DENIS JACOB MACHADO

Comparative Genomics and the Evolution of Amphibian Chemical Defense

Genômica Comparada e a Evolução da Defesa Química em Anfíbios

São Paulo

2018

Denis Jacob Machado

Comparative Genomics and the Evolution of Amphibian Chemical Defense

Genômica Comparada e a Evolução da Defesa Química em Anfíbios

Versão corrigida

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de Concentração: Bioinformática

Orientador: Prof. Taran Grant, Ph.D.

São Paulo

2018

FICHA CATALOGRÁFICA

Machado, Denis Jacob

Genômica Comparada e a Evolução da Defesa Química em Anfíbios / Denis Jacob Machado, [orient.] Taran Grant. São Paulo: 2018.

188 p.

Orientador: Grant, Taran

Tese (Doutorado) – Universidade de São Paulo

Orientador: Prof. Dr. Taran Grant

Programa Interunidades de Pós-Graduação em Bioinformática

Área de concentração: Bioinformática

1. Bioinformática 2. Genômica Comparada 3. Defesa Química 4. Rãs-de-veneno I. Grant, Taran, orientador. II. Universidade de São Paulo III. Genômica Comparada e a Evolução da Defesa Química em Anfíbios.

Elaborada pelo Serviço de Informação e Biblioteca "Prof. Carlos Benjamin de Lyra", IME-USP

Às minorias na ciência, por sua contribuição que não pode evitar senão brilhar além de esfera acadêmica e iluminar todo o universo social.

Aos benfeitores anônimos, por terem mais medo de não ajudar a quem possa merecer do que de serem traídos.

A todos que compartilharam comigo um pedaço de suas vidas, por serem uma família.

Ao Laboratório de Anfíbios, por ser uma casa.

À Regiane Lima, por ser um futuro.

DEDICO

AGRADECIMENTOS

Ao Prof. Dr. Taran Grant (Universidade de São Paulo, Instituto de Biociências) pela orientação e oportunidade de participar de seu grupo de pesquisa.

Ao Prof. Dr. Daniel Janies (University of North Carolina at Charlotte, Department of Bioinformatics and Genomics), pela parceria e colaboração que ajudaram a diminuir a distância entre a pesquisa de base em organismos não-modelo e a pesquisa aplicada em biologia computacional e bioinformática.

Ao Programa Interunidades de Pós-Graduação em Bioinformática e sua secretaria, pela atenção e suporte.

Ao Museu de Zoologia da Universidade de São Paulo (MZUSP), cuja infraestrutura permitiu a instalação do cluster de computadores ACE (FAPESP Proc. No. 2012/10000-5), usado largamente para análise de dados apresentados nesta dissertação.

Ao Department of Bioinformatics and Genomics da University of North Carolina at Charlotte (UNC Charlotte) e ao David H. Murdock Research Institute (DHM-RI) por sua colaboração em sequenciamento e análise de dados genômicos.

Ao Prof. Dr. Edmond Brodie Jr. (Utah State University, Department of Biology), pela parceria, colaboração, e exemplo de vida.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo suporte financeiro (Proc. Nos. 2013/05858-8 e 2015/18654-2).

“People may say I can't sing, but no one can ever say I didn't sing.”

Florence Foster Jenkins

ABSTRACT

Chapter 1 of this dissertation brings a discussion about the current state of basic and applied research on the evolution and biology of amphibians in general, and on amphibian chemical defense in particular. This introductory chapter also advocates comparative genomics as a strategy to increase our understanding of amphibian chemical defense and evolution, which is made possible thanks to high-throughput sequencing technologies but remains challenging mainly due to specific challenges of working with non-model organisms, including the lack of diverse and well-curated genomic databases. The remaining chapters are the first stepping-stones in the direction of a more extensive line of investigation that ultimately leads towards a multidisciplinary work that aims to enhance the transfer of knowledge between basic and applied research. Chapter 2 addresses the challenges in de novo assembly of mitogenomes of frogs when no reference sequence is available, and resources are limited. Chapter 3 introduces a new ad hoc mapping strategy to test the circularization of novel mitogenomes, using alignment scores and a new per-position sequence coverage value (which we named "connectivity") to assess the quality of the inferred circularization. Chapter 4 address the assembly of the draft nuclear genomes of *Scaphiopus holbrookii* and *Phyllobates terribilis*, with focus on the homology-based and ab initio annotation of protein-coding genes and the proposal of new phylogenetic markers. Chapter 5 presents the de novo assembly of repetitive DNA in the nuclear genome of frogs, with new insights on the role of repeats in the variation of genome size in amphibians and new phylogenetic markers based on these genomic elements. Chapter 6 is a reply to Tarvin *et al.* (2016), including a review of the current knowledge about alkaloid defense in poison frogs. In combination, these chapters add to our infant but growing knowledge of amphibian genomics and chemical defense and strengthen the communication between basic research in non-model organisms and cutting-edge bioinformatic methods.

RESUMO

No capítulo 1 desta tese, eu reviso a literatura especializada sobre biologia de anfíbios, especialmente os estudos associados à defesa química. Eu defendo a genômica comparada como uma estratégia promissora para novas pesquisas nessa área. Esta estratégia se torna possível graças ao advento de tecnologia de sequenciamento de DNA de alta eficiência, mas enfrenta muitos desafios (*e.g.*, falta de referências, acessibilidade a grandes centros de sequenciamento, disponibilidade de computadores e pessoal especializado) para se sedimentar e ser aproveitada por pesquisadores interessados em pesquisa de base com organismos não-modelo. Deste modo, os demais capítulos são os primeiros passos em direção a uma linha de pesquisa mais abrangente que tem como objetivo final aproximar a pesquisa de base em biologia à pesquisa aplicada em bioinformática, usando anfíbios como modelo. O capítulo 2 apresenta uma estratégia de montagem de novos genomas mitocondriais na ausência de referências e usando uma quantidade mínima de recursos. O capítulo 3 introduz uma nova ferramenta para testar a circularização destes genomas. O capítulo 4 descreve o sequenciamento, montagem, e anotação dos genomas parciais de *Scaphiopus holbrookii* e *Phyllobates terribilis*, com ênfase em novos marcadores para filogenética. O capítulo 5 trata da montagem de novo de repetições no DNA nos genomas de *S. holbrookii* e *P. terribilis*, assim como nos genomas de *Melanophryniscus moreirae* e *Hyloxalus subpunctatus*, trazendo uma discussão sobre o papel dessas repetições sobre aumento do tamanho genômico em anfíbios e seu possível uso como marcadores "Hennigianos". Finalmente, usamos dados da anotação preliminar dos genomas de *S. holbrookii* e *P. terribilis* para revisar o atual estado de conhecimento sobre resistência a alcalóides em rãs-de-veneno em resposta ao artigo de Tarvin *et al.* (2016), o primeiro estudo a tentar demonstrar as bases genéticas da defesa química em rãs-de-veneno. Em conjunto, espero que estes capítulos sirvam somem tanto ao nosso conhecimento em genômica e defesa química de anfíbios quanto para o esforço de aproximar a pesquisa de base com organismos não-modelo das ferramentas mais avanças disponíveis em bioinformática.

INDEX

1. INTRODUCTION

The gap between basic and applied research

Garnett & Christidis (2017) argued that taxonomists and phylogeneticists are arbitrarily creating and modifying the formal classification of organisms with no regard to the consequences to other sciences and communities that rely on the assumption that species are fixed entities. According to the authors, taxonomists and phylogeneticists should not have the final say on species classification if such delineations are not entirely free from arbitrary decisions. Instead, they advocate that deliberations must draw on expertise beyond taxonomy, phylogenetics, morphology, systematics, and genetics. Lawyers, anthropologists, and sociologists (and possibly others) should also participate in taxonomy decisions given the possible legal and social repercussions of taxonomic decisions.

It is indisputable that efforts to minimize discrepancies between taxonomy and applied conservation efforts would be beneficial to all. However, taxonomy is not a service provider for conservation biologists or policymakers. It is instead an independent biological discipline and, as any scientific discipline, hypotheses are its cornerstone (Lambertz 2017).

The comment published by (Garnett & Christidis 2017) revitalized the debate about the role of taxonomy and its importance to conservation efforts, as well as other applications. However, the fact that the authors focused more in questioning the hypotheses erected or falsified by taxonomists instead of focusing on how scientists test these hypotheses serve to show a broader and more profound problem: the gap between basic and applied research. For even when we recognize the relevance of the results of some studies in basic research (such as taxonomic classifications), this is often not meet with a proportional amount of respect (regarding recognition and accessibility to funds) or with the acceptance of these disciplines as independent and fundamental fields of science.

To continue illustrating the importance of conceptualizing this dissertation within the context of the gap between basic and applied research, let's now look at the example of the Nobel Prize Award in 2008. That year, the Nobel Prize in Chemistry was awarded jointly to Osamu Shimomura, Martin Chalfie, and Roger Y. Tsien for the discovery and development of the green fluorescent protein (GFP). By mutating GFP in various ways, scientists can create

differently colored fluorescent proteins that make the protein gene products fluoresce at different wavelengths. The GFPs made possible, for the first time, for us to study the location of multiple proteins in the cell at the same time, and therefore became rapidly popular. On the other hand, the history behind their discovery of GFPs is not as well-known.

The discovery of GFPS came out of the curiosity of Shimomura, who was interested in understanding why certain jellyfish were a striking green fluorescent color. It was only later that Douglas Prasher cloned GFP for the first time and suggested that its usage as a fluorescent tag. Since Prasher was unable to raise funds to develop GFP in this way, Chalfie and Tsien took the task. Given this history and context, Tsien used his Nobel Prize acceptance speech to draw attention to how "funding was difficult to obtain for basic research on obscure organisms like the jellyfish that was the source of GFP." Tsien added that he hoped the award would reinforce "recognition of the importance of basic science as the foundation for practical benefits to our health and economies." During the ceremony,
Tsien became aware that Prasher (who kept finding problems financing his basic research) left science and was working as a shuttle driver. Tsien decided to pay for Prasher to attend the Nobel celebrations in Stockholm and offered him a job as a senior scientist in his lab (see Parrington 2015 for additional details).

Unfortunately, Garnett & Christidis (2017) is only one example that indicates that the gap has not closed since 2008. However, there are frequent successful attempts to enhance the transfer of knowledge between basic and applied research that demonstrate the importance of such endeavors. For instance, Schneider *et al.* (2017) took advantage of well-established theories about the hypothesis of homology and inference of evolutionary relationships that are popular and well understood amongst phylogeneticists interested in non-model organisms and applied them to study the evolution of Flaviviridae. Flaviviridae is a family of ssRNA positive-strand viruses which include Dengue, Zika, Chikungunya, Yellow Fever, among other flaviviruses. By avoiding spurious alignments on the flavivirus polyprotein genome and using outgroups to test the monophyly of the ingroup and the orientation of nucleotide transformations on the tree, Schneider *et al.* (2017) proposed a new template for the evolution of flaviviruses which has an immediate impact on how we understand the epidemiology of these viruses.

While the examples above show how a broad range of applied disciplines, from species conservation to human health, can benefit from basic research, the latter would also immensely benefit from cutting-edge technology which is often more readily available to

other fields of science. Within this context, this dissertation is ultimately an effort to approach state-of-the-art tools in computational biology and bioinformatics to the study of the evolution of amphibian chemical defense, as well as amphibian taxonomy and phylogenetics.

Amphibians as a research model

Amphibians (Gymnophiona, Caudata, and Anura; see **Figure 1**) are a conspicuous and ecologically important component of the world's vertebrate fauna, with over 7,800 species worldwide (Frost, 2017; accessed on March 27, 2018) and at least 1,080 species in Brazil according to Segalla *et al.* (2016), a number that continues to increase every year (*e.g.*, Orrico *et al.* 2014, Ferreira *et al.* 2015, Pinheiro *et al.* 2016, Dias *et al.* 2017). Unfortunately, the rapid growth in knowledge of amphibian diversity (taxonomists described 395 new species of amphibians from 2014 until the end of 2017) is coincident with a massive and global decline in amphibians populations (Alford and Richards 1999, Houlahan *et al.* 2000, Young *et al.* 2001, Stuart 2004). The extinction rate of amphibians is currently estimated to be four orders-of-magnitude higher than the background extinction rate (Alroy 2015), and at least 6.9% of all frog species may be lost within the next century, even if there is no acceleration in the growth of environmental threats. This may be due to a variety of reasons, including habitat loss and fragmentation (Green & Muths 2005, Collins & Halliday 2005), global environmental changes (Donnelly & Crump 1998, Blaustein & Kiesecker 2002, Heyer 2003, Licht 2003), and emerging pathogens (Collins & Storfer 2003, Rosenblum *et al.* 2010).

Despite the specific causes, a general message from amphibians is that we may have little time to stave off a possible worldwide extinction event and do basic research will allow us to think of conservation strategies and learn as much from them while we can (Wake & Vredenburg 2008). Also, even while species of amphibians are disappearing, the current discovery rate of about 100 new species per year indicates that we are still very much in the discovery phase of documenting amphibian diversity. As such, it is urgent that the scientific community redouble its efforts to gather information on amphibians biology to understand the full extent of their diversity and develop strategies to stem the decline and extinction of these species.

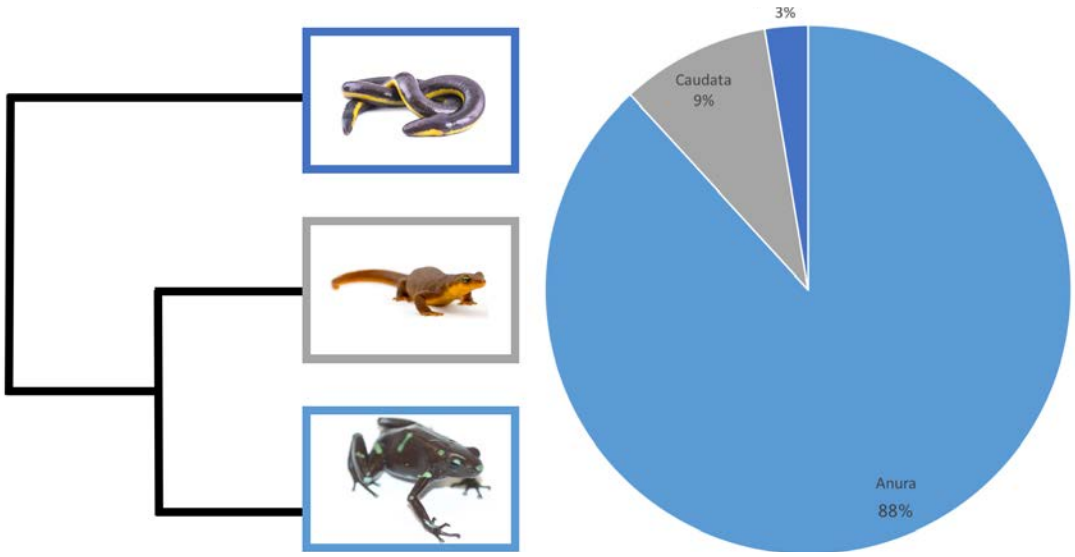


Figure 1 – The three orders of Amphibia with the number of species in each (according to Frost, 2017; on March 27, 2018). Photos taken from www.amphibiaweb.org.

At this time, many valid research routes could be taken to enhance our understanding of amphibian diversification. Novel strategies to study amphibian systematics, for example, are vital as they provide the foundation for all conservation efforts (Kim & Byrne 2006; also remember Garnett & Christidis 2017). Another valid approach is to focus on amphibian defensive compounds given their importance in diverse aspects of amphibian biology, as well as possible applications to human health.

Chemical defense in amphibians

Different types of chemical defense are found in a variety of animals in the form of complex adaptations to avoid predation and parasites, also acting against pathogens (e.g., fungi, viruses, bacterias). Amphibians, for example, are protected by an exocrine defense system composed of cutaneous poison glands (Toledo & Jared 1995). These glands are specialized cells that secrete a variety of defensive chemicals, defined as substances that are produced to reduce the risk of bodily harm by another organism (Berenbaum 1995). In fact, chemical defense in amphibians involves such a diverse group of substances that Roseghini *et al.* (1976: p. 31) stated that "The amphibian skin may be regarded as an enormous storehouse of biogenic amines and active polypeptides. Indeed, no other vertebrate or invertebrate tissue can

compare with amphibian cutaneous tissue in regard to variety and concentration of these active compounds."

The secretions of the amphibian skin are believed to function as a critical component of the innate immune system in defending against pathogens and parasites (Rivas *et al.* 2009a, Conlon 2011) and are also involved in complex anti-predator mechanisms (Brodie *et al.* 1991) that science remains to unveil fully. For example, the antimicrobial peptides (AMPs) are cationic, amphipathic and α-helical peptides that represent the second major group and the largest class of frog skin peptides. Until recently, scientists believed that AMPs had an exclusive antimicrobial activity. However, five years ago (König *et al.* 2012) proposed a new name for these molecules, cytolysins, together with a new hypothesis for their functions. By observing that cytolysins are distributed sporadically (*i.e.*, non-universally) across Anura and that these molecules typically co-occur with neurotoxic peptides, König and collaborators started proposing the hypothesis that cytolysins have an important role delivering neuroactive peptides to the endocrine and nervous system of the predator.

Other substances with a more clear role in the defense against predators are amphibian are steroids (better know as bufadienolides), which occurs exclusively in the anuran family Bufonidae (although Daly *et al.* 1987 reported trace levels of bufadienolide-like compounds in certain frogs of the family Dendrobatidae). They are cardiotoxic substances that the prominent paratoid macroglands biosynthesizes and stores. Recently, the combination of modern methods and increase sampling effort allowed researchers to demonstrate the function of some bufadienolides as activators of CIC-3 chloride (Cl^-) channels with antitumor activities (Liu *et al.* 2013). Also, (Bókony *et al.* 2017) suggested that bufadienolide production may serve to mitigate risks posed by competitors, including aggression, cannibalism or disease. Therefore, bufadienolides are currently considered intriguing candidates for multi-purpose defenses that may provide protection not only against predators but also against competitors.

But not all defensive chemicals are synthesized by the amphibian organism. For example, approximately 150 species of brightly colored, primarily diurnal anurans are capable of sequestering and secreting defensive lipophilic alkaloids. The lipophilic alkaloids compose a group of neurotoxins that occur in only lineages within Amphibia (Daly *et al.* 2005, Grant *et al.* 2006, Flórez-Rodríguez *et al.* 2010). Given their lipophilic characteristic, these alkaloids cross cell membranes and may permeate the blood-brain barrier. Outside the frog's body, these alkaloids act as neurotoxins, frequently targeting voltage-gated sodium (Na^+) channels

and enabling them to open persistently (*e.g.*, the neurotoxins batrachotoxin – BTX – and pumilotoxin – PTX).

The large volume of applied research focusing on lipophilic alkaloids highlights its importance to science and development. Many investigators used lipophilic alkaloids to learn how Na+ channels work, the differences between types of channels, and of the identity of some chemical constituents of the channels (Strichartz *et al.* 1987). There is also a large volume of pharmacological research on poison frog alkaloids, and some of these compounds have extremely promising applications in the development of new anesthetics (*e.g.*, epibatidine) or understanding neuromuscular functions (*e.g.*, pumiliotoxins, batrachotoxins, and izidines) (Daly *et al.* 2005a). The work of experts in anuran defensive chemicals has resulted in a fertile field of artificial synthesis of dendrobatid alkaloids and design of therapeutic agents based on their structures (Savitzky & Saporito 2012; Toyooka *et al.* 2002). Nevertheless, detailed *in vitro* assays of their function are scarce but for a few compounds (BTX, PTX 251D, epibatidine) and progress in studying these chemicals, in general, has slowed both by stricter legal regulations in response to concerns regarding bioprospecting for commercial drugs (Angerer 2011) and the lack of information on the basic biology and evolution of the animals that secrete them.

Anuran species that sequester and secret lipophilic alkaloids form a polyphyletic assemblage referred to as "poison frogs." The poison frogs are formed by approximately 150 species in eight lineages of five anuran families, including Bufonidae (*Melanophryniscus*), Dendrobatidae (independently derived in *Epipedobates*, *Ameerega*, and *Dendrobatinae*), Eleutherodactylidae (*Eleutherodactylus*), Mantellidae (*Mantella*), and Myobatrachidae (*Pseudophryne*) (see **Figure 2**).

	Antifungal	**Antibiotic**	**High toxicity**	**Low or moderate toxicity**	**Total**
Compounds	67	118	71	139	395
Compounds unique to amphibians	37	108	42	82	269

Table 1 – Summary of information from Santos *et al.* (2016: Table 21.1), showing the number of compounds with antifungal, antibiotic, and/or toxic activity in amphibians (unknown states were not computed).

Saporito et al. (2012: p. 160) stated that "more than 850 lipophilic alkaloids, organized into more than 20 structural classes, have been detected in the skin of poison frogs, a number that apparently reflects the large diversity of alkaloids present in arthropods." Santos et al. (2016: Table 21.1) say that the known defensive chemicals in poison frogs of the family

Dendrobatidae sum up to 525 compounds, including 337 unique compounds. These substances include 24 different structural classes and 395 (269 unique) chemicals with antifungal, antibiotic, or toxic activity (see **Table 1**).

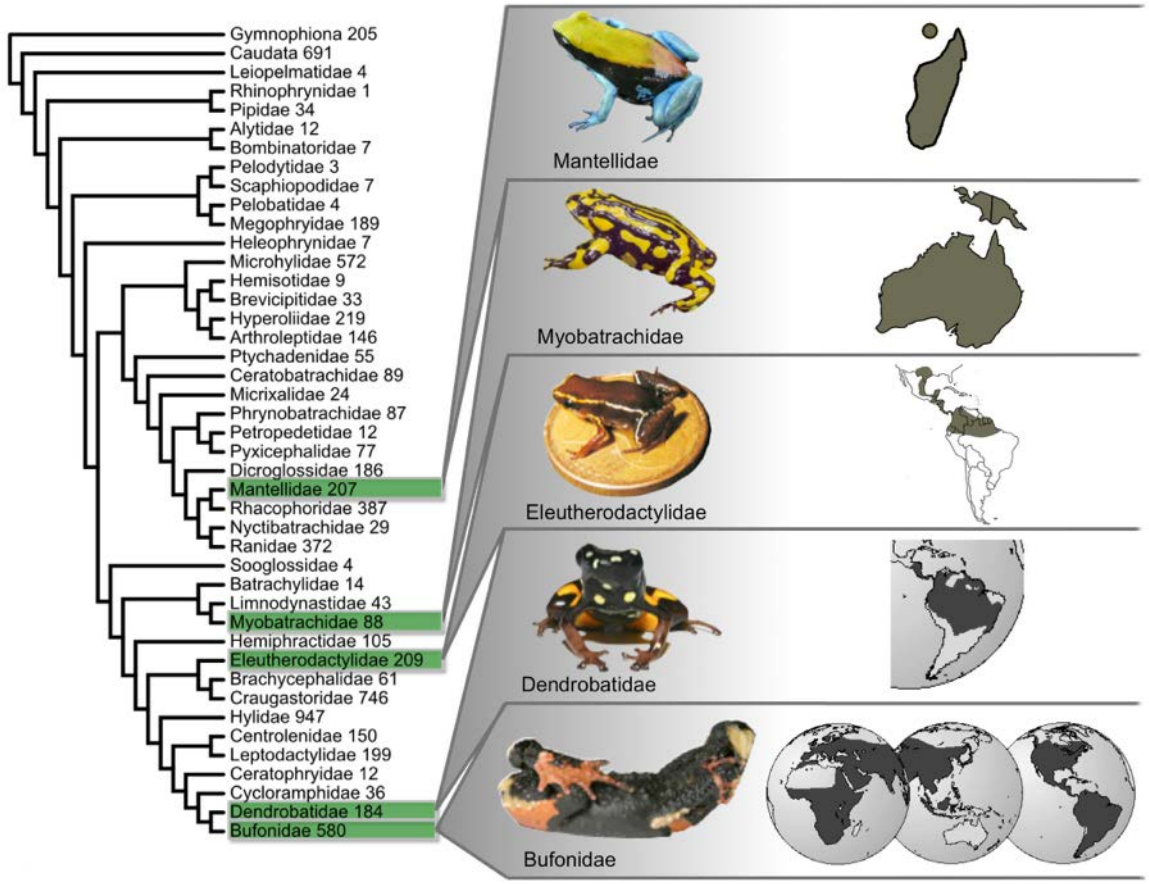


Figure 2 – The poison frogs (modified from Frost *et al.* 2006, Padial *et al.* 2014, and Frost 2015).

The diversity (*i.e.*, richness, composition, and abundance) of lipophilic alkaloids in poison frogs is extremely variable among individuals, populations, and species (*e.g.*, Saporito *et al.* 2011, Grant *et al.* 2012), but the causes of this variation are not clear. Recent studies are piecing together the ecological puzzle of alkaloid variation in alkaloid-sequestering frogs (*e.g.*, Grant *et al.* 2012). These efforts are facing difficulties because, although a vast literature exists on amphibian defensive chemicals, most research has focused on chemistry and the search for natural products, with much little being known about the evolution of amphibian chemical defense.

There are two key aspects in the study of defense mechanisms using lipophilic alkaloids that deserve most attention: toxin sequestration and physiological resistance. Poison

frogs typically sequester lipophilic alkaloids from dietary sources (Daly *et al.* 1994, Clark *et al.* 2012, Hantak *et al.* 2013, Saporito *et al.* 2009, 2011, Raspotnig *et al.* 2011), specially ants and mites. The mechanisms involved in toxin absorption are unclear, and even the nature of the dietary sources of lipophilic alkaloids in poison frogs remains a major research challenge for chemical ecologists (Daly *et al.* 2000). Although it is believed that most lipophilic alkaloids in poison frogs comes from ants and mites, other sources might be involved. Recent reports of trace amounts of alkaloids in unfertilized eggs suggest that dendrobatid mothers may provide both food and chemical defense through parental care (Stynoski *et al.* 2014a, 2014b). Once food with lipophilic alkaloids is ingested, various tissues can be able to sequester these chemicals, and it is possible that the toxin will be bound and transported by some element in the blood, passing through various tissues to eventually accumulate in the skin. However, the identity of such alkaloid transporters in poison frogs remains unknown. In fact, little is know about any proteins that bind neurotoxins in general, except for saxitoxin, a shellfish alkaloid neurotoxin that targets sodium channels and is bound by saxiphilin in the plasma. Once more genetic information becomes available for poison frogs, molecular evolution methods could be employed to determine candidate genes that may play a similar role as saxiphilin in the binding and transport of alkaloid toxins. For now, the mechanisms used by poison frogs for alkaloid uptake, including capture, transport, and accumulation in dermal granular glands, remains a mystery (Santos *et al.* 2016).

To add to our lack of understanding of this system, we also have very little information on resistance to lipophilic alkaloids in poison frogs. After toxin sequestration, it has been demonstrated that alkaloids became anatomically widespread in the poison frog *Melanophryniscus simplex* (Grant *et al.* 2012), indicating that physiological resistance evolved in *M. simplex* (Daly *et al.* 2000) and possibly in other poison frogs. Many alkaloids were found in *M. simplex* (Grant *et al.* 2012) and many other bufonids and dendrobatids disruption-channel activity or neurotransmitter-receptor binding in nerve and muscle cell (Daly *et al.* 1999), but only in a few cases the genetic modifications related to alkaloid insensitivity were studied (*e.g.*, (Wang & Wang 2017). One typical example is the physiological resistance to the lipophilic alkaloid batrachotoxin (BTX) in *Phyllobates terribilis* (Myers *et al.* 1978a). The resistance mechanism, in this case, appears to be due to modification of the regulatory site controlling a voltage-dependent NA+ channel activation and permeability, thus preventing binding by BTX (Daly *et al.* 1980, Wang & Wang 1999, Wang *et al.* 2006, Hanifin 2010, Wang & Wang 2017).

Resistance to lipophilic alkaloids has been investigated only in *Phyllobates terribilis* (*e.g.*, Wang & Wang 2017). The evidence suggests that *P. terribilis* resists to alkaloids such as batrachotoxin due to a modification of the regulatory site controlling voltage-dependent sodium (Na^+) channel activation and permeability, thus preventing binding by the neurotoxin. However, poison frog alkaloids have diverse mechanisms of action (Daly *et al.* 1999, 2005a), and Grant *et al.* (2012) showed that lipophilic alkaloids are anatomically beyond the skin glands. Therefore, specialists argue that physiological resistance and uptake evolved in tandem and that we are only scratching the surface of the complicated poison frog resistance mechanisms to these toxins.

Beyond peptides, bufadienolides, and lipophilic alkaloids, the list of defensive chemicals in amphibians also include biogenic amines, proteins, other types of steroids, and some volatiles (*e.g.*, Daly *et al.* 1987, Daly 2004, Daly *et al.* 2005, Pukala *et al.* 2006), as well as hydrophilic alkaloids such as the famous neurotoxin tetrodotoxin (TTX). The TTX is better know for the widely studied case of physiological resistance: the tolerance this chemical that evolved multiple times in garter snakes (*Thamnophis*), in a predator-prey arms race against newts of the genus *Taricha* (Geffeney *et al.* 2002). Current evidence suggests that physiological resistance emerged through specific mutations in a functional region of a TTX-sensitive NA^+ channel gene (Nav1.4) that alters the channel pore reducing TTX binding affinity (Geffeney *et al.* 2005, Geffeney & Ruben 2006, Feldman *et al.* 2009).

Understandably, natural products discovery has oriented most studies on these defensive chemicals. Hence, it is no surprise that scientists have already screened close to 545 amphibian species for bioactive compounds by the end of 2012 (not including Dendrobatidae and Mantellidae with their alkaloid-containing skin secretions; see König *et al.* 2015 and references therein). Nevertheless, such studies are severely limited not only in their ability to explain their findings (*e.g.*, without understanding of amphibian phylogeny and ecology, variation in the kinds and amounts of defensive chemicals is unintelligible), but also the efficiency of their amphibian sampling (*e.g.*, phylogenetic trees provide roadmaps for natural products discovery; for example, see Smith & Wheeler 2006, Saslis-Lagoudakis *et al.* 2012, and Garnatje *et al.* 2017).

Currently, the underlying molecular data required to test the presence of the core genetic elements of amphibian chemical defense in the common ancestor of all crown-group anurans is not available. More than that, we currently don't have enough information to identify all those elements across the amphibian tree of life and understand their evolution.

Recent advances in DNA sequencing technologies make a genomics approach to the evolution of chemical defense in poison frogs feasible. The following sections are focused on how the newest advances in DNA sequencing technologies promise to scale up the amount of evidence available for evolutionary studies, leading amphibian systematics and the evolutionary study of chemical defense in amphibians one step further.

Recent advances in DNA sequencing

The rise of first-generation DNA sequencing technologies in the 1970's promoted the generation of overlapping genomic regions using DNA enriched for a single locus (see França *et al.* 2002 for a review on early sequencing methods). First-generation sequencing was first developed by Sanger & Coulson (1975) and Sanger *et al.* (1977) (the chain-termination method, commonly known as Sanger sequencing) and in parallel by Maxam & Gilbert (1977) (a chemical sequencing method). However, it was only in the 1980's that the polymerase chain reaction (PCR) and the use of a heat-stable DNA polymerase from *Thermus aquaticus* (Taq polymerase) arrived. These discoveries were responsible to finally making sequencing reactions (cycle-sequencing) with reduced amounts of DNA template compared to isothermal enzymes possible (Mullis *et al.* 1986, Mullis & Faloona 1987).

Sanger sequencing ultimately prevailed over the chemical sequencing method because it was less technically complex and more amenable to being scaled up (Schadt *et al.* 2010). Sanger sequencing has been the dominant method of directly sequencing DNA, and has dominated the DNA sequencing market for nearly past 30 years (Varshney *et al.* 2009). It revolutionized many fields of molecular biology and allowed monumental accomplishments including the completion of the (IHGSC 2004).

First-generation sequencing (1G) methods have many limitations which are primarily related to its reliance on the visualization of the distribution of fluorescent dyes at the terminal ends of products for base calling. These limitations restrict Sanger technology to a single template, and to gathering data one locus at a time (Carstens *et al.* 2012). Further restrictions include the requirement of high DNA concentration, the short read length (less than 1000 nucleotides per sample), and the incapability to sequence some regions. Finally, although innovations in Sanger technologies have increased the number of samples sequenced per machine at the same time, the length of reads remains unchanged, and the costs are still high (approximately $2 per run per machine).

The need to overcome Sanger sequencing constraints has catalyzed the development high-throughput sequencing (HTS), starting in the 2000's with next-generation sequencing (NGS) or, more precisely, second-generation sequencing (2G). The 2G platforms aggregates various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. Another important characteristic of 2G platforms is that they can sequence libraries of the template instead of a single fragment of DNA isolated via PCR (Carstens *et al.* 2012). The specificities of 2G, as well as newer, less established methods such as third-generation (3G) and fourth-generation (4G) sequencing, are beyond the scopes of this dissertation but the interested readers can refer to the specialized literature for more information (Metzker 2010, McCormack *et al.* 2012b, Ku & Roukos 2013, Rhoads & Au 2015). It suffices to say that HTS offers a significant advance in molecular data obtainment once it makes the cost per raw megabase of DNA sequence more than three orders of magnitude less expensive than in 1G methods (see **Figure 3**). Also, each instrument run can produce more than one billion short reads with 25-1000 bases each, allowing generation of complete genomes within hours/ days (Niedringhaus *et al.* 2011).

The National Center for Biotechnology Information's (NCBI) "GenBank and WGS Statistics" website (available at https://www.ncbi.nlm.nih.gov/genbank/statistics/, last accessed on Sep. 11, 2017) provides a snapshot of the overall impacts of HTS in science. According to the webpage, the number of nucleotide bases in whole genome shotgun (WGS) projects (see "Whole Genome Shotgun Submissions", available at https://www.ncbi.nlm.nih.gov/genbank/wgs/, last accessed on Sep. 11, 2017) surpassed the number of bases in GenBank (Benson *et al.* 2017) in August 2005 and the number of WGS sequences exceeded the number of sequences in GenBank nine years later, in August 2014. Currently, there are 2.24 times more WGS sequences than GenBank sequences, and the number of WGS bases is 9.21 times greater than the number of bases in GenBank. See **Figure 4**.

The scale and efficiency of HTS is providing unprecedented progress in a variety of fields, including genomic structural analysis, the study of proteins and nucleic acids interactions, molecular epidemiological analysis, forensic analysis, etc. (*e.g.*, Didelot *et al.* 2012, König *et al.* 2012b, Veltman & Brunner 2012, Wilson *et al.* 2013). Yet the capacity to generate the data significantly outpaces our ability to analyze it (Nekrutenko & Taylor 2012). Each area of research will now face particular challenges to integrate HTS data into its

evidential basis. In the following sections, we will discuss, respectively, the application of HTS to phylogenetic systematics and comparative genomics.

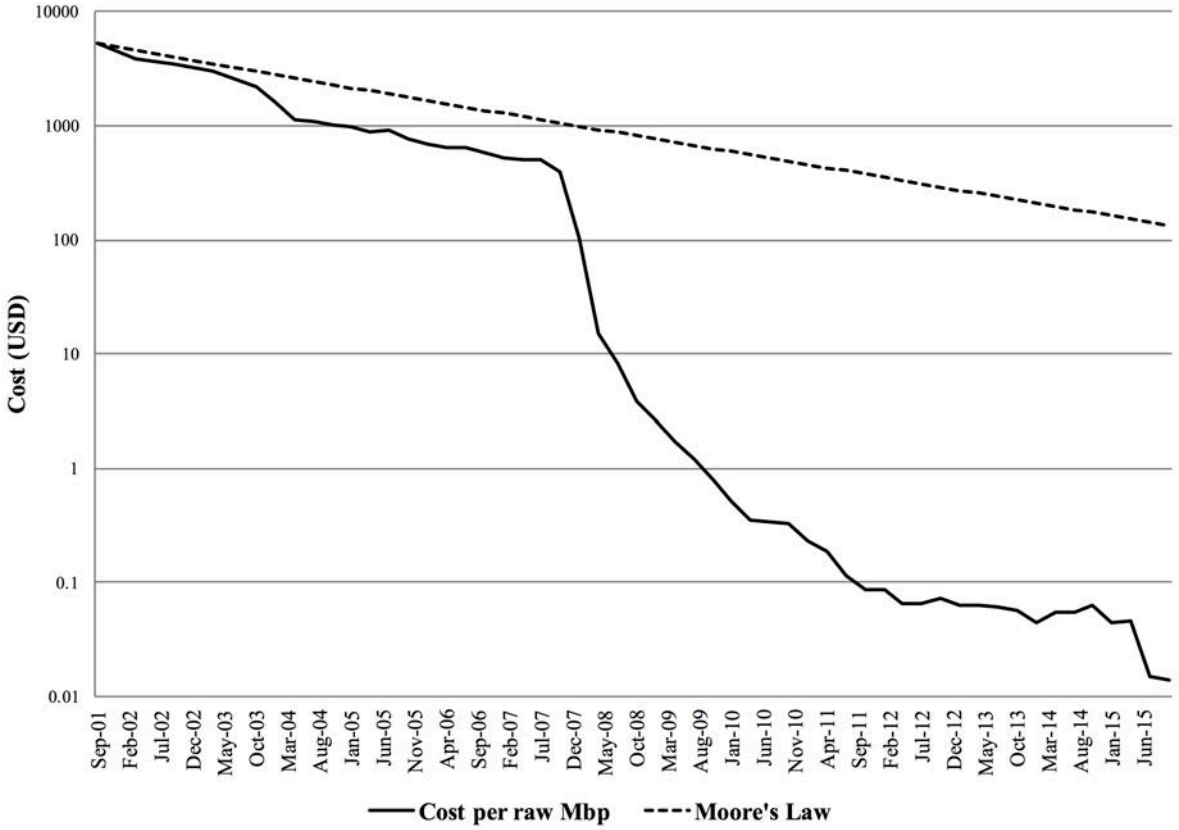


Figure 3 – The evolution of the cost per raw megabase of DNA sequence in comparison to the Moore's Law, which describes a long-term trend in the computer hardware industry that involves the doubling of "compute power" every two years. This graphic illustrates the paradigm shift in disciplines that rely on DNA data, from data poor to data rich, and the difficulties in processing all the DNA information available. Source of data: https://www.genome.gov/sequencingcostsdata/.

The need and challenges of integrating HTS to phylogenetics

Modern phylogenetics was spawned by Sanger sequencing and by PCR approaches applied to mitochondrial DNA in the late 1980's, followed by the adoption of nuclear sequence data during the 1990's. As a result, using multiple loci to infer the history of different taxa has become the baseline in phylogenetics (Brito & Edwards 2009). However, accessibility to a small number of genes has restricted most phylogeneticists. Each of these genes can evolve in radically different ways, and their phylogenetic signals and substitution processes may diverge drastically from one another. From this emerges a trend toward amassing larger data sets to include more informative sites and increase nodal support (Brito & Edwards 2009 see Smith *et al.* 2013).

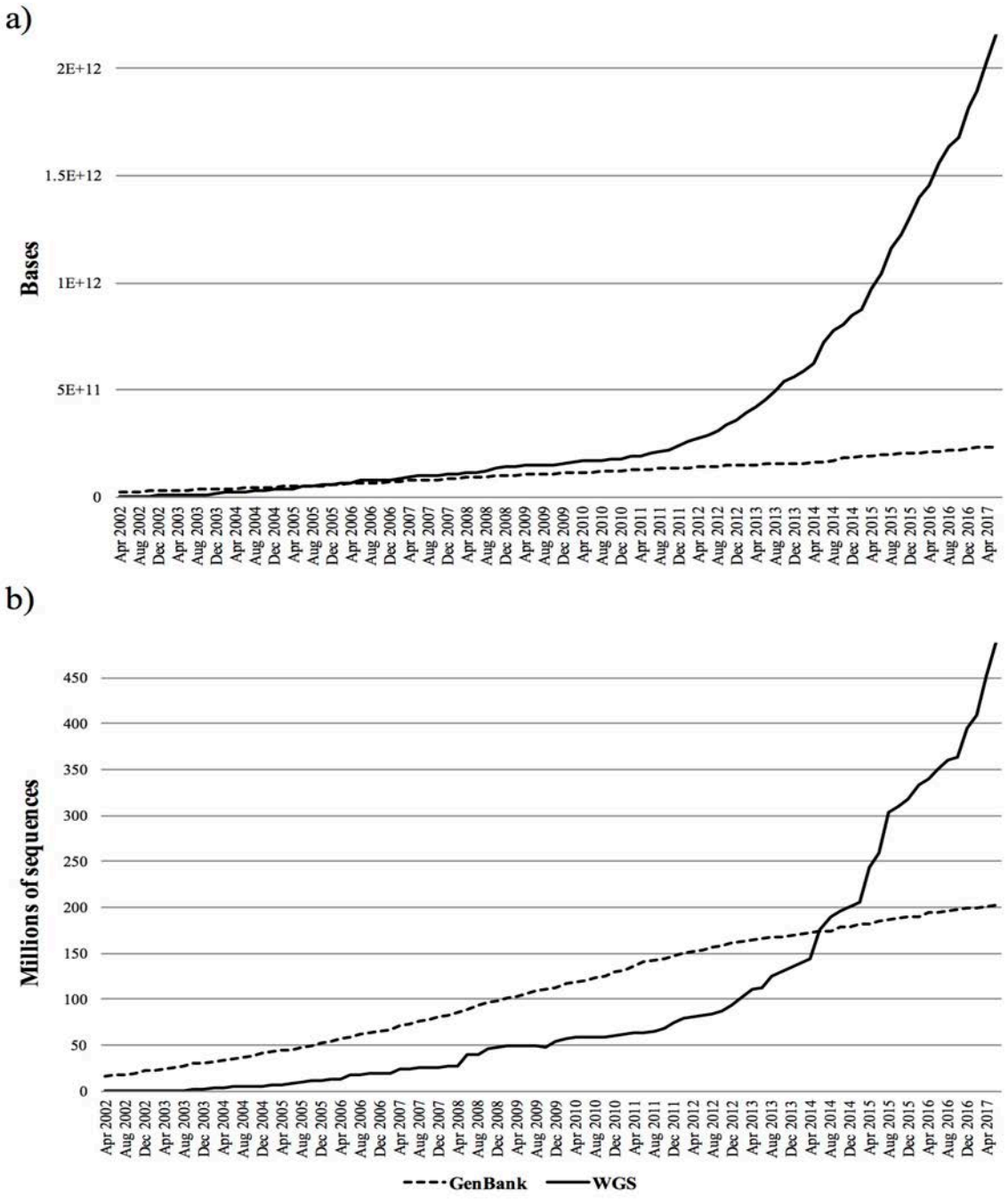


Figure 4 – The number of (a) bases and (b) sequences in GenBank in comparison with whole genome shotgun (WGS) projects, since April 2002 until June 2017. Source of data: https://www.ncbi.nlm.nih.gov/genbank/statistics/.

As scientists recognize the limits working with a small number of genes, which traditional sequencing methods impose, (*e.g.*, Edwards & Beerli 2000), they become increasingly frustrated with time and cost expenses associated with gathering data on a locus-by-locus basis. Under these circumstances, it is understandable that phylogeneticists have

been looking toward HTS with a large interest as a potential means to abridge the steps of multilocus data generation into a more cost-effective procedure (Carstens *et al.* 2012, Mccormack *et al.* 2012a). Nevertheless, Mardis (2008) argued that HTS has been slow to take root in phylogenetics compared to other fields like metagenomics and disease genetics, and most experts would say that the gap between the techniques used for applied research and the methods employed for phylogenetics is still considerable.

The general problem is that phylogenetic analysis requires that homologous character states be identified to infer transformation events between them (Kluge & Grant 2006, Grant & Kluge 2009). However, the generation of homologous sequences is not as highly targeted or straightforward in NGS as it is in traditional Sanger sequencing methods. With NGS, DNA is not necessarily enriched for single locus via PCR-based amplification, although this is one possible application, but for many loci through a variety of methods involving reduction of the size of the genome. One alternative is to analyze only a subset of the data, either by reducing genomes to a few specific genes (*e.g.*, Herniou *et al.* 2001) or analyzing random fragments (*e.g.*, Vishnoi *et al.* 2010). Still, this results in the exclusion of large amounts of data and defeats the purpose of sequencing whole genomes. In theory, dynamic homology analysis (Wheeler 2006) allows whole genome phylogenetic analysis without prior alignment by just increasing the classes of events to include genome-level transformations (*e.g.*, rearrangements, horizontal transfers, inversions, fragment indels and duplications). But the computational cost of dynamic homology analysis makes it impossible to apply to more than a handful of terminals. Most methods disregard homology altogether and construct trees based on measures of overall genome similarity, such as shared gene content (Snel *et al.* 1999), genome blast distance (Henz *et al.* 2005), and feature frequency profiles (Sims *et al.* 2009). However, such approaches are phenetic and inherit all of the problems of that failed research program. Finally, although theoretical information approaches that employ data compression techniques were presented (Giancarlo *et al.* 2009, Nalbantoglu *et al.* 2010), existing methods still require unrealistic assumptions and are too computationally expensive to be feasible in phylogenetic relevant analyses of hundreds or thousands of genomes. The challenge, therefore, is to take advantage of NGS in large scale phylogenetic studies without wasting resources by excluding most of the data or resorting to phenetic methods of analysis.

Further difficulties emerge when phylogeneticists focus their research on non-model organisms. In these situations, scientists face uncertainty about which sample preparation methods and analyses are appropriate for different research questions at various evolutionary

timescales contributes, which ultimately adds to the lag between cutting-edge HTS and phylogenetics (Mccormack *et al.* 2012b). Shortening the distance between HTS and phylogenetics will, therefore, require concentrating efforts on non-model organisms and homology problems. Given the ecological importance of amphibians (Halliday 2008), their role in pharmacology and toxicology (Daly *et al.* 2000), and the availability of a sound basis for taxonomic and phylogenetic work (Frost *et al.* 2006a), the advantages of using this group as a model became evident.

As knowledge of amphibian species diversity has increased, so has our ability to extract and analyze biological information for phylogenetic analysis. Nevertheless, the identification of the optimal phylogenetic solution for a given dataset is one of the most computationally challenging problems known, despite the continuous developing of new analytical tools and theoretical approaches (*e.g.*, Stamatakis 2006, Drummond *et al.* 2006, Wheeler 2006, Goloboff *et al.* 2008, Wheeler *et al.* 2015 p. 5, Nguyen *et al.* 2015, Höhna *et al.* 2016). Similarly, the genomic revolution has substantially increased the size of phylogenetic datasets. Modern amphibian systematics studies usually include 1–10 gene regions and 1–10 kilo bases (kb) of DNA sequence data. Although this entails an increase in the evidential basis of amphibian systematics of several orders of magnitude (*e.g.* compare Ford & Cannatella 2013 and Frost *et al.* 2006), it remains an absurdly small sample of the 1.7 Gb and 20,000 protein-coding genes that comprise the genome of *Xenopus tropicalis* (Gray, 1864) (Hellsten *et al.* 2010a), suggesting that our current understanding may yet be radically overhauled. Therefore, the importance of employing HTS technologies for whole genome sequencing of selected taxa becomes evident, and identifying candidate gene regions for phylogenetic analysis, as well as designing probes for direct amplicon HTS, becomes the obvious next move that will substantially increase the number of loci available to the scientific community.

The promises of comparative genomics in amphibian chemical defense

In the previous sections, we reviewed, in general lines, the current state of basic and applied research on the evolution and biology of amphibians in general, and on amphibian chemical defense in particular. Such studies provide evidence for how physiological resistance evolved in specific cases and present general guidelines and questions to be addressed through more comprehensive research. Whereas chemical defense in poison frogs evolved multiple times

independently, have the genetic modifications implicated in physiological resistance evolved through evolutionary convergence at the molecular level? And if so, to which extent natural selection and evolutionary constraints are involved in convergent evolution of physiological resistance to lipophilic alkaloids in poison frogs?

A comparative genomics strategy is promising to start addressing these gaps in our understanding about amphibian chemical defense. Comparative analysis of genome sequences is a major part of the effort of finding functional parts of genome sequences (Hardison 2003). Thanks to HTS, it is possible to rapidly compare genomes of both close and distantly related organisms and identify germline and somatic variants of interest, such as insertions and deletions (indels), copy number variants (CNVs), single nucleotide polymorphisms (SNPs), and other structural variations. In one hand, comparing the genomes of distantly related organisms allow us to identify the core set of proteins shared by these organisms as well as sequences that are more likely to be functional given the signature of purifying selection (*e.g.*, (Clamp *et al.* 2003). On the other hand, comparing genomes of closely related organisms allow us asking what sequences account for unique features of organisms (*e.g.*, Stein *et al.* 2003) and have managed to identifying genetic changes associated with specific phenotypic traits (*e.g.*, toxin sequestration strategies and different mechanisms of tolerance to lipophilic alkaloids), discovering positively selected genes that may be related to evolutionary adaptation, and identifying expansion and contraction of relevant gene families, among other applications (*e.g.*, Hardison 2003).

The comparison of the nuclear genomes of poison frogs and other batrachians have a great potential to help to identify both new DNA markers for phylogenetic systematics and genetic changes associated with the development of genetic mechanisms related to chemical defense (*i.e.*, lipophilic alkaloids sequestration and resistance), which will provide new directions for future research. Hence, despite the enormous effort involved in building a bridge between basic research in non-model organisms and cutting edge DNA sequencing technology, it promises a high payback. Each chapter of the current dissertation is, therefore, a stepping-stone in the direction of a larger line of investigation that ultimately leads towards enhancing the communication between basic and applied research, using amphibians as our selected model.

Organelle genomes are a major component of the total genome content of the eukaryotic cell. Mitochondrial DNA (mtDNA) sequences are essential sources of information for a broad range of studies, from population genetics to phylogenetics, ultimately improving

our knowledge on the evolution of both genomes and organisms (*e.g.*, Hancock-Hanser *et al.* 2013, Darrin Hulsey *et al.* 2013, Bertrand *et al.* 2015). However, assembling the complete mitochondrial genomes (mitogenomes) of non-model organisms, especially when reference sequences are lacking, can be challenging. More than that, although sequencing costs have dropped, basic research might still find certain constraints in fundings for both sequence and data analysis. Hence, C**hapter 2** aims to propose feasible solutions for basic researchers assembling novel mitogenomes, leveraging as much as possible of sequence data that is already available, and using minimal amounts of computational resources.

Once the novel mitochondrial genome is available, there are many strategies for annotation and comparative analysis (*e.g.*, Bernt *et al.* 2013, Laslett & Canbäck 2008, Lowe and Eddy 1997, Schattner *et al.* 2005). However, studies addressing how to infer the completeness of those sequences are lacking. A quick survey of the recent specialized literature (searching https://www.scopus.com and https://scholar.google.com.br for "complete mitochondrial genome" within the last three years) shows that many authors infer circularity through visual inspection of reads at the ends of the assembly (*e.g.*, Gan *et al.* 2014, Grau *et al.* 2015, Vacher *et al.* 2016) or using other more convoluted methods of visual inspection (*e.g.*, (Cong and Grishin 2016a). Nevertheless, this practices can lead to erroneous inferences of sequence completeness especially if reference sequences are not available for comparison. Hence, **Chapter 3** introduces a new *ad hoc* mapping strategy to test for assembly circularization, using Bowtie2 alignment scores and a new per-position sequence coverage value (which we named "connectivity") to assess the quality of the inferred circularization.

Chapter 4 address the assembly of the draft nuclear genomes of the eastern spadefoot toad (*Scaphiopus holbrookii*) and the golden poison frog (*Phyllobates terribilis*), with focus on the homology based and *ab initio* annotation of protein coding genes and the development of possible probes for phylogenetic analysis of amphibians. Once this genomic data is consolidate, it will advance comparative studies on the genetic bases of toxicity (including mechanisms involved in sequestration, biosynthesizes, biotransformation, and resistance) in amphibians. Until them, these new genomic data from amphibians are immediate sources of phylogenetic markers to investigate amphibian diversification at different scales.

Chapter 5 address the *de novo* assembly of repetitive DNA in the nuclear genome of frogs, proposing new insights on the role of repeats in the variation of genome size in amphibians and new "Hennigian" characters for phylogenetic investigations.

While all the previous chapters discuss original DNA data and the methods used to analyses it, **Chapter 6** is a reply to the first work to investigate the genetic bases of alkaloid resistance in poison frogs by Tarvin *et al.* (2016).

In combination, the next chapters are also an attempt to provide guidelines for new frog genome projects, from nucleotide extraction to gene annotation.

Taxon sampling

Through the chapters of this thesis, many frog species will serve as source material for genomic DNA extractions, sequence assembly, and bioinformatics analyses. Given licenses and the availability of biological material, taxa selection followed two guidelines. First, favor pairs of amphibian species with and without lipophilic alkaloids that are relatively close to each other in terms of phylogenetic relationships, sponsoring current and future research on comparative genomic analyses seeking to find the genetic bases of alkaloid sequestration and resistance. Second, favor amphibians that are likely diploids and for each the genome size and be estimated based on direct observation or their phylogenetic relationships. **Figure 5** shows the species that we selected following the guidelines above.

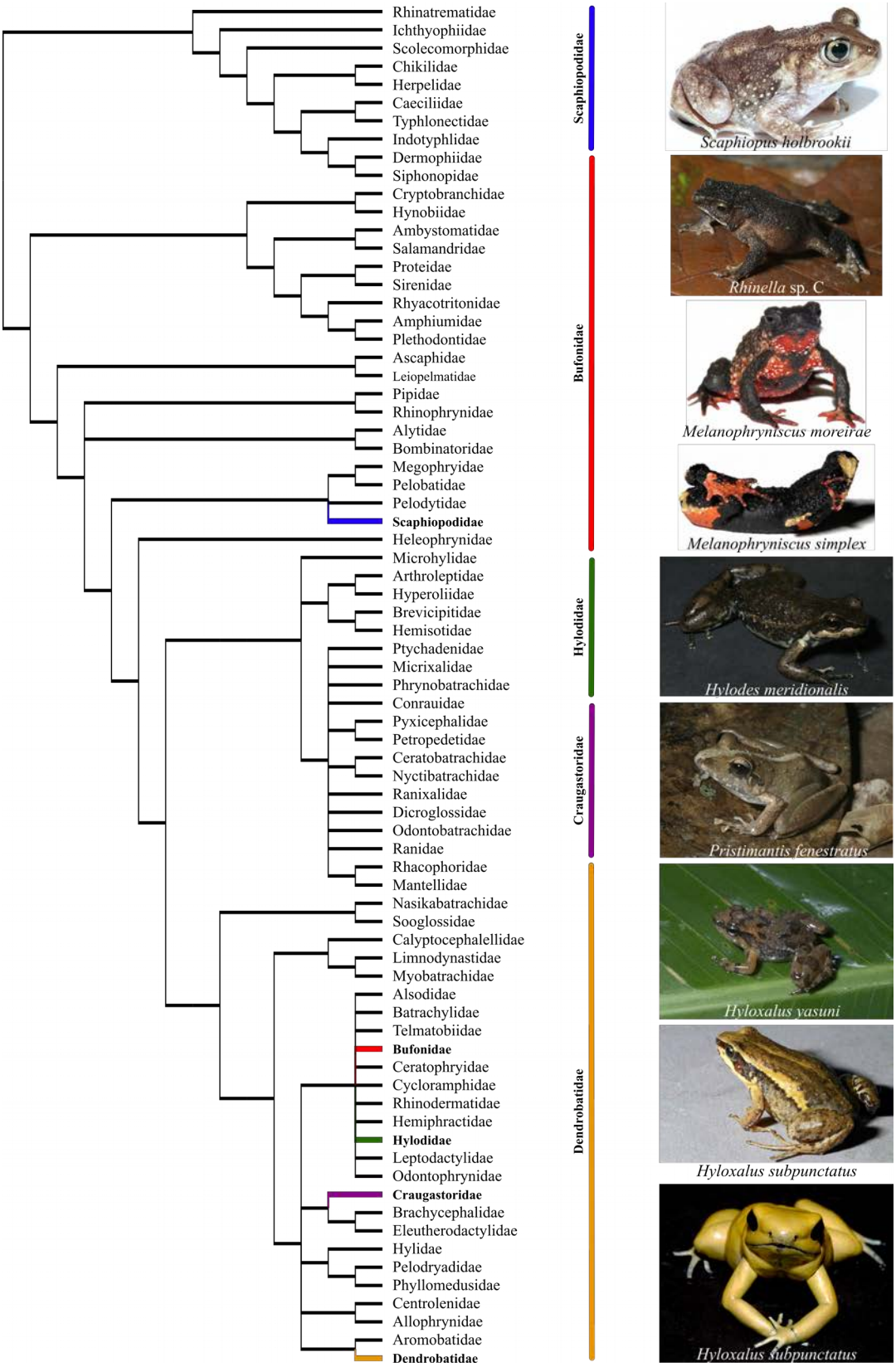


Figure 5 – The phylogenetic position of selected taxa in the cladogram of families of Anura (based on Frost *et al.* 2006, Pyron & Wiens 2011, Padial *et al.* 2014, and Grant *et al.* 2017).

2. MITOGENOME ASSEMBLY FROM GENOMIC MULTIPLEX LIBRARIES

Background

Most vertebrate mitochondrial genomes (mitogenomes) are about 15–22 kbp, double-stranded, circular DNAs that encode a set of 37 genes (two rRNAs, 13 proteins and 22 tRNAs), as well as a major non-coding region (control region, CR) that accounts for much of the mitogenome size variation (Gissi *et al.* 2008). Mitochondrial DNA (mtDNA) sequences have applications in a wide range of studies, from population genetics to phylogenetics, ultimately improving our knowledge on the evolution of both genomes and organisms (*e.g.*, Hancock-Hanser *et al.* 2013b, Darrin Hulsey *et al.* 2013b, Bertrand *et al.* 2015). Until the 2000s, only a few model organisms have had the molecular biology of their mitochondrial systems studied (Boore 1999). More recently, next-generation sequencing (NGS) and advances in bioinformatics tools have enabled the analysis of mitogenomes to extend to non-model organisms on an unprecedented scale (Mardis 2008).

The specialized literature has proposed numerous methods for rapidly assembling mitogenomes directly from shotgun sequencing (*e.g.*, (Cameron 2014, Gan *et al.* 2014b, Lounsberry *et al.* 2015). These methods are intended for fast recovery, assembly and annotation of mitogenomes as the primary research objective. For example, Gan *et al.* (2014) provide a detailed protocol for the fastest recovery, assembly, and annotation of mitogenome using the MITOBIM software (Hahn *et al.* 2013), the MITOS (Bernt *et al.* 2013a) annotation web service and data from the Illumina MiSeq platform. However, in addition to studies designed specifically to capture mitogenomic sequences, whole-genome sequencing, targeted amplicon sequencing and hybrid enrichment approaches also capture mitogenomic reads as by-catch, albeit with significantly lower coverage and quality. Hence an efficient bioinformatics pipeline is required to extract and assembly mitogenomes from limited data.

Here, we add to the methods for harvesting complete mitogenomes from whole-genome multiplex libraries sequenced using the Illumina HiSeq platform and compare the performance of different assembly strategies when read number and quality are limited. As test data, we present novel, near-complete mitogenomes from five South American frog species of the families Bufonidae, Craugastoridae, Dendrobatidae and Hylodidae.

Material and methods

Taxon selection and data archiving

We sequenced five South American species of frogs from four families: the torrent frog *Hylodes meridionalis* (Mertens, 1927) (Hylodidae), the rocket frog *Hyloxalus yasuni* Paez-Vacas, Coloma, & Santos, 2010 (Dendrobatidae), the rain frog *Pristimantis fenestratus* (Steindachner, 1864) (Craugastoridae) and the red-belly toad *Melanophryniscus simplex* Caramaschi and Cruz, 2002 and an undescribed species of beaked toad (*Rhinella acrolopha* group *sensu* Grant & Bolıvar-G 2014) that we refer to as *Rhinella sp*. C. (Bufonidae).

