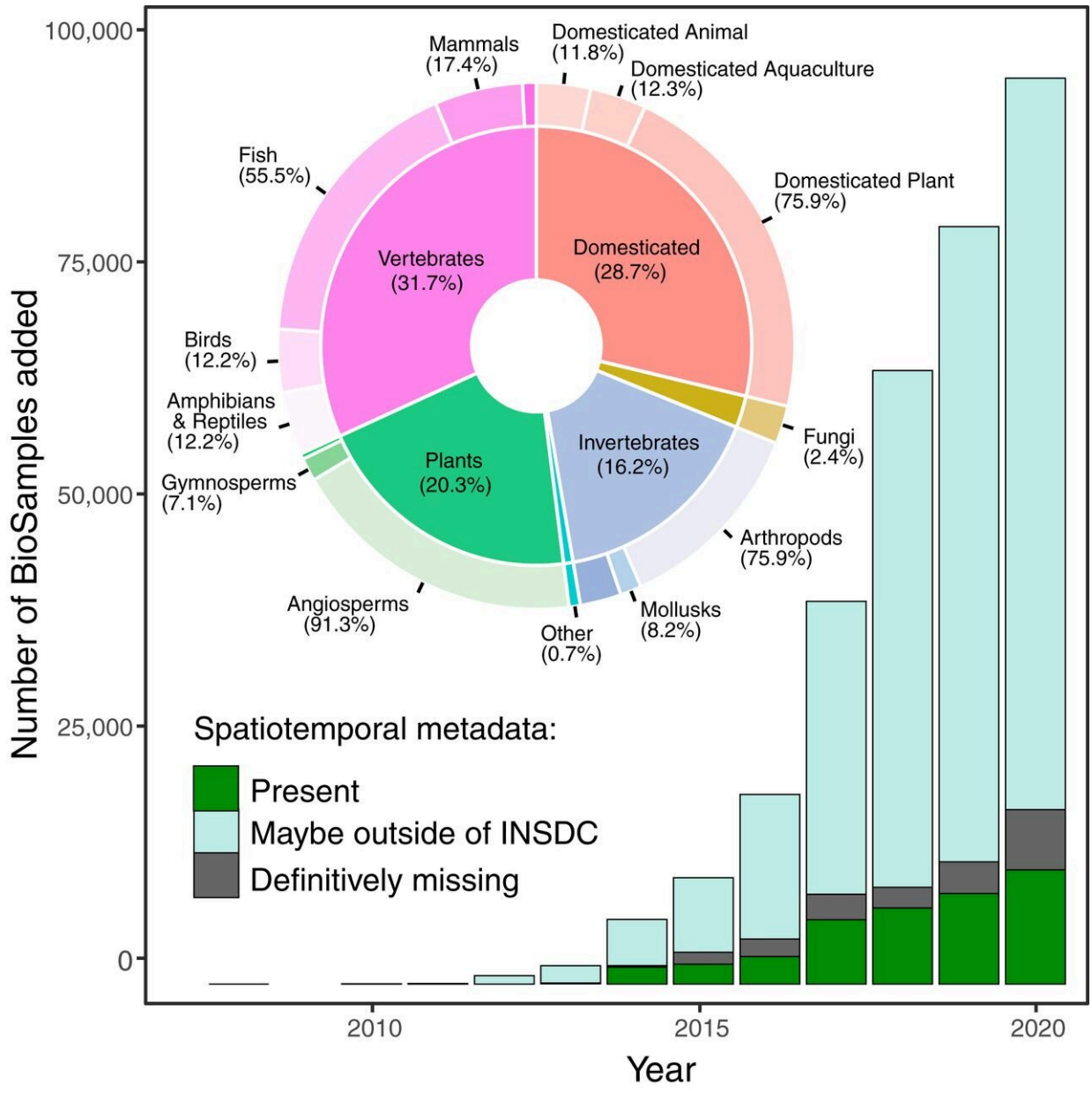
Biodiversity needs better data archiving

August 27 2021



Genomic-level sequence data are being added to the INSDC at an exponential

rate across eukaryotic taxa. Colors represent the status of spatiotemporal metadata (latitude/longitude and collection year) for each individual (BioSample, n = 327,577, see SI Appendix, Appendices S1–S3). (Inset) Taxonomic breakdown of BioSamples. Percentages in outer rings sum to corresponding inner-ring totals. Unlabeled inner-ring slices correspond to “other” for the outer-ring taxa. Credit: DOI: 10.1073/pnas.2107934118

Missing metadata—data that provide information about other data—might not sound like a big deal, but it's a costly problem that's hindering humanity's plans to protect the planet's biodiversity. An international team of researchers has audited the largest global repository for storing genetic sequence data to see if the entries included basic metadata needed to make them useful for monitoring genetic diversity. They found that more than half of the datasets were missing that metadata.

