Briefings in Functional Genomics, 19(5–6), 2020, 337–338

doi: 10.1093/bfgp/elaa016
Advance Access Publication Date: 4 September 2020
Letter to the Editor

Depositing annotated sequences in GenBank: there needs to be a better way

David Roy Smith

Corresponding author: D.R. Smith, Department of Biology, University of Western Ontario, 1151 Richmond Street, London N6A 5B7, Canada.
Tel.: (519) 661 2111; E-mail: dsmit242@uwo.ca

Abstract

Submitting sequences to the National Center for Biotechnology Information (NCBI) is an integral part of research and the publication process for many disciplines within the life sciences, and it will only become more important as sequencing technologies continue to improve. Here, I argue that the available infrastructure and resources for uploading data to NCBI—especially the associated annotations of eukaryotic genomes—are inefficient, hard to use and sometimes just plain bad. This, in turn, is causing some researchers to forgo annotations entirely in their submissions. The time is overdue for the development of sophisticated, user-friendly software for depositing annotated sequences in GenBank.

Key words: Bankit; GenBank; NCBI; sequin; tbl2asn

Main text

If I were to stroll through the biology department where I work and ask people at random if they or someone from their lab groups have used next-generation sequencing data in the past year, most would answer yes. Indeed, the use of molecular sequencing technologies has become mainstream across the life sciences and beyond. Consequently, more and more researchers from diverse fields are depositing data in the National Center for Biotechnology Information (NCBI) and its allied repositories, including the DNA Data Bank of Japan and the European Bioinformatics Institute. This is good. But I would argue that the available infrastructure and resources for uploading these data, especially eukaryotic genome annotations, are inefficient, hard to use and sometimes just plain bad.

My first experience with submitting sequences to GenBank came in 2005 as part of an undergraduate thesis project involving the assembly of the scallop mitochondrial DNA (mtDNA). To get this genome sequence into GenBank, I used a standalone NCBI-developed software called Sequin. It was slow, frequently crashed, and had a black-and-white aesthetic akin to late-1980s Mac applications. But Sequin did have a graphical user interface (GUI), allowing a bioinformatics-newbie, like myself, to use it. Over a 5-h period, I entered into Sequin the coordinates of the 40 or so genes from my scallop mitogenome sequence. This process was slower than you might think as every coding region needed to be inputted one a time (*cox1*, *cob*, ...) and many contained multiple annotations (e.g. gene + tRNA + anticodon sequence).

During my PhD and postdoc, I sequenced dozens of more organelle genomes, all of which I deposited to NCBI using Sequin. Some of these entries took a few days to prepare, such as large chloroplast genomes with hundreds of genes, but because I was only ever submitting one to a few genomes at a time, the task was never insurmountable, just tedious. Of course, I learned a few tricks to streamline my submissions (e.g. adding source modifiers to the fasta file), but for the most part, I was still entering the coordinates of each annotation by hand. Computer-savvy readers might roll their eyes at this point, knowing that there are easier ways to input genome annotations, such as uploading GFF3 or GTF files. But like many biologists, my computer-programing and command-line skills are rudimentary, forcing me to point and click my way around most bioinformatics problems.

I continued using Sequin until 2019 when NCBI phased it out. Today, if someone wants to submit an organelle genome (or most other types of sequences) to GenBank they must use Bankit, an online GUI submission tool, or tbl2asn, a complex command-line-based program. (There is also an online tool called the Submission Portal for depositing certain barcoding and viral genome sequences.) I have tried to use tbl2asn on multiple occasions, hoping to modernize my submission skills, but still struggle with the basics of this counter-intuitive software, even

David Roy Smith is an associate professor of biology at the University of Western Ontario. He studied genome evolution of eukaryotic microbes and can be found online at www.arrogantgenome.com and @arrogantgenome.

after slogging through the online instruction manual. Thus, I am stuck with Bankit for depositing my organelle genome data. It is essentially a pared-down version of Sequin, but in no ways has it made preparing my submissions faster or easier; if anything, Bankit has slowed me down because, unlike Sequin, it requires an internet connection. What's more, major improvements in sequencing technologies have greatly increased the frequency and volume of organelle genome data that I send to GenBank. For example, my collaborators and I recently described 71 yeast mitochondrial genomes [1]. It took the lead author more than a week of full-time work to prepare these sequences for NCBI using Bankit. It did not help that some of the mtDNAs contained large numbers of introns, resulting in dozens of intron-exon annotations for some genes.

