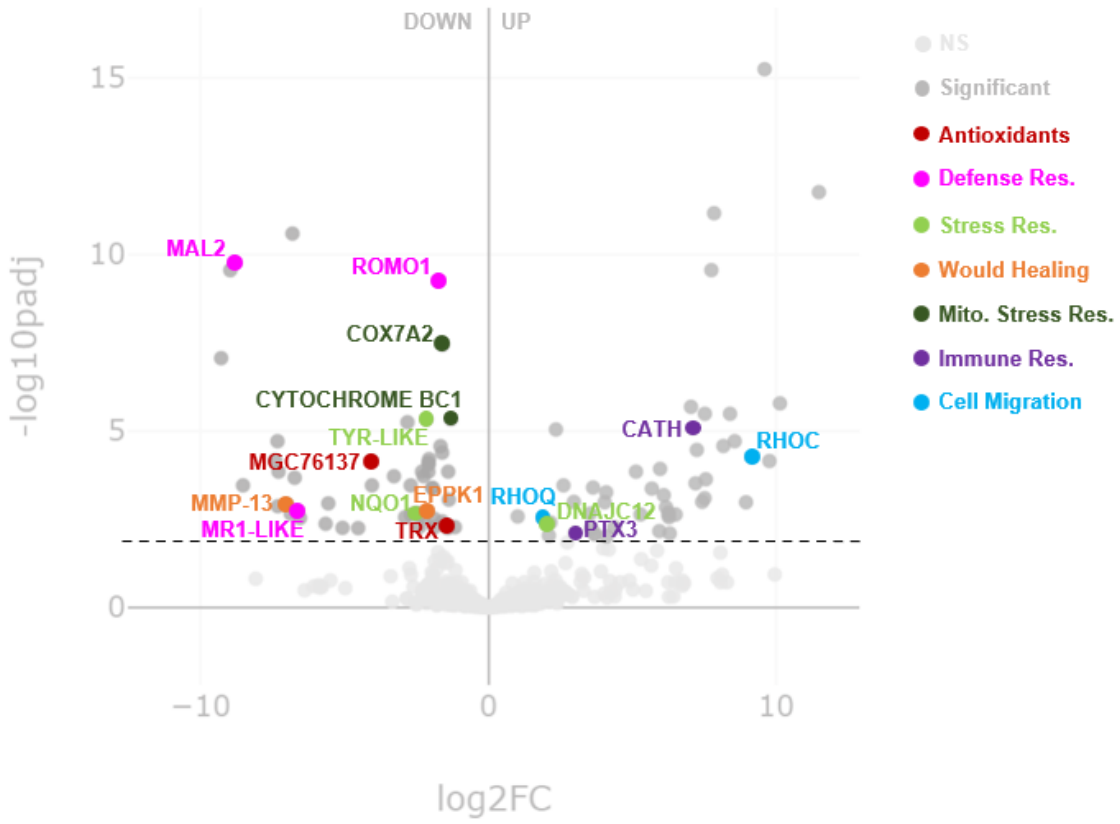

Figure 7: Differential Expression (DE) Analysis Comparing Native to Invasive *E. coqui*.

Nonsignificant transcripts under p<0.05 are depicted in light gray. Significant (p>0.5) transcripts are depicted either in dark gray or in non-light gray color. Transcripts of interest and the gene ontology groups of antioxidants (red), defense response (defense res.; pink), stress response (defense res.; light green), would healing (orange), mitochondrial stress response (mito. stress res.; dark green), immune response (immunes res.; purple) and cell migration (light blue) are labeled, and color coded accordingly. DE was conducted via DESeq2 with default parameters. Low read counts under ten were filtered prior to analysis. 102 total DE genes are depicted.

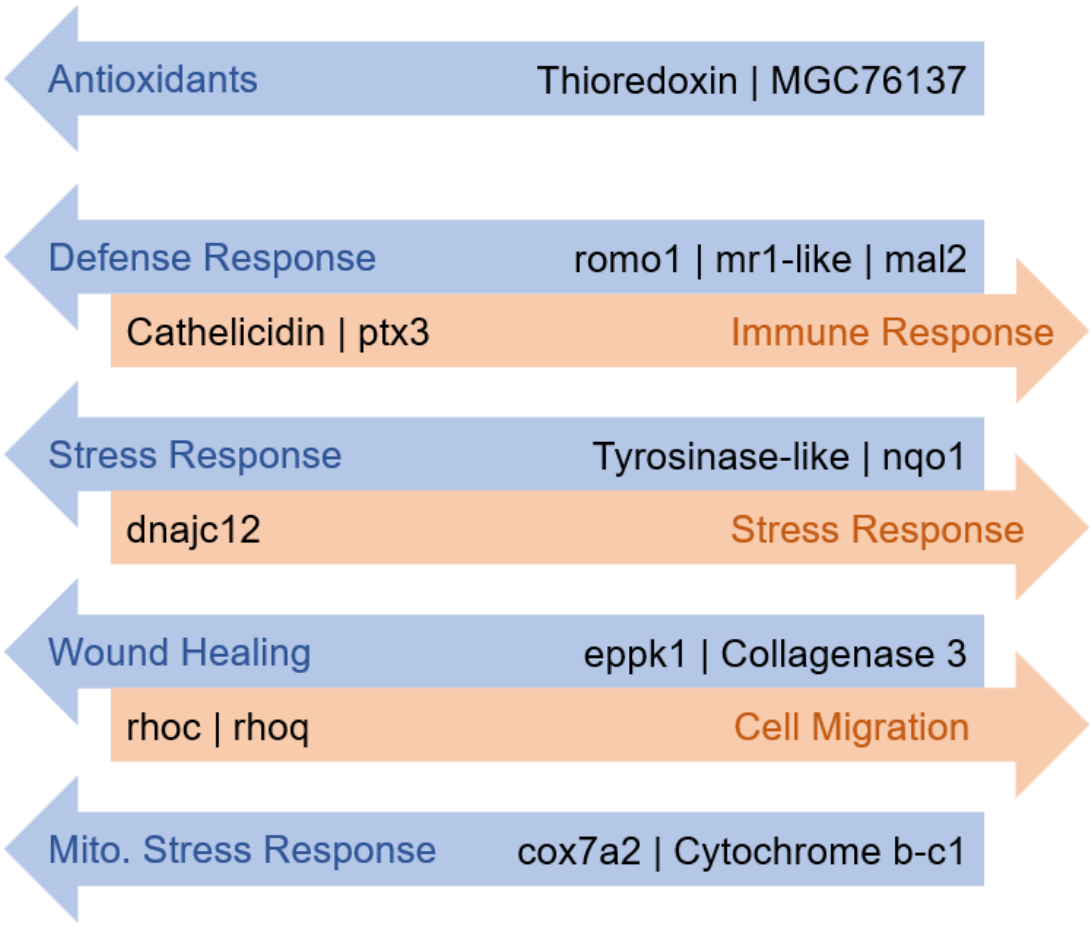

Figure 8: Genes of Interest in Native *E. coqui*.

DESEQ2 expression analysis results. Overexpressed and underexpressed genes in PR coqui are represented in orange and blue respectively.

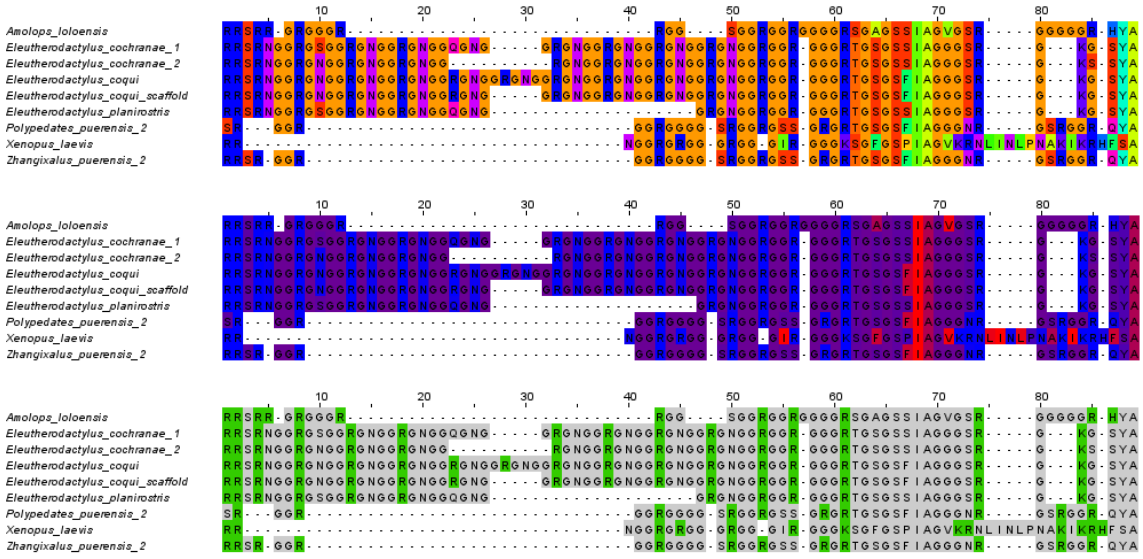

Figure 9: Anuran Cathelicidin Antimicrobial Peptide Alignment.

Protein alignment between our newly discovered glycine-rich cathelicidin antimicrobial peptide (AMP) from *E. coqui* and known anuran cathelicidin AMP sequences. The bottom alignment highlights hydrophobic residues in red, slightly hydrophobic in purple, and hydrophilic residues in blue. The top alignment is a traditional protein alignment without the hydrophobicity overlay.

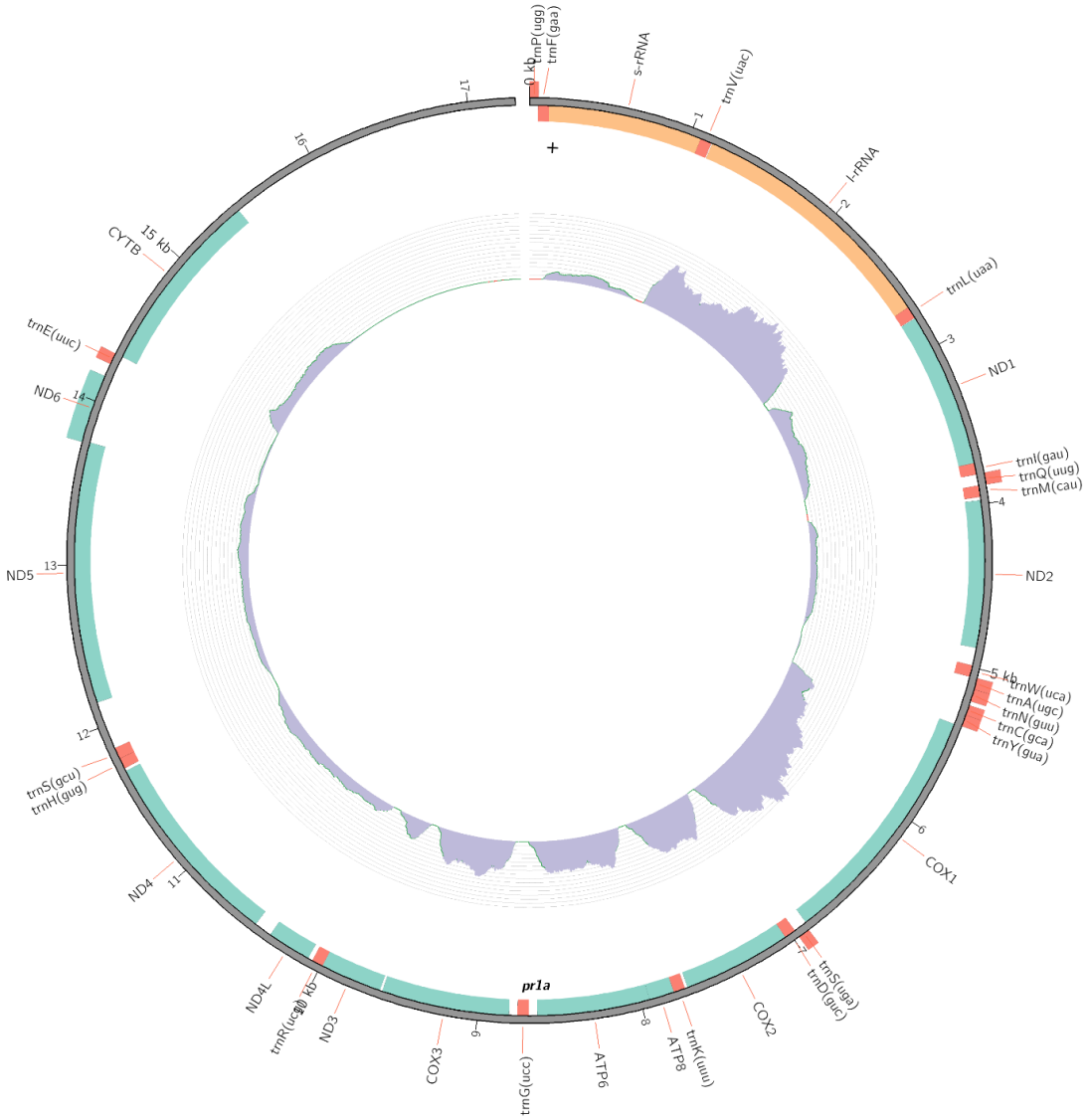

Figure 10: Draft Mitochondrial Genome of Native *E. coqui*.

