Opinion

EMBO reports

A short guide to genetic data mining

David R Smith*

With social distancing measures still in place throughout the world and the second wave of COVID-19 building up, there has never been a better time for data mining. Most universities and academic research labs, including my own, are operating at reduced capacity and efficiency, if at all. In many cases, experiments and fieldwork have been postponed, graduate students' timelines have been extended by months or years, and remaining grant funds have been stretched to their absolute limit. This is all the more reason to hunker down in your home office and take advantage of the prodigious amounts of free and easily accessible online data, especially in genomics. Trust me, you do not have to be a computer whiz to do this, and the end results can yield exciting new findings.

For nearly two decades, I have primarily made my way in science—from graduate student to postdoc to PI—by siphoning off and parsing together other people's data. I have done this with no programming skills and only a rudimentary ability to work in a Unix environment. Admittedly, my entry into the domain of data mining was a rocky one. I began my PhD PCRing and sequencing my way through mitochondrial genomes. In between experiments, I would sit in my cubicle at the back wall, surveying online material about my organisms of interest, green algae. One day, I stumbled upon a data bank at the National Center for Biotechnology Information (NCBI) called the Trace Archive, which housed Sanger sequencing reads from various published and unpublished genome projects. The interface of the Trace of Archive allowed users to easily blast against the raw sequencing information for individual species, from a diversity of lineages, and download the hits.

Soon I was spending all my spare time doing exactly that—blasting mitochondrial and chloroplast genes against the database to stockpile organelle-derived reads, which I assembled into contigs on my laptop. Once I had put together a small piece of organelle DNA, I used it to fish out more and more reads until the entire genome was assembled. Note: I could have easily used the same technique for mining other kinds of genomes, such as those from viruses or prokaryotes. In short time, I had amassed organelle genomes from a variety of species, none of which had been explored before. Unbeknownst to my supervisor, I deposited the sequences in GenBank and then submitted a paper describing one of them to a journal with myself as the sole author. A few weeks later, I got an email from GenBank saying they had received a complaint about my entries and were deleting them, thus, forcing me to withdraw my manuscript. I had naively ignored Rule #1 of data mining: If you use other researchers' data to assemble a genome, gene, or even a small piece of noncoding DNA, you must submit that sequence to GenBank as a Third Party Annotation (TPA), describing in detail the data used to create it. Many GenBank authors ignore this rule (preparing a TPA can be onerous), and most of these errors still go unreported.

Realizing that it was likely the primary authors of the reads I used who complained to GenBank, I reached out to them describing my findings, which I hoped to resubmit after revising my entries as TPAs. This was their reply: "Dear Mr. Smith. Using data generated by others, publicly available or not, and publishing on it would not be an approach that I would advise. Working on a project with a larger group of collaborators is now considered a better measure that one will be able to succeed in genomic science than writing and publishing a paper by oneself. You could make valuable connections to others in the community by taking this approach rather than potentially alienating them. I recommend that you wait to resubmit the manuscript pending further discussions of attribution, acknowledgments, or co-authorship". This takes us to Rule #2: Whether you are a student or a PI, it is usually better to collaborate with the creators of the data you are mining rather than go behind their backs.

In only a short while, I had formally withdrawn a research paper, been reprimanded by NCBI, and angered a team of prominent scientists. When I eventually told my supervisor what had happened, he laughed out loud and said, "Smitty, you are off to a great start. Keep it up". Indeed, I did not let a few setbacks deter me. I smoothed things over with GenBank and the PIs and continued to assemble organelle DNAs from the Trace Archive. Ultimately, about half of my thesis was made up of mined data, and I am still collaborating with some of those "angry" scientists, some of whom have gone on to become master data miners themselves.

