Measuring quality control of sequences of the mitochondrial barcode gene Cytochrome Oxidase subunit I (COX1) in highly diverse groups: The Poison Frog (Dendrobatidae) Case

Authors: James Tu[1], Rochelle Avezki[2] Mentor: Juan C Santos[3]

[1]The Kew Forest School, [2]Townsend Harris High School, [3]St. John's University

URBAN BARCODE RESEARCH PROGRAM
Supported by the Pinkerton Foundation & Science Sandbox

CSH Cold Spring Harbor Laboratory DNA LEARNING CENTER

Abstract: Species identification by amplifying and sequencing segments of DNA can reveal the identity of unknown samples of individuals. DNA barcoding has served immensely in cataloging biodiversity. One of the most common barcodes are fragments of the mitochondrial gene, COX1, readily available in large public databases, including NCBI. A crucial step for DNA barcoding is making sure that reference sequences are of high quality. However, as such databases continue to grow, the accuracy and reliability of the data can decline as a consequence of variance in the quality of the sequences deposited by the researchers doing the actual barcoding. In order to address this problem, our team created a framework of quality control (QC) of the COX1 gene in the GenBank database, using poison frogs (Dendrobatidae) as an example. After collecting 1248 sequences and their metadata, we compared different levels of barcode quality using year of submission, sequencing technology, taxonomic placement, ambiguities, and completeness to determine misidentifications and contaminants. Our results proved that some sequences were misidentified as different genera of dendrobatid frogs if compared to the rest that were correctly assigned. Moreover, some sequences assigned as poison frogs were even worse contaminants of distant organisms (e.g., wasps or fish).

Introduction

DNA barcodes are a standardized short sequence of DNA from 400-800 base pairs that should, in theory, be characterized for every species on our planet, even humans. Barcodes have been compiled alongside consortiums containing reference libraries of known species, such as the International Barcode of Life (iBOL) [5].

Molecular barcodes are crucial to track and discover the diversity of organisms on Earth, such that around 9 million animals, plants, and other living organisms have been barcoded in iBOL [5]. However, as these databases have grown, taxonomic misidentification has caused the reliability of public DNA to diminish. Likewise, with the pressing declines of biodiversity, DNA barcode sequences are becoming ever more important in studying, tracking, and improving the precision of species identification.

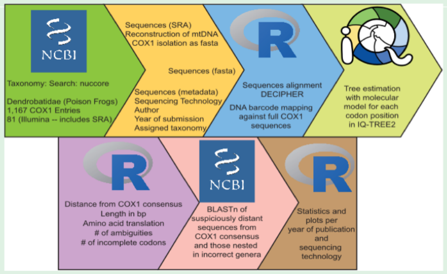
The Consortium for the Barcode of Life (CBOL) is an international initiative devoted to developing DNA barcoding as a global standard for the identification of biological species and uses the ideal barcoding gene, cytochrome c oxidase subunit I [2]. Reasons why the COX1 gene is used is because amplifying the gene is easy since cells can have hundreds of mitochondria, the gene is present in most eukaryotic cells, and is highly conserved in species that use mitochondria for energy, in accordance with the endosymbiotic hypothesis.

Misidentification of species can best be explained as a type of false-positive error which signifies that the genetic evidence may have been altered at the site of extraction or checked for the wrong organism in a lab. These alterations can be caused by mislabeling evidence, contamination of samples during PCR, and PCR-based errors such as chimeric sequences [8]. Typically, researchers revise each other's barcodes by updating the taxonomy, thus creating a separate obsolete taxonomy and increasing the QC of the database altogether.

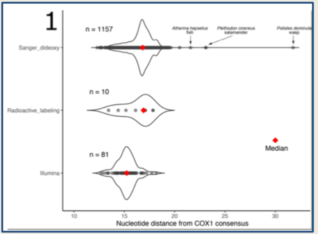
There are many types of DNA sequencing methods, but in this study, we are investigating three: Sanger sequencing, Radiolabelling, and Illumina. The oldest one is Radiolabelling, which radioactively labels nucleotides for the detection of specific nucleic acid sequences, but is very tedious [3]. Sanger sequencing is also an older technology and can be performed manually or automated by the same three steps: PCR with ddNTPs, separation by gel electrophoresis, reading the gel results to determine the sequence of the DNA. Sanger sequencing can only sequence one DNA fragment at a time while Illumina is able to sequence millions of fragments in parallel [4]. Illumina also has the advantage of a higher mutation resolution and higher sensitivity to detect lower-frequency variants [4]. With the rise of NGS technologies such as Illumina, DNA barcoding faces the issue of becoming obsolete, as such technologies are more efficient and are able to provide more complete genomic data than DNA barcoding at a rapidly declining price