MitoZ and Circos depiction of the mitochondrial genome of *E. coqui* from Puerto Rico (PR). The sample name is *pr1A*, the first *E. coqui* collected by our team in PR. Inner blue circle denotes levels of read mapping from raw Illumina reads.

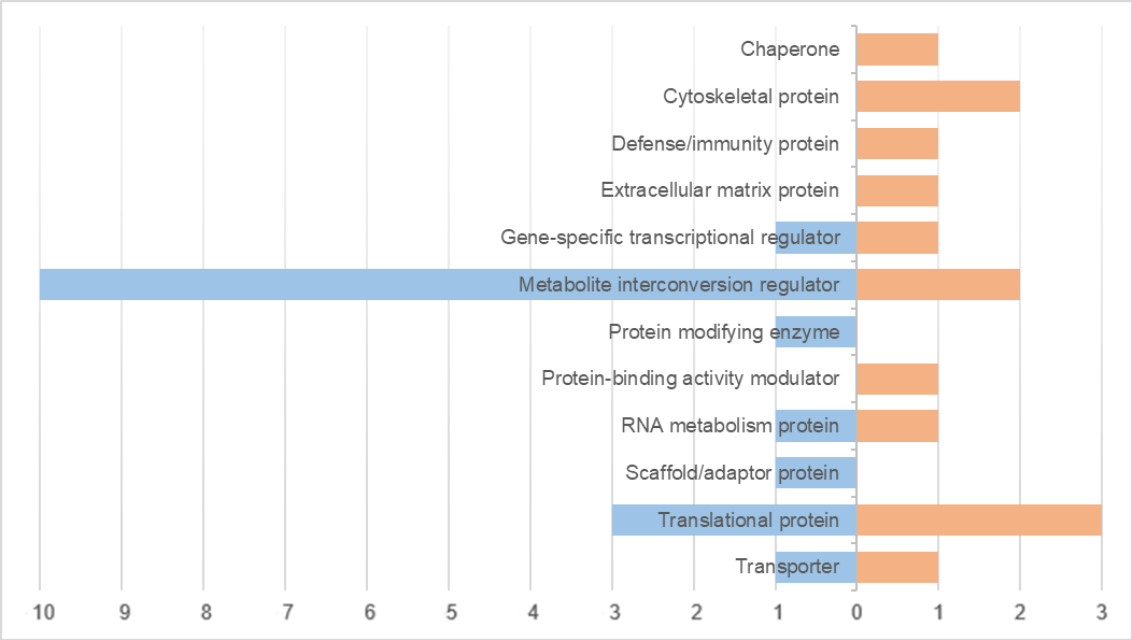

Figure 11: Functional Classification via Protein Class of Invasive and Native *E. coqui*.

Panther17.0 function classification of DESEQ2 expression data. Transcript count number for each protein class denoted on the x-axis. Native *coqui* are represented in orange and invasive *coqui* are represented in blue.

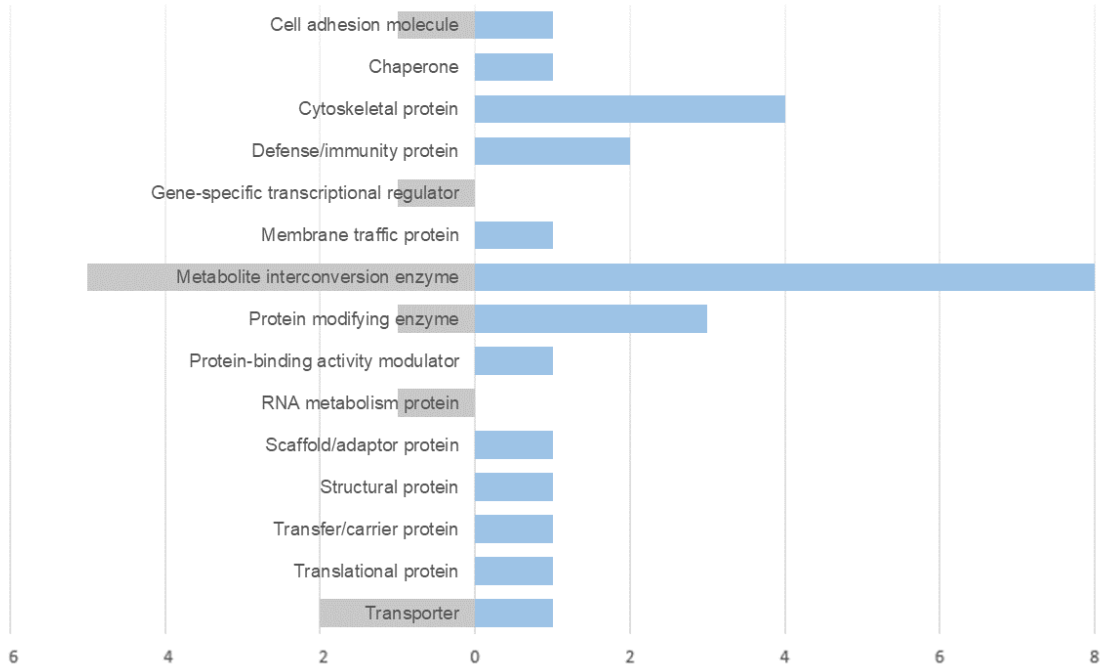

Figure 12: Functional Classification via Protein Class of Invasive *E. johnstonei* and Invasive *E. coqui*.

Panther17.0 function classification of DESEQ2 expression data. Transcript count number for each protein class denoted on the x-axis. Invasive *E. johnstonei* are represented in gray and invasive *E. coqui* are represented in blue.

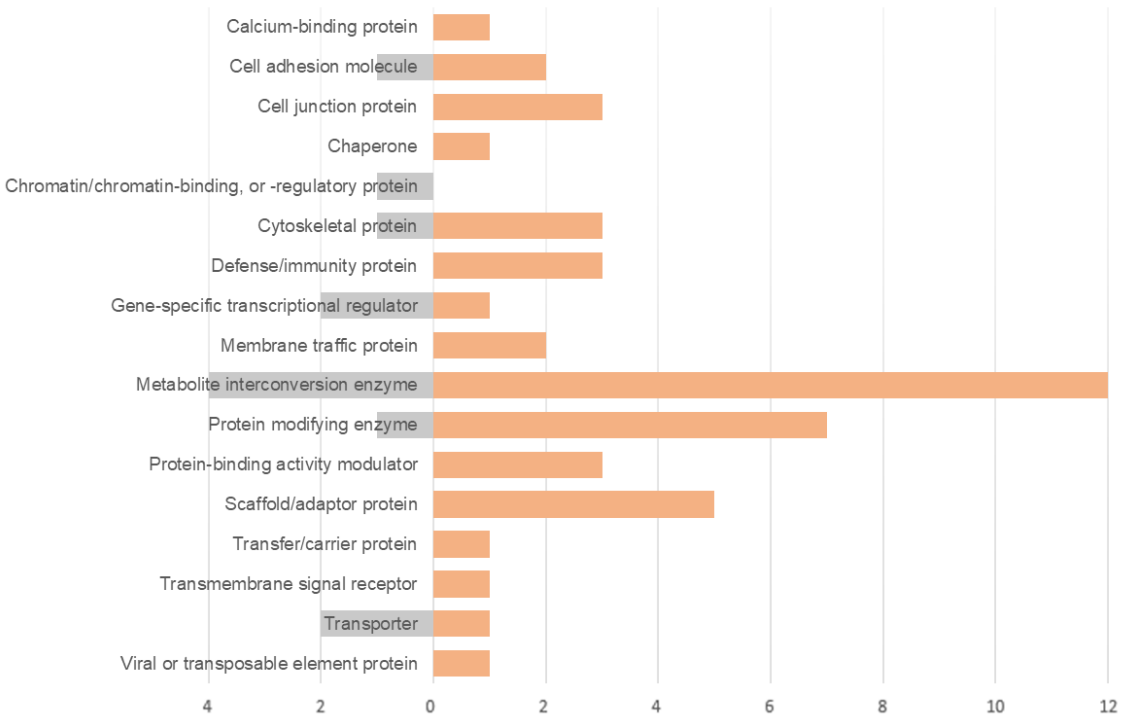

Figure 13: Functional Classification via Protein Class of Invasive *E. johnstonei* and Native *E. coqui*.

Panther17.0 function classification of DESEQ2 expression data. Transcript count number for each protein class denoted on the x-axis. Invasive *E. johnstonei* are represented in gray and native *E. coqui* are represented in orange.

CHAPTER 3

ANURAN DEPARTURE FROM TRADITIONAL VERTEBRATE HEMOGLOBIN PEPTIDE STRUCTURE POINTS MANY ANURAN SPECIES TOWARDS HEMATOLOGICAL CONSEQUENCES

3.1 INTRODUCTION

3.1.1 Anuran Circulatory and Respiratory System

Anurans utilize a combination of a three-chambered heart and simple lungs along with buccal pumping and cutaneous respiration to circulate oxygen and carbon dioxide throughout the body [131]. Buccal pumping involves pumping air into alveoli-absent lungs where gas exchange can occur directly from the lung tissue; however, with a three-chambered heart, the blood that circulates to cells is always a mixture of oxygenated and deoxygenated. For non-direct developing anurans, the tadpole stage contains gills that aid in gas exchange until they mature into a terrestrial morphology, where the gills disappear. Like other ectothermic vertebrates, anurans' respiratory and circulatory systems are not completely separated; therefore, response to hypoxic conditions can be mediated by not only increased ventilation, but also with increased heart rate or vascular tone modifications [132]. Adaptations for oxygen acquisition provide a low stress mechanism for gas exchange in anurans, especially as they enter low oxygen conditions (i.e., oxygen deficient waterways, high elevations, and high temperatures).

3.1.2 Hemoglobin has a Central Role in Oxygen Transport and Delivery

Hemoglobin plays a key role in gas exchange throughout the body, carrying oxygen to the cells and removing carbon dioxide in complex organisms. Larger, more complex, organisms (i.e., vertebrates) require a method of gas exchange that simple diffusion from the epidermal skin layer cannot supplant. Thus, the ever important role of gas exchange is fulfilled primarily by hemoglobin. Vertebrate hemoglobin contains four globular protein subunits interacting to form a quaternary protein structure. In adult hemoglobin, the four subunits are typically hemoglobin alpha 1 (HBA1), hemoglobin alpha 2 (HBA2), hemoglobin beta 1 (HBB1), and hemoglobin beta 2 (HBB2). Each globular protein contains a non-protein prosthetic heme group (four heme groups total), which will ultimately associate iron, attracting its own oxygen molecule in a 1:1 ratio. Carbon dioxide binds with the amine groups attached to the heme groups, allowing hemoglobin to act as a carrier for both oxygen and carbon dioxide simultaneously.

Ectothermic vertebrates do not utilize blood flow for body temperature regulation and therefore cannot regulate their own body temperature, relying on the environment and surroundings instead to modulate their temperature via radiation, conduction, convection, and/or evaporation. An interesting trade-off is that because of the inability to regulate body temperature, ectothermic vertebrates tend to be at cooler body temperatures increasing the oxygen affinity of hemoglobin [133]. As hemoglobin becomes oxygenated it transitions to a ligated state (R state) from the deoxy state (T state) [134]. The T state will have higher oxygen affinity due to the decrease of allosteric interactions compared to the R state [133-134]. In addition, the T state leads to an altered distribution of protons in red blood cells, increasing the pH and subsequently also further increasing oxygen

affinity [134]. Higher oxygen affinity of hemoglobin greatly benefits ectothermic vertebrates in hypoxic/anoxic conditions.

3.1.3 Cytoglobin is Ubiquitous but Poorly Understood

As opposed to the heteromeric quaternary structure of hemoglobin, located in erythrocytes, cytoglobin (CYGB) forms a homodimer and is found in all tissue types [135]. CYGB is believed to bind one heme group per monomer, for a total of two heme groups. Whereas the role of hemoglobin is well understood, the role(s) of CYGB is not fully characterized. CYGB has been shown to defend against oxidative stress, CYGB may protect against hypoxia, and may have many more speculated roles such as facilitation of the diffusion of oxygen to the mitochondria [136-138]. One aspect of cytoglobin that is known for certain is that it is upregulated in scenarios of hypoxia; therefore, it must have an important role regarding hypoxia resilience in anurans [137].