Total DNA extraction and sequencing

Muscle tissue samples were stored in 70% ethanol at -20 °C for several months or years. Materials were separately pooled for DNA extraction using the AGENCOURT DNAdvanceTM Genomic DNA Isolation Kit. Total genomic libraries were prepared using a NEBNext DNA Library Prep Master Mix (Neb #E6040S) and sequenced using an Illumina HiSeq 2000TM at the multiuser high-throughput sequencing facility of the University of São Paulo Luiz de Queiroz College of Agriculture. Libraries were distributed in two lanes of a standard Illumina HiSeq 2000 flow cell and sequenced using the high-throughput module. However, each lane also received an unknown number of additional libraries, reducing the expected total number of reads from approx. 250,000,000 to 38,903,325 (approx. 75% fewer) paired-end reads of 100 bp.

We chose the Illumina platform because it produces high-quality data for various scales of analysis at costs that have decreased substantially relative to other second-generation sequencing instruments (*e.g.* 454/Roche and SOLiD; see Mardis 2013). Among Illumina platforms, the Genome Analyzer IIx (GA IIx) is less automatable and produces fewer data than the MiSeq and HiSeq systems, which are therefore preferred by most researchers interested in large-scale analysis. Gan *et al.* (2014) selected the Illumina MiSeq over the HiSeq platform due to its reduced run time and more tractable data. Nevertheless, the HiSeq platform is the system of preference in numerous research projects targeting elements of nuclear DNA (nuDNA), such as microsatellite analysis (Castoe *et al.* 2012) and whole-genome sequencing (Sun *et al.* 2015a).

Computational resources

All *in silico* procedures were executed using "ACE", an SGI rackable computer cluster housed in the Museum of Zoology of the University of São Paulo. Selected servers had four 2.3 GHz Operon CPUs with 16 cores each and 256 or 516 GB of memory. After optimization, we were able to reconstruct genomes using a single core and ca. 20 GB of memory. The software environment in ACE consists of a SUSE Linux Enterprise Server with SGI Performance Suite, SGI Management Center and PBS Pro Job Scheduler.

Quality control

As stated by Yang *et al.* (2013: 14), 'to get reliable result[s] in downstream analysis, it is necessary to remove low-quality reads, avoiding mismatches in read mapping and false paths during genome assembly'. Due to its function versatility and run-time efficiency, we selected the HTQC toolkit (Yang *et al.* 2013) to perform read quality assessment and filtration. The complete quality control protocol is described below and the step-by-step procedures are given in Protocol S2.1 (see **Supplementary material**).

Raw reads from each pair were pre-processed using a series of Unix commands and a package of home-made Python scripts (PATO-FU). The programs ht-stat, ht-filter and ht-trim are components of the HTQC toolkit and were employed as follows: the summary of the sequencing read quality was generated with ht-stat. In order for tile selection to be automated and repeatable, we post-processed the ht-stat results using a homemade Python script (selectTiles.py). Tile removal followed criteria derived from the HTQC guidelines: (i) more than 50% of the reads have quality score below 10; (ii) <10% of the reads have quality >30; and (iii) more than 50% of the reads have quality below 20. Selected tiles were removed with ht-filter. Remaining reads were trimmed with ht-trim, removing low-quality bases from reads' heads or tails. Finally, short reads were removed with ht-filter and the quality of filtered reads was evaluated using FastQC (Andrew 2010). Only paired-end filtered reads were used for assembly.

Mitogenome assembly

We analyzed the filtered reads of *Hylodes meridionalis* using three assembly strategies: (1) mapping against a reference mtDNA genome ('reference based'); (2) *de novo*;

and (3) baiting and iterative mapping. To implement each strategy, we selected the best-commented and most frequently used software in the specialized literature.

Reference-based assembly (1) was performed using Bowtie2 v2.2.3 (Langmead & Salzberg 2012). The mitogenome of the Tibetan toad *Bufo tibetanus* (NCBI accession number NC 020048; Wang *et al.* 2013), which is currently a junior synonym of *B. gargarizans* (for taxonomic comments see (Frost 2017), was selected as reference due to its completeness and phylogenetic position and the reliability of the long PCR-based amplification method used to sequence it.

For *de novo* sequence assembly (2) we used the programs SOAP-denovo2 v2.04 (Luo *et al.* 2012), ABySS v1.5 (Simpson *et al.* 2009) and Velvet v1.2.10 (Zerbino & Birney 2008b). SOAPdenovo2 v2.04 was run with average insert sizes of 150, 200 and 250 bp. ABySS and Velvet were run for all k-mer sizes from 21 to 63, with incremental steps of 2. BLAT (Kent 2002) was used to map contigs and scaffolds against the reference genome of *B. tibetanus*.

For the baiting and iterative mapping strategy (3), we used Mira v4.0 (Chevreux *et al.* 1999) and a modified version of MITObim.pl v1.6 (Hahn *et al.* 2013). This strategy has two main steps (Hahn *et al.* 2013). First, reads are mapped against a reference sequence in Mira, effectively generating a new reference based on the most conserved regions. New reads with overlap are then iteratively fished from the read-pool and mapped against the previous reference using MITObim. Each iteration in MITObim expands the novel reference sequence until reaching a stationary number of reads. This approach only returns a single-padded consensus sequence in the end, but sequences can be connected by "N" to indicate that the fragments are not connected by reads are probably not contiguous.

Four baiting and iterative mapping strategies were employed: (i) mapping to the complete mitogenome of a closely related species (*B. tibetanus*); (ii) mapping to the mitogenome of a more distantly related species (a salamander, *Rhyacotriton variegatus*; NCBI accession number NC 006331; Mueller *et al.* 2004); (iii) baiting with a barcode seed (the cytochrome C oxidase subunit I [*cox1*] gene sequence from *B. tibetanus*, NCBI accession number NC 020048, 5533–7044 bp) with the *de novo* option off; and (iv) same as (iii) but with the *de novo* option on. Only consensus sequences with average coverage >20 and average quality >80 were accepted. If more than one consensus sequence was recovered, the longest one was chosen for further analysis.

The optimal mitogenome assembly strategy was selected according to the number of reads used, total ungapped sequence size, average coverage and consensus quality. This

strategy was then applied to assemble mitogenomes using the libraries of *Hyloxalus yasuni*, *Pristimantis fenestratus*, *Melanophryniscus simplex* and *Rhinella sp*. C. The complete bioinformatics protocol for assembly is available in Protocol S2.2.

Mitogenome annotation and comparison

Assemblies in CAF format were parsed using a homemade Python script (parseCaf.py) to extract DNA data and evaluate the coverage and quality of each mtDNA element. Preliminary *de novo* mitogenome annotation used the mitochondrial genome annotation server MITOS (Bernt *et al.* 2013) with default parameters. Additional search and validation of tRNA sequences were performed using ARWEN (Laslett & Canbäck 2008) and tRNAscan-SE (Lowe and Eddy 1997, Schattner *et al.* 2005). Automated annotation was confirmed and edited manually by comparison to published anuran mitogenomes (**Table S2.1**). The control region (CR), which typically lies between cytochrome B (*cytb*) and the LTPF tRNA cluster in neobatrachians (Zhang *et al.* 2013), was annotated using sequence similarity searching with BLAST using default parameters (Altschul *et al.* 1990).

Results

Software

Home-made Python scripts (PATO-FU, selectTiles and parseCaf) are available at http://www.ib.usp.br/grant/anfibios/researchSoftware.html and https://gitlab.com/MachadoDJ/ under the GNU General Public License version 3.0 (GPL-3.0). We modified the MITObim original script so it would create manifest files for MIRA pointing to a directory in a local file system in a cluster environment. Modifications to MITOBIM allow multiple mitogenomes to be reconstructed simultaneously using the same compute node. The modified MITObim script is available at http://www.ib.usp.br/grant/anfibios/researchHPC.html.

Quality control results

Comparison of quality reports before and after quality control shows major improvements in per base/tile sequencing quality and over-represented sequences (see **Table S2.1**). Some filtered sequence files still failed per base sequence content and k-mer content tests; however, according to the FASTQC help page (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/; last access: December 16, 2017), libraries derived from random priming will nearly always show k-mer bias, and sequences subjected to aggressive trimming are more likely to present per base sequence content bias. Quality control took <1 h total computation time using ACE, with <5 min hands-on time.

Comparison of assembling strategies

Reference-based genome assembly using Bowtie2 failed to align mtDNA sequence reads to the *B. tibetanus* reference genome. Likewise, BLAT mapping failed to find mtDNA sequences within contigs and scaffolds generated using the de novo sequencing strategy with ABySS, SOAPdenovo2 and Velvet. Only the baiting and iterative mapping strategy with Mira and MITObim succeeded in assembling mtDNA sequences.

We were able to assemble mtDNA for all three variations of the baiting and iterative mapping protocol. However, the consensus sequences generated using a barcode seed with the '–denovo' option in MITObim did not pass our minimum quality criteria. The remaining assembled consensus sequences were compared according to the number of reads used, total ungapped sequence size, average coverage and consensus quality (see **Table 2**). The N50 and N90 values are incalculable because only one contig remains in the last iteration. The longest ungapped consensus sequences that passed minimum quality criteria were achieved by mapping to the complete mitogenome of the more closely related species (*B. tibetanus*).

Assembly of mitogenomes using MIRA and MITObim took variable amounts of time depending on the reference used and the number of iterations required by MITObim. However, mitogenome assembly using the complete frog mitogenome as reference required fewer iterations and <3-h computation time, with <5 min hands-on time.

Species	Strat.	Iterations	Reads (x2)			Ungapped	Avg.	Avg.
			Raw	*Filtered*	*Used*	consensus size (bp)	coverage	quality
Hylodes meridionalis	1	18	8,608,779	7,745,168	4,389	16,166	26.88	81
	2	49			4,122	15,651	20.96	60
	3	84			3,389	12,079	29.66	87
	4	37			3,261	13,205	23.78	79
Hylodes yasuni	1	35	6,894,772	6,160,445	4,650	16,052	28.28	80
	2	29			4,625	15,946	23.06	64
	3	133			3,324	10,330	33.44	87
	4	Fail	-	-	-	-	-	-
M. simplex	1	8	7,958,678	7,166,358	3,717	16,498	23.2	81
	2	49			3,261	13,633	17.3	55
	3	81			2,503	10,258	26.09	87
	4	38			1,690	7,404	22.07	69
P. fenestratus	1	17	4,714,625	4,213,416	23,566	17,892	130.46	87
	2	47			16,890	15,880	76.13	63
	3	118			17,483	15,966	107.52	88
	4	Fail	-	-	-	-	-	-
Rhinella sp. C.	1	13	9,874,464	8,738,815	4,765	17,050	28.79	84
	2	39			4,081	15,879	20.58	62
	3	68			1,700	6,912	25.91	86
	4	36			1,641	7,608	0	83

Table 2 – Baiting and iterative mapping assembly statistics. Strategies: (1) closely related mitogenome (the Tibetan toad, *Bufo tibetanus*, NCBI accession number NC_020048); (2) distantly related genome (a salamander, *Rhyacotriton variegatus*, NCBI accession number NC_006331); (3) barcode seed (the COI gene sequence of *B. tibetanus*, NCBI accession number NC_020048, 5533–7044 bp), *de novo* option off; (4) same as previous, *de novo* option on. The chosen mitogenomic sequences for each species are highlighted. See gene order information in Table S2.2.

Mitogenomic sequences and gene rearrangements

We recovered the nearly complete mitogenome of all five species of frogs, including the standard 13 protein-coding genes, 2 ribosomal subunits and 21–22 tRNAs (Fig. 1). The number of reads used for assembling mitogenomes and the size of each ungapped consensus sequenced are shown in **Table 2**. We also recovered partial CR sequences for all five mitogenomes. The partial CR of *Hyloxalus yasuni* was recovered in a single, contiguous sequence with all coding genes. The CR sequences of the remaining four mitogenomes were recovered as non-contiguous sequences, with fragments varying from 145 bp in *Rhinella sp.* C. to 2302 bp in *P. fenestratus*.

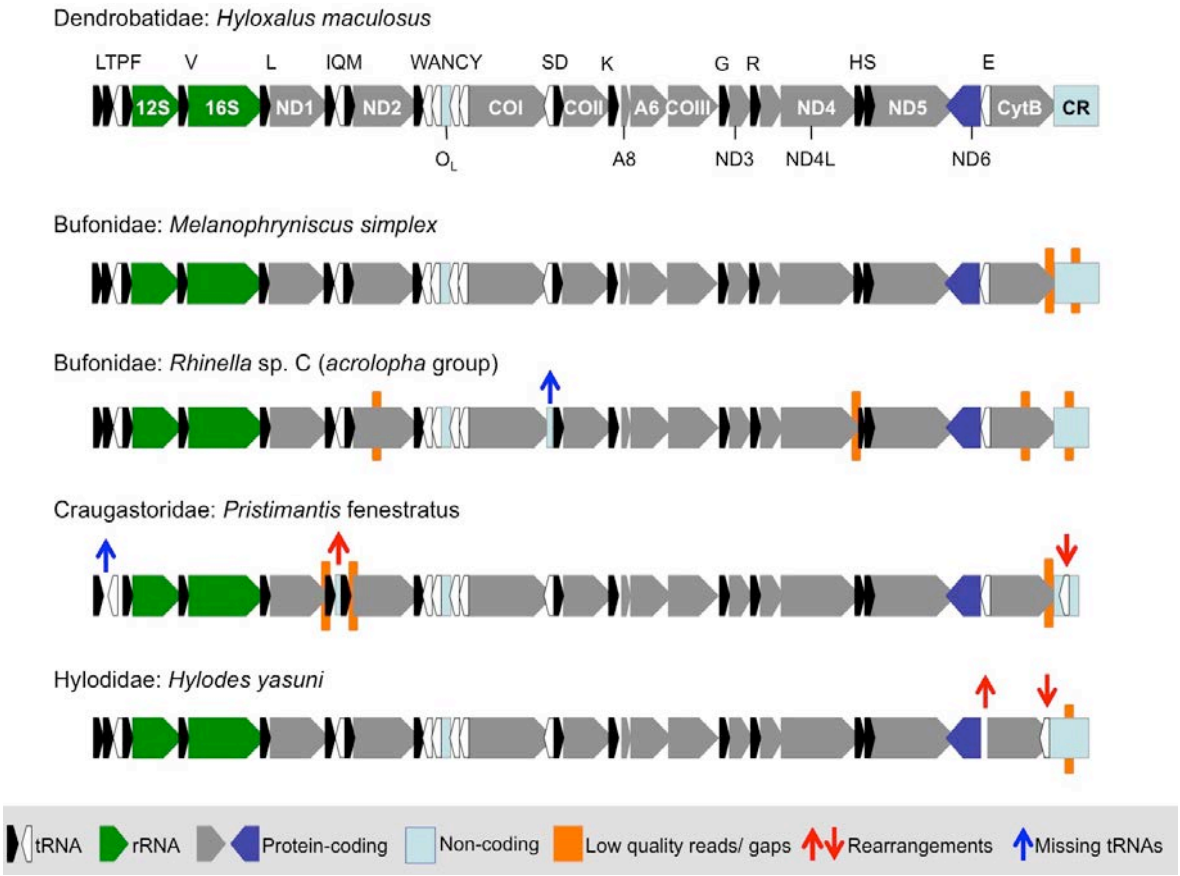


Figure 6 – Gene order and orientation for mitogenomes of five species of South American frogs. The mitogenome of *Hyloxalus yasuni* has all the expected elements in the most common gene order in Neobatrachia. The remaining mitogenomes follow alphabetical order (family: genus). Graphical representation shows elements pointed in the corresponding direction in the mitogenome. See gene order information in Table S2.2.

Screening with parseCaf allowed us to identify only a few poorly sequenced (<109 coverage, quality <40) regions in all mitogenomes, in most cases associated with homopolymeric regions (poly-G or poly-C sequences). In the mitogenome of *M. simplex*, there is a poorly sequenced fragment at the 5′ end of CytB that resulted in a small duplication that was removed manually in the final assembly. In the *Rhinella sp.* C. mitogenome, we found three poorly sequenced regions: two small regions inside the ND2 and CytB gene sequences and one region between ND4 and tRNA-H. The stop codon for ND4 and a fragment of approx. 20 bp of the tRNA-H sequence could not be assembled. These regions were also edited manually and 'Ns' were included in the final assembly. Finally, we found two poorly sequenced regions in the mitogenome of *P. fenestratus*: a small fragment immediately before the tRNA-I sequence and another fragment just after the tRNA-M sequence.

Most genes in the five mitogenomes we report are transcribed from the H-strand,

exceptions being ND6 and eight tRNA genes (**Figure 6**; also see **Table S2.2**), as described in other anurans (Irisarri *et al.* 2012, Zhang *et al.* 2013). The gene arrangement in *Hyloxalus yasuni* and *Melanopryniscus simplex* mitogenomes follows the most common order of Neobatrachia (Zhang *et al.* 2005, Kurabayashi & Sumida 2013). In the case of *Rhinella sp.* C., the gene arrangement also matches the arrangement found in most neobatrachian anurans, except that we were unable to find the tRNA-S2 gene. Although there is a non-coding region in the expected position of this tRNA (*i.e.* just before tRNA-D), the sequence has low similarity with the tRNA-S2 sequence from other anurans and we were unable to predict its secondary structure.

One novel tRNA gene rearrangement was observed in *Hylodes meridionalis*, in which the tRNA-E is located between CytB and the major non-coding region rather than the typical neobatrachian location between ND6 and Cyt B (Fig. 1). Given the high coverage and quality of this fragment, this unique pattern is unlikely to be an artifact of assembly. Similarly, in *P. fenestratus* we found a new arrangement in the LTPF tRNA cluster, the tRNA-T was not recovered, and the IQM gene cluster is modified such that tRNA-Q is absent and now occurs inside the control region.

Additional details on base composition and other features of the four mitogenomes presented here can be found in **Table S2.3**.

Discussion

Numerous studies have employed the baiting and iterative mapping using MIRA and MITObim (*e.g.* (Doyle *et al.* 2014, Grau *et al.* 2015). Most of these studies share a set of characteristics: species were sequenced one at a time; sequencing the mitochondrial genome was at least one of the main objectives; and the number and quality of the sequence reads were high. However, when libraries are multiplexed (*e.g.* several libraries of different species are sequenced simultaneously) and/or genomic DNA samples have been enriched for particular loci (see Jones & Good 2015), read number will decrease substantially, with possible negative effects on overall read quality. The methods described here can be used to assemble organellar genomes in this latter scenario.

Our results show that even a low number of reads can be enough to provide high coverage for most of the mitochondrial genome, allowing organellar genomes to be extracted and assembled as by-catch from any Illumina HiSeq machine run using total genomic

libraries, even when libraries are multiplexed. The strategy presented here might also be effective for other technologies, since MIRA and MITOBIM also accept Ion Torrent and 454 data as input. We note that MIRA's manual suggests that the program may not be suitable for data sets with more than 20–40 million reads and that in some cases it may be necessary to randomly sample reads from the original pool, but read number will be naturally reduced for multiplex libraries.

Mitogenomes assembled by mapping to a more closely related reference mitogenome (in this case, another anuran) were longer and required fewer iterations than those generated by mapping to a more distantly related mitogenome (a salamander), and the assemblies obtained by mapping to a complete mitogenome generated longer consensus sequences than by using barcode seeds. However, no other significant differences were observed in sequence order and composition when reference sequences were changed. It should be noted that the anuran reference mitogenome we employed is more closely related to our four test species than is the salamander reference mitogenome, but it is deeply nested within the family Bufonidae and is, therefore, not especially close to any of our test species (Frost *et al.* 2006). Consequently, we suggest choosing references based first on sequence length and second on phylogenetic proximity.

Mitochondrial DNA has historically been the molecule of choice to address problems in phylogenetics and population genetics. The availability of complete or partial mitogenomes from different species provides a unique model to understand mechanisms of genome evolution (Gissi *et al.* 2008). Several genome features, such as molecular evolutionary rates, gene content, gene order and secondary structure of RNAs, can be explored in a phylogenetic context, but the utility of these data sets is fully dependent on taxon sampling (Boore 1999, Gissi *et al.* 2008). By using an optimized in silico strategy to recover mitogenomes from NGS data, the available mitogenome data set can be efficiently increased and can enable comparative genomic analysis.

Frog mitogenomics has been a slow yet steadily growing field of research. At the time the work on this chapter was finished and submitted for publication (Machado *et al.* 2016a), there were 192 complete and 109 partial mitogenomes of different species of Amphibia (Gymnophiona, Caudata and Anura) available in NCBI's Organelle Genome Resources database (Wolfsberg *et al.* 2001), and only 83 complete and 56 near-complete (>14 000 bp) mitogenomes of anurans of 22 different families. This constitutes a very small proportion of the 7703 known species of amphibians and 6784 species of Anura (Frost 2017; accessed on

Sep. 21st, 2017). The five new mitogenomes presented here represented, by the time of publication, three families and five genera for which mitogenomes were unknown previously.

Among vertebrates, amphibian mitogenomes have the greatest variation in gene order. Gene rearrangements are present in all orders of amphibian. For example, San Mauro *et al.* (2006) found rearrangements in the WANCY tRNA cluster in the caecilian genus *Siphonops*, Mueller & Boore (2005) found rearrangements in ND6-tRNA-E and WANCY tRNA cluster in plethodontid salamanders, and Kurabayashi *et al.* (2008) reported high frequency of genomic reorganization in the mitochondria of members of the anuran family Mantellidae. The increasing number of mitogenomes available for this group contributed to overturning the accepted view that mitochondrial gene organization in vertebrates was stable (Boore 1999, Saccone *et al.* 1999, Gissi *et al.* 2008).

Even though the available anuran mitogenomes are a small sample of the diversity of frogs, numerous mitochondrial gene rearrangements have already been reported for frogs. Irisarri *et al.* (2012) found new arrangements for the ND5 gene and ND6-tRNA-E cluster in the neobatrachian frogs *Lechriodus melanopyga* (Limnodynastidae) and *Heleophryne regis* (Heleophrynidae) and also reported modifications in the tRNA clusters of neobatrachians. Zhang *et al.* (2013) and Xia *et al.* (2014) also found several different gene orders for Neobatrachia that are mainly associated with tRNA clusters LTPF, WANCY, and IQM and the occurrence of pseudogenes. Here, we sequenced five new genomes and found three different gene arrangements associated with tRNA clusters, one in *Hylodes* and two in *Pristimantis*.

The mitogenomes presented here should contribute to future phylogenetic analyses of Amphibia and help improve understanding of the evolution of mitochondrial gene order arrangement in this taxon. At this point, however, the taxonomic and phylogenetic significance of these rearrangements is unclear and requires comparison with additional mitogenomes of closely related frogs.

Conclusion

We have reported the first mitogenomic sequences for the anuran families Craugastoridae and Hylodidae and the genera *Hylodes*, *Hyloxalus*, *Pristimantis*, *Melanophryniscus* and *Rhinella*. The mitogenomes of *M. simplex* and *Rhinella sp.* C. are the first mitogenomes of Neotropical bufonids. *Melanoprhyniscus* is the sister group of all other bufonids (*e.g.* Peloso *et al.* 2012),

making the mitogenome of *M. simplex* especially important for studies of mitochondrial evolution in this large, nearly cosmopolitan family.

By employing the baiting and iterative mapping strategy tested herein, workers can assemble organelle genomes as by-catch for use in comparative studies. Our results demonstrate that even a low number of reads can be sufficient to assemble high-quality mitogenomes, making any Illumina HiSeq run using libraries prepared with total genomic DNA extractions a potential source of organelle assemblies.

List of supplementary material

- Protocol S2.1 – Protocol for quality control
- Protocol S2.1 – Protocol for sequence assembly
- Table S2.1 – Summary of FastQC statistics
- Table S2.2 – Detailed mitogenome annotations
- Table S2.3 – Base composition and other features of the mitogenomes

All supplementary material is available upon request via the email machadodj@usp.br. After this thesis is provided a DOI by USP's Digital Library, you will be able to search for these materials at DRYAD (https://datadryad.org). Additionally, a compressed file containing all the Supplementary Material of this Ph.D. dissertation can be downloaded from http://www.ib.usp.br/grant/anfibios/datasets/Machado2018.zip.

3. A NEW STRATEGY TO INFER SEQUENCE CIRCULARITY

Background

There are currently 7,763 species of Anura (Frost 2017; accessed on December 19, 2017), the vast majority of which have not yet had their mitochondrial genomes (mitogenomes) sequenced and studied. At the time this manuscript was written, GenBank (https://www.ncbi.nlm.nih.gov/genbank/) listed partial mitogenomes for 107 species from 71 genera and 35 families and complete mitogenomes for 238 species from 76 genera and 27 families. Increasing the diversity of studied frog mitogenomes not only would improve our understanding of mitogenome evolution and provide important information for studies ranging from phylogenetics and population genetics to genomic evolution (*e.g.* Mueller and Boore 2005, Bertrand *et al.* 2015, Peng *et al.* 2015), but also would reduce the lack of reference sequences that hampers the analysis of novel mitogenomes in terms of sequence assembly and circularity inference (*i.e.* validation of sequence completeness).

To overcome the challenges in assembling anuran mitogenomes when no closely related reference is provided, Machado *et al.* (2016) optimized a strategy to efficiently reconstruct high-quality mitogenomes directly from genomic reads using the baiting and iterative mapping approach proposed by Hahn *et al.* (2013). Machado *et al.* (2016) validated the efficiency of this strategy as a means of assembling organelle genomes as by-catch from short genomic sequence reads sequenced using high-throughput sequencing technology even when the total number of reads is low and the reference belongs to distantly related taxa (*i.e.* different species, family, or even order). Both the strategy and the partial mitogenomes provided by Machado *et al.* (2016) have been successfully incorporated into the specialized literature (*e.g.* Anmarkrud & Lifjeld 2016, Vacher *et al.* 2016, Yuan *et al.* 2016).

Here we employ the same procedures outlined by Machado *et al.* (2016) to expand the diversity of sequenced anuran mitogenomes. Further, given that the majority of mitogenomes are circular, we propose a strategy to assess mitogenome completeness by testing the circularization of the assembled mitogenome. This procedure can also be applied to other circular genomes [*e.g.* chloroplasts, plasmids, covalently closed circular DNA (cccDNA) from

viruses, and circular bacterial chromosomes], although applications of our strategy to non-mitochondrial sequences will be discussed in detail elsewhere.

Material and methods

Whole genomic DNA sequencing

In order to increase the diversity of complete mitogenomes from undersampled clades, we selected four species of frogs from which to sequence mitogenomes. The mitogenome of *Scaphiopus holbrookii* (Harlan, 1835) is the first of the family Scaphiopodidae and that of the dendrobatid poison frog *Phyllobates terribilis* Myers, Daly and Malkin, 1978 is the first for its genus. The complete mitogenomes of the bufonid *Melanophryniscus moreirae* (Miranda-Ribeiro, 1920) and the dendrobatid *Hyloxalus subpunctatus* (Cope, 1899) are the first of their genera, although partial mitogenomes that lacked portions of the control region (CR) were recently published for *M. simplex* (GenBank accession KT221611.1) and *H. yasuni* (GenBank acession KT221612.1) by Machado *et al.* (2016).

Species	GenBank	SRA	BioSample
Scaphiopus holbrookii (Scaphiopodidae)	KY962390	To be provided	SAMN07271246
Melanophryniscus moreirae (Bufonidae)	KY962391	To be provided	SAMN07271247
Hyloxalus subpunctatus (Dendrobatidae)	KY962392	To be provided	SAMN07271248
Phyllobates terribilis (Dendrobatidae)	KY962393	To be provided	SAMN07271249

Table 3 – Taxa analyzed in the present study with accession numbers to NCBI's data bases. SRA number were not yet available at the time this document was printed, but records will be locatable online with the BioSamples.

Whole genomic DNA samples were extracted from muscle and liver samples using the DNeasy Blood & Tissue kit (Qiagen). Libraries were prepared using TruSeq Nano DNA Library Prep kit and Nextera Mate Pair (Illumina) and sequenced on Illumina HiSeq 2000/2500 machines. Library preparation and DNA sequencing of *M. moreirae* and *S. holbrookii* was performed by Macrogen Inc., Korea. Library preparation and DNA sequencing of *H. subpunctatus* and *P. terribilis* was performed by the David H. Murdock Research Institute (DHMRI). Details on mitogenomes, sequencing experiments, and specimen vouchers are reported in NCBI's GenBank, and BioSample databases (**Table 3**).

Quality control

Post-sequencing quality control was performed using the detailed guidelines provided by Machado *et al.* (2016) with some modifications. Specifically, adapter trimming for mate-pair sequences was performed using NxTrim v0.3.0-alpha (O'Connell *et al.* 2015) and all filtered reads were analyzed with FastUniq v1.1 (Xu *et al.* 2012) to remove putative PCR duplications. The overall quality of all sequence reads was evaluated before and after post-sequencing quality control using FastQC (Andrew 2010).

Mitogenome assembly

Mitogenomes were assembled using MIRA v4.0.2 (Chevreux *et al.* 1999) and MITObim v1.8 (Hahn *et al.* 2013) following the baiting and iterative strategy using reference genomes from different genera or families, as discussed by Machado *et al.* (2016). The complete mitogenome of *Pelodytes* cf. *punctatus* II-2011 (accession no. NC_020000.1; Pelodytidae) was used as reference for the assembly of the *S. holbrookii* mitogenome, *Bufo tibetanus* (accession no. NC_020048; Bufonidae) was used as reference for *M. moreirae*, and *Anomaloglossus baeobatrachus* (accession no. NC_030054; Aromobatidae) was used as reference for *H. subpunctatus* and *P. terribilis*.

Only sequences identified as paired-end reads after quality control were used for assembly. The interleaved paired-end sequence read file from *P. terribilis* was the largest (> 150 GB disk size with ~500 M reads). Assuming mtDNA reads have a random distribution of occurrence within sequenced libraries, analyzing only a fraction of the paired-end reads should provide adequate information to assemble the mitogenome. Therefore, we divided the reads of *P. terribilis* into three files of up to 52 GB disk size and ~170 M read pairs, ultimately reducing computational requirements and assembly run-time. We validated this strategy by comparing the three scaffolds.

Inference of circularity

A quick survey of the recent specialized literature (searching https://www.scopus.com and https://scholar.google.com.br for "complete mitochondrial genome" within the last three years) shows that many authors infer circularity through visual inspection of reads at the ends of the assembly (*e.g.* Gan *et al.* 2014a, Grau *et al.* 2015, Vacher *et al.* 2016) or using other

more convoluted method of visual inspection (*e.g.* Cong & Grishin 2016). Based on our own unpublished data and the material published herein, we have observed that MITObim assemblies can produce sequences flanked by erroneous sequences that seem to have resulted from spurious assembly of repetitive fragments. Unless the ends of the assembly overlap and there are reads that map to both ends, it can be difficult to visually detect circularity—even if the entire mitogenome was assembled correctly. A few programs, such as circules.py (distributed with MITObim), can find putative circular sequences based on k-mer overlap at a given minimum distance. However, these programs do not provide statistics to help the user judge the overall quality of the results or allow different assemblies to be compared. More elegant solutions are available that check circular assembly by homology searches using BLAST (Altschul *et al.* 1990) and comparing the size of the assembled genome to the reference (*e.g.* Soorni *et al.* 2017), but they are limited to specific pipelines, sequencing technology, and availability of closely related reference genomes.

We divided the problem of testing for assembly circularization into two parts: The first part of the problem is to find putative overlapping sequences and use the original sequence reads to validate the circularization. To track these putative overlapping sequences, we devised a strategy that searches for identical "words" (*i.e.* continuous text strings) at a minimum distance from each other (**Figure 7a**). Once a putative mtDNA sequence is found, it is flipped and rewritten so the ends are adjacent to each other in the middle of the sequence (**Figure 7b**). In the second part of the problem, we use Bowtie2 (Langmead *et al.* 2009) to map the original paired-end reads to the flipped fragment. Next, we acquire quality metrics from the assembly. These metrics include sequence similarity, coverage, and average alignment score. We also calculated a modified per-position sequence coverage value (which we named "connectivity") in which sequence reads that start or end at a position are excluded from the coverage calculation of that position. This allows us to quantify the number of reads that support the position of a particular nucleotide in relation to its two adjacent nucleotides (*e.g.* in the sequence fragment "ACT", the connectivity of "C" ignored reads starting in or ending in "C", and considers only reads that align to the entire fragment "ACT"). If connectivity is above a minimum threshold, we conclude that the sequence is contiguous (**Figure 7c**).

Bowtie2 can align short reads quickly and efficiently and the remaining operations can be executed in linear time, making the entire process feasible using standard personal

computers. This allowed us to test the sensitivity of our strategy to multiple k-mer and mtDNA sizes.

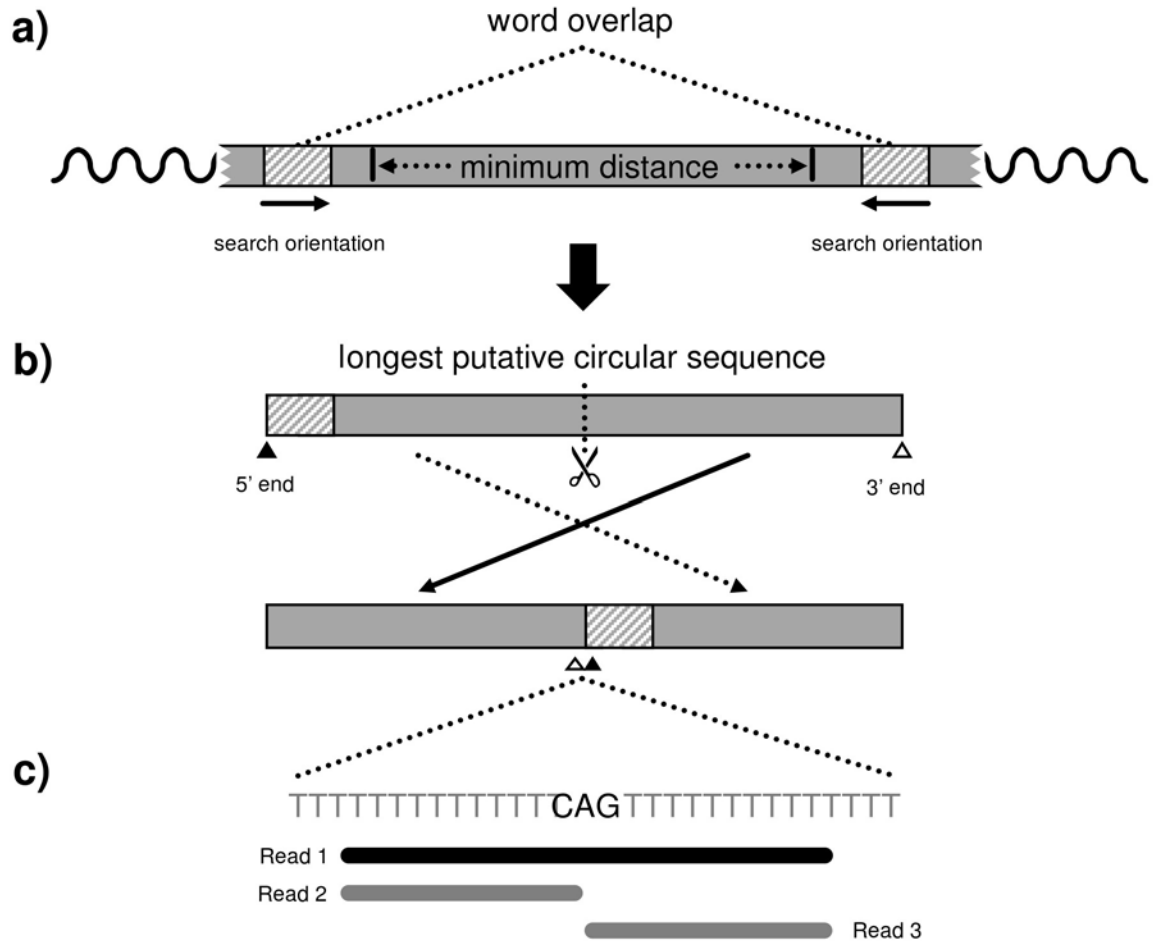


Figure 7 – Main steps of our strategy to infer circularity. **a)** We search for words of a specified length, from the end to the middle of the scaffold, with the condition that they are at a minimum distance from each other. **b)** The longest putative circular sequence found for each word size is flipped so the 5 prime and 3 prime ends will be adjacent to each other in the middle of the fragment. **c)** Original sequence reads are re-mapped against the flipped putative circular sequence. All the mapped reads (represented by reads 1–3) contribute to the average alignment score. For each nucleotide, only the reads that support its position in relation to the two adjacent nucleotides (represented by read 1) are counted to determine the contiguity coverage.

To validate our strategy, we randomly added or deleted nucleotides in 50 bp fragments at both ends of the proposed circular sequence. Random deletions and additions with 1 and 5% chance were performed, iterating 100 times per operation (deletion or addition of nucleotides) and probability of modification (total of 400 iterations per species). We then flipped sequences to make their ends adjacent to each other at the middle of the sequence. Finally, we used the flipped and modified sequences to compare the results of each iteration

based on the same quality metrics described above and ranked them by the distance to the flipped sequence with no modifications (modification probability of 0%) in terms of similarity, connectivity, and average alignment scores.

Mitogenome annotation

We parsed the DNA in CAF format using a Python script (parseCaf.py described in Machado *et al.* 2016) to extract DNA data and evaluate the coverage and quality of each mtDNA element. Our preliminary *de novo* mitogenome annotations were performed using the mitochondrial genome annotation server MITOS (Bernt *et al.* 2013) with default parameters. Additional search and validation of tRNA sequences were performed using ARWEN (Laslett & Canbäck 2008) and tRNAscan-SE (Lowe & Eddy 1997, Schattner *et al.* 2005).

Automated annotation was confirmed and edited manually by comparison to published anuran mitogenomes of closely related taxa. The CR, which typically lies between cytochrome b (mt-cyb) and the LTPF tRNA cluster in neobatrachians (mt-tl1, mt-tt, mt-tp, and mt-tf) (Zhang *et al.* 2013), was annotated using sequence similarity searching with BLAST using default parameters (Altschul *et al.* 1990).

We compared our complete mitogenome sequences from *Melanophryniscus moreirae* and *Hyloxalus subpunctatus* with partial mitochondrion genome sequences available for *M. simplex* and *H. yasuni* (GenBank accession no. KT221611 and KT221612), respectively, using the progressiveMauve whole genome alignment algorithm (Darling *et al.* 2010) available in Geneious version 8.1.9 (Kearse *et al.* 2012) as an additional verification step for the annotations and gene arrangement of these sequences, considering that gene order is not suspected to vary at this level of divergence.

Computational resources

Assemblies were executed on the high-performance computing clusters ACE and Steelhead. ACE is composed of 12 quad-socket AMD Opteron 6376 16-core 2.3-GHz CPU, 16MB cache, 6.4 GT/s compute nodes (= 768 cores total), eight with 128 GB RAM DDR3 1600 MHz (16 x 8GB), two with 256 GB (16 x 16GB), and two with 512 GB (32 x 16GB), and QDR 4x InfiniBand (32 GB/s) networking, and is housed at the Museum of Zoology of the University of São Paulo (MZUSP). Steelhead comprises five high-memory machines (Dell R815 - 64 AMD cores per node, 512–768 GB RAM each) and a separate computer

cluster with 25 nodes, each with 16 CPUs and is housed at the University of North Carolina at Charlotte. Inferences of circularity and sequence annotation were performed on a MacBook Pro (Retina, Mid 2012), 2.6 GHz Intel Core i7, 16 GB 1600 MHz DDR3. Using these resources, assembling each mitogenome in parallel, followed by annotation in sequential fashion, was performed in 1–2 days of computer and user time.

Results

Software

The AWA (the Tupi word for 'round') package comprises all the Python programs used for inferring circularity. Specifically, awa-trim is used to find putative circular sequences and awa-map is used to validate the circularization and provide basic statistics of the quality of the assembly. AWA is available at http://www.ib.usp.br/grant/anfibios and GitLab (https://gitlab.com/MachadoDJ/awa) under the GNU General Public License version 3.0 (GPL-3.0). A Wiki page with detailed user instructions and examples is available at https://gitlab.com/MachadoDJ/awa/wikis/home.

Inference of circularity

The four mitogenome assemblies passed the circularization tests with average alignment scores of -2.89 to -0.29 (the Bowtie2 alignment score is ≤ 0 in end-to- end mode and the quality of the alignment is directly proportional to the alignment score). With 5% chance of adding or deleting a nucleotide at the ends of the sequence, no permutation passed the circularization test. Likewise, with 1% chance of adding a random nucleotide, no sequence was considered circular. False positives for circularity only occurred under a 1% chance of deletion and were limited to 4% of the permutations with alignment scores 1.95–14.93 times worse than the observed scores, so we expect false positives to be easy to detect. In case different putative circular sequences are obtained with different word sizes, we suggest using the contiguity coverage and alignment scores to choose the optimal circularization. For additional details on these experiments see supplemental material (**Table S3.1**).

Mitogenomic sequences and gene rearrangements

The mitogenomes of *H. subpunctatus*, *M. moreirae*, *P. terribilis*, and *S. holbrookii* have 16,751, 18,005, 17,702, and 16,881 bp, respectively. The final average coverage reported by MITObim is, respectively, 871.75, 196.58, 2277.35, and 1326.69X. These mitogenomes have gene contents similar to those of other vertebrates, including 13 protein-coding genes, 22 transfer RNA (tRNA) genes, 2 ribosomal RNA (rRNA) genes, and 1 CR. Base composition of each mitogenome is shown in **Table 4**. As in other vertebrates, most mitochondrial genes are encoded on the heavy strand, except for eight tRNA genes (mt-tp, mt-tq, mt-ta, mt-tn, mt-tc, mt-ty, mt-ts2, and mt-te) and mt-nd6 (NAD6).

	Base pairs	Overall base composition				
Species		*A*	*C*	*G*	*T*	*GC*
Scaphiopus holbrookii (Scaphiopodidae)	16,881	32.40%	24.90%	20.10%	22.60%	44.90%
Melanophryniscus moreirae (Bufonidae)	18,005	30.30%	24.20%	14.20%	31.30%	38.40%
Phyllobates terribilis (Dendrobatidae)	17,702	28.30%	26.00%	14.60%	31.00%	40.60%
Hyloxalus subpunctatus (Dendrobatidae)	16,751	26.90%	27.60%	14.90%	30.60%	42.40%

Table 4 – Number of base pairs and nucleotide composition of the new mitogenomes.

Gene order in the mitochondria of *H. subpunctuatus*, *M. moreirae*, and *P. terribilis* is identical to that of other mitogenomes of Bufonidae and Dendrobatidae. The mitogenome of *S. holbrookii* is the first of the family Scaphiopodidae and matches the reference sequences available for the closely related families Pelobatidae (accession no. NC_008144) and Pelodytidae (accession no. NC_020000). Differences in gene order between *S. holbrookii* and the other three genomes are: 1) the mt-rnr1 (12S RNA) gene is preceded by mt-tf in *S. holbrookii* and by mt-tl1 + mt-tt + mt-tp + mt-tf in the other three mitogenomes; 2) the mt-nd5 (NAD5) gene is preceded by mt-th + mt-ts1 + mt-tl1 in *S. holbrookii* and by mt-th + mt-ts1 only in the other genomes; 3) The CR begins immediately after the cyb (cytochrome b) gene in all but *S. holbrookii*, which has mt-cyb and CR flanking mt-tt + mt-tp. See **Figure 8**.

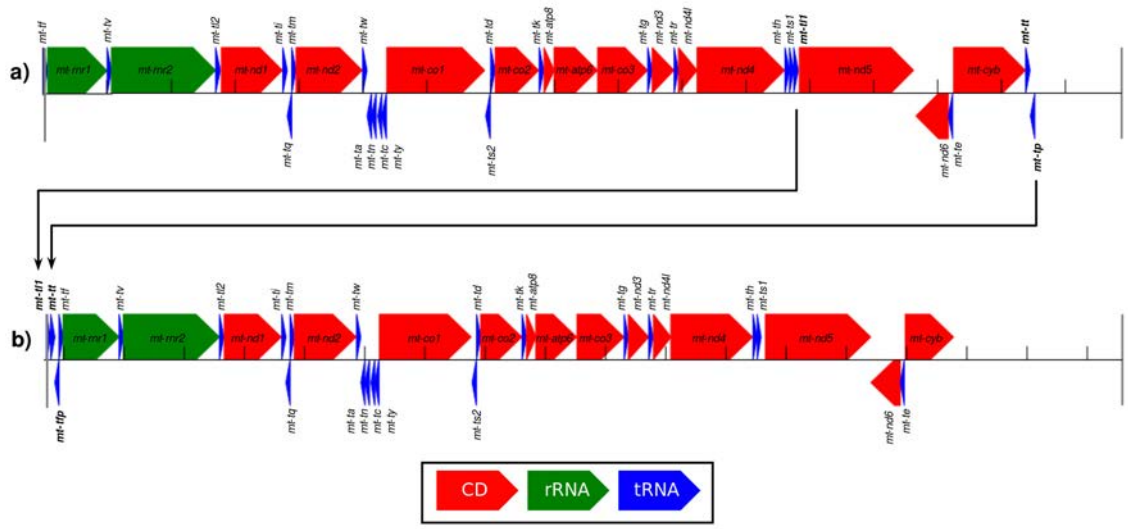


Figure 8 – Genome arrangement in the mitochondrial genome of a) *Scaphiopus holbrookii* and b) *Hyloxalus subpunctatus*, *Melanoprhyniscus moreirae*, and *Phyllobates terribilis*.

Whole genome alignments between the complete mitogenome *H. subpunctatus* and the partial mitogenome of *H. yasuni* revealed identical gene order and overall similarity of 78.2%, with 12,597 identical sites. The similarity between the complete mitogenome of *M. moreirae* and the partial mitogenome of *M. simplex* was higher, 82.4% with 14,017 identical sites, and the gene arrangements were also identical. Most differences between these sequences were concentrated in their CR, as expected. The CR of our complete mitogenomes are 1,374 and 2,599 bp long for *H. subpunctatus* and *M. moreirae*, respectively. The partial CR of *H. yasuni* is 663 bp and the partial CR of *M. simplex* is 515 bp long.

Discussion

The four new mitogenomes presented here represent the first complete mtDNA sequences for each of the four genera, *Hyloxalus*, *Melanophryniscus*, *Phyllobates*, and *Scaphiopus*. They also represent the first complete mitogenome of Scaphiopodidae and of important clades inside Bufonidae and Dendrobatidae.

As expected based on the phylogenetic relationships and prior information on mitochondrial diversification in anurans (*e.g.* Irisarri *et al.* 2012), the mitogenome of *S. holbrookii* respects the vertebrate consensus mitochondrial gene order, while the other

mitogenomes agree with what has been proposed as a modification in gene order (mt-th, mt-ts1, mt-nd5, mt-nd6, mt-te, mt-cyb, CR, mt-tl1, mt-tt, mt-tp, mt-tf) in the Neobatrachia lineage (Sumida *et al.* 2001; also see discussion in Xia *et al.* 2014). These findings suggest that the methods applied here produce reliable results.

The test of genome completeness followed our new approached based on word search and read mapping with Bowtie2. This allowed us to infer that sequences were contiguous based on overlapping words on the scaffolds as well as high quality reads mapped against the putative mitogenome with an average alignment score lower than -2.9. This automated approach to infer sequence circularity is further supported by permutation tests, which found only 2% false positives in all iterations. Since false positives had an overall alignment score 1.95–14.93 times worse than the best scores, authors should be able to use poor alignment scores (-3 or lower) as indications that the sequence should be reviewed and curated manually.

Conclusion

In this study, we present the first complete mitogenome of the family Scaphiopodidae and the genera *Hyloxalus*, *Melanophryniscus*, and *Phyllobates*. This increases in 1.68%, 5.26%, and 3.70% the number of anuran species, genera, and families, respectively, for which complete mitogenome sequences are known. Our approach for testing the completeness of circular DNA assemblies (presented here as a Python package named AWA) is time efficient and not computationally intensive. The test for mitogenome completeness can be done within minutes on a standard personal computer even when the file is large (*i.e.* 50–100 GB), and it both enables reproducibility of the tests of completeness and minimize human error.

List of supplementary material

- Table S3.1 – Results of all permutation tests

All supplementary material is available upon request via the email machadodj@usp.br. After this thesis is provided a DOI by USP's Digital Library, you will be able to search for these materials at DRYAD (https://datadryad.org). Additionally, a compressed file containing all the Supplementary Material of this Ph.D. dissertation can be downloaded from http://www.ib.usp.br/grant/anfibios/datasets/Machado2018.zip.

4. DRAFT NUCLEAR GENOME ASSEMBLIES OF THE GOLDEN POISON FROG AND THE EASTERN SPADEFOOT TOAD WITH EMPHASIS ON NEW PHYLOGENETIC MARKERS

Background

Because all biology is connected, new information from model organisms such as mice (*Mus musculus*), nematode worms (*Caenorhabditis elegans*), fruit flies (*Drosophila* spp.), zebrafish (*Danio regio*), chicken (*Gallus gallus*), and frogs (*Xenopus laevis* and *X. tropicalis*), among others, are often the best tools to discover the molecular mechanisms fundamental to life, thereby providing a shortcut to understanding different aspects of the human biology (Wheeler & Brändli 2009). However, we have to move beyond arbitrarily chosen model organisms to detect generalities in trait relationships between environment and evolvability (McGuigan and Sgrò 2009, Wolkovich *et al.* 2014) and close the gaps in our understanding of life on the planet (Richards 2015).

In the long-term, the increase in genomic information from diverse branches of the tree of life promises to increase our understanding of ecosystems by enhancing our ability to predict the metabolic capacity of trophic levels within food chains and biomes and determines the rates of material transfer between them (Falkowski *et al.* 2008). It might also provide new insights on the molecular warfare between attacking species and defending immune systems, and help to unlock the pharmacology of natural environments, furthering bio-prospection with critical medical applications (*e.g.*, Vonk *et al.* 2013, Sanggaard *et al.* 2014). In the medium-term, novel amphibian genomes can be used to study genome structure and function. Finally, in the short-term, increase genomic data from non-model organisms provides additional markers for phylogenetic systematics, allowing broad taxonomic surveys of essential clades of the tree of life and the addressing of mattering biological questions.

Amphibians are one of the most important and exciting branches of the animal tree of life for which genomic data is lacking. Instead of the mechanical defenses of amniotes, amphibians have evolved a vast array of defensive chemicals as protection against infectious diseases, parasites, and predators. They also developed physiological and morphological adaptations alongside the richest reproductive diversity of any tetrapod group, which allow them to thrive in heterogeneous terrestrial environments and is partially responsible for their recognition as evolutionary innovators. Nevertheless, only 4 of the 7,763 species of

amphibian (see Frost 2017; accessed on December 19, 2017) had their nuclear genome analyzed.

The first complete genome of a frog was from the western (or Congolese) clawed frog, *Xenopus tropicalis* Fischberg, Colombelli, and Picard, 1982 (Anura: Pipidae), with 1.7 billion base pairs in 10 chromosomes containing 21,378 protein-coding genes and 2,600 RNA genes (Hellsten *et al.* 2010). The western clawed frog is an important model for vertebrate development that combines experimental advantages of the African clawed frog, *Xenopus laevis* (Daudin, 1802), with more tractable genetics (*i.e.*, a diploid genome with relatively small size). The most striking features of this genome included its remarkable shared synteny with human and chicken over main parts of large chromosomes, broken by lineage-specific chromosome fusions and fissions, mainly in the mammalian lineage. Also, more than one-third of its nuclear genome consists of transposable elements which, curiously, are composed of a majority of DNA transposons.

The second was the whole-genome sequence of the Tibetan frog, *Nanorana parkeri* (Stejneger, 1927) (Anura: Dicroglossidae), with a genome size of 2.3 Gb containing 18,958 protein coding genes and 1,262 RNA genes (Sun *et al.* 2015). Most of the difference between the genome sizes of *X. tropicalis* and *N. parkeri* is due to transposable elements, emphasizing the importance of repetitive DNA to the amphibian genome size (see **Chapter 5**). With this genomic data, scientists were able to observe considerable conserved whole-genome synteny among anuran genomes that diverge over 250 Ma, which indicates a slow rate of DNA structural evolution in frogs. Furthermore, multigenome synteny blocks show that amphibians have fewer interchromosomal rearrangements than mammals despite a comparable rate of intrachromosomal rearrangements.

Last year, the third complete anuran genome was published for the allotetraploid African clawed frog (Session *et al.* 2016). This genome of almost 3 billion base pairs (slightly smaller to the human genome size) has 31,644 protein-coding genes and 3,018 RNA genes. This publication was a significant step forward in the analysis of large non-diploid genomes, which are common among amphibians.

Finally, this year we saw the publication of the draft genome of the North American bullfrog, *Lithobates catesbeianus* (Shaw, 1802) (Anura: Ranidae), an invasive species in several countries (Liu & Li 2009), including Brazil (Both *et al.* 2011). The bullfrog genome has 5.8 Gb with predicted 22,000 protein-coding genes and 6,223 candidate long noncoding

RNAs (lncRNAs) and will serve as a representative Ranid genomic resource because it is consistently diploid and has the widest global distribution of any true frog.

Despite the recent advances in amphibian genomics, there are currently no genomic resources for 53 of the 56 families of Anura (but see comments on anuran mitogenomics in **Chapters 2** and **3**). The lack of genomic information for amphibians hampers the scientific understanding of these animals in many ways. Two examples are more pertinent to this work. First, the investigation of the genetic bases of toxicity (including mechanisms involved in sequestration, biosynthesizes, biotransformation, and resistance) would greatly benefit from consolidated genomic data for comparative studies. Second, new genomic data from amphibians are immediate sources of phylogenetic markers to investigate amphibian diversification at different scales.

It has already been demonstrated by different groups of researchers that highly-conserved or ultra-conserved anchor regions of animal genomes can be harvested from comparative genomic studies to design hundreds or thousands of probes for target enrichment. These methods allow phylogeneticists to leverage from technologies of high-throughput sequencing (HTS) of DNA to rapidly (~2 weeks) acquire phylogenetic data with low cost (~1% of the cost of traditional Sanger sequencing) (Lemmon *et al.* 2012, McCormack *et al.* 2012a, 2012b, Faircloth *et al.* 2015). Nevertheless, target enrichment of highly-conserved elements (HCEs) or ultra-conserved elements (UCEs) using HTS have a few downsides. The strategy requires large quantities of high-quality genomic DNA. Also, the enriched DNA fragments are less conserved towards the 5' and 3' ends. Although this provides additional phylogenetic information at different levels of divergence, it also poses challenges on the process of aligning and trimming the sequences.

Given the lack of reference genomes and specific difficulties involved in working with some groups of animals, researchers focused on particular clades of non-model organisms are often forced to devise new strategies for their specific needs. For instance, Boyd *et al.* (2017) introduced a new approach (named “target restricted assembly”) to sequence and assemble data from very small tissue samples of parasitic lice (Phthiraptera: Philopteridae: *Columbicola*) preserved on ethanol, from which the extraction of large quantities of DNA would be challenging. The technique relies on assembling a selection of hundreds or possibly thousands of orthologous gene sequences from whole-genome sequence data representing dozens of species or more. This strategy provided trees that were much more robust than the ones obtained using only a few genes. However, the know genome sizes of the order

Phthiraptera are small (C-value of 0.11 pg; see Johnston *et al.* 2007) and therefore required a relatively modest sequence depth. For large genome sizes, however, restricted target assembly would be much less efficient and require much more sequencing effort to acquire a reasonable coverage.

In amphibians, the genome size is often much larger than the human genome size, with an abundance of repetitive DNA (see **Chapter 5**). Additionally, tissue samples from which to extract DNA can be limited by animal size (*e.g.*, Biju *et al.* 2007, Lehr & Catenazzi 2009, Das & Haas 2010), the difficulty to collect samples from specific areas (*e.g.*, restricted access, difficulty in acquiring permits, field trip costs), or even due to accelerate extinction rate of some amphibian populations (Mccallum 2007, Alroy 2015). In combination with the high value of museum specimens for systematics and biogeography investigations (Burrell *et al.* 2015, Hykin *et al.* 2015, Besnard *et al.* 2016), these factors create a demand for specific methods to integrate HTS to amphibian phylogenetics using the smallest tissue samples possible (*i.e.*, using little amounts input DNA).

To date, herpetologists rely on only a handful of gene regions to perform phylogenetic studies of amphibians (*e.g.*, eight markers in Frost *et al.* 2006, 11 markers in Blotto *et al.* 2012, 22 markers in Padial *et al.* 2014, 15 markers in Grant *et al.* 2017). And even though 512 HCEs probes (Lemmon *et al.* 2012) and 5,472 UCEs probes (Faircloth *et al.* 2012) have been designed for vertebrates and tetrapods, respectively, only a single amphibian genome (*Xenopus tropicalis*) was used in each case.

Here, we leverage on the preliminary assemblies and draft annotations of the genomes of the eastern spadefood toad (*Scaphiopus holbrookii*) and the golden poison frog (*Phyllobates terribilis*), as well as on the published complete genomes of three frogs (*X. tropicalis*, *X. laevis*, and *Nanorana parkeri*) and one lizard (*Anolis carolinensis*, used as outgroup) to designed UCE probes specifically for anurans. However, since UCE markers depend on large amounts of well preserved DNA and their use depends on specific strategies of DNA amplification, we also explored our data to unveil new intra and inter-exonic markers for the phylogenetic analysis of frogs. These new markers are presented in the form of non-degenerated and ultra-conserved primer pairs, which can be amplified using a variety of methods, possibly with more loosen DNA requirements.

Material and methods

The materials and methods described here are supplemented by detailed bioinformatic protocols (see **Supplementary material**).

Next-generation sequencing and quality control

Details on specimens and DNA sequencing data for *S. holbrookii* and *P. terribilis* are available in **Appendix S4.1** and **S4.2**., respectivelly The protocols for DNA extraction, library preparation, and quality control of raw sequence data for *S. holbrookii* and *P. terribilis* and have been described in **Chapter 3**. Details on the assembly and analysis of repetitive DNA is provided on **Chapter 5**.

Computational resources

Originally we were expecting to execute most computational analysis using "ACE", a FAPESP-funded SGI cluster housed in the Museu de Zoologia da Universidade de São Paulo (MZUSP) that entered production in October 2013. ACE uses the SUSE Linux Enterprise operational system and is composed of 12 quad-socket AMD Opteron 6376 16-core 2.3-GHz CPU, 16MB cache, 6.4 GT/s compute nodes (768 cores total), eight with 128 GB RAM DDR3 1600 MHz (16 x 8GB), two with 256 GB (16 x 16GB), and two with 512 GB (32 x 16GB), and QDR 4x InfiniBand (32 GB/s) networking (FAPESP Proc. No. 2012/10000-5). High memory vnodes were supposed to be available during the duration of this project. Additional local computational resources included a new high-memory server, "Heket", with 256GB RAM and additional 120GB cache have been installed at ACE's rack at MZUSP and will be entirely available for this project. However, due to electrical problems and limited internet bandwidth, data transmission and computer usage were severely limited according to directives from MZUSP. Fortunately, we also had access to facilities, equipment and other resources from Professor Daniel Janies' laboratory in the Bioinformatics and Genomics Department (BIG) at the University of North Carolina (UNC) at Charlotte. At The Bioinformatics Services Division in Kannapolis, NC, I was granted limited access to a 544-core Linux cluster based on hex-core processors with 4.5 TB RAM. This also included a 276 TB Lustre file system and a 65 TB Compellent Storage Solution and access to the staff who are consulted on the project. At the University Research Computing (URC) group at UNC

Charlotte, I had limited access to high-performance computing clusters and services to support the research mission of the University. This included 5 high memory machines (Dell R815 - 64 AMD core per node, 512-768 GB RAM each) and the Steelhead computer cluster (25 nodes, each with 16 CPUs).

Raw sequencing reads and project data acquired before 2016 are stored at TOT, a FAPESP-funded storage server with a current capacity of approximately 20 TB housed at the Institute of Biosciences of the University of São Paulo (IBUSP). All additional data generated during the execution of this project is stored at UNCC's servers and are being currently being transferred to our servers in Brazil. Data used in the manuscripts resulting from my doctoral dissertation project will be available upon publication.