Being mentored by Dr. Santos (St. John's University), who studies amphibians, our team became particularly interested in researching amphibians and the relevance of COX1 as a barcode in a well-known and biodiverse group, poison frogs (Dendrobatidae). By analyzing the quality of the collected data we to create a framework for quality control in online databases such as Genbank using COXI as an example, so that we may help to improve the quality of sequences in these databases for future researchers. Also, we plan to determine the improvements in the quality of COX1 barcodes Dendrobatidae in large, thought to be reliable, gene banks over the span of 23 years since the submitted first barcode in 1998.

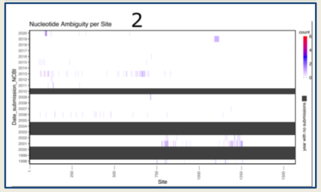
Materials & Methods


NCBI
Taxonomy: Search: nuccore
Dendrobatidae (Poison Frogs)
1,167 COX1 Entries
81 (Illumina -- includes SRA)
Sequences (SRA)
Reconstruction of mtDNA
COX1 isolation as fasta
Sequences (fasta)
Sequences (metadata)
Sequencing Technology
Author
Year of submission
Assigned taxonomy
R
Sequences alignment
DECIPHER
DNA barcode mapping against full COX1 sequences
IQ
Tree estimation with molecular model for each codon position in IQ-TREE2
R
Distance from COX1 consensus
Length in bp
Amino acid translation
# of ambiguities
# of incomplete codons
NCBI
BLASTn of suspiciously distant sequences from COX1 consensus and those nested in incorrect genera
R
Statistics and plots per year of publication and sequencing technology

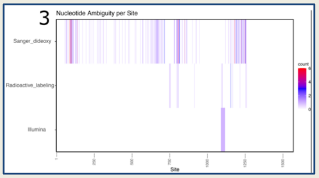
1



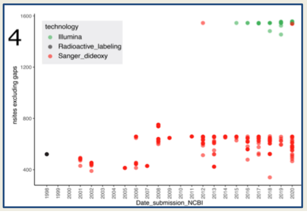
2



3



4



Results and Discussion

Our research provided us with a general overview of problematic sequences submitted to the GenBank as a result of misidentification. Even though most sequences were of good quality, few were problematic. For the context of better understanding our results, we referred to some general guidelines that suggest for barcode researchers to upload trace files (electropherogram trace files) to BOLD (Barcode of Life Data) in association with the reference (submission) number in the GenBank to better monitor QC of reference sequences used as barcodes. For instance, trace files are necessary for records to qualify for barcode status and also to provide quality control for sequences. These standards (used on animal COX1 sequences) include a minimum sequence length of 500bp, less than 1% ambiguous bases, the presence of two trace files, a minimum of low trace quality status, and the presence of a country specification in the record as set out by the CBOL [6]. However, of the 1248 sequences that we collected, 12.7%, or 158 of them, had a length less than 500bp and 5 sequences had more than 1% ambiguous bases. Another database, iBOL or International Barcode of Life, generates a phylogenetic tree and marks any outgroup for review as part of their QC. We found that Sanger sequencing is more commonly used than Illumina due to the fact that it is much cheaper. In research institutions where the cost of research is high (e.g., many developing countries) using Illumina might be expensive for many samples as well as the costs of storage of collected DNA samples [1]. However, the more expensive and detailed technology of Illumina is proved to be superior since it produces fewer ambiguities overall and most of its data fall within the consensus sequence as seen on Figs. 1-5.

Our findings are key because they support that the large amount of data generated by many researchers contributes to the variability of the quality of DNA barcodes. This affects the reliability of such reference sequences, its role to catalog and identify biodiversity, and conservation efforts. We suggest that all researchers determine the level of QC before using any DNA barcode for species identification. Likewise, it is important to assure what we have now about identified species is correct and reliable.

While our research focused on shortcomings of COX1 sequences used for DNA barcoding of poison frogs, it would be interesting to sees studies beyond this group as well as how individual variability affects species delimitation –the process of determining which groups of individual organisms constitute different populations of a single species and which constitute different species–which has been used to be one of the crucial shortcomings of DNA barcoding [7]. Lastly, a future study might suggest if Illumina or other NGS technology might finally overcome the more simple and less informative Sanger sequencing of ~500bp piece of mtDNA for cataloging diversity on the planet.