3.1.4 Hypoxia and the Hypoxia Inducible Pathway

Hypoxia is a state of oxygen deficiency that impairs metabolic rate, greatly jeopardizing the survivability of the organism. In response to hypoxic/anoxic conditions, animals will initiate the hypoxia inducible pathway via upregulation of the hypoxia inducible factor (HIF). HIF is a transcription factor created from the heterodimerization of either HIF1A, HIF2A (EPAS1), or HIF3A with HIF1B (ARNT) [139]. HIF1B is constitutively expressed whereas HIF alpha subunits are upregulated during moments of low oxygen availability [139]. HIF1A is the primary, and essential, component of the hypoxic response leading to a cascade of upregulated genes to mitigate the detrimental effects of hypoxia [140]. HIF2A can bind to the same consensus hypoxia response element (HRE) as HIF1A but HIF2A only upregulates genes involved in angiogenesis

and pluripotent stem cell maintenance [139]. HIF3A was found to inhibit HIF1A expression while at the same time upregulating HIF2A; however, HIF3A is still poorly understood [140].

The HIF-1 (HIF1A/HIF1B heterodimer) pathway modulates the transcription of hundreds of genes pertaining to metabolism, vascularization, vascular tone, tissue growth, and cell longevity, to name a few, in cases of low oxygen availability [141-142]. Under hypoxic conditions an organism will favor anaerobic respiration for energy production and initiate self-preservation processes to prolong short-term survival. In our study we focus on a subset of proteins regulated by HIF-1 under (1) anaerobic respiration, (2) vascular tone, and (3) apoptosis inhibition.

HIF-1 will upregulate transcription of enzymes pertaining to glycolysis necessary for processing carbohydrates such as aldolase, fructose-bisphosphate A (ALDOA) in catalytic step four of glycolysis, and those necessary for processing glycerol from fats and completing carbohydrate processing such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in step six, and enolase 1 (ENO1) in step nine. HIF-1 also increases transcription of proteins that regulate vascular tone as either a vasodilator such as the inducible nitric oxide synthase (iNOS) enzyme or as a vasoconstrictor mediating protein such as endothelin 1 (EDN1).

3.1.5 Anaerobic Respiration

In oxygen deficient environments, cellular oxygen supply decreases leading to a shift from aerobic respiration to anaerobic respiration in anurans. This leads to more glycolysis to compensate for the lack of energy, unfortunately not producing nearly enough energy compared to aerobic respiration and that energy produced by glycolysis is also

accompanied with the undesirable byproduct of lactic acid. Glycolysis is not sufficient to preserve life in anurans for extended amounts of time. As aforementioned, HIF-1 upregulates glycolytic enzymes to increase the amount of anaerobic respiration and subsequent energy production.

3.1.6 Vascular Tone

As anuran circulatory and respiratory systems are not completely separate, alterations to vascular tone can help improve survival outlook in hypoxic settings. Vasodilation mediating proteins such as iNOS widen blood vessels [143]. The widening of the vessels allows for an increase in blood flow which helps circulate more oxygen throughout the body if there is oxygen available. If severe oxygen deficiency persists, vasoconstrictor mediating proteins such as EDN1 allow for the narrowing of blood vessels prioritizing internal organs with available oxygen.

3.1.7 Apoptosis Inhibition

As cells remain in hypoxic conditions for extended periods of time, they will elicit programmed cell death. Apoptosis inhibitors, like BCL2, act to prevent cells from initiating programmed cell death by inhibiting BH3 mediated pro-apoptotic effector oligomerization, which would otherwise release a cascade of pro-apoptotic molecules [144-145]. HIF1 is known to upregulate BCL2 in response to low oxygen availability, which in turn would prolong the lifespan of cells in oxygen deficient environments.

3.1.8 Our Hypothesis and Direction

We analyze *de novo* protein reconstructions and alignments pertaining to oxygen transport, delivery, and conservation systems from 151 species of anurans to characterize any signs of significant modification to peptide structure. We monitor gene-wide

selection events as well as peptide-specific essential motifs and their proximity which tend to be highly conserved due to their functional importance. We hypothesize that proteins pertaining to oxygen transport, delivery, and conservation systems will demonstrate high levels of positive selection across anurans as they are all characterized in literature to be exhibit various levels of hypoxia/anoxia resilience [146-147].

3.2 MATERIALS AND METHODS

3.2.1 Dataset Selection and Criteria

Our broad-scale anuran study utilizes 151 species of anurans and 353 distinct RNA-seq datasets from local and online sources (Supplementary Table S4). Of the chosen 353 anuran datasets, 87 were from the National Center for Biotechnology Information (NCBI; i.e. SRR), 5 from the European Molecular Biology Laboratory, European Bioinfomatics Institute (EMBL-EBI; i.e. ERR), 22 NCBI reference transcripts, 5 NCBI genomic references and the remaining 234 were local datasets (Supplementary Table S4). SRR and ERR Sequence Read Archive (SRA) files were constrained to transcriptomic Illumina RNA-seq datasets. As our transcripts of interest are not constrained to a particular tissue type, SRA files were selected independent of tissue type. To ensure high sequence coverage, SRA files were chosen based on proximity to 5.2 billion bases total, where files with more bases were favored at a factor of five versus files containing less than 5.2 billion bases. Local files stemmed from laboratory whole transcriptome sequencing, assembly, and annotation.

The genomic counterpart of our experiment utilizes fourteen species of anurans from their respective genomic datasets on the National Center for Biotechnology Information (NCBI) genome database (Table 6). We utilized exon 2 from an annotated *Xenopus tropicalis* (*X. tropicalis*) HBA1 nucleotide sequence (NM_203529.1) as our query against all anuran genomes (35 genomes total, representing 29 unique species) on the NCBI (as of 10/7/2022) with online NCBI BLAST. We downloaded all aligned sequence results from BLAST as a FASTA file.

3.2.2 Reference Genes for SRA datasets

Full length coding sequences (CDS) were obtained directly from the NCBI nucleotide database. *Xenopus tropicalis* (*X. tropicalis*) was chosen as a reference for HBA1 (NM_203529.1), HBA2 (NM_001142252.1), HBB1 (NM_203528.2), HBB2 (NM_001004899.1), CYGB (NM_001006869.1), HIF1A (NM_001011165.2), HIF2A (NM_001005647.1), HIF1B (NM_001123453.1), ALDOA (NM_001005643.1), GAPDH (NM_001004949.2), EDN1 (XM_002932664.5) and iNOS (NM_001126513.1). *Xenopus laevis* (*X. laevis*) was chosen as the reference for enolase 1 (ENO1; NM_203813.1). *Bufo bufo* (*B. bufo*) was chosen as the reference for B-cell lymphoma-2 (BCL2; XM_040431930.1). References were split into 80 base-pair (bp) fragments with a 40 bp overlap. Any end sequence shorter than 40 bp was removed. List of fragments (LoF) generated from fragmenting the CDS were compared to original CDS to ensure full representation.

3.2.3 Single Gene Harvesting from SRA datasets

As whole transcriptome assembly tends to fragment or completely miss transcripts, we elicited a single gene reconstruction approach to ensure accurate full-length transcript assembly for our SRA datasets [32]. SRA toolkit's blastn vdb [47-48] was utilized to align each LoF against each SRA accession number separately, resulting in a subset list of SRA read IDs. The subset read IDs were then used with SRA toolkit's fastq-dump [48] to extract all matching reads from the SRA dataset. Extracted reads were assembled with rnaSPAdes [3; 23] under default settings. Final transcripts were validated online via a NCBI blastn against the NCBI non-redundant nucleotide (nt/nr) database [148].

3.2.4 Whole Transcriptome Assembly

Whole transcriptome assembly for non-SRA samples was conducted as follows: Total RNA was extracted from various tissue (skin, gut, liver, brain, muscle, tongue, intestine, eggs, eyes, and whole head) using the Trizol total RNA extraction [86]. Total RNA was quantified with nanodrop, integrity determined with bioanalyzer, and kept at -80C. Only samples with RIN>7.0 were used for mRNA isolation, directional RNAseq library preparation, and Illumina 150 bp paired-end sequencing. Briefly, after the QC procedures using bioanalyzer, mRNA was isolated using oligo(dT) beads and rRNA was removed using the Ribo-Zero kit. The isolated mRNA was used for cDNA synthesis and 150 bp pair-end sequenced directionally.

Raw sequence reads (Supplementary Table S4) were processed with Pincho v0.1 [6; 87]. We selected the recommended high quality full run of the pipeline, which automates (a) removal of Illumina adapter sequences with Trimmomatic [7-8], (b) error correction with Rcorrector [9-10], (c) assembly with trans-ABySS [27-28] and rnaSPAdes [3; 23] at five computationally generated k-mer sizes, and TransLig [29-30] at default settings, (d) consensus assembly generation with TransRate [32-33], (e) redundancy reduction with CD-HIT [34-35], (f) assembly assessment with BUSCO [38-39] and BUSCO's eukaryota_odb10.2019-11-20 dataset, and finally (g) annotation against UniProt: Swiss-Prot with BLASTx [40-42].

3.2.5 Gene Alignment and Phylogenetic Trees

We aligned assembled transcripts with Clustal Omega [149] in JalView [150] and manually trimmed areas before the start codon and after the stop codon if necessary. Stop

codons were removed and the resulting transcripts were exported in FASTA format. Phylogenetic trees of the resulting transcripts were constructed on TimeTree [151].

The genomic counterpart of the study required all nucleotide sequences be converted into all three possible reading frames for the positive strand. We then translated our nucleotide sequences to transcripts and removed all transcripts with in-line stop codons and subsequently also removing their nucleotide precursors. We aligned the surviving nucleotide sequences with Clustal Omega in JalView. We also translated nucleotide sequences to transcripts again and realigned with Clustal Omega and sorted sequences by pairwise identity. Transcripts were divided into groups and the group that closely matched *X. tropicalis* HBA1 exon 1 was kept for analysis. *X. tropicalis* HBA1 exon 2 was then BLASTED against surviving transcripts to choose only one transcript per species that had the highest percent identity. Sequences were manually trimmed to conform with HYPHYs [152] parameters (i.e. nucleotide sequences must be divisible by three and must have complete codons). A phylogenetic tree of the resulting sequences was constructed on TimeTree. Two species were not present on TimeTree and were excluded from the study (*Phrynoglossus myanhessei* and *Platyplectrum ornatum*) leaving only fourteen species for analysis.

3.2.6 Maximum Likelihood (ML) Analysis

Sites under episodic positive diversifying selection were determined with the Hyphy's Mixed Effects Model of Evolution (MEME; $p<0.05$) [153], Fixed Effects Likelihood (FEL; $p<0.05$) [154] and Fast Unconstrained Bayesian Approximation (FUBAR; $PP>0.95$) [155]. Gene-wide episodic diversifying selection was determined with Branch-Site Unrestricted Statistical Test for Episodic Diversification (BUSTED;

$p<0.05$) [156]. FEL ($p<0.05$) was utilized for site-wise purifying selection. Branch-wise episodic diversifying selection was determined with Adaptive Branch-Site Random Effects Likelihood (ABSREL; $p<0.05$) [157].

3.2.7 Hematological Consequence

We utilized the Globin Gene Server [158] to determine which missense mutations of HBA 1 and 2 are linked with disorders in humans. Mutations with no known disorder were not included in the study. In total, 44 variants were utilized, spanning 37 distinct sites on HBA. HBA 1 and 2 sequence alignments were merged prior to analysis for hematological consequence. We included 216 distinct species (chordates, 1459 HBA 1 and 2 sequences) from the NCBI protein database in a separate analysis from our anuran dataset. Sequences that were shorter or longer than 142 amino acids were excluded from the study. The number of occurrences of each mutation was determined in Microsoft Excel.

3.3 RESULTS

3.3.1 Oxygen Transport and Delivery Globins

Hemoglobin transcripts demonstrated a high number of sites under episodic diversifying selection (Fig. 14, Table 7). Adult hemoglobin subunits HBA1 (14.79%), HBA2 (17.61%) and HBB2 (10.88%) exhibited higher percentages of positive selection when compared to juvenile hemoglobin subunits HBA3 (6.34%) and HBA5 (8.39%; Table 6). Adult subunit HBB1 (1.36%) was an exception, exhibiting the least amount of episodic diversifying selection (Table 7). ABSREL predicted episodic diversifying selection in the following families: Bufonidae, Dendrobatidae and Hylidae across HBA1, HBA2 and HBB2; Craugastoridae across HBA1 and HBA2; Ascaphidae and Pelobatidae in HBA1; Aromobatidae and Eleutherodactylidae in HBA2; and Bombinatoridae in HBB2 (Fig. 15). Site-wise episodic diversifying selection affected sites neighboring the proximal *histidine* less than the distal *histidine* (Fig. 14, Table 8). In 7.13% of the cases between hemoglobin subunits, the distal histidine was replaced by glutamine or leucine (Table 8). In 0.89% of the cases between hemoglobin subunits the proximal histidine was replaced by a tyrosine (Table 8).