"This work is an important wake-up call to evolutionary biologists, molecular ecologists and the biodiversity community at large that while we are doing a great job of archiving genetic sequence data, we need to greatly improve the metadata attached to them if we want to be able to monitor the evolutionary health of natural populations in the face of accelerating climate change," said Eric Crandall, senior author on the study and an assistant research professor of biology at Penn State.

According to the researchers, every individual plant or animal has thousands of genes in its genome that help it to adapt and survive in its unique environment. Organisms with lots of genetic diversity are very adaptable, while those that lack genetic diversity are more vulnerable to changing conditions, such as warming and drying environments, the appearance of an invasive species and poor health resulting from inbreeding.

"Genetic diversity affects the health of species, which in turn affects the health of ecosystems," said Rachel Toczydlowski, a postdoctoral researcher at Michigan State University (MSU), and lead author on the paper. "Having diversity across all these levels is critical for a healthy planet."

Researchers, therefore, want to know how much genetic diversity is in a given place at a given time to understand the health of those organisms and their environment. Tracking changes in genetic diversity over time would also let ecologists forecast how ecosystems will fare in the future and prepare accordingly. Conservationists, for example, could use the information to determine which organisms would be best suited to launch successful restoration efforts in disrupted ecosystems. But that goal can be met only if the available data are complete.

To get an idea of how much metadata, such as when and where a sample was collected, was missing, the team surveyed thousands of data sets from the International Nucleotide Sequence Database Collection—the largest data repository of its kind—representing more than 325,000 individual organisms from nearly 17,000 different species. The researchers found that 86% of these samples were missing important metadata.

The findings appeared Aug. 16 in the journal *Proceedings of the National Academy of Sciences*.

"Researchers spend incredible amounts of time and money to generate genomic sequence data, and these data can provide novel insights into basically every field of biology, from conservation to ecology to behavior to evolution," said Gideon Bradburd, an assistant professor of integrative biology at MSU. "But, if the context of the data—the location and time at which individuals are sampled—is dissociated from these genetic resources, they become much less useful, especially for

conservation monitoring."

There's the time that's spent obtaining permits to collect samples, then traveling to field sites and then actually tracking down the samples in the wild. And all of that is before researchers return to the lab to extract the DNA they want to sequence, which costs about $50 per sample.

That may not sound like much, but when added up over all the samples from this study that researchers cannot reuse in future analyses because of missing metadata, the sum is in the tens of millions of dollars.

"Almost every photo that people take with their smartphones contains metadata that describes the time and place the photo was taken, so it comes as a surprise that expensive genetic sequence data do not have similar information attached," said Crandall. "The system for providing these metadata is difficult to learn quickly, and currently there just aren't enough incentives for researchers to spend their valuable time on this."

There is good news, though. Undergraduate and graduate students on the team were able to find a good chunk of that missing metadata published elsewhere in the scientific literature.

"They were able to resurrect about 20,000 individual samples that couldn't have been used in future conservation monitoring otherwise," Toczydlowski said. And the fact that these students were able to contribute is, in itself, a silver lining.

When the pandemic struck, the team started discussing what they should do with grant money that was set to expire and had been earmarked for attending conferences. With travel and gatherings off the table, the team pivoted and put the money toward enlisting graduate students to track down missing metadata about when, where and how samples used to generate genetic sequence data were collected. After reading through

associated scientific publications and contacting their authors, the students were still unable to locate the missing metadata for 67% of the datasets they worked on.

"Raw genomic data in public repositories are inimitable historical resources—analogous to natural history museums—for the most fundamental level of biodiversity," said Crandall. "However, reuse of genomic sequences also minimally requires information about the spatial and temporal context of the sampled organism. Without appropriate archival practices that maintain links between genotypes, place, and time, these growing genomic resources will have limited real-world impact on genetic diversity surveillance."

More information: Rachel H. Toczydlowski et al, Poor data stewardship will hinder global genetic diversity surveillance, *Proceedings of the National Academy of Sciences* (2021). DOI: 10.1073/pnas.2107934118

Provided by Pennsylvania State University