Technology table:

	Sanger	Radiolabelling	Illumina
# of entries	1157	10	81
Mean sequence length	607.35	520.90	1544.41
# of sites with ambiguities	231	20	28
# of incomplete codons	1127	10	0
Mean # of codons	201.86	171.10	514.68

Errors Table (by BLASTn & R):
Wrong Organism:

Submitted organism on NCBI	Year; technology submitted; length bp	BLASTn results	% Identical	On tree	Gap percent distance per total length
Allobates brunneus DQ502910.1	2006; Sanger; 658 bp	*Polistes dominula* (european insect) MK091919.1	99.70%	Very far	31.76% (highest and far from consensus); 0 amb
Allobates femoralis DQ502733.1	2006; Sanger; 658 bp	*Atherina hepsetus* (Mediterranean sand smelt-fish) GQ352685.1	99.39%	Very far	21.58% (not that far from consensus); 0 amb
Allobates niputidea KP149233.1	2014; Sanger, 658 bp	*Fundulus blairae* (fish) JN026619.1	82.21%	Very far	16.17% (not that far from consensus); 0 amb
Hyloxalus elachyhistus DQ502723.1	2006; Sanger; 658 bp	*Plethodon cinereus* (salamander) MG421642	98.48%	Very far	23.10% (not that far from consensus); 0 amb
Hyloxalus elachyhistus DQ502722.1	2006; Sanger; 658 bp	*Plethodon cinereus* (salamander) MG421642	98.48%	Very far	23.10% (not that far from consensus); 0 amb

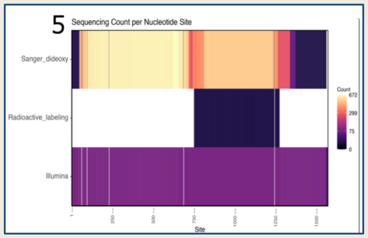
Wrong Genera & Species:

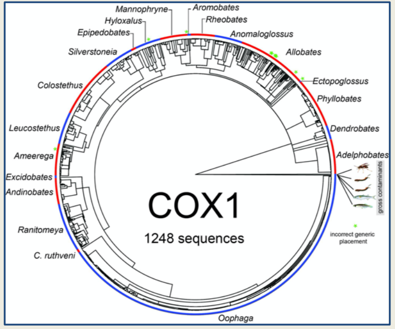
Submitted organism on NCBI	Year; technology submitted; length bp	BLASTn Results	% Identical
Allobates caeruleodactylus DQ502814.1	2006; Sanger; 658 bp	*Dendrobates auratus* DQ502751.1	100%
Allobates caeruleodactylus DQ502899.1	2006; Sanger; 658 bp	*Dendrobates auratus* DQ502751.1	100%
Allobates undulatus DQ502755.1	2006; Sanger, 658 bp	*Colostethus sp. Cuyabeno* DQ502897.1	99.54%
Allobates zaparo DQ502732.1	2006; Sanger; 658 bp	*Ameerega parvula* MW042032.1	95.14%
Aromobates saltuensis DQ502935.1	2006; Sanger; 658 bp	*Colostethus sp. Magdalena* DQ502934.1	100%
Andinobates claudiae DQ502754.1	2006; Sanger; 658 bp	*Colostethus sp. Cuyabeno* DQ502888.1	99.85%
Dendrobates auratus DQ502751.1	2006; Sanger; 658 bp	*Allobates caeruleodactylus* DQ502887.1	99.85%
Hyloxalus sylvaticus DQ502721.1	2006; Sanger; 658 bp	*Allobates femoralis* DQ502716.1	98.48%