Assembly of the genome of the eastern spadefoot toad and the golden poison frog

The *S. holbrookii* genome has 26 chromosomes and a C-value varying from 1.34 to 1.41 depending on the author (Goin *et al.* 1968, Sexsmith 1968, Olmo 1973), which is relatively small for amphibians. Poison frogs like *P. terribilis*, on the other hand, have much larger genome sizes. According to the Animal Genome Size Database (http://www.genomesize.com), the genome size of anurans of the family Dendrobatidae shows a great variance, with C-Values ranging from 2.98 (*Mannophryne trinitatis*) to 8.95 (*Dendrobates tinctorius*). Therefore, different strategies were selected to assemble the genome of the eastern spadefoot toad and the golden poison frog, meeting the specifics challenges posed by each project.

Our original intention was to compare the efficiency of different assembly strategies, including both reference-based methods and de novo assembly methods such as SOAPdenovo (Xie *et al.* 2014), Velvet (Zerbino & Birney 2008) and ABySS (Simpson *et al.* 2009). We spent significant amounts of time compiling and executing different assemblers but after a series of failed runs and according to suggestions from programmers, professors and Ph.D. staff members from both BiG and the Bioinformatics Services Division in Kannapolis (UNC, USA), we decided to that computer resources and program specifications made ABySS a suitable assembler for the assembly of the genome of *S. holbrookii*, which has a relatively small genome size for amphibians.

The ABySS program allows the computation of the assembly of multiple libraries in a parallel environment. The most important parameter is the k-mer length (*k*), that must be specified manually every time the program is executed. Testing all possible values for all

parameters is unfeasible for our data. Therefore, we limited other parameter variations to the minimum overlap between unitigs (*m*) and the minimum number of pairs required to join two contigs (*n*). We tested 33 values of k (21 to 85, with increments of 2), three values of m (20, 30, and 50), and 6 values of n (4, 7, 10, 20, 30, and 40).

Selection of the best assemblies took several assembly variables into account. Zerbino & Birney (2008) recommended choosing the assembly that produces the highest N50 (*i.e.*, shortest sequence length at 50% of the genome). The N50 scaffold length was designed to measure the contiguity of an assembly and it became a very popular metric to gauge the quality of a genome assembly. Continued reliance on this measure has attracted criticism (*e.g.* Narzisi & Mishra 2011) and others have proposed alternative metrics such as "normalized N50" (Mäkinen *et al.* 2012) to address some of the criticisms. Following the recommendations on the first Assemblathon papers (Earl *et al.* 2013, Bradnam *et al.* 2013), I decided to evaluate assemblies not only based on the N50 metric, but also on total sequence size, the number of sequences, and minimum and maximum sequence length, respectively.

The draf nuclear genome of *P. terribilis* was assemble at UNC in partnership with Dr. Robert Reid using the MaSuRCA genome assembler (Zimin *et al.* 2013) following the gerenal guidelines applied to the 22-Gb loblolly pine genome (Zimin *et al.* 2014). Several of MaSuRCA's its innovations were developed o handle the demands of very large genomic projects. The key idea in MaSuRCA is to reduce high-coverage paired-end reads to a much smaller and more concise set of "super-reads." Applied to our data, the MaSuRCA assembler can be conceptually divided into the same three phases described in Zimin *et al.* (2014), including: 1) corrects errors in the Illumina reads, using the QuORUM error corrector (Marçais *et al.* 2015); 2) educes the short and highly redundant Illumina paired-end reads to a concise set of super-reads; and 3) assemble the super-reads together with filtered read pairs from the longer diploid libraries.

Annotation

The complete protocol for sequence annotation of the draft genomes of the eastern spadefoot toad and the golden poison frog is provided in **Protocol S4.1**. In summary, we used RepeatMasker (Tarailo-Graovac & Chen 2009) to mask repetitive DNA sequences based on the available database of *X. tropicalis*. Masked scaffolds were them submitted to annotation using different strategies, combining *ab initio* and homology-based annotation strategies.

We used the GeneMark program (Borodovsky & Lomsadze 2011) to perform *ab initio* annotation with self-training algorithm. The Blast suit was used to search for matches to genes from *X. tropicalis*, *X. laevis* (available at http://www.xenbase.org/other/static/ftpDatafiles.jsp, version 9.0), *Nanorana parkeri* (available at http://gigadb.org/dataset/100132, version 2015-02-11), *Anolis carolinensis* (available at http://www.ensembl.org/Anolis_carolinensis version 2.0) and *Homo sapiens* (downloaded from Uniprot/ SwissProt on September 7, 2017).

In the absence of RNA-Seq data for this species, we are utilizing an homology-based method to predict protein-coding genes in the draft genome based on Sun *et al.* (2015). First, we mapped protein sequences of *X. tropicalis*, *X. laevis*, *N. parkeri*, *A. carolinensi*s, and *Homo sapiens* to the draft assembly (scaffolds only) using TBlastN (Kent 2002). Second, we filtered the aligned sequences and query proteins, and passed them to GeneWise (Birney 2000, Birney & Clamp 2004) to obtain accurate spliced alignments. We also used Augustus with a specific training set composed of manually revised genes from *N. parkeri* (BioProject ID: PRJNA344660; 780 genes for training, 403 for testing). Annotations were parsed using a series of homemade Python programs described in Protocol S4.1 and available with the **Supplementary material** of this chapter.

Evaluating assembly completeness

For the current stage of our genome projects, which consists of draft assemblies and annotations which are still awaiting for additional sequencing data, we decided for a simple approach to evaluate the completeness levels using BUSCO v2 (Simão *et al.* 2015).

BUSCO provides quantitative measures for the assessment of genome assembly, gene set, and transcriptome completeness, based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from the Hierarchical Catalog of Orthologs (OrthoDB) v9.1 (Zdobnov *et al.* 2016). Genes that make up the BUSCO sets for each major lineage are selected from orthologous groups with genes present as single-copy orthologs in at least 90% of the species. While allowing for rare gene duplications or losses, this establishes an evolutionarily-informed expectation that these genes should be found as single-copy orthologs in any newly-sequenced genome. The evolutionary expectation means that if the BUSCOs cannot be identified in a genome assembly or annotated gene set, it is possible that the sequencing and/or assembly and/or annotation approaches have failed to capture the complete expected gene content.

WeI ran the selected scaffolds from ABySS through BUSCO using existing Augustus species gene finding parameters set to humans (closest available species) and the vertebrate lineage database from OrthoDB.

Identifying UCE loci

The **Protocol S4.2** describes the step-by-step process used to identify UCE loci and design probes to target them. The original software and revised pipeline required for replicating this protocol are available with the **Supplementary material** that accompanies this chapter.

The strategy we used for the identification of UCE loci is based on the phyluce package (Faircloth 2017) and was designed to provided the first UCE probe set specific to amphibians, taking into account their know repetitive DNA. Selected genomic data include scaffolds (multifasta files with repetitive DNA masked as lowecase nucleotides) from *Anolois carolinensis* (used as outogroup), *X. tropicalis*, *X. laevis*, *N. parkeri*, *S. holbrookii*, and *P. terribilis*. Analysis parameters are the same as described in phyluce's Tutorial IV (http://phyluce.readthedocs.io/en/latest/tutorial-four.html).

Identifying new intra- and inter-exon markers

We took the draft annotated genomes of *S. holbrookii* and *P. terribilis* to search for new photogenic markers inside exons, or in exon-intron-exon fragments. First, we applied different programs from the CD-HIT package (Li *et al.* 2001, Li & Godzik 2006) to cluster exons both within and between species to extract the most representative sequences that are more likely to represent fragments of homologous genes. Second, we used original Python programs and PRIMER3 (Untergasser *et al.* 2012) to find stable regions (100% identical in both genomes) flanking variable target regions (bellow 75% identity) of 600 base pairs or less. The step-by-step bioinformatic protocol os provided in **Protocol S4.3**, and original programs are available with the **Supplementary material** of this chapter.

Results

The draft genome of the eastern spadefoot toad

We completed the assembly of the draft genome of *S. holbrookii* using 1,667,663,446 Illumina sequence reads (101-151 bp) using a *de novo* strategy implemented in the software ABySS (Simpson *et al.*, 2009) and the computational resources made available to us at UNC Charlotte and MZUSP. The draft assembly has approx. 2.5 million sequences, totaling 856,300,000 bp (60.73% of the expected genome size) with N50 of 28,615. This result is expected due to the highly repetitive content of amphibian genomes. Qualitative analysis with BUSCO v2 (Simão *et al.*, 2015) indicate that the draft assembly has most of the expected genes for vertebrates. BUSCO analyses are based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from the Hierarchical Catalog of Orthologs (OrthoDB) v9.1 (Zdobnov *et al.*, 2016). From 2748 (69.5%) complete BUSCOs found, 2703 (68.4%) were single-copy and only 45 (1.1%) where duplicated. The search also found 738 (18.7%) fragmented BUSCOs, with only 464 (11.8%) missing.

The draft genome of the golden poison frog

The statistics of assembly of the draft genome of the golden poison frog using the MaSuRCA genome assembler (Zimin *et al.* 2013) and the computer clusters at UNC Charlotte are very satisfactory given the available data. The assembly has 442,103 sequences, with a total of 3,817,837,047 residues. Sequence lengths vary from 1,000 to 5,752,821 base pairs, with average length of 8635.63 base pairs. The N50 is 11,421.

BUSCO v2 (Simão *et al.*, 2015) indicate that the draft assembly has most of the expected genes for vertebrates, and is more complete that the draft assembly of *S. holbrookii*. From 3394 (85.4%) complete BUSCOs found, 3363 (84.6%) were single-copy and only 32 (0.8%) where duplicated. The search also found 183 (4.6%) fragmented BUSCOs, with only 397 (10%) missing.

Annotations

Preliminary annotations focused on protein-coding genes. Our analysis identified 10,114 genes in the draft genome of *S. holbrookii* and 12,214 genes in the draft genome of *P.*

terribilis, excluding numerous incomplete genes due to scaffold fragmentation. Since we are not reporting here the results of analyses performed by our collaborators, details on the annotations of these genomes will be left for a later publication. However, preliminary data is available upon direct request and further comments on the annotation of RNAs and repetitive DNA are given on **Chapter 5**.

UCE markers

We identified 47,925 ultra-conserved loci in the frog genomes and designed a total of 377,193 probes that are immediately available for phylogenetic studies of amphibians. These include ~95 K probes for *X. tropicalis*, ~11 K probes for *X. laevis*, ~92 K probes for *N. parkeri*, ~90 K probes for *S. holbrookii*, and ~ 88 K probes for *P. terribilis* (see **Supplementary material**). Previous studies using a variety of genomes from vertebrates, including a single amphibian (*X. tropicalis*) resulted in no more than ~1 K probes aligning efficiently to the frog genome. Aligning this probes to the draft assemblies of *P. terribilis* and *S. holbrookii* showed that, respectively, ~40.1 and 10.2% of them would capture regions of repetitive DNA, which is undesirable. The new probes are 300 times more numerous and do not capture regions of repetitive DNA in any of the genomes we analyzed.

Inter and intra-exon markers

We present an efficient bioinformatics pipeline that recovered 678 intra-exon markers with 300 to 600 base pairs (average 501.6 base pairs) that diverge 25 to 95% (average 75%) from each other among the genomes of *S. holbrookii* and *P. terribilis*. We further provide an efficient bioinformatics pipeline that recovered 216 intra-exon markers with 90 to 600 base pairs (average 150.3 base pairs) that diverge 45 to 82% (average 62%) from each other among the genomes of *S. holbrookii* and *P. terribilis*.

Discussion

The assembly of the complete genomes of eastern spadefoot toad (*Scaphiopus holbrookii*) and golden poison frog (*Phyllobates terribilis*) genomes are part of a broader effort of the Laboratório de Anfíbios (Instituto de Biociências, Universidade de São Paulo). These projects are under the supervision of Professor Taran Grant, with financial support from the Fundação

de Amparo à Pesquisa do Estado de São Paulo (FAPESP, Proc. No. 2012/10000-5). Our ongoing objective is to unveil the genetic basis of alkaloid sequestration and resistance in amphibians by conduction comparative genomic studies of poison frogs and their non-toxic relatives. Other whole-genome projects currently underway include the Maldonado redbelly toad (*Melanophryniscus moreirae*) and the Bogota rocket frog (*Hyloxalus subpunctatus*), from which draft assemblies were also completed by our collaborators and us. Due to their quality, current stage of assembly, and the contribution of our collaborators, we decided these works would be better presented elsewhere.

These genome projects await additional sequencing data to be completed. At that time, gene functions will be assigned according to the best match of their alignments using BlastP to the SwissProt and TrEMBL databases(Consortium 2014). Motifs and domains of genes will be determined by InterProScan (Zdobnov & Apweiler 2001) against protein databases including ProDom, PRINTS, Pfam, SMART, PANTHER, and PROSITE. Gene Ontology (GO) IDs for each gene will be obtained from the corresponding SwissProt and TrEMBL entries. All genes will be aligned against the KEGG proteins to detect possible gene pathways.

In the meanwhile, the genomic information at hand is an immediate source of information for phylogenetic analysis. The new phylogenetic markers proposed here can potentially unlock new information on the evolution of amphibians which has been limited due to the relatively small amounts of DNA information used in amphibian phylogenetics to date. Even considering the increase in the evidential basis of the amphibian systematics of several orders of magnitude since the 1990s (*e.g.*, Ford & Cannatella 1993, Frost *et al.* 2006a, Grant *et al.* 2017), the mass of molecular information at the disposal of herpetologists remains an absurdly small. This discrepancy is made more clear if we compared the customarily 1-10 gene regions and 1-10 kilobases of DNA sequence used in most amphibian phylogenetic studies and the 1.7 Gb and 20,000 protein-coding genes that comprise the genome of *X. tropicalis*. Within this context, it is clear that the many thousands of new UCE probes for anurans (377,193 probes for 47,925 conserved loci) are a significant step forward. Nevertheless, amplification of UCE markers require large amounts of high-quality DNA samples and pose challenges in sequence alignment since sequence ends that are very divergent from each other must be trimmed. Still, some of the limitations posed by UCE probes can be overcome with other types of markers, such as the intra- and inter-exonic markers (678 and 216 intra- and inter-exon markers, respectively), which require less input DNA and are easier to process given that the ends of the fragments are the most stable

regions. Also, these two types of markers can be amplified with a more extensive variety of methods, from multiplex PCR (Taly *et al.* 2012) to molecular inversion probes (Mamanova *et al.* 2010).

Finally, the revised and original bioinformatic pipelines we wrote can be used to increase the value of preliminary steps of genomic studies with non-model organisms as they allow draft genome assemblies to be used to find more phylogenetic markers with a minimum computational requirement (*e.g.*, the pipelines can be executed in 2 to 6 hours using a personal laptop with 2.6 GHz Intel Core i7, 16 GB 1600 MHz DDR3, and 750 GB of disk space).

Conclusion

The new phylogenetic markers (678 and 216 intra- and inter-exon markers, respectively) and molecular probes (377,193 probes for 47,925 conserved loci) are ready to be used in large-scale phylogenetic studies combined to high-throughput DNA sequencing technology. Furthermore, the revised and original bioinformatic pipelines give an additional value to the preliminary steps of genomic studies with non-model organisms as they allow researchers to acquire new phylogenetic markers as by-catch of draft genome assemblies.

List of supplementary material

- Appendix S4.1 – Specimen and DNA sequence information for *Scaphiopus holbrookii*
- Appendix S4.2 – Specimen and DNA sequence information for *Phyllobates terribilis*
- Protocol S4.1 – Modified phyluce's tutorial IV
- File S4.1 – UCE probes for Anura (377,193 probes in multifasta format)
- Original and modified programs used for Protocol S4.1 (available at GitLab, https://gitlab.com/MachadoDJ/Modified_phyluce_Tutorial_IV)
- Protocol S4.2 – Revision of the original tutorial of phyluce for identifying UCE loci and designing baits to target them
- Protocol S4.3 – New intra and inter–exon phylogenetic markers for Anura
- Files S4.2 and S4.3 – Primers for inter and intra-exon markers (multifasta file containing target sequence, forward, and reverse)
- File S4.4 and S4.5 – Tables of synonyms for *P. terribilis* and *S. holbrookii*

- Original programs used for Protocol S4.2 (available at GitLab, https://gitlab.com/MachadoDJ/findExonicMarkers)

All supplementary material, including scaffolds and annotations for the draft genome assembly of *P. terribilis* and *S. holbrookii*, is available upon request via the following email address: machadodj@usp.br. Once we publish these draft genomes, we will make sequences, annotation, and raw data available at the appropriated public databases. After this thesis is provided a DOI by USP's Digital Library, you will be able to search for these materials at DRYAD (https://datadryad.org). Additionally, a compressed file containing all the Supplementary Material of this Ph.D. dissertation can be downloaded from http://www.ib.usp.br/grant/anfibios/datasets/Machado2018.zip.

5. REPETITIVE DNA IN THE ANURAN GENOME

Background

The total amount of DNA in the nucleus has both biological and ecological consequences that affect the distribution and persistence of biodiversity. Genome evolution in eukaryotes has been driven by a number of processes, including the breakage and rejoining of different chromosomes (translocations), gene and segmental duplication, the shuffling of functional domains in exons, and gene conversion. However, our knowledge of most of the process leading to variation of the genome size is sketchy and awaits basic research that allows comparative genomic studies among diverse clades.

The most common methods to measure DNA amounts are flow cytometry, Feulgen absorbance cytophotometry (densitometry) and static cell fluorometry (for a review on cytochemestry methods to measure DNA, see Greilhuber 2008). These methods provide the a measure of DNA which is often represented as the "C-value" or “holoploid genome size”. The C-value can be defined as the DNA content of the whole chromosome complement (with chromosome number *n*) and is characteristic of the organism (irrespective of the degree of generative polyploidy, aneuploidies, etc.) (Greilhuber 2005). The C-value is often given in picograms of DNA (1 pg = 978 Mbp; Doležel *et al.* 2003). For diploid organisms, we can use the terms C-value and genome size interchangeably. For polyploid organisms, however, the C-value might include several haploid genomes. Since many amphibians are polyploid and we do know the karyotype of most of them, the next few paragraphs will focus on C-values, and the reader is advised to exert care while interpreting the relationship between genome size and C-value.

Observed variation of C-values in amphibians

The Animal Genome Database (Gregory 2017) is a comprehensive catalog of animal genome size data. Searching for all amphibians on the database returned 928 results on September 4, 2017. These include expected "C-values" of more than 450 species of amphibians belonging to three orders, 29 families, 136 genera (see **Table S5.1** and corresponding references in **Table S5.2**). **Figure 9** shows violin plots that summarize this

data. The plots allow us to see that Caudata is the order of amphibians with greater C-values, of which the smallest C-value is still bigger than the maximum expected C-value for Anura.

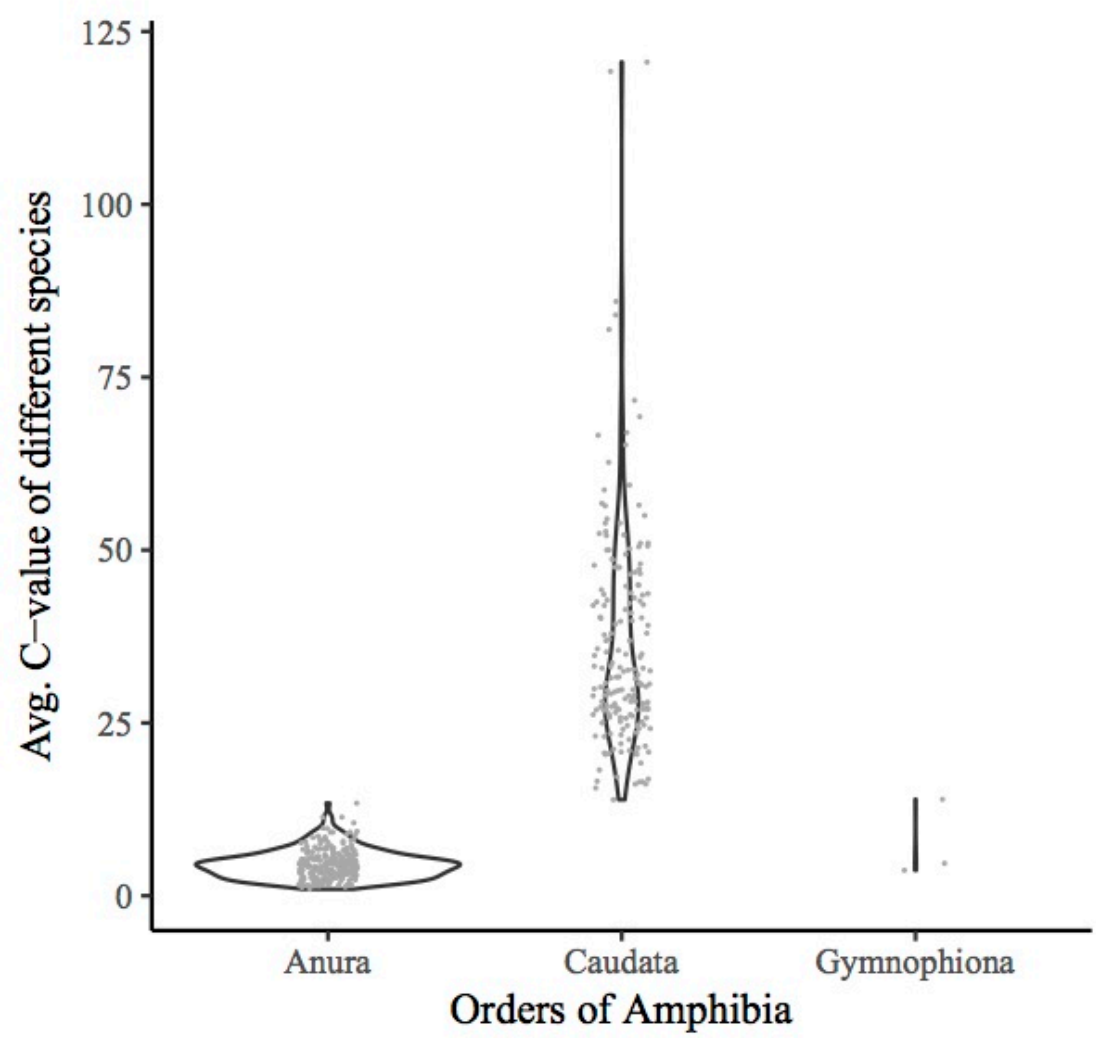


Figure 9 – C-values (pg) of different species of amphibians divided into its three orders: Anura, Caudata, and Gymnophiona. The violin plots represent the probability density at different values. Dots superimposed on each plot represent different species of the corresponding order. Source: http://genomesize.com. Access date: Sep. 4, 2017.

The records at the Animal Genome Database show only four studies proposing the C-values of three different species of Caeciliidae: *Gymnopis multiplicata*, 3.7 pg; *Geotrypetes seraphini*, 4.60 pg; and *Siphonops annulatus*, 13.95 pg. Although very few information is available regarding the expected C-values within Gymnophiona, the data is enough to draw two hypothesis which the evidence available for other amphibians will corroborate. The first hypothesis is that there can be high levels of variation in C-value within the same family of

amphibians, which in Caeciliidae can reach at least 377.03%. The second hypothesis is that the difference in chromosome number is not necessarily proportional to the variation in C-value. For the second hypothesis, we rely on the observation that the chromosome number in *Ge. seraphini* (38) is at least 146.15% bigger than the chromosome number in *Gy. multiplicata* (24-26), but the C-value of *Ge. seraphini* is only 124.22% heavier than *Gy. multiplicata*.

The C-values in Caudata ranges from 13.895 pg to 120.565 pg, with an average of 36.573 pg. The C-values in Anura ranges from 0.95 pg to 13.4 pg, with an average of 4.549 pg. Therefore, although the percentage of variation in C-value is greater in frogs than in newts (up to 1,410.53% Anura against 868.34% in Caudata), newts have a wider range of C-values which are always heavier than in frogs. If we focus on different families of amphibians instead of their orders (see **Figure 10**), we see that the same applies to any pair of frog and newt families.

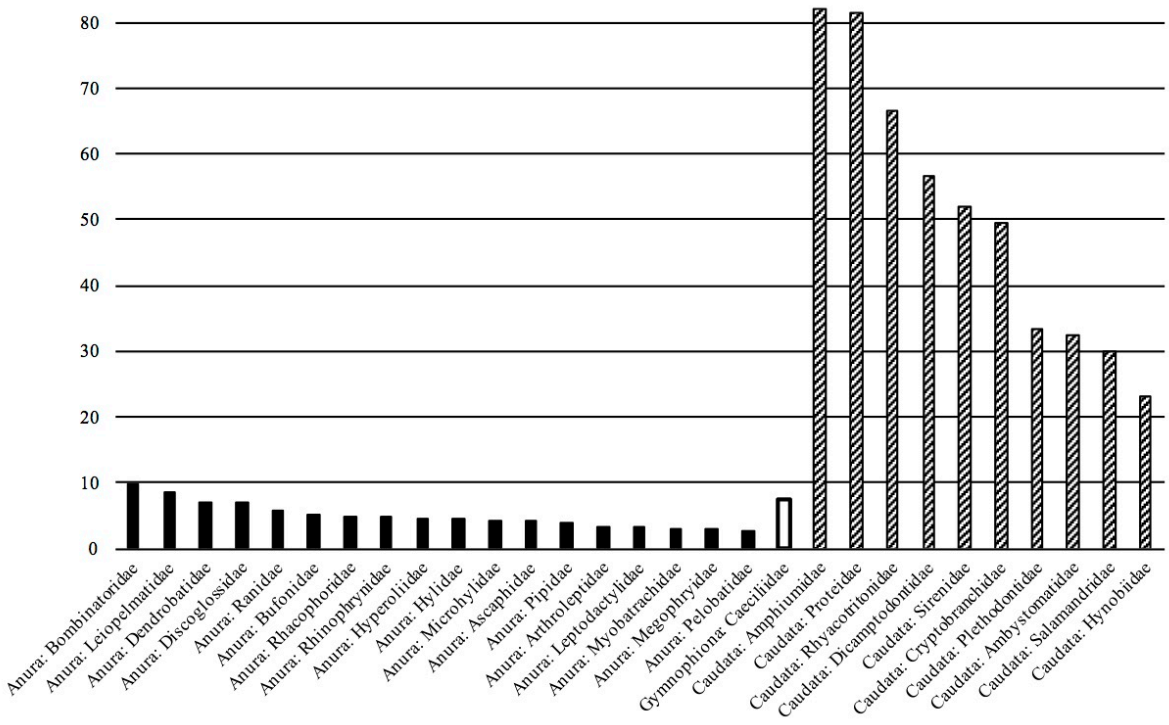


Figure 10 – Average C-value (pg) for 29 of the 75 families of amphibians, ordered by order and highest value. Source: http://genomesize.com. Access date: Sep. 4, 2017.

There is considerable variation in C-values within different families of Anura and Caudata, just as observed for Gymnophiona. The change in C-values within different families of Anura is as low as 120% in Hyperoliidae and as high as 921.05% in Myobatrachidae (the

largest C-value is 4.45 times bigger than the smallest, on average, for anuran families). The variation in C-values within different families of Caudata is as low as 118.33% in Cryptobranchidae and as high as 459.87% in Plethodontidae (the largest C-value is 2.31 times bigger than the smallest, on average, for anuran families).

Also as observed in Gymnophiona, changes in C-values do not accompany proportional variations in the chromosome number within Anura and Caudata. For example, the minimum and maximum C-value in newts of the family Ambystomatidae is 24 pg and 48 pg, respectively. However, every species of this family have the same chromosome number, 28. Moreover, the individuals of the family Amphiumidae, which also has a chromosome number of 28, possess more than two times the same amount of DNA as any species of Ambystomatidae.

The C-value paradox from the perspective of amphibian genomics

Thomas (1971) coined the term "C-value paradox" to refer to the emerging evidence at the time that studies of the amount of DNA in the cells of different species seemed to bear no relationship to their complexity as organisms. For example, take the Fugu pufferfish – a species prized as a delicacy in Japanese restaurants but so toxic that, if prepared incorrectly, it can rapidly result in depth (Buerk 2012). The genome of the Fugu pufferfish is unusually compact, with a haploid genome that contains c. 400 million bp. This is a much smaller amount of DNA than what the human haploid genome has. The female and male haploid genome size in humans is 3.203×10^9 and 3.147×10^9 bp, respectively. Given that the length of one nucleotide is estimated to be c. 0.34 nm, there is approximately 28 cm of DNA per somatic nucleus in Fugu and over 2 m in humans. However, more extraordinary than the tremendous difference in the size of the Fugu pufferfish and the human genome is that both encode roughly the same number of genes, which has significant implications for the discussion of how much of our genome is functional (*e.g.*, Buerk 2012, "The ENCODE Project Consortium" 2012, Brendan & Maher 2012, Doolittle 2013, Graur 2017).

In amphibians, the huge variation in genome size, which sometimes occur with no variation in chromosome number, forces us to look at the C-value paradox with a different perspective. For the amount of DNA in the cells of various species of amphibians vary tremendously despite their similar complexity as organisms. Unfortunately, the lack of basic research inside this group hampers our ability to ask questions about the extent of genome conservation and genome functionality within Gymnophiona, Caudata, and Anura. Still, *de*

novo assembly of repetitive DNA in the chromosomes of anurans of assorted clades and with different C-values allows us to investigate at least a few aspects of the mechanisms behind genome size variation in these animals.

Repeat elements in the chromosomal DNA

A repeat is defined here as the consensus of all copies of the same segment of DNA that appears multiple times in the genome in identical or near-identical form. There are many types of repeats (Cordaux & Batzer 2009), such as multicopy genes (*e.g.*, tRNAs, rRNAs, and snRNAs), integrated viruses (*e.g.*, DNA viruses, Caulimoviridae viruses), or simple repeats that arise via gene duplications (*i.e.*, microsatellites, < 10 bp; minisatellites, 10–100 bp; and satellites, > 100 bp). However, perhaps the most well-known type of repetitive element in the DNA are the transposable elements (TEs).

The TEs are endogenous mobile DNA elements that are ubiquitous amid eukaryotes. Alongside the ability to move across their host genome, some types of TEs can increase their copy number while active and therefore densely populate the chromosomes. Therefore, amplification of TEs and polyploidy (particularly in some plant lineages) are currently the two most recognized mechanisms for increasing genome sizes and generating evolutionary novelties in eukaryotes (Baidouri & Panaud 2013). For instance, the TEs constitute 25–40% of the mammalian genome (Cordaux & Batzer 2009) and 85% of the maize genome (Schnable *et al.* 2009). Thus, the scientific community has acknowledged TEs as "parasitic DNA" for a long time. Currently, however, numerous studies have established their sound biological impact on the structure, function, and evolution of eukaryotic genomes (Kobayashi 2004, Jones & Gellert 2004, Feschotte 2008, Hollister & Gaut 2009).

There are multiple classifications available for TEs. Wicker *et al.* (2007) divides them into two categories according to their move and amplification mechanisms. Class II elements or transposons move via a “cut–and–paste mechanism” and do not use RNA intermediates in their transposition process. These transposons are wholly DNA-based elements found in both prokaryotes and eukaryotes that can directly relocate autonomously via recombination. Inserted copies can excise themselves and insert into new loci, making use of their transposase gene. Consequently, copy numbers do not always increase, although the single excised DNA strand may employ another transposon sequence from the opposite

strand to repair itself. The insertion loci are therefore somewhat ephemeral. Also, class II elements appear quite adept at jumping between species (*i.e.*, frequently undergoes horizontal transfer between host genomes) (Schaack *et al.* 2010).

Class I elements or retrotransposons (RTs) replicate and relocate indirectly via an RNA intermediate using a "copy–and–paste mechanism," undergoing fewer horizontal transferences than TEs from class II. The name of the RTs comes from the RNA-mediated transposition (or retroposition) and denote the reverse flow of genetic information that occurs from RNA back to chromosomal DNA. Retroposition is made possible thanks to reverse transcriptase (RTase), an enzyme that generates complementary DNA (cDNA) from an RNA template during reverse transcription. The repeat element can or cannot encode RTase, which divides class I TEs into two subgroups: retroelements and retroposons.

Retroelements are retrotransposons that encode RTase and are thus said to be autonomous repeat elements regarding their transposition activity. Retroelements include Gypse and Copia type long terminal repeats (LTRs) retrotransposons, non-LTR retrotransposons or long interspersed elements (LINEs), and retroviruses, among other repeat elements.

Retroposons are non-autonomous retrotransposons that do not encode RTase. They include processed retropseudogenes and short interspersed elements (SINEs). The SINEs are often confused with LINEs, but they are quite different from each other. Besides the inability to code for RTase, SINEs differs from LINES due to their relatively short length (70–500 bp) and larger copy number (often over 100 total copies in the eukaryotic genome). Thus, SINEs greatly outnumber other repetitive elements and are of increasing interest to systematists because of their exceptional diagnostic power for establishing common ancestry among taxa and straightforward analysis once properly characterized (Shedlock & Okada 2000).

Different authors might propose slightly different definitions or classifications of repeats in the chromosomal DNA (*e.g.*, Kazazian 2004). This is a reflection of our incomplete but increasing understanding of the relationships among transposons, retroelements, and retroposons. For instance, retroviruses are thought to have evolved from LTR-retrotransposons by the acquisition of env genes, and LTR retrotransposons are in turn believed to have evolved from LINEs by the addition of long terminal repeats. LINEs encode for RTase and are typically moderately or severely truncated at the 5' end, which suggests that a RTase encoded by a LINE must recognize the 3' end of the RNA template for first strand synthesis. Most SINEs are derived from tRNA and are believed to recombine and interact

functionally with corresponding LINEs, leading to the acquisition of their retropositional activity (Kazazian 2004, Feschotte 2008, Chalopin *et al.* 2015).

Class I elements as "Hennigian" markers

It is clear that there is a balance between detrimental effects on the individual and long-term beneficial effects on a species provoked by mobile and repetitive elements in the chromosomal DNA. But the number of unanswered questions greatly surpass the amount of knowledge we have. Therefore, most scientists studying mobile elements are focused on fundamental questions about how they contribute to genome functionality and evolution: How are they associated with speciation events? Are they "junk DNA"? Did they ever have a function? To which extent they contribute to genome malleability?

There is, however, another interesting line of investigation concerning mobile elements which is interested in their usage as "Hennigian" markers. The term Hennigian marker comes from Willi Hennig, the German entomologist who founded the cladistic methodology that currently dominates phylogenetics. In cladistics, the shared occurrence of a derived (apomorphic) character-states (synapomorphies) is used to form monophyletic groups or clades. Within this context, A Hennigan marker is a term that betokens inheritable characters which are phylogenetically informative (*i.e.*, high level of synapomorphy) at the same time that they result in minimal amounts of error or incongruence during character codification (*i.e.*, low level of homoplasy; for a glossary of phylogenetic terms see https://gitlab.com/MachadoDJ/ybyra/wikis/Glossary).

Most molecular phylogenetic analyses of non-model organisms focus only on a few sequences of DNA and rely primarily on single nucleotide substitutions and insertions or deletions (indels). Although the value of these type of data is undeniable, it has many caveats which include convergent evolution of nucleotide bases, differing substitution rates among sites and lineages, saturation of mutations at variable sites, nonindependent substitutions among sites, and functional constraints at the molecular level, among others.

The advent of the genomic era has brought the opportunity to consider other types of information embedded in DNA sequences and find better Hennigian markers. In this context, rare genomic changes, such as intron indels, signature sequences, mitochondrial and chloroplast gene order changes, gene duplications, genetic code changes, and integration of class I repeat elements provide a suite of complementary markers with enormous potential for molecular systematics.

Class I elements are rarely pruned to horizontal transfer between host genomes (Schaack *et al.* 2010) and it is improbable that different species will have the same class I element in the same locus by chance alone. Also, secondary loss of these elements, especially among closely related species, is expected to be uncommon (Shedlock & Okada 2000). Therefore, class I elements have the potential to be suitable Hennigian markers as they represent hereditary characteristics which arise due to common evolutionary processes and are little prone extensive convergency, parallel evolution, or subsequent losses.

Since the 1990's, class I elements such as SINE insertions have been successfully implemented as a method to determine the order of divergence of relatively closely related species, demonstrating that these repetitive elements are phylogenetically informative and possess a low degree of homoplasy (Kido *et al.* 1991, Murata *et al.* 1993). Almost a decade later, Shedlock & Okada (2000) and Rokas & Holland (2000) advocated the use of retroposons such as SINEs and LINEs in the phylogenetic analyses of different orders of eukaryotes. The hugest drawback with using mobile elements as phylogenetic markers is that prior genomic information concerning retrotransposons is required for primer design, obstructing universal applications. Nevertheless, the specialized literature continues to show the utility of mobile elements as molecular markers as they are applied to disentangle phylogenetic conflicts in different branches of the tree of life (Poczai *et al.* 2013, Gallus *et al.* 2015, Doronina *et al.* 2017).

Any descriptive or exploratory data analysis of repeats in the chromosomal DNA has the potential to unveil new molecular markers in the forms of primers for target mobile elements. These primers are useful for many different target enrichment strategies, such as multiplex PCR (Taly *et al.* 2012), molecular inversion probes (MITs; see Mamanova *et al.* 2010), or Relay-PCR™ (technology made available by LC Sciences' VariantPRO Targeted Sequencing), that generates amplified material to be sequenced on a next-generation sequencing platform.

Exploratory analysis of repeat assemblies from four anuran genomes

In this chapter, an explanatory analysis is used to add to our understanding of the relationship of C-values and repetitive elements in the genome of four different anuran species.

Scaphiopus holbrookii (Harlan 1835) (Scaphiopodidae), popularly known as solitary frog or eastern spadefoot toad, among other English names, has a chromosome number of 26

(Olmo 1973) and reported C-value ranging from 1.34 to 1.41 pg, or 1.31 to 1.38 Gbp (Goin *et al.* 1968, Sexsmith 1968, Olmo 1973). The holoploid genome size of *S. holbrookii* is among the three smallest in Anura, losing only to *Limnodynastes ornatus* (929.1 Mbp) and *S. couchii* (1.05 Gbp).

Melanophryniscus moreirae (Miranda-Ribeiro 1920) (Bufonidae), popularly know as Maldonado redbelly toad, has a chromosome number of 22 and C-value of 2.94 pg (2.88 Gbp) (Beçak *et al.* 1970), which is about twice as large as the holoploid genome of *S. holbrookii* but smaller than the average for anurans (4.99 pg).

Hyloxalus subpunctatus (Cope 1899) (Dendrobatidae: Hyloxalinae), commonly known as the Bogota rocket frog, and *Phyllobates terribilis* Myers, Daly, and Malkin, 1978 (Dendrobatidae: Dendrobatinae), popularly known as the golden poison frog, have unknown chromosome numbers and C-values. But other anurans that have been unequivocally identified as dendrobatids (*Dendrobates tinctorius* and *Ameerega trivittatus*) have C-values ranging from 8.49 to 8.95 pg (8.30 to 8.75 Gbp, assuming their genome is diploid) (Camper *et al.* 1993), which is way above the average C-values of anurans and close to current upper limit for the order (*Ceratophrys ornata* has an octaploid genome with C-value of 13.4 pg; Horner and Macgregor 1983).

At the current state of knowledge of the evolution of amphibian genomes, the process of *de novo* assembling and annotating repeats should unveil valuable new information even in the absence of closely related reference genomes and karyotypes for all the selected species. Within this context; this chapter has the following general objects: (i) Describe repeat assemblies as a way to uncover new repeats and repeat polymorphisms; (ii) Use repeat annotation to reveal which classes of mobile elements are more abundant in each genome; (iii) Calculate the number of copies of each repeat and evaluate to which extent it is feasible to expect that repeats alone contribute to the variation of genome size in amphibians; and (iv) Propose candidate markers for phylogenetics.

Material and methods

The materials and methods described here are supplemented by detailed bioinformatic protocols described in Protocol S5.1.

Next-generation sequencing and quality control

The protocols for DNA extraction, library preparation, and quality control of raw sequence data for *S. holbrookii*, *M. moreirae*, *H. subpunctatus*, and *P. terribilis* have been described in chapter 3. Using BlastN (Altschul *et al.* 1990b), any repetitive DNA sequences that aligned against the corresponding complete mitochondrial genome (described in chapter 3) were removed before any other analysis was performed.

Repeat assembly

Paired-end (PE) and single-end (SE) reads were assembled separately with REPdenovo v0.1.9 (Chu *et al.* 2016). The number of reads and base pairs used for each species is summarized in **Table 5**.

Species	Library	Sequence reads	Base pairs
Scaphiopus holbrookii	Paired-end reads (x2)	981,558,462	94,223,438,454
	Single-end reads	666,720,836	9,301,615,402
Melanophryniscus moreirae	Paired-end reads (x2)	95,451,108	7,481,704,364
	Single-end reads	290,345,206	26,184,560,460
Phyllobates terribilis	Paired-end reads (x2)	686,475,576	72,691,278,270
	Single-end reads	213,335,759	21,511,100,580
Hyloxalus subpunctatus	Paired-end reads (x2)	238,563,156	22,787,987,820
	Single-end reads	98,034,297	10,541,009,366

Table 5 – Input DNA data for REPdenovo.

The scaffolds and contigs from independent assemblies of PE and SE sequence reads were merged in Geneious v8.1.9, using the built-in Geneious de novo assembly algorithm and custom parameters aimed to concatenate identical repeats in a conservative manner possible. Minimum overlap was set to 21 identical characters, with a total ambiguities representing no more than 2 nucleotides.

REPdenovo provides coverage estimation for scaffolds from PE assemblies but not for SE assemblies. Also, REPdenovo may overestimate coverage according to the programmer (Chong Chu, personal communication, Feb. 2, 2017). Therefore, after sequence concatenation, unique and aligned sequences must have their depth estimated some other way.

Repeat coverage

The coverage of each assembled repeat correlates with its copy number. The strategy used here to estimate the number of copies of each repeat has two steps. First, Bowtie2 v2.3.3 (Langmead & Salzberg 2012b) builds an index of all sequences and maps PE and SE sequence reads to it using very sensitive local settings. Bowtie2 does not report coverage estimation, but SAMtools v1.3.1 (Li *et al.* 2009) can covert the SAM alignment to BAM format so coverage can be calculated using the genomeCoverageBed program that comes with BEDtools2 v2.26.0 (Quinlan & Hall 2010). To convert sequence coverage to copy number, single copy genes also had their coverage calculated this way (see list of genes and their accession numbers on **Table 6**). The number of copies of each repeat was calculated as the average coverage of a repeat divided by the repeat length divided by a factor, which is the maximum ratio of average coverage and sequence length for single-copy genes.

	Accession number			
Gene	*S. holbrookii*	*M. moreirae*	*H. subpunctatus*	*P. terribilis*
bdnf	AB612074.1	AB612060.1	HQ290611.1	HQ290643.1
bmp2	NM_001015963.1	HQ291008.1	HQ291034.1	HQ291066.1
ntf3	NM_001016370.1	—	HQ290791.1	HQ290823.1
pomc	NM_001011318.1	AF194966.1	HQ290851.1	HQ290883.1
rag1	AB612071.1	KF666223.1	DQ503405.1	DQ503358.1
rho	AY323738.1	—	DQ503279.1	DQ503244.1
siah1	DQ282710.1	XM_018553010.1	GQ366167.1	GQ366169.1
slc8a1	JQ626831.1	JQ626831.1	JF703249.1	JQ626814.1
tyr	AB612073.1	KX026238.1	HQ290911.1	HQ290943.1
zeb2	XM_012971006.1	XM_018554053.1	HQ290671.1	HQ290703.1
cxcr4	—	DQ306494.1	—	—

Table 6 – Selected single copy genes used to normalized the calculation of the number of copies of repeats.

Repeat masking

Repeat masking is the process of annotating the repeat sequences and later mask (*e.g.*, replace by lower case base pairs) the regions that were successfully annotated. The process of repeat masking used the program RepeatMasker (Tarailo-Graovac & Chen 2009) and a repeat database from *X. tropicalis*.

Searching for candidate primer pairs

Primer3 (Untergasser *et al.* 2012) was used to find primers with the help of a homemade Python script, findPrimersForRepeats.py. The homemade Python program takes a multifasta file with repetitive DNA masked (lowercase) with RepeatMasker.

Primers for different types of satellites were estimated with the following parameters:

- Selected types: Chap4sat_Xt, MSAT2_XT, MSAT4_XT, REM2b_Xt, Sat1_Xt, Tc1Sat1_Xt, URR1a_Xt, and simple repeats with units with four or more nucleotides
- Maximum repeat length: 150 bp
- Primer TM: 58–62 ºC (optimal = 60 ºC)
- Maximal TM difference between left and right primers: 2 ºC
- Primer GC content: 40–60% (optimal = 50%)
- Product length: 80–150 bp

Primers for different retrotransposons were estimated with the following parameters:

- Selected types: SINE, MIR, L1, CIN4, L2, CR1, REX, and Penelope
- Maximum repeat length: 300 bp
- Primer TM: 57–63 ºC (optimal = 60 ºC)
- Maximal TM difference between left and right primers: 2 ºC
- Primer GC content: 20–80% (optimal = 50%)
- Product length: repeat length + 36 bp to repeat length + 80 bp

Repeat similarity

Due to the limitations imposed by de novo repeat assembly and homology-based annotation methods, the percentage of masked repeats is too small for any meaningful calculation of pairwise TE analysis or divergence time analysis. However, a general pairwise similarity analysis is possible through the comparison of words or substrings. The similarity among masked and unmasked repeats was estimated with a homemade Python script called wordSimilarity.py, which calculates the number of identical substrings (words or *k*-mers) of a particular length (*k*) among different multifasta files.

Percentage of repeats per genome

Due to incomplete sampling of sequence repeats (originated from wet-lab protocols as well as filters implemented *in silica*), the copy number of each repeat describe the minimum expected number of copies of that repeat in the haploid genome. Therefore, the diversity and abundance of repetitive DNA sequence elements (defined as the sum of the length of each repeat multiplied by its copy number) are also minimum expected values. Consequently, the estimation of the prevalence of repetitive DNA sequences in each genome relies on the independent analysis of the percentage of a subsample of sequence reads (excluding those that map to the mitochondrial genome) that map to the assembled repeat sequences using Bowtie2 with very sensitive local settings. Our strategy relies on the following assumptions: (i) After filtering out mitochondrial DNA sequences, the remaining sequence reads represent a random sample of the chromosomal DNA of each frog species; (ii) The ease of mappability of every sequence read is similar; (iii) All repetitive DNA sequences in our samples were correctly assembled.

Realistically, it would be unwise to assume that we entirely met all the above assumptions, and many additional observations are required to make more robust predictions. However, at the current stage of knowledge, the methods listed here should allow an adequate exploratory analysis of the data available.

Results

Quantification of repetitive elements

The total amount of assembled repeats, masked or not, is shown in **Table 7**. Although the quantities shown in this table represent minimal expect values, the amount and diversity of input sequence reads from *S. holbrookii* and *M. moreirae* in respect to the maximum expected hoploid genome size for these species is very high, and it is likely that the repeat assemblies for theses genomes are very close to complete.

Species	Repetitive DNA (bp)	†C-value (pg)	‡Percentage of expected genome size			Ratio of unmasked repeats
			All	*Unmasked*	*Masked*	
Scaphiopus holbrookii	1.8E+08	1.41[1]	13.05%	11.62%	1.43%	92.61%
Melanophryniscus moreirae	1.2E+09	2.90[2]	41.20%	37.23%	3.97%	92.40%
Hyloxalus subpunctatus	2.2E+09	8.95[3]	25.13%	21.98%	3.12%	91.34%
Phylobates terribilis	2.2E+09	8.95[3]	24.87%	21.93%	2.91%	90.19%

Table 7 – The total amount of repeats assembled (defined as the sum of the length of each repeat multiplied by its copy number), masked or not, and their proportion in relation each other and to the maximum expected holoploid genome size. †Maximum C-value reported in the literature for the species or its family. ‡Holoploid genome size assuming maximum C-value and diploid genomes. [1]Sexsmith (1968); [2]Beçak *et al.* (1970); [3]Camper *et al.* (1993).

Table 8 shows the expected prevalence of repeats in the four frog genomes in the form Expectedof the percentage of reads mapped to repeats.

Species	Input sequence reads†	Reads mapped to repeats	
		Total	*Percentage*
Scaphiopus holbrookii	1,648,279,298	231,091,512	14.02%
Melanophryniscus moreirae	385,796,314	230,551,322	59.76%
Hyloxalus subpunctatus	336,597,453	257,040,818	76.36%
Phyllobates terribilis	899,811,335	748,902,123	83.23%

Table 8 – Percentage of reads mapped to assembled repeats for each frog genome.

Qualification of repetitive elements

Table 9 indicates the prevalence of different repeats annotated with RepeatMasker using *X. tropicalis* as reference species. A few observations are worth noting here. In general, retrotransposons dominate the annotations and seem to increase in number as the genome size increases, specially LINEs and LTR elements, although the number of DNA transposons considerably outpasses the number of retroelements in *M. moreirae*. The portion of annotated SINEs is meager in all genomes, ranging from 0.01 to 0.04%, and the number of Penelope retroelements always surpasses the number of SINEs. Among DNA transposons, Tc1/ IS630/ Pogo elements are the most abundant. Other striking observations arise from comparing the annotated repeats in *H. subpunctatus* and *P. terribilis*, which are the species that have that are phylogenetic closest to each other. Low complexity repeats sum up 0.02% of the repeats in *H. subpunctatus* and 0.46% in *P. terribilis*, although they occupy 0.23% in both *S. holbrookii* and M. moreirae. Bel/ Pao retroelements are much rarer in all other species (up to 0.06%) than in *H. subpunctatus* (0.51%).

Element	*S. holbrookii*	*M. moreirae*	*H. subpunctatus*	*P. terrilis*
Retroelements	5.44%	3.99%	9.06%	8.88%
Retroelements - SINEs	0.02%	0.01%	0.03%	0.04%
Retroelements - Penelope	0.04%	0.01%	0.12%	0.10%
Retroelements - LINEs	4.89%	1.75%	2.92%	3.51%
Retroelements - LINEs - CRE/SLACS	-	-	-	-
Retroelements - LINEs - CRE/SLACS - L2/CR1/Rex	4.67%	1.65%	2.33%	2.75%
Retroelements - LINEs - CRE/SLACS - R1/LOA/Jockey	-	-	-	-
Retroelements - LINEs - CRE/SLACS - R2/R4/NeSL	-	-	-	-
Retroelements - LINEs - CRE/SLACS - RTE/Bov-B	-	-	-	-
Retroelements - LINEs - CRE/SLACS - L1/CIN4	0.16%	0.07%	0.37%	0.37%
Retroelements - LTR elements	0.52%	2.24%	6.11%	5.33%
Retroelements - LTR elements - BEL/Pao	0.06%	0.00%	0.51%	0.04%
Retroelements - LTR elements - Ty1/Copia	0.01%	0.02%	0.27%	0.17%
Retroelements - LTR elements - Gypsy/DIRS1	0.37%	1.84%	5.02%	4.78%
Retroelements - LTR elements - Gypsy/DIRS1 - Retroviral	0.06%	0.38%	0.25%	0.22%
DNA transposons	1.23%	6.22%	4.41%	4.70%
DNA transposons - hobo-Activator	0.01%	1.71%	1.12%	1.00%
DNA transposons - Tc1-IS630-Pogo	1.15%	3.86%	2.87%	3.46%
DNA transposons - En-Spm	-	-	-	-
DNA transposons - MuDR-IS905	-	-	-	-
DNA transposons - PiggyBac	0.06%	0.29%	0.22%	0.13%
DNA transposons - Tourist/Harbinger	0.00%	0.11%	0.03%	0.01%
DNA transposons - Other (Mirage, P-element, Transib)	-	-	-	-
Rolling-circles	-	-	-	-
Unclassified repeats	0.01%	0.02%	0.16%	0.04%
Total interspersed repeats	6.68%	10.23%	13.63%	13.62%
Small RNAs	0.72%	0.32%	0.39%	0.29%
Satellites	0.00%	0.02%	0.07%	0.06%
Simple repeats	1.26%	2.48%	1.27%	1.84%
Low complexity repeats	0.23%	0.23%	0.02%	0.46%

Table 9 – Percentage of repeats masked with RepeatMasker (using *Xenopus tropicalis* database).

Similarities in repeat composition

Table 10 shows the overall similarity among masked and unmasked repeats of the four selected frog genomes. In general, the results reflect the expectation that closely related species (*H. subpunctatus* and *P. terribilis*) share more substrings (in the form of words or 11-mers) with each other than with *S. holbrookii* or *M. moreirae*. Surprisingly, the repeats from *Scaphiopus* seem more similar to *Hyloxalus* and *Phyllobataes* than to *Melanophryniscus*.

Primer selection

The number of candidate primer pairs for satellites and other retrotransposons (LINEs, SINEs, and Penelope elements) is 267 and 775, respectively. Details are available in Table S5.3.

a) Unique 11-mers, masked

	Sh	Mm	Hs	Pt
Sh	—			
Mm	10.33%	—		
Hs	13.96%	13.37%	—	
Pt	14.68%	13.15%	26.58%	—

c) All 11-mers, masked

	Sh	Mm	Hs	Pt
Sh	—			
Mm	3.60%	—		
Hs	4.76%	5.62%	—	
Pt	5.37%	5.73%	13.74%	—

b) Unique 11-mers, unmasked

	Sh	Mm	Hs	Pt
Sh	—			
Mm	37.74%	—		
Hs	67.79%	37.70%	—	
Pt	68.40%	37.60%	68.66%	—

d) All 11-mers, unmasked

	Sh	Mm	Hs	Pt
Sh	—			
Mm	10.45%	—		
Hs	19.57%	10.03%	—	
Pt	20.88%	10.21%	20.83%	—

Table 10 – The similarity of masked and unmasked repeats among different frog genomes, measured according to the number of substrings of eleven letters (11-mer). **a)** Unique 11-mers in the masked repeats; **b)** Unique 11-mers in the unmasked repeats; **c)** All 11-mers in the masked repeats; **d)** All 11-mers in the unmasked repeats. **Sh**: *Scaphiopus holbrookii*; **Mm**: *Melanoprhyniscus moreirae*; **Hs**: *Hyloxalus subpunctatus*; **Pt**: *Phyllobates terribilis*.

Discussion

The number of observations available hampers the predictive power of our analysis, but the observed values in **Table 7** and **8** are consistent with each other. Moreover, they indicate that the assembled repeats of *S. holbrookii* (13.05% of the genome size) and *M. moreirae* (41.20%) is close to our estimation of the contribution of repeats to the total genome size in these frogs (14.02% and 59.76%, respectively). Therefore, it is reasonable within our observations to assert that genome size in amphibians is dependent on the prevalence of repetitive elements at least to a great degree. In fact, if about 14% of the genome of *S. holbrookii* is composed of repeats (see **Table 7)**, and insertion and expansion of repeats is responsible for all the variation in genome size in amphibians, them the prevalence of repeats in the genomes of *M. moreirae* and the selected dendrobatids (*H. subpunctatus* and *P. terribils*) would be around 58 and 86%, respectively, which is very close to the estimations presented here.

Of course, these are rough estimations based on preliminary data and depend on assumptions such that as that the genomes of H. subpunctatus and P. terribils are diploid and have C-values close to 8.95. Nevertheless, if we consider only the estimated amount of repeats in S. holbrookii and M. moreirae (for which the C-value and chromosome number are

known), and assume repeats can explain all the difference in genome size between these two species, we would conclude these two genomes have about 1 Gbp of nonrepetitive DNA. Interestingly, this is the same amount of nonrepetitive DNA found in the genomes of *Xenopus tropicalis* (Hellsten *et al.* 2010) and *Nanorana parkeri* (Sun *et al.* 2015) and is very similar to the smallest holoploid genome sizes in anura (see **Table S5.1**).

Additional repeat libraries are needed to annotate as much more repeats as possible in each of the four genomes studied here, which is a requirement for an in-depth analysis of the evolution of repeat elements in amphibians. Also, incomplete assemblages of repeats also make it hard to draw comparisons among the different anuran genomes. However, the low degree of similarity observed among 11-mer contents (**Table 10**) might be indicative that repeats insertions are more relevant to anuran genome size than repeat expansion. The frequent insertion of new repeats could also explain why so few repeats can be masked using only *X. tropicalis* repeat database.

In addition to the information described above concerning the evolution of anuran genome size, this exploratory analysis also yielded candidate phylogenetic markers (**Table S5.3**) which are ready to wet-lab testing, along with an easy protocol to obtain them (see **Protocol S5.1**). Considering the number of candidate primers found with only one-sixth of the repeats masked, comparison with more Repbase libraries (Jurka 1998, 2000, Jurka *et al.* 2005, Bao *et al.* 2015) will likely result in even larger numbers of markers.

Conclusion

We described new information that can sponsor further predictive and mechanistic studies of genome size variation in eukaryotes. Our observations, in combination with the published genomic data from X. tropicalis and N. parkeri, suggest that amphibian holoploid genomes might consist of about 1 Gbp of non-repetitive DNA sequences and that variation in anuran genome sizes in diploid species is chiefly if not wholly caused by insertion and expansion of different repeats. Investigations of repeats in the diploid genomes of anurans in the extremes of the variation of amphibian genome sizes are likely to shed more light on this matter. As we prepare this chapter for publication, additional efforts to assemble and annotate the repeats of the four anurans studied here might also unveil new details, mainly related to the divergence and abundance of different TE families.

Also, given the promising usage of some mobile elements as genetic markers for phylogenetic analysis, we proposed a list of candidate primer pairs for target enrichment as well as a detailed protocol to facilitate other researchers expanding on this list. If works such as Murata *et al.* (1993) and Gallus *et al.* (2015) are any indications of the power of repeat elements in phylogenetics, even if a small percentage of these markers hold after wet-lab tests we should expect them to provide meaningful information regarding the relationships of different clades of amphibians, specially those closely related to *Scaphiopus*, *Melaniphryniscus*, *Hyloxalus*, and *Phyllobates*.

List supplementary material

- Protocol S5.1 – Protocol for analysis of repeat elements in the chromosomal DNA
- Table S5.1 – Known C-values and chromosome numbers in Amphibia (source: Gregory, 2017; URL: http://genomesize.com; last accessed: Sep. 5, 2017)
- Table S5.2 – List of references in Table S5.1 (source: Gregory, 2017; URL: http://genomesize.com; last accessed: Sep. 5, 2017)
- Table S5.3 – List of candidate primer pairs.

All supplementary material is available upon request via the email machadodj@usp.br. After this thesis is provided a DOI by USP's Digital Library, you will be able to search for these materials at DRYAD (https://datadryad.org). Additionally, a compressed file containing all the Supplementary Material of this Ph.D. dissertation can be downloaded from http://www.ib.usp.br/grant/anfibios/datasets/Machado2018.zip.

6. KILLIFISHES ARE NOT POISON FROGS

Background

Voltage-gated sodium channels (Na_V) are crucial elements of action potential initiation and propagation in excitable cells because they are responsible for the initial depolarization of the membrane. These channels consist of an intensely processed α subunit that is ~260 kDa and is enough for functional expression. However, the kinetics and voltage dependence of channel gating are altered by auxiliary β subunits (Catterall 2000). With important implications to human medical research, much of our understanding of how Na_v operates comes from studies on animal systems.

One famous system that provided much information on Na_V functionality is the evolution of tetrodotoxin (TTX) resistance in the predator-prey arms-race between rough-skin newts (*Taricha granulosa*) and garter snakes (*Thamnophis* spp.). The TTX molecule is a small hidrophilic alkaloid. This potent neurotoxin has a particular and well-studied mode of action (Moczydlowski 2013): exposure to submicromolar concentrations of TTX causes paralysis and death by disrupting the initiation and propagation of action potentials (APs) in peripheral nerves and skeletal muscle (Narahashi 1972). The antipredator role of TTX in salamander is central to the coevolution between the TTX-bearing newts and garter snakes that eat them (Brodie 1968; Brodie III *et al.* 2005; Hanifin & Brodie 2008; Feldman *et al.* 2009, 2010; Hanifin & Gilly 2014). By studying the adaptive history and molecular basis of TTX-resistance in this system and other taxa (*e.g.*, fish, frogs, mollusks, and flatworms), scientists unveiled the specific regions (at the level of amino acid residues) of Na_V that regulate TTX-binding. These discoveries were greatly depended on solid field and laboratory work, involving experiments that range from the observation of intracellular action potentials (Aps) in salamanders before and after TTX injection (Hanifin & Gilly 2014) to mobility measurements of garter snakes that received different levels of TTX injections (Feldman *et al.* 2010).

