Science & Society

EMBO reports

Don't just dump your data and run

Authors should submit as much experimental information as possible when uploading sequence data

Matheus Sanitá Lima & David Roy Smith

If you have, in any way, been involved with genetic research over the past 10 years, then you have likely heard of the Sequence Read Archive (SRA), which is jointly housed at the National Center for Biotechnology (NCBI), the DNA Data Bank of Japan (DDBJ), and the European Bioinformatics Institute (EBI). And if you regularly work with genome or transcriptome sequence information, then you have probably extracted data from and/or deposited data into the SRA. For those who are unfamiliar with it, the SRA is an international public online archive for next-generation sequencing (NGS) data, which was established about a decade ago under the guidance of the International Nucleotide Sequence Database Collaboration (INSDC) [1,2]. Despite nearly being shut down in 2011 [3], it has grown at a staggering rate over the past 10 years. As of September 1, 2017, it housed over five quadrillion (10^{15}) open-access bases of NGS data, coming from thousands of different species and spanning the entire gamut of cellular and viral life. It contains DNA- and RNA-sequencing (DNA-seq and RNA-seq) reads of every kind, from bisulfite-seq to strand-specific RNA-seq to single-cell DNA-seq, and it accepts reads from every type of NGS platform, be it Illumina, Ion Torrent, or PacBio sequencing. In other words, the SRA is a crucial and central resource in the fast-paced and increasingly important domain of contemporary genetic research.

> *"... the SRA is a crucial and central resource in the fast-paced and increasingly important domain of contemporary genetic research."*

The Sequence Read Archive

The SRA can be easily accessed and searched via the NCBI (https://www.ncbi.nlm.nih.gov/sra), DDBJ (http://trace.ddbj.nig.ac.jp/dra/index_e.html), and EBI (http://www.ebi.ac.uk/ena/submit/read-submission) websites. Once there, you will find yourself at a sequencing-read superstore. With a decent Wi-Fi connection, a couple of keyword searches, and a few clicks of the trackpad, you can quickly download NGS experiments from your favorite model species, and thousands of non-model species, in anywhere from 5 minutes to a few hours, depending on the size and number of data sets you are interested in. If you are new to the SRA, one of the easiest and fastest ways to start exploring it is via the Taxonomy Database at NCBI, which contains a curated classification and nomenclature of all organisms in the data bank (https://www.ncbi.nlm.nih.gov/taxonomy). Simply enter a strain, species, or broader group name in the search bar and once you have clicked on the result tick the "SRA experiments" box at the top of the screen to see all the available projects for your organism(s) of interest. For example, if you are an algal buff and had searched the word "Chlorophyta", you would have found that there are over 3,000 different SRA experiments for green algae, including more than 1,800 for the model unicell *Chlamydomonas reinhardtii*.

Exploring the SRA might be straightforward, but getting your own NGS experiments into the archive can be complicated and tedious. This is not surprising given that an SRA submission involves the uploading of very large files and creating a summary of those files. As many bioinformaticians can attest, depositing reads into the SRA is much more time-consuming and requires many more steps than, for instance, submitting a set of annotated gene sequences to GenBank, and it also entails the onerous task of creating a BioSample and BioProject—a summary and online record of biological source material and data related to a single initiative. Perhaps this is why members of the INSDC recently went out of their way to remind scientists to submit their raw sequencing reads to the SRA [4].

Thankfully, many researchers do upload their NGS data to the SRA, partly because most journals require a database accession number as a condition of publication. But pressure from journals cannot be the only incentive as there are a large number of unpublished experiments in the SRA, some of which will likely never get published by the authors who initially generated them. Published or not, an SRA project can be a major asset and an important resource for the scientific community, provided it is properly annotated.

> *"Published or not, an SRA project can be a major asset and an important resource for the scientific community, provided it is properly annotated."*

Big data, little methods

Recently, we were mining data from the SRA to study transcription in mitochondria and chloroplasts. Specifically, we used publically available eukaryotic RNA-seq experiments to reconstruct complete or near-complete organelle genome sequences. The SRA provided us with ample data to carry out our analyses in a diversity of

Department of Biology, University of Western Ontario, London, ON, Canada. E-mail: dsmit242@uwo.ca
DOI 10.15252/embr.201745118 | Published online 27 October 2017

species, which allowed us to document the widespread occurrence of pervasive organelle transcription across the eukaryotic domain [5]. Our study on organelle transcription, which ultimately formed the bulk of an MSc thesis, reinforces the utility of the SRA for both large and small research groups (we represent the latter). Apart from the price of a computer and a commercial bioinformatics software suite—and significant time investment, of course—the research project cost us nothing. We did, however, encounter some setbacks when trying to determine the protocols used to generate the various RNA-seq data sets employed in our analysis. In short, we were confronted with an SRA annotation issue. We had used hundreds of RNA-seq experiments generated from different laboratory groups, often using very different protocols. Some of these experiments contained detailed and meticulous information on the growth conditions, RNA isolation and purification techniques, library preparation, and sequencing methods. Other experiments, unfortunately, had little or no accompanying details about how they were generated, leaving us guessing about the underlying experimental procedures.

"Well-annotated nucleotide sequence information will only help to advance science, promote data sharing and collaboration, and increase the influence and reach of your research."