Hemoglobin alpha 1 exon 2 demonstrated a singular site under episodic diversifying selection (Fig. 16). Site 26 of HBA1 exon 2 was predicted to be under episodic diversifying selection by both HYPHY's FEL ($p<0.007$) and FUBAR ($PP>0.96$). FEL predicted 19 sites under negative selection, meaning roughly 29% of exon 2 to be under purifying selection ($p<0.05$; sites 2, 6, 7, 10, 12, 13, 16, 27, 30, 36, 41, 45, 46, 51, 52, 57, 63, 64, and 66). HYPHY's ABSREL predicted *Xenopus tropicalis* and *Phyllomedusa bahiana* branches were under episodic diversifying selection ($p<0.01$).

Cytoglobin transcripts demonstrated 3.91% episodic diversifying selection (Table 6). Sites neighboring the proximal and distal histidine in cytoglobin generally exhibited purifying selection and an overall lack of positive selection (Fig. 14). Negative selection was detected in 57.54% of cytoglobin sites, which is the second highest among the hemoglobin subunits (Table 7).

3.3.2 Hypoxia Inducible Factors

Of all the HIFs tested, only HIF1A was found by BUSTED to be under gene-wide episodic diversifying selection (Fig. 14). HIF1A was under slight positive selection at 4.58% of total sites, while HIF2A was at 1.25% and HIF1B was at 0.63% of total sites under positive selection (Table 7). HIF1A was also under the greatest amount of purifying selection when compared to HIF2A and HIF1B, with 71.57% of its sites under negative selection compared to 48.64% and 54.04% for HIF2A and HIF1B respectively (Table 7).

3.3.3 Vascular Tone

Neither EDN1, at 1.35% positively selected sites, nor iNOS, at 0.89% positively selected sites, were predicted by BUSTED to exhibit gene-wide episodic diversifying selection. iNOS experienced a greater magnitude of negative selection than EDN1 with the percentages being 60.11% and 24.66% respectively (Fig. 14, Table 7). This provides us with sufficient evidence to conclude that both vascular tone peptides are highly conserved.

3.3.4 Anaerobic Respiration

ALDOA and GAPDH both exhibited transcript-wide evidence of episodic diversifying selection via BUSTED but ENO1 did not (Fig. 14). Only 1.65% of the sites

in ALDOA were under episodic diversifying selection, while GAPDH had 5.46% sites under positive selection (Table 7). All three transcripts were observed to have purifying selection in above 50% of their sites (ALDOA with 54.12%, GAPDH with 63.14% and ENO1 with 56.46%; Table 7). Upon inspection of ALDOA protein alignments, sites in ALDOA that are responsible for binding and catalysis of fructose 1,6-bisphosphate (Ser271, Gly272, Lys41, Arg42, Arg303 and Lys229) were highly conserved, showing no signs of positive selection [159]. A site known to drop ALDOA activity was also void of positive selection [159].

3.3.5 Apoptosis Inhibition

Under 0.67% of the sites in BCL2 exhibited episodic diversifying selection, with BUSTED being unable to predict transcript-wide episodic diversifying selection (Table 7, Fig. 14). The overwhelming majority of sites across BCL2 transcripts were not under positive nor negative selection (Table 7). BCL2 was subsequently more conserved than all under transcripts in the study.

3.3.6 Hematological Disorders in Hemoglobin Alpha

Of 37 sites monitored for missense mutations correlating with hematological disorder in our anuran HBA dataset consisting of 82 distinct species (92 sequences), we noted 4 sites (site 41, 75, 82, and 116) with missense mutations correlating to hematological consequence in humans (Miyano, Chapel Hill, Nigeria, and J-Tongariki respectively; Table 9). Hematological disorders include erythrocytosis in 1 species, middle anemia in 8 species, and microcytosis and hemochromatosis in 67 species (76 sequences; Table 9). In total 67 of 82 species are predicted to exhibit a hematological disorder (Table 9).

The NCBI protein sequences for HBA of 216 chordate species yielded only 21 species with missense mutations corresponding to hematological disorder (Table 9). The Shenyang mutant (site 27) causing anemia in humans was only found in *Rattus norvegicus,* Hopkins-II mutant (site 113) causing minor hemolysis was found in 19 species (11 hummingbirds: *Archilochus alexandri, Boissonneaua matthewsii, Chalcostigma ruficeps, Chalcostigma stanleyi, Colibri coruscans, Eriocnemis luciani, Haplophaedia aureliae, Heliodoxa leadbeateri, Pterophanes cyanopterus, Schistes geoffroyi*, and *Selasphorus platycercus*, and 8 perching birds: *Cinclodes albiventris, Conirostrum cinereum, Diglossa brunneiventris, Diglossa glauca, Notiochelidon cyanoleuca, Notiochelidon murina, Tangara nigroviridis,* and *Tangara vassorii*), and the J-Tongariki mutant (site 116) causing erythrocytosis was found in 2 species (*Varanus komodoensis* and *Orycteropus afer afer;* Table 9).

3.4 DISCUSSION

Arguably one of the most interesting results is how much episodic diversifying selection was observed in adult hemoglobin transcripts. Not only was there significant evidence of positive selection, but there were also signs of positive selection neighboring crucial components implicated in binding oxygen. Distal histidine residues and their neighboring amino acids are particularly necessary to form the right shape for holding oxygen on the heme group of which would be attached to the proximal histidine residue. We observed significant variation in neighboring residues to both the proximal and distal histidine as well as changes to the histidine residues. The distal histidine was often mutated to glutamine, which is common among the literature for distal histidine mutations; however, the distal histidine was never recorded to be mutated to a leucine as we noted in our alignments. In addition, the proximal histidine, which is not recorded to have an alternative residue, has been recorded as a tyrosine in some of our samples. We speculate that anurans exhibit high amounts of positive selection and pointed selection directed at these sites of heme binding and oxygen affinity as an adaptation of encountering areas of low oxygen availability.

We sought out genomic HBA1 exon 2 transcripts to help support our high levels of positive selection in hemoglobin. In the genomic dataset we were only able to locate a single site of episodic diversifying selection adjacent to the distal histidine residue. We further note that among the sixteen tested anurans, the amino acids at site 26 differ greatly in their helix propensity leading us to believe that this amino acid substitution could lead to altered protein structure and therefore a change in oxygen binding affinity.

The HIF-1 complex is responsible for most of the metabolic response to hypoxia, so it was interesting to observe greater amounts of episodic diversifying selection in HIF1A than for HIF2A, which has a much smaller role in hypoxia response. The HIF-1 complex recognizes specific HRE motifs upstream of the genes it upregulates; therefore, the large amount of purifying selection is justified.

Glycolytic enzymes ALDOA and ENO1 exhibited high conservation and no significant change to crucial enzymatic portions of their structures, while GAPDH exhibited significant signs of episodic selection but again, not on any positions crucial to its enzymatic activity. It is possible that the changes could alter the shape of these enzymes overall and affect the potency via reaction rate; however, the structure could not be determined at this time.

In the case of proteins that mediate vascular tone, EDN1 and iNOS exhibited high sequence conservation across tested anuran species. As EDN1 is a ligand, we expected little to no change in overall peptide sequence and the results supported our prediction. Ligands must conform to a particular shape to associate with their respective receptor. Any change in peptide sequence risks altering the shape and making the ligand inept in cell signaling for vasoconstriction. In the same regard, the iNOS enzyme also exhibited very little episodic diversifying selection.

Apoptosis regulator BCL2 exhibited the least amount of positive and negative selection; however, during gene expression, only a portion of the BCL2 transcript will be translated into a functional protein. The portion which will translate into a protein in anurans is currently unknown. Therefore, it will be difficult to determine the exact location of areas of interest, which would be the BH3-binding domain, within the anuran

BCL2 which we could monitor for signs of positive selection or conservation. We instead monitor signs of positive selection across the entirety of the transcript, even if only part of it will ultimately be translated and we concluded that BCL2 is also highly conserved. As an essential inhibitor of apoptosis, high sequence conservation is again expected.

A vast number of anurans (~82%) tested contained at least one mutation that coincides with hematological disorder in human HBA sequences compared to only <10% of the species with HBA sequences in the NCBI protein database. This disparity demonstrates how peculiar anuran HBA sequences are. Some anurans have up to 3 different mutations simultaneously. If we were to assume the human hematology applies to anurans, microcytosis and hemochromatosis would be the most common in anurans when compared to all other species in the NCBI dataset. Microcytosis is a condition in which the red blood cells (RBCs) are smaller than normal, while hemochromatosis is a condition where the RBCs are paler than normal. Anuran RBCs are roughly 3 times larger than human RBCs. Small RBCs provide more efficient gas exchange and circulation, which is ideal for organisms with a fast metabolism and high energy demands; however, anurans typically have slower metabolisms and lower energy demands [160]. Due to this fundamental difference, anuran RBCs can afford to be larger without detrimental side effects; however, the smaller the RBCs can become, the more efficiency they can gain. This leads us to believe that if anurans do exhibit microcytosis as they are predicted to, it would subsequently aid their oxygen transport and delivery systems. Of 44 mutations that lead to hematological disorder in humans, only 3 were found among the NCBI protein database for HBA and only 4 were found among our anuran dataset showing that both groups generally avoid these missense mutations.

There is a great amount of research on anurans that have adapted to environments with extreme hypoxic conditions. *Rana amurensis* has adapted extreme hypoxia tolerance to survive in areas with no oxygen for months [161]. *Telmatobius culeus*, which lives 3800 meters above sea level, exhibited adaptations allowing higher oxygen affinity and smaller blood cells than frogs at sea level [162]. There are even frogs that seek out hypoxic conditions to aid in hibernation in the colder climates, such as *Cyclorana alboguttata* [163]. There is research supporting the idea that all frogs are hypoxia/anoxia resilient, but there is also research that proposes that amphibians are anoxia-intolerant [164]. Overall, all studies on hypoxia in anurans can benefit from ML analysis to determine the direction of protein evolution.

3.5 CONCLUSION AND FUTURE DIRECTIONS

Episodic diversifying selection aggregates towards transport and delivery of oxygen instead of hypoxia mediated processes. Among these positively selected sites in transport systems, we note several potentially significant alterations neighboring the proximal and distal *histidine* positions of hemoglobin subunits, potentially affecting oxygen-binding affinity. Hemoglobin subunits are difficult to reconstruct and properly annotate due to their high sequence similarity amongst subunits. HBA1 and HBA2 are nearly identical and therefore it is difficult to determine which is alpha 1 and which is alpha 2 forcing us to rely on BLAST evalues and alignments as the sole determinants of sequence identity. The same holds true for HBB1 and HBB2. We included the genomic counterpart to validate our findings on HBA1 exon 2 on a smaller scale with sequences that were, without a doubt, HBA1. Although we did not find the same levels of episodic diversifying selection, we did find that the positive selection was still located adjacent to the zones crucial for oxygen binding (the distal histidine). This proves to us that these sequences are changing in anurans and that these changes can potentially have significant consequences and/or benefits for the species.

We noted a significant number of HBA sequences from our tested anurans contained mutations that signify hematological consequence in humans. There is no literature available that links these disorders to anurans, and it would be difficult to assess most hematological disorders in anurans; however, hematological consequences such as microcytosis could be determined with a simple blood smear. We could compare the sizes of the red blood cells among species of the same genus, those which have the mutation and those that do not. This way we can observe if those with the mutation have

smaller red blood cells than the others in their genus (i.e., *Bufo tibetanus* vs. *Bufo bufo*, *Rhinella arenarum* vs. *Rhinella margaritifera*, and *Litoria rubella* vs. *Litoria fallax*).

An unavoidable drawback is the unequal representation among the anuran families, with some groups being overrepresented and others left with only one species to represent the entire family. Current online transcriptomic datasets heavily favor some families over others. Our study focuses on only 8 of 44 families of anurans (Scaphiopodidae, Bombinatoridae, Pipidae, Limnodynastidae, Leptodactylidae, Bufonidae, Hylodidae, and Ranidae). It will take years for the underrepresented families to appear in online databases, but the outlook is promising. We are currently working towards creating bioinformatic tools to better reconstruct hemoglobin transcripts; however, perhaps full-length sequencing for hemoglobin subunits may be the best remedy for the problem. We are also currently working towards adding more transcriptomic data pertaining to unique anuran species to online repositories, SRA, in the hopes that future research may continue with less uncertainty pertaining to family representation. We contribute all aligned sequences used in this study in supplemental materials for future research and directions.