One way to streamline the submission process is simply to deposit genome sequences without any annotations, which is permitted under NCBI regulations. In the case of the 71 yeast mitogenomes, such an approach would have reduced the submission time from days to hours. But there is one obvious drawback to this tactic: anyone accessing the data would be left staring at blank sequences. In many instances, the annotations can be as important as the sequences themselves. The research paper describing the yeast mitochondrial DNAs, for instance, focused on the abundance and location of introns, meaning detailed and publicly available data on the exon-intron boundaries were crucial to the study [1]. Nevertheless, many authors are choosing to forgo annotations entirely. A recent high-profile paper in *Nature Microbiology* presented mitochondrial genome sequences from a range of under-sampled eukaryotic lineages [2]. But of the 15 complete mtDNAs deposited in GenBank as part of this study (accession numbers MK188935-MK188947 and MN082144-MN082145), only two contained annotations, despite that the paper focused largely on variation in gene content [2]. I could provide many other examples of published genome data without annotations, both from my own field of organelle genomics as well as from other fields, such as bacterial genomics.

Who can blame these authors for streamlining their submissions when adding detailed annotations can take days or even weeks? Although, at the end of the day, the responsibility of providing well-annotated data falls on those who generated those data. The time is overdue for the development of sophisticated, user-friendly software for adding annotated sequences into GenBank. Even a commercial option would be welcomed, but it is not ideal. The past decade has seen major improvements in the design, speed and capabilities of user-friendly bioinformatics software suites, both open-source and commercial [3, 4]. I use these types of programs nearly everyday in my research, and one of my favorite features is the ability to search and import data directly from GenBank via the software. For example, using the commercial program Geneious (Biomatters Ltd.), I can quickly access all the available mtDNAs in NCBI, download these data, and then extract specific annotations. However, to the best of my knowledge, a seamless and straightforward system to do the opposite—upload annotated sequences from a user-friendly bioinformatics platform to GenBank—is still unavailable. To be fair, Geneious has developed a GenBank submission plugin that connects to Bankit and allows users to upload sequences to NCBI. But in my experience, this approach is just as tedious as using Bankit directly because the annotation tools in Geneious are not designed to easily match the NCBI Feature Key, often resulting in numerous errors and warnings during the submission process. And, again, this plugin is only available to users who have purchased the software.

As one reviewer of this letter pointed out, there already exists a fast and sophisticated solution for annotating and preparing bacterial genomes for GenBank submissions called the Prokaryotic Genome Annotation Pipeline (PGAP). My experiences with PGAP are limited, and I will note that the stand-alone package of this NCBI-developed software, which is freely available from GitHub, is not user-friendly, requiring Linux or a compatible container technology, a Common Workflow Language, and about 30 GB of supplemental data. A more straightforward approach is to have NCBI apply PGAP to your data, which is currently an available service for GenBank submitters. The reviewer said it best: 'I have already submitted several hundred bacterial genomes to GenBank in one submission in less than an hour.' Hopefully, NCBI will develop PGAP-like services for eukaryotic genomes. But if PGAP is so efficient why are there still tens of thousands of unannotated prokaryotic genomes in GenBank?

We live in a world of increasingly seamless integration between our technological devices. Whether I am having a Zoom meeting, submitting a manuscript to a journal or developing a lecture with PowerPoint, I have come to expect a certain level of quality, usability, and cross-platform support. Given the billions of dollars being spent and invested in genome sequencing technologies, why is there not an efficient means for submitting annotated data to GenBank? When researchers choose to add annotations to their submissions, we all win. For that to happen, we need easy-to-use software solutions, including accessible stand-alone tools for annotating all types of genomes. As it stands, neither Bankit nor tbl2asn fit the bill.

Key points

- Submitting annotated genome sequences to GenBank can be time consuming and tedious, especially for those who are not bioinformatics experts.
- In order to expedite submissions, some researchers are uploading unannotated genomes to GenBank, which can hinder or slow future work.
- To overcome these issues, there needs to be improved, user-friendly bioinformatics software catered to efficient data deposition.

Funding

This work was supported by a Discovery Grant to DRS from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. Lee DK, Hsiang T, Lachance MA, Smith DR. Do *Metschnikowia* yeasts have the strangest mitochondrial genomes of all fungi? *Curr Biol* 2020;**30**:R800–R801.
2. Wideman JG, Monier A, Rodríguez-Martínez R, *et al*. Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists. *Nat Microbiol* 2020;**5**:154–165.
3. Smith DR. Buying in to bioinformatics: an introduction to commercial sequence analysis software. *Brief Bioinform* 2015; **16**:700–709.
4. Perkel JM. Democratizing bioinformatics. *Nature* 2017;**543**: 137–138.