The Trace Archive has been supplanted by the Sequence Read Archive (SRA), which houses all types of sequencing data—from Illumina to PacBio to Oxford Nanopore—from all types of biological systems. Like many PIs studying genomics, I have an ongoing obsession with the SRA, one that has only grown stronger during the lockdown. I check it almost daily and maintain detailed lists of species and lineages for which I am hoping data become available. When they do, I pounce, downloading the

Department of Biology, University of Western Ontario, London, ON, Canada
*Corresponding author. Tel: +1 519 661 2111, ext. 86482; E-mail: dsmit242@uwo.ca
DOI 10.15252/embr.202051845 | EMBO Reports (2020) e51845

raw sequences and staying up late into the night analyzing them. Sometimes, it is a single piece of missing data that I'm searching for, something I need for completing a larger project or solving a nagging riddle, such as a gene sequence from a specific organism for a phylogeny. Other times, I'm looking to glean detailed information on genome architecture. Whatever it is, I have discovered that it is a good rule—call it Rule #3—to have clear, well-defined goals. That is not to say that sometimes I do not randomly explore GenBank; it is just that these undirected wanderings do not often bear fruit.

Navigating the SRA and NCBI's other databases is not as straightforward as one might expect. My preferred method is to use the Taxonomy Browser. In brief, I enter my lineage of interest—for instance, Chlamydomonadales, but users can search any group from across the tree of life. This takes me to an index of known species for that group (there are >750 chlamydomonadaleans). I select the databases I am interested in (there are more than 25 to choose from), such as the SRA, hit "Go", and, voila, I can see all the species and strains for which datasets are available. For instance, on August 12, 2020, Illumina HiSeq 400 reads were uploaded for the green alga *Volvox aureus*. This is great. I have an ongoing interest in the chloroplast genomes of *Volvox* species, and *V. aureus* would make an excellent addition to my growing dataset. All I need to do is collect the reads and attempt to assemble the chloroplast DNA. But other users might want to assemble a plasmid from a bacterium, a viral genome, or a specific region of nuclear DNA, for example—or perhaps explore the plethora of bisulfite sequencing data in the SRA to better understand methylation patterns.

Rule #4: Whatever type of molecular sequence data you are mining, do not underestimate the need for sophisticated software for downstream analyses. Not being a computer expert, I favor programs with a graphical user interface (GUI) and intuitive design, but many experienced researchers will be comfortable using command-line-based ones, most of which are open-source. Some of the GUI bioinformatics software suites I use are free, and others are commercial and require a significant financial investment. In the end, it does not matter what programs you use, provided they do what you want them to do.

One of the most important things I have learned about data mining is to persevere. It is unlikely I will sit down at my computer this evening and uncover a cache of unusual genome sequences resulting in a paper. It is more likely that, over time and with significant effort, I will collect small crumbs of information, which add up to a valuable and impactful contribution to my field. I have also learned—and this is the fifth and final rule—to look in unexpected places. For instance, when searching for reads to assemble genomes, I used to avoid RNA sequencing (RNA-seq) data, assuming they would give me only fragmented, incomplete assemblies. One day, however, I surveyed an assortment of transcriptomic experiments and I discovered that I could reconstruct entire organelle genomes from RNA-seq alone, including ones with long intergenic regions. RNA-seq reads can also be used for assembling entire prokaryotic genomes and large segments of nuclear chromosomes. This discovery had a significant impact on my research because for green algae, and various other lineages, there are more RNA-seq datasets in NCBI than DNA-seq ones. A colleague of mine recently found that bacterial environmental sequencing experiments, especially ones targeting cyanobacteria, are an untapped reservoir for surveying eukaryotic diversity because the bacterial primers often amplify chloroplast DNA. No matter your organism or system of study, there is likely a data bank out there that is ideal for your research, but you might have to think out of the box to find it.

Despite the pandemic, we are lucky to live in a time when we can access so much information from the safety of our living rooms, not to mention the increasing power and capabilities of entry-level computers. I hope my own journey and these five simple rules serve you well and provide an entry point into the world of genetic databases, whether you are student or a weathered researcher. The sheer volume of available data can be intimidating, but do not be fooled into thinking that a worthwhile project needs to be equally massive in scope. Yes, big data is the catch phrase of the day, but sometimes even small discoveries from clever online explorations can yield interesting insights. Happy mining.

Acknowledgements

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

Conflict of interest

The author declares that he has no conflict of interest.