Another system that might shed light the function of Na_V is the evolution of chemical defense of poison-dart frogs (Dendrobatidae) with high levels of alkaloid toxicity. Since the toxins on amphibian skins are part of their innate immune systems and play a significant role

in defending against pathogens, parasites, and predators (Brodie *et al.* 1991, Rivas *et al.* 2009b, Conlon 2011), understating the mechanisms of actions of these neurotoxins might also bring valuable information on amphibian biology and evolution. However, studying alkaloid resistance in poison frogs poses additional and specific challenges in comparison with the traditional TTX model. For instance, the poison frog system is much more complicated than the TTX system because it involves some 1,400 lipophilic alkaloids of 24 structural classes (Daly *et al.* 2005, Grant *et al.* 2006, Santos *et al.* 2016) and we know almost nothing about their binding sites or activity. There is little evidence that resistance to lipophilic alkaloids in poison frogs is centered on Na_V (although evidence suggest that these defensive chemicals could affect ion channels; see Daly & Spande 1986). Also, the level of autoresistance in poison frogs is mostly unkwon.

Despite the difficulties in working with the dart-poison frog system, Tarvin *et al.* (2016) presented an article that was the first to discuss the genetic basis of autoresistance in frogs with alkaloid defenses. Using protein-docking models, the authors predict that three significant classes of alkaloids found in poison frogs (histrionicotoxins–HTX, pumiliotoxins–PTX, and batrachotoxins–BTX) bind to similar sites in the highly conserved inner pore of the muscle voltage-gated sodium channel, Scn4a (also known as NaV1.4). The authors included a list of 49 residues predicted to interact with seven poison frog alkaloids of the classes above and six types of amino acid replacements that they predicted to provide resistance by decreasing target-site sensitivity to the neurotoxins. The six types of amino acid replacements, identified according to the modifications in relation to the rat Scn4a (NCBI accession number NM_013178.1), are: S429A, I433V, A446D, A446E, V1583I, and N1584T. According to the authors, these modifications that are exclusive to poison frogs in Dendrobatidae except for a distantly related alkaloid-defended frog from Madagascar, *Mantella aurantiaca*.

Given the new data that became available to us on the Scn4a of the eastern spadefoot toad, *Scaphiopus holbrookii*, and the alkaloid-defended golden poison frog, *Phyllobates terribilis* (see **Chapter 4**), we decided to replicate the work of Tarvin *et al.* (2016) with our new data. Here we report the minor inconsistencies and discrepant results we found, and challenge some of the hypothesis put forth by those authors.

Material and methods

Close inspection of Tarvin *et al.* (2016) revealed minor errors. We list the less striking inconsistencies in **Appendix S6.1**.

Taxonomy

According to the most recent and well supported taxonomy for amphibians, the species named *Bufu nubulifer* and *Dendrobates captivus* in Tarvin *et al.* (2016) will be treated here as *Incilius nebulifer* and *Excidobates captivus*, respectivelly.

Novel Scn4a sequence data

During the draft assembly and annotation of the draft genomes of *Scaphiopus holbrookii* and *Phyllobates terribilis* (see **Chapter 4**), we identified Scn4a sequences in both species (data will become available upon publication). The average sequence depth of the scaffolds of both species, re-checked using Bowtie2 (Langmead & Salzberg 2012) and SamTools (Li *et al.* 2009), with coverage above 100x and average identity to the consensus above 95% (key positions related to the residues associated by Tarvin *et al.* 2016 to alkaloid resistance had 100% identity and maximum quality). From these scaffolds, we used Primer3 (Untergasser *et al.* 2012) to design species-specific primers for the DI, DII, DIII, and DIV regions of Scn4a. The PCR amplification and Sanger sequencing using those primers are currently underway and will be used to check the sequences before publication on GenBank.

Aditional Na_V sequences from online databases

Tarvin *et al.* (2016: p. 1,068) predicted that poison frogs are somewhat resistant to certain alkaloids because they have six types of amino acid replacements in the Scn4a (sodium channel protein, type 4, alpha – Nav1.4) inner pore that are exclusive to posion-dart frogs (Dendrobatidae) except for a distantly related alkaloid-defended frog from Madagascar, *Mantella aurantiaca*. Protein-docking models and comparative phylogenetics support the role of these replacements in alkaloid resistance." These six replacements are S429A, I433V, A446D, A446E (DI-S6 region), V1583I, and N1584T (DIV-S6 region). The numbers correspond to the positions in the Scn4a of rats (NCBI accession number NM_013178.1).

However, the outgroup on Tarvin *et al.* (2016) consists of only five species: West Indian Ocean coelacanth (*Latimeria chalumnae*), red junglefowl (*Gallus gallus*), humans (*Homo sapiens*), house mice (*Mus musculus*), and Norway rats (*Rattus norveticus*). Since the authors do not mention how or if they searched for the modifications above on the other available sequences of Scn4a on public databases, we downloaded all NaV amino acid sequences from GenBank and UniProt (total of 37,914 entries) using combinations following search terms: "SCN", "NAV", "sodium channel", and "voltage-gated". The local database created this way included Scn4a and other Na_V paralogs.

Multiple sequence alignment

Different programs with various alignment parameters were used to align original sequences with the dataset from Tarvin *et al.* (2016) and selected sequences from the GenBank and UniProt database. Although less conserved regions of the Scn4a showed some degree of sensitivity to the alignment method used, all methods agreed on the alignment of the positions in the DI-, DII-, DIII-, and DIV-S6 regions. These is a list of the alignment programs tested: Muscle (Edgar 2004), Mafft (Katoh *et al.* 2005), Clustal W (Larkin *et al.* 2007), and the Geneious Alignment (Geneious version 8.1.9, http://www.geneious.com, Kearse *et al.* 2012).

Replication of docking analyses

A detailed protocol for the replication of Tarvin *et al.*'s (2016) docking analysis is given on **Protocol S6.1**. Most significant modifications to the original analysis include the inclusion of new receptor (Na_V) and ligand (alkaloid) models.

Replication of the docking analyses employed ten receptor models (see **Table 11**), including a new "complete" model based on the Scn4a of the blue poison frog, *Dendrobates tinctorius*. The "complete" receptor model include all modifications *D. tinctorius* have in relation to the rat reference in any of the 49 positions that Tarvin *et al.* 2016 associate to alkaloid activity. New models also incorporate residues found in killifishes, chicken, and salamanders that differ from the rat reference on one or more of those 49 positions. The other two new receptor models were based on the modifications found in the DI- to DIV-S6 regions of the Scn4a of turquoise killifishes and chickens that are on positions predicted by Tarvin *et al.* (2016) to be associated with alkaloid binding.

Model Name	423	424	429*	433*	445*	782	1262	1276*	1287*	1565	1569	1581*	1583*	1584*
Epipedobates					D									
Ameerega					D								I	
Hyloxalus													I	
Excidobates				V	E									
Dendrobates				V	D								I	
Killifish					D				V			V	I	
Chicken								A						
Salamander								A					I	
Complete	L	V		V	D	V	M	A		A	V		I	
Phyllobates			A	V	D								I	T

(*) Positions associated by Tarvin et al. with modifications that could lead to alkaloid resistance.

Table 11 – Receptor models used during the replication of docking analyses. Only positions that had AA modifications in relation to the rat reference are shown. Models in PDB format are available with the **Supplementary material** (**File S6.1**).

Ligand models included all alkaloids used by Tarvin *et al.* (2016) plus samandarin (ChemSpider ID: 107738) and samandarine (ChemSpider ID: 107738), two potent alkaloid neurotoxins found in the fire salamander.

Docking analysis was executed on AutoDock VINA version 1.1.2 (Trott & Olson 2009). The free Gibbs energy (affinity, indicated by ΔG) reported by the program is given in $kcal \times mol^{-1}$ and was converted to $kJ \times mol^{-1}$. Basics of thermodynamics say that a reaction with negative free energy is spontaneous. In other words, the smaller the ΔG, the higher the requested energy to break that binding between two molecules. The affinity variation (ΔΔG) of mutation model in relation to the receptor model if defined as $\Delta\Delta G = \Delta G_{reference} - \Delta G_{mutated}$. Defined this way, a more negative ΔΔG corresponds to decreased affinity of the mutated model in relation to the reference model. Please note that this definition of ΔΔG is the opposite as the one used in Tarvin *et al.* (2016).

Statistical analysis followed the guidelines of Tarvin *et al.* (2016), using the Wilcoxon-Mann-Whitney test to compared the affinity of different mutated receptor to the original "Walker 1" model (Walker *et al.* 2012: Database S1, homology model for Scn4a). We also used the same statistical test to compared the affinity variation between the models based on the alkaloid-defended *D. tinctorius* and the models based on killifishes and chickens.

Results

Revision of AA residue positions

It is likely that the modifications refereed by Tarvin *et al.* (2016) as A446D/ E are a numbering error and for now on we will refer to these same modifications as A445D/ E.

All the AA residues in the Scn4a of *S. holbrookii* match the sequence of frogs of the family Pipidae, as the western clawed frog, *Xenopus tropicalis*.

Mutations specific to the golden poison frog

We were unable to locate the residues reported by Tarvin *et al.* (2016) for *P. terribili*s on our original sequences of Scn4a for this species. We will amplify primers designed by Tarvin *et al.* (2016) and primers designed from our original data to confirm whether or not the residues 429A and 1584T can be found on the Scn4a of the golden poison frog.

Exclusivity of residues associated to alkaloid resistance

Tarvin *et al.* (2016: p. 1,069) declares that "All five sites [*i.e.*, DI-S6 429, 433 and 445 in DIV-S6 1583 and 1584] are intriguing because these residues are highly conserved among vertebrates, and the replacements are unique to poison frogs and *Mantella*." However, all the replacements that are not exclusive to the sequences of *P. terribilis* used in their article can be found in the Scn4a of other non-amphibian, non-toxic animals. The number of Scn4a, Scn4a-like, and other sodium channel proteins with these modifications is large. Bellow, we will mention only a few examples. A complete list of the matches we found (which is still not a complete census of the available protein databases) is available on the **Supplementary material** (**Tables S6.1** and **S6.2**).

433V + 445E—The modification I433V is predicted to have at least three origins. It is associated with the origin of chemical defense in *Phyllobates* + *Dendrobates*. In *P. terribilis*, Tarvin *et al.* (2016) report this modification in association with A429A, A445D, V1583I, and N1584T. In *D. ticntorius*, this modifications happens with A445D and V1583I. Finally, this modifications occurs with A445E in *Excidobates captivus*.

The phylogenetic tree impacts the conclusions regarding this modification since Tarvin *et al.* (2016) use phylogenetic correlation for additional evidence that this replacement is

associated with alkaloid resistance, I433V, but it increased the binding affinity of PTX 251D and had no significant effect on other alkaloid dockings, according to the authors.

Tarvin *et al.* (2016) state that all three species-level patterns including this replacement (*D. tinctorius*, *E. captivus*, and *P. terribilis*) did not have a significantly higher PTX 251D binding affinity than the Scn4a model without mutations. Also, the full *D. tinctorius* species model (including replacements I433V and A446D) shows a more significant decrease in BTX, aPTX 267A, and aPTX 323B binding affinity than predicted for V1583I alone.

The I433V modification is found on the Scn4a of other animals such as rainbow trouts (XP_021477990), Maylandia zebras (XP_014268546), and northern pikes (XP_019901179), which also have the A445E modification (*i.e.*, matching the full *E. captivus* model). Although, considering the animal tree of life, perhaps in this case we should call these modifications V433I and E445A instead. To avoid confusion, we will address the modifications as <position in the rat Scn4a><AA> residues from now on.

The 433V residue is additionally found in other sodium channel proteins which have DI-S6 AA sequences very similar to rats or poison frogs. For example, the residues 433V and 445E are found on Scn3a of humans (NP_001075145), rats (NP_037251), mice (NP_001342095), and zebrafish (NP_001038387). These residues are further located on the Scn8a of humans (NP_055006), rats (NP_062139), mice (NP_035453), zebrafish (NP_001038648), and many others.

The Scn3a ("sodium channel protein type 3 subunit alpha"), as the Scn4a gene, mediates the voltage-dependent sodium ion permeability of excitable membranes. Assuming opened or closed conformations in response to the voltage difference across the membrane, the protein forms a sodium-selective channel through which Na+ ions may pass in accordance with their electrochemical gradient. In humans, the Scn3a gene is mainly expressed in the forebrain, neocortex, and frontal cortex (Bgee dataBase for Gene Expression Evolution, Ensembl ID ENSG00000153253). Not only all isoforms of the Scn3a in humans have the residues 433V and 445E, but it matches 100% of the AA of the DI-S6 region of the Scn4a of *E. captivus*.

The Scn8a ("sodium channel protein type 8 subunit alpha") also has similar functions as the Scn3a and Scn8a genes. In macrophages and melanoma cells, isoform 5 of Scn8a may participate in the control of podosome and invadopodia formation. In humans, the Scn8a is mainly expressed in the frontal cortex, corpus callous, middle temporal gyrus, primary visual

cortex, and Broadmann (1909) area 23 (Bgee, ENSG00000196876). The human Scn8a DI-S6 only differs from the Scn4a DI-S6 of *E. captivus* by 1 AA (427V).

445D + 1583I—According to Tarvin *et al.*, 445D has at least five origins within amphibians and evolved independently in *Mantella*, which has PTX defense convergent with dendrobatids. Predictions that this replacement provides alkaloid resistance relies on phylogenetic correlations. *Ameerega parvula* has increased resistance predicted by the presence of 445D. The presence of 445D is only marginally correlated increase levels of PTX, but it collaborates to resistance in the full model of *D. tinctorius*. Tarvin *et al.* argue that the presence of 445D or 445E in lineages with high PTX/ BTX defense could support their model of autoresistance evolution.

The DI-S6 AA sequences of *M. aurantiaca* and *A. parvula* are equal and identical (including the 445D residue) to the corresponding fragment on the Scn4a of the Atlantic salmon (A0A1S3NG10_SALSA) and the Carolina anole (G1KMZ7_ANOCA). Also, the Scn4a of the turquoise killifish (XP_015818233) has not only the 445D residue but also the 1583I residue (*i.e.*, similar to *A. parvula*).

Th 1583I residue was not significantly correlated with origins of defense. It evolved at least five times in amphibians. Full models with a combination of modifications provided better results than this modification alone. Nevertheless, the docking models of Tarvin *et al.* predict that resistance to alkaloids is conferred by 429A and 1583I replacements in Scn4a, which decrease alkaloid binding affinity. The authors also state that the 1583I residue decreased the binding affinity of BTX, HTX, and all five PTXs. The 429A and 1583I residues were the only ones that significantly lowered alkaloid binding affinity. Moreover, Tarvin *et al.* (2016) suspect that the 15831I provide broad alkaloid resistance.

In total, the 1583I residue evolved three times in clades of alkaloid defended poison frogs and, unexpectedly, twice in undefended clades *Hyloxalus* and *Silverstoneia*. As we said before, the 1583I residue is also found in with 445D in the turquoise killifish. In addition to that, it is seen on the Scn4a of the fire salamander (A0A0A0Y2X1_SALSL), which had its Scn4a sequence studies many times before in the context of toxin resistance. The 1583I residue in the fire salamander has never been designated as a special modification that provided chemical resistance to the fire salamander, and it does not happen on the Scn4a of other more toxic Caudata such as the rough-skin newt (A0A0A0Y7I0_TARGR). The modification 1583I residue is further observed in the sodium channel proteins of frogs that do

not have alkaloids (*e.g.*, *Xenopus laevis* – A0A1L8HA74_XENLA) and other types of sodium channels proteins of various animals (*e.g.*, whale shark – XP_020387692, copepods – A0A125R3Q8_9MAXI).

It is intriguing that, from 7 species in Tarvin *et al.* (2016) and two other species that we found (the turquoise killifish and the fire salamander) with 1583I residue, 2 are frogs with no alkaloids and two more are animals with either no alkaloids or an entirely different kind of toxins.

Docking analyses

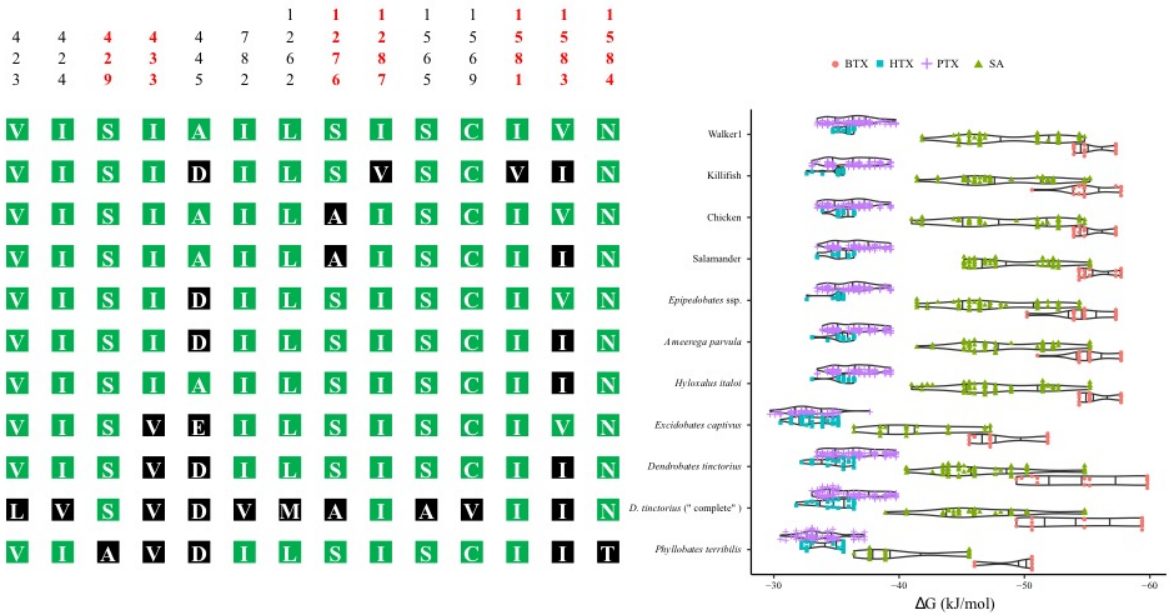


Figure 11 – Graphical summary of the docking analyses in AutoDock Vina. Modified residues and shown on the left, with modifications to the rat reference indicated in black. On the right, the dot plots are shown on top of violin plots that indicate de density of data with lines representing the 0.25, 0.5, and 0.75 quartiles. High-resolution image (**Figure S6.1**) as well as a violin plot for different PTX alkaloids (**Figure S6.2**) are in the **Supplementary material**.

Figure 11 provides a graphical summary of the docking analyses in AutoDock Vina, and the main results from Wilcoxon-Mann-Whitney tests are provided in **Table 12**.

The results we obtained from the Wilcoxon-Mann-Whitney tests differ from those presented in Tarvin *et al.* (2016) the following ways:

Wilcoxon-Mann-Whiteney test results, including test–statistics (W), p–values, and ΔΔG (kJ/ mol).

		BTX	HTX	PTX251D	PTX307A	PTX323A	aPTX267A	aPT 323B	San	Sae
Chicken	W	117.5	104	103.5	115	141	109.5	120	112.5	115.5
	[a] ***p*–value**	0.845	0.732	0.713	0.931	0.228	0.908	0.653	1.000	0.908
	[b] **ΔΔG**	**-0.056**	0.084	0.112	**-0.028**	**-0.251**	0.000	**-0.028**	0.000	**-0.056**
Killifish	W	95	163	55	117.5	145	95.5	6.5	89.5	144.5
	[a] ***p*–value**	0.467	0.032	0.014	0.845	0.169	0.439	**0.000**	0.315	0.182
	[b] **ΔΔG**	0.195	**-0.446**	0.474	**-0.056**	**-0.363**	0.056	0.614	0.084	**-0.223**
Salamander	W	87.5	91.5	36	110	152	66.5	0	67.5	128.5
	[a] ***p*–value**	0.299	0.371	**0.001**	0.930	0.094	0.044	**7.092E-07**	0.050	0.504
	[b] **ΔΔG**	0.279	0.139	0.892	0.028	**-0.307**	0.279	0.753	0.251	7.105E-15
Ameerega parvula	W	75	106	0	100.5	152	49	0	67.5	133
	[a] ***p*–value**	0.120	0.798	**0.000**	0.614	0.093	**0.004**	**7.092E-07**	0.050	0.391
	[b] **ΔΔG**	0.418	0.056	1.255	0.084	**-0.307**	0.335	0.753	0.251	**-0.028**
***Epipedobates* spp.**	W	112.5	161	98.5	112.5	132	119.5	120	97.5	136.5
	[a] ***p*–value**	1.000	0.040	0.557	1.000	0.418	0.762	0.653	0.510	0.296
	[b] **ΔΔG**	0.000	**-0.418**	0.112	0.000	**-0.167**	**-0.028**	**-0.028**	0.084	**-0.418**
Hyloxalus italoi	W	75	83.5	36	120	160	56	0	67.5	135
	[a] ***p*–value**	0.120	0.215	**0.001**	0.758	0.042	0.012	**0.000**	0.050	0.347
	[b] **ΔΔG**	0.418	0.223	0.920	**-0.028**	**-0.363**	0.307	0.781	0.251	**-0.139**
Dendrobates tinctorius	W	115	78	60	165	111	14	0	202.5	152.5
	[a] ***p*–value**	0.933	0.139	0.024	0.024	0.966	**0.000**	**0.000**	**0.000**	0.094
	[b] **ΔΔG**	0.335	0.251	0.586	**-0.335**	0.000	0.837	1.450	**-0.864**	**-1.924**
***D. tinctorius* (complete)**	W	120	97.5	55	175	115	10.5	0	202.5	152.5
	[a] ***p*–value**	0.766	0.530	0.014	**0.008**	0.931	**0.000**	**0.000**	**0.000**	0.094
	[b] **ΔΔG**	**0.112**	**0.139**	**0.586**	**-0.446**	**-0.028**	**0.697**	**1.422**	**-0.864**	**-1.924**
Excidobates captivus	W	225	203.5	202.5	225	225	181	214.5	225	225
	[a] ***p*–value**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.003**	**0.000**	**0.000**	**0.000**
	[b] **ΔΔG**	**-6.051**	**-1.227**	**-1.227**	**-5.159**	**-4.769**	**-1.060**	**-2.398**	**-5.605**	**-9.370**
Phyllobates terribilis	W	225	178	198.5	225	225	222.5	222	225	225
	[a] ***p*–value**	**0.000**	**0.006**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	[b] **ΔΔG**	**-5.605**	**-0.753**	**-1.032**	**-3.402**	**-4.573**	**-1.143**	**-1.450**	**-6.860**	**-9.314**

[a] Light grey cells indicate *p*–value between 0.05 and 0.01. Dark grey cells indicate *p*–value smaller than 0.01.

[b] $\Delta\Delta G = \Delta G_{Walker1} - \Delta G_{model}$. Negative values indicate a receptor with less affinity to the ligand). Darker cells indicate negative values.

Table 12 – Wilcoxon-Mann-Whiteney test results, including test–statistics (W), p–values, and ΔΔG (kJ/ mol). BTX = batrachotoxin; HTX = histrionicotoxin; PTX = pumiliotoxin; San = samandarin; Sae = samandarine.

1. Generic models for the single AA replacements 429A, 433V, and 445E: These models were not included since we never observe any of these residues solely.
2. The *Epipedobates* spp. (445D) receptor model: These modifications did not significantly alter resistance in Tarvin *et al.* (2016). In our analysis, however, this residue increases the affinity of the receptor to PTX (pumiliotoxin 251D, aPTX 267A, and aPTX 323B). *Ameeraga parvula* has lesser levels of PTX, but higher levels of HTX and no observed BTX toxicity. This residue is also observed in *P. terribilis* (no PTX), *D. tinctorius* (major levels of PTX), and in many species of *Epipedobates*

(which have various levels of PTX toxicity, no BTX, and minor levels of HTX only in *E. darwinwallacei*).

3. The *Hyloxalus italoi* (1583I) receptor model: Tarvin *et al.* reported that the residue 1583I by itself would reduce the affinity of the receptor to all ligands. In our analysis, the affinity of the modified receptor was only significantly different from the Walker1 model for the ligands pumiliotoxin 251D, aPTX 267A, aPTX 323B (increased affinity), and PTX 323A (increased resistance). There is no observation of BTX, HTX, or PTX in *H. italoi* or *Silverstoneia flotator*, two species included in Tarvin *et al.*'s (2016) analyses that have the 1584I residue alone. *Ameerega bilinguis* also has the residue 1583I by itself and has minor levels of HTX with no BTX or PTX toxicity.
4. The *Ameerega parvula* (445D + 1583I) receptor model: In Tarvin *et al.*, these combined model significantly increased resistance to BTX, HTX, and PTX. In our analysis, it increased the affinity of the receptor to pumiliotoxin 251D, aPTX 267A, and aPTX 323B. *Ameerega parvula* has minor levels of PTX, major levels of HTX, and no BTX.
5. The *Excidobates captivus* (I433V + 445E) receptor model: In Tarvin *et al.* (2016), this model was not significantly more or less resistant to toxins than the Walker1 model. Herein, this model was significantly more resistant to all ligands (including samandarin and samandarine), showing the highest ΔG variation of all models (see **Table 14**). *Excidobates captivus* has no observed BTX, HTX, or PTX toxicity.

	***p*–value < 0.05, incresed resistance**	***p*–value < 0.05, more affinity**	**Avg. ΔΔG**
Salamander	0	3	0.257
Ameerega parvula	0	3	0.313
Killifish	1	2	0.037
Chicken	0	0	-0.025
Hyloxalus italoi	1	3	0.263
***Epipedobates* spp.**	1	0	-0.096
Dendrobates tinctorius	2	3	0.037
***D. tinctorius* (complete)**	2	3	-0.034
Excidobates captivus	9	0	-4.096
Phyllobates terribilis	9	0	-3.793

Table 13 – Terminals ordered according to, first, the number of ligand with significant worst affinity and, second, by the number of ligands with significant better affinity. More negative ΔΔG (kJ/ mol) indicate mutations that increased the ΔG of the model. The average ΔΔG in kJ/ mol for all permutations of models and ligands is given on right column.

6. The *Dendrobates tinctorius* (433V + 445D + 1583I) receptor model: In Tarvin *et al.* (2016), this model was not significantly more or less resistant to toxins than the Walker1 model. In our replication, this model displayed increased resistance to PTX 307A and samandarin, but increase affinity to pumiliotoxin 251D, aPTX 267A, and aPTX 323B. *Dendrobates tinctorius* has no BTX but has significant levels of HTX and PTX. It is interesting to note that the average ΔΔG of this model (0.0371822222) was closest to the average ΔΔG of the killifish model (0.037182222) and higher (less resistant) than that of the chicken model (-0.024788148). Additional Wilcoxon-Mann-Whitney tests comparing the *Dendrobates tinctorius* model with chicken and killifish models (see **Table 15**) showed that, in comparison to the *D. tinctorius* model, the chicken model was statistically more resistant to pumiliotoxin 251D, aPTX 267A, and aPTX 323B at the same time its affinity to PTX 307A and samandarin increased. On the other hand, the killifish model was more resistant to HTX, aPTX 267A, and aPTX 323B and had more affinity to both samandarin and samandarine than the *D. tinctorius* model.

		BTX	HTX	PTX251D	PTX307A	PTX323A	aPTX267A	aPTX323B	San	Sae
Chicken	**W**	112	141.5	160	65	141	209	225	22.5	70
	***p*–value**	1.000	0.213	**0.042**	**0.042**	0.230	**0.000**	**0.000**	**0.000**	0.075
	ΔΔG	**-0.390**	**-0.167**	**-0.474**	0.307	**-0.251**	**-0.837**	**-1.478**	0.864	1.868
Killifish	**W**	115	28	119.5	155	80	20	17	220	167.5
	***p*–value**	0.933	**0.000**	0.779	0.070	0.171	**0.000**	**0.000**	**0.000**	**0.021**
	ΔΔG	**-0.139**	**-0.697**	**-0.112**	0.279	**-0.363**	**-0.781**	**-0.837**	0.948	1.701

Table 14 – Wilcoxon-Mann-Whiteney test results comparing the chicken and killifish models to the *Dendrobates tictorius* (simple model with 433V + 445D + 1583I), including test–statistics (W), p–values, and ΔΔG (kJ/ mol). BTX = batrachotoxin; HTX = histrionicotoxin; PTX = pumiliotoxin; San = samandarin; Sae = samandarine.

The **Supplementary material** for this chapter includes all the receptor models (**File S6.1**) and details of the statistical analysis, including complete results table and scripts for execution in R (**File S6.2**).

Discussion

Of the six types of amino acid replacements in the Scn4a inner pore that Tarvin *et al.* (2016) associated with alkaloid resistance, two are exclusive to the golden poison frog (*P. terribilis*)

and were not observed in the sequences we assemble for this species. The other four replacements are observed, solely or in some combination, in the Scn4a of animals that are not resistant to poison frog alkaloids or in which resistance would be unexpected (*e.g.*, killifishes, trouts, salmons, chickens, anole lizards, and zebras). Furthermore, these same replacements and sometimes the entire DI- or DIV-S6 regions of the poison frogs that possess them are seen in the equivalent regions of other Na_V paralogs of animals such as zebrafishes, whale-sharks, mice, rats, and humans. These observations break the phylogenetic exclusivity of these AA residues to alkaloid-defended frogs and weaken the hypothesis that they confer decrease affinity to the ligand for being present in animals with no known resistance to these neurotoxins.

Results of the replication of docking analyses also contradicted the hypothesized cause-and-effect relationship between the six AA residues identified by Tarvin *et al.* (2016) and increased poison frog alkaloid resistance. Comparing mutated models to the rat (Walker1) reference model, our results indicate that the receptor models based on the Rio Santiago poison frog (*E. captivus*) (I433V, 445E) and the golden poison frog (429A, 433V, 445D, 1583I, and 1584T) showed the most conspicuous decrease in affinity. Yet, even their affinity values dramatically overlap with those from Walker1 and do not become positive enough to justify expecting a biological effect. In other words, we argue that the statistically significant variation observed between mutated and reference model in binding affinity does not necessarily hint a difference in biological effect.

One indication that the statistical difference between binding affinities of different receptor models to the ligands does not necessarily correlate to a biological effect comes from the docking analyses with samandarin and samandarine, two potent lipophilic alkaloid neurotoxins found in fire salamanders (*Salamandra salamandr*a). The fire salamander receptor model does not show statistical differences in binding affinity to these ligands in comparison to the Walker1 model, but the models based on the Rio Santiago and blue poison frogs (*D. captivus*) are statistically more resistant to them. If the AA residues we analyzed are indeed providing resistance to alkaloids, it would be possible but not expected for a toxic species to do not resist their toxin while other species are resistant to them without having them.

Also, take the example from receptor models based on chickens, turquoise killifishes, and the incredibly toxic blue poison frog. None of these models were significantly more resistant to BTX, HTX, or PTX in our replication of the docking analyses. However, if we

were to compare and discuss them using the same statistical criteria as Tarvin *et al.* (2016), the killifish model would be more resistant than the blue poison frog model to HTX, aPTX 267A, and aPTX 323B. On the other hand, the chicken model would be more resistant to PTX 251D, aPTX 267A, and aPTX 323B. Given that the blue poison frog has major amounts of HTX and PTX, as indicated by Tarvin *et al.* (2016: Fig. 2), this interpretation would be nonsensical, pointing further towards the idea that the affinity values observed are too negative and too similar for us to expect they indicate real variation in binding affinity.

Conclusion

Given the importance of defensive compounds in diverse aspects of amphibian biology, studies of chemical defense are essential to understanding amphibian diversification. Therefore, the contributions of Tarvin *et al.* (2016) are welcome not only for the new data they bring to light but also due to the debate they foster.

For example, Tarvin *et al.* (2017) studied the defensive alkaloid epibatidine, a nicotinic acetylcholine receptor (nAChR) agonist that is lethal at microgram doses. Epibatidine shares a highly conserved binding site with acetylcholine, making it difficult to evolve resistance yet maintain nAChR function. Electrophysiological assays of human and frog nAChR allowed Tarvin *et al.* (2017) to illustrate how resistance to agonist toxins can evolve and indicate some genetic changes that drive organisms near an adaptive peak of chemical defense. However, just as in the case of TTX, epibatidine has a well-known biding activity, and the researchers had direct observations of the binding effects of the alkaloid to the receptor in different experimental conditions.

However, the mechanisms behind BTX, HTX, and PTX are much less understood. For instance, we lack bioassays for many species of alkaloid-protected frogs, and the diversity of alkaloids on the skin of most dendrobatids is still waiting to be investigated. Therefore, it is perhaps not completely surprising that we could not replicate the results and corroborate the hypothesis of the very first study to first to try to demonstrate the genetic basis of autoresistance in frogs with alkaloid defenses. We hope that this serves as an incentive for more researchers to contribute to the research on amphibian chemical defense and that this also serves as a small example of how new genomic data, even from preliminary studies, can contribute to the field.

List of supplementary material

- Appendix S6.1 – Minor problems found on Tarvin *et al.* (2016), with comments
- Protocol S6.1 – Docking analyses
- Figure S6.1 – Graphical summary (Figure 9, high resolution)
- Figure S6.2 – Violin plots for docking of PTX alkaloids
- Table S6.1 – List of sequences of voltage-gated sodium channels available in GenBank or UniProt which possess residues 429A, 433V 445D, and/ or 445E.
- Table S6.2 – List of sequences of voltage-gated sodium channels available in GenBank or UniProt which possess residues 1583I and/ or 1584T.
- File S6.1 – Receptor models in PDB format and ligand models in MOL2 format
- File S6.2 – Statistical analysis (following methodology chosen by Tarvin *et al.* 2016)

All supplementary material is available upon request via the email machadodj@usp.br. After this thesis is provided a DOI by USP's Digital Library, you will be able to search for these materials at DRYAD (https://datadryad.org). Additionally, a compressed file containing all the Supplementary Material of this Ph.D. dissertation can be downloaded from http://www.ib.usp.br/grant/anfibios/datasets/Machado2018.zip.

7. CONCLUSION

Thomas Robert Cech is an American chemist who shared the 1989 Nobel Prize in Chemistry with Sidney Altman for their independent discovery of catalytic properties of RNA. There is a quote from Cech that appears in many books on genetics, from Berman's "Rare Diseases and Orphan Drugs" to Parrington's "The Deeper Genome." The quote is: "Because all biology is connected, one can often make a breakthrough with an organism that exaggerates a particular phenomenon, and later explore the generality." This is the same biological connection shared among all living things that guarantee that unveiling new genomic data from different branches of the animal tree of life increases our understanding of genome evolution and, consequently, sponsor our comprehension of the mechanisms that make us human.

It is striking to notice that, although all biology is connected, biological research is not unified. This disconnection partially explains why, although phylogenetic systematics has advanced dramatically in the last decade, attempts at integrating cutting-edge DNA sequencing technology to the study of amphibian evolution are still few and sparsed if we take into account all questions that are still begging for answers (*e.g.*, Peloso *et al.* 2016, Portik *et al.* 2016, Rodríguez *et al.* 2017). In fact, although McCormack & Faircloth (2013) have affirmed that next-generation phylogenetics (*i.e.*, the integration of high-throughput sequencing technology to studies in phylogenetic systematics) took root, most labs are still far from reaching this reality.

Among the many alternatives to sponsor next-generation phylogenetics, the scientific community can invest in the sequence and analysis of the genomes of more organisms and share the technology developed along the way, shading light to the single tree of life that might one day unify all researchers studying it.

Currently, the absence of genome references in many branches of the tree of life is not only "slowing research into specific questions; it is precluding a complete description of the molecular underpinnings of biology necessary for a true understanding of life on our planet" (Richards 2015: p. 414). To improve genome sequencing and survey a broader diversity of species, we need to take advantage of new sequencing technologies that enable cost-efficient nucleotide sequencing and support initiatives in both big science and small genome projects. Furthermore, scientists need to continue seeking collaborations that shorten the distance

between basic and applied research and advance de novo reference genome sequencing not as an end, but rather as a foundational necessity for biological and medical research.

Given the above, and recognizing the important place amphibians have in the tetrapod diversification, the present work aimed to add to the ongoing efforts to understand animal genomes in general and specific traits of anurans in particular. The first chapter summarized the current state of the research on the evolution and biology of amphibians in general, and on amphibian chemical defense in particular, conceding that poison frogs are a key component of the hidden pharmacology of the planet. We also advocated that comparative genomics is a strategy to increase our understanding of amphibian chemical defense and evolution. Therefore, the remaining chapters represented the first steps we took in that direction. On the second and third chapters, we addressed the challenges in the *de novo* assembly of mitochondrial genomes in the absence of genome references while providing new methods to facilitate the acquisition of such data. In the fourth and fifth chapters, we summarized the advances we made in the assembly of the genomes of the eastern spadefoot toad (*Scaphiopus holbrookii*) and the golden poison frog (*Phyllobates terribilis*), including comments on the assembly of repetitive DNA of these frogs, the Bogotá rocket frog (*Hyloxalus subpunctatus*), and the Maldonado redbelly toad (*Melanophryniscus moreirae*). These are incomplete genome assemblies and annotations, but they pave the way for later, thorough comparative analysis. We also used this data to propose new, ready-to-use phylogenetic markers and complete pipelines to generate them. Finally, on the last chapter, we used a few sequences from the genomes of *P. terribilis* and *S. holbrookii* to examine the work and conclusions of Tarvin *et al.* (2016), falsifying some of the hypothesis put forth by these authors and contributing to our understanding of alkaloid resistance in poison-dart frogs. In combination, these chapters add to our infant but growing knowledge of amphibian genomics and chemical defense and strengthen the communication between basic research in non-model organisms and cutting-edge bioinformatic methods.

Perspectives

The genomes of *M. moreirae*, *H. terribilis*, *S. holbrookii*, and *P. terribilis* are ongoing projects that await additional sequences from the sequencing facilities (*e.g.*, Macrogen and DHM-RI). The draft genome of *S. holbrookii* is the closest to be ready for publication, and we intend to submit it in the following year. The draft genome of *P. terribilis* will be concluded in

partnership with the laboratory of Professor Marcus Kronforst (The University of Chicago) and the laboratory of Professor Daniel Janies (University of North Carolina at Charlotte – UNC Charlotte). The draft genome of *H. subpunctatus* will also be a conjoint work between Professor Janie's laboratory and our Laboratório de Anfíbios (Instituto de Biociências, Universidade de São Paulo – USP).

Our partnership with Professor Janies started with a research internship financed by the Fundação de Apoio à Pesquisa do Estado de São Paulo (FAPESP, Proc. No. 2015/18654-2) in 2016-2017 and already led to a collaboration with the David H. Murdock Research Institute (DHM-RI), which resulted in a 150% increase of the DNA sequence data available for this thesis during 2016. This collaboration permitted us to use the computational resources at UNC Charlotte, which was indispensable to the conclusion of this work. The Laboratório de Anfíbios intend to maintain this fructiferous collaboration in our future endeavors in amphibian genomics.

Our laboratory also plans to collaborate with Dr. Ralph A. Saporito (John Carrol University) and Professor Edmund Brodie Jr. (Utah State University) on the transcriptomic analysis of populations of the rough-skin newt (*Taricha granulosa*) that posses different levels of tetrodotoxin (TTX) toxicity. Preliminary data on this project (not shown here) will be presented to FAPESP shortly as a post-doctoral research proposal to unveil the genetic bases of TTX acquisition and resistance.

Furthermore, the Laboratório de Anfíbios will immediately start to test different strategies enrich to use the new phylogenetic markers proposed here using both fresh and museum tissue samples. Our goal is to provide the interested scientific community with new tools to study amphibian phylogenetics on a large scale.

Publications

The student authored or co-authored the following research articles during the development of this doctoral thesis:

Machado D. J. (2015) Denis Jacob Machado. YBYRÁ facilitates comparison of large phylogenetic trees. *BMC Bioinformatics*. **16(1)**: 204–204. DOI: 10.1186/s12859-015-0642-9.

Machado D. J., Lyra M. L., Grant T. (2016) Mitogenome assembly from genomic multiplex libraries: Comparison of strategies and novel mitogenomes for five species of frogs. *Molecular Ecology Resources*. **16(3)**: 686–693. DOI: 10.1111/1755-0998.12492.

Grant T., Rada M., Anganoy–Criollo M., Batista A., Dias P. H., Jeckel A. M., Machado D. J., Rueda-Almonacid J. V (2017) The pylogenetic systematics of dart-poison frogs and their relatives revisited (Anura: Dendrobatoidea). *South American Journal of Herpetology*. **12 (special issue)**: S1–S90. DOI: 10.2994/SAJH-D-17-00017.1.

Machado D. J., Janies D., Brouwer C., Grant T. (2018) Four new complete mitochondrial genomes of frogs and a new strategy to infer circularity. *Ecology and Evolution*. **8(8)**: 4011-4018. DOI: 10.1002/ece3.3918

In addition to the articles above, we list 4 manuscripts that are under preparation and that have been already presented in international conferences. All the innovative methods and strategies presented in these works were inspired by specific challenges faced during the development of the doctoral thesis:

Schneider A. de. B., Machado D. J., Lambodhar D., Janies D. (2017) Flavivirus Phylogeny Revisited: In search of the Orthologs. (Oral presentation) *5th International Quest for Orthologs Meeting*, **Los Angeles–USA**. Financial support: FAPESP & Department of Genomics and Bioinformatics–UNC Charlotte.

Machado D. J., Castroviejo-Fisher S., Grant T. (2016) Evidence of absence treated as absence of evidence: the effects of gaps in standart maximum likelihood analysis. (Oral presentation) *35th Annual Meeting of the Willi Hennig Society and XII Reunión Argentina de Cladística y Biogeografia*, **Buenos Aires–Argentina**. Financial support: FAPESP.

Machado D. J., Marques F. P. de L., Grant T. (2016) Direct Measures of Support for Maximum Likelihood. (Oral presentation) *35th Annual Meeting of the Willi Hennig Society and XII Reunión Argentina de Cladística y Biogeografia*, **Buenos Aires–Argentina**. Financial support: FAPESP.

Dias P. H. S., Machado D. J. (2014) Phylogenetic analysis of transformation series composed of ordered sequences. (Oral presentation) *X Congreso Latinoamericano de Zoologia*, **Cartajena das Indias–Colombia**. Financial support: FAPESP.

Machado D. J., Marques F. P. de L. (2013) On the use of iterative pass as a refinement strategy. (Oral presentation) *XXXII Willi Hennig Meeting*, **University of Rostock, Rostock–Germany**. Financial support: FAPESP and Willi Hennig Society (Hennig Award and Kurt Milton Pickett Award).

Chapters 2 contains material published in the journal Molecular Ecology Resources. **Chapter 3** includes results from a manuscript accepted in Ecology and Evolution (see above). **Chapters 4** and **5** are will be submitted as a single manuscript on novel phylogenetic markers for amphibians, awaiting the results from tests at the molecular laboratory. **Chapter 6** will soon be submitted as a letter to Molecular Ecology and Evolution. These document contains original material only, and all analyses and result interpretation were performed by the doctoral candidate.

REFERENCES

Alford, R. A., & Richards, S. J. (1999). Global amphibian declines: a problem in applied ecology. *Annual review of Ecology and Systematics*, *30*(1), 133-165.

Alroy, J. (2015). Current extinction rates of reptiles and amphibians. *Proceedings of the National Academy of Sciences*, *112*(42), 13003-13008.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

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

Angerer, K. (2011). Frog tales–on poison dart frogs, epibatidine, and the sharing of biodiversity. *Innovation: The European Journal of Social Science Research*, *24*(3), 353-369.

Anmarkrud, J. A., & Lifjeld, J. T. (2017). Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. *Molecular ecology resources*, *17*(2), 334-341.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11.

Beçak, W., Beçak, M. L., Schreiber, G., Lavalle, D., & Amorim, F. O. (1970). Interspecific variability of DNA content in Amphibia. *Cellular and Molecular Life Sciences*, *26*(2), 204-206.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic acids research*, *45*(Database issue), D37.

Berenbaum, M. R. (1995). The chemistry of defense: theory and practice. *Proceedings of the National Academy of Sciences*, *92*(1), 2-8.

Bernt, M., Braband, A., Schierwater, B., & Stadler, P. F. (2013). Genetic aspects of mitochondrial genome evolution. *Molecular phylogenetics and evolution*, *69*(2), 328-338.

Bertrand, B., Alburaki, M., Legout, H., Moulin, S., Mougel, F., & Garnery, L. (2015). MtDNA COI−COII marker and drone congregation area: An efficient method to establish and monitor honeybee (Apis mellifera L.) conservation centres. *Molecular ecology resources*, *15*(3), 673-683.

Besnard, G., Bertrand, J. A., Delahaie, B., Bourgeois, Y. X., Lhuillier, E., & Thébaud, C. (2015). Valuing museum specimens: high-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura). *Biological Journal of the Linnean Society*, *117*(1), 71-82.

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and genomewise. *Genome research*, *14*(5), 988-995.

Birney, E., & Durbin, R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Research*, *10*(4), 547-548.

Blaustein, A. R., & Kiesecker, J. M. (2002). Complexity in conservation: lessons from the global decline of amphibian populations. *Ecology letters*, *5*(4), 597-608.

Blotto, B. L., Nunez, J. J., Basso, N. G., Ubeda, C. A., Wheeler, W. C., & Faivovich, J. (2013). Phylogenetic relationships of a Patagonian frog radiation, the Alsodes+ Eupsophus clade (Anura: Alsodidae), with comments on the supposed paraphyly of Eupsophus. *Cladistics*, *29*(2), 113-131.

Bókony, V., Üveges, B., Móricz, Á. M., & Hettyey, A. Competition induces increased toxin production in toad larvae without allelopathic effects on heterospecific tadpoles. *Functional Ecology*.

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic acids research*, *27*(8), 1767-1780.

Borodovsky, M., Mills, R., Besemer, J., & Lomsadze, A. (2003). Prokaryotic gene prediction using GeneMark and GeneMark. hmm. *Current protocols in bioinformatics*, 4-5.

Both, C., Lingnau, R., Santos-Jr, A., Madalozzo, B., Lima, L. P., & Grant, T. (2011). Widespread occurrence of the American bullfrog, Lithobates catesbeianus (Shaw, 1802)(Anura: Ranidae), in Brazil. *South American Journal of Herpetology*, *6*(2), 127-134

Boyd, B. M., Allen, J. M., Nguyen, N. P., Sweet, A. D., Warnow, T., Shapiro, M. D., ... & Johnson, K. P. (2017). Phylogenomics using Target-Restricted Assembly Resolves Intrageneric Relationships of Parasitic Lice (Phthiraptera: Columbicola). *Systematic biology*, syx027.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., ... & Chitsaz, H. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 10.

Brendan Maher. (2012, September 6). Fighting about ENCODE and junk [Blog]. Retrieved September 4, 2017, from http://blogs.nature.com/news/2012/09/fighting-about-encode-and-junk.html

Brito, P. H., & Edwards, S. V. (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, *135*(3), 439-455.

Brodie, E. D. (1968). Investigations on the Skin Toxin of the Red-Spotted Newt, Notophthalmus viridescens viridescens. *American Midland Naturalist*, *80*(1), 276–280. doi:10.2307/2423618

Brodie Jr, E. D., Formanowicz Jr, D. R., & Brodie III, E. D. (1991). Predator avoidance and antipredator mechanisms: distinct pathways to survival. *Ethology Ecology & Evolution*, *3*(1), 73-77.

Brodie III, E. D., Feldman, C. R., Hanifin, C. T., Motychak, J. E., Mulcahy, D. G., Williams, B. L., & Brodie Jr, E. D. (2005). Parallel arms races between garter snakes and newts involving tetrodotoxin as the phenotypic interface of coevolution. *Journal of Chemical Ecology*, *31*(2), 343–356. doi:10.1007/s10886-005-1345-x

Brodie Jr, E. D., Formanowicz Jr, D. R., & Brodie III, E. D. (1991). Predator avoidance and antipredator mechanisms: distinct pathways to survival. *Ethology Ecology & Evolution*, *3*(1), 73-77.

Buerk, R. (2012). Fugu: The fish more poisonous than cyanide. *BBC News*. Retrieved from http://www.bbc.com/news/magazine-18065372

Burrell, A. S., Disotell, T. R., & Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *Journal of Human Evolution*, *79*, 35–44. doi:10.1016/j.jhevol.2014.10.015

Cameron, S. L. (2014). How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research. *Systematic Entomology*, *39*(3), 400-411.

Camper, J. D., Ruedas, L. A., Bickham, J. W., & Dixon, J. R. (1993). The relationship of genome size with developmental rates and reproductive strategies in five families of neotropical bufonoid frogs. *Genetics (Life Science Advances)*, *12*, 79–87.

Carstens, B., Lemmon, A. R., & Lemmon, E. M. (2012). The promises and pitfalls of next-generation sequencing data in phylogeography. *Systematic Biology*, *61*(5), 713-715.

Castoe, T. A., Poole, A. W., de Koning, A. J., Jones, K. L., Tomback, D. F., Oyler-McCance, S. J., ... & Pollock, D. D. (2012). Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PloS one*, *7*(2), e30953. doi:10.1371/journal.pone.0030953

Catterall, W. A. (2000). From Ionic Currents to Molecular Mechanisms. *Neuron*, *26*(1), 13–25. doi:10.1016/S0896-6273(00)81133-2

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J. N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, *7*(2), 567–580. doi:10.1093/gbe/evv005

Chevreux, B., Wetter, T., & Suhai, S. (1999, October). Genome sequence assembly using trace signals and additional sequence information. In *German conference on bioinformatics* (Vol. 99, pp. 45-56).

Chu, C., Nielsen, R., & Wu, Y. (2016). REPdenovo: Inferring De Novo repeat motifs from short sequence reads. *PLoS ONE*, *11*(3), 1–17. doi:10.1371/journal.pone.0150719

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., ... & Down, T. (2003). Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Research*, *31*(1), 38-42.

Clark, V. C., Harinantenaina, L., Zeller, M., Ronto, W., Rocca, J., Dossey, A. T., ... & Shaw, C. (2012). An endogenous bile acid and dietary sucrose from skin secretions of alkaloid-sequestering poison frogs. *Journal of natural products*, *75*(3), 473-478. doi:10.1021/np200963r

Collins, J. P., & Halliday, T. (2005). Forecasting changes in amphibian biodiversity: aiming at a moving target. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*(1454), 309-314.

Collins, J. P., & Storfer, A. (2003). Global amphibian declines: sorting the hypotheses. *Diversity and distributions*, *9*(2), 89-98.

Cong, Q., & Grishin, N. V. (2016). The complete mitochondrial genome of *Lerema accius* and its phylogenetic implications. *PeerJ*, *4*, e1546.

Conlon, J. M. (2011). The contribution of skin antimicrobial peptides to the system of innate immunity in anurans. *Cell and tissue research*, *343*(1), 201-212.

Cope, E. D. (1899). Contributions to the herpetology of New Granada and Argentina, with descriptions of new forms. *Philadelphia Museums. Science Bulletin*, *1*, 1–19.

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*(10), 691–703. doi:10.1038/nrg2640

D. Green, & Erin Muths. (2005). Health evaluation of amphibians in and near Rocky Mountain National Park (Colorado, USA). *Alytes*, *22*, 109–129.

Daly, J. W. (2004). Marine toxins and nonmarine toxins: convergence or symbiotic organisms?. *Journal of natural products*, *67*(8), 1211-1215. doi:10.1021/np040016t

Daly, J. W., Garraffo, H. M., & Spande., T. F. (1999). Alkaloids from amphibian skins. In S. W. Pelletier (Ed.), *Alkaloids: chemical and biological perspectives* (pp. 1–161). New York: Pergamon.

Daly, J. W., Garraffo, H. M., Spande, T. F., Decker, M. W., Sullivan, J. P., & Williams, M. (2000). Alkaloids from frog skin: the discovery of epibatidine and the potential for developing novel non-opioid analgesics. *Natural product reports*, *17*(2), 131-135. doi:10.1039/a900728h

Daly, J. W., Myers, C. W., Warnick, J. E., & Albuquerque, E. X. (1980). Levels of batrachotoxin and lack of sensitivity to its action in poison-dart frogs (*Phyllobates*). *Science*, *208*(4450), 1383-1385.

Daly, J. W., Myers, C. W., & Whittaker, N. (1987). Further classification of skin alkaloids from neotropical poison frogs (Dendrobatidae), with a general survey of toxic/noxious substances in the amphibia. *Toxicon*, *25*(10), 1023-1095.
Daly, J. W., Secunda, S. I., Garraffo, H. M., Spande, T. F., Wisnieski, A., & Cover, J. F. (1994). An uptake system for dietary alkaloids in poison frogs (Dendrobatidae). *Toxicon*, *32*(6), 657–663.
Daly, J. W., & Spande, T. F. (1986). Amphibian alkaloids: chemistry, pharmacology, and biology. In *Alkaloids: chemical and biological perspectives* (Vol. 4, pp. 1–274). New York: John Wiley & Sons.
Daly, J. W., Spande, T. F., & Garraffo, H. M. (2005). Alkaloids from amphibian skin: a tabulation of over eight-hundred compounds. *Journal of Natural Products*, *68*(10), 1556–75. doi:10.1021/np0580560
Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement., *5*(6), e11147–e11147. doi:10.1371/journal.pone.0011147
Darrin Hulsey, C., Keck, B. P., Alamillo, H., & O'Meara, B. C. (2013). Mitochondrial genome primers for Lake Malawi cichlids. *Molecular ecology resources*, *13*(3), 347-353. doi:10.1111/1755-0998.12066
David B. Wake, & Vance T. Vredenburg. (2008). Are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proceedings of the National Academy of Sciences*, *105*(1), 11466–11473. doi:10.1073/pnas.0801921105
Dias, I. R., Haddad, C. F. B., Argólo, A. J. S., & Orrico, V. G. D. (2017). The 100th: An appealing new species of Dendropsophus (Amphibia: Anura: Hylidae) from northeastern Brazil. *PLoS ONE*, *12*(3), 1–20. doi:10.1371/journal.pone.0171678
Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. a., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing, *13*(9), 601–612. doi:10.1038/nrg3226
Doležel, J., Bartoš, J., Voglmayr, H., & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry Part A*, *51A*(2), 127–128. doi:10.1002/cyto.a.10013
Donnelly, M. A., & Crump, M. L. (1998). Potential effects of climate change on two neotropical amphibian assemblages. *Climatic Change*, *39*(2), 541–561.
Doolittle, W. F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences*, *110*(14), 5294–5300.
Doronina, L., Churakov, G., Kuritzin, A., Shi, J., Baertsch, R., Clawson, H., & Schmitz, J. (2017). Speciation network in Laurasiatheria: retrophylogenomic signals. *Genome Research*, *27*(6), 997–1003. doi:10.1101/gr.210948.116
Doyle, J. M., Katzner, T. E., Bloom, P. H., Ji, Y., Wijayawardena, B. K., & DeWoody, J. A. (2014). The genome sequence of a widespread apex predator, the golden eagle (Aquila chrysaetos). *PloS one*, *9*(4), e95599. doi:10.1371/journal.pone.0095599
Drummond, A. J., Ho, S. Y., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS biology*, *4*(5), e88. doi:10.1371/journal.pbio.0040088
Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., ... & Nguyen, N. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research*, *21*(12), 2224-2241. doi:10.1101/gr.126599.111
Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. doi:10.1093/nar/gkh340

Edwards, S. V., & Beerli, P. (2000). Perspective: gene divergence, population divergence, and the variance incoalescence time in phylogeographic studies. *Evolution*, *54*(6), 1839-1854.

El Baidouri, M., & Panaud, O. (2013). Comparative Genomic Paleontology across Plant Kingdom Reveals the Dynamics of TE-Driven Genome Evolution. *Genome Biology and Evolution*, *5*(5), 954–965. doi:10.1093/gbe/evt025

Faircloth, B. C. (2017). Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution*, *8*(9), 1103–1112. doi:10.1111/2041-210X.12754

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among hymenoptera. *Molecular Ecology Resources*, *15*(3), 489–501. doi:10.1111/1755-0998.12328

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–726. doi:10.1093/sysbio/sys004

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, *320*(5879), 1034–1039. doi:10.1126/science.1153213

Feldman, C. R., Brodie, E. D., & Pfrender, M. E. (2009). The evolutionary origins of beneficial alleles during the repeated adaptation of garter snakes to deadly prey., *106*(32), 13415–20. doi:10.1073/pnas.0901224106

Feldman, C. R., Brodie, E. D., & Pfrender, M. E. (2010). Genetic architecture of a feeding adaptation: garter snake (Thamnophis) resistance to tetrodotoxin bearing prey. *Proceedings. Biological Sciences / The Royal Society*, *277*(1698), 3317–25. doi:10.1098/rspb.2010.0748

Ferreira, R. B., Faivovich, J., Beard, K. H., & Pombal, J. P. (2015). The First Bromeligenous Species of Dendropsophus (Anura: Hylidae) from Brazil's Atlantic Forest. *PLoS ONE*, *10*(12). doi:10.1371/journal.pone.0142893

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397–405. doi:10.1038/nrg2337

Flórez-Rodríguez, A., Carling, M. D., & Cadena, C. D. (2011). Reconstructing the phylogeny of "Buarremon" brush-finches and near relatives (Aves, Emberizidae) from individual gene trees. *Molecular Phylogenetics and Evolution*, *58*(2), 297-303. doi:10.1016/j.ympev.2010.11.012

Ford, L. S., & Cannatella, D. C. (1993). The Major Clades of Frogs. *Herpetological Monographs*, *7*, 94–117.

Ford, L. S., & Cannatella, D. C. (1993). The major clades of frogs. *Herpetological Monographs*, 94-117.

Franca, L. T., Carrilho, E., & Kist, T. B. (2002). A review of DNA sequencing techniques. *Quarterly reviews of biophysics*, *35*(2), 169-200. doi:10.1017/S0033583502003797

Frost, D. R., Grant, T., & Mendelson III, J. R. (2006). Ollotis Cope, 1875 is the oldest name for the genus currently referred to as Cranopsis Cope, 1875 (Anura: Hyloides: Bufonidae). Ollotis Cope, 1875 es el más viejo nombre para el género designado actualmente como Cranopsis Cope, 1875 (Anura: Hyloides: Bufonidae). *Copeia.*, *2006*(3), 558.