3.6 TABLES AND FIGURES

3.6.1 TABLES

Table 6. Genomic Dataset ID and Partial Hemoglobin alpha 1 Exon 2 Sequences

Species	GenBank ID	HBA1 Partial Exon 2 Protein Sequence	HBA1 Partial Exon 2 Nucleotide Sequence
Bombina variegata	CAJPDP011885949.1	----- PQTKTYFPNFDFSKNSAQ IKAHGKKVVDALTEAVNH LDNIEGCLSKLSDLHAFNL RVDPGN--	---------- CCCCAGACAAAAACCTACTT CCCAAACTTTGACTTCAGCA AGAACTCAGCCCAAATC AAGGCTCATGGCAAGAAGG TGGTGGATGCCCTGACTGA GGCTGTTAATCACCTGGAC AACATCGAAGGCTGC CTGTCTAAACTGAGTGACCT TCATGCCTTCAATCTGAGAG TGGATCCTGGCAAC------
Engystomops pustulosus	WNYA01000008.1	MFLSNPQTKTYFPKFDCS KDSAHVKSHGKKVLDALT ETVKHLDNIDGALCKLSE LHAYDMRVDPGNFP	ATGTTCCTTTCCAATCCTCA AACCAAGACCTACTTTCCAA AGTTTGATTGTTCCAAGGAT TCTGCCCATGTT AAATCTCACGGCAAGAAAG TACTTGATGCTCTGACTGAA ACAGTAAAACACCTGGACA ACATTGATGGAGCC CTTTGTAAGCTAAGTGAACT CCATGCTTACGACATGAGA GTGGACCCAGGAAATTTCC CA
Glandirana rugosa	BLSH010115312.1	MFVCHPQTKTYFPNFDF HANSPNLKSHGKKVMNA LTEVVKHLDHPEAALSHL SDLHAYSLRVDPGN--	ATGTTCGTTTGCCATCCTCA AACCAAGACTTACTTTCCCA ATTTTGACTTCCACGCAAAT TCTCCTAACCTA AAGAGTCATGGCAAGAAGG TAATGAATGCTCTGACTGAA GTGGTCAAACATTTGGACC ACCCTGAGGCCGCC TTGAGCCATCTGAGTGATCT CCATGCCTATAGCTTGAGG GTGGATCCTGGCAAC------
Hymenochirus boettgeri	JAACNH010000009.1	MFLSNPQTKTYFPKFDFH KDSAQMKAHGKKVVDAL TEASNHLDNIGGTLSKLS DLHAHDMRVDPGNF-	ATGTTCCTTTCAAACCCCCA GACCAAAACCTACTTTCCCA AATTTGACTTCCACAAGGAT TCAGCACAGATG AAGGCTCATGGGAAGAAAG TGGTGGATGCCCTAACTGA GGCTTCAAACCATCTGGAC AACATTGGAGGAACC CTGAGCAAACTTAGCGACC TCCATGCTCATGATATGAGA GTGGATCCTGGCAACTTC---
Limnodynastes dumerilii	WWET01002829.1	MFICHPQTKTYFPDFDFH KDSPHILKHGKKVLDAITE	ATGTTCATTTGCCATCCTCA GACCAAGACTTACTTCCCTG

		ASKHLDNIEGALDKLSDL HAFKLRVDPGN--	ATTTTGATTTCCACAAGGAT TCTCCTCATATT CTGAAGCATGGCAAGAAGG TTCTTGATGCTATCACTGAA GCTTCCAAACACTTGGACAA CATTGAGGGAGCC CTGGATAAACTGAGTGATCT ACATGCCTTCAAACTGAGA GTGGATCCTGGCAAC------
Lithobates catesbeianus	LIAG020380476.1	MFLCHPQTKTYFPNFDFH ANSAHLKNHGKKVMNAL TDAVKHLDHPEASLSSLS DLHAFTLRVDPGN--	ATGTTCCTTTGCCATCCTCA AACCAAGACTTACTTTCCCA ATTTTGACTTCCACGCAAAT TCTGCTCACCTA AAGAATCATGGCAAGAAGG TAATGAATGCTCTGACTGAT GCAGTCAAACATCTGGACC ACCCTGAGGCCTCA TTGAGCTCATTGAGTGATCT CCATGCCTTTACCTTGAGG GTGGATCCCGGCAAT------
Phyllomedusa bahiana	JAODAL010001895.1	- FLCSPRTKTYFPNFDFHC NSPQLKAHGKKVIDALTD AVKHLDNIDAALDKLSSL HAFDLRVDPGNFP	--- TTCCTCTGCAGTCCCAGGA CCAAGACCTACTTCCCCAAT TTTGACTTCCATTGCAATTC TCCTCAGTTG AAGGCTCACGGCAAGAAAG TAATTGATGCTCTGACTGAT GCTGTAAAACATCTGGACAA CATTGATGCAGCC CTGGATAAGCTGAGTAGCC TCCATGCTTTTGATCTGAGA GTGGATCCTGGAAACTTCC CA
Rana temporaria	CAJIML010007699.1	MFLCHPQTKTYFPTFDFH ANSAQLKGHGKKVMNAL TEVVKHLDHPEASLSHLS DLHAFTLRVDPGN--	ATGTTCCTTTGCCACCCTCA AACCAAGACATACTTTCCCA CTTTTGACTTCCACGCAAAT TCTGCTCAACTA AAGGGTCATGGCAAGAAGG TAATGAATGCTCTGACTGAA GTGGTCAAACATCTGGACC ACCCTGAGGCCTCA TTGAGCCATCTGAGTGATCT CCATGCCTTTACCTTGAGG GTGGATCCTGGCAAC------
Rhinella marina	ONZH01021842.1	MFLSNPQTKTYFPAFDFH KDSPHIRGHGKKVVDALT EASKHLDNIDGALNKLSEI HAFDLRVDPGNFP	ATGTTTCTTTCCAATCCCCA GACCAAGACCTACTTCCCT GCTTTTGACTTCCACAAGGA CTCTCCTCATATT AGGGGTCATGGCAAGAAAG TAGTTGATGCTCTGACTGAA GCTTCAAAACACCTGGACA ACATTGATGGAGCT CTGAATAAGCTGAGTGAAAT CCATGCTTTTGACCTGAGA GTGGATCCTGGAAACTTCC CA
Scaphiopus holbrookii	VKOB010760974.1	MFNACPQTKTYFPKFDC SKESAHVRAHGKKVFDA LTEAANHLDNIPGCLSKL SDLHAYDLRVDPGN--	ATGTTCAATGCGTGCCCTCA GACCAAGACTTACTTCCCCA AATTTGACTGCAGCAAGGA ATCTGCTCATGTC AGAGCTCATGGTAAGAAGG TATTTGATGCCCTGACAGAG GCTGCAAACCACCTGGACA ACATCCCAGGATGC CTGAGCAAGCTCAGTGACC TCCATGCTTATGATTTGAGA GTGGATCCTGGCAAT------
Spea bombifrons	VKNZ011483092.1	- FISSPQTKTYFPKFDFHK DSPHIRAHGKKVFDALTE	--- TTTATCTCAAGCCCTCAGAC CAAGACTTACTTCCCCAAAT

		AANHLDNIPGCLSKLSDL HAYDLRVDPGN--	TTGACTTCCACAAGGACTCT CCTCATATC AGAGCTCATGGTAAGAAGG TATTTGATGCCCTGACGGA GGCTGCAAACCACCTGGAC AACATCCCAGGATGC CTGAGCAAGCTCAGTGACC TCCATGCTTATGATCTGAGA GTGGATCCTGGCAAT------
Xenopus borealis	JAMBMQ010000018.1	MFLINPKTKTYFPNFDFH HNSKQISAHGKKVVDALN EAANHLDNIAGSLSKLSD LHAYDLRVDPGNFP	ATGTTCTTGATTAATCCTAA AACCAAAACCTACTTTCCTA ATTTTGACTTCCACCACAAT TCAAAACAAATC AGTGCTCATGGCAAGAAAG TTGTGGATGCTCTGAATGAA GCTGCCAACCACTTGGATA ACATTGCTGGAAGC CTGAGCAAGCTGAGTGACC TCCATGCCTATGACTTGAGA GTGGATCCGGGCAACTTTC CA
Xenopus laevis	LYTH01182431.1	MFIVNPKTKTYFPSFDFH HNSKQISAHGKKVVDALN EASNHLDNIAGSMSKLSD LHAYDLRVDPGNFP	ATGTTCATAGTCAACCCCAA GACCAAAACTTACTTCCCTA GTTTTGACTTCCACCACAAT TCAAAACAGATC AGTGCTCATGGCAAGAAAG TTGTGGATGCTCTGAATGAA GCTTCCAACCATTTGGATAA CATCGCTGGAAGC ATGAGCAAGCTGAGTGACC TCCATGCCTATGACCTGAG AGTGGACCCTGGCAACTTC CCA
Xenopus tropicalis	JABAHN010000009.1	MFMCAPKTKTYFPDFDF SEHSKHILAHGKKVSDAL NEACNHLDNIAGCLSKLS DLHAYDLRVDPGNFP	ATGTTCATGTGTGCTCCCAA GACCAAAACCTACTTTCCTG ATTTTGACTTCAGCGAACAT TCAAAACACATC TTGGCTCATGGCAAGAAAG TTTCGGATGCTCTGAATGAG GCTTGCAACCATCTGGACA ACATTGCCGGATGC CTGTCCAAGCTGAGTGACC TCCATGCCTATGACCTGAG AGTGGATCCAGGCAACTTC CCA

Table 7. HYPHY Site-wise Selection Summary

	Species	*Sites*	*Positive Sites*	*Negative Sites*	*Neutral Sites*	*% Positive Selection*	*% Negative Selection*
HBA1	20	142	21	65	56	14.79	45.77
HBA2	72	142	25	70	47	17.61	49.30
HBA3	24	142	9	45	88	6.34	31.69
HBA5	76	143	12	84	47	8.39	58.74
HBB1	14	147	2	36	109	1.36	24.49
HBB2	50	147	16	83	48	10.88	56.46
CYGB	111	179	7	103	69	3.91	57.54
HIF1A	111	830	38	594	198	4.58	71.57
HIF2A	52	882	11	429	442	1.25	48.64
HIF1B	28	792	5	428	359	0.63	54.04
BCL2	36	299	2	69	228	0.67	23.08
ALDOA	38	364	6	197	161	1.65	54.12
ENO1	32	441	13	249	179	2.95	56.46
GAPDH	54	293	16	185	92	5.46	63.14
EDN1	18	223	3	55	165	1.35	24.66
iNOS	53	564	5	339	220	0.89	60.11

Table 8. Average Unconserved Sites (%) in Hemoglobin Distal and Proximal Zones

	4	*3*	*2*	*1*	*0*	*1*	*2*	*3*	*4*
Distal Zone	9.97	0	29.16	26.52	7.13	0	5.32	1.65	0.55
Proximal Zone	0	5.88	6.79	8.22	0.89	0.22	11.20	19.47	0

Table 9. Hematological Consequence within Hemoglobin Alpha 1

Hb Name	Hematology	Mutation	Occurrences in Anurans (of 82 species)	Occurrences in NCBI Database (of 216 species)
Shenyang	anemia	27E		1
Hopkins-II	minor hemolysis	113R		19
Nigeria	microcytosis and hemochromatosis	82C	22	
J-Tongariki	microcytosis and hemochromatosis	116D	65	2
Miyano	erythrocytosis	41S	1	
Chapel Hill	mild anemia	75G	8	
Total			**67 / 82**	**22 / 216**

3.6.2 FIGURES AND FIGURE LEGENDS

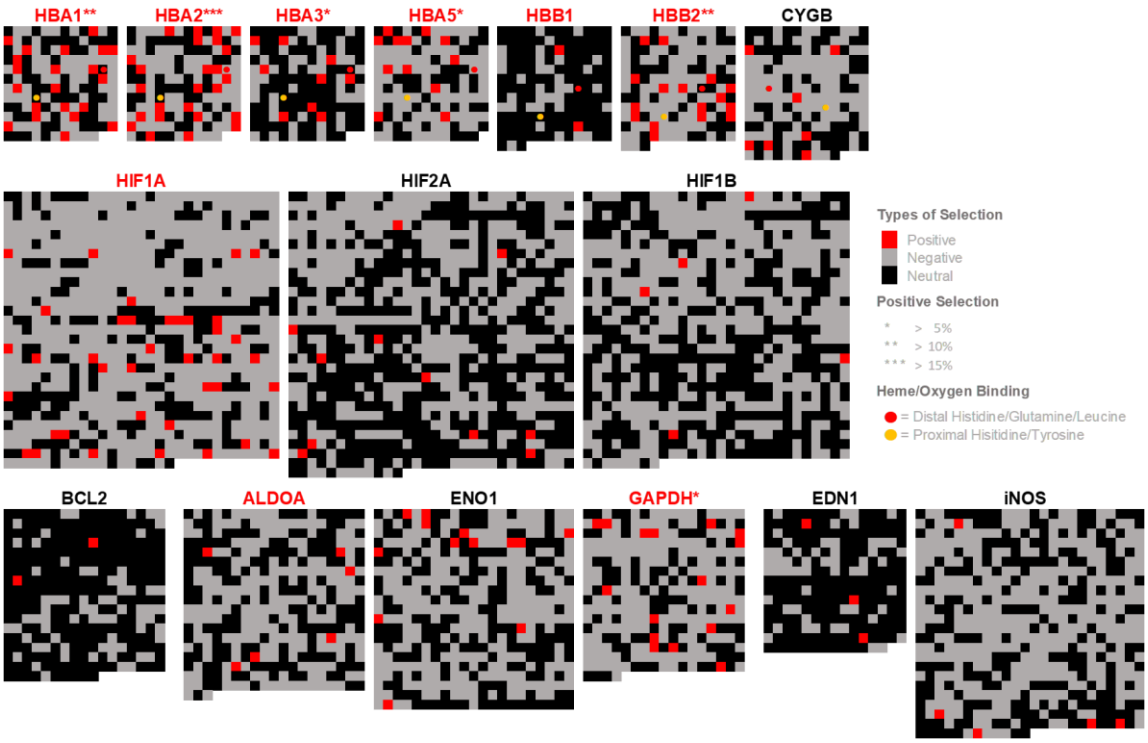

Figure 14. Signs of Episodic Diversifying Selection in Oxygen Transport, Consumption and Delivery Systems.