When an NGS project is submitted to the SRA, it must contain certain basic details about the strain, species, or population that was used and the sequencing technologies that were employed. However, it appears to be at the author's discretion to include a summary of the methods, despite the fact that the SRA provides space for such a summary in both the "Design" and "Study" sections of the entry. A quick scan of the SRA reveals many submissions with exemplary methods. Sometimes even a concise statement describing the study can make a big difference. Take, for example, SRA accession SRX2788293, an RNA-seq experiment for the green alga *Dunaliella tertiolecta*, which includes the following under study design: "Cells grown in continuous culture at 40 μE with low dilution rate. When cells reached steady state, light intensity was increased to 400 μE. Two hours after light intensity increased, RNA was extracted using RNeasy Mini Kit (Qiagen, Germany) and was converted to cDNA library using Illumina TruSeq Stranded Total RNA LT Sample Prep Kit with Ribo-Zero Plant". But it is also easy to find SRA experiments with absolutely no specifics about how the sample was collected and prepared.

One of the referees who evaluated our meta-analysis of organelle transcription asked: "Is it possible to decipher from the protocol description in the SRA database if the data sets you used were prepared with poly-A selection? If so, please discuss the differences in RNA-seq mapping success for the experiments with and without poly-A selection". This was an excellent suggestion, but we were unable to carry out the referee's request because, as already noted, most of the SRA studies we employed contained no methods section.

One could argue that instead of relying on the SRA we could have just read the Methods and Materials from the primary research articles for the various data sets we used. But in certain cases, the SRA data we employed had not yet been published. Moreover, it would have taken a lot of time and energy to look up the individual papers for hundreds of different experiments, many of which were behind a paywall, which goes against the purpose of an open-access data bank like the SRA. In our opinion, it is much more efficient, fair, and useful to have the methods directly linked to the SRA entry. In many ways, the experiments being deposited in the SRA can be as important and impactful as the primary research papers presenting the data.

The importance of genetic database entries

In today's publish-or-perish academic landscape, one can understand why researchers would rush through the often slow and tiresome task of uploading their genetic data to an online data bank. Surely, it is the peer-reviewed papers that matter most and where our energy should be invested. However, one should not underestimate the growing significance of online archives in science—and daily life in general. A typical refereed publication employing NGS data may be read by a few hundred people and cited a couple dozen times throughout its academic lifespan. But the NGS experiment used in that work could get integrated into many different research projects and in turn have a much larger impact than the initial study. This is particularly true for data generated from large-scale sequencing initiatives, such as the 1000 Plants Genome Project [6] or the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) [7], but even a small NGS data set can have a long shelf life.

"... more and better information on methods is not only helpful for users of the SRA, but it benefits science in general if any publication of experiments contains as much information as possible."

Genome papers exemplify the growing importance of sequence repositories: They used to be widely read and represent milestones in the scientific literature, but now they have become mundane and formulaic [8]. Scientists who are truly interested in investigating a new genome sequence are arguably better served by going directly to the annotated entry in NCBI rather than by reading the primary paper, especially if it is a genome report. Similarly, a small error in a genome paper, such as the mislabeling of an annotation on a genomic map, would likely cause fewer problems and less confusion than if that mislabeling were found in the online sequence. Whether or not a genetic database entry is as impactful as a publication is beside the point. Well-annotated nucleotide sequence information will only help to advance science, promote data sharing and collaboration, and increase the influence and reach of your research.

In certain respects, SRA annotation issues touch upon the broader and ongoing debate in science about reproducibility—often referred to as the "reproducibility crisis" [9]. Whether or not this crisis is real, most scientists would agree that providing as much information as possible about their experiments greatly helps others to reproduce and build upon published results. In a recent commentary in *Nature* "A long journey to reproducible results", the authors highlight

how "improved reproducibility often comes from pinning down methods" [10]. They describe how two cancer labs spent more than a year trying to understand inconsistencies: "It took scientists working side by side on the same tumor biopsy to reveal that small differences in how they isolated cells —vigorous stirring versus prolonged gentle rocking—produced different results" [10]. In other words, more and better information on methods is not only helpful for users of the SRA, but it benefits science in general if any publication of experiments contains as much information as possible. So, do not just dump your genetic data online and run. Take the time and trouble to accurately and rigorously characterize them in whatever sequence archive you are using.

Before we start sounding too self-righteous, we should come clean and admit that the senior author of this article has submitted his fair share of data into the SRA without providing a detailed protocol for those entries. It was not until he started mining large amounts of RNA-seq data from the SRA that he finally saw the proverbial Illumina light at the end of the annotation tunnel and asked forgiveness for all of his sins. Thankfully, he is now a reformed bioinformatician and is looking forward to developing a clean SRA record in the future.

Acknowledgements

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

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Kodama Y, Shumway M, Leinonen R (2011) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* 40: D54 – D56
2. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R *et al* (2010) The European nucleotide archive. *Nucleic Acids Res* 39: D28 – D31
3. Lipman D, Flicek P, Salzberg S, Gerstein M, Knight R, GB Editorial Team (2011) Closure of the NCBI SRA and implications for the long-term future of genomics data storage. *Genome Biol* 12: 402
4. Blaxter M, Danchin A, Savakis B, Fukami-Kobayashi K, Kurokawa K, Sugano S, Roberts RJ, Salzberg SL, Wu CI (2016) Reminder to deposit DNA sequences. *Science* 352: 780
5. Sanitá Lima M, Smith DR (2017) Pervasive, genome-wide transcription in the organelle genomes of diverse plastid-bearing protists. *G3 (Bethesda)* 7: 3789 – 3796
6. Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M *et al* (2014) Data access for the 1,000 Plants (1KP) project. *Gigascience* 3: 17
7. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ *et al* (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* 12: e1001889
8. Smith DR (2013) Death of the genome paper. *Front Genet* 4: 72
9. Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature* 533: 452
10. Lithgow GJ, Driscoll M, Phillips P (2017) A long journey to reproducible results. *Nature* 548: 387