Frost, D. R. (2017). Amphibian Species of the World: An Online Reference [Database]. Retrieved September 4, 2017, from http://research.amnh.org/herpetology/amphibia/index.html
Frost, D. R. (2015). Amphibian Species of the World: An Online Reference [Database]. Retrieved in 2015 from http://research.amnh.org/herpetology/amphibia/index.html
Frost, D. R., Grant, T., Faivovich, J., Bain, R. H., Haas, A., Haddad, C. F., ... & Raxworthy, C. J. (2006). The amphibian tree of life. *Bulletin of the American Museum of natural History*, 1-291. doi:10.1206/0003-0090(2006)297
Gallus, S., Janke, A., Kumar, V., & Nilsson, M. A. (2015). Disentangling the Relationship of the Australian Marsupial Orders Using Retrotransposon and Evolutionary Network Analyses. *Genome Biology and Evolution*, *7*(4), 985–992. doi:10.1093/gbe/evv052
Gan, H. M., Schultz, M. B., & Austin, C. M. (2014). Integrated shotgun sequencing and bioinformatics pipeline allows ultra-fast mitogenome recovery and confirms substantial gene rearrangements in Australian freshwater crayfishes. *BMC evolutionary biology*, *14*(1), 19. doi:10.1186/1471-2148-14-19
Garnatje, T., Peñuelas, J., & Vallès, J. (2017). Ethnobotany, Phylogeny, and 'Omics' for Human Health and Food Security. *Trends in Plant Science*, *22*(3), 187–191. doi:10.1016/j.tplants.2017.01.001
Garnett, S. T., & Christidis, L. (2017a). Taxonomy anarchy hampers conservation. *Nature*, *546*(7656), 25–27. doi:10.1038/546025a
Geffeney, S., Brodie, E. D., & Ruben, P. C. (2002). Mechanisms of adaptation in a predator-prey arms race: TTX-resistant sodium channels. *Science*, *297*(5585), 1336-1339. doi:10.1126/science.1074310
Geffeney, S. L., Fujimoto, E., Brodie, E. D., & Ruben, P. C. (2005). Evolutionary diversification of TTX-resistant sodium channels in a predator–prey interaction. *Nature*, *434*(7034), 759-763. doi:10.1038/nature03392
Geffeney, S. L., & Ruben, C. (2006). The structural basis and functional consequences of interactions between tetrodotoxin and voltage-gated sodium channels. *Marine Drugs*, *4*(3), 143–156.
Giancarlo, R., Scaturro, D., & Utro, F. (2009). Textual data compression in computational biology: a synopsis. *Bioinformatics*, *25*(13), 1575-1586. doi:10.1093/bioinformatics/btp117
Gissi, C., Iannelli, F., & Pesole, G. (2008). Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity*, *101*(4), 301-320. doi:10.1038/hdy.2008.62
Goin, O. B., Goin, C. J., & Bachmann, K. (1968). DNA and amphibian life history. *Copeia*, *1968*(3), 532–540.
Goloboff, P. A., Farris, J. S., & Nixon, K. C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics*, *24*(5), 774-786. doi:10.1111/j.1096-0031.2008.00217.x
Grant, T., Colombo, P., Verrastro, L., & Saporito, R. A. (2012). The occurrence of defensive alkaloids in non-integumentary tissues of the Brazilian red-belly toad *Melanophryniscus simplex* (Bufonidae). *Chemoecology*, *22*(3), 169-178. doi:10.1007/s00049-012-0107-9
Grant, T., Frost, D. R., Caldwell, J. P., Gagliardo, R. O. N., Haddad, C. F., Kok, P. J., ... & Wheeler, W. C. (2006). Phylogenetic systematics of dart-poison frogs and their relatives (Amphibia: Athesphatanura: Dendrobatidae). *Bulletin of the American Museum of natural History*, 1-262. doi:http://dx.doi.org/10.1206/0003-0090(2006)299[1:PSODFA]2.0.CO;2

Grant, T., & Kluge, A. G. (2009). Perspective: Parsimony, explanatory power, and dynamic homology testing. *Systematics and Biodiversity*, *7*(4), 357-363. doi:10.1017/S147720000999017X

Grant, T., Rada, M., Anganoy-Criollo, M., Batista, A., Dias, P. H., Jeckel, A. M., … Rueda-Almonacid, J. V. (2017). Phylogenetic systematics of dart-poison frogs and their relatives revisited (Anura: Dendrobatoidea). *South American Journal of Herpetology*, *12*(s1), S1–S90.

Grau, J. H., Nuñez, J. J., Plötner, J., & Poustka, A. J. (2016). The complete mitochondrial genome of *Telmatobufo australis* (Amphibia: Anura: Calyptocephalellidae). *Mitochondrial DNA Part A*, *27*(4), 2817-2818. doi:10.3109/19401736.2015.1053082

Graur, D. (2017). An upper limit on the functional fraction of the human genome. *Genome Biology and Evolution*, 1–21. doi:10.1093/gbe/evx121

Gregory, T. R. (2017). Animal Genome Database [Database]. Retrieved September 19, 2017, from http://www.genomesize.com.

Greilhuber, J. (2005). The Origin, Evolution and Proposed Stabilization of the Terms "Genome Size" and "C-Value" to Describe Nuclear DNA Contents. *Annals of Botany*, *95*(1), 255–260. doi:10.1093/aob/mci019

Greilhuber, J. (2008). Cytochemistry and C-values: The Less-well-known World of Nuclear DNA Amounts. *Annals of Botany*, *101*(6), 791–804. doi:10.1093/aob/mcm250

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic acids research*, *41*(13), e129-e129. doi:10.1093/nar/gkt371

Halliday, T. R. (2008). Why amphibians are important. *International Zoo Yearbook*, *42*(1), 7-14. doi:10.1111/j.1748-1090.2007.00037.x

Hancock−Hanser, B. L., Frey, A., Leslie, M. S., Dutton, P. H., Archer, F. I., & Morin, P. A. (2013). Targeted multiplex next−generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, *13*(2), 254-268. doi:10.1111/1755-0998.12059

Hanifin, C. T. (2010). The chemical and evolutionary ecology of tetrodotoxin (TTX) toxicity in terrestrial vertebrates. *Marine drugs*, *8*(3), 577-593. doi:10.3390/md8030577

Hanifin, C. T., & Brodie, E. D. (2008). Phenotypic mismatches reveal escape from arms-race coevolution. *PLoS Biology*, *6*(3), e60–e60. doi:10.1371/journal.pbio.0060060

Hanifin, C. T., & Gilly, W. F. (2014). Evolutionary history of a complex adaptation: Tetrodotoxin resistance in salamanders. *Evolution; International Journal of Organic Evolution*, 1–38. doi:10.1111/evo.12552

Hantak, M. M., Grant, T., Reinsch, S., Mcginnity, D., Loring, M., Toyooka, N., & Saporito, R. A. (2013). Dietary alkaloid sequestration in a poison frog: an experimental test of alkaloid uptake in *Melanophryniscus stelzneri* (Bufonidae). *Journal of chemical ecology*, *39*(11-12), 1400-1406. doi:10.1007/s10886-013-0361-5

Hardison, R. C. (2003). Comparative genomics. *PLoS biology*, *1*(2), e58. doi:10.1371/journal.pbio.0000058

Harlan, R. (1835). *Medical and Physical Researches, Or Original Memoirs in Medicine: Surgery, Physiology, Geology, Zoology and Comparative Anatomy...* LR Bailey.

Hellsten, U., Harland, R. M., Gilchrist, M. J., Hendrix, D., Jurka, J., Kapitonov, V., … Rokhsar, D. S. (2010). The genome of the Western clawed frog *Xenopus tropicalis*. *Science*, *328*(5978), 633–636. doi:10.1126/science.1183670

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., & Schuster, S. C. (2004). Whole-genome prokaryotic phylogeny. *Bioinformatics*, *21*(10), 2329-2335. doi:10.1093/bioinformatics/bth324

Herniou, E. A., Luque, T., Chen, X., Vlak, J. M., Winstanley, D., Cory, J. S., & O'Reilly, D. R. (2001). Use of whole genome sequence data to infer baculovirus phylogeny. *Journal of virology*, *75*(17), 8117-8126. doi:10.1128/JVI.75.17.8117

Heyer, W. R. (2003). Ultraviolet-B and amphibia. *AIBS Bulletin*, *53*(6), 540-541. doi:10.1641/0006-3568(2003)053[0540:UAA]2.0.CO;2

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., … Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology*, *65*(4), 726–736. doi:10.1093/sysbio/syw021

Hollister, J. D., & Gaut, B. S. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research*, *19*(8), 1419–1428. doi:10.1101/gr.091678.109

Horner, H. A., & Macgregor, H. C. (1983). C value and cell volume: their significance in the evolution and development of amphibians. *Journal of Cell Science*, *63*, 135–146.

Houlahan, J. E., Findlay, C. S., Schmidt, B. R., Meyer, A. H., & Kuzmin, S. L. (2000). Quantitative evidence for global amphibian population declines. *Nature*, *404*(6779), 752-755. doi:10.1038/35008052

Hykin, S. M., Bi, K., & McGuire, J. A. (2015). Fixing Formalin: A Method to Recover Genomic-Scale DNA Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing. *Plos One*, *10*(10), e0141579. doi:10.1371/journal.pone.0141579

Indraneil Das, & Alexander Haas. (2010). New species of Microhyla from Sarawak: Old World's smallest frogs crawl out of miniature pitcher plants on Borneo (Amphibia: Anura: Microhylidae). *Zootaxa*, *2571*, 37–52.

"International Human Genome Sequencing Consortium," & others. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945.

Irisarri, I., San Mauro, D., Abascal, F., Ohler, A., Vences, M., & Zardoya, R. (2012). The origin of modern frogs (Neobatrachia) was accompanied by acceleration in mitochondrial and nuclear substitution rates. *BMC genomics*, *13*(1), 626. doi:10.1186/1471-2164-13-626

Johnston, J. S., Yoon, K. S., Strycharz, J. P., Pittendrigh, B. R., & Clark, J. M. (2007). Body Lice and Head Lice (Anoplura: Pediculidae) Have the Smallest Genomes of Any Hemimetabolous Insect Reported to Date. *Journal of Medical Entomology*, *44*(6), 1009–1012. doi:10.1603/0022-2585(2007)44[1009:BLAHLA]2.0.CO;2

Jones, J. M., & Gellert, M. (2004). The taming of a transposon: V(D)J recombination and the immune system. *Immunological Reviews*, *200*(1), 233–248. doi:10.1111/j.0105-2896.2004.00168.x

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185-202. doi:10.1111/mec.13304

Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. *Current opinion in structural biology*, *8*(3), 333-337. doi:10.1016/S0959-440X(98)80067-5

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends in genetics*, *16*(9), 418-420. doi:10.1016/S0168-9525(00)02093-X

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1–4), 462–467. doi:10.1159/000084979

Katoh, K., Kuma, K., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, *33*(2), 511–8. doi:10.1093/nar/gki198

Kazazian, H. H. (2004). Mobile elements: drivers of genome evolution. S*cience*, *303*(5664), 1626-1632. doi:10.1126/science.1089670

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... & Thierer, T. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647-1649. doi:10.1093/bioinformatics/bts199

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. doi:10.1101/gr.229202

Kido, Y., Aono, M., Yamaki, T., Matsumoto, K., Murata, S., Saneyoshi, M., & Okada, N. (1991). Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. *Proceedings of the National Academy of Sciences*, *88*(6), 2326–2330. doi:10.1073/pnas.88.6.2326

Kim, K. C., & Byrne, L. B. (2006). Biodiversity loss and the taxonomic bottleneck: emerging biodiversity science. *Ecological Research*, *21*(6), 794–810. doi:10.1007/s11284-006-0035-7

Kluge, A. G., & Grant, T. (2006). From conviction to anti−superfluity: old and new justifications of parsimony in phylogenetic inference. *Cladistics*, *22*(3), 276-288.

Kobayashi, S. (2004). Retrotransposon-Induced Mutations in Grape Skin Color. *Science*, *304*(5673), 982–982. doi:10.1126/science.1095011

König, E., Bininda-Emonds, O. R., & Shaw, C. (2015). The diversity and evolution of anuran skin peptides. *Peptides*, *63*, 96-117. doi:10.1016/j.peptides.2014.11.003

König, E., Zhou, M., Wang, L., Chen, T., Bininda-Emonds, O. R., & Shaw, C. (2012). Antimicrobial peptides and alytesin are co-secreted from the venom of the Midwife toad, Alytes maurus (Alytidae, Anura): Implications for the evolution of frog skin defensive secretions. *Toxicon*, *60*(6), 967-981. doi:10.1016/j.toxicon.2012.06.015

Ku, C. S., & Roukos, D. H. (2013). From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. *Expert review of medical devices*, *10*(1), 1-6. doi:10.1586/erd.12.63

Kurabayashi, A., & Sumida, M. (2013). Afrobatrachian mitochondrial genomes: genome reorganization, gene rearrangement mechanisms, and evolutionary trends of duplicated and rearranged genes. *BMC genomics*, *14*(1), 633. doi:10.1186/1471-2164-14-633

Kurabayashi, A., Sumida, M., Yonekawa, H., Glaw, F., Vences, M., & Hasegawa, M. (2008). Phylogeny, recombination, and mechanisms of stepwise mitochondrial genome reorganization in mantellid frogs from Madagascar. *Molecular Biology and Evolution*, *25*(5), 874-891. doi:10.1093/molbev/msn031

Lambertz, M. (2017). Taxonomy: retain scientific autonomy. *Nature*, *546*(7660), 600–600. doi:10.1038/546600b

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359. doi:10.1038/nmeth.1923

Langmead, B., & Salzberg, S. L. (2012b). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. doi:10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, *10*(3), R25. doi:10.1186/gb-2009-10-3-r25

Larkin, M. a, Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. a, McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947–8. doi:10.1093/bioinformatics/btm404

Laslett, D., & Canbäck, B. (2007). ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics*, *24*(2), 172-175. doi:10.1093/bioinformatics/btm573

Lehr, E., & Catenazzi, A. (2009). A New Species of Minute Noblella (Anura: Strabomantidae) from Southern Peru: The Smallest Frog of the Andes. *Copeia*, *2009*(1), 148–156. doi:10.1643/CH-07-270

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Systematic Biology*, *61*(5), 727–744. doi:10.1093/sysbio/sys049

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. doi:10.1093/bioinformatics/btl158

Li, W., Jaroszewski, L., & Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, *17*(3), 282–283.

Licht, L. E. (2003). Shedding light on ultraviolet radiation and amphibian embryos. *BioScience*, *53*(6), 551-561. doi:10.1641/0006-3568(2003)053[0551:SLOURA]2.0.CO;2

Liu, J., Zhang, D., Li, Y., Chen, W., Ruan, Z., Deng, L., ... & Luo, H. (2013). Discovery of bufadienolides as a novel class of ClC-3 chloride channel activators with antitumor activities. *Journal of medicinal chemistry*, *56*(14), 5734-5743. doi:10.1021/jm400881m

Liu, X., & Li, Y. (2009). Aquaculture Enclosures Relate to the Establishment of Feral Populations of Introduced Species. *PLoS ONE*, *4*(7), e6199. doi:10.1371/journal.pone.0006199

Lounsberry, Z. T., Brown, S. K., Collins, P. W., Henry, R. W., Newsome, S. D., & Sacks, B. N. (2015). Next−generation sequencing workflow for assembly of nonmodel mitogenomes exemplified with North Pacific albatrosses (Phoebastria spp.). *Molecular ecology resources*, *15*(4), 893-902. doi:10.1111/1755-0998.12365

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, *25*(5), 955-964. doi:10.1093/nar/25.5.955

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., ... & Tang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, *1*(1), 18. doi:10.1186/2047-217X-1-18

Machado, D. J., Lyra, M. L., & Grant, T. (2016). Mitogenome assembly from genomic multiplex libraries: comparison of strategies and novel mitogenomes for five species of frogs. *Molecular Ecology Resources*, *16*(3), 686–693. doi:10.1111/1755-0998.12492

Magno V. Segalla, Ulisses Caramaschi, Carlos Alberto Gonçalves Cruz, Taran Grant, Célio F.B. Haddad, Paulo Christiano de Anchietta Garcia, … José A. Langone. (2016). Brazilian Amphibians: List of Species. *Herpetologia Brasileira*, *5*(2), 34–46.

Mäkinen, V., Salmela, L., & Ylinen, J. (2012). Normalized N50 assembly metric using gap-restricted co-linear chaining. *BMC Bioinformatics*, *13*(1), 255–255. doi:10.1186/1471-2105-13-255

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., … Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, *7*(2), 111–118. doi:10.1038/nmeth.1419

Marçais, G., Yorke, J. A., & Zimin, A. (2015). QuorUM: An Error Corrector for Illumina Reads. *PLOS ONE*, *10*(6), e0130821. doi:10.1371/journal.pone.0130821

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, *9*, 387-402. doi:10.1146/annurev.genom.9.081307.164359

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual review of analytical chemistry*, *6*, 287-303. doi:10.1146/annurev-anchem-062012-092628

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560-564.

Mccallum, M. L. (2007). Amphibian decline or extinction? Current declines dwarf background extinction rate. *Journal of Herpetology*, *41*(3), 483–491.

McCormack, J. E., & Faircloth, B. C. (2013). Next-generation phylogenetics takes root. *Molecular Ecology*, *22*(1), 19–21. doi:10.1111/mec.12050

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012a). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis, *22*(4), 746–754. doi:10.1101/gr.125864.111

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2012b). Molecular Phylogenetics and Evolution Applications of next-generation sequencing to phylogeography and phylogenetics. doi:10.1016/j.ympev.2011.12.007

McGuigan, K., & Sgrò, C. M. (2009). Evolutionary consequences of cryptic genetic variation. *Trends in Ecology & Evolution*, *24*(6), 305–311. doi:10.1016/j.tree.2009.02.001

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, *11*(1), 31-46. doi:10.1038/nrg2626

Miranda-Ribeiro, A. de. (1920). Os brachycephalus do Museu Paulista (com tres espeies novas). *Revista Do Museu Paulista*, *12*, 307–316.

Moczydlowski, E. G. (2013). The molecular mystique of tetrodotoxin. *Toxicon*, *63*, 165–183. doi:10.1016/j.toxicon.2012.11.026

Mueller, R. L., & Boore, J. L. (2005). Molecular mechanisms of extensive mitochondrial gene rearrangement in plethodontid salamanders. *Molecular Biology and Evolution*, *22*(10), 2104-2112. doi:10.1093/molbev/msi204

Mueller, R. L., Macey, J. R., Jaekel, M., Wake, D. B., & Boore, J. L. (2004). Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(38), 13820-13825. doi:10.1073/pnas.0405785101

Mullis, K. B., & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, *155*, 335–350.

Mullis, K., Faloona, F., Scharf, S., Saiki, R. K., Horn, G. T., & Erlich, H. (1986, January). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 51, pp. 263-273). Cold Spring Harbor Laboratory Press.

Murata, S., Takasaki, N., Saitoh, M., & Okada, N. (1993). Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs)

as temporal landmarks of evolution. *Proceedings of the National Academy of Sciences*, *90*(15), 6995–6999.
Myers, C. W., Daly, J. W., & Malkin, B. (1978). A dangerously toxic new frog (Phyllobates) used by Emberá Indians of western Colombia, with discussion of blowgun fabrication and dart poisoning. Bulletin of the AMNH; v. 161, article 2.
Nalbantoglu, Ö. U., Russell, D. J., & Sayood, K. (2009). Data compression concepts and algorithms and their applications to bioinformatics. *Entropy*, *12*(1), 34-52. doi:10.3390/e12010034
Narahashi, T. (1972). Mechanism of action of tetrodotoxin and saxitoxin on excitable membranes. In *Federation proceedings* (Vol. 31, p. 1124).
Narzisi, G., & Mishra, B. (2011). Comparing de novo genome assembly: the long and short of it. *PloS One*, *6*(4), e19175–e19175. doi:10.1371/journal.pone.0019175
Nekrutenko, A., & Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, *13*(9), 667-672. doi:10.1038/nrg3305
Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. doi:10.1093/molbev/msu300
Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., & Barron, A. E. (2011). Landscape of next-generation sequencing technologies. *Analytical chemistry*, *83*(12), 4327-4341. doi:10.1021/ac2010857
O'Connell, J., Schulz-Trieglaff, O., Carlson, E., Hims, M. M., Gormley, N. A., & Cox, A. J. (2015). NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics*, *31*(12), 2035-2037. doi:10.1093/bioinformatics/btv057
Olmo, E. (1973). Quantitative variations in the nuclear DNA and phylogenesis of the Amphibia. *Caryologia*, *25*(April 2015), 43–68. doi:10.1080/00087114.1973.10796525
Orrico, V. G. D., Peloso, P. L. V., Sturaro, M. J., Da Silva-Filho, H. F., Neckel-Oliveira, S., Gordo, M., … Haddad, C. F. B. (2014). A new "Bat-Voiced" species of Dendropsophus Fitzinger, 1843 (Anura, Hylidae) from the Amazon Basin, Brazil. *Zootaxa*, *3881*(4), 341–361. doi:10.11646/zootaxa.3881.4.3
Padial, J. M., Grant, T., & Frost, D. R. (2014). Molecular systematics of terraranas (Anura: Brachycephaloidea) with an assessment of the effects of alignment and optimality criteria. *Zootaxa*, *3825*(1), 1-132. doi:10.11646/zootaxa.3825.1.1
Parrington, J. (2015). *The deeper genome: why there is more to the human genome than meets the eye* (First edition). Oxford, United Kingdom: Oxford University Press.
Peloso, P. L., Faivovich, J., Grant, T., Gasparini, J. L., & Haddad, C. F. (2012). An extraordinary new species of Melanophryniscus (Anura, Bufonidae) from southeastern Brazil.
Peloso, P. L. V., Frost, D. R., Richards, S. J., Rodrigues, M. T., Donnellan, S., Matsui, M., … Wheeler, W. C. (2016). The impact of anchored phylogenomics and taxon sampling on phylogenetic inference in narrow-mouthed frogs (Anura, Microhylidae). *Cladistics*, *32*(2), 113–140. doi:10.1111/cla.12118
Peng, M. S., Fan, L., Shi, N. N., Ning, T., Yao, Y. G., Murphy, R. W., ... & Zhang, Y. P. (2015). DomeTree: a canonical toolkit for mitochondrial DNA analyses in domesticated animals. *Molecular ecology resources*, *15*(5), 1238-1242. doi:10.1111/1755-0998.12386
Pinheiro, P. D., Pezzuti, T. L., Leite, F. S., Garcia, P. C., Haddad, C. F., & Faivovich, J. (2016). A New Species of the *Hypsiboas pulchellus* Group from the Serra da

Mantiqueira, Southeastern Brazil (Amphibia: Anura: Hylidae). *Herpetologica*, *72*(3), 256-270. doi:10.1655/Herpetologica-D-15-00062.1

Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J. P., & Hyvönen, J. (2013). Advances in plant gene-targeted and functional markers: a review. *Plant Methods*, *9*(1), 6. doi:10.1186/1746-4811-9-6

Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular Ecology Resources*, *16*(5), 1069–1083. doi:10.1111/1755-0998.12541

Pukala, T. L., Bowie, J. H., Maselli, V. M., Musgrave, I. F., & Tyler, M. J. (2006). Host-defence peptides from the glandular secretions of amphibians: structure and activity. *Natural product reports*, *23*(3), 368-393. doi:10.1039/b512118n

Pyron, A. R., & Wiens, J. J. (2011). A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, *61*(2), 543–583. https://doi.org/10.1016/j.ympev.2011.06.012

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. doi:10.1093/bioinformatics/btq033

Raspotnig, G., Norton, R. a, & Heethoff, M. (2011). Oribatid mites and skin alkaloids in poison frogs. *Biology Letters*, *7*(4), 555–556; discussion 557. doi:10.1098/rsbl.2010.1113

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, *13*(5), 278-289. doi:10.1016/j.gpb.2015.08.002

Richards, S. (2015). It's more than stamp collecting: how genome sequencing can unify biological research. *Trends in Genetics*, *31*(7), 411–421. doi:10.1016/j.tig.2015.04.007

Rivas, L., Luque-Ortega, J. R., & Andreu, D. (2009). Amphibian antimicrobial peptides and Protozoa: Lessons from parasites, *1788*(8), 1570–1581. doi:10.1016/j.bbamem.2008.11.002

Rodríguez, A., Burgon, J. D., Lyra, M., Irisarri, I., Baurain, D., Blaustein, L., … Vences, M. (2017). Inferring the shallow phylogeny of true salamanders (*Salamandra*) by multiple phylogenomic approaches. *Molecular Phylogenetics and Evolution*, *115*, 16–26. doi:10.1016/j.ympev.2017.07.009

Rokas, A., & Holland, P. W. H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution*, *15*(11), 454–459. doi:10.1016/S0169-5347(00)01967-4

Roseghini, M., Erspamer, V., & Endean, R. (1976). Indole-, imidazole-and phenyl-alkylamines in the skin of one hundred amphibian species from Australia and Papua New Guinea. *Comparative Biochemistry and Physiology Part C: Comparative Pharmacology*, *54*(1), 31-43. doi:10.1016/0306-4492(76)90022-8

Rosenblum, E. B., Voyles, J., Poorten, T. J., & Stajich, J. E. (2010). The deadly chytrid fungus: a story of an emerging pathogen. *PLoS Pathogens*, *6*(1), e1000550. doi:10.1371/journal.ppat.1000550

Biju, S. D., Van Bocxlaer, I., Giri, V. B., Roelants, K., Nagaraju, J., & Bossuyt, F. (2007). A new nightfrog, Nyctibatrachus minimus sp. nov.(Anura: Nyctibatrachidae): the smallest frog from India. *Current Science*, 854-858.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., & Reyes, A. (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*, *238*(1), 195-209. doi:10.1016/S0378-1119(99)00270-X

San Mauro, D., Gower, D. J., Zardoya, R., & Wilkinson, M. (2005). A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. *Molecular Biology and Evolution*, *23*(1), 227-234. doi:10.1093/molbev/msj025

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, *94*(3), 441IN19447-446IN20448. doi:10.1016/0022-2836(75)90213-2

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, *74*(12), 5463-5467. doi:10.1073/pnas.74.12.5463

Sanggaard, K. W., Bechsgaard, J. S., Fang, X., Duan, J., Dyrlund, T. F., Gupta, V., … Wang, J. (2014). Spider genomes provide insight into composition and evolution of venom and silk. *Nature Communications*, *5*(May), 3765. doi:10.1038/ncomms4765

Santos, E. S., Jamieson, I. G., Santos, L. L., & Nakagawa, S. (2013). Low genetic and morphological differentiation between an introduced population of dunnocks in New Zealand and an ancestral population in England. *Biological invasions*, *15*(1), 185-197. doi:10.1007/s10530-012-0278-x

Santos, J. C., Tarvin, R. D., & O'Connell, L. A. (2016). A review of chemical defense in poison frogs (Dendrobatidae): Ecology, pharmacokinetics, and autoresistance. In B. A. Schulte, T. E. Goodwin, & M. H. Ferkin (Eds.), *Chemical Signals in Vertebrates* (pp. 305–337). Retrieved from http://link.springer.com/10.1007/978-0-387-73945-8

Santos, R. R., Leonardi, S. B., Caorsi, V. Z., & Grant, T. (2010). Directional orientation of migration in an aseasonal explosive-breeding toad from Brazil. *Journal of Tropical Ecology*, *26*(4), 415-421. doi:10.1017/S0266467410000180

Saporito, R. A., Donnelly, M. A., Spande, T. F., & Garraffo, H. M. (2012). A review of chemical ecology in poison frogs. *Chemoecology*, *22*(3), 159-168. doi:10.1007/s00049-011-0088-0

Saporito, R. A., Norton, R. A., Andriamaharavo, N. R., Garraffo, H. M., & Spande, T. F. (2011). Alkaloids in the mite Scheloribates laevigatus: further alkaloids common to oribatid mites and poison frogs. *Journal of chemical ecology*, *37*(2), 213-218. doi:10.1007/s10886-011-9914-7

Saporito, R. A., Spande, T. F., Garraffo, H. M., & Donnelly, M. A. (2009). Arthropod alkaloids in poison frogs: a review of the dietary hypothesis. *Heterocycles*, *79*(1), 277-297. doi:10.3987/REV-08-SR(D)11

Saslis-Lagoudakis, C. H., Savolainen, V., Williamson, E. M., Forest, F., Wagstaff, S. J., Baral, S. R., … Hawkins, J. A. (2012). Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proceedings of the National Academy of Sciences*, *109*(39), 15835–15840. doi:10.1073/pnas.1202242109

Savitzky, A. H., & Saporito, R. A. (2012). Sequestration of defensive toxins by tetrapod vertebrates: contributions in memory of John W. Daly. doi:10.1007/s00049-012-0111-0

Schaack, S., Gilbert, C., & Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution*, *25*(9), 537–546. doi:10.1016/j.tree.2010.06.001

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing., *19*(R2), R227–40. doi:10.1093/hmg/ddq416

Schattner, P., Brooks, A. N., & Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs., *33*(Web Server issue), W686–9. doi:10.1093/nar/gki366

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., … Wilson, R. K. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, *326*(5956), 1112–1115. doi:10.1126/science.1178534

Schneider, A. de B., Machado, D. J., Lambodhar Damodaran, & Daniel Janies. (2017). Flavivirus Phylogeny Revisited: In search of the Orthologs. Presented at the Quest For Orthologs conference, Los Angeles.

Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., … Veenstra, G. J. C. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature*, *538*(7625), 1–15. doi:10.1038/nature19840

Sexsmith, L. E. (1968). *DNA values and karyotypes of amphibians*.

Shedlock, A. M., & Okada, N. (2000). SINE insertions: powerful tools for molecular systematics. *BioEssays*, *22*(2), 148–160. doi:10.1002/(SICI)1521-1878(200002)22:2<148::AID-BIES6>3.0.CO;2-Z

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M., Sima, F. A., … Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. doi:10.1093/bioinformatics/btv351

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, *19*(6), 1117-1123. doi:10.1101/gr.089532.108

Sims, G. E., Jun, S. R., Wu, G. A., & Kim, S. H. (2009). Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences*, *106*(40), 17077-17082. doi:10.1073/pnas.0909377106

Smith, J. V., Braun, E. L., & Kimball, R. T. (2012). Ratite nonmonophyly: independent evidence from 40 novel loci. *Systematic Biology*, *62*(1), 35-49. doi:10.1093/sysbio/sys067

Smith, W. L., & Wheeler, W. C. (2006). Venom evolution widespread in fishes: A phylogenetic road map for the bioprospecting of piscine venoms. *Journal of Heredity*, *97*(3), 206–217. doi:10.1093/jhered/esj034

Snel, B., Bork, P., & Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nature genetics*, *21*(1), 108-110. doi:10.1038/5052

Soorni, A., Haak, D., Zaitlin, D., & Bombarely, A. (2017). Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC genomics*, *18*(1), 49. doi:10.1186/s12864-016-3412-9

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688-2690. doi:10.1093/bioinformatics/btl446

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., ... & Coulson, A. (2003). The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS biology*, *1*(2), e45. doi:10.1371/journal.pbio.0000045

Strichartz, G., Rando, T., & Wang, G. K. (1987). An integrated view of the molecular toxinology of sodium channel gating in excitable cells. *Annual Review of Neuroscience*, *10*, 237–267. doi:10.1146/annurev.neuro.10.1.237

Stuart, S. N. (2004). Status and Trends of Amphibian Declines and Extinctions Worldwide. *Science*, *306*(5702), 1783–1786. doi:10.1126/science.1103538

Stynoski, J. L., Shelton, G., & Stynoski, P. (2014a). Maternally derived chemical defences are an effective deterrent against some predators of poison frog tadpoles (Oophaga pumilio). *Biology letters*, *10*(5), 20140187. doi:10.1098/rsbl.2014.0187

Stynoski, J. L., Torres-Mendoza, Y., Sasa-Marin, M., & Saporito, R. A. (2014b). Evidence of maternal provisioning of alkaloid−based chemical defenses in the strawberry poison frog *Oophaga pumilio*. *Ecology*, *95*(3), 587-593. doi:10.1890/13-0927.1

Sumida, M., Kanamori, Y., Kaneda, H., Kato, Y., Nishioka, M., Hasegawa, M., & Yonekawa, H. (2001). Complete nucleotide sequence and gene rearrangement of the mitochondrial genome of the Japanese pond frog *Rana nigromaculata*. *Genes & genetic systems*, *76*(5), 311-325. doi:10.1266/ggs.76.311

Lee, M. Y., Jeon, H. S., Min, M. S., & An, J. (2017). Sequencing and analysis of the complete mitochondrial genome of Hyla suweonensis (Anura: Hylidae). *Mitochondrial DNA Part B*, *2*(1), 126-127. doi:10.3109/19401736.2015.1122767

Sun, Y.-B., Xiong, Z.-J., Xiang, X.-Y., Liu, S.-P., Zhou, W.-W., Tu, X.-L., … Zhang, Y.-P. (2015). Whole-genome sequence of the Tibetan frog Nanorana parkeri and the comparative evolution of tetrapod genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), E1257-62. doi:10.1073/pnas.1501764112

Taly, V., Pekin, D., Abed, A. E., & Laurent-Puig, P. (2012). Detecting biomarkers with microdroplet technology. *Trends in Molecular Medicine*, *18*(7), 405–416. doi:10.1016/j.molmed.2012.05.001

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, (SUPPL. 25). doi:10.1002/0471250953.bi0410s25

Tarvin, R. D., Santos, J. C., O'Connell, L. A., Zakon, H. H., & Cannatella, D. C. (2016). Convergent substitutions in a sodium channel suggest multiple origins of toxin resistance in poison frogs. *Molecular Biology and Evolution*, *33*(4), 1068–1081. doi:10.1093/molbev/msv350

"The ENCODE Project Consortium." (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74. doi:10.1038/nature11247

Thomas, C. A. (1971). the Genetic Organization of Chromosomes 3027. *Annual Review of Genetics*, *5*(1), 237–256.

Toledo, R. D., & Jared, C. (1995). Cutaneous granular glands and amphibian venoms. *Comparative Biochemistry and Physiology Part A: Physiology*, *111*(1), 1-29. doi:10.1016/0300-9629(95)98515-I

Toyooka, N., Okumura, M., & Nemoto, H. (2002). Stereodivergent process for the synthesis of the decahydroquinoline type of dendrobatid alkaloids. *The Journal of organic chemistry*, *67*(17), 6078-6081. doi:10.1021/jo025929b

Trott, O., & Olson, A. J. (2009). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, NA-NA. doi:10.1002/jcc.21334

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115–e115. doi:10.1093/nar/gks596

Vacher, J.-P., Fouquet, A., Holota, H., & Thébaud, C. (2016). The complete mitochondrial genome of *Anomaloglossus baeobatrachus* (Amphibia: Anura: Aromobatidae), *1*(1), 338–340. doi:10.1080/23802359.2016.1172053

Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in biotechnology*, *27*(9), 522-530. doi:10.1016/j.tibtech.2009.05.006

Veltman, J. A., & Brunner, H. G. (2012). *De novo* mutations in human genetic disease. *Nature Reviews Genetics*, *13*(8), 565-575. doi:10.1038/nrg3241

Vishnoi, A., Roy, R., Prasad, H. K., & Bhattacharya, A. (2010). Anchor-based whole genome phylogeny (ABWGP): A tool for *inferring evolutio*nary relationship among closely related microorganims. PLoS One, 5(11), e14159. doi:10.1371/journal.pone.0014159Vonk, F. J., Casewell, N. R., Henkel, C. V., Heimberg, A. M., Jansen, H. J., McCleary, R. J. R., … Richardson, M. K. (2013). The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. Proceedings of the National Academy of Sciences, 110(51), 20651–20656. doi:10.1073/pnas.1314702110

Walker, J. R., Novick, P. A., Parsons, W. H., McGregor, M., Zablocki, J., Pande, V. S., & Du Bois, J. (2012). Marked difference in saxitoxin and tetrodotoxin affinity for the human nociceptive voltage-gated sodium channel (Nav1.7). *Proceedings of the National Academy of Sciences*, *109*(44), 18102–18107. doi:10.1073/pnas.1206952109

Wang, S. Y., Mitchell, J., Tikhonov, D. B., Zhorov, B. S., & Wang, G. K. (2006). How batrachotoxin modifies the sodium channel permeation pathway: computer modeling and site-directed mutagenesis. *Molecular pharmacology*, *69*(3), 788-795. doi:10.1124/mol.105.018200.channels

Wang, S. Y., & Wang, G. K. (2017). Single rat muscle Na+ channel mutation confers batrachotoxin autoresistance found in poison-dart frog Phyllobates terribilis. *Proceedings of the National Academy of Sciences*, *114*(39), 10491-10496. doi:10.1073/pnas.1707873114

Wang, S. Y., & Wang, G. K. (1999). Batrachotoxin-resistant Na+ channels derived from point mutations in transmembrane segment D4-S6. *Biophysical Journal*, *76*(6), 3141-3149. doi:10.1016/S0006-3495(99)77465-5

Wang, X., Wang, Y., Yue, B., Zhang, X., & Liu, S. (2013). The complete mitochondrial genome of the Bufo tibetanus (Anura: Bufonidae). *Mitochondrial DNA*, *24*(3), 186-188. doi:10.3109/19401736.2012.744978

Wheeler, G. N., & Brändli, A. W. (2009). Simple vertebrate models for chemical genetics and drug discovery screens: Lessons from zebrafish and *Xenopus*. *Developmental Dynamics*, *238*(6), 1287–1308. doi:10.1002/dvdy.21967

Wheeler, W. C. (2006). Dynamic homology and the likelihood criterion. *Cladistics*, *22*(2), 157-170.

Wheeler, W. C., Lucaroni, N., Hong, L., Crowley, L. M., & Varón, A. (2015). POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. *Cladistics*, *31*(2), 189-196.

Wilson, M. R., Allard, M. W., & Brown, E. W. (2013). The forensic analysis of foodborne bacterial pathogens in the age of whole-genome sequencing. *Cladistics*, *29*(4), 449–461. doi:10.1111/cla.12012

Wolfsberg, T. G., Schafer, S., Tatusov, R. L., & Tatusova, T. A. (2001). Organelle genome resources at NCBI. *Trends in biochemical sciences*, *26*(3), 199-203. doi:10.1016/S0968-0004(00)01773-4

Wolkovich, E. M., Cook, B. I., & Davies, T. J. (2014). Progress towards an interdisciplinary science of plant phenology: building predictions across space, time and species diversity. *New Phytologist*, *201*(4), 1156–1162. doi:10.1111/nph.12599

Xia, Y., Zheng, Y., Miura, I., Wong, P. B., Murphy, R. W., & Zeng, X. (2014). The evolution of mitochondrial genomes in modern frogs (Neobatrachia): nonadaptive evolution of mitochondrial genome reorganization. *BMC genomics*, *15*(1), 691. doi:10.1186/1471-2164-15-691Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., … Wang, J. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, *30*(12), 1660–1666. doi:10.1093/bioinformatics/btu077

Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., ... & Chen, S. (2012). FastUniq: a fast de novo duplicates removal tool for paired short reads. *PloS one*, *7*(12), e52249. doi:10.1371/journal.pone.0052249

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., ... & Zhu, B. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC bioinformatics*, *14*(1), 33. doi:10.1186/1å471-2105-14-33

Young, B. E., Lips, K. R., Reaser, J. K., Ibanez, R., Salas, A. W., Cedeño, J. R., ... & Munoz, A. (2001). Population declines and priorities for amphibian conservation in Latin America. *Conservation Biology*, *15*(5), 1213-1223.

Yuan, Z. Y., Zhou, W. W., Chen, X., Poyarkov Jr, N. A., Chen, H. M., Jang-Liaw, N. H., ... & Kuzmin, S. L. (2016). Spatiotemporal diversification of the true frogs (genus Rana): a historical framework for a widely studied group of model organisms. *Systematic biology*, *65*(5), 824-842. doi:10.1093/sysbio/syw055

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., … Kriventseva, E. V. (2016). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, 1–15. doi:10.1093/nar/gkw1119

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, *18*(5), 821-829. doi:10.1101/gr.074492.107

Zhang, P., Liang, D., Mao, R. L., Hillis, D. M., Wake, D. B., & Cannatella, D. C. (2013). Efficient sequencing of anuran mtDNAs and a mitogenomic exploration of the phylogeny and evolution of frogs. *Molecular Biology and Evolution*, *30*(8), 1899-1915. doi:10.1093/molbev/mst091

Zhang, P., Zhou, H., Chen, Y. Q., Liu, Y. F., & Qu, L. H. (2005). Mitogenomic perspectives on the origin and phylogeny of living amphibians. *Systematic Biology*, *54*(3), 391-400. doi:10.1080/10635150590945278

Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marcais, G., … Langley, C. H. (2014). Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. *Genetics*, *196*(3), 875–890. doi:10.1534/genetics.113.159715

Zimin, A. V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669–2677. doi:10.1093/bioinformatics/btt476

PROTOCOL S2.1

Bioinformatics Protocol For Quality Control

Quality Control

Commented bioinformatics protocol using *Hylodes meridionalis* raw shotgun sequencing reads.

All homemade Python scripts are available at http://www.ib.usp.br/grant/anfibios/researchSoftware.html and at https://gitlab.com/MachadoDJ under the GNU General Public License version 3.0 (GPL-3.0).

Illumina HiSeq raw shotgun sequencing reads

Sequences were delivered in several unsorted files:

- Hyl_meridionalis_TGACCA_L003_R1_001.fastq.gz
- Hyl_meridionalis_TGACCA_L003_R1_002.fastq.gz
- Hyl_meridionalis_TGACCA_L003_R1_003.fastq.gz
- Hyl_meridionalis_TGACCA_L003_R2_001.fastq.gz
- Hyl_meridionalis_TGACCA_L003_R2_002.fastq.gz
- Hyl_meridionalis_TGACCA_L003_R2_003.fastq.gz

Concatenate raw sequencing reads

Files were concatenated using common UNIX command lines:

```
$ cat Hyl_meridionalis_TGACCA_L003_R1* > unsorted_R1.fastq.gz
$ cat Hyl_meridionalis_TGACCA_L003_R2* > unsorted_R2.fastq.gz
```

These processes took less than one second to and used up to 7 Mb of RAM.

Pre-processing reads

Using PATO-FU, we sorted all concatenated files using chunk sizes of approximately 10% of the number of total reads (completed in approx. 50 min using up to 13 Gb of RAM, in average):

```
$ pato-fu -S --stanza_size 4 --chunk_size 1000000 -z -i \
  unsorted_R1.fastq.gz -o sorted_R1.fastq.gz
$ pato-fu -S --stanza_size 4 --chunk_size 1000000 -z -i \
  unsorted_R1.fastq.gz -o sorted_R1.fastq.gz
```

Sorting sequence reads with PATO-FU used approx. 580% less memory (RAM) and took approx. 490% more time than sorting with common Linux executables. We recommend using PATO-FU to sort reads if the number of reads is high and/or the amount of available memory is low.

Sequencing order and pairing was checked using PATO-FU (completed in approx. 9 min in average using up to 24 Mb of RAM):

```
$ pato-fu --checker --stanza_size 4 --checker_opt 2 -z –i \
      sorted_R1.fastq.gz sorted_R2.fastq.gz > check.txt
```

Quality evaluation of pre-processed raw reads with FASTQC v0.11.2

The following command line was executed in one of ACE's vnodes using the template PBS script available at http://www.ib.usp.br/grant/anfibios/researchHPC.html.

```
$ fastqc sorted_R1.fastq.gz > fastqc_R1.out
$ fastqc sorted_R2.fastq.gz > fastqc_R2.out
```

The execution time was approx. 2 minutes in avegare and these processes used up to 535 Mb of RAM.

Summarize sequencing reads quality with ht-stat v0.90.7

```
$ ht-stat -P -t 4 -z -i sorted_R1.fastq.gz sorted_R2.fastq.gz -o htstat >
htstat.log
```

The execution time was approx. 2 minutes in average and this process used up to 265 Mb of RAM.

Automatize tile selection

The program "selectTiles.py" automatizes the selection of tiles to be removed after running "ht-stat", following criteria based in the HTQC guidelines:

- More than 50% of the reads have quality score bellow 10
- Less than 10% of the reads have quality greater than 30
- Most reads have quality bellow 20

Command line to execute selectTiles.py is simple and just requires the user to point to the ht-stat results directory:

```
$ selectTiles -d htstat > tiles.txt
```

Tiles selected this way can be removed with "ht-filter" using the "–-filter tile" argument.
Command execution took less than one second and used up to 16 Mb of RAM.

Remove bad bases from reads head or tail with ht-trim v0.90.7

```
$ ht-trim -z -i sorted_R1.fastq.gz -o trimmed_R1.fastq.gz
$ ht-trim -z -i sorted_R2.fastq.gz -o trimmed_R2.fastq.gz
```

The execution time of these processes were approx. 2 minutes in average and they used up to 265 Mb of RAM.

\Filter sequences by length with ht-filter v0.90.7

```
$ ht-filter -P --filter length -z -i trimmed_R1.fastq.gz
trimmed_R2.fastq.gz -o filtered
```

The execution time of ht-filter was approx. 13 min in average and the process used up to 16 Mb of RAM.

In the case of *Hylodes meridionalis*, aprox. 5.6% of the reads were discarded during quality control. Approx. 9.9% of the remaining reads were filtered as single end sequencing reads and were ignored for the remaining of the analysis.

Quality evaluation of filtered raw reads with FASTQC v0.11.2

```
$ fastqc filtered_1.fastq.gz > fastqc_filtered_1.out
$ fastqc filtered_2.fastq.gz > fastqc_filtered_2.out
```

Final quality evaluatin took less than one and a half minutes using up to 529 Mb of RAM.

Comparison of quality reports before and after quality control show major improvements in per base/tile sequencing quality and over-represented sequences (see **Table S1**). However, some filtered sequence files still failed per base sequence content and k-mer content. We must note that, according to FASTQC help page (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/; last access: May 4, 2015), libraries that derive from random priming will nearly always show k-mer bias. Also according to the FASTQC help page, sequences subjected to aggressive trimming are more likely to present per base sequence content bias.

PROTOCOL S2.2

Bioinformatics Protocol For Sequence Assembly

Sequence assembly

We employed the filtered reads of Hylodes meridionalis for three different assemblage approaches: 1. reference-based (using Bowtie2); 2. de novo (using ABySS, SOAP2 and Velvet); and 3. baiting and iterative mapping (using MIRA and MITObim).

Assembling algorithms were executed in ACE's vnodes using the template PBS script available at http://www.ib.usp.br/grant/anfibios/researchHPC.html.

Reference based sequence assembly with Bowtie2 v2.2.3

We selected the mitogenome of Bufo tibetanus as reference. The first step is indexing this reference mitogenome with "bowtie2-build":

```
$ bowtie2-build -f reference/reference.fa Btibetanus
```

Instead of extracting the sequence files, users may prefer to use FIFO special files:

```
$ mkfifo filtered_1.fifo
$ mkfifo filtered_2.fifo
$ nohup zcat filtered_1.fastq.gz > filtered_1.fifo &
$ nohup zcat filtered_2.fastq.gz > filtered_2.fifo &
```

Aligning sequencing reads with "bowtie2":

```
$ bowtie2 -x Btibetanus -q --phred33 -1 filtered_1.fifo -2 filtered_2.fifo
-S ref_align.sam -I 100 -X 500 -p 8 -t > align.out 2> align.err
```

Using "samtools" to convert SAM file to BAM:

```
$ samtools view -bS ref_align.sam > ref_align.bam
```

Sort:

```
$ samtools sort ref_align.bam ref_align_sorted.bam
```

Generate variant calls in VCF format:

```
$ samtools mpileup -uf reference/reference.fa ref_align_sorted.bam |
bcftools view -bvcg - > ref_align.raw.bcf
```

View alignment:

```
$ bcftools view ref_align.raw.bcf
```

After multiples trials and tweaking with the input parameters, we were unable to assemble to assemble the mitogenome of Hylodes meridionalis using this strategy.

***De novo* assembly with ABySS v1.5**

Genome assembly with ABySS can be achieved with a single command line:

```
$ abyss-pe -C k21 np=64 j=64 k=21 n=10 N=10 name=mtdna_k21 lib='pe1' pe1='../filtered_1.fastq.gz ../filtered_1.fastq.gz' ABYSS_OPTIONS=--no-chastity
```

We repeated this for every k-mer size between 21 and 63, with a step of 2. ABySS stats file for each k-mer size was produced with the "abyss-fac" command. For example, for k-mer size 21, the command line would be the following:

```
$ abyss-fac k21/mtdna_k21-contigs.fa > mtdna_k21-stats.txt
```

The coverage (represented by the "sum" variable in the statistics file) and N50 value can be used to select the best k-mer size. We used BLAT (blatSrc35) to align scaffolds from all k-mer sizes to the reference mitogenome of Bufo tibetanus to extract potencial mtDNA fragments.

```
$ blat -t=dna -q=dna -oneOff=0 -stepSize=11 -minMatch=2 -minScore=30 -minIdentity=90 -maxGap=2 -repMatch=1024 reference.fa mtdna_scaffolds.fa output.psl
```

The pslScore.pl Perl script was used to replicate the percent identity and score calculations produced by the web-based BLAT (available at https://genome.ucsc.edu/cgi-bin/hgBlat):

```
$ pslScore.pl output.pl
```

None of the scaffolds mapped to the reference mitogenome. We considered this assembly approach ineffective for our dataset.

Note: We had difficulties using heads in the CASAVA 1.8+ format. It seems that at least some versions of ABySS requires a pair of reads to be named with the suffixes /1 and /2 to identify the first and second read, or the reads may be named identically. Therefore, we had to edit sequence headers using PATO-FU so that sequence identifiers would match the format previous to CASAVA 1.8+.

***De novo* assembly with SOAPdenovo2 v2.04**

Example configuration file:

```
#maximal read length
max_rd_len=100
[LIB]
```

```
#average insert size (we tested the values 150, 200 and 250)
avg_ins=2150
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used
asm_flags=3
#in which order the reads are used while scaffolding
rank=1
# cutoff of pair number for a reliable connection
pair_num_cutoff=3
#minimum aligned length to contigs for a reliable read location
map_len=32
#a pair of fastq files
q1=filtered_1.fastq
q2=filtered_2.fastq
```

Example command line to execute "SOAPdenovo"

```
$ SOAPdenovo-63mer all -s my.config -o graphOutput -K 63 -k 63 -p 32 -a 16
-d 1 -R -D 1 -M 1 -e 0 -z 9000000000
```

We mapped all the scaffolds to the reference plastid genome using "BLAT" (see above). None of the scaffolds mapped to the reference mitogenome. We considered this assembly approach ineffective for our dataset.

Note: We tested average insert sizes of 150, 200 and 200, modifying the "avg_ins" parameter in the configuration file.

De novo assembly with Velvet v1.2.10

The primary parameter options (K, -exp_cov, -cov_cutoff) for the "Velvet" de novo sequence assembler were optimized using the Perl script "VelvetOptimiser.pl", testing every k-mer size between 21 and 63, with a step of 2.

```
$ perl -I VelvetOptimiser.pl -s 21 -e 65 -f '-shortPaired -fastq
filtered_1.fastq filtered_2.fastq' -t 64
```

The k-mer size 31 was selected. Hashing was executed in "velveth".

```
$ velvethvelvet_output/ 31 -fastq -shortPaired filtered_1.fastq
filgtered_2.fastq
```

Finally, the assembly was executed with "velvetg".

```
$ velvetg velvet_output/ -exp_cov 2 -cov_cutoff 0.3777216
```

We mapped all the scaffolds to the reference plastid genome using "BLAT" (see above). None of the scaffolds mapped to the reference mitogenome. We considered this assembly approach was ineffective for our dataset. None of the contigs mapped to the reference mitogenome. We considered this assembly approach ineffective for our dataset.

Four different baiting and iterative mapping strategies were employed giving time constrains and available references. (1) Mapping to a complete mitochondrial genome of a closely related taxa; (2) mapping to a distantly related mitochondrial genome; (3) baiting with a barcode seed from a closely related taxa; (4) baiting with a barcode seed from a closely related taxa, de novo option on.

Also, we modified the "MITObim_1.6.pl" original Perl script so it would create manifest files for "Mira" pointing to a directory in local file system in ACE's vnodes. Modified "MITObim" program is available at http://www.ib.usp.br/grant/anfibios/researchHPC.html respecting the original MIT license.

First, we used "Mira" to map the filtered reads to the reference. This is an example of the manifest files used in this step:

```
project = Mitogenome_1
job=genome,mapping,accurate
parameters = -DI:tmp=/LFS/mira,-NW:cmrnl=warn,SOLEXA_SETTINGS
readgroup
is_reference
data = reference.fa
strain = Reference
readgroup = reads
data = filtered_*.fastq
template_size = 100 300
segment_placement = ---> <---
technology = solexa
strain = Mitogenome
```

When using a computer cluster, you must create a temporary directory for Mira at the Local File System (which in this case is "/LFS/mira"). This will be passed on to our modified MITObim script using the argument "--lfspath" (see bellow).

Mira can be executed by calling this manifest file:

```
$ /apps/mira manifest.conf
```

Mira was executes in 45 to 60 min using up to up to 75 Gb of RAM.

Iterative mapping using MITObim can be executed in a single command line:

```
$ MITObim.pl -start 1 -end 1000 -sample Mitogenome -ref Reference-readpool
interleaved.fastq   -maf   Mitogenome_1_assembly/   Mitogenome_1_d_results/
Mitogenome_1_out.maf  --pair  --readlength  150  --insert  300  --kbait  31
--clean --mirapath /apps/ --lfspath /LFS/Hbocagei1/
```

Each MITObim iteration took one to 5 min using up to 1.5 Gb of RAM.

We repeated the steps above using both a more closely (the frog Bufo tibetanus, NCBI accession number NC_020048) and a more distantly related (the salamander Rhiacotriton variegatus, NCBI accession number NC_006331) mitogenome as reference. We also used barcode seeds (COX I gene from the B. tibetanus genome, 5533-7044 bp) with and without the "--denovo" option in MITObim.

MITObim can be executed several times using the same Mira output files. In this case, we suggest the user to create separate directories for each MITObim analysis.

Supplementary material

Mitogenomic sequences assembled with different reference sequences aligned to each other with no base-to-base variation. Differences in the mitogenomic sequences assembled were restricted to sequence length and number of reads assembled, which also affects average coverage and quality.

Table S2.1 - General FASTQC statistics for all sequencing reads before and after quality control.

Basic statistics	**Raw**	**Filtered**
Basic Statistics	PASS	PASS
Per base sequence quality	FAIL	PASS
Per tile sequence quality	FAIL	WARN
Per sequence quality scores	PASS	PASS
Per base sequence content	FAIL	FAIL
Per sequence GC content	FAIL	FAIL
Per base N content	PASS	PASS
Sequence Length Distribution	PASS	WARN
Sequence Duplication Levels	FAIL	PASS
Overrepresented sequences	FAIL	WARN
Adapter Content	PASS	PASS
Kmer Content	FAIL	FAIL

Table S2.2 - Gene order. +/- sines indicate hot and cold chains, respectively. Gene start and and positions in the unppadded mitogenome are inside the parenthesis.

Hylodes meridionalis	*Hyloxalus* m*aculosus*	*M. simplex*	*P. fenestratus*	*Rhinella* sp. C.
+trnL1 (44-115)	+trnL1 (1-71)	+trnL1 (70-141)	+trnL1 (108-178)	+trnL1 (56-127)
+trnT (116-188)	+trnT (73-143)	+trnT (142-213)	-trnP (187-255)	+trnT (128-199)
-trnP (188-256)	-trnP (143-211)	-trnP (213-281)	+trnF (319-385)	-trnP (199-266)
+trnF (256-323)	+trnF (211-278)	+trnF (281-348)	+rrnS (386-1319)	+trnF (266-334)
+rrnS (325-1262)	+rrnS (279-1214)	+rrnS (349-1281)	+trnV (1318-1387)	+rrnS (335-1264)
+trnV (1260-1328)	+trnV (1212-1281)	+trnV (1279-1347)	+rrnL (1389-2991)	+trnV (1262-1330)
+rrnL (1329-2922)	+rrnL (1282-2873)	+rrnL (1348-2949)	+trnL2 (2990-3062)	+rrnL (1331-2924)
+trnL2 (2922-2994)	+trnL2 (2878-2950)	+trnL2 (2949-3021)	+nad1 (3105-4058)	+trnL2 (2924-2996)
+nad1 (3010-3945)	+nad1 (2960-3901)	+nad1 (3040-3975)	+trnI (4093-4172)	+nad1 (3009-3950)
+trnI (3953-4023)	+trnI (3909-3980)	+trnI (3983-4053)	+trnM (4223-4292)	+trnI (3958-4028)
-trnQ (4023-4093)	-trnQ (3980-4050)	-trnQ (4053-4123)	+nad2 (4421-5329)	-trnQ (4028-4097)
+trnM (4093-4161)	+trnM (4050-4118)	+trnM (4123-4191)	+trnW (5335-5403)	+trnM (4097-4165)
+nad2 (4162-5190)	+nad2 (4119-5147)	+nad2 (4192-5220)	-trnA (5405-5473)	+nad2 (4166-4765)
+trnW (5197-5266)	+trnW (5153-5222)	+trnW (5226-5295)	-trnN (5474-5546)	+nad2 (4800-5195)
-trnA (5267-5335)	-trnA (5223-5291)	-trnA (5297-5365)	-trnC (5574-5628)	+trnW (5200-5269)
-trnN (5336-5408)	-trnN (5292-5364)	-trnN (5366-5438)	-trnY (5630-5696)	-trnA (5270-5338)
-trnC (5434-5496)	-trnC (5391-5454)	-trnC (5465-5528)	+cox1 (5689-7236)	-trnN (5339-5411)
-trnY (5497-5566)	-trnY (5455-5522)	-trnY (5529-5598)	-trnS2 (7241-7312)	-trnY (5436-5496)
+cox1 (5568-7100)	+cox1 (5515-7053)	+cox1 (5600-7132)	+trnD (7313-7381)	-trnY (5497-5566)
-trnS2 (7114-7184)	-trnS2 (7071-7141)	-trnS2 (7146-7216)	+cox2 (7382-8053)	+cox1 (5568-7097)
+trnD (7185-7252)	+trnD (7143-7211)	+trnD (7218-7286)	+trnK (8064-8131)	+trnD (7186-7253)
+cox2 (7247-7930)	+cox2 (7212-7889)	+cox2 (7288-7971)	+atp8 (8132-8290)	+cox2 (7254-7928)
+trnK (7941-8012)	+trnK (7900-7971)	+trnK (7976-8047)	+atp6 (8287-8961)	+trnK (7942-8013)
+atp8 (8013-8171)	+atp8 (7972-8130)	+atp8 (8049-8207)	+cox3 (8966-9748)	+atp8 (8000-8173)
+atp6 (8168-8845)	+atp6 (8118-8804)	+atp6 (8189-8881)	+trnG (9750-9817)	+atp6 (8155-8847)
+cox3 (8851-9633)	+cox3 (8810-9592)	+cox3 (8887-9669)	+nad3 (9803-10156)	+cox3 (8853-9635)
+trnG (9635-9703)	+trnG (9594-9662)	+trnG (9671-9739)	+trnR (10158-10225)	+trnG (9637-9705)
+nad3 (9701-10042)	+nad3 (9660-10001)	+nad3 (9740-10078)	+nad4l (10227-10523)	+nad3 (9703-10044)
+trnR (10044-10112)	+trnR (10003-10070)	+trnR (10080-10148)	+nad4 (10520-11866)	+trnR (10046-10113)
+nad4l (10114-10410)	+nad4l (10094-10372)	+nad4l (10138-10449)	+trnH (11872-11939)	+nad4l (10153-10410)
+nad4 (10407-11765)	+nad4 (10369-11727)	+nad4 (10446-11807)	+trnS1 (11940-12006)	+nad4 (10407-11762)
+trnH (11771-11838)	+trnH (11733-11802)	+trnH (11811-11878)	+nad5 (12032-13792)	+trnS1 (11843-11909)
+trnS1 (11839-11905)	+trnS1 (11803-11869)	+trnS1 (11879-11945)	-nad6 (13817-14320)	+nad5 (11945-13684)
+nad5 (11939-13627)	+nad5 (11914-13683)	+nad5 (12028-13770)	-trnE (14322-14390)	-nad6 (13729-14220)
-nad6 (13732-14229)	-nad6 (13688-14179)	-nad6 (13772-14275)	+cob (14393-15517)	-trnE (14218-14285)
+cob (14296-15429)	-trnE (14181-14248)	-trnE (14261-14328)	-trnQ (17853-17924)	+cob_0 (14287-15162)
-trnE (15567-15634)	+cob (14251-15378)	+cob_0 (14331-15380)		+cob_1 (15309-15422)
		+cob_1 (15784-15996)		+trnH (16015-16068)

Table S2.3 - Base composition and other features of mitochondrial genomes.

Species	Family	GenBank	Size
Pristimantis fenestratus	Strabomantidae	KT221610.1	17889
Melanophryniscus simplex	Bufonidae	KT221611.1	16338
Hyloxalus yasuni	Dendrobatidae	KT221612.1	16052
Rhinela sp.	Bufonidae	KT221613.1	17045
Hylodes meridionalis	Hylodidae	KT221614.1	16166

Table S2.3 - Continued.

Species	GC-content	AT/GC ratio	A-content
Pristimantis fenestratus	0.39236402258371067	15,486,536,543,667,100	0.3007993739169322
Melanophryniscus simplex	0.39135757130615745	15,552,080,075,070,300	0.3059125964010283
Hyloxalus yasuni	0.4261774233740344	13,464,405,788,627,300	0.2742960378769001
Rhinela sp.	0.3906717512466999	1,559,693,647,694,840	0.2948078615429745
Hylodes meridionalis	0.38846962761350984	1,574,203,821,656,050	0.2941976988741804

Table S2.3 - Continued.