Sites under episodic diversifying selection are determined with MEME ($p<0.05$), FEL ($p<0.05$) and FUBAR ($PP>0.95$) and are highlighted in red. Gene names are highlighted in red if supported by BUSTED ($p<0.05$) or black if not supported by BUSTED ($p>0.05$). Sites under purifying selection are determined with FEL and are highlighted in light gray. Sites not determined to be under selection are highlighted in black. Sequences are fasta-wrapped with each square representing one site. Distal histidine/glutamine/leucine sites are denoted with a red circle and proximal histidine/tyrosine are denoted with an orange circle.

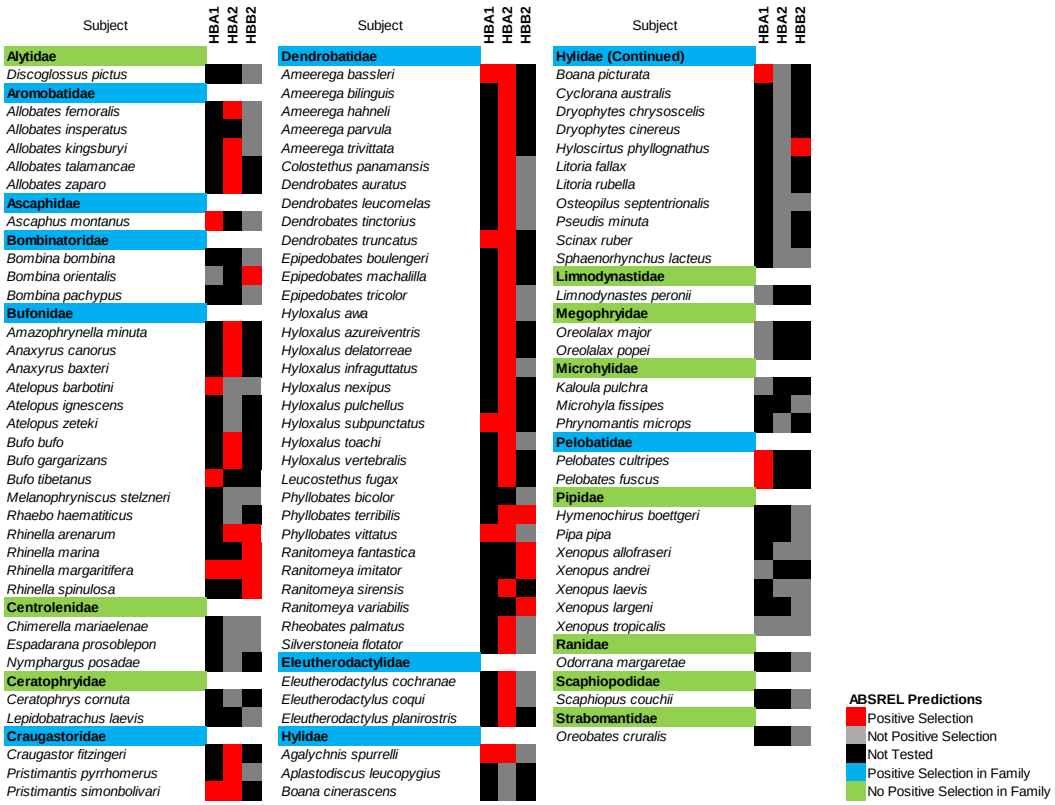

Figure 15. ABSREL Results for Hemoglobin.

Species under episodic diversifying selection are determined with ABSREL ($p<0.05$) and are highlighted in red. Sites not listed under positive selection are denoted in gray, while sites not tested are denoted in black. Positive selection within each family is denoted in light blue for present or green for absent.

Figure 16. Hemoglobin Alpha 1, Partial Exon 2 Conservation Alignment Demonstrates Signs of Episodic Diversifying Selection Adjacent to Distal Histidine.

Partial protein alignment of hemoglobin alpha 1 exon 2 across 14 distinct species of anurans. Amino acids are color coded via clustalx default coloration against amino acid conservation discoloration (white). Sites which contain conservation above 30% are in color while sites that do not have at least 30% conservation are uncolored. Site 26, denoted with a green arrowhead, was found to be under positive episodic diversifying selection under both HYPHY's FEL ($p<0.007$) and FUBAR ($PP>0.96$). Distal histidine sites, site 27, are denoted with a red arrowhead and proximal histidine sites, site 56, are denoted with an orange arrowhead. Sites under purifying selection predicted by FEL ($p<0.05$) are denoted by blue circles.

REFERENCES

1. Martin, J.; Bruno, V.M.; Fang, Z.; Meng, X.; Blow, M.J.; Zhang, T.; Sherlock, G.; Snyder, M.; Wang, Z. (2010). Rnnotator: An automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genom. 11*, 663.
2. MacManes, M.D. (2018). The Oyster River Protocol: A multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ. 6*, e5428.
3. Bushmanova, E.; Antipov, D.; Lapidus, A.; Prjibelski, A.D. (2019). rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience. 8*, giz100.
4. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.. 29*, 644–652.
5. Kannan, S.; Hui, J.; Mazooji, K.; Pachter, L.; Tse, D. (2016). Shannon: An Information-Optimal de novo RNA-Seq Assembler. *bioRxiv.* 39230.
6. Pincho. Pincho (Version 0.1). 2021. Available online: https://github.com/RandyOrtiz/Pincho/releases/tag/v01 (accessed on 6 June 2021).
7. Bolger, A.M.; Lohse, M.; Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics. 30*, 2114–2120.
8. Bolger, A.M.; Lohse, M.; Usadel, B. (2014). Trimmomatic. *Bioinformatics.* Version 0.39. Available online: http://www.usadellab.org/ cms/?page=trimmomatic (accessed on 6 June 2021).

9. Song, L.; Florea, L. (2015). Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience. 4*, 48.
10. Song, L.; Florea, L. (2015). Rcorrector. *GigaScience.* Version 1.0.4. Available online: https://github.com/mourisl/Rcorrector/ releases/tag/v1.0.4 (accessed on 6 June 2021).
11. Simpson, J.T.; Wong, K.; Jackman, S.D.; Schein, J.E.; Jones, S.; Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Res. 19*, 1117–1123.
12. Simpson, J.T.; Wong, K.; Jackman, S.D.; Schein, J.E.; Jones, S.J.; Birol, I. (2009). ABySS. *Genome Res.* Version 2.2.4. Available online: https://github.com/bcgsc/abyss/releases/tag/2.2.4 (accessed on 6 June 2021).
13. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*; Lawrence Berkeley National Lab.: Berkeley, CA, USA, 2014.
14. Bushnell, B. BBMap. Version 38.86. Available online: https://sourceforge.net/projects/bbmap/files/ (accessed on 6 June 2021).
15. Liu, J.; Li, G.; Chang, Z.; Yu, T.; Liu, B.; McMullen, R.; Chen, P.; Huang, X. (2016). BinPacker: Packing-Based De novo Transcriptome Assembly from RNA-seq Data. *PLoS Comput. Biol. 12*, e1004772.
16. Liu, J.; Li, G.; Chang, Z.; Yu, T.; Liu, B.; McMullen, R.; Chen, P.; Huang, X. (2016). BinPacker. *PLoS Comput. Biol.* Version 1.0. Available online: https://sourceforge.net/projects/transcriptomeassembly/files/ (accessed on 6 June 2021).

17. Peng, Y.; Leung, H.C.M.; Yiu, S.-M.; Lv, M.-J.; Zhu, X.-G.; Chin, F.Y.L. (2013). IDBA-tran: A more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics. 29*, i326–i334.
18. Peng, Y.; Leung, H.C.; Yiu, S.M.; Lv, M.J.; Zhu, X.G.; Chin, F.Y. (2013). IDBA-tran. *Bioinformatics.* Version 1.1.3. Available online: https://github.com/loneknightpy/idba/releases/tag/1.1.3 (accessed on 6 June 2021).
19. Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics. 31*, 1674–1676.
20. Li, D.; Liu, C.M.; Luo, R.; Sadakane, K.; Lam, T.W. (2015). MEGAHIT. *Bioinformatics* Version 1.2.9. Available online: https: //github.com/voutcn/MEGAHIT/releases/tag/v1.2.9 (accessed on 6 June 2021).
21. Schulz, M.H.; Zerbino, D.R.; Vingron, M.; Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics. 28*, 1086–1092.
22. Schulz, M.H.; Zerbino, D.R.; Vingron, M.; Birney, E. (2012). Oases. *Bioinformatics.* Version 0.2.09. Available online: https: //github.com/dzerbino/oases/releases/tag/0.2.09 (accessed on 6 June 2021).
23. Bushmanova, E.; Antipov, D.; Lapidus, A.; Prjibelski, A.D. (2019). rnaSPAdes. *GigaScience.* Version 3.14.1. Available online: https://github.com/ablab/spades/releases/tag/v3.14.1 (accessed on 6 June 2021).

24. Kannan, S.; Hui, J.; Mazooji, K.; Pachter, L.; Tse, D. (2016). Shannon Cpp. *bioRxiv*. Version 0.4.0. Available online: https: //github.com/bx3/shannon_cpp/releases/tag/v0.4.0 (accessed on 6 June 2021).
25. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.; Pham, S.; Prjibelski, A.D.; et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol. 19*, 455–477.
26. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. (2012). SPAdes. *J. Comput. Biol. 19*, 455–477.
27. Robertson, G.; Schein, J.; Chiu, R.; Corbett, R.; Field, M.; Jackman, S.D.; Mungall, K.; Lee, S.; Okada, H.M.; Qian, J.Q.; et al. (2010). *De novo* assembly and analysis of RNA-seq data. *Nat. Methods. 7*, 909–912.
28. Robertson, G.; Schein, J.; Chiu, R.; Corbett, R.; Field, M.; Jackman, S.D.; Mungall, K.; Lee, S.; Okada, H.M.; Qian, J.Q.; et al. (2010). TransABySS. *Nat. Methods.* Version 2.0.1. Available online: https://github.com/bcgsc/transabyss/releases/tag/2.0.1 (accessed on 6 June 2021).
29. Liu, J.; Yu, T.; Mu, Z.; Li, G. (2019). TransLiG: A de novo transcriptome assembler that uses line graph iteration. *Genome Biol. 20*, 81.
30. Liu, J.; Yu, T.; Mu, Z.; Li, G. (2019). TransLiG. *Genome Biol.* Version 1.3. Available online: https://sourceforge.net/projects/ transcriptomeassembly/files/TransLiG/ (accessed on 6 June 2021).

31. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. (2011). Trinity. *Nat. Biotechnol.* Version 2.11.0. Available online: https://github.com/trinityrnaseq/trinityrnaseq/releases/ tag/v2.11.0 (accessed on 6 June 2021).
32. Smith-Unna, R.; Boursnell, C.; Patro, R.; Hibberd, J.M.; Kelly, S. (2016). TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Res. 26*, 1134–1144.
33. Smith-Unna, R.; Boursnell, C.; Patro, R.; Hibberd, J.M.; Kelly, S. (2016). TransRate. *Genome Res.* Version 1.0.3. Available online: https://github.com/blahah/transrate/releases/tag/v1.0.3 (accessed on 6 June 2021).
34. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics. 28*, 3150–3152.
35. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. (2012). CD-HIT. *Bioinformatics.* Version 4.8.1. Available online: https://github.com/ weizhongli/cdhit/releases/tag/V4.8.1 (accessed on 6 June 2021).
36. Kim, D.; Langmead, B.; Salzberg, S.L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods. 12*, 357–360.
37. Kim, D.; Langmead, B.; Salzberg, S.L. (2015). HISAT2. *Nat. Methods*. Version 2.1.0. Available online: http://daehwankimlab.github. io/hisat2/download/#version-hisat2-210 (accessed on 6 June 2021).

38. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics. 31*, 3210–3212.

39. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. (2015). BUSCO. *Bioinformatics.* Version 4.0.1. Available online: https://gitlab.com/ezlab/busco/-/releases/4.0.1 (accessed on 6 June 2021).

40. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol. 215*, 403–410.

41. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. (1990). NCBI BLAST. *J. Mol. Biol.* Version 2.3.0+. Available online: https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.3.0/ (accessed on 6 June 2021).

42. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. (1990). NCBI BLAST. *J. Mol. Biol.* Version 2.10.0+. Available online: https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.10.0/ (accessed on 6 June 2021).

43. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol. 34*, 525–527.

44. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. (2016). kallisto. *Nat. Biotechnol.* Version 0.46.1. Available online: https: //pachterlab.github.io/kallisto/download (accessed on 6 June 2021).

45. Li, B.; Fillmore, N.; Bai, Y.; Collins, M.; Thomson, J.A.; Stewart, R.; Dewey, C.N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol. 15*, 553.

46. Li, B.; Fillmore, N.; Bai, Y.; Collins, M.; Thomson, J.A.; Stewart, R.; Dewey, C.N. RSEM. (2014). *Genome Biol. Version* 1.3.1. Available online: https://github.com/deweylab/RSEM/releases/tag/v1.3.1 (accessed on 6 June 2021).

47. Leinonen, R.; Sugawara, H.; Shumway, M. (2011). on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.*, *39*, D19–D21.

48. SRA Toolkit Development Team. SRA-Tools. 2014. Version 2.11.0. Available online: https://github.com/ncbi/sra-tools (accessed on 6 June 2021).

49. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. (2009). 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

50. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. (2009). 1000 Genome Project Data Processing Subgroup. Samtools. *Bioinformatics. Version* 1.10. Available online: https://github.com/samtools/samtools/ releases/tag/1.10 (accessed on 6 June 2021).

51. Shen, W.; Le, S.; Li, Y.; Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE.*, *11*, e0163962.

52. Shen, W.; Le, S.; Li, Y.; Hu, F. (2016). SeqKit. *PLoS ONE*. Version 0.16.0. Available online: https://bioinf.shenwei.me/seqkit/ download/ (accessed on 6 June 2021).
53. SRA Database. Available online: https://www.ncbi.nlm.nih.gov/sra (accessed on 17 August 2020).
54. Rogers, R.L.; Zhou, L.; Chu, C.; Márquez, R.; Corl, A.; Linderoth, T.; Freeborn, L.; MacManes, M.D.; Xiong, Z.; Zheng, J.; et al. (2014). Genomic takeover by transposable elements in the Strawberry poison frog. *Mol. Biol. Evol*, *35*, 2913–2927.
55. Andrews, S. (2010). FastQC. Babraham. *Bioinformatics* Version 0.11.9. Available online: https://www.bioinformatics.babraham.ac. uk/projects/download.html#fastqc (accessed on 6 June 2021).
56. Francis, W.R.; Christianson, L.M.; Kiko, R.; Powers, M.L.; Shaner, N.C.; Haddock, S.H.D. (2013). A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genom*, *14*, 167.
57. Illumina. Estimating Sequencing Coverage. Pub. No. 770-2011-022. 2014. Available online: https://www.illumina.com/ documents/products/technotes/technote_coverage_calculation.pdf (accessed on 6 June 2021).
58. Beard, K. H., Vogt, K. A., & Kulmatiski, A. (2002). Top-down effects of a terrestrial frog on forest nutrient dynamics. *Oecologia*, *133*(4), 583-593.
59. Barrantes-Madriga, J., Parallada, M. S., Alvarado, G., & Chaves, V. J. A. (2019). Distribution and invasion progress of Eleutherodactylus coqui (Anura:

Eleutherodactylidae) introduced in Costa Rica. *Phyllomedusa: Journal of Herpetology*, *18*(1), 101-107.

60. AmphibiaWeb. 2022. <https://amphibiaweb.org> University of California, Berkeley, CA, USA. Accessed 10 March 2022.
61. Beard, K. H., Price, E. A., & Pitt, W. C. (2009). Biology and Impacts of Pacific Island Invasive Species. 5. Eleutherodactylus coqui, the Coqui Frog (Anura: Leptodactylidae) 1. *Pacific Science*, *63*(3), 297-316.
62. IUCN. 2021. The IUCN Red List of Threatened Species. Version 2021-3. https://www.iucnredlist.org. Accessed on 10 March 2022.
63. Callery, E. M., Fang, H., & Elinson, R. P. (2001). Frogs without polliwogs: evolution of anuran direct development. *BioEssays*, *23*(3), 233-241.
64. Townsend, D. S., & Stewart, M. M. (1994). Reproductive ecology of the Puerto Rican frog Eleutherodactylus coqui. *Journal of Herpetology*, 34-40.
65. Mesquita, A. F., Lambertini, C., Lyra, M., Malagoli, L. R., James, T. Y., Toledo, L. F., ... & Becker, C. G. (2017). Low resistance to chytridiomycosis in direct-developing amphibians. *Scientific Reports*, *7*(1), 1-7.
66. Ellepola, G., Pie, M. R., Pethiyagoda, R., Hanken, J., & Meegaskumbura, M. (2022). The role of climate and islands in species diversification and reproductive-mode evolution of Old World tree frogs. *Communications biology*, *5*(1), 1-14.
67. Burrowes, P. A., Joglar, R. L., & Green, D. E. (2004). Potential causes for amphibian declines in Puerto Rico. *Herpetologica*, *60*(2), 141-154.
68. Gründler, M. C., Toledo, L. F., Parra-Olea, G., Haddad, C. F., Giasson, L. O., Sawaya, R. J., ... & Zamudio, K. R. (2012). Interaction between breeding habitat and

elevation affects prevalence but not infection intensity of Batrachochytrium dendrobatidis in Brazilian anuran assemblages. *Diseases of aquatic organisms*, *97*(3), 173-184.

69. Beard, K. H., & Pitt, W. C. (2005). Potential consequences of the coqui frog invasion in Hawaii. *Diversity and Distributions*, *11*(5), 427-433.
70. Woolbright, L. L., Hara, A. H., Jacobsen, C. M., Mautz, W. J., & Benevides, F. L. (2006). Population densities of the coqui, Eleutherodactylus coqui (Anura: Leptodactylidae) in newly invaded Hawaii and in native Puerto Rico. *Journal of Herpetology*, *40*(1), 122-126.
71. Woolbright, L. L. (1996). Disturbance influences long-term population patterns in the Puerto Rican frog, Eleutherodactylus coqui (Anura: Leptodactylidae). *Biotropica*, 493-501.
72. Reagan, D. P., & Waide, R. B. (Eds.). (1996). *The food web of a tropical rain forest*. University of Chicago Press. Pg 290.
73. Beard, K. H. (2007). Diet of the invasive frog, Eleutherodactylus coqui, in Hawaii. *Copeia*, *2007*(2), 281-291.
74. Sin, H., Beard, K. H., & Pitt, W. C. (2008). An invasive frog, Eleutherodactylus coqui, increases new leaf production and leaf litter decomposition rates through nutrient cycling in Hawaii. *Biological Invasions*, *10*(3), 335-345.
75. Woolbright, L. L. (1989). Sexual dimorphism in Eleutherodactylus coqui: selection pressures and growth rates. *Herpetologica*, 68-74.

76. Beard, K. H., & Pitt, W. C. (2006). Potential predators of an invasive frog (Eleutherodactylus coqui) in Hawaiian forests. *Journal of Tropical Ecology*, *22*(3), 345-347.

77. Formanowicz Jr, D. R., Stewart, M. M., Townsend, K., Pough, F. H., & Brussard, P. F. (1981). Predation by giant crab spiders on the Puerto Rican frog Eleutherodactylus coqui. *Herpetologica*, 125-129.

78. Rollins-Smith, L. A., Reinert, L. K., & Burrowes, P. A. (2015). Coqui frogs persist with the deadly chytrid fungus despite a lack of defensive antimicrobial peptides. *Diseases of Aquatic Organisms*, *113*(1), 81-83.

79. Marr, S. R., Mautz, W. J., & Hara, A. H. (2008). Parasite loss and introduced species: a comparison of the parasites of the Puerto Rican tree frog,(Eleutherodactylus coqui), in its native and introduced ranges. *Biological invasions*, *10*(8), 1289-1298.

80. Smith, R. L., Beard, K. H., & Koons, D. N. (2018). Invasive coqui frogs are associated with greater abundances of nonnative birds in Hawaii, USA. *The Condor: Ornithological Applications*, *120*(1), 16-29.

81. Velo-Antón, G., Burrowes, P. A., Joglar, R. L., Martínez-Solano, I., Beard, K. H., & Parra-Olea, G. (2007). Phylogenetic study of Eleutherodactylus coqui (Anura: Leptodactylidae) reveals deep genetic fragmentation in Puerto Rico and pinpoints origins of Hawaiian populations. *Molecular Phylogenetics and Evolution*, *45*(2), 716-728.

82. Peacock, M. M., Beard, K. H., O'NEILL, E. M., Kirchoff, V. S., & Peters, M. B. (2009). Strong founder effects and low genetic diversity in introduced populations of Coqui frogs. *Molecular Ecology*, *18*(17), 3603-3615.

83. Laslo, M. (2019). *Evolutionary conservation of endocrine-mediated development in the direct-developing frog, Eleutherodactylus coqui* (Doctoral dissertation, Harvard University).

84. Westrick, S. E., Laslo, M., & Fischer, E. K. (2022). The Natural History of Model Organisms: The big potential of the small frog Eleutherodactylus coqui. *Elife*, *11*, e73401.

85. Mudd, A. B. (2019). *Comparative genomics and chromosome evolution*. University of California, Berkeley.

86. "Trizol Reagent User Guide - Pub. No. MAN0001271 - Rev. A.0." *ThermoFisher*, Invitrogen, 2016, https://tools.thermofisher.com/content/sfs/manuals/trizol_reagent.pdf.

87. Ortiz, R., Gera, P., Rivera, C., & Santos, J. C. (2021). Pincho: a modular approach to high quality DE novo transcriptomics. *Genes*, *12*(7), 953.

88. Buchfink B., Reuter K., Drost H. G. (2021). "Sensitive protein alignments at tree-of-life scale using DIAMOND", *Nature Methods* **18**, 366–368 doi:10.1038/s41592-021-01101-x

89. Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

90. Buchfink, B., Reuter, K., & Drost, H. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND (Version 2.0.11) [Computer software]. https://doi.org/10.1038/s41592-021-01101-x. Available online: https://github.com/bbuchfink/diamond/releases/tag/v2.0.6 (accessed on 6 June 2021).

91. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature methods, 14(4), 417–419. https://doi.org/10.1038/nmeth.4197

92. Patro, R. (2017). Salmon. Version 1.5.2. Available online: https://github.com/COMBINE-lab/salmon/releases/tag/v1.5.2 (accessed on 6 June 2021).

93. Love, M.I., Huber, W. & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 https://doi.org/10.1186/s13059-014-0550-8

94. Kucukural, A., Yukselen, O., Ozata, D.M. *et al.* (2019). DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics* **20,** 6 https://doi.org/10.1186/s12864-018-5362-x

95. Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, *8*(8), 1551-1566.

96. Alford, M. A., Baquir, B., Santana, F. L., Haney, E. F., & Hancock, R. (2020). Cathelicidin Host Defense Peptides and Inflammatory Signaling: Striking a

Balance. *Frontiers in microbiology*, *11*, 1902. https://doi.org/10.3389/fmicb.2020.01902

97. Chen, J., Lin, Y. F., Chen, J. H., Chen, X., & Lin, Z. H. (2021). Molecular characterization of cathelicidin in tiger frog (Hoplobatrachus rugulosus): Antimicrobial activity and immunomodulatory activity. *Comparative biochemistry and physiology. Toxicology & pharmacology : CBP*, *247*, 109072. https://doi.org/10.1016/j.cbpc.2021.109072
98. Wang, Y., Ouyang, J., Luo, X., Zhang, M., Jiang, Y., Zhang, F., Zhou, J., & Wang, Y. (2021). Identification and characterization of novel bi-functional cathelicidins from the black-spotted frog (Pelophylax nigromaculata) with both anti-infective and antioxidant activities. *Developmental and comparative immunology*, *116*, 103928. https://doi.org/10.1016/j.dci.2020.103928
99. Gao, F., Xu, W. F., Tang, L. P., Wang, M. M., Wang, X. J., & Qian, Y. C. (2016). Characteristics of cathelicidin-Bg, a novel gene expressed in the ear-side gland of Bufo gargarizans. *Genetics and molecular research : GMR*, *15*(3), 10.4238/gmr.15038481. https://doi.org/10.4238/gmr.15038481
100. Waterhouse A. M., Procter J. B., Martin D. M. A., Clamp M., Barton G. J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191. doi:10.1093/bioinformatics/btp033
101. Meng, G., Li, Y., Yang, C., & Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic acids research*, *47*(11), e63-e63.

102. Meng, G., Li, Y., Yang, C., & Liu, S. (2019). MitoZ. Version 3.4. Available online: https://github.com/linzhi2013/MitoZ/releases/tag/3.4 (accessed on 6 June 2021).

103. Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, *35*(11), 1026-1028.

104. Davis, M. A. (2003). Biotic globalization: does competition from introduced species threaten biodiversity?. *Bioscience*, *53*(5), 481-489.

105. Pimm, S. L., Russell, G. J., Gittleman, J. L., & Brooks, T. M. (1995). The future of biodiversity. *Science*, *269*(5222), 347-350.