Same Genera, Different Species

Mannophryne collaris MT627188.1	2020; Illumina; 15,434 bp	*Mannophryne lamarcai* EU380897.1	98.40%
Ameerega picta KU494334.1	2006; Sanger; 658 bp	*Ameerega flavopicta* DQ502831.1	90.25%
Colostethus imbricolus MF614309.1	2017; Sanger; 658 bp	*Colostethus panamansis* DQ502871.1	91.95%
Colostethus imbricolus DQ502881.1	2006; Sanger; 658 bp	*Colostethus panamansis* DQ502875.1	97.42%
Colostethus panamansis DQ502875.1	2006; Sanger; 658 bp	*Colostethus imbricolus* DQ502881.1	97.42%
Oophaga histrionica KX574024.1	2016; Sanger; 618 bp	*Oophaga sylvatica* KX574040.1	100%
Oophaga lehmanni MH589839.1	2018; Sanger; 523 bp	*Oophaga anchicayensis* MH589798.1	100%
Oophaga lehmanni MH589836.1	2018; Sanger; 523 bp	*Oophaga anchicayensis* MH589798.1	100%
Oophaga lehmanni MH589815.1	2018; Sanger; 523 bp	*Oophaga anchicayensis* MH589800.1	99.43%
Oophaga lehmanni MH589811.1	2018; Sanger; 523 bp	*Oophaga anchicayensis* MH589798.1	100%
Oophaga lehmanni MH589809.1	2018; Sanger; 523 bp	*Oophaga anchicayensis* MH589800.1	99.81%
Oophaga lehmanni KF582755.1	2013; Sanger; 423 bp	*Oophaga histrionica* KF582702.1	100%
Oophaga lehmanni KF582754.1	2013; Sanger; 423 bp	*Oophaga histrionica* KF582666.1	100%
Oophaga occultator MH589859.1	2018; Sanger; 523 bp	*Oophaga sylvatica* KX574029.1	100%
Oophaga occultator MH589860.1	2018; Sanger; 523 bp	*Oophaga sylvatica* KX574033.1	100%
Oophaga solanensis MH589897.1	2018; Sanger; 523 bp	*Oophaga histrionica* KX574024.1	100%
Oophaga solanensis MH589896.1	2018; Sanger; 523 bp	*Oophaga histrionica* KX74024.1	100%
Oophaga solanensis MH589895.1	2018; Sanger; 523 bp	*Oophaga histrionica* KX574024.1	100%
Oophaga solanensis MH589877.1	2018; Sanger; 523 bp	*Oophaga histrionica* KX574024.1	100%
Oophaga solanensis MH589875.1	2018; Sanger; 523 bp	*Oophaga histrionica* KX574024.1	100%
Oophaga solanensis MH589874.1	2018; Sanger; 523 bp	*Oophaga histrionica* KX574024.1	100%
Ranitomeya reticulata DQ502827.1	2006; Sanger; 658 bp	*Ranitomeya ventrimaculata* DQ502792.1	91.19%
Ranitomeya ventrimaculata AF412432.1	2001; Sanger; 487 bp	*Ranitomeya imitator* AF412428.1	99.39%
Ranitomeya ventrimaculata DQ502792.1	2006; Sanger; 658 bp	*Ranitomeya reticulata* DQ502827.1	91.19%

5





References

[1]Borisenko, A. V., Sones, J. E., & Hebert, P. D. N. (2009). The front-end logistics of DNA barcoding: Challenges and prospects. *Molecular Ecology Resources*, *9*, 27–34. https://doi.org/10.1111/j.1755-0998.2009.02629.x

[2]*CBOL | iBOL*. (2019, September 13). Retrieved January 13, 2021, from https://www.ibol.org/phase1/cbol/ Introduction to the Fundamentals of Time Series Data and Analysis. *Aptech*. https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/

[3]*DNA and RNA Labeling | Radiolabeled Nucleotides*. (n.d.). PerkinElmer. Retrieved May 19, 2021, from https://www.perkinelmer.com/lab-products-and-services/application-support-knowledgebase/radiometric/dna-rna-labeling.html

[4]*NGS vs. Sanger Sequencing*. (n.d.). Retrieved January 13, 2021, from https://www.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html

[5]Pennisi, E. (2019, June 6). *$180 million DNA 'barcode' project aims to discover 2 million new species*. Science | AAAS. https://www.sciencemag.org/news/2019/06/180-million-dna-barcode-project-aims-discover-2-million-new-species

[6]Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited—Do identification errors arise in the lab or in the sequence libraries? *PloS One*, *15*(4), e0231814. https://doi.org/10.1371/journal.pone.0231814

[7]Rannala, B., & Yang, Z. (2020). *Species Delimitation*. 19.

[8]Vilgalys, R. (2003). Taxonomic misidentification in public DNA databases. *New Phytologist*, *160*(1), 4–5. https://doi.org/10.1046/j.1469-8137.2003.00894.x

Acknowledgements

We would like to thank our mentor, Juan C. Santos, and his student, Juan David Carvajal Castro, from St. John's University for their guidance and advice throughout the course of our project. We would also like to thank the Cold Spring Harbor Laboratory DNA Learning Center and Urban Barcoding Research Program for providing us the enriching opportunity and resources to conduct scientific research. We would also like to thank the NCBI for making available the samples used for our project.