Species	C-content	G-content	T-content
Pristimantis fenestratus	0.2651349991614959	0.12722902342221476	0.30683660349935715
Melanophryniscus simplex	0.25009181050312157	0.14126576080303588	0.3027298322928143
Hyloxalus yasuni	0.2785322701221032	0.14764515325193123	0.29952653874906554
Rhinela sp.	0.24218245819888531	0.1484892930478146	0.3145203872103256
Hylodes meridionalis	0.24044290486205616	0.14802672275145368	0.31733267351230976

Table S2.3 - Continued.

Species	Protein coding-genes	tRNA coding-genes	rRNA coding-genes
Pristimantis fenestratus	13	21	2
Melanophryniscus simplex	13	22	2
Hyloxalus yasuni	13	22	2
Rhinela sp.	13	22	2
Hylodes meridionalis	13	22	2

TABLE S3.1

Inference of circularity: resuls of permutation tests

Table S3.1 – Inference of circularity: results of all permutation tests.

Species	Indel Prob.	Iteration	Contiguity	Coverage	Connectivity	Quality	Score	Similarity	Prob. Mod.
S. holbrookii	0	0	True	790.93	774.5	37.0633	-0.2960	100	0.00%
M. moreirae	0	0	True	137.5	134.14	35.7815	-0.9039	100	0.00%
H. subpunctatus	0	0	True	97.53	95.44	36.0021	-1.4185	100	0.00%
P. terribilis	0	0	True	6033.78	5812.65	36.7310	-2.8147	98	0.00%
M. moreirae	0.01	4	True	59.64	57.4	34.8347	-2.8976	97	1.00%
S. holbrookii	0.01	1	True	56.24	54.86	34.7100	-4.1828	100	1.00%
H. subpunctatus	0.01	3	True	87.33	84.6	35.9878	-4.5423	100	1.00%
S. holbrookii	0.01	2	True	761.05	745.03	37.1123	-4.7021	100	1.00%
S. holbrookii	0.01	3	True	34.35	33.31	34.8008	-5.2704	100	1.00%
M. moreirae	0.01	2	True	131.93	128.3	35.7931	-5.3080	100	1.00%
M. moreirae	0.01	1	True	107.84	104.64	36.2125	-5.4454	100	1.00%
P. terribilis	0.01	2	True	4373.78	4189.01	36.7182	-5.4992	100	1.00%
H. subpunctatus	0.01	1	True	164.3	160.29	36.0379	-5.6545	100	1.00%
H. subpunctatus	0.01	2	True	81.68	79.77	35.9890	-6.2284	100	1.00%
H. subpunctatus	0.01	4	True	17.34	16.64	35.6092	-6.5575	100	1.00%
M. moreirae	0.01	3	True	81.29	78.46	35.9761	-6.9222	100	1.00%
P. terribilis	0.01	1	True	147.81	143.47	36.9234	-7.3826	100	1.00%
S. holbrookii	0.01	4	True	132.16	127.35	35.1886	-7.5551	100	1.00%
P. terribilis	0.01	3	True	81.38	78.17	36.3840	-8.0851	100	1.00%
P. terribilis	0.01	5	True	79.74	76.52	36.3877	-9.0115	100	1.00%
H. subpunctatus	0.01	34	False	0.22	0.17	3.6350	0.0000	10	1.00%
H. subpunctatus	0.01	78	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	80	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	82	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	84	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	86	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	88	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	90	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	102	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	104	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	106	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	108	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	110	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	112	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	114	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	116	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	118	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	120	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	122	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	124	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	126	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	128	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	130	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	132	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	134	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	136	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	138	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	140	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	142	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	144	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	146	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	148	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	150	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	152	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	154	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	156	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	157	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	158	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	159	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	160	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	161	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	162	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	163	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	164	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	165	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	166	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	167	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	168	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	169	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	170	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	171	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	172	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	173	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	174	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	175	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	176	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	177	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	178	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	179	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	180	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	181	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	182	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	183	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	184	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	185	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	186	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	187	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	188	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	189	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	190	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	191	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	192	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	193	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	194	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	195	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	196	False	0	0	0.0000	0.0000	0	1.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
H. subpunctatus	0.01	197	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	198	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	199	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.01	200	False	0	0	0.0000	0.0000	0	1.00%
H. subpunctatus	0.05	1	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	2	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	3	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	4	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	5	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	6	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	8	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	10	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	12	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	14	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	16	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	18	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	20	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	22	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	24	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	26	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	28	False	0.03	0.01	0.6600	0.0000	2	5.00%
H. subpunctatus	0.05	30	False	0.05	0.03	0.9900	0.0000	3	5.00%
H. subpunctatus	0.05	32	False	0.09	0.07	1.6500	0.0000	5	5.00%
H. subpunctatus	0.05	34	False	0.11	0.09	2.0000	0.0000	6	5.00%
H. subpunctatus	0.05	109	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	111	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	113	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	114	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	115	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	116	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	117	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	118	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	119	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	120	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	121	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	122	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	123	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	124	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	125	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	126	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	127	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	128	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	129	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	130	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	131	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	132	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	133	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	134	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	135	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	136	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	137	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	138	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	139	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	140	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	141	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	142	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	143	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	144	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	145	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	146	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	147	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	148	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	149	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	150	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	151	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	152	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	153	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	154	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	155	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	156	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	157	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	158	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	159	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	160	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	161	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	162	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	163	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	164	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	165	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	166	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	167	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	168	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	169	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	170	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	171	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	172	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	173	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	174	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	175	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	176	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	177	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	178	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	179	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	180	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	181	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	182	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	183	False	0	0	0.0000	0.0000	0	5.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
H. subpunctatus	0.05	184	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	185	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	186	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	187	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	188	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	189	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	190	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	191	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	192	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	193	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	194	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	195	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	196	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	197	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	198	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	199	False	0	0	0.0000	0.0000	0	5.00%
H. subpunctatus	0.05	200	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.01	147	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	148	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	149	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	150	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	151	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	152	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	153	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	154	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	155	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	156	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	157	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	158	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	159	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	160	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	161	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	162	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	163	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	164	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	165	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	166	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	167	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	168	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	169	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	170	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	171	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	172	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	173	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	174	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	175	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	176	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	177	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	178	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	179	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	180	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	181	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	182	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	183	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	184	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	185	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	186	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	187	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	188	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	189	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	190	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	191	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	192	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	193	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	194	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	195	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	196	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	197	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	198	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	199	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.01	200	False	0	0	0.0000	0.0000	0	1.00%
M. moreirae	0.05	1	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	2	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	3	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	4	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	5	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	6	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	7	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	8	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	9	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	10	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	11	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	12	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	13	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	14	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	15	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	17	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	110	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	112	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	114	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	115	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	116	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	117	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	118	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	119	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	120	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	121	False	0	0	0.0000	0.0000	0	5.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
M. moreirae	0.05	122	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	123	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	124	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	125	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	126	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	127	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	128	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	129	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	130	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	131	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	132	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	133	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	134	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	135	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	136	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	137	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	138	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	139	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	140	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	141	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	142	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	143	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	144	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	145	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	146	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	147	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	148	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	149	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	150	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	151	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	152	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	153	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	154	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	155	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	156	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	157	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	158	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	159	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	160	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	161	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	162	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	163	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	164	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	165	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	166	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	167	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	168	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	169	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	170	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	171	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	172	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	173	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	174	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	175	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	176	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	177	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	178	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	179	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	180	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	181	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	182	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	183	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	184	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	185	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	186	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	187	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	188	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	189	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	190	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	191	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	192	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	193	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	194	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	195	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	196	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	197	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	198	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	199	False	0	0	0.0000	0.0000	0	5.00%
M. moreirae	0.05	200	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.01	27	False	0.25	0.16	3.8827	0.0000	11	1.00%
S. holbrookii	0.01	29	False	0.25	0.16	3.8827	0.0000	11	1.00%
S. holbrookii	0.01	33	False	0.22	0.14	1.0300	0.0000	3	1.00%
S. holbrookii	0.01	35	False	0.38	0.3	1.7663	0.0000	5	1.00%
S. holbrookii	0.01	37	False	0.38	0.3	1.7663	0.0000	5	1.00%
S. holbrookii	0.01	39	False	0.62	0.54	2.8350	0.0000	8	1.00%
S. holbrookii	0.01	41	False	0.62	0.54	2.8350	0.0000	8	1.00%
S. holbrookii	0.01	43	False	0.62	0.54	2.8350	0.0000	8	1.00%
S. holbrookii	0.01	115	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	122	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	123	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	125	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	126	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	127	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	128	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	129	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	130	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	131	False	0	0	0.0000	0.0000	0	1.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
S. holbrookii	0.01	132	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	133	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	134	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	135	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	136	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	137	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	138	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	139	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	140	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	141	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	142	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	143	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	144	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	145	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	146	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	147	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	148	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	149	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	150	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	151	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	152	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	153	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	154	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	155	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	156	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	157	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	158	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	159	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	160	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	161	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	162	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	163	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	164	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	165	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	166	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	167	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	168	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	169	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	170	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	171	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	172	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	173	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	174	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	175	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	176	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	177	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	178	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	179	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	180	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	181	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	182	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	183	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	184	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	185	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	186	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	187	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	188	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	189	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	190	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	191	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	192	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	193	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	194	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	195	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	196	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	197	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	198	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	199	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.01	200	False	0	0	0.0000	0.0000	0	1.00%
S. holbrookii	0.05	1	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	2	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	3	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	4	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	5	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	7	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	105	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	107	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	109	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	111	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	113	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	114	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	115	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	116	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	117	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	118	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	119	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	120	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	121	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	122	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	123	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	124	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	125	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	126	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	127	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	128	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	129	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	130	False	0	0	0.0000	0.0000	0	5.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
S. holbrookii	0.05	131	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	132	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	133	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	134	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	135	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	136	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	137	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	138	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	139	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	140	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	141	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	142	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	143	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	144	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	145	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	146	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	147	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	148	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	149	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	150	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	151	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	152	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	153	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	154	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	155	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	156	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	157	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	158	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	159	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	160	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	161	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	162	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	163	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	164	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	165	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	166	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	167	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	168	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	169	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	170	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	171	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	172	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	173	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	174	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	175	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	176	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	177	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	178	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	179	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	180	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	181	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	182	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	183	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	184	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	185	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	186	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	187	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	188	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	189	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	190	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	191	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	192	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	193	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	194	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	195	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	196	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	197	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	198	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	199	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.05	200	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.01	159	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	163	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	165	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	166	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	167	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	168	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	169	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	170	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	171	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	172	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	173	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	174	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	175	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	176	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	177	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	178	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	179	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	180	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	181	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	182	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	183	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	184	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	185	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	186	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	187	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	188	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	189	False	0	0	0.0000	0.0000	0	1.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
P. terribilis	0.01	190	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	191	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	192	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	193	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	194	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	195	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	196	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	197	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	198	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	199	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.01	200	False	0	0	0.0000	0.0000	0	1.00%
P. terribilis	0.05	1	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	2	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	3	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	4	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	5	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	7	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	9	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	11	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	13	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	107	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	109	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	111	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	113	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	115	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	117	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	118	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	119	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	120	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	121	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	122	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	123	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	124	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	125	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	126	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	127	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	128	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	129	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	130	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	131	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	132	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	133	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	134	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	135	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	136	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	137	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	138	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	139	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	140	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	141	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	142	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	143	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	144	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	145	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	146	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	147	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	148	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	149	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	150	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	151	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	152	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	153	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	154	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	155	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	156	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	157	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	158	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	159	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	160	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	161	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	162	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	163	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	164	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	165	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	166	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	167	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	168	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	169	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	170	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	171	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	172	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	173	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	174	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	175	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	176	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	177	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	178	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	179	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	180	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	181	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	182	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	183	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	184	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	185	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	186	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	187	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	188	False	0	0	0.0000	0.0000	0	5.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
P. terribilis	0.05	189	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	190	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	191	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	192	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	193	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	194	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	195	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	196	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	197	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	198	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	199	False	0	0	0.0000	0.0000	0	5.00%
P. terribilis	0.05	200	False	0	0	0.0000	0.0000	0	5.00%
S. holbrookii	0.01	45	False	0.74	0.62	3.1908	-0.0185	9	1.00%
S. holbrookii	0.01	47	False	1.02	0.74	3.5337	-0.0631	10	1.00%
S. holbrookii	0.01	49	False	1.35	1.02	3.8813	-0.0932	11	1.00%
S. holbrookii	0.01	51	False	2.41	2.04	4.9417	-0.1601	14	1.00%
H. subpunctatus	0.01	32	False	0.25	0.19	3.2650	-0.2000	9	1.00%
S. holbrookii	0.05	91	False	1.98	1.64	4.2791	-0.3510	13	5.00%
H. subpunctatus	0.01	28	False	0.21	0.16	1.8250	-0.3600	5	1.00%
S. holbrookii	0.05	79	False	1.03	0.88	5.0249	-0.3816	14	5.00%
S. holbrookii	0.05	93	False	2.8	2.46	5.8815	-0.3924	18	5.00%
S. holbrookii	0.05	89	False	1.49	1.19	3.0083	-0.3959	9	5.00%
S. holbrookii	0.05	95	False	3.39	3.02	7.1017	-0.3999	22	5.00%
H. subpunctatus	0.01	30	False	0.29	0.21	2.5507	-0.4408	7	1.00%
S. holbrookii	0.05	81	False	1.65	1.47	6.5222	-0.4714	18	5.00%
S. holbrookii	0.05	77	False	0.85	0.7	3.2734	-0.5302	9	5.00%
H. subpunctatus	0.05	70	False	0.67	0.58	5.3441	-0.5434	15	5.00%
M. moreirae	0.05	35	False	1.16	1	3.5966	-0.5452	10	5.00%
S. holbrookii	0.01	63	False	3.44	3.13	4.2833	-0.5504	13	1.00%
S. holbrookii	0.01	65	False	3.44	3.13	4.2833	-0.5504	13	1.00%
S. holbrookii	0.01	61	False	2.51	2.21	3.3520	-0.5509	10	1.00%
S. holbrookii	0.05	73	False	0.5	0.4	1.8650	-0.6000	5	5.00%
S. holbrookii	0.05	75	False	0.6	0.5	2.2440	-0.6000	6	5.00%
H. subpunctatus	0.01	72	False	0.14	0.12	3.3900	-0.6429	13	1.00%
S. holbrookii	0.05	17	False	0.67	0.57	2.5695	-0.7369	8	5.00%
S. holbrookii	0.05	23	False	0.61	0.46	1.6838	-0.7455	5	5.00%
S. holbrookii	0.05	44	False	1.16	1.01	2.9585	-0.7566	11	5.00%
H. subpunctatus	0.05	36	False	0.45	0.38	4.0907	-0.8000	12	5.00%
S. holbrookii	0.05	13	False	0.37	0.28	1.6125	-0.8190	5	5.00%
S. holbrookii	0.05	15	False	0.37	0.28	1.6125	-0.8190	5	5.00%
S. holbrookii	0.01	111	False	0.07	0	0.3371	-0.8571	1	1.00%
S. holbrookii	0.01	113	False	0.07	0	0.3371	-0.8571	1	1.00%
S. holbrookii	0.01	117	False	0.07	0	0.3371	-0.8571	1	1.00%
S. holbrookii	0.01	119	False	0.07	0	0.3371	-0.8571	1	1.00%
S. holbrookii	0.01	121	False	0.07	0	0.3371	-0.8571	1	1.00%
S. holbrookii	0.05	9	False	0.14	0.07	0.6600	-0.8571	2	5.00%
S. holbrookii	0.05	11	False	0.21	0.14	0.9614	-0.8571	3	5.00%
S. holbrookii	0.05	85	False	1.14	0.88	4.3120	-0.8602	12	5.00%
S. holbrookii	0.05	25	False	1.31	1.13	3.0246	-0.8971	9	5.00%
S. holbrookii	0.05	39	False	0.31	0.23	1.4804	-0.9571	5	5.00%
H. subpunctatus	0.01	70	False	0.14	0.1	3.0367	-1.0000	11	1.00%
S. holbrookii	0.05	41	False	0.56	0.47	2.4762	-1.0496	8	5.00%
S. holbrookii	0.05	43	False	0.83	0.74	3.4695	-1.0664	11	5.00%
S. holbrookii	0.05	45	False	1.16	1.01	4.4851	-1.0966	14	5.00%
S. holbrookii	0.05	84	False	8.57	7.31	11.5376	-1.5078	35	5.00%
S. holbrookii	0.01	71	False	4.5	4.14	6.1861	-1.5137	18	1.00%
S. holbrookii	0.01	69	False	4.14	3.78	5.5531	-1.5293	16	1.00%
S. holbrookii	0.05	82	False	9.59	8.3	11.2653	-1.5416	34	5.00%
S. holbrookii	0.01	67	False	3.83	3.47	5.2344	-1.5524	15	1.00%
S. holbrookii	0.05	76	False	6.69	5.17	7.2600	-1.6843	22	5.00%
M. moreirae	0.05	74	False	3.14	2.7	8.3997	-1.7399	23	5.00%
H. subpunctatus	0.01	68	False	0.16	0.12	3.0300	-1.8000	9	1.00%
H. subpunctatus	0.05	48	False	0.57	0.38	3.1461	-1.8174	8	5.00%
S. holbrookii	0.05	74	False	4.8	3.47	4.9364	-1.9176	15	5.00%
S. holbrookii	0.05	78	False	8.89	7.31	8.9386	-1.9936	27	5.00%
H. subpunctatus	0.01	66	False	0.15	0.11	2.8900	-2.0000	8	1.00%
S. holbrookii	0.05	80	False	14.16	12.32	10.6624	-2.0559	32	5.00%
M. moreirae	0.05	26	False	5.56	4.69	5.1533	-2.0695	14	5.00%
M. moreirae	0.05	24	False	4.69	3.83	4.7959	-2.0844	13	5.00%
H. subpunctatus	0.01	64	False	0.11	0.07	2.3133	-2.1429	6	1.00%
M. moreirae	0.05	29	False	1.96	1.72	4.5838	-2.2162	12	5.00%
M. moreirae	0.01	104	False	0.74	0.54	2.3780	-2.5527	6	1.00%
S. holbrookii	0.01	100	False	26.17	23.06	8.6297	-2.5597	23	1.00%
S. holbrookii	0.01	98	False	26.03	22.94	8.2918	-2.6307	22	1.00%
S. holbrookii	0.05	86	False	6.29	5.01	11.5084	-2.6772	37	5.00%
S. holbrookii	0.01	96	False	17.64	14.89	7.5516	-2.7065	20	1.00%
M. moreirae	0.05	94	False	8.93	8.17	8.5750	-2.7091	23	5.00%
M. moreirae	0.01	106	False	0.54	0.41	2.0385	-2.7949	5	1.00%
M. moreirae	0.01	108	False	0.54	0.41	2.0385	-2.7949	5	1.00%
H. subpunctatus	0.05	13	False	1.44	1.2	3.8877	-2.8729	9	5.00%
S. holbrookii	0.01	94	False	17.38	14.65	6.8788	-2.8968	18	1.00%
H. subpunctatus	0.05	35	False	4.68	4.29	6.7960	-2.9859	18	5.00%
M. moreirae	0.05	92	False	6.75	5.97	7.8112	-3.0172	21	5.00%
H. subpunctatus	0.05	63	False	5.64	5.12	8.8216	-3.0233	22	5.00%
S. holbrookii	0.01	46	False	9.18	7.53	9.0651	-3.0513	25	1.00%
S. holbrookii	0.01	44	False	8.96	7.32	8.7323	-3.1121	24	1.00%
S. holbrookii	0.05	58	False	20.41	18.04	11.2337	-3.1519	30	5.00%
M. moreirae	0.01	92	False	3.54	3.15	10.0830	-3.2023	28	1.00%
M. moreirae	0.01	7	False	20.31	19.22	21.9239	-3.2123	60	1.00%
M. moreirae	0.01	102	False	0.75	0.54	2.6780	-3.2337	7	1.00%
H. subpunctatus	0.05	31	False	3.21	2.89	5.3602	-3.2499	14	5.00%
H. subpunctatus	0.05	33	False	3.21	2.89	5.3602	-3.2499	14	5.00%
S. holbrookii	0.01	42	False	8.54	6.9	8.0533	-3.2621	22	1.00%
M. moreirae	0.05	23	False	3.97	3.48	5.1010	-3.2850	14	5.00%
H. subpunctatus	0.01	48	False	0.76	0.66	9.7875	-3.2994	25	1.00%
M. moreirae	0.05	65	False	3.23	2.86	9.0278	-3.3315	23	5.00%
S. holbrookii	0.01	36	False	11.4	9.63	8.4777	-3.3670	22	1.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
H. subpunctatus	0.05	29	False	2.59	2.3	4.6470	-3.4078	12	5.00%
H. subpunctatus	0.01	46	False	0.64	0.54	8.6825	-3.4306	22	1.00%
S. holbrookii	0.01	34	False	9.84	8.27	8.1219	-3.4312	21	1.00%
P. terribilis	0.05	37	False	10.2	9.23	7.0650	-3.4652	20	5.00%
H. subpunctatus	0.01	44	False	0.62	0.52	8.3325	-3.4710	21	1.00%
M. moreirae	0.05	63	False	3	2.64	8.5409	-3.5072	22	5.00%
H. subpunctatus	0.01	42	False	0.54	0.44	7.9675	-3.5492	20	1.00%
H. subpunctatus	0.05	61	False	5.04	4.57	8.5631	-3.5525	21	5.00%
M. moreirae	0.01	85	False	8.87	7.96	11.8329	-3.5620	34	1.00%
H. subpunctatus	0.05	86	False	2.3	2.04	9.8970	-3.5777	26	5.00%
S. holbrookii	0.01	32	False	12.68	10.72	7.8327	-3.5829	20	1.00%
M. moreirae	0.05	111	False	0.37	0.24	1.5700	-3.5832	4	5.00%
M. moreirae	0.01	87	False	11.21	10.21	13.4935	-3.5978	39	1.00%
M. moreirae	0.05	50	False	8.47	7.68	7.6853	-3.5994	20	5.00%
P. terribilis	0.05	53	False	7.4	6.72	7.4033	-3.6103	21	5.00%
M. moreirae	0.05	39	False	2.34	2.11	5.7007	-3.6592	14	5.00%
P. terribilis	0.05	35	False	7.88	7.08	6.3577	-3.6849	16	5.00%
S. holbrookii	0.01	88	False	23.04	20.01	14.1366	-3.6925	38	1.00%
S. holbrookii	0.01	95	False	0.88	0.73	6.5292	-3.7000	17	1.00%
H. subpunctatus	0.01	40	False	0.48	0.38	6.9375	-3.7149	17	1.00%
M. moreirae	0.05	77	False	8.16	7.27	10.6586	-3.7555	28.999	5.00%
M. moreirae	0.01	89	False	11.32	10.31	13.7625	-3.7642	40	1.00%
P. terribilis	0.05	33	False	7.08	6.34	6.0036	-3.7766	15	5.00%
M. moreirae	0.05	59	False	2.05	1.73	7.5302	-3.7778	20	5.00%
H. subpunctatus	0.01	38	False	0.46	0.37	6.5875	-3.7824	16	1.00%
S. holbrookii	0.01	86	False	18.09	15.45	13.4307	-3.7993	36	1.00%
M. moreirae	0.05	22	False	1.37	1.09	3.6378	-3.8084	9	5.00%
S. holbrookii	0.01	93	False	0.76	0.61	6.1833	-3.8158	16	1.00%
M. moreirae	0.01	61	False	9.86	9	11.8716	-3.8280	34	1.00%
M. moreirae	0.05	72	False	2.4	1.98	9.0792	-3.8528	25	5.00%
P. terribilis	0.01	4	False	750.55	665.51	26.2274	-3.8644	68	1.00%
S. holbrookii	0.01	84	False	17.67	15.03	13.1200	-3.8717	35	1.00%
S. holbrookii	0.01	30	False	9.16	7.68	6.8105	-3.8807	17	1.00%
H. subpunctatus	0.01	36	False	0.38	0.31	6.2300	-3.9020	15	1.00%
S. holbrookii	0.01	31	False	0.34	0.25	4.2371	-3.9021	11	1.00%
S. holbrookii	0.01	102	False	12.34	10.2	7.7820	-3.9035	22	1.00%
M. moreirae	0.01	59	False	8.66	7.88	11.1497	-3.9458	32	1.00%
S. holbrookii	0.01	82	False	17.25	14.69	12.7654	-3.9482	34	1.00%
S. holbrookii	0.05	20	False	0.78	0.65	3.1955	-3.9615	9	5.00%
M. moreirae	0.05	98	False	5.68	5.3	11.0472	-3.9756	28.999	5.00%
S. holbrookii	0.01	91	False	0.58	0.49	5.4885	-3.9971	14	1.00%
H. subpunctatus	0.01	75	False	5.02	4.39	7.7828	-4.0018	19	1.00%
S. holbrookii	0.05	18	False	0.52	0.39	2.5355	-4.0096	7	5.00%
S. holbrookii	0.01	101	False	1.3	1.12	7.8418	-4.0119	20	1.00%
S. holbrookii	0.01	80	False	16.91	14.35	12.4319	-4.0290	33	1.00%
M. moreirae	0.05	69	False	3.6	2.8	9.8214	-4.0384	28	5.00%
S. holbrookii	0.01	99	False	1.24	1.08	8.1025	-4.0400	21	1.00%
S. holbrookii	0.01	103	False	1.26	1.09	7.5293	-4.0559	19	1.00%
S. holbrookii	0.01	105	False	1.2	1.03	6.8693	-4.0612	17	1.00%
S. holbrookii	0.01	28	False	6.73	5.56	6.0888	-4.0636	15	1.00%
M. moreirae	0.05	48	False	7.43	6.65	7.6299	-4.0661	22	5.00%
S. holbrookii	0.05	16	False	0.28	0.19	1.9019	-4.1270	5	5.00%
M. moreirae	0.05	21	False	1.75	1.4	3.5173	-4.1561	9	5.00%
S. holbrookii	0.05	56	False	14.95	12.83	9.8070	-4.1640	25	5.00%
M. moreirae	0.05	113	False	0.1	0.04	0.5942	-4.1667	1	5.00%
S. holbrookii	0.01	78	False	14.35	11.88	11.7171	-4.1852	31	1.00%
M. moreirae	0.01	100	False	0.77	0.56	3.2380	-4.1869	9	1.00%
M. moreirae	0.05	60	False	2	1.75	3.9994	-4.1869	10	5.00%
S. holbrookii	0.05	83	False	1.68	1.5	7.6322	-4.2136	19	5.00%
H. subpunctatus	0.05	87	False	7.99	7.17	12.5775	-4.2369	33	5.00%
P. terribilis	0.01	61	False	3.89	3.18	6.7607	-4.2399	15	1.00%
S. holbrookii	0.01	97	False	1.04	0.88	7.1642	-4.2500	18	1.00%
S. holbrookii	0.01	40	False	9.21	7.47	8.7410	-4.2671	25	1.00%
M. moreirae	0.05	75	False	6.98	6.06	9.8699	-4.2859	27	5.00%
M. moreirae	0.05	96	False	4.58	4.22	9.2674	-4.2870	24	5.00%
M. moreirae	0.05	90	False	3.96	3.21	7.9476	-4.2891	19	5.00%
S. holbrookii	0.05	61	False	2.04	1.85	8.3625	-4.2947	23	5.00%
M. moreirae	0.05	73	False	6.35	5.44	9.0891	-4.3107	27	5.00%
H. subpunctatus	0.05	54	False	3.34	2.99	6.6853	-4.3229	18	5.00%
S. holbrookii	0.05	60	False	20.68	18.55	11.3132	-4.3276	32	5.00%
H. subpunctatus	0.05	77	False	3.95	3.36	11.5588	-4.3455	28.999	5.00%
H. subpunctatus	0.01	135	False	1.97	1.69	5.2047	-4.3535	13	1.00%
M. moreirae	0.05	93	False	4.13	3.24	12.2195	-4.3575	37	5.00%
S. holbrookii	0.01	38	False	9	7.26	8.3910	-4.3836	24	1.00%
P. terribilis	0.05	89	False	7.66	6.8	11.1783	-4.3926	28.999	5.00%
S. holbrookii	0.05	35	False	0.85	0.74	4.4776	-4.3986	12	5.00%
M. moreirae	0.01	101	False	9.6	8.62	11.1292	-4.4012	32	1.00%
P. terribilis	0.05	25	False	5.3	4.71	6.7224	-4.4078	17	5.00%
M. moreirae	0.05	64	False	2.37	2.09	6.1788	-4.4254	17	5.00%
H. subpunctatus	0.05	84	False	1.85	1.62	7.7430	-4.4398	20	5.00%
S. holbrookii	0.05	71	False	0.51	0.4	8.1230	-4.4500	23	5.00%
H. subpunctatus	0.05	85	False	6.37	5.68	11.1152	-4.4562	28.999	5.00%
H. subpunctatus	0.05	102	False	1.12	0.87	6.8311	-4.4626	17	5.00%
H. subpunctatus	0.01	105	False	8.83	8.26	9.8050	-4.4844	26	1.00%
H. subpunctatus	0.05	100	False	1.39	1.13	9.4211	-4.4942	23	5.00%
S. holbrookii	0.05	97	False	4.65	4.06	8.7691	-4.5012	27	5.00%
M. moreirae	0.05	51	False	2.19	1.74	5.0472	-4.5118	12	5.00%
P. terribilis	0.05	31	False	5.27	4.6	5.2744	-4.5183	13	5.00%
H. subpunctatus	0.01	133	False	1.69	1.42	4.8637	-4.5241	12	1.00%
M. moreirae	0.01	88	False	5.9	5.39	11.9611	-4.5265	34	1.00%
M. moreirae	0.01	73	False	9.17	8.37	11.4583	-4.5272	34	1.00%
P. terribilis	0.05	85	False	4.71	3.92	8.9020	-4.5386	21	5.00%
P. terribilis	0.01	96	False	749.22	674.05	10.6237	-4.5450	25	1.00%
S. holbrookii	0.05	33	False	0.74	0.63	4.0958	-4.5682	11	5.00%
M. moreirae	0.01	71	False	8.67	7.87	11.1017	-4.5717	33	1.00%
H. subpunctatus	0.01	73	False	3.87	3.3	7.0469	-4.5769	17	1.00%
M. moreirae	0.01	99	False	9.69	8.7	11.4496	-4.5811	33	1.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
S. holbrookii	0.05	59	False	0.87	0.76	6.4010	-4.5926	18	5.00%
P. terribilis	0.01	98	False	906.75	826.11	10.9878	-4.6223	26	1.00%
P. terribilis	0.01	65	False	5.13	4.35	7.8301	-4.6277	18	1.00%
P. terribilis	0.05	87	False	5.55	4.8	9.3386	-4.6319	24	5.00%
H. subpunctatus	0.05	59	False	1.22	0.93	4.6041	-4.6344	11	5.00%
H. subpunctatus	0.01	107	False	7.69	7.12	9.1251	-4.6351	24	1.00%
P. terribilis	0.01	37	False	15.78	14.15	12.5901	-4.6664	32	1.00%
M. moreirae	0.05	108	False	0.03	0	0.3367	-4.6667	1	5.00%
H. subpunctatus	0.05	52	False	2.04	1.76	5.2521	-4.6710	14	5.00%
H. subpunctatus	0.01	10	False	3.1	2.85	13.8536	-4.6739	38	1.00%
P. terribilis	0.05	91	False	7.14	6.35	9.8062	-4.6768	24	5.00%
M. moreirae	0.01	112	False	0.62	0.48	2.6114	-4.6964	6	1.00%
M. moreirae	0.01	114	False	0.62	0.48	2.6114	-4.6964	6	1.00%
M. moreirae	0.01	90	False	5.85	5.35	11.7913	-4.7151	34	1.00%
H. subpunctatus	0.01	131	False	1.42	1.16	4.5133	-4.7155	11	1.00%
H. subpunctatus	0.01	137	False	1.42	1.16	4.5133	-4.7155	11	1.00%
H. subpunctatus	0.01	141	False	1.42	1.16	4.5133	-4.7155	11	1.00%
H. subpunctatus	0.05	11	False	1.42	1.16	4.5133	-4.7155	11	5.00%
P. terribilis	0.05	80	False	81.68	62.13	10.2749	-4.7209	25	5.00%
H. subpunctatus	0.05	75	False	2.48	2.02	9.7555	-4.7223	24	5.00%
P. terribilis	0.01	36	False	1578.08	1455.89	14.2097	-4.7227	36	1.00%
M. moreirae	0.05	62	False	2.02	1.76	4.5594	-4.7317	12	5.00%
H. subpunctatus	0.05	82	False	1.62	1.4	7.0170	-4.7588	18	5.00%
P. terribilis	0.01	94	False	826.11	748.81	10.2469	-4.7598	24	1.00%
M. moreirae	0.01	51	False	7.76	6.96	10.9678	-4.7780	31	1.00%
M. moreirae	0.05	99	False	9.78	8.56	13.3891	-4.7782	40	5.00%
P. terribilis	0.01	92	False	748.91	673.8	9.8797	-4.7860	23	1.00%
H. subpunctatus	0.05	83	False	8.44	7.55	13.9240	-4.7863	37	5.00%
M. moreirae	0.05	49	False	3.51	3.09	8.8724	-4.7955	23	5.00%
P. terribilis	0.05	29	False	3.97	3.35	4.5711	-4.8035	11	5.00%
M. moreirae	0.01	69	False	8.11	7.35	10.4821	-4.8089	31	1.00%
S. holbrookii	0.01	68	False	20.42	17.99	13.6280	-4.8099	37	1.00%
H. subpunctatus	0.01	8	False	2.88	2.65	13.5033	-4.8121	37	1.00%
S. holbrookii	0.05	64	False	18.8	16.77	11.7405	-4.8178	32	5.00%
H. subpunctatus	0.01	109	False	7.95	7.37	9.1518	-4.8189	24	1.00%
M. moreirae	0.01	97	False	8.83	7.86	11.1067	-4.8244	32	1.00%
H. subpunctatus	0.05	96	False	1.19	1.05	8.2857	-4.8345	20	5.00%
M. moreirae	0.05	97	False	9.14	8.05	13.0525	-4.8630	39	5.00%
M. moreirae	0.01	67	False	7.61	6.85	10.1411	-4.8672	30	1.00%
P. terribilis	0.05	51	False	7.86	7.14	7.1108	-4.8736	20	5.00%
M. moreirae	0.01	57	False	9.11	8.28	12.3680	-4.8767	33	1.00%
M. moreirae	0.01	86	False	5.27	4.82	10.6335	-4.8812	30	1.00%
S. holbrookii	0.01	74	False	15.93	14.04	13.3145	-4.8815	35	1.00%
M. moreirae	0.01	98	False	0.44	0.3	2.5323	-4.8846	7	1.00%
P. terribilis	0.05	43	False	8.42	7.58	6.8866	-4.8905	19	5.00%
S. holbrookii	0.05	52	False	6.39	4.89	5.9215	-4.9099	14	5.00%
H. subpunctatus	0.01	129	False	1.16	0.91	4.1571	-4.9321	10	1.00%
P. terribilis	0.01	63	False	4.43	3.65	7.4807	-4.9347	17	1.00%
S. holbrookii	0.01	8	False	33.37	30.57	16.9193	-4.9367	49	1.00%
P. terribilis	0.01	34	False	1577.87	1455.69	13.0625	-4.9377	33	1.00%
H. subpunctatus	0.01	50	False	0.54	0.45	6.7925	-4.9386	18	1.00%
S. holbrookii	0.01	26	False	12.44	10.89	8.2601	-4.9438	23	1.00%
M. moreirae	0.01	55	False	8.54	7.73	12.0032	-4.9613	32	1.00%
H. subpunctatus	0.05	81	False	7.17	6.31	12.8568	-4.9773	34	5.00%
M. moreirae	0.01	49	False	6.54	5.77	10.2321	-4.9915	28.999	1.00%
S. holbrookii	0.01	76	False	12.34	10.73	11.7496	-4.9926	31	1.00%
H. subpunctatus	0.01	94	False	0.07	0.04	1.0633	-5.0000	2	1.00%
H. subpunctatus	0.01	98	False	0.07	0.04	1.0633	-5.0000	2	1.00%
H. subpunctatus	0.05	108	False	0.11	0.07	1.0433	-5.0000	2	5.00%
H. subpunctatus	0.05	110	False	0.07	0.03	0.6933	-5.0000	1	5.00%
H. subpunctatus	0.05	112	False	0.03	0	0.3433	-5.0000	0	5.00%
M. moreirae	0.01	65	False	7.35	6.62	9.8461	-5.0032	28.999	1.00%
M. moreirae	0.05	46	False	3.99	3.42	6.1890	-5.0063	17	5.00%
M. moreirae	0.05	88	False	2.16	1.64	6.0895	-5.0130	14	5.00%
H. subpunctatus	0.01	103	False	7.73	7.14	9.8651	-5.0183	26	1.00%
M. moreirae	0.05	61	False	2.61	2.25	8.3820	-5.0311	23	5.00%
P. terribilis	0.01	58	False	850.35	755.04	9.8830	-5.0349	27	1.00%
S. holbrookii	0.01	66	False	16.23	14.06	12.5540	-5.0420	34	1.00%
M. moreirae	0.01	32	False	10.23	9.31	15.9697	-5.0430	45	1.00%
H. subpunctatus	0.01	51	False	10.08	9.46	13.3650	-5.0447	36	1.00%
M. moreirae	0.01	53	False	7.99	7.21	11.6489	-5.0489	31	1.00%
H. subpunctatus	0.01	45	False	8.23	7.59	13.3391	-5.0595	34	1.00%
M. moreirae	0.01	110	False	0.48	0.34	2.2457	-5.0612	5	1.00%
M. moreirae	0.01	116	False	0.48	0.34	2.2457	-5.0612	5	1.00%
S. holbrookii	0.05	57	False	1.5	1.32	5.4948	-5.0628	15	5.00%
M. moreirae	0.05	37	False	1.44	1.23	4.8955	-5.0640	12	5.00%
P. terribilis	0.01	59	False	5.56	4.67	8.2442	-5.0852	21	1.00%
S. holbrookii	0.01	72	False	19.75	16.81	13.3218	-5.0987	37	1.00%
M. moreirae	0.01	47	False	5.9	5.18	9.8528	-5.1105	28	1.00%
H. subpunctatus	0.01	43	False	7.79	7.19	12.9709	-5.1110	33	1.00%
S. holbrookii	0.05	62	False	14.94	13.16	10.3965	-5.1155	28	5.00%
S. holbrookii	0.01	64	False	15.71	13.54	12.1967	-5.1185	33	1.00%
P. terribilis	0.01	38	False	1217.64	1116.06	10.2733	-5.1355	28	1.00%
P. terribilis	0.05	49	False	6.46	5.78	6.3987	-5.1530	18	5.00%
S. holbrookii	0.01	57	False	3.09	2.71	5.2746	-5.1669	15	1.00%
S. holbrookii	0.01	59	False	3.09	2.71	5.2746	-5.1669	15	1.00%
S. holbrookii	0.05	68	False	10.61	8.79	11.5765	-5.1679	33	5.00%
H. subpunctatus	0.01	16	False	4.42	4.14	15.4397	-5.1731	41	1.00%
H. subpunctatus	0.01	125	False	0.91	0.71	3.7995	-5.1805	9	1.00%
H. subpunctatus	0.01	127	False	0.91	0.71	3.7995	-5.1805	9	1.00%
H. subpunctatus	0.01	139	False	0.91	0.71	3.7995	-5.1805	9	1.00%
H. subpunctatus	0.01	143	False	0.91	0.71	3.7995	-5.1805	9	1.00%
H. subpunctatus	0.01	145	False	0.91	0.71	3.7995	-5.1805	9	1.00%
H. subpunctatus	0.01	14	False	4.33	4.05	15.0697	-5.1904	40	1.00%
M. moreirae	0.01	45	False	5.49	4.79	9.4862	-5.2049	27	1.00%
H. subpunctatus	0.05	50	False	1.21	0.96	4.1972	-5.2157	11	5.00%
S. holbrookii	0.05	66	False	15.11	13.04	12.0449	-5.2265	35	5.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
S. holbrookii	0.05	47	False	3.81	3.59	8.6569	-5.2304	25	5.00%
H. subpunctatus	0.05	79	False	5.48	4.67	11.4095	-5.2364	30	5.00%
H. subpunctatus	0.05	104	False	0.28	0.21	2.4533	-5.2381	6	5.00%
H. subpunctatus	0.05	45	False	1.83	1.66	5.3542	-5.2413	13	5.00%
S. holbrookii	0.01	70	False	16.81	14.63	12.6182	-5.2467	35	1.00%
P. terribilis	0.01	67	False	5.18	4.4	7.8455	-5.2535	18	1.00%
H. subpunctatus	0.01	12	False	4.14	3.86	14.7145	-5.2554	39	1.00%
M. moreirae	0.05	91	False	2.01	1.45	10.3097	-5.2565	32	5.00%
S. holbrookii	0.05	87	False	0.77	0.57	3.8373	-5.2761	11	5.00%
H. subpunctatus	0.05	92	False	2.01	1.77	8.5064	-5.2879	23	5.00%
M. moreirae	0.01	23	False	5.16	4.52	9.7718	-5.2885	28	1.00%
M. moreirae	0.01	43	False	5.18	4.51	9.1485	-5.2922	26	1.00%
S. holbrookii	0.01	62	False	13.54	11.53	11.4756	-5.2950	31	1.00%
M. moreirae	0.01	63	False	6.89	6.16	9.2257	-5.2965	27	1.00%
S. holbrookii	0.05	50	False	4.45	3.09	4.5396	-5.3246	10	5.00%
P. terribilis	0.05	88	False	357.29	315.87	9.8971	-5.3335	25	5.00%
M. moreirae	0.01	5	False	46.97	44.91	34.7469	-5.3342	97	1.00%
S. holbrookii	0.01	89	False	0.77	0.65	5.9006	-5.3349	14	1.00%
H. subpunctatus	0.01	17	False	4.55	3.95	10.3244	-5.3402	28.999	1.00%
M. moreirae	0.05	95	False	6.44	5.4	11.3848	-5.3421	34	5.00%
P. terribilis	0.01	44	False	1684.1	1559.02	12.3415	-5.3426	33	1.00%
M. moreirae	0.05	47	False	2.91	2.52	7.5746	-5.3525	19	5.00%
P. terribilis	0.01	42	False	1559.07	1438.67	11.9727	-5.3707	32	1.00%
M. moreirae	0.01	21	False	4.83	4.19	9.4182	-5.3800	27	1.00%
S. holbrookii	0.01	60	False	13.04	11.04	11.1098	-5.3834	30	1.00%
S. holbrookii	0.01	53	False	2.71	2.33	4.9243	-5.3837	14	1.00%
S. holbrookii	0.01	55	False	2.71	2.33	4.9243	-5.3837	14	1.00%
H. subpunctatus	0.05	41	False	1.66	1.49	4.9925	-5.3930	12	5.00%
H. subpunctatus	0.05	43	False	1.66	1.49	4.9925	-5.3930	12	5.00%
M. moreirae	0.01	41	False	4.79	4.12	8.7828	-5.3962	25	1.00%
S. holbrookii	0.05	31	False	0.32	0.22	3.0670	-5.4000	7	5.00%
H. subpunctatus	0.05	51	False	3.28	2.85	6.1277	-5.4021	17	5.00%
H. subpunctatus	0.01	52	False	0.48	0.39	7.1675	-5.4042	19	1.00%
H. subpunctatus	0.05	25	False	2.99	2.68	6.9488	-5.4056	19	5.00%
H. subpunctatus	0.01	49	False	8.9	8.25	12.9654	-5.4243	34	1.00%
S. holbrookii	0.01	25	False	0.23	0.21	5.0600	-5.4286	14	1.00%
H. subpunctatus	0.01	15	False	4.23	3.65	9.9612	-5.4350	28	1.00%
M. moreirae	0.05	102	False	0.41	0.33	4.9565	-5.4378	13	5.00%
H. subpunctatus	0.01	13	False	3.95	3.37	9.6041	-5.4511	27	1.00%
M. moreirae	0.05	19	False	0.55	0.33	1.8767	-5.4596	4	5.00%
S. holbrookii	0.01	58	False	13.71	11.57	10.7384	-5.4606	28.999	1.00%
M. moreirae	0.01	19	False	4.51	3.92	9.0150	-5.4682	26	1.00%
H. subpunctatus	0.01	123	False	0.71	0.56	3.4375	-5.4686	8	1.00%
H. subpunctatus	0.05	9	False	0.71	0.56	3.4375	-5.4686	8	5.00%
P. terribilis	0.05	47	False	5.11	4.44	5.6783	-5.4915	16	5.00%
M. moreirae	0.05	42	False	2.67	2.3	5.1289	-5.4929	14	5.00%
M. moreirae	0.05	44	False	2.67	2.3	5.1289	-5.4929	14	5.00%
H. subpunctatus	0.05	57	False	0.76	0.51	3.1976	-5.5039	7	5.00%
S. holbrookii	0.01	21	False	0.21	0.19	4.6900	-5.5385	13	1.00%
S. holbrookii	0.01	23	False	0.21	0.19	4.6900	-5.5385	13	1.00%
M. moreirae	0.05	67	False	3.88	3.39	9.5481	-5.5401	27	5.00%
M. moreirae	0.05	86	False	1.59	1.07	4.9569	-5.5457	11	5.00%
H. subpunctatus	0.05	106	False	0.22	0.15	2.0967	-5.5556	5	5.00%
P. terribilis	0.05	86	False	272.58	236.33	8.0232	-5.5593	20	5.00%
S. holbrookii	0.01	56	False	12.06	10.06	10.3825	-5.5740	28	1.00%
M. moreirae	0.05	33	False	1.14	0.96	4.0403	-5.5861	12	5.00%
H. subpunctatus	0.01	41	False	7.19	6.6	12.6009	-5.6026	32	1.00%
M. moreirae	0.01	17	False	4.03	3.43	8.2697	-5.6099	24	1.00%
H. subpunctatus	0.01	47	False	8.13	7.52	12.5805	-5.6464	33	1.00%
P. terribilis	0.05	41	False	12.87	11.87	8.6223	-5.6566	24	5.00%
M. moreirae	0.01	83	False	14.65	13.56	14.8510	-5.6568	44	1.00%
M. moreirae	0.05	81	False	3.6	3.19	8.0021	-5.6610	21	5.00%
P. terribilis	0.05	23	False	3.05	2.66	5.6327	-5.6633	15	5.00%
S. holbrookii	0.01	54	False	11.57	9.58	10.0327	-5.6789	27	1.00%
P. terribilis	0.01	100	False	656.3	593.34	10.2391	-5.6931	26	1.00%
S. holbrookii	0.01	87	False	0.67	0.55	5.2746	-5.7018	12	1.00%
P. terribilis	0.01	55	False	3.84	3.16	6.4683	-5.7065	16	1.00%
M. moreirae	0.05	84	False	1.54	1.04	4.6029	-5.7078	10	5.00%
P. terribilis	0.01	57	False	5.14	4.24	7.9142	-5.7081	20	1.00%
P. terribilis	0.05	39	False	10.7	9.78	7.5285	-5.7184	21	5.00%
S. holbrookii	0.05	55	False	1.14	0.96	4.0032	-5.7201	11	5.00%
S. holbrookii	0.05	48	False	4.48	4.07	6.2106	-5.7202	21	5.00%
P. terribilis	0.01	90	False	826.18	748.88	10.2892	-5.7229	23	1.00%
M. moreirae	0.01	81	False	13.9	12.89	14.5017	-5.7341	43	1.00%
M. moreirae	0.05	104	False	0.3	0.23	3.9725	-5.7361	11	5.00%
P. terribilis	0.01	32	False	1337.78	1223.2	14.6911	-5.7460	38	1.00%
S. holbrookii	0.05	37	False	0.72	0.57	2.8607	-5.7493	7	5.00%
S. holbrookii	0.01	85	False	0.6	0.48	4.9203	-5.7676	11	1.00%
H. subpunctatus	0.01	147	False	0.56	0.43	3.0862	-5.7725	7	1.00%
P. terribilis	0.05	73	False	8.62	7.71	12.9161	-5.7813	33	5.00%
S. holbrookii	0.01	52	False	11.09	9.37	9.6737	-5.7892	26	1.00%
H. subpunctatus	0.05	94	False	1.27	1.1	7.1174	-5.8013	19	5.00%
H. subpunctatus	0.05	90	False	3.08	2.73	12.4441	-5.8051	35	5.00%
M. moreirae	0.05	79	False	8.48	7.64	9.2157	-5.8069	24	5.00%
S. holbrookii	0.01	83	False	0.48	0.38	4.2100	-5.8154	9	1.00%
H. subpunctatus	0.05	65	False	1.6	1.32	6.6412	-5.8257	18	5.00%
P. terribilis	0.01	88	False	602.03	532.61	9.5758	-5.8383	21	1.00%
P. terribilis	0.01	40	False	1683.97	1558.9	11.1955	-5.8597	30	1.00%
H. subpunctatus	0.01	115	False	1.7	1.43	4.8301	-5.8782	12	1.00%
M. moreirae	0.01	39	False	3.26	2.63	7.3260	-5.8869	21	1.00%
S. holbrookii	0.05	14	False	0.75	0.63	4.0894	-5.9000	11	5.00%
P. terribilis	0.01	43	False	17.81	16.17	13.4222	-5.9055	36	1.00%
P. terribilis	0.05	78	False	297.12	257.86	12.8085	-5.9084	33	5.00%
P. terribilis	0.05	63	False	5.68	4.98	8.8641	-5.9189	22	5.00%
P. terribilis	0.05	93	False	4.64	4.03	9.4227	-5.9199	26	5.00%
P. terribilis	0.01	86	False	601.92	532.51	9.1977	-5.9228	20	1.00%
M. moreirae	0.05	54	False	9.53	8.78	8.0428	-5.9421	23	5.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
M. moreirae	0.01	77	False	11.3	10.38	12.8165	-5.9464	38	1.00%
P. territilis	0.05	76	False	186.07	156.14	10.9329	-5.9555	27	5.00%
S. holbrookii	0.01	50	False	9.37	7.73	8.9645	-5.9573	24	1.00%
P. territilis	0.05	69	False	4.19	3.59	9.2871	-5.9593	23	5.00%
P. territilis	0.01	8	False	835.71	745.58	25.7774	-5.9603	68	1.00%
M. moreirae	0.01	79	False	11.96	11.01	13.4553	-5.9652	40	1.00%
P. territilis	0.01	102	False	721.36	655.97	9.8574	-5.9861	25	1.00%
P. territilis	0.01	41	False	16.76	15.16	13.0535	-5.9886	35	1.00%
H. subpunctatus	0.01	97	False	8.84	8.05	10.7811	-5.9913	28.999	1.00%
H. subpunctatus	0.01	96	False	0.04	0.01	0.7067	-6.0000	1	1.00%
S. holbrookii	0.01	19	False	0.15	0.13	3.6300	-6.0000	10	1.00%
P. territilis	0.05	71	False	7.2	6.35	11.8275	-6.0130	30	5.00%
P. territilis	0.05	94	False	888.05	800.06	11.4751	-6.0133	28	5.00%
H. subpunctatus	0.01	33	False	6.23	5.58	11.9206	-6.0279	30	1.00%
M. moreirae	0.05	57	False	4.4	3.94	7.8945	-6.0317	21	5.00%
M. moreirae	0.01	37	False	2.99	2.39	7.0048	-6.0333	20	1.00%
M. moreirae	0.05	106	False	0.18	0.13	2.6600	-6.0417	7	5.00%
P. territilis	0.05	68	False	349.3	298.95	8.0112	-6.0425	21	5.00%
P. territilis	0.01	12	False	1173.68	1067.98	25.0313	-6.0443	66	1.00%
H. subpunctatus	0.01	101	False	9.45	8.82	10.9457	-6.0481	28	1.00%
H. subpunctatus	0.05	64	False	0.78	0.71	5.6843	-6.0491	15	5.00%
P. territilis	0.01	6	False	661.99	580.55	25.4243	-6.0513	67	1.00%
M. moreirae	0.05	55	False	2.3	1.98	6.0988	-6.0704	16	5.00%
P. territilis	0.01	39	False	15.75	14.2	12.6801	-6.0731	34	1.00%
H. subpunctatus	0.05	7	False	0.43	0.33	2.7285	-6.0998	6	5.00%
S. holbrookii	0.01	48	False	7.94	6.43	8.6164	-6.1018	23	1.00%
H. subpunctatus	0.01	99	False	9	8.24	11.5267	-6.1070	28.999	1.00%
S. holbrookii	0.01	92	False	25.75	22.61	12.0497	-6.1125	32	1.00%
H. subpunctatus	0.05	72	False	1.21	1.09	6.7813	-6.1137	18	5.00%
H. subpunctatus	0.05	74	False	1.21	1.09	6.7813	-6.1137	18	5.00%
P. territilis	0.05	79	False	8.53	7.56	13.8145	-6.1232	37	5.00%
P. territilis	0.01	10	False	968.9	874.47	24.6791	-6.1296	65	1.00%
P. territilis	0.01	53	False	2.94	2.35	5.3994	-6.1308	13	1.00%
M. moreirae	0.01	75	False	10.4	9.49	12.4353	-6.1313	37	1.00%
H. subpunctatus	0.01	113	False	1.43	1.17	4.4712	-6.1395	11	1.00%
P. territilis	0.01	51	False	2.54	2.01	5.0421	-6.1455	12	1.00%
H. subpunctatus	0.01	95	False	7.61	6.88	10.0629	-6.1645	27	1.00%
H. subpunctatus	0.01	69	False	4.49	3.8	8.8164	-6.1694	24	1.00%
H. subpunctatus	0.01	31	False	5.49	4.84	11.2073	-6.1797	28	1.00%
H. subpunctatus	0.01	39	False	7.91	7.25	13.3751	-6.1814	34	1.00%
M. moreirae	0.05	52	False	8.03	7.29	7.3249	-6.1829	21	5.00%
M. moreirae	0.01	35	False	2.63	2.05	6.6283	-6.1859	19	1.00%
P. territilis	0.05	27	False	3.84	3.29	5.0062	-6.1896	13	5.00%
M. moreirae	0.01	13	False	18.72	17.66	21.1557	-6.1901	60	1.00%
P. territilis	0.05	30	False	828.68	740.15	6.5660	-6.1937	16	5.00%
P. territilis	0.05	75	False	8.08	7.3	13.6597	-6.1938	31	5.00%
M. moreirae	0.01	25	False	5	4.4	9.8801	-6.1945	28.999	1.00%
M. moreirae	0.01	93	False	10.65	9.71	13.0242	-6.1986	39	1.00%
P. territilis	0.05	56	False	110.93	87.66	10.1150	-6.1993	26	5.00%
M. moreirae	0.01	96	False	3.23	2.93	6.7306	-6.2254	18	1.00%
M. moreirae	0.01	118	False	0.2	0.06	1.5836	-6.2286	3	1.00%
M. moreirae	0.01	78	False	7.82	7.26	12.1472	-6.2434	34	1.00%
M. moreirae	0.01	11	False	18.34	17.32	20.8323	-6.2536	59	1.00%
S. holbrookii	0.01	90	False	22.61	19.61	11.3363	-6.2544	30	1.00%
H. subpunctatus	0.05	103	False	1	0.8	4.2014	-6.2608	11	5.00%
P. territilis	0.05	99	False	3.72	3.06	8.8243	-6.2778	24	5.00%
M. moreirae	0.01	95	False	11.89	10.91	13.4927	-6.2779	40	1.00%
M. moreirae	0.01	76	False	7.31	6.84	11.8179	-6.2840	33	1.00%
P. territilis	0.01	64	False	1053.17	950.15	11.0029	-6.2846	30	1.00%
P. territilis	0.05	97	False	4.48	3.82	10.9921	-6.2904	30	5.00%
P. territilis	0.05	67	False	3.33	2.77	8.2209	-6.2916	20	5.00%
M. moreirae	0.01	91	False	10.09	9.21	12.6804	-6.2974	38	1.00%
H. subpunctatus	0.05	23	False	2.48	2.18	6.9430	-6.3012	19	5.00%
H. subpunctatus	0.05	97	False	2.72	2.28	8.2320	-6.3042	19	5.00%
S. holbrookii	0.05	12	False	0.55	0.47	3.4936	-6.3136	9	5.00%
P. territilis	0.01	54	False	1098.76	1001.33	11.7839	-6.3169	32	1.00%
P. territilis	0.01	52	False	1001.41	908.67	11.4184	-6.3301	31	1.00%
P. territilis	0.01	49	False	2.01	1.54	4.3078	-6.3331	10	1.00%
P. territilis	0.01	18	False	1080.57	974.74	21.3841	-6.3373	57.999	1.00%
H. subpunctatus	0.05	88	False	2.1	1.79	10.2777	-6.3375	28.999	5.00%
P. territilis	0.01	50	False	908.75	819.6	11.0517	-6.3396	30	1.00%
P. territilis	0.01	46	False	738.26	660.62	10.3191	-6.3490	28	1.00%
M. moreirae	0.01	33	False	2.35	1.77	6.2904	-6.3530	18	1.00%
H. subpunctatus	0.05	17	False	1.9	1.65	4.5045	-6.3607	11	5.00%
P. territilis	0.05	46	False	1393.75	1287.25	11.6616	-6.3648	28.999	5.00%
M. moreirae	0.05	38	False	1.38	1.11	3.7433	-6.3665	10	5.00%
M. moreirae	0.05	40	False	1.38	1.11	3.7433	-6.3665	10	5.00%
P. territilis	0.05	74	False	61.57	44.67	7.6142	-6.3702	18	5.00%
M. moreirae	0.01	9	False	17.78	16.75	21.0508	-6.3730	60	1.00%
P. territilis	0.01	62	False	1053.07	950.06	10.6289	-6.3737	28.999	1.00%
P. territilis	0.05	42	False	635.03	558.53	6.1479	-6.3754	15	5.00%
H. subpunctatus	0.01	37	False	7.01	6.34	12.6431	-6.3954	32	1.00%
P. territilis	0.05	40	False	558.53	487.71	5.7858	-6.4108	14	5.00%
P. territilis	0.01	47	False	4.26	3.45	6.5319	-6.4300	15	1.00%
P. territilis	0.05	96	False	1071	977.85	12.9320	-6.4327	33	5.00%
P. territilis	0.01	56	False	1435.37	1314.87	12.5020	-6.4344	34	1.00%
H. subpunctatus	0.01	117	False	1.17	0.92	4.1135	-6.4364	10	1.00%
H. subpunctatus	0.01	119	False	1.17	0.92	4.1135	-6.4364	10	1.00%
H. subpunctatus	0.05	47	False	1.6	1.21	4.6662	-6.4414	13	5.00%
H. subpunctatus	0.05	49	False	1.6	1.21	4.6662	-6.4414	13	5.00%
P. territilis	0.05	38	False	487.71	421.58	5.4217	-6.4516	13	5.00%
H. subpunctatus	0.01	29	False	5.16	4.58	10.8134	-6.4658	28	1.00%
P. territilis	0.01	16	False	1080.57	974.76	21.2929	-6.4716	57.999	1.00%
S. holbrookii	0.05	100	False	8.58	7.19	8.1492	-6.4721	22	5.00%
H. subpunctatus	0.05	105	False	0.57	0.46	2.7672	-6.4773	8	5.00%
H. subpunctatus	0.01	35	False	6.62	5.95	12.2836	-6.4809	31	1.00%
M. moreirae	0.05	45	False	3.37	2.99	8.6398	-6.4887	23	5.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
S. holbrookii	0.05	69	False	1.36	1.28	12.3200	-6.5000	33	5.00%
S. holbrookii	0.01	14	False	25.53	23.1	12.9439	-6.5005	38	1.00%
S. holbrookii	0.05	54	False	24.24	21.26	9.4654	-6.5014	24	5.00%
H. subpunctatus	0.05	101	False	2.09	1.76	7.3939	-6.5040	19	5.00%
H. subpunctatus	0.05	73	False	1.71	1.45	8.5075	-6.5107	20	5.00%
M. moreirae	0.01	31	False	2.06	1.53	5.9121	-6.5108	17	1.00%
P. terribilis	0.05	54	False	190.23	160.69	8.7242	-6.5184	21	5.00%
P. terribilis	0.05	70	False	402.64	349.51	9.4692	-6.5283	24	5.00%
M. moreirae	0.05	53	False	2.33	1.94	5.4388	-6.5301	13	5.00%
H. subpunctatus	0.01	26	False	0.84	0.75	9.0717	-6.5308	26	1.00%
P. terribilis	0.01	60	False	950.06	850.23	9.8815	-6.5321	27	1.00%
P. terribilis	0.01	111	False	9.68	9.06	10.0215	-6.5368	28	1.00%
M. moreirae	0.05	34	False	1.3	1.12	4.8310	-6.5497	14	5.00%
P. terribilis	0.01	33	False	23.05	21.14	16.5006	-6.5607	44	1.00%
H. subpunctatus	0.05	21	False	1.93	1.7	6.2286	-6.5650	17	5.00%
P. terribilis	0.01	30	False	1822.38	1686.39	15.4599	-6.5733	42	1.00%
S. holbrookii	0.01	75	False	0.63	0.52	4.9715	-6.5813	16	1.00%
M. moreirae	0.01	74	False	7.74	7.22	11.5012	-6.5836	32	1.00%
P. terribilis	0.05	95	False	6.4	5.65	10.9202	-6.5860	30	5.00%
H. subpunctatus	0.05	15	False	1.65	1.43	4.1469	-6.5941	10	5.00%
S. holbrookii	0.01	81	False	0.45	0.33	3.8583	-6.6095	8	1.00%
M. moreirae	0.05	89	False	2.55	2.24	10.6472	-6.6103	30	5.00%
H. subpunctatus	0.01	27	False	4.57	4.02	10.0848	-6.6104	26	1.00%
M. moreirae	0.01	72	False	7.23	6.8	11.1719	-6.6378	31	1.00%
P. terribilis	0.05	83	False	8.28	7.27	11.8321	-6.6491	31	5.00%
H. subpunctatus	0.01	23	False	4.11	3.57	9.3503	-6.6529	24	1.00%
M. moreirae	0.05	71	False	4.49	3.64	7.8403	-6.6537	23	5.00%
P. terribilis	0.01	109	False	9.06	8.46	9.6495	-6.6541	27	1.00%
P. terribilis	0.01	113	False	9.06	8.46	9.6495	-6.6541	27	1.00%
P. terribilis	0.01	115	False	9.06	8.46	9.6495	-6.6541	27	1.00%
P. terribilis	0.05	45	False	7.86	7.06	6.8794	-6.6728	18	5.00%
P. terribilis	0.05	61	False	5.54	4.77	7.5039	-6.6766	19	5.00%
S. holbrookii	0.05	96	False	10.48	8.83	7.8124	-6.6784	21	5.00%
H. subpunctatus	0.05	99	False	2.92	2.44	8.9262	-6.6795	23	5.00%
H. subpunctatus	0.01	65	False	4.54	3.94	8.5783	-6.6859	23	1.00%
M. moreirae	0.05	70	False	4.49	3.99	9.7942	-6.6865	25	5.00%
H. subpunctatus	0.01	22	False	3.36	3.14	15.1529	-6.6883	42	1.00%
P. terribilis	0.01	28	False	1686.39	1554.61	14.7124	-6.6884	40	1.00%
H. subpunctatus	0.05	68	False	0.86	0.76	6.7071	-6.6891	19	5.00%
P. terribilis	0.05	72	False	79.4	60.41	9.1402	-6.7026	23	5.00%
P. terribilis	0.05	65	False	2.5	2.08	7.1449	-6.7127	17	5.00%
P. terribilis	0.01	35	False	23.87	21.8	16.8528	-6.7147	45	1.00%
H. subpunctatus	0.01	5	False	56.21	54.02	35.9661	-6.7232	98	1.00%
S. holbrookii	0.01	24	False	15.79	14.05	10.9487	-6.7249	33	1.00%
M. moreirae	0.05	27	False	6.91	6.35	6.6188	-6.7390	20	5.00%
S. holbrookii	0.01	6	False	65.25	61.72	24.4954	-6.7488	71	1.00%
H. subpunctatus	0.05	95	False	2.23	1.8	7.1538	-6.7543	16	5.00%
P. terribilis	0.05	20	False	513.2	456.18	6.1148	-6.7553	14	5.00%
H. subpunctatus	0.05	98	False	1.47	1.31	9.7646	-6.7654	23	5.00%
M. moreirae	0.05	85	False	2.49	2.26	9.5201	-6.7665	24	5.00%
P. terribilis	0.01	48	False	1001.28	908.58	10.6736	-6.7667	28.999	1.00%
H. subpunctatus	0.01	111	False	4.27	3.87	7.7702	-6.7686	21	1.00%
P. terribilis	0.01	26	False	1426.65	1305.08	14.3587	-6.7721	39	1.00%
S. holbrookii	0.05	10	False	0.39	0.31	2.9061	-6.7722	7	5.00%
H. subpunctatus	0.01	25	False	4.3	3.75	9.7399	-6.7739	25	1.00%
P. terribilis	0.01	107	False	8.46	7.88	9.2891	-6.7761	26	1.00%
H. subpunctatus	0.01	121	False	0.92	0.72	3.7559	-6.7778	9	1.00%
M. moreirae	0.05	83	False	2.37	2.08	7.8063	-6.7781	21	5.00%
P. terribilis	0.05	14	False	136.83	97.21	3.2795	-6.7839	7	5.00%
P. terribilis	0.05	92	False	488.65	434.14	10.4190	-6.7893	25	5.00%
P. terribilis	0.05	36	False	199.85	155.98	3.5983	-6.8048	8	5.00%
M. moreirae	0.05	20	False	0.64	0.48	2.7292	-6.8177	6	5.00%
S. holbrookii	0.05	49	False	0.74	0.56	4.5443	-6.8179	10	5.00%
M. moreirae	0.01	50	False	12.59	11.74	13.6851	-6.8193	36	1.00%
H. subpunctatus	0.01	67	False	4.56	3.89	8.9292	-6.8253	24	1.00%
H. subpunctatus	0.01	20	False	3.05	2.83	14.0429	-6.8447	39	1.00%
P. terribilis	0.05	81	False	8.74	7.76	12.6723	-6.8521	35	5.00%
M. moreirae	0.05	87	False	2.01	1.85	9.4059	-6.8579	26	5.00%
M. moreirae	0.01	70	False	7.22	6.79	10.8219	-6.8591	30	1.00%
H. subpunctatus	0.05	80	False	1.74	1.53	8.9008	-6.8660	22	5.00%
H. subpunctatus	0.05	39	False	3.28	2.97	7.6359	-6.8725	21	5.00%
M. moreirae	0.01	48	False	11.74	10.9	12.9580	-6.8915	34	1.00%
M. moreirae	0.01	68	False	6.8	6.38	10.4543	-6.8944	28.999	1.00%
P. terribilis	0.05	77	False	6.94	6.15	12.3515	-6.9043	34	5.00%
P. terribilis	0.01	91	False	9.16	8.41	9.8435	-6.9053	27	1.00%
P. terribilis	0.01	95	False	9.47	8.64	10.5604	-6.9075	28	1.00%
P. terribilis	0.01	93	False	9.3	8.5	10.2057	-6.9189	27	1.00%
M. moreirae	0.05	43	False	2.7	2.34	7.5950	-6.9193	20	5.00%
H. subpunctatus	0.05	19	False	1.28	1.09	4.7198	-6.9235	13	5.00%
P. terribilis	0.05	8	False	42.32	28.94	2.1501	-6.9250	4	5.00%
P. terribilis	0.05	98	False	649.55	579.67	13.0425	-6.9304	35	5.00%
H. subpunctatus	0.01	18	False	2.79	2.57	13.3491	-6.9403	37	1.00%
P. terribilis	0.01	87	False	8.41	7.71	9.1114	-6.9446	25	1.00%
P. terribilis	0.05	52	False	112.14	90.35	8.8059	-6.9519	20	5.00%
H. subpunctatus	0.05	66	False	0.68	0.59	6.3961	-6.9624	16	5.00%
H. subpunctatus	0.01	63	False	5.9	5.29	11.4586	-6.9680	31	1.00%
M. moreirae	0.05	36	False	0.86	0.71	3.0586	-6.9767	8	5.00%
P. terribilis	0.05	100	False	362.96	319.56	7.5782	-6.9804	19	5.00%
P. terribilis	0.05	18	False	352.13	303.89	5.0182	-6.9987	11	5.00%
H. subpunctatus	0.05	37	False	2.97	2.66	7.2684	-7.0000	20	5.00%
H. subpunctatus	0.01	149	False	0.17	0.11	1.6933	-7.0000	3	1.00%
M. moreirae	0.05	100	False	5.45	5.06	11.0366	-7.0031	31	5.00%
P. terribilis	0.01	89	False	8.5	7.79	9.4748	-7.0065	26	1.00%
S. holbrookii	0.05	101	False	2.91	2.48	5.2479	-7.0124	17	5.00%
H. subpunctatus	0.01	11	False	34.78	33.03	24.6991	-7.0301	69	1.00%
M. moreirae	0.05	31	False	1.95	1.73	4.8078	-7.0384	15	5.00%
H. subpunctatus	0.01	62	False	0.28	0.24	6.9867	-7.0500	19	1.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
M. moreirae	0.01	46	False	10.42	9.62	12.5543	-7.0503	33	1.00%
P. terribilis	0.01	14	False	792.49	712.92	23.9634	-7.0561	65	1.00%
H. subpunctatus	0.01	6	False	10.3	9.8	27.5561	-7.0908	78	1.00%
M. moreirae	0.05	101	False	5.15	4.65	4.7497	-7.0992	14	5.00%
S. holbrookii	0.01	10	False	29.51	26.68	15.2589	-7.1044	45	1.00%
S. holbrookii	0.01	18	False	22.16	20.04	10.8449	-7.1060	32	1.00%
H. subpunctatus	0.01	21	False	3.73	3.21	8.6288	-7.1138	24	1.00%
S. holbrookii	0.01	12	False	27.3	24.77	15.4894	-7.1292	46	1.00%
M. moreirae	0.01	44	False	9.62	8.82	11.8271	-7.1365	31	1.00%
H. subpunctatus	0.05	78	False	1.83	1.64	7.8308	-7.1443	20	5.00%
H. subpunctatus	0.01	9	False	35.64	33.9	24.3401	-7.1453	68	1.00%
S. holbrookii	0.05	51	False	1.03	0.86	6.6669	-7.1502	15	5.00%
S. holbrookii	0.05	53	False	0.98	0.85	4.8791	-7.1639	12	5.00%
S. holbrookii	0.01	20	False	20.83	18.78	11.1369	-7.1650	33	1.00%
H. subpunctatus	0.01	19	False	3.53	3.02	8.2559	-7.1658	23	1.00%
S. holbrookii	0.01	17	False	0.68	0.63	6.3810	-7.1741	18	1.00%
P. terribilis	0.05	44	False	1175.84	1077.2	10.0097	-7.1745	25	5.00%
H. subpunctatus	0.01	93	False	9.38	8.66	10.9271	-7.1751	28.999	1.00%
P. terribilis	0.01	24	False	1306.47	1188.98	17.6327	-7.1776	48	1.00%
P. terribilis	0.05	102	False	236.9	198.11	6.1301	-7.1782	15	5.00%
P. terribilis	0.05	48	False	1502.42	1393.7	12.0341	-7.1822	28.999	5.00%
M. moreirae	0.01	38	False	11.18	10.28	15.9994	-7.1852	46	1.00%
S. holbrookii	0.01	16	False	19.81	17.83	9.8545	-7.1915	28.999	1.00%
S. holbrookii	0.01	22	False	19.66	17.68	11.3961	-7.1962	34	1.00%
P. terribilis	0.01	85	False	7.31	6.66	8.7532	-7.2018	24	1.00%
P. terribilis	0.05	104	False	106.25	81.05	3.2532	-7.2038	7	5.00%
P. terribilis	0.01	22	False	1306.28	1188.8	17.2496	-7.2061	47	1.00%
H. subpunctatus	0.05	46	False	0.74	0.54	5.9397	-7.2092	17	5.00%
H. subpunctatus	0.05	40	False	1.16	1.06	6.9863	-7.2105	20	5.00%
M. moreirae	0.01	15	False	8.94	8.12	12.4445	-7.2175	35	1.00%
H. subpunctatus	0.01	91	False	8.66	7.99	10.2400	-7.2303	27	1.00%
H. subpunctatus	0.01	55	False	6.03	5.49	11.8935	-7.2370	33	1.00%
P. terribilis	0.01	31	False	19.09	17.34	14.7100	-7.2399	39	1.00%
H. subpunctatus	0.05	27	False	2.47	2.11	5.4743	-7.2413	16	5.00%
H. subpunctatus	0.01	89	False	7.99	7.33	9.5304	-7.2537	25	1.00%
H. subpunctatus	0.01	60	False	0.25	0.21	6.6300	-7.2632	18	1.00%
M. moreirae	0.01	6	False	44.93	43.06	26.6347	-7.2648	75	1.00%
M. moreirae	0.01	34	False	10.28	9.38	15.2819	-7.2658	44	1.00%
P. terribilis	0.01	140	False	28.94	18.06	1.7891	-7.2685	3	1.00%
P. terribilis	0.01	144	False	28.94	18.06	1.7891	-7.2685	3	1.00%
P. terribilis	0.01	20	False	1188.98	1076.48	16.8819	-7.2725	46	1.00%
S. holbrookii	0.05	98	False	9.95	8.32	7.8323	-7.2745	22	5.00%
S. holbrookii	0.01	73	False	0.57	0.47	4.6682	-7.2778	15	1.00%
M. moreirae	0.05	68	False	3.31	2.92	8.0270	-7.2872	20	5.00%
S. holbrookii	0.05	42	False	1.03	0.91	7.8223	-7.2968	25	5.00%
M. moreirae	0.01	36	False	10.53	9.63	15.6363	-7.3052	45	1.00%
S. holbrookii	0.05	94	False	15.75	13.73	8.6885	-7.3082	23	5.00%
M. moreirae	0.01	42	False	8.28	7.59	11.0700	-7.3182	28.999	1.00%
P. terribilis	0.05	59	False	4.69	3.97	7.1559	-7.3289	19	5.00%
S. holbrookii	0.05	99	False	5.75	5.1	10.1357	-7.3294	31	5.00%
S. holbrookii	0.01	79	False	0.28	0.18	2.8300	-7.3333	5	1.00%
H. subpunctatus	0.01	87	False	7.75	7.02	9.9358	-7.3343	27	1.00%
P. terribilis	0.05	55	False	4.17	3.68	4.9843	-7.3372	14	5.00%
S. holbrookii	0.05	46	False	1.91	1.5	4.3003	-7.3444	14	5.00%
M. moreirae	0.01	8	False	42.93	41.33	24.9513	-7.3481	69	1.00%
S. holbrookii	0.05	103	False	0.28	0.14	1.4095	-7.3571	3	5.00%
H. subpunctatus	0.01	85	False	7.02	6.32	9.2539	-7.3698	26	1.00%
S. holbrookii	0.01	107	False	0.89	0.77	5.2136	-7.3778	15	1.00%
S. holbrookii	0.01	104	False	10.35	8.34	7.6628	-7.4400	21	1.00%
P. terribilis	0.05	84	False	150.52	125.33	12.3317	-7.4427	30	5.00%
H. subpunctatus	0.01	53	False	5.42	4.89	11.5099	-7.4469	31	1.00%
P. terribilis	0.01	79	False	7.61	6.73	10.8559	-7.4484	28	1.00%
H. subpunctatus	0.01	83	False	6.47	5.79	8.9040	-7.4515	25	1.00%
H. subpunctatus	0.05	69	False	2.12	1.84	8.7164	-7.4517	21	5.00%
P. terribilis	0.01	45	False	7.19	6.07	9.5057	-7.4531	25	1.00%
S. holbrookii	0.01	109	False	0.77	0.66	4.5700	-7.4578	14	1.00%
H. subpunctatus	0.01	24	False	1.34	1.2	11.3096	-7.4587	32	1.00%
H. subpunctatus	0.05	38	False	0.8	0.72	5.5283	-7.4600	16	5.00%
H. subpunctatus	0.01	57	False	8.07	7.46	13.9708	-7.4606	38	1.00%
H. subpunctatus	0.01	7	False	42.83	40.93	32.9710	-7.4637	92	1.00%
P. terribilis	0.05	10	False	189.51	155.15	4.7483	-7.4918	11	5.00%
P. terribilis	0.05	12	False	189.51	155.15	4.7483	-7.4918	11	5.00%
M. moreirae	0.05	25	False	4.23	3.76	4.9326	-7.5107	15	5.00%
P. terribilis	0.01	66	False	389.14	336.88	8.0773	-7.5296	22	1.00%
S. holbrookii	0.05	88	False	4.14	3.13	10.9173	-7.5344	34	5.00%
H. subpunctatus	0.05	89	False	6.63	5.93	11.5932	-7.5355	32	5.00%
M. moreirae	0.05	103	False	3.74	3.31	3.7580	-7.5467	11	5.00%
P. terribilis	0.01	125	False	2.57	2.28	5.6224	-7.5490	15	1.00%
S. holbrookii	0.05	8	False	0.1	0.07	1.8233	-7.5556	5	5.00%
P. terribilis	0.05	32	False	401.16	349.31	5.7871	-7.5593	16	5.00%
H. subpunctatus	0.01	79	False	5.26	4.67	7.8414	-7.5651	22	1.00%
M. moreirae	0.01	103	False	6.38	5.67	9.6150	-7.5672	27	1.00%
M. moreirae	0.01	94	False	3.54	3.13	9.3515	-7.5700	28	1.00%
M. moreirae	0.05	78	False	4.54	4.11	9.4485	-7.5705	28	5.00%
P. terribilis	0.05	57	False	4.4	3.82	6.0768	-7.5751	17	5.00%
P. terribilis	0.01	83	False	7.62	6.95	9.8587	-7.5805	27	1.00%
P. terribilis	0.01	25	False	25.44	23.41	16.6504	-7.5832	46	1.00%
M. moreirae	0.01	84	False	10.05	9.43	13.5529	-7.5907	38	1.00%
H. subpunctatus	0.01	61	False	4.94	4.38	10.7496	-7.6046	28.999	1.00%
P. terribilis	0.01	126	False	342.84	298.73	7.8389	-7.6077	21	1.00%
M. moreirae	0.01	29	False	3.58	2.96	8.4821	-7.6098	25	1.00%
H. subpunctatus	0.01	81	False	5.41	4.76	8.1888	-7.6188	23	1.00%
H. subpunctatus	0.01	71	False	5.67	5	9.1526	-7.6208	25	1.00%
P. terribilis	0.05	28	False	803.22	707.84	8.3949	-7.6217	23	5.00%
H. subpunctatus	0.01	151	False	0.11	0.07	1.3600	-7.6250	2	1.00%
P. terribilis	0.01	104	False	224.34	196.04	8.1968	-7.6358	21	1.00%
M. moreirae	0.01	22	False	14.39	13.47	15.3473	-7.6405	43	1.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
P. terribilis	0.05	34	False	349.31	301.79	5.4250	-7.6434	15	5.00%
S. holbrookii	0.01	15	False	0.92	0.86	10.1985	-7.6483	28.999	1.00%
H. subpunctatus	0.05	62	False	4.47	4.2	7.7365	-7.6524	22	5.00%
P. terribilis	0.01	117	False	2.28	2.01	5.2727	-7.6568	14	1.00%
P. terribilis	0.01	119	False	2.28	2.01	5.2727	-7.6568	14	1.00%
P. terribilis	0.01	121	False	2.28	2.01	5.2727	-7.6568	14	1.00%
P. terribilis	0.01	123	False	2.28	2.01	5.2727	-7.6568	14	1.00%
H. subpunctatus	0.05	67	False	1.95	1.7	7.7021	-7.6594	20	5.00%
P. terribilis	0.05	82	False	124.89	102.07	10.8867	-7.6660	26	5.00%
P. terribilis	0.01	77	False	6.92	6.1	10.1168	-7.6662	26	1.00%
H. subpunctatus	0.05	76	False	1.31	1.19	6.7655	-7.6909	17	5.00%
M. moreirae	0.01	82	False	8.91	8.29	12.8174	-7.6917	36	1.00%
M. moreirae	0.05	58	False	8.34	7.7	7.0003	-7.6927	20	5.00%
M. moreirae	0.05	76	False	3.54	3.13	7.7329	-7.7007	23	5.00%
M. moreirae	0.01	14	False	27.11	25.86	20.9555	-7.7095	57.999	1.00%
M. moreirae	0.05	109	False	3.31	2.92	3.4357	-7.7127	10	5.00%
P. terribilis	0.01	114	False	298.73	259.86	7.4688	-7.7139	20	1.00%
H. subpunctatus	0.05	71	False	2.23	1.95	8.5066	-7.7143	22	5.00%
P. terribilis	0.01	81	False	6.62	5.95	9.4771	-7.7203	25	1.00%
H. subpunctatus	0.05	60	False	3.93	3.66	6.9995	-7.7213	20	5.00%
P. terribilis	0.01	142	False	18.06	10.05	1.4276	-7.7354	2	1.00%
P. terribilis	0.01	146	False	18.06	10.05	1.4276	-7.7354	2	1.00%
P. terribilis	0.01	148	False	18.06	10.05	1.4276	-7.7354	2	1.00%
H. subpunctatus	0.01	54	False	0.42	0.34	6.0450	-7.7597	17	1.00%
M. moreirae	0.01	20	False	13.49	12.6	14.9727	-7.7720	42	1.00%
H. subpunctatus	0.01	56	False	0.43	0.35	6.4150	-7.7730	18	1.00%
M. moreirae	0.01	107	False	4.3	3.65	7.6145	-7.7840	20	1.00%
H. subpunctatus	0.01	58	False	0.44	0.36	6.6850	-7.7850	19	1.00%
M. moreirae	0.01	105	False	5.67	4.97	8.9667	-7.7852	25	1.00%
M. moreirae	0.05	32	False	7.84	7.26	5.6723	-7.7943	16	5.00%
P. terribilis	0.05	21	False	1.66	1.38	4.1448	-7.8025	9	5.00%
S. holbrookii	0.05	72	False	7.97	6.29	10.9691	-7.8044	31	5.00%
P. terribilis	0.01	19	False	20.25	18.49	14.7996	-7.8110	41	1.00%
P. terribilis	0.01	124	False	259.86	224.25	7.1028	-7.8156	19	1.00%
P. terribilis	0.01	75	False	6.1	5.35	9.3947	-7.8310	24	1.00%
M. moreirae	0.05	41	False	1.75	1.47	5.9241	-7.8354	16	5.00%
P. terribilis	0.01	23	False	23.63	21.58	16.2706	-7.8386	45	1.00%
M. moreirae	0.01	27	False	2.96	2.44	7.8008	-7.8549	23	1.00%
P. terribilis	0.01	106	False	224.29	196.02	7.8148	-7.8647	20	1.00%
M. moreirae	0.05	105	False	2.92	2.56	3.1116	-7.8660	9	5.00%
S. holbrookii	0.05	92	False	10.28	8.87	7.5988	-7.8693	20	5.00%
P. terribilis	0.05	26	False	618.77	533.55	7.6655	-7.8753	21	5.00%
M. moreirae	0.01	128	False	0.64	0.46	2.6466	-7.8813	6	1.00%
M. moreirae	0.01	130	False	0.64	0.46	2.6466	-7.8813	6	1.00%
S. holbrookii	0.05	102	False	8.03	6.72	6.2639	-7.8826	16	5.00%
P. terribilis	0.01	21	False	20.88	19.1	15.1687	-7.8850	42	1.00%
P. terribilis	0.01	73	False	5.61	4.87	9.0232	-7.8871	23	1.00%
P. terribilis	0.05	16	False	109.42	80.98	2.8413	-7.8937	5	5.00%
M. moreirae	0.01	18	False	12.67	11.78	14.6189	-7.8941	41	1.00%
H. subpunctatus	0.01	77	False	3.98	3.42	7.1124	-7.8970	20	1.00%
S. holbrookii	0.01	77	False	0.68	0.55	5.2999	-7.9157	15	1.00%
H. subpunctatus	0.01	153	False	0.07	0.03	1.0000	-7.9167	1	1.00%
H. subpunctatus	0.01	155	False	0.07	0.03	1.0000	-7.9167	1	1.00%
P. terribilis	0.01	110	False	224.25	196	6.7348	-7.9223	18	1.00%
P. terribilis	0.01	112	False	224.25	196	6.7348	-7.9223	18	1.00%
P. terribilis	0.01	116	False	224.25	196	6.7348	-7.9223	18	1.00%
P. terribilis	0.01	120	False	224.25	196	6.7348	-7.9223	18	1.00%
P. terribilis	0.01	122	False	224.25	196	6.7348	-7.9223	18	1.00%
P. terribilis	0.01	128	False	224.25	196	6.7348	-7.9223	18	1.00%
P. terribilis	0.01	130	False	224.25	196	6.7348	-7.9223	18	1.00%
P. terribilis	0.01	108	False	224.26	196	7.0348	-7.9262	19	1.00%
H. subpunctatus	0.05	58	False	2.37	2.12	4.8007	-7.9270	14	5.00%
M. moreirae	0.01	80	False	8.29	7.7	12.1379	-7.9274	34	1.00%
P. terribilis	0.01	71	False	5.13	4.41	8.6460	-7.9422	22	1.00%
P. terribilis	0.05	90	False	383.24	340.36	10.9318	-7.9569	28	5.00%
S. holbrookii	0.05	90	False	10.09	8.57	11.7183	-7.9575	38	5.00%
H. subpunctatus	0.05	56	False	2.12	1.87	4.4459	-7.9583	13	5.00%
M. moreirae	0.01	28	False	17.37	16.14	16.2630	-7.9636	45	1.00%
S. holbrookii	0.05	28	False	1.48	1.38	4.7656	-7.9833	14	5.00%
M. moreirae	0.01	30	False	17.62	16.39	16.6174	-7.9860	46	1.00%
S. holbrookii	0.05	30	False	1.38	1.28	4.4736	-7.9956	13	5.00%
S. holbrookii	0.05	32	False	1.38	1.28	4.4736	-7.9956	13	5.00%
S. holbrookii	0.01	114	False	0.01	0	0.2400	-8.0000	1	1.00%
S. holbrookii	0.01	116	False	0.02	0.01	0.5900	-8.0000	2	1.00%
S. holbrookii	0.01	118	False	0.02	0.01	0.5900	-8.0000	2	1.00%
S. holbrookii	0.01	120	False	0.01	0	0.2400	-8.0000	1	1.00%
S. holbrookii	0.01	124	False	0.01	0	0.2400	-8.0000	1	1.00%
S. holbrookii	0.05	6	False	0.04	0.03	1.2400	-8.0000	4	5.00%
P. terribilis	0.01	118	False	196	170.59	6.3687	-8.0065	17	1.00%
P. terribilis	0.01	132	False	196	170.59	6.3687	-8.0065	17	1.00%
P. terribilis	0.01	134	False	196	170.59	6.3687	-8.0065	17	1.00%
P. terribilis	0.05	24	False	533.55	451.78	7.2948	-8.0209	20	5.00%
S. holbrookii	0.01	5	False	29.68	28.8	30.1165	-8.0233	88	1.00%
S. holbrookii	0.05	26	False	1.18	1.08	3.8486	-8.0256	11	5.00%
P. terribilis	0.01	105	False	11.52	10.83	10.6029	-8.0276	28.999	1.00%
P. terribilis	0.01	27	False	20.97	19.02	15.4737	-8.0624	43	1.00%
S. holbrookii	0.05	24	False	0.98	0.88	3.3036	-8.0667	9	5.00%
S. holbrookii	0.05	63	False	2.33	2.14	8.9907	-8.0712	25	5.00%
H. subpunctatus	0.01	74	False	0.14	0.12	3.5400	-8.0714	13	1.00%
M. moreirae	0.01	16	False	12.96	12.11	14.4688	-8.0721	40	1.00%
P. terribilis	0.01	136	False	170.59	147.68	6.0066	-8.0853	16	1.00%
P. terribilis	0.01	138	False	170.59	147.68	6.0066	-8.0853	16	1.00%
M. moreirae	0.01	56	False	5.99	5.57	10.1075	-8.0879	26	1.00%
M. moreirae	0.01	54	False	5.58	5.15	10.0231	-8.1133	26	1.00%
P. terribilis	0.01	17	False	18.64	17.01	14.4409	-8.1448	40	1.00%
P. terribilis	0.01	82	False	956.95	866.54	12.5054	-8.1613	32	1.00%
S. holbrookii	0.05	22	False	0.68	0.58	2.4896	-8.1667	6	5.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
S. holbrookii	0.05	70	False	4.64	3.56	9.2594	-8.1670	26	5.00%
P. terribilis	0.01	103	False	10.97	10.2	10.9617	-8.1883	30	1.00%
M. moreirae	0.05	107	False	2.2	1.84	2.4591	-8.1936	7	5.00%
P. terribilis	0.01	101	False	11.06	10.28	11.3251	-8.2001	31	1.00%
P. terribilis	0.01	84	False	1245.04	1145.86	12.8598	-8.2051	33	1.00%
M. moreirae	0.01	66	False	6.48	6.09	12.1615	-8.2110	31	1.00%
P. terribilis	0.01	69	False	3.77	3.2	7.2189	-8.2246	18	1.00%
M. moreirae	0.01	117	False	2.22	1.92	3.6159	-8.2385	10	1.00%
M. moreirae	0.01	127	False	2.22	1.92	3.6159	-8.2385	10	1.00%
M. moreirae	0.01	129	False	2.22	1.92	3.6159	-8.2385	10	1.00%
M. moreirae	0.01	133	False	2.22	1.92	3.6159	-8.2385	10	1.00%
M. moreirae	0.01	135	False	2.22	1.92	3.6159	-8.2385	10	1.00%
P. terribilis	0.01	99	False	10.23	9.38	11.6630	-8.2467	31	1.00%
P. terribilis	0.05	58	False	134.75	109.12	10.0431	-8.2829	27	5.00%
M. moreirae	0.01	10	False	28.68	27.4	22.1927	-8.2866	62	1.00%
M. moreirae	0.01	12	False	29.22	27.93	22.8450	-8.2924	64	1.00%
M. moreirae	0.01	26	False	15.16	14.16	15.8574	-8.3028	44	1.00%
P. terribilis	0.01	80	False	956.82	866.43	12.1339	-8.3061	31	1.00%
P. terribilis	0.01	78	False	779.53	698.27	11.7754	-8.3141	30	1.00%
S. holbrookii	0.05	19	False	0.54	0.48	2.9350	-8.3333	9	5.00%
S. holbrookii	0.05	21	False	0.72	0.66	3.8883	-8.3333	12	5.00%
M. moreirae	0.05	82	False	3.99	3.73	9.5191	-8.3370	25	5.00%
M. moreirae	0.01	123	False	1.92	1.66	3.2865	-8.3424	9	1.00%
M. moreirae	0.01	131	False	1.92	1.66	3.2865	-8.3424	9	1.00%
M. moreirae	0.01	137	False	1.92	1.66	3.2865	-8.3424	9	1.00%
M. moreirae	0.01	139	False	1.92	1.66	3.2865	-8.3424	9	1.00%
M. moreirae	0.01	141	False	1.92	1.66	3.2865	-8.3424	9	1.00%
M. moreirae	0.01	119	False	1.66	1.4	2.9557	-8.3462	8	1.00%
M. moreirae	0.01	121	False	1.66	1.4	2.9557	-8.3462	8	1.00%
M. moreirae	0.01	125	False	1.66	1.4	2.9557	-8.3462	8	1.00%
M. moreirae	0.01	109	False	4.53	3.98	6.9988	-8.3498	19	1.00%
M. moreirae	0.05	80	False	3.23	2.97	8.0487	-8.3504	21	5.00%
P. terribilis	0.01	97	False	9.59	8.76	11.2958	-8.3510	30	1.00%
P. terribilis	0.01	150	False	10.05	3.85	1.0657	-8.3709	1	1.00%
P. terribilis	0.01	152	False	10.05	3.85	1.0657	-8.3709	1	1.00%
P. terribilis	0.01	154	False	10.05	3.85	1.0657	-8.3709	1	1.00%
P. terribilis	0.01	156	False	10.05	3.85	1.0657	-8.3709	1	1.00%
M. moreirae	0.05	28	False	5.69	4.99	5.1780	-8.3722	15	5.00%
S. holbrookii	0.05	40	False	0.56	0.45	6.6448	-8.3793	21	5.00%
P. terribilis	0.05	64	False	345.46	303.94	9.8217	-8.3807	24	5.00%
H. subpunctatus	0.05	42	False	0.28	0.24	4.9200	-8.3810	14	5.00%
H. subpunctatus	0.05	44	False	0.28	0.24	4.9200	-8.3810	14	5.00%
M. moreirae	0.05	30	False	4.46	3.93	3.4708	-8.3948	10	5.00%
M. moreirae	0.01	64	False	6.83	6.4	11.3793	-8.4013	28.999	1.00%
M. moreirae	0.05	56	False	4.55	4.04	4.8192	-8.4108	14	5.00%
P. terribilis	0.01	15	False	16.67	15.09	13.7120	-8.4180	38	1.00%
P. terribilis	0.05	106	False	72.55	50.99	3.7212	-8.4315	8	5.00%
M. moreirae	0.01	62	False	6.07	5.73	11.0242	-8.4447	28	1.00%
M. moreirae	0.01	120	False	0.46	0.28	2.3288	-8.4595	5	1.00%
M. moreirae	0.01	132	False	0.46	0.28	2.3288	-8.4595	5	1.00%
S. holbrookii	0.01	13	False	2.15	2.06	14.2335	-8.4675	42	1.00%
P. terribilis	0.01	76	False	779.4	698.16	11.4038	-8.4731	28.999	1.00%
S. holbrookii	0.01	106	False	1.67	1.24	4.6006	-8.4754	11	1.00%
M. moreirae	0.05	66	False	2.68	2.43	6.3222	-8.4772	15	5.00%
M. moreirae	0.01	60	False	5.75	5.41	10.6523	-8.4848	27	1.00%
S. holbrookii	0.01	108	False	1.54	1.11	4.2898	-8.5061	10	1.00%
M. moreirae	0.01	113	False	2.24	1.93	4.2259	-8.5095	12	1.00%
M. moreirae	0.01	24	False	13.92	12.93	15.4255	-8.5116	43	1.00%
S. holbrookii	0.01	11	False	1.97	1.88	13.4713	-8.5131	40	1.00%
M. moreirae	0.01	52	False	5.56	5.12	10.6960	-8.5136	28	1.00%
M. moreirae	0.01	58	False	5.43	5.1	10.2795	-8.5278	26	1.00%
P. terribilis	0.05	22	False	302.05	244.9	6.1869	-8.5575	17	5.00%
P. terribilis	0.05	66	False	218.84	180.75	8.1228	-8.5619	22	5.00%
S. holbrookii	0.01	9	False	1.79	1.7	12.7369	-8.5635	38	1.00%
P. terribilis	0.05	62	False	263.06	226.23	8.3393	-8.5839	20	5.00%
M. moreirae	0.01	115	False	1.94	1.67	3.8965	-8.6186	11	1.00%
M. moreirae	0.01	40	False	9.65	8.87	13.4750	-8.6222	39	1.00%
M. moreirae	0.01	111	False	3.45	3.06	6.3445	-8.6311	17	1.00%
P. terribilis	0.01	72	False	622.58	551.61	10.6796	-8.6485	27	1.00%
P. terribilis	0.01	74	False	866.33	779.1	11.0142	-8.6863	28	1.00%
S. holbrookii	0.01	110	False	0.98	0.73	3.4341	-8.6925	9	1.00%
S. holbrookii	0.01	7	False	4.54	4.35	22.1421	-8.7155	66	1.00%
P. terribilis	0.05	108	False	33.37	17.08	2.9909	-8.7491	7	5.00%
P. terribilis	0.05	19	False	0.98	0.81	3.0798	-8.7551	6	5.00%
P. terribilis	0.05	101	False	0.79	0.55	2.8687	-8.8007	6	5.00%
H. subpunctatus	0.05	107	False	0.08	0.04	1.0100	-8.8056	1	5.00%
S. holbrookii	0.05	104	False	3.34	2.34	4.4955	-8.8462	10	5.00%
P. terribilis	0.01	13	False	15.81	14.26	14.1842	-8.8897	39	1.00%
P. terribilis	0.05	60	False	161.22	132.63	7.6439	-8.9077	18	5.00%
S. holbrookii	0.01	112	False	0.15	0	0.5621	-8.9643	1	1.00%
P. terribilis	0.01	70	False	483.85	418.01	8.8260	-8.9815	24	1.00%
H. subpunctatus	0.01	76	False	0.01	0	0.3700	-9.0000	0	1.00%
H. subpunctatus	0.01	92	False	0.01	0	0.3700	-9.0000	0	1.00%
H. subpunctatus	0.01	100	False	0.01	0	0.3700	-9.0000	0	1.00%
P. terribilis	0.05	112	False	0.06	0.04	1.1650	-9.0000	3	5.00%
P. terribilis	0.05	114	False	0.04	0.02	0.7750	-9.0000	2	5.00%
P. terribilis	0.05	116	False	0.02	0	0.3800	-9.0000	1	5.00%
P. terribilis	0.05	105	False	0.14	0.05	0.6540	-9.0444	1	5.00%
P. terribilis	0.01	29	False	21.42	19.46	15.8713	-9.0725	44	1.00%
H. subpunctatus	0.05	93	False	4.37	3.83	9.7699	-9.1258	25	5.00%
P. terribilis	0.01	68	False	483.76	417.93	8.4460	-9.1932	23	1.00%
P. terribilis	0.01	11	False	14.83	13.26	13.8165	-9.2192	38	1.00%
M. moreirae	0.01	122	False	0.28	0.22	2.0027	-9.2306	4	1.00%
M. moreirae	0.01	126	False	0.28	0.22	2.0027	-9.2306	4	1.00%
M. moreirae	0.01	134	False	0.28	0.22	2.0027	-9.2306	4	1.00%
M. moreirae	0.01	136	False	0.28	0.22	2.0027	-9.2306	4	1.00%
M. moreirae	0.01	124	False	0.22	0.16	1.6693	-9.3433	3	1.00%