106. Didham, R. K., Tylianakis, J. M., Hutchison, M. A., Ewers, R. M., & Gemmell, N. J. (2005). Are invasive species the drivers of ecological change?. *Trends in ecology & evolution*, *20*(9), 470-474.

107. Somma, L. A. (2021). Geographic Distribution: Eleutherodactylus coqui (Common Coqui). USA: Florida.

108. Roedder, D. (2009). 'Sleepless in Hawaii'-does anthropogenic climate change enhance ecological and socioeconomic impacts of the alien invasive Eleutherodactylus coqui Thomas 1966 (Anura: Eleutherodactylidae)?. *North-Western Journal of Zoology*, *5*(1).

109. Kaiser, B. A., & Burnett, K. M. (2006). *Economic impacts of E. coqui frogs in Hawaii* (No. 379-2016-21643).

110. *Coqui Information.* (2022). State of Hawaii, Plant Industry Division. https://hdoa.hawaii.gov/pi/ppc/cm/coqui-information/

111. Kraus, F., Campbell, E. W., Allison, A., & Pratt, T. (1999). Eleutherodactylus frog introduction to Hawaii. *Herpetological Review*, *30*(1), 21-25.

112. *Eleutherodactylus coqui.* (2022) iNaturalist. https://www.inaturalist.org/taxa/22454-Eleutherodactylus-coqui

113. Tuttle, N. C., Beard, K. H., & Al-Chokhachy, R. (2008). Aerially applied citric acid reduces the density of an invasive frog in Hawaii, USA. *Wildlife research*, *35*(7), 676-683.

114. Sin, H., & Radford, A. (2007). Coqui frog research and management efforts in Hawaii.

115. Beachy, J. R., Neville, R., Arnott, C., Veitch, C. R., Clout, M. N., & Towns, D. R. (2011). Successful control of an incipient invasive amphibian: Eleutherodactylus coqui on O'ahu, Hawai'i. *Island invasives: eradication and management*, 140-147.

116. Pitt, W. C., & Doratt, R. E. (2008). Dermal toxicity of sodium bicarbonate to control Coqui frogs, Eleutherodactylus coqui, in Hawaii.

117. *Management of an invasive species in Hawaii: the coqui frog.* (2007). United States Department of Agriculture, Research, Education & Economics Information System. https://reeis.usda.gov/web/crisprojectpages/0204184-management-of-an-invasive-species-in-hawaii-the-coqui-frog.html

118. Hara, A. H., Jacobsen, C. M., Marr, S. R., & Niino-DuPonte, R. Y. (2010). Hot water as a potential disinfestation treatment for an invasive anuran amphibian, the coqui frog, Eleutherodactylus coqui Thomas (Leptodactylidae), on potted plants. *International Journal of Pest Management*, *56*(3), 255-263.

119. Owais, M., Ansari, M. A., Ahmad, I., Zia, Q., Pierard, G., & Chauhan, A. (2010). Innate immunity in pathogenesis and treatment of dermatomycosis. In *Combating Fungal Infections* (pp. 347-371). Springer, Berlin, Heidelberg.

120. Nardo, A.D., Vitiello, A., Gallo, R.L. (2003) Cutting Edge: Mast Cell Antimicrobial Activity is Mediated by Expression of Cathelicidin Antimicrobial Peptide. *The Journal of Immunology,* 170:2274-2278.

121. Tossi, A., D'Este, F., Skerlavaj, B., & Gennaro, R. (2017). Structural and functional diversity of cathelicidins. *Wang, G., Ed.; CABI: Wallingford, UK*, 20-48.

122. Zaiou, M., & Gallo, R. L. (2002). Cathelicidins, essential gene-encoded mammalian antibiotics. *Journal of molecular medicine*, *80*(9), 549-561.

123. Zanetti, M. (2005). The role of cathelicidins in the innate host defenses of mammals. *Current issues in molecular biology*, *7*(2), 179-196.

124. Xhindoli, D., Pacor, S., Benincasa, M., Scocchi, M., Gennaro, R., & Tossi, A. (2016). The human cathelicidin LL-37—A pore-forming antibacterial peptide and host-cell modulator. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1858*(3), 546-566.

125. König, E., Bininda-Emonds, O. R., & Shaw, C. (2015). The diversity and evolution of anuran skin peptides. *Peptides*, *63*, 96-117.

126. Rosenblum, E. B., Poorten, T. J., Settles, M., & Murdoch, G. K. (2012). Only skin deep: shared genetic response to the deadly chytrid fungus in susceptible frog species. *Molecular ecology*, *21*(13), 3110–3120. https://doi.org/10.1111/j.1365-294X.2012.05481.x

127. Mu, L., Zhou, L., Yang, J., Zhuang, L., Tang, J., Liu, T., ... & Yang, H. (2017). The first identified cathelicidin from tree frogs possesses anti-inflammatory and partial LPS neutralization activities. *Amino Acids*, *49*(9), 1571-1585.

128. Bottazzi, B., Vouret-Craviari, V., Bastone, A., De Gioia, L., Matteucci, C., Peri, G., ... & Mantovani, A. (1997). Multimer formation and ligand recognition by the long pentraxin PTX3: similarities and differences with the short pentraxins C-reactive protein and serum amyloid P component. *Journal of Biological Chemistry*, *272*(52), 32817-32823.

129. Garlanda, C., Hirsch, E., Bozza, S., Salustri, A., De Acetis, M., Nota, R., ... & Mantovani, A. (2002). Non-redundant role of the long pentraxin PTX3 in anti-fungal innate immune response. *Nature*, *420*(6912), 182-186.

130. Younus, H. (2019). Oxidoreductases: overview and practical applications. *Biocatalysis*, 39-55.

131. Wang, T., Hedrick, M. S., Ihmied, Y. M., & Taylor, E. W. (1999). Control and interaction of the cardiovascular and respiratory systems in anuran amphibians. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *124*(4), 393-406.

132. Wang, T. O. B. I. A. S., & Hicks, J. W. (1996). The interaction of pulmonary ventilation and the right-left shunt on arterial oxygen levels. *The Journal of experimental biology*, *199*(10), 2121-2129.

133. Weber, R. E., & Jensen, F. B. (1988). Functional adaptations in hemoglobins from ectothermic vertebrates. *Annual Review of Physiology*, *50*(1), 161-179.

134. Yonetani, T., & Tsuneshige, A. (2003). The global allostery model of hemoglobin: an allosteric mechanism involving homotropic and heterotropic interactions. *Comptes rendus biologies*, *326*(6), 523-532.

135. Burmester, T., Ebner, B., Weich, B., & Hankeln, T. (2002). Cytoglobin: a novel globin type ubiquitously expressed invertebrate tissues. *Molecular biology and evolution*, *19*(4), 416-421.

136. Nishi, H., Inagi, R., Kawada, N., Yoshizato, K., Mimura, I., Fujita, T., & Nangaku, M. (2011). Cytoglobin, a novel member of the globin family, protects kidney fibroblasts against oxidative stress under ischemic conditions. *The American journal of pathology*, *178*(1), 128-139.

137. Fordel, E., Geuens, E., Dewilde, S., Rottiers, P., Carmeliet, P., Grooten, J., & Moens, L. (2004). Cytoglobin expression is upregulated in all tissues upon hypoxia: an in vitro and in vivo study by quantitative real-time PCR. *Biochemical and biophysical research communications*, *319*(2), 342-348.

138. Pesce, A., Bolognesi, M., Bocedi, A., Ascenzi, P., Dewilde, S., Moens, L., ... & Burmester, T. (2002). Neuroglobin and cytoglobin. *EMBO reports*, *3*(12), 1146-1151.

139. Dengler, V. L., Galbraith, M. D., & Espinosa, J. M. (2014). Transcriptional regulation by hypoxia inducible factors. *Critical reviews in biochemistry and molecular biology*, *49*(1), 1-15.

140. Forristal, C. E., Wright, K. L., Hanley, N. A., Oreffo, R. O., & Houghton, F. D. (2010). Hypoxia inducible factors regulate pluripotency and proliferation in human

embryonic stem cells cultured at reduced oxygen tensions. *Reproduction (Cambridge, England)*, *139*(1), 85.

141. Ziello, J. E., Jovin, I. S., & Huang, Y. (2007). Hypoxia-Inducible Factor (HIF)-1 regulatory pathway and its potential for therapeutic intervention in malignancy and ischemia. *The Yale journal of biology and medicine*, *80*(2), 51.
142. Semenza, G. L. (2007). Hypoxia-inducible factor 1 (HIF-1) pathway. *Science's STKE*, *2007*(407), cm8-cm8.
143. Smith, C. J., Santhanam, L., Bruning, R. S., Stanhewicz, A., Berkowitz, D. E., & Holowatz, L. A. (2011). Upregulation of inducible nitric oxide synthase contributes to attenuated cutaneous vasodilation in essential hypertensive humans. Hypertension (Dallas, Tex. : 1979), 58(5), 935–942. https://doi.org/10.1161/HYPERTENSIONAHA.111.178129
144. Adams, C. M., Clark-Garvey, S., Porcu, P., & Eischen, C. M. (2019). Targeting the Bcl-2 family in B cell lymphoma. *Frontiers in oncology*, *8*, 636.
145. Tzifi, F., Economopoulou, C., Gourgiotis, D., Ardavanis, A., Papageorgiou, S., & Scorilas, A. (2012). The role of BCL2 family of apoptosis regulator proteins in acute and chronic leukemias. *Advances in hematology*, *2012*.
146. Bickler, P. E., & Buck, L. T. (2007). Hypoxia tolerance in reptiles, amphibians, and fishes: life with variable oxygen availability. *Annual review of physiology*, *69*(1), 145-170.
147. Bickler, P. E., & Buck, L. T. (2007). Hypoxia tolerance in reptiles, amphibians, and fishes: life with variable oxygen availability. *Annual review of physiology*, *69*(1), 145-170.

148. National Center for Biotechnology Information (NCBI)[Internet]. (1998). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Available online: https://www.ncbi.nlm.nih.gov/ (accessed on 16 August 2022).

149. Sievers, F., & Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. In *Multiple sequence alignment methods* (pp. 105-116). Humana Press, Totowa, NJ.

150. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189-1191.

151. Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular biology and evolution*, *34*(7), 1812-1819.

152. Kosakovsky Pond, S. L., Poon, A. F., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., ... & Muse, S. V. (2020). HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. *Molecular biology and evolution*, *37*(1), 295-299.

153. Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS genetics*, *8*(7), e1002764.

154. Kosakovsky Pond, S. L., & Frost, S. D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, *22*(5), 1208-1222.

155. Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution*, *30*(5), 1196-1205.

156. Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., ... & Kosakovsky Pond, S. L. (2015). Gene-wide identification of episodic selection. *Molecular biology and evolution*, *32*(5), 1365-1371**.**

157. Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular biology and evolution*, *32*(5), 1342-1353.

158. Hardison, R. C., Chui, D. H., Giardine, B., Riemer, C., Patrinos, G. P., Anagnou, N., ... & Wajcman, H. (2002). HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. *Human mutation*, *19*(3), 225-233.

159. Dalby, A., Dauter, Z., & Littlechild, J. A. (1999). Crystal structure of human muscle aldolase complexed with fructose 1, 6-bisphosphate: mechanistic implications. *Protein Science*, *8*(2), 291-297.

160. Smith, H. M. (1925). Cell size and metabolic activity in Amphibia. *The Biological Bulletin*, 48(5), 347-378.

161. Shekhovtsov, S. V., Bulakhova, N. A., Tsentalovich, Y. P., Zelentsova, E. A., Yanshole, L. V., Meshcheryakova, E. N., & Berman, D. I. (2020). Metabolic response of the Siberian wood frog Rana amurensis to extreme hypoxia. *Scientific reports*, *10*(1), 1-11.

162. Hutchison, V. H., Haines, H. B., & Engbretson, G. (1976). Aquatic life at high altitude: respiratory adaptations in the Lake Titicaca frog, Telmatobius culeus. Respiration physiology, 27(1), 115–129. https://doi.org/10.1016/0034-5687(76)90022-0

163. Rossi, G. S., Cramp, R. L., Wright, P. A., & Franklin, C. E. (2020). Frogs seek hypoxic microhabitats that accentuate metabolic depression during dormancy. *Journal of Experimental Biology*, *223*(2), jeb218743.

164. Shekhovtsov, S. V., Bulakhova, N. A., Tsentalovich, Y. P., Zelentsova, E. A., Yanshole, L. V., Meshcheryakova, E. N., & Berman, D. I. (2020). Metabolic response of the Siberian wood frog Rana amurensis to extreme hypoxia. *Scientific reports*, *10*(1), 1-11.

Vita

Name	*Randy Ortiz*
Degrees	*Bachelor of Arts, Vassar College, Poughkeepsie NY, Major: Biology*
Date Graduated	*May 2014*
Other Degrees	*Master of Science, St. John's University, Queens NY, Major: Biology*
Date Graduated	*May 2018*