GB Editorial Team *Genome Biology* 2011, **12**:402
http://genomebiology.com/2011/12/3/402

Genome Biology

EDITORIAL

Closure of the NCBI SRA and implications for the long-term future of genomics data storage

GB Editorial Team*

The National Center for Biotechnology Information (NCBI) in the US recently announced that, as a result of budgetary constraints, it would no longer be accepting submissions to its Sequence Read Archive (SRA) and that over the course of the next year or so it would slowly phase out support for this database (http://www.ncbi.nlm.nih.gov/sra). There seems to be a certain amount of confusion in the community about what effect this decision will have. At *Genome Biology* we feel that the free availability of data is an important concept for science, so we asked the views of various interested people on what the short-term implications of this announcement will be, and also how they envisaged the future of data storage in the long term. These people include those involved in the running of the databases (David Lipman (DL) from the NCBI and Paul Flicek (PF) from the European Bioinformatics Institute (EBI)) and users of the data stored in the database as well as data producers (Steven Salzberg (SS) from the University of Maryland, Mark Gerstein (MG) from Yale University and Rob Knight (RK) from the University of Colorado).

1. Why did the SRA close? How widely used by the community was it?

DL: NCBI was facing budgetary constraints and presented a range of options to the National Institutes of Health (NIH) leadership, who chose to phase out the SRA along with other resources. One factor in making the determination was the understanding that because the raw sequence data within the SRA are processed into derived forms in order to answer the underlying biological questions, as methods mature, the SRA was seen as a transitional resource. The SRA primarily has been used by a relatively small community of project analysts and researchers working on methods development in genome scale research projects.

PF: The SRA isn't closing. It started as a joint venture between the NCBI and the EBI, so the NCBI ceasing to accept submissions doesn't meant that the SRA is closing, merely changing and the European Nucleotide Archive (ENA) at EMBL-EBI will remain. The NCBI's decision was based on budgetary constraints. It should be noted that most people don't realize that storage space is only a minor fraction of the budget of the database; the bulk of the cost is associated with the staff who maintain the database, process the submissions, develop the software and so on.

SS: From the outside, it appears that the SRA is closing because of NIH budgetary considerations. One problem is that the amount of sequence being generated is growing at an extraordinary rate, probably faster than increases to the budget. My group uses the SRA a lot. Due to the nature of our work, we rely on it maybe more than others. We download data reasonably frequently, but because of the size of the datasets we try not to do it too often.

RK: The SRA was widely disliked by a lot of users, in particular because it was hard to get data. Partly that was because of poor standards for metadata associated with the data entries. This makes it hard to find the samples you were looking for. It wasn't set up for projects that were generating many samples at a time, and multiplexing with barcoded samples was also not supported. This made it particularly unsuitable for metagenomics data. It's possible that other communities, such as the cancer genomics community, had better experiences.

MG: I don't really know the details. I've heard some speculation that it might be a bit of brinkmanship.

2. What are the alternatives now to the SRA?

DL: Our partners in Europe at the EBI and in Japan at DDBJ will continue to archive raw sequence data in their SRA repositories.

PF: Well, the ENA [via the EBI].

SS: GEO can be used for RNA-seq data. For whole genome sequencing, the alternatives are a little unclear, but it may be that groups that are generating the sequences will have to store the data themselves. Funding agencies may have to consider funding not just the sequencing projects but storage of the resulting data too.

*Correspondence: editorial@genomebiology.com

RK: For metagenomics data, there are a number of community-led databases such as the Metagenomics Analysis Server (MG-RAST, http://metagenomics.anl.gov/), Integrated Microbial Genomes/Metagenomics (IMG/M, http://img.jgi.doe.gov/cgi-bin/m/main.cgi), Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA, http://camera.calit2.net/) and Visualization and Analysis of Microbial Population Structures (VAMPS, http://vamps.mbl.edu/). Other communities probably have their own databases.

MG: We've heard that the ENA will remain open. We've also heard that the NCBI will continue to accept submissions from some of the large established projects, such as ENCODE, at least in the near future.

3. Will other repositories/alternatives provide a suitable replacement for archiving short read data, now and in the future?

DL: The NIH institutes are investigating alternatives to the SRA for archiving sequence read data for its grantees.

PF: The EBI will continue to accept submissions. In order to cope with the increase in submission numbers, we're working on extending ENA's 'ecosystem' model with various groups or organizations acting as brokers, or a single submission pipeline, for data submission. The EBI is working on implementing 'reference-based compression', which will drastically reduce the amount of disk space per stored sequenced base and hence the cost of that storage. Some communities haven't been well served by the way the SRA was organized, with the metagenomics community being a good example. The EBI is working on ways to address that.

SS: If the ENA's SRA is going to be stable, that would be a good alternative. The 1000 Genomes project has been investigating using the cloud. One alternative might potentially be, rather than keeping the sequence data, as sequencing is getting so cheap, to instead just store the DNA and re-sequence it as needed. Certainly for bacterial genomes, that's currently a feasible option. This also avoids the problem of changing formats for digital storage. DNA won't change, but computer disks will.

RK: The EBI's SRA has a better data submission pipeline than NCBI's, and I understand the EBI is keen to involve communities to ensure the database is better tailored to individuals' needs. In the long term it is probably better to just store the samples and then resequence them with improved technology.

MG: It's extremely important that there is a proper archive to put things in. The European archive could be a good place to store data, but it seems to me that it is not good for the US not to have a national archive. It seems strange to me that we would have a situation where the US is paying to generate all these data - probably the majority of genomics data is coming from the US at present - but where it's not prepared to meet the cost of archiving the data.

4. Should we store short reads at all?

DL: As the performance of next-generation sequencing machines continues to improve in terms of speed, cost, accuracy, and length, and as computational processing continues to improve, the need to access the underlying reads decreases. This will vary depending on the application (such as RNA-seq, metagenomics, cancer genomics, and so on). For all of these applications, however, there needs to be more attention focused on the specifications/guidelines/requirements of the derived