Continuation of **Table S3.1** 1

Species	Indel Prob.	Iteration	Contiguity	Cov.	Connectivity	Qual.	Score	Similarity	Prob. Mod.
M. moreirae	0.01	138	False	0.22	0.16	1.6693	-9.3433	3	1.00%
M. moreirae	0.01	140	False	0.22	0.16	1.6693	-9.3433	3	1.00%
M. moreirae	0.01	142	False	0.22	0.16	1.6693	-9.3433	3	1.00%
P. terribilis	0.05	110	False	1.25	0.08	1.9000	-9.3436	4	5.00%
S. holbrookii	0.05	67	False	2.81	2.64	12.4298	-9.3529	34	5.00%
P. terribilis	0.01	162	False	3.85	0.94	0.7021	-9.4080	0	1.00%
P. terribilis	0.05	6	False	3.85	0.94	0.7021	-9.4080	0	5.00%
H. subpunctatus	0.05	91	False	3.62	3.19	8.3152	-9.4345	21	5.00%
P. terribilis	0.01	9	False	11.97	10.68	13.0661	-9.4905	36	1.00%
M. moreirae	0.01	144	False	0.16	0.11	1.3310	-9.5125	3	1.00%
P. terribilis	0.01	7	False	51.07	48.66	36.0940	-9.5163	98	1.00%
P. terribilis	0.05	103	False	0.26	0.11	1.3790	-9.5222	2	5.00%
S. holbrookii	0.05	38	False	0.25	0.17	5.7938	-9.5278	18	5.00%
M. moreirae	0.05	18	False	0.08	0.05	1.0467	-9.5556	2	5.00%
H. subpunctatus	0.05	55	False	1.07	0.93	6.1290	-9.5848	17	5.00%
S. holbrookii	0.05	65	False	2.56	2.39	10.7714	-9.5967	30	5.00%
S. holbrookii	0.05	27	False	1.28	1.1	4.0898	-9.6296	12	5.00%
S. holbrookii	0.05	29	False	1.28	1.1	4.0898	-9.6296	12	5.00%
P. terribilis	0.01	127	False	1	0.81	3.4065	-9.6987	6	1.00%
H. subpunctatus	0.01	59	False	8.28	7.65	13.9908	-9.7033	38	1.00%
P. terribilis	0.05	50	False	243.61	218.51	11.9788	-9.7798	31	5.00%
H. subpunctatus	0.05	53	False	0.81	0.69	5.4282	-9.9961	15	5.00%
M. moreirae	0.01	146	False	0.02	0	0.3600	-10.0000	1	1.00%
M. moreirae	0.05	16	False	0.02	0	0.3600	-10.0000	1	5.00%
S. holbrookii	0.05	34	False	0.16	0.15	5.2500	-10.0000	16	5.00%
S. holbrookii	0.05	36	False	0.15	0.14	4.9100	-10.0000	15	5.00%
S. holbrookii	0.05	106	False	0.11	0.05	2.3080	-10.0000	5	5.00%
S. holbrookii	0.05	108	False	0.05	0.04	1.6400	-10.0000	3	5.00%
S. holbrookii	0.05	110	False	0.05	0.04	1.6400	-10.0000	3	5.00%
S. holbrookii	0.05	112	False	0.03	0.02	0.9400	-10.0000	1	5.00%
P. terribilis	0.05	17	False	0.5	0.38	2.0209	-10.0444	3	5.00%
P. terribilis	0.01	158	False	0.94	0	0.3502	-10.1596	0	1.00%
P. terribilis	0.01	160	False	0.94	0	0.3502	-10.1596	0	1.00%
P. terribilis	0.01	164	False	0.94	0	0.3502	-10.1596	0	1.00%
P. terribilis	0.01	129	False	0.51	0.39	2.3306	-10.6571	3	1.00%
P. terribilis	0.01	131	False	0.51	0.39	2.3306	-10.6571	3	1.00%
P. terribilis	0.01	133	False	0.51	0.39	2.3306	-10.6571	3	1.00%
P. terribilis	0.01	135	False	0.39	0.27	2.0681	-10.7667	3	1.00%
P. terribilis	0.01	137	False	0.39	0.27	2.0681	-10.7667	3	1.00%
P. terribilis	0.01	139	False	0.39	0.27	2.0681	-10.7667	3	1.00%
P. terribilis	0.01	145	False	0.39	0.27	2.0681	-10.7667	3	1.00%
P. terribilis	0.01	151	False	0.39	0.27	2.0681	-10.7667	3	1.00%
P. terribilis	0.01	153	False	0.18	0.1	1.0302	-10.8333	1	1.00%
P. terribilis	0.01	141	False	0.27	0.17	1.7164	-10.9200	2	1.00%
P. terribilis	0.01	143	False	0.27	0.17	1.7164	-10.9200	2	1.00%
P. terribilis	0.01	147	False	0.27	0.17	1.7164	-10.9200	2	1.00%
P. terribilis	0.01	149	False	0.27	0.17	1.7164	-10.9200	2	1.00%
M. moreirae	0.01	143	False	0.01	0	0.2900	-11.0000	0	1.00%
M. moreirae	0.01	145	False	0.01	0	0.2900	-11.0000	0	1.00%
P. terribilis	0.01	155	False	0.1	0.05	0.6840	-11.0000	0	1.00%
P. terribilis	0.01	157	False	0.05	0	0.3280	-11.0000	0	1.00%
P. terribilis	0.01	161	False	0.05	0	0.3280	-11.0000	0	1.00%
P. terribilis	0.05	15	False	0.1	0.05	0.6840	-11.0000	0	5.00%

End of **Table S3.1**

Appendix S4.1

Specimen and DNA data for the assembly of the draft genome of *Scaphiopus holbrookii*

Specimen

- Code: 5-DJ04
- Species: *Scaphiopus holbrookii* (Harlan 1835) (Anura: Scaphiopodidae)
- Common names: Eastern spadefoot, solitary frog, Holbrook's spadefoot, hermit spadefoot, among others
- Specimen: female, adult
- Extraction method: DNeasy Blood & Tissue Kit - QIAGEN

The adult female was acquired via pet trade, fixated using liquid nitrogen and stored in 100% alcohol at -20 °C. A total of 3 genomic DNA extractions from liver and muscle were made.

Expected genome size

The frog *Scaphiopus holbrookii* has 26 chromosomes and a C-value varying from 1.34 to 1.41 depending on the author (Olmo 1973; Goin et al. 1968; Sexsmith 1968).

Sample DNA quality

- DNA Concentration: 114.194 ng/ µL
- Purity (A260/A): 1.89
- Volume: 86 µL
- Total Amount of DNA: 9.821 µg

Raw data report

Experiment overview

Steps of the experiment include sample preparation, library quality control, clustering, sequencing, and data analysis. Base call files (in BCL format) were converted into FASTQ format in the last step.

Data Analysis

The Illumina HiSeq sequencing machine generates raw images utilizing the HiSeq Control Software (HCS v2.2.38) for system control and base calling through an integrated primary analysis software caller Real Time Analysis (RAT v1.18.61.0). The BCL binaries were converted into FASTQ utilizing Illumina package bcl2fastq v1.8.4.

FASTQ format

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores, usually using PHRED scores, which are logarithmically related to the probability of an error.

Mate paired-end library (3kb)

- Preparation kit: Nextera Mate Pair
- Sequencing machine: Illumina HiSeq 2000
- Number of lanes: 2
- Total bases: 78,220,059,838
- Read Count[1]: 774,456,038
- GC content: 40.16%
- AT content: 59.84%
- Q20[2]: 95.37%
- Q30[3]: 87.57%

Paired-end (5-DJ04_shotgun)

- Preparation kit: TrueSeq nano
- Sequencing machine: Illumina HiSeq 2000
- Number of lanes: 2
- Total bases: 90,213,948,208
- Read Count: 893,207,408
- GC content: 39.76%
- AT content: 60.25%
- Q20: 96.79%
- Q30: 90.16%

FastQC report

We used FastQC v0.11.2 (Andrew 2010) to perform some basic quality checks in different steps of the quality control procedure. FastQC report (**Table S4.1.2**) for raw sequence reads show high adapter content and sequence supplication levels, which are the two most important problems These problems were addressed during the *in silico* quality control procedure (see bellow). FastQC also report imbalanced *k-mers* but this is expected from libraries derived from random priming.

[1] The expected number of paired end reads per lane for the Illumina HiSeq 2000 is 300-400M.

[2] Percentage of bases with Phred quality score above 20. the minimum expected is 90%.

[3] Percentage of bases with Phred quality score above 20. the minimum expected is 85%.

Table S4.1.1 - Raw data FastQC report

	Mate pairs 1	Mate pairs 2	Shotgun 1	Shotgun 2
Adapter content	Fail	Fail	Pass	Pass
Basic statisctics	Pass	Pass	Pass	Pass
Kmer content	Fail	Fail	Fail	Fail
Overrepresented sequences	Pass	Pass	Pass	Pass
Per base N content	Pass	Pass	Pass	Pass
Per base sequence content	Warn	Warn	Warn	Warn
Per base sequence quality	Pass	Pass	Pass	Pass
Per sequence GC content	Pass	Pass	Pass	Pass
Per sequence quality scores	Pass	Pass	Pass	Pass
Per tile sequence quality	Warn	Pass	Pass	Pass
Sequence duplication levels	Fail	Fail	Warn	Warn
Sequence length distribution	Pass	Pass	Pass	Pass

Quality control

Adapter removal for shotgun reads

The presence of poor quality or technical sequences such as adapters in NGS data can easily result in sub-optimal downstream analyses. Therefore, althought FastQC reports do not indicate any adaptor contamination for shotgun read sequences, adaptor removal was performed using Trimmomatic v0.32 (Bolger et al. 2014) and the adapter file "TruSeq3-PE-2.fa" provided with the program.

Example command line used to execute Trimmomatic:

```
$ `which java` -jar trimmomatic-0.35.jar PE -threads 64 -phred33 \
  -trimlog trimmomatic_shotgun.log 5-DJ04_shotgun_1.fastq.gz \
  5-DJ04_shotgun_2.fastq.gz shotgun_paired_1.fastq.gz\
  shotgun_unpaired_1.fastq.gz shotgun_paired_2.fastq.gz \
  shotgun_unpaired_2.fastq.gz \
  ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 \
  LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Adapter trimming for mate pair reads

Mate pair library sequencing enables the generation of libraries with insert sizes in the range of several kilobases (Kb) that can provide detailed information about genomic regions that are separated by large distances.

The composition of library templates from a mate pair experiment will include reads with both forward-reverse and reverse-forward orientation, as well as single and reads that lacks the adapter region and therefore cannot be immediately categorized. Downstream adaptor trimming is therefore required to separate reads according where the adapter lies in the read.

NxTrim v0.3.0-alpha (O'Connell et al. 2014) converts raw NMP reads into four "virtual libraries":

- **MP:** a set of known mate pairs having an outward-facing relative orientation and an effective genomic distance (EGD) whose distribution mirrors the size distribution of the circularized DNA
- **Unknown:** A set of read pairs for which the adapter could not be found within either read
- **PE:** a set of paired-end reads, having an inward-facing relative orientation and an EGD whose distribution mirrors the size distribution of the sequenced templates
- **SE:** a set of single reads

Example command line used to execute NxTrim:

```
$ nxtrim --separate -1 5-DJ04_3kb_1.fastq.gz -2 \
  5-DJ04_3kb_2.fastq.gz -O nxtrimmed
```

Initial quality assessment

Result sequences from Trimmomatic and NxTrim where processed through the High-Throughput Quality Control (HTQC) toolkit v0.90.7 (Yang et al. 2013), which consists of six programs (ht-stat, ht-stat-draw.pl, ht-tile-filter, ht-trim, ht-qual-filter, and ht-length-filter) for generation of graphic reports and reads quality assessment, trimming and filtration.

After adaptor removal, quality reports were generated with ht-stat. Results were used as input to the homemade Python program selectTiles.py[4] for automated selection of tiles that might be removed. All tiles pass the default parameters of selectedTiles.py.

Example command line used to execute ht-stat with paired-end reads:

```
$ ht-stat -P -t 32 -z -i ${PE1} ${PE2} \
  -o htqc_shot_pe > htqc_shot_pe.log 2>&1 &
```

Example command line used to execute ht_stat with single-end reads:

```
$ ht-stat -S -t 32 -z -i ${SE1} \
  -o htqc_shot_se1 > htqc_shot_se1.log 2>&1 &
```

Example command line used to execute selectTiles.py:

```
$ python selectTilesHTQC.py -d htqc_shot_pe > tiles_pe.txt &
```

Trimming by quality

An extensive evaluation of read trimming effects on illumine NGS data analysis shows that trimming is shown to increase the quality and reliability of the analysis, with concurrent gains in terms of execution time and computational resources needed (Del Fabbro et al. 2013).

The program ht-trim was used to cut low quality bases at the beginning or the end of the reads until the quality score reached the default threshold.

[4] Available at http://www.ib.usp.br/grant/anfibios/researchSoftware.html and https://gitlab.com/MachadoDJ/selectTiles.

Example command line used to execute ht-trim:

```
$ ht-trim -z -i shotgun_paired_1.fastq.gz \
  -o httrim_pe1.fastq.gz > httrim_pe1.log
```

Filtering by length

Sequences were filtered by length using ht-filter and default parameters.

Example command line for paired-end reads:

```
$ ht-filter -P --filter length -z \
  -i httrim_pe1.fastq.gz httrim_pe2.fastq.gz \
  -o htfilter_pe > htfilter_pe.log
```

Example command line for single-end reads:

```
$ ht-filter -S --filter length -z -i httrim_se1.fastq.gz \
  -o htfilter_se1 > htfilter_se1.log
```

De novo *duplicates removal for paired short reads*

The presence of duplicates introduced by PCR amplification in paired short reads from next-generation sequencing platforms might have a serious impact on research applications, such as scaffolding in whole genome sequencing and discovering large-scale genome variations. To remove these duplications, I used FastUniq v1.1 (Xu et al. 2012), which identifies duplicates by comparing sequences between read pairs and does not require complete genome sequences as prerequisites.

Example command line:

```
$ fastuniq -i files.txt -t q -o fastuniq_pe1.fastq \
  -p fastuniq_pe2.fastq -c 0
```

QC report

After trimmed sequences were subjected to quality control procedures with HTQC, we observed an improvement in adapter content but not in sequence duplication levels. Sequence duplication levels improved with FastUniq (see **Table S4.1.3**).

Table S4.1.2 - Summary of the FastQC report after quality control.

	After HTQC		After FastqUniq	
	Nextera Mate Pair	**Paired-end**	**Nextera Mate Pair**	**TruSeq Nano**
Per base sequence content	Pass or Warn	Pass or Warn	Pass	Warn
Per sequence GC content	Pass	Pass	Pass	Pass
Adapter Content	Pass	Pass	Pass	Pass
Basic Statistics	Pass	Pass	Pass	Pass
Per tile sequence quality	Pass	Pass	Pass	Pass
Per base sequence quality	Pass	Pass	Pass	Pass
Sequence Duplication Levels	Pass or Fail	Pass or Warn	Pass or Warn	Pass
Sequence Length Distribution	Warn	Warn	Warn	Warn
Per base N content	Pass	Pass	Pass	Pass
Kmer Content	Warn or Fail	Pass or Fail	Warn or Fail	Fail
Overrepresented sequences	Pass	Pass	Pass	Pass
Per sequence quality scores	Pass	Pass	Pass	Pass

After quality control, FastQC keeps reporting *k-mers* with positionally biased enrichment. However, libraries which derive from random priming will nearly always show *k-mer* bias at the start of the library due to an incomplete sampling of the possible random primers. Also, it is expected that read lengths will vary after the quality control procedures specified above and warnings related to sequence length distribution can therefore be ignored.

Only 92.9% of the reads were maintained after quality control with HTQC, including single end reads. The total number of reads dropped to 36.57% of the number of raw sequence reads after analyses in FastUniq (single end reads not included). Approximately 58% of the reads analyzed with FastUniq were discarded (see **Table** S4.1.**4**).

Table S4.1.3 - Summary statistics of sequence read files after quality control with HTQC and FastUniq.

Program	HTQC						FastqUniq			
Library	Nextera mate pairs				TruSeq Nano		Nextera mate pairs			TruSeq Nano
Type	Mate pairs	Paired-end	Unknown	Single end	Paired-end	Single end	Paired-end	Mate pairs	Unkown	Paired-end
Total sequences	134,293,404	72,812,844	388,446,506	81,247,500	854,820,768	17,643,177	20,852,234	44,962,662	172,436,118	371,653,724
GC	40	40	39 or 40	39 or 41	39	39-41	41	41	41	39
Min. sequence length	50	50	50	50	50	50	50	50	50	50
Max. sequence length	101	101	101	89 or 101	101	101	101	101	101	101

Manual editing

Unfortunately, the sequence headers have to be manually edited to run the *de novo* sequence assembly programs. This was done using the program sed to replace " 1:N:0:"(including leading space) by "/1" and " 2:N:0:"(including leading space) by "/2".

Example command line for sed:

