Comparisons of 4.7 million mtDNA sequences show GenBank is reliable for animal IDs

October 21 2019

For Matt Leray (on the left in this photo) , postdoctoral fellow at the Smithsonian Tropical Research Institute in Panama, being able to rely on GenBank to accurately identify coral reef organisms is essential for studies of reef health. Credit: Sean Mattson, STRI

Did a murderer walk through the room? Did a shark just swim by? Is this a poisonous mushroom? Which reef species are lost when the coral dies? These questions can potentially be answered quickly and cheaply based on tiny samples of DNA found in the environment. But identifying DNA requires a trustworthy library of previously identified DNA sequences for comparison. Smithsonian scientists and their colleagues analyzed more than 4.7 million animal DNA sequences from GenBank, the most commonly used tool for this purpose, and discovered that animal identification errors are surprisingly rare—but sometimes quite funny.

"We wanted to use GenBank to identify DNA from ocean water samples as we evaluate the health of coral reefs and other marine ecosystems, but we were concerned by reports questioning the accuracy of the data there," said Matthieu Leray, post-doctoral fellow at the Smithsonian Tropical Research Institute (STRI). "In our sequence comparisons we found fewer errors than people had predicted, which is a very good news, because monitoring programs and conservation efforts increasingly rely on analysis of environmental DNA."

The reliability of data in GenBank, the virtual library maintained by the U.S. National Center for Biotechnology Information at the National Institutes of Health, where geneticists deposit DNA sequences from all living creatures, has been questioned in the past. An article entitled "Can You Bank on GenBank?" published in *Trends in Ecology and Evolution* in 2003, referred to studies showing that half of human mitochondrial DNA sequences contained errors, and that there were significant differences in sequences deposited for fruit flies. Another article reported that 12 of 51 species of the highly poisonous fungus, Amanita, were misidentified.

"We assumed that we would find lots of errors when we started the study," said Nancy Knowlton, of the Smithsonian National Museum of Natural History (NMNH).

"Some people think that GenBank is just a data dump," said Leray. "No one checks to see if the data are entered correctly. Researchers just upload their sequence data and they don't have to deposit a specimen anywhere in particular, so if there is a question, there may be no way to go back to the source to find out if a sequence is correct. We needed to be sure that GenBank was a good tool to use to identify marine organisms in our samples, so we decided to find out."

With colleagues from Academia Sinica and The George Washington University, Leray and Knowlton estimated the proportion of sequences with incorrect genus, family, order, class and phylum names. Overall, less than 1 percent of the sequences were mislabeled. They identified certain groups of animals that are particularly problematic and some of the potential sources of error like mislabeling and contamination from humans, rodents, lab animals, food, mosquitos and pets like dogs and cats.

"For example, when you enter sequence data, at some point there is a drop-down menu giving choices of different species. Some people evidently just clicked on the wrong species, the one above or below the name of the species they were trying to enter. This part of the process could be fixed to lower the error rate even further."

Direct DNA identification is a fast, low cost way to answer many questions about the environment, and GenBank is a reliable tool to use to identify the source of the DNA. The authors concluded: "Our encouraging results suggest that the rapid uptake of DNA-based approaches is supported by a bioinformatic infrastructure capable of assessing both the losses to biodiversity caused by global change and the

effectiveness of conservation efforts aimed at slowing or reversing these losses."

The Smithsonian Tropical Research Institute, headquartered in Panama City, Panama, is a unit of the Smithsonian Institution. The institute furthers the understanding of tropical biodiversity and its importance to human welfare, trains students to conduct research in the tropics and promotes conservation by increasing public awareness of the beauty and importance of tropical ecosystems.

The study is published in *Proceedings of the National Academy of Sciences*.

More information: Matthieu Leray el al., "GenBank is a reliable resource for 21st century biodiversity research," *PNAS* (2019). www.pnas.org/cgi/doi/10.1073/pnas.1911714116

Provided by Smithsonian Tropical Research Institute