```
$ sed –i '' 's/ 1:N:0:/\/1/g' fastuniq_pe1.fastq
```

Expected coverage

If we calculate the expected coverage as the number of base pairs available divided by genome length, reads after HTQC could provide a coverage of up to 110x and reads after FastqUniq could provide a coverage of 43x, approximately.

Appendix S4.2

Specimen and DNA data for the assembly of the draft genome of *Phyllobates terribilis*

Specimen

- Code: 1b2 (no collection code assigned to tissue, hind leg, 94.6 µg of DNA, A260/ A280 ratio is 1.779) and 1a1 (no collection code assigned to tissue, hind leg, 43.7 µg of DNA, A260/ A280 ratio is 1.865)
- Species: *Phyllobates terribilis* Myers, Daly, and Malkin, 1978 (Anura: Dendrobatidae)
- Common name: Golden Poison Frog
- Sex and stage of life: undefined sex, immature
- Extraction method: DNeasy Blood & Tissue Kit - QIAGEN

Expected genome size

According to the Animal Genome Size Database (http://www.genomesize.com), the genome size of anurans of the family Dendrobatidae shows a great variance, with C-Values ranging from 2.98 (*Mannophryne trinitatis*) to 8.95 (*Dendrobates tinctorius*).

Sample DNA quality

Sample quality passed all steps of quality control at both Macrogen and DHM-RI before proceeding to library preparation. Neither facility reported any problems during library preparation and sequencing.

Raw data report

Macrogen–USA

The DNA sample was used to prepare four mate-paired libraries with different insert sizes (gel-free, 3kb, 5kb, and 8kb). Sequencing was performed using the Illumina HiSeq2500 system and other related instrumentation and software with 150bp paired-end sequencing method. The four libraries for each specimen were pooled and sequenced across 6 lanes of a common flow cell. This resulted in 263,828,699, 240,701,531, 247,545,514, and 252,654,530 read pairs for the four different insert sizes (gel-free, 3kb, 5kb, and 8kb, respectively), with GC content varying between 44 and 46% (with an average of 45.25%). These reads passed most quality filters of FastQC except per base "N" content (50% pass and 50% issued warnings), sequence duplication levels (25% issued warnings and 75% failed), overexpressed sequences (25% pass, 50% passed with warnings, and and 25% failed), adapter content (100% failed), and kmer content (87.5% failed). This is to be expected in from Illumina sequence reads derived from random priming and that include Nextera adapters. Also, overrepresented sequences and duplications may be due to the fact that fragmentation is not completely random and there are a large numbers or natural occurring repetitions in the frog genome.

DHM-RI

The project of "Whole Genomic Sequencing of Samples from Frogs" (UNCC-166) with DHM-RI had the objective to generate sequencing libraries for sequencing of DNA derived from two species of frogs. Libraries were sequenced using the Illumina HiSeq2500 platform. The quality of DNA samples was verified using the Quant-It Picogreen dsDNA Assay Kit (ThermoFisher P11496). DNA samples were used to generate 4 mate pair genomic DNA libraries. The first library was constructed using the Nextera Mate Pair Library Prep Kit and the gel-free protocol with no size selection. The same DNA sample will be used to generate other 3 size-selected libraries using the Nextera Mate Pair Library Prep Kit with gel size selection. The fragment sizes for the 3 mate pair libraries were 3kb, 5kb and 10kb. A 125bp paired end sequencing run was performed on the Illumina HiSeq2500 instrument using a total of 3 lanes of a regular flow cell. This generated 210,072,773, 253,169,616, 218,733,982, and 173,314,326 sequence read pairs for gel-free, 3kb, 5kb, and 10kb libraries, respectively.

The sequence read files had a GC content reaching from 43 to 46% (average equal to 44.8%) and a total of 532,681,735 read pairs. Read files passed most quality control checks.

A total of 41.6% of the files passed the "per tile sequence quality" filter with a warning, but I could only observe mildly effects on a very small number of tiles for a small number of cycles, which does not draw significant suspicion to the quality of sequence reads in these files. 41.6% of the files failed to pass the "per base sequence content" filter, but low quality bases will be removed during the next steps of quality control (see bellow). 37.5% of the files resulted in warnings for "sequence duplication level" and 41.6% failed the same filter. Duplications might be caused by PCR error or they might be real duplications in the frog genome, and will be dealt with accordingly during quality control. 77% of the files passes the "overrepresented sequences" filter with warnings and 100% of the files failed the "adapter content" and "kmer content" filters. This results are expected in from Illumina sequence reads because they are derived from derive from random priming and include Nextera adapters. Also, overrepresented sequences may be due to the fact that fragmentation is not completely random.

Quality control

Mate pair library sequencing enables the generation of libraries with insert sizes in the range of several kilobases (Kb) that can provide detailed information about genomic regions that are separated by large distances. But they require careful quality control involving adapter trimming, base quality trimming, and duplicate removal, as explained in the following paragraphs.

The composition of library templates from a mate pair experiment will include reads with both forward-reverse and reverse-forward orientation, as well as single and reads that lacks the adapter region and therefore cannot be immediately categorized. Downstream adaptor trimming is therefore required to separate reads according where the adapter lies in the read.

Trimming according to base quality increases the quality and reliability of the analysis, with concurrent gains in terms of execution time and computational resources needed (Del Fabbro et al. 2013).

Finally, the presence of duplicates introduced by PCR amplification in paired short reads from next-generation sequencing platforms might have a serious impact on research applications, such as scaffolding in whole genome sequencing and discovering large-scale genome variations.

Quality control using NxTrim software

I used two strategies for trimming sequences and removing duplications. The first strategy uses NxTrim v0.3.0-alpha (O'Connell et al. 2014) to separate reads into four different categories according to the adapter position (mate pairs, unknow – which are mostly mate pairs, paired-end, and single end sequence reads). The output files from NxTrim are them processed with the High-Throughput Quality Control (HTQC) toolkit v0.90.7 (Yang et al. 2013) to trim bases with poor quality and remove reads that are too short. Finally, I used FastUniq v1.1 (Xu et al. 2012) to remove duplications.

Macrogen

After NxTrim, HTQC, and FastUniq, mate and paired-end reads passed most of the filters of FastQC (Andrew 2010), with the following exceptions: 16.7% of the files passed the "per base sequence content" filter with a warning; 100% of the files passed the "sequence length distribution" filter with a warning (which is expected due to sequence trimming); 29.2% of the files passed the "sequence duplication filter" with a warning (which is expected since a small percentage of PCR duplications is kept, and some duplications may be due to real repetitions in the frog genome); and 100% of the files failed to pass the filter of "kmer content". This shows major improvements in relation to the quality of raw sequence files and is a good result for sequences from Illumina platforms.

After quality control, mate paired reads constitute 22.1% of the original reads, 18.2% were categorized as "unknown" (which are mostly mate paired reads with some contamination by paired end reads), and 21.9% of the raw reads we categorized as paired end reads (see details in **Table S4.2.1**).

Table S4.2.1 - Quantity of reads in each category (mate pair, paired end, and unknown) after quality control pipeline with NxTrim, HTQC, and FastUniq.

Library or insert size	Type of read	Read pairs	Percentage of the respective raw
Gel free	mate pair	79,791,955	30.2%
Gel free	paired-end	80,938,948	30.7%
Gel free	unknown (mostly mp)	50,796,769	19.3%
3kb	mate pair	34,357,461	14.3%
3kb	paired-end	31,856,945	13.2%
3kb	unknown (mostly mp)	36,408,897	15.1%
5kb	mate pair	52,972,036	21.4%
5kb	paired-end	51,632,883	20.9%
5kb	unknown (mostly mp)	46,932,165	19.0%
8kb	mate pair	55,110,212	21.8%
8kb	paired-end	55,568,071	22.0%
8kb	unknown (mostly mp)	48,421,754	19.2%

Single end reads passed most of the 12 FastQC filters, except sequence length distribution (warning due to different sequence lengths, which are a result of trimming process), sequence duplication levels (a warning was issued probably due to a combination of factors, which might include PCR duplications, bias during fragmentation and primer attachment, and/ or real duplications on the frog genome), and kmer content (single end reads failed this criteria, but this is expected in from Illumina sequence reads derived from random priming). A total of 138,788,265 single end reads were obtained, representing 6.91% of the original sequence reads from both left and right pairs. The GC content of single end reads is 46% and sequence length varies from 50 to 151bp.

DHM-RI

Most files passed FastQC filters, with only a few exceptions: 18,75% of the files issued warnings for the “per tile quality filter” filter; 52.78% of the files issued warnings for the “per base sequence content” filter, and 13.89% failed; 37.5% of the files issued warnings of the “per sequence GC content” filter, and 9.03% failed it; 100% of the files issued warnings for the “sequence length distribution” filter; 8.33% of the files issued warnings for the “sequence duplication levels” filter; 0.69% of the files issued warnings for the “overrepresented sequences” filter; and 22.22% of the files issued warnings for the “per base sequence content” filter, and 77.78% failed it.

Filtered reads had an average 46.3% GC contend (min. 43 and max 49%) and read length varying from 50 to 125bp. The quality control pipeline with NxTrim resulted in 127,258,537 mate paired-end, 99,112,845 unknown (mostly mate paired-end), and 123,240,941 paired-end sequence reads (see details in **Table S4.2.2**).

Table S4.2.2 - Quantity of reads in each category (mate pair, paired end, and unknown) after quality control pipeline with NxTrim, HTQC, and FastUniq.

Library/ insert size	Type	Read pairs	Percentage of the raw reads
Gel free	mate pair	58163788	0.276874471
Gel free	paired-end	57369659	0.273094215
Gel free	unknown (mostly imp)	50220524	0.239062508
3kb	mate pair	32280976	0.127507307
3kb	paired-end	30824045	0.121752545
3kb	unknown (mostly imp)	21597513	0.085308472
5kb	mate pair	25585719	0.116971852
5kb	paired-end	24061647	0.110004156
5kb	unknown (mostly imp)	18789523	0.085901252
10kb	mate pair	11228054	0.064784339
10kb	paired-end	10985590	0.063385355
10kb	unknown (mostly imp)	8505285	0.049074333

Quality control using NextClip software

NextClip (Leggett et al. 2014) is a tool for analyzing reads from LMP libraries, generating a comprehensive quality report and extracting good quality trimmed and de-duplicated reads. Differently from NxTrim, which retains the 3'-wards portion of the read to improve coverage and *de novo* assembly quality (by means of re-interpreting the 3' side of the adapter, together with the other half of the read pair, as a standard paired-end read), NextClip work by trimming the adapter and everything to the 3' side of it, retaining only the portion of the read that lies to the 5' side of the adapter. NextClip outputs reads in up to 5 categories (mate paired-end sequence reads are stored in categories A, B, and C while reads stored in category D should be treated as unreliable since there is no way to tell if they are true mate paired-end reads or not).

Macrogen

In general, NextClip kept much fewer sequences as mate pairs than NxTrim. Also, the amount of reads considered unreliable by NextClip (that can be treated as single end reads) was higher than the amount of single end reads produced by NxTrim (probably because NxTrim would categorize great part of those same reads as paired end). See **Table S4.2.3** for details.

Table S4.2.3 - Amount of sequence reads from Macrogen categorized as either mate pairs or unreliable (single end) by NextClip.

Library/ Insert size	Total usable read pairs (NextClip, 50kb)	Percentage of usable reads	Unreliable reads	Percentage of unreliable reads (single end)
Gel free	80,633,002	30.56%	72,176,936	27.36%
3kb	28,221,594	11.72%	31,988,771	13.29%
5kb	48,852,831	19.73%	49,289,840	19.91%
8kb	51,570,823	20.41%	54,019,689	21.38%

DHM-RI

In general, NextClip kept much fewer sequences as mate pairs than NxTrim. Also, the amount of reads considered unreliable by NextClip (that can be treated as single end reads) was higher than the amount of single end reads produced by NxTrim (probably because NxTrim would categorize great part of those same reads as paired end). See **Table S4.2.4** for details.

Table S4.2.4 - Amount of sequence reads from DHM-RI categorized as either mate pairs or unreliable (single end) by NextClip.

Library/ insert size	**Usable pairs**	**Usable pairs (percentage)**	**Unreliable pairs**	**Unreliable pairs (percentage)**
Gel free	54,275,758	25.84%	64,961,351	30.92%
3kb	26,102,724	10.31%	28,336,062	11.19%
5kb	19,564,719	8.94%	20,974,018	9.59%
10kb	5,362,561	3.09%	5,159,852	2.98%

PROTOCOL S4.1.

Annotation of draft genomes

Original python programs are available in the "Other Python programs" section of the software page of the Laboratório de Anfíbios (http://www.ib.usp.br/grant/anfibios).

Dependencies

- Python 3.5+
- Perl 5.2+
- RepeatMasker: http://www.repeatmasker.org/
- Augustus: http://augustus.gobics.de/
- BLAST Command Line Applications (see manual at https://www.ncbi.nlm.nih.gov/books/NBK279690/)
- GeneMark: http://exon.gatech.edu/GeneMark/
- Original python programs (available at http://www.ib.usp.br/grant/anfibios/researchSoftware.html):
  - outfmt6_to_gtf.py
  - parse_annotations.py
  - parse_exons.py
  - sort_annotations.py
  - sudoParallelBlast.py

RepeatMasker

I used RepeatMasker to annotate all repetitive DNA sequences in the contigs and scaffolds assembled with REPdenovo.

run_repeatmasker.sh

```
#!/bin/bash
set -o errexit

# Configure the shebang according to the location of the bash
# executable in your system.

NCPU=64 # Change according to the number of CPUs to be used.
SP="xenopus tropicalis" # I used X. tropicalis as reference.

function main { # Define the main function
    RepeatMasker repeats.fasta \
        -species "${SP}" \
        -s -par ${NCPU} -engine crossmatch -frag 20000 \
        -nocut -ali -inv -small -xsmall -poly -ace -gff \
        -xm -excln
}

# Execute functions and close.
```

```
main
exit
```

GeneMark

GeneMark performs annotation of novel genomic sequences using a self-training algorithm.

Ab initio annotation using self-training algorithm

```
$ gmes_petap.pl --ES --sequence <masked multifasta>\
    --soft_mask 100 --cores <number of cpus>
```

This command line takes the sequences masked with repeat masker as input. Only lowercase repeats longer than a specified threshold (100) are considered.

BLAST

We compared all sequences from *Phyllobates terribilis* and *Scaphiopus holbrookii* with a genes (amino acids) from *Xenopus tropicalis*, *X. laevis* (available at http://www.xenbase.org/other/static/ftpDatafiles.jsp, version 9.0), *Nanorana parkeri* (available at http://gigadb.org/dataset/100132, version 2015-02-11), *Anolis carolinensis* (available at http://www.ensembl.org/Anolis_carolinensis/Info/Index/, version 2.0) and *Homo sapiens* (downloaded from Uniprot/ SwissProt on September 7, 2017).

We prepared the nucleotide database containing 1,341,855 sequences from *P. terribilis* and *S. holbrookii* with the following command line:

Prepare protein database

```
$ makeblastdb -dbtype nucl -title "Nucleotide database: masked
scaffolds" -parse_seqids -in scf_masked.fasta -out scf_db
```

Once the database was created, we used the homemade Python program sudoParallelBlast.py (see comments in the program) to prepare PBS scripts to run tBlastN and best match for each protein in the nucleotide database.

Prepare PBS scripts and execute them

```
$ cd <work directory with protein sequences and nucleotide db>
$ ./sudoParallelBlast.py
$ for s in script*.pbs ; do qsub ${s} ; done
```

Augustus

We trained Augustus a selected and manually reviewed database of genes from *ab initio* annotations from *Scaphiopus holbrookii* (280 genes for training, 200 genes for validation) as well as published genes from *Xenopus tropicalis* (BioProject ID: PRJNA205740; 500 genes for training, 3,210 genes for testing) and *Nanorana parkeri* (BioProject ID: PRJNA344660; 780 genes for training, 403 for testing). Genes were reviewed in Geneious to guarantee nucleotide sequences were all in the forward direction and that the genes contained no gaps. The *N. parkeri* training set was selected as the best training set due to the higher sensitivity and specificity in annotating genes from the training set as well as from the other two species.

Example command line for training

```
$ etraining --species=nparkeri NparkeriGenes.gb.train
```

Example command line for Augustus testing

```
$ augustus --species=nparkeri NparkeriGenes.gb.test
```

Once *N. parkeri* was selected as the best training set, we optimized prediction accruracy by adjusting the meta parameters in the species configuration file.

Optimizing meta parameters with optimize_augustus.pl

```
$ optimize_augustus.pl –species=nparkeri NparkeriGenes.gb
```

Once optimization was done, we re-trained Augustus with the meta parameters set by the program optimize_augustus.pl and splited the scaffolds of *S. holbrookii* and *N. parkeri* in several parts to facilitate parallel annotation.

Example command line for Augustus

```
$ augustus --species=nparkeri part1.fasta > part1.gff
```

Parsing and merging annotations

First, we converted all output tables from tBlastN to the GTF (GFF 2) format using the program outfmt6_to_gtf.py. Then we concatenated all GTF files (from Augustus, GeneWise, and RepeatMasker) and parsed all annotations using parse_annotations.py.

Merging annotations

```
$ python3 poutfmt6_to_gtf.py -i <table in output format 6> \
    --source “tBlastN”
$ cat *.gtf *.gff > annotations.gtf.merged
$ python3 parse_annotations.py -f <scaffolds in fasta format> \
    -g annotations.gtf.merged -o <output files prefix>
```

The program parse_annotations.py will merge annotations from different programs and categorize them into three groups:

- *Category 1*: non-overlapping annotations (ignoring repeats)
- *Category 2*: overlapping annotations (the range of the longest annotation contains the range of all other overlapping annotations)
- *Category 3*: overlapping annotations (the range of the longest annotation does not contain the range of all other overlapping annotations)

Annotations of each category are printed into separate files, which can be concatenated (e.g., using cat) and sorted (e.g., using sort_annotations.py) if needed.

PROTOCOL S4.2

Revision of the phyluce tutorial for identifying UCE loci and designing baits to target them

Original phyluce's tutorial is available at http://phyluce.readthedocs.io/en/latest/#. This revision was concluded on December 8, 2017, when the current phyluce was on v1.5.0. The phyluce package was written by Brant C. Faircloth. The interested reader is referred to the following publications:

Faircloth (2017) Brant C. Faircloth. Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution* **8(9)**: 1103–1112. doi: 10.1111/2041-210X.12754.

Faircloth et al. (2015) Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among hymenoptera. *Molecular Ecology Resources* **15(3)**: 489–501. doi: 10.1111/1755-0998.12328.

Faircloth et al. (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* **61(5)**: 717–726. doi: 10.1093/sysbio/sys004.

Dependencies

- OS: Unix based (phyluce is not supported on Windows)
- Anaconda or Miniconda (see Conda at https://conda.io/docs/user-guide/install/index.html)
- phyluce (see original documentation at http://phyluce.readthedocs.io/en/latest/installation.html)
- SAMtools (should become available after phyluce installation with CONDA)
- BEDtools (should become available after phyluce installation with CONDA)
- modify_headers_4_phyluce.py (homemade Python script, available at GitLab or upon email request)
- ART (art_illumina v2016.06.05)
- BLAT (Src35)
- Stampy (available ar http://www.well.ox.ac.uk/project-stampy)
- phyluce_probe_slice_sequence_from_genomes (modified from original, lines 267–273, skip questions; available at GitLab or upon email request)

Original and modified software, including complete Bash script for phyluce's tutorial IV, are available at: https://gitlab.com/MachadoDJ/Modified_phyluce_Tutorial_IV.

How to execute

Open a terminal window, move to the working directory, and run the following command lie:

```
Terminal window at working directory
nohup bash pipeline.sh > stdout.txt 2> stderr.txt &
```

- **nohup**: nohup is a POSIX command to ignore the HUP (hangup) signal. The HUP signal is, by convention, the way a terminal warns dependent processes of logout. Output that would normally go to the terminal goes to a file called nohup.out if it has not already been redirected.
- **bash** Bash is a Unix shell and command language.
- **pipeline.sh**: Example pipeline (available at https://gitlab.com/MachadoDJ/Modified_phyluce_Tutorial_IV/blob/master/phyluce_pipeline.sh)
- **> stdout.txt**: Writes standard output into stdout.txt
- **> stderr.txt**: Writes standard error into stderr.txt
- **&**: This trailing ampersand directs the shell to run the command in the background

How to edit this pipeline for your own usage

The pipeline.sh is a commented bash script that can be modified with the help of the comments therein. As I reuse this pipeline in the future, I intend to update those comments and possibly write a program with multiple arguments to automate the process of editing the pipeline.

PROTOCOL S4.3

New intra and inter–exon phylogenetic markers for Anura

Dependencies

- Python 3.5+
- CD-HIT: http://weizhongli-lab.org/cd-hit/
- PRIMER3: http://primer3.sourceforge.net/
- fa_encrypt.py: https://gitlab.com/MachadoDJ/FASTX/blob/master/fa_encrypt.py
- parseCDHIT.py: https://gitlab.com/MachadoDJ/findExonicMarkers/blob/master/parseCDHIT.py
- parseCDHITEST2.py: https://gitlab.com/MachadoDJ/findExonicMarkers/blob/master/parseCDHITEST2.py
- findPrimersForGenes.py: https://gitlab.com/MachadoDJ/findExonicMarkers/blob/master/findPrimersForGenes.py

Clustering exons using CD-HIT-EST and CD-HIT-EST-2D

CD-HIT is a very widely used program for clustering and comparing protein or nucleotide sequences. CD-HIT is very fast and can handle extremely large databases. CD-HIT helps to significantly reduce the computational and manual efforts in many sequence analysis tasks and aids in understanding the data structure and correct the bias within a dataset.

The CD-HIT package has CD-HIT, CD-HIT-2D, CD-HIT-EST, CD-HIT-EST-2D, CD-HIT-454, CD-HIT-PARA, PSI-CD-HIT, CD-HIT-OTU, CD-HIT-LAP, CD-HIT-DUP and over a dozen scripts.

- CD-HIT (CD-HIT-EST) clusters similar proteins (DNAs) into clusters that meet a user-defined similarity threshold.
- CD-HIT-2D (CD-HIT-EST-2D) compares 2 datasets and identifies the sequences in db2 that are similar to db1 above a threshold.
- CD-HIT-454 identifies natural and artificial duplicates from pyrosequencing reads.

- CD-HIT-OTU clusters rRNA tags into OTUs
- CD-HIT-DUP identifies duplicates from single or paired Illumina reads
- CD-HIT-LAP identifies overlapping reads

CD-HIT is currently maintained by the Dr. Li's group (http://weizhongli-lab.org/cd-hit/) at J Craig Venter Institute.

Preparation of sequence reads

The output from CD-HIT only prints up to 20 characters of the sequence ID. Therefore, I replaced all sequence IDs for an alphanumeric code using the homemade Python script named fa_encrypt.py. UUID of 16 characters are given to each sequence, so they can all be identified in the output of CH-HIT (I mean the log files, full descriptors are used on the multi-fasta output files).

```
Encrypt headers with fa_encrypt.py
$ python3 fa_encrupt.py -f <fasta file> -m 0 -o <output prefix>
Decrypt headers with fa_encrypt.py
$ python3 fa_encrupt.py -f <fasta file> -s <synonyms file> \
    -m 1 -o <output prefix>
```

Notes about CD-HIT-EST

CD-HIT-EST clusters a nucleotide dataset into clusters that meet a user-defined similarity threshold, usually a sequence identity. The input is a DNA/RNA dataset in fasta format and the output are two files: a fasta file of representative sequences and a text file of list of clusters. Since eukaryotic genes usually have long introns, which cause long gaps, it is difficult to make full-length alignments for these genes. So, CD-HIT-EST is good for non-intron containing sequences like EST.

```
Example command line for CD-HIT-EST
$ cd-hit-est -i <input fasta> -o <output filename> \
    -c <sequence identity threshold> -n <word size>
```

Notes about CD-HIT-EST-2D

CD-HIT-EST-2D compares 2 nucleotide datasets (db1, db2). It identifies the sequences in db2 that are similar to db1 at a certain threshold. The input are two DNA/RNA datasets (db1, db2) in fasta format and the output are two files: a fasta file of sequences in db2

that are not similar to db1 and a text file that lists similar sequences between db1 and db2. For same reason as CD-HIT-EST, CD-HIT-EST-2D is good for non-intron containing sequences like EST.

Example command line for CD-HIT-EST-2D

```
$ cd-hit-est-2d -i <input fasta 1> -i1 <input fasta 2> \
    -c <sequence identity threshold> -n <word size>
```

Thresholds and word sizes

For both CD-HIT-EST and CD-HIT-EST-1D, it is suggested that the word size should be chosen taken the following table into consideration:

-n <word size>	**-c <sequence identity threshold>**
8, 9, 10	0.90 ~ 1.0
7	0.88 ~ 0.90
6	0.85 ~ 0.88
5	0.80 ~ 0.85
4	0.75 ~ 0.80

Please check CD-HIT manual for additional details.

Clustering exons

First, I clustered exons from the same species at different identity thresholds using CD-HIT-EST to extract representative sequences at each identity level:

Example command lines for CH-HIT-EST

```
$ cd-hit-est -i renamed_pt_pterribilis_exons.fasta \
    -o pterribilis_cluster100.fasta -c 1.0 -n 10 -M 512000 \
    -T 64 -B 0 -g 1
$ cd-hit-est -i renamed_pt_pterribilis_exons.fasta \
    -o pterribilis_cluster95.fasta -c 0.95 -n 8 -M 512000 \
    -T 64 -B 0 -g 1
$ cd-hit-est -i renamed_pt_pterribilis_exons.fasta \
    -o pterribilis_cluster80.fasta -c 0.80 -n 5 -M 512000 \
    -T 64 -B 0 -g 1
$ cd-hit-est -i renamed_pt_sholbrookii_exons.fasta \
    -o sholbrookii_cluster100.fasta -c 1.0 -n 10 -M 512000 \
    -T 64 -B 0 -g 1
$ cd-hit-est -i renamed_pt_sholbrookii_exons.fasta \
    -o sholbrookii_cluster95.fasta -c 0.95 -n 8 -M 512000 \
    -T 64 -B 0 -g 1
$ cd-hit-est -i renamed_pt_sholbrookii_exons.fasta \
    -o sholbrookii_cluster80.fasta -c 0.80 -n 5 -M 512000 \
    -T 64 -B 0 -g 1
```

Then, different combinations of order os species (*Phyllobates terribiis* vs. *Scaphiopus holbrookii*, or vice-versa) and identity thresholds were used to cluster exons between species with CD-HIT-EST-2D:

Example command lines for CD-HIT-EST-2D

```
$ cd-hit-est-2d -i renamed_pt_pterribilis_exons.fasta \
    -i2 renamed_sh_sholbrookii_exons.fasta -o pt-sh-100.fasta \
    -c 1.0 -n 10  -M 512000 -T 64 -B 0 -g 1
$ cd-hit-est-2d -i renamed_pt_pterribilis_exons.fasta \
    -i2 renamed_sh_sholbrookii_exons.fasta -o pt-sh-95.fasta \
    -c 0.95 -n 8  -M 512000 -T 64 -B 0 -g 1
$ cd-hit-est-2d -i renamed_pt_pterribilis_exons.fasta
    -i2 renamed_sh_sholbrookii_exons.fasta -o pt-sh-80.fasta
    -c 0.80 -n 5  -M 512000 -T 64 -B 0 -g 1
$ cd-hit-est-2d -i renamed_sh_sholbrookii_exons.fasta \
    -i2 renamed_pt_pterribilis_exons.fasta -o pt-sh-100.fasta \
    -c 1.0 -n 10  -M 512000 -T 64 -B 0 -g 1
$ cd-hit-est-2d -i renamed_sh_sholbrookii_exons.fasta \
    -i2 renamed_pt_pterribilis_exons.fasta -o pt-sh-95.fasta \
    -c 0.95 -n 8  -M 512000 -T 64 -B 0 -g 1
$ cd-hit-est-2d -i renamed_sh_sholbrookii_exons.fasta
    -i2 renamed_pt_pterribilis_exons.fasta -o pt-sh-80.fasta
    -c 0.80 -n 5  -M 512000 -T 64 -B 0 -g 1
```

We parsed CD-HIT results using two independent homemade Python scripts, parseCDHITEST.py and parseCDHITEST2.py. With this strategy, we managed to merge merging clusters within each species as well as between them, and extract the better representative sequence from each cluster using CD-HIT parameters. This strategy follows a conservative rationale that ignores some valid exon pairs that are homologous to reduce the chance of comparing two non-homologous exons.

Example command lines for parseCDHITEST.py

```
$ python3 parseCDHITEST.py \
    --source1 <multifasta with renamed exons from species 1> \
    --source2 <multifasta with renamed exons from species 2> \
    --representatives1 <.clstr for sp. 1 from cd-hit-est> \
    --representatives2 <.clstr for sp. 2 from cd-hit-est> \
    --assembly1 <multifasta with scaffolds from species 1> \
    --assembly2 <multifasta with scaffolds from species 2>
```

The program parseCDHITEST.py will return multifasta files with the representatives from each cluster. The description of each entry include the cluster identification number and its size (i.e., how many exons are included in the same clusters). The program parseCDHITEST2.py will read these representative sequences along with the synonym files (in tab-separated format) and original scaffolds (in multifasta format) to calculate statistics and call PRIMER3.

Example command lines for parseCDHITEST.py

```
$ python3 parseCDHITEST2.py \
    --input1 <representatives for sp. 1 from parseCDHITEST.py> \
    --input2 <representatives for sp. 2 from parseCDHITEST.py> \
    --synonym1 <.tsv file with synonyms for exons in sp. 1> \
    --synomym2 <.tsv file with synonyms for exons in sp. 2> \
    --assembly1 <multifasta with scaffolds from species 1> \
    --assembly2 <multifasta with scaffolds from species 2>
```

Primer extraction

The program findPrimersForGenes.py takes the configuration files written by parseCDHITEST.py and uses them to run PRIMER3 and propose the primers for intra and inter-exon primers.

Example command lines for findPrimersForGenes.py

```
$ python3 findPrimersForGenes.py \
    --config <Path to the directory of configuration files> \
    --output <Output prefix> \
    --primer3 <Path to the primer3_core executable> \
    -t <.tsv with sequence ID, target start, end, and identity>
```

Note that findPrimersForGenes.py requires a working installation of Primer3 (tested with libprimer3 release 2.3.7).

PROTOCOL S5.1

Analyses of Repetitive DNA

Repeat assembly with REP*denovo*

Preparation of sequence reads

Left (first pair) sequence headers should end in "/1". Righ (second pair) sequence headers should end in "/2". This formation increases the chance that different programs will be able to parse the sequence files.

Paired-end reads should be in two different files, and the nth read pair in one file should correspond to the nth read pair in the other file.

Single end reads can be concatenated in a single multi-fastq file.

It is suggested that all sequence files are trimmed (Trimmomatic, NxTrim), filtered (HTQC tookit), and de-duplicated (FastUniq) before assembly.

Dependencies

- Python v2.7+: https://www.python.org/download/releases/2.7/
- REPdenovo (TERefiner and ContigsMerger v0.1.9.): https://github.com/Reedwarbler/REPdenovo
- JellyFish v2.2.6: https://github.com/gmarcais/Jellyfish/releases
- Velvet v1.2.10: https://www.ebi.ac.uk/~zerbino/velvet/
- BWA v0.7.12: http://bio-bwa.sourceforge.net/
- SAMtools v1.3.1: http://samtools.sourceforge.net/

Configuration

REPdenovo assemblies of PE and SE sequence reads separately. Hence, REPdenovo requires two configuration files and two text files with the path to the fastq files to be used.

The files_pe.txt file has the path to the PE files, one per line, in the first column. The second, third, and fourth column must contain the group number (starting with 1; files in the same group pair to each other), mean insert-size, and insert-size standard-derivation, respectively.

files_pe.txt

```
/path/to/pe1.fastq 1 400 200
/path/to/pe2.fastq 1 400 200
```

The files_se.txt file has the path to a single fastq file with the SE reads. For SE reads, there should be an “-1 -1 -1” string after the file path.

files_se.txt

```
/path/to/se.fastq -1 -1 -1
```

The configuration file for PE (configurations_pe.txt) and SE (configuration_se.txt) assembly are similar. The only difference is the value of the OUTPUT_FOLDER parameter, which indicates the location that the user wants to save the output files. Variables

BWA_PATH to REFINER_PATH (lines 17–21) have to contain the path to the corresponding executables of “GLOBAL”, if the program was installed in the default location and its available to the user’s path.

configuration_pe.txt

```
MIN_REPEAT_FREQ 10
RANGE_ASM_FREQ_DEC 2
RANGE_ASM_FREQ_GAP 0.8
K_MIN 23
K_MAX 63
K_INC 2
READ_LENGTH 126
GENOME_LENGTH 9000000000
MIN_CONTIG_LENGTH 126
ASM_NODE_LENGTH_OFFSET -1
IS_DUPLICATE_REPEATS 0.85
COV_DIFF_CUTOFF 0.5
MIN_SUPPORT_PAIRS 20
MIN_FULLY_MAP_RATIO 0.2
TR_SIMILARITY 0.85
TREADS 64
BWA_PATH GLOBAL
SAMTOOLS_PATH GLOBAL
JELLYFISH_PATH /path/to/JELLYFISH_226/bin/
VELVET_PATH /path/to/VELVET_1210/
REFINER_PATH /path/to/TERefiner_1
CONTIGS_MERGER_PATH /path/to/REPdenovo/ContigsMerger
OUTPUT_FOLDER ./output_pe/
VERBOSE 1
```

Execution

The Bash script bellow will execute REPdenovo and assembly repetitive DNA from PE and SE reads. Assembling PE reads occurs in two steps: assembly and scaffolding. The scaffolding step is only for PE reads.

run_repdenovo.txt

```
#!/usr/bin/bash
set -o errexit
#
# Please change the shebang line to match the location of Bash
# in your system.
# Lines that require editing to match your system requirements
# are indicated with an arrow ("<——").
#
# Set variables
OMP_NUM_THREADS=63 # <——
OMP_THREAD_LIMIT=64 # <——
REPdenovo_path="/path/to/REPdenovo" # <——
#
function assembly_pe { # Defines new function: assembly_pe
# Assembly
printf "> Assembly started — PE [ `date` ]\n"
`which python` ${REPdenovo_path}/main.py -c Assembly \
```

```
    -g configuration_pe.txt -r files_pe.txt
wait
# Scaffolding
printf "> Scaffolding started [ `date` ]\n"
`which python` ${REPdenovo_path}/main.py -c Scaffolding \
    -g configuration_pe.txt -r files_pe.txt
wait
printf "> Finished – PE [ `date` ]\n"
}
#
function assembly_se { # Defines new function: assembly_se
printf "> Assembly started – PS [ `date` ]\n"
`which python` ${REPdenovo_path}/main.py -c Assembly \
    -g configuration_se.txt -r files_se.txt
wait
printf "> Finished – SE [ `date` ]\n"
}
#
# Execute functions and close
assembly_pe
assembly_se
exit
```

Once assembly is finished, there will be two main output files in the output directory:

- contigs.fa: mutifasta file which contains the constructed repeats
- X_contig_pairs_info.txt_cov_info_with_cutoff.txt: text file which contains the repeat coverage information

The columns in the X_contig_pairs_info.txt_cov_info_with_cutoff.txt file correspond to repeat ID, repeat length, and the repeat average coverage.

In theory, the information on this last file could be used to estimate the average number of copies of each repeat. However, our initial calculations suggested that the total amount of repetitive DNA was sometimes two times bigger than the expected genome size. I wrote the programmer of REPdenovo, Chong Chu (personal communication, Feb. 2, 2017), who said that this error could be related to the "bwa mem -a" option in REPdenono, which reports all the possible alignments and causes overestimation of repeat coverage. This forced me to find an alternative solution to estimate the number of copies of each repeat.

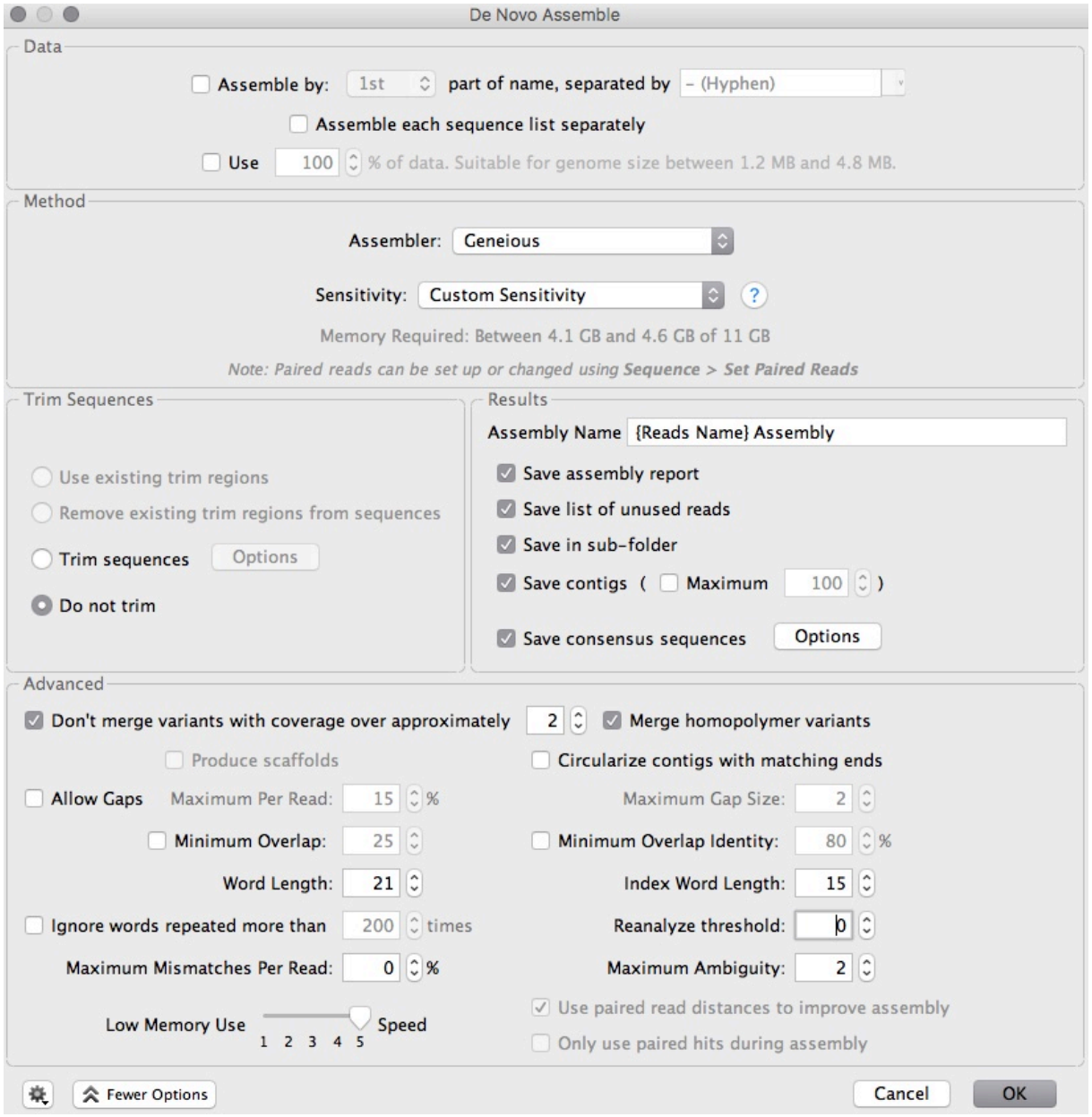
Estimating the number of copies of each repeat

Dependencies

- Geneious v8.1.9: https://www.geneious.com
- Bowtie2 v2.2.3: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- SAMtools v1.3.1: https://sourceforge.net/projects/samtools/files/samtools/1.3.1/
- BAMtools v2.4.1: https://github.com/fd00/yacp/tree/master/bamtools
- BEDtools2 v2.26.0: https://github.com/arq5x/bedtools2/releases
- genomeCoverage.py: http://www.ib.usp.br/grant/anfibios/researchSoftware.html
- polishStats.py: http://www.ib.usp.br/grant/anfibios/researchSoftware.html

Merging contigs and scaffolds

The scaffolds and contigs from independent PE and SE analysis were merged in Geneious v8.1.9, using the built-in Geneious *de novo* assembly algorithm and the parameters specified in the window bellow.



Geneious requires a paid license, but there are free programs that can perform the assembly of unpaired sequences with 100 bp or more. No matter which approach is used to combine different scaffolds and contigs, it is key to choose parameters that are very restrictive, avoiding the creation of chimeric sequences. Geneious de novo assembly parameters above required a word match of at least 21 letters, with 100% sequence identity.

The resulting consensus sequences and all the unaligned sequences are combined. I renamed all the sequence headers to have a short ID in the format "repeat<num-

ber>," while saving the original headers on a separate synonyms file with two columns corresponding to the original and the modified header.

Estimating repeat coverage

I mapped the same input sequences used for repeat assembly against the processed repeats and a set of selected, curated and single-copy genes that are available in GenBank (see table bellow).

Acession number			
Scaphiopus holbrookii	*Melanophryniscus moreirae*	*Hyloxalus subpunctatus*	*Phyllobates terribilis*
AB612074.1	AB612060.1	HQ290611.1	HQ290643.1
NM_001015963.1	HQ291008.1	HQ291034.1	HQ291066.1
NM_001016370.1	—	HQ290791.1	HQ290823.1
NM_001011318.1	AF194966.1	HQ290851.1	HQ290883.1
AB612071.1	KF666223.1	DQ503405.1	DQ503358.1
AY323738.1	—	DQ503279.1	DQ503244.1
DQ282710.1	XM_018553010.1	GQ366167.1	GQ366169.1
JQ626831.1	JQ626831.1	JF703249.1	JQ626814.1
AB612073.1	KX026238.1	HQ290911.1	HQ290943.1
XM_012971006.1	XM_018554053.1	HQ290671.1	HQ290703.1
—	DQ306494.1	—	—

I performed read mapping using Bowtie2 v2.3.3, building an index for the repeats and single-copy genes together and them mapping all the original sequence reads to this index. I used SAMtools v1.3.1 to convert the SAM alignments to BAM format and sort the BAM file. Finally, I used BEDtools2 v2.26.0 to estimate the coverage of each sequence in the index.

estimate_read_coverage.sh

```
#!/usr/bin/bash
set -o errexit

# The fist line is the shebang and must be modified to match the
# location of the Bash executable in your system.

module load bowtie2/2.3.3 bedtools2/2.26.0 samtools/1.3.1 bamtools/2.4.1
# The line above is an example of how to make the programs that are used here
available to the user in a system that has modules. Just make sure the programs
# are correctly install and that they are accessible to the user.

NCPU=64 # Change to the number of CPUs to be used.

function ix { # This function build the index.
    bowtie2-build reference.fasta repeats.fasta my_index
    wait
}

function bt2 { # This function maps the sequence reads to the index.
    bowtie2 -x my_index -p ${NCPU} -1 pe1.fastq -2 pe2.fastq -U se.fastq \
        -S alignment.sam --no-unal --local --very-sensitive-local
    wait
}

function cov { # This function sorts the alignments and calculates the coverage.
    samtools view -Sb alignment.sam > alignment.bam
    wait
    samtools sort alignment.bam alignment.sorted
    wait
```

```
    genomeCoverageBed -ibam alignment.sorted.bam > my_coverage.txt
    wait
}

# Execute the functions and quit.
ix
bt2
cov
exit
```

Histograms from genomeCoverageBed (that comes with BEDtools2) were parsed using a home-made Python script named estimateCopyNumber.py to calculate the average coverage of each fragment and estimate the number of copies of each repeat sequence by comparing their coverage with the coverage of single-copy genes. Calculation proceeds as follows:

1. Calculate the average coverage per sequence from the genomeCoverageBed output histogram
2. Calculate the factor *F*, which is the maximum ratio of average coverage and sequence length for single-copy genes
3. Calculate the expected copy number of each repeat by dividing the average coverage of the repeat by its length and by factor *F*.

$$Number\ of\ copies = \frac{\frac{Avg.\ coverage}{Length}}{Factor}$$

Run "python3 estimateCopyNumber.py --help" for details on how to execute this program. See comments therein for additional information.

Note that estimateCopyNumber.py requires the following:

- The histogram from genomeCoverageBed
- List of repeats that match the mitochondrial genome (can be retrieved from BlastN searches, not shown here)
- The multifasta file with the repeats assembled with REPdenovo (using renamed, simplified headers is advisable)
- The IDs for the single-copy genes

Find primers for repetitions

Dependencies

- Unix system with Perl v5.8.0 or higher installed
- Sequence Search Engine
  - Cross_Match: http://www.phrap.org
- Tandem Repeat Finder (TRF): http://tandem.bu.edu/trf/trf.html
- Repeat Databases
  - Dfam libraries: http://www.dfam.org
  - RepBase RepeatMasker Edition: http://www.girinst.org
- RepeatMasker: http://www.repeatmasker.org/RMDownload.html
- Primer3 v2.3.7: https://sourceforge.net/projects/primer3/
- findPrimersForRepeats.py: http://www.ib.usp.br/grant/anfibios/researchSoftware.html

Repeat annotation with RepeatMasker

I used RepeatMasker to annotate all repetitive DNA sequences in the contigs and scaffolds assembled with REPdenovo.

run_repeatmasker.sh

```
#!/bin/bash
set -o errexit

# Configure the shebang according to the location of the bash
# executable in your system.

NCPU=64 # Change according to the number of CPUs to be used.
SP="xenopus tropicalis" # I used X. tropicalis as reference.

function main { # Define the main function
    RepeatMasker repeats.fasta \
        -species "${SP}" \
        -s -par ${NCPU} -engine crossmatch -frag 20000 \
        -nocut -ali -inv -small -xsmall -poly -ace -gff \
        -xm -excln
}

# Execute functions and close.
main
exit
```

Find primers

I used the homemade Python script named findPrimersForRepeats.py to find primers for selected types of satellites or transposable elements (TEs). This program takes a MULTIFASTA file with repetitive DNA masked (lowercase) with RepeatMasked. It also takes an annotation file in GFF format from RepeatMasker. The findPrimersForRepeats.py is a helper script to execute Primer3, that must be correctly installed in your system.

Built-in configuration for satellites:

- Selected types: Chap4sat_Xt, MSAT2_XT, MSAT4_XT, REM2b_Xt, Sat1_Xt, Tc1Sat1_Xt, URR1a_Xt, and simple repeats with units with four or more nucleotides
- Maximum repeat length: 150 bp
- Primer TM: 58–62 ºC (optimal = 60 ºC)
- Maximal TM difference between left and right primers: 2 ºC
- Primer GC content: 40–60 % (optimal = 50 %)
- Product length: 80–150 bp

Built-in configuration for TEs:

- Selected types: SINE, MIR, L1, CIN4, L2, CR1, REX, and Penelope
- Maximum repeat length: 300 bp
- Primer TM: 57–63 ºC (optimal = 60 ºC)
- Maximal TM difference between left and right primers: 2 ºC
- Primer GC content: 20–80 % (optimal = 50 %)
- Product length: repeat length + 36 bp to repeat length + 80 bp

Run “python3 findPrimersForRepeats.py --help” for details on how to execute this program. See comments therein for additional information.

Word count

A homemade Python script called wordSimilarity.py was used to count identical substrings of DNA sequence of size *n* among a certain number of multifasta files. The program is available at http://www.ib.usp.br/grant/anfibios/researchSoftware.html and can be executed as so:

Executing wordSimilarity.py

```
$ python3 wordSimilarity.py --input file1.fasta file2.fasta …
fileN.fasta --size 11
```

PROTOCOL S6.1

Docking analysis

Model selection

Tarvin *et al*. (2016: Table S5) list 49 residues of "Walker1" (Walker et al. 2012: Database S1, homology model for Scn4a) predicted by protein docking analyses to interact with seven poison frog alkaloids. Herein, we call "primary alkaloid resistance candidate" (PARC) any modification in these 49 residues. To filter the PARC list further, the authors aligned the DI-S6, DII-S6, DIII-S6, and DIV-S6 regions of the Scn4a gene of 38 species of animals (14 outgroup terminals and 24 Dendrobatidae terminals) and selected all the PARCs that were exclusive to Dendrobatidae (even if they were not exclusive to frogs with alkaloids). From now on, a residue selected this way by Tarvin *et al*. will be referred to as "secondary alkaloid resistance candidate" (SARC).

Using the strategy above, Tarvin *et al*. found 6 SARCs in five positions: 429A, 433V, 445D/E, 1583I, and 1584T. Two SARCs, 429A and 1584T, were found only in the Scn4a of *Phyllobates terribilis* and were not recovered in our draft assembly of the nuclear genome of this species of poison frog. The other SARCs amount to 7 modified receptor models (MRMs): 4 MRMs with a single AA replacement and 3 species-specific MRMs with multiple AA replacement patterns. From those 7 MRMs, we selected the ones with SARC patterns observed in Dendrobatidae (e.g., the MRM with SARC 445E and no other SARC was not included in our analysis because no poison frog posses this modified residue solely).

Aside from the MRMs from Tarvin *et al*., we also created new modified receptor models (NMRMs) for other species of non-poison frogs with SARCs that were not included by Tarvin *et al*. and that disrupt the phylogenetic pattern of these residues as described by the authors. This way, we created NMRMs for chickens, fire salamanders, and turquoise killifishes using all the corresponding PARCS and SARCS found in the Scn4a of these species. We also created the "complete" Dendrobates NMRM including all the observed aminoacid modifications between rats and *Dendrobates tinctorius*.

Model Name	423	424	429*	433*	445*	782	1262	1276*	1287*	1565	1569	1581*	1583*	1584*
Epipedobates					D									
Ameerega					D								I	
Hyloxalus													I	
Excidobates				V	E									
Dendrobates				V	D								I	
Killifish					D				V			V	I	
Chicken								A						
Salamander								A					I	
Complete	L	V		V	D	V	M	A		A	V		I	
Phyllobates			A	V	D								I	T

(*) Positions associated by Tarvin et al. with modifications that could lead to alkaloid resistance.

Ligands

The following are the seven original selected ligands and their ChemSpider IDs (searchable at http://www.chemspider.com), plus samandarin and samandarine used in our analysis.

- Batrachotoxin–BTX: Molecular Formula: C31H42N2O6. Average mass: 538.675 Da. Monoisotopic mass: 538.304260 Da. ChemSpider ID: 10310314
- aPTX 267A: Molecular Formula: C16H29NO2. Average mass: 267.407 Da. Monoisotopic mass: 267.219818 Da. ChemSpider ID: 4580699
- PTX 307A: Molecular Formula: C19H33NO2. Average mass: 307.471 Da. Monoisotopic mass: 307.251129 Da. ChemSpider ID: 9154941
- PTX 323A: Molecular Formula: C19H33NO3. Average mass: 323.470 Da. Monoisotopic mass: 323.246033 Da. ChemSpider ID: 4941919
- aPTX 323B: Molecular Formula: C19H33NO3. Average mass: 323.470 Da. Monoisotopic mass: 323.246033 Da. ChemSpider ID: 4518097
- Pumiliotoxin–PTX 251D: Molecular Formula: C16H29NO. Average mass: 251.408 Da. Monoisotopic mass: 251.224915 Da. ChemSpider ID: 4944741
- Histrionicotoxin–HTX: Molecular Formula: C19H25NO. Average mass: 283.408 Da. Monoisotopic mass: 283.193604 Da. ChemSpider ID: 4941928
- Samandarin: Molecular Formula: C19H31NO2. Average mass: 305.455 Da. Monoisotopic mass: 305.235474 Da. ChemSpider ID: 107738
- Samandarine: Molecular Formula: C19H31NO2. Average mass: 305.455 Da. Monoisotopic mass: 305.235474 Da. ChemSpider ID: 107738

Original methods

The following is described in Tarvin *et al*. (2016: p.1,078).
Protocol

1. Python Molecule Viewer (PMV; see Sanner 1999)
1.1. Set up docking parameters
1.2. Add Gasteiger PEOE partial charges to the ligands (alkaloids; BTX, HTX, and 5 PTX alkaloids)
1.3. Add Kollman United Atom charges to the receptor (protein model)
2. Autodock Vina (Trott & Olson 2010):
2.1. Searching a 3D rectangular prism centered in the middle of the protein that encompassed both inner and outer regions of the channel pore (pore loop and S6) with the exhaustiveness parameter set to 10,000.
2.2. For each permutation of mutation pattern and alkaloid, retain the top three docking positions from five different runs to obtain N = 15
3. Chimera (see Pettersen *et al*. 2004):
3.1. Mutate the Nav1.4 model from Walker *et al*. (2012: Dataset S1) to contain each single AA replacement as well as unique species-specific multiple AA replacement patterns found in DI-S6 and DIV-S6 of dendrobatids
4. Reran the docking analysis (2) to determine the effect of single and multiple AA replacement patterns on alkaloid binding affinity
Interpretation of results

If the Gibbs free energy (ΔG; converted from kcal to kJ) of the new predicted docking site increased (became more positive/ less negative), then the alkaloid had a lower affinity for the mutated model, suggesting that the replacement pattern provides decreased sensitivity to that alkaloid.

Statistics

Nonparametric unpaired Wilcoxon rank-sum tests in R v3.0.2 (R Core Team 2013) comparing the free energy of alkaloid docking to ancestral ("Walker1") and derived Scn4a models. The authors reported test statistic (W) and its P-value (P).

Modifications on the original methods

Protocol

1. Download receptor (Scn4a model from Walker *et al*. 2016: Dataset S1), save as PDB

2. Download ligands from ChemSpider in MOL format

3. Use babel program to convert files in MOL to MOL2

Convert MOL to MOL2

```
$ for i in *.mol ; do obabel ${i} -O ${i}2 ; done
```

4. Open the receptor PDB file in Chimera (Pettersen *et al*. 2004), perform AA changes according to several models

4.1. On Chimera, residues are changed using the structure editing > rotamers option (there is no "mutation" on the options menu). The most likely (higher probability) rotamer option is selected if it does not have any conflict with any other residues

4.2. Perform clash and contact analysis after each rotamer modification using default parameters. Is any clashes or contacts were found, performed minimization analysis using default parameters to correct them, and keeping all unselected (non-problematic AA) fixed. If minimization cannot correct the clades and contacts, select another rotamer option and repeat. Note: Clashes and contacts were found and corrected in 433C, 1622M, and 1583I. No conflicts remained in the modified AA on the final homology model.

4.3. Save modified receptor models in PDB format

5. Open each receptor model in Python Molecule Viewer (PMV; Sanner 1999)

5.1. Add polar hydrogen atoms

5.2. Add Kollman United Atom charges

5.3. Initiate AutoDock Tools > Grid > select macromolecule and save PDBQT file

5.4. Set grid options for 3D rectangular prism centered in the middle of the protein that encompassed both inner and outer regions of the channel pore (pore loop and S6) using the following parameters: X-dimension = 50; Y-dimension = 52; Z-dimension = 52; Spacing (angstrom) = 1.000; X center = 95.585; Y center = 93.512; Z center = 23.795.

5.5. Read ligands, check the hydrogen atoms adding them if needed, compute Gasteiger changes, use the Torsion tree option to check the rotatable bonds (i.e., make sure rotatable bonds were loaded correctly), export PDBQT file.

6. Run docking analysis in AutoDock Vina (Trott & Olson 2010).

6.1. Edit the following configuration file according to each permutation of receptor and ligand (performed using a Bash script):

Template Bash script for AutoDock VINA

```
receptor = <receptor>.pdbqt
ligand = <ligand>.pbqt
out = out_<ligand>.pbqt
log = log_<ligand>.txt
center_x = 95.585
center_y = 93.512
center_z = 23.795
size_x = 50
```

```
size_y = 52
size_x = 52
exhaustiveness = 10000
cpu = <number of available CPUs>
```

6.2. Execute each configuration file (performed using a Bash script):

```
Template command line for AutoDock VINA
$ vina --config <ligand>.conf
```

6.3. Retain the Gibbs free energy (ΔG) of the three best docking sites (lowest RMSD values). Convert ΔG from kcal/mol to kJ/mol.

Interpretation of results

Basics of thermodynamics say that a reaction with negative free energy is spontaneous. In other words, the smaller the ΔG, the higher the requested energy to break that binding between two molecules. Therefore, negative binding free energy values during molecular docking mean that ligand bounds to the receptor. However, the binding affinities reported by Vina are only predictive and rely on an energy function somewhat empirical. Vina is valuable for predicting a binding mode of a small ligand, but ranking two bound models is really up to the user appreciation.

*Affinity—*Δ*G*: As a rule of thumb, binding affinity is considered high for models with binding ΔG equal to or smaller than –6 kcal/mol or –25.104 kJ/mol (e.g., Zaidi *et al*. 2013). However, the ΔG as a measure of affinity is usually not trusted on its own and requires cross-validation with other methods, often involving correlation analysis with activity scores.

*Affinity variation–*Δ*ΔG*: It is hard to find theoretical material indicating which value of ΔΔG is biologically significant, and these might be very specific to the system at hand. In Zaidi *et al*. (2013), a comprehensive insight into the binding of hippuric acid (HA) to human serum albumin (HSA), the ΔΔG between high affinity and low-affinity sites of HA-HSA vary between 5,69024 and 6,217424 kJ/mol. However, in view of the errors typically associated with ligand docking in programs such as AutoDock Vina (including the potential pitfalls in handling the flexibility of the receptor), a ΔΔG value of 2kcal/mol or 8.383kJ/mol may not represent a conclusive difference and many authors prefer to correlate affinity score with affinity data before considering the ΔΔG significant (i.e., the ΔΔG is considered significant if there is a strong correlation between affinity and activity in both targets, and the ΔΔG is superior to the combined root-mean-square deviation).

Therefore, the mere statistical significance of ΔΔG does not suffice for justifying claims of a biologically significant difference in affinity, especially if the ΔΔG is small (i.e., a few decimals).

Statistics

We repeated the statistical analysis of the original article, and also compared the affinity of the chicken and killifish models with *D. tinctorius*. However, we disagree that test results that are statistically significant (i.e., *p*–value < 0.05) are necessarily an indication of biologically significant differences in receptor affinity to the ligands. All models with statistically significant Wilcoxon-Mann-Whitney tests showed an overlapping range of ΔGs and all results were more negative than -6 kcal/ mol. Also, following Tarvin *et al*.'s methods

and interpretation of statistical results, we could conclude that the killifish and chicken models are more resistant to alkaloides than the *D. tinctorius* model, which is obviously nonsensical.

Appendix S6.1

Specific comments on Tarvin *et al*. (2016)

Taxonomy— *Bufu nebulifer* = *Incilius nebulifer*, *Dendrobates captivus* = *Excidobates captivus*.

Residue identification: A446D/ E = A445D/ E.

ABSTRACT—"We predict that poison frogs are somewhat resistant to these compounds because they have six types of amino acid replacements in the $Na_V1.4$ inner pore that are absent in all other frogs except for a distantly related alkaloid-defended frog from Madagascar, *Mantella aurantiaca*. Protein-docking models and comparative phylogenetics support the role of these replacements in alkaloid resistance."

Comment—Of the six amino acid replacements, four are found in the sodium channels of other organism (e.g., mouse, whale shark, zebrafish, and humans). These four replacements appear in other types of Na_V but also in Na_V 1.4 of distantly related organisms. The remaining two replacements are exclusive to *P. terribilis*, but could not be found in our draft assembly of the genome of *P. terribilis*.

Page 1,069—"We focus on the skeletal muscle VGSC, Na_V 1.4, which is one of three Nav1 paralogs expressed outside of the central nervous system in frogs and is likely exposed to relatively high levels of alkaloids (Zakon 2012; McGlothlin *et al*. 2014)."

Comment—Although the authors focus on a particular paralog of Na_V1, the AA mutations on the specified positions would confer resistance to alkaloids to any of sodium channel proteins in which they were found. Also, the alkaloids seem to be widely distributed in the frog's body and there is no special reason to assume that the central nervous system would not need to resist to high concentrations of alkaloids.

Page 1,069—"(...) and propose similar binding sites for HTX and PTX, which were previously unknown."

Comment—Actually, only two new binding sites are proposed, based on mutations that were only observable in the sequences of Nav1.4 of *P. terribilis* sequenced and assembled by the authors.

Page 1,077—"We reviewed the alignment site by site to identify residues that were highly conserved in non-dendrobatid frogs and other vertebrates, but showed patterns of AA replacements that were associated with chemical defense in dendrobatids."

Comments—The authors do not explain how broad were their searchers, but indicate that only the sequences listed on the article and its supplementary material were compared to each other. That might explain why the authors missed other Nav1.4 AA sequences with the same mutations as the poison frogs on UniProt (Swiss-Prot and TrEMBL) and GenBank (RefSeq-Prot).

Page 1,077—"A reference molecular phylogeny of amphibians and outgroups was inferred for the 39 species in figure 1B using previously published mitochondrial sequences (see supplementary table S6 and figure S1, Supplementary Material online, for accession numbers and full phylogeny)."

Comment—The tree on the supplementary material is very polytomyc and do not match the tree on Figures 1 and 2. Also, these topologies do not match the trees on the specialized literature. Tree topology might affect ancestral state reconstruction and dating.

Page 1,077—Evolutionary and ecological analysis.

Comment—The Pagel's test and the tests of phylogenetic correlation using the pgls function in the R "caper" package are affected by the presence of AA sequences of animals which are not resistant to alkaloids and possess the same mutations as poison frogs.

Page 1,069—"All five sites [*i.e.*, 429, 433 and 446 in DI-S6 and 1583 and 1584 in DIV-S6] are intriguing because these residues are highly conserved among vertebrates, and the replacements are unique to poison frogs and Mantella".

Comment—This is not correct. First, position 446 is not shown in the paper. DI-S6 goes up to position 445 in all the sequences shown by the authors in the paper or in the supplementary material. This is most likely a counting error, and we will refer to position 446 in DI-S6 as position 445 instead. Second, see comments below:

1) 429A and 1584T: are unique to the material of *Phyllobates terribilis* sequenced by the authors. In our draft assembly of the *P. terribilis*, we could not find these mutations.

2) 433V: Occur in the DI-S6 region of Scn4a in *P. terribilis*, *D. tinctorius*, and *D. captivus*. The 31 AA fragment of the DI-S6 is identical to 31 aa fragments of the Scn2a-like protein in *Rhincodon typus* (XP_018409738), Scn3a in *Rattus norvegicus* (P08104) and *Homo sapiens* (Q9NY46-4), and Scn1a in *Mus musculus* (AAH23034.1) and *Cricetulus griseus* (EGW09057.1); among others. The mutation is also found in the DI-S6 of the scn4a of *Nothobranchius furzeri* (A0A1A8UIU5) and *Anilios bituberculatus* (ANP22541.1).

3) 445D and 433V: These residues occur in *Anolis carolinensis* (XP_016850007), *Salmo salar* (XP_0160836333), and *Danio regio* (Q2XVR3), among others.

4) 446E: In Tarvin *et al*. 2016, I446E is exclusive to *De. captivus*. It is also found on our scaffolds of *Ph. terribilis*. Furthermore, the 31 AA fragment of Scn4a DI-S6 from *De. captivus* is identical to a 31 AA fragment of *Na. parkeri* Scn2a-like (XP_018409738) and *Rattus norveticus* Scn3a (P08104).

5) 1583I: Occur in *Salamandra salamandra*, but not other more toxic genera of newts. Occur in sodium channel proteins of frogs that do not have alkaloids (e.g., *Xenopus laevis* - A0A1L8HA74_XENLA). Occur in other types of sodium channels proteins of various animals (e.g., whale shark - XP_020387692, copepods - A0A125R3Q8_9MAXI).

Page 1,070—"Finally, selection analyses did not identify any sites under positive selection, but three sites that were not associated with chemical defense were identified to be under strong purifying selection, namely: V422 (...), A439 (...), and V1582 (...)."

Comment—See comments below.

1) V422 is invariable in Tarvin *et al*. (2016), but a quick search on UniProt and GeneBank revealed Nav1.4 sequences from different groups of animals (including fish, mollusks, and arthropods) which have different AA on this position (e.g., XP_014770743, XP_021341526, XP_013120659, XP_002427248).

2) A439: This residue does not exist. In all the sequences in Tarvin *et al*. (2016), position 439 has a V. It is possible that the authors were referring to residue A438, which is extremely stable even amongst distantly related organisms.

Page 1.073 (Docking Analysis, also see p. 1,078)—The authors modified only the amino acids in positions 429, 433, 446, 1583, and 1584. However, there are many other differences when we compare the Nav1.4 of poison frogs and rat (e.g., 419, 423, 424, 774, 777, 779, 782, 1263, and 1276). Later they state that their results

suggest "synergistic increases in alkaloid resistance." It seems odd that, if combinations of residues affect the result, the authors will ignore the possible effects of the remaining residues that are not identical to the Nav1.4 of the rat. Also, if only five positions were modified, none of the models represent the full species model of any poison frog.

FIGURES

TABLES

SUPPLEMENTARY MATERIAL

GLOSSARY

BTX	Batrachotoxin
HTS	High-throughput [DNA] sequencing
HTX	Histrionicotoxin
NGS	Next-generation [DNA] sequencing
PTX	Pumiliotoxin
Sae	Samandarine
San	Samandarin
TTX	Tetrodotoxin
WGS	Whole-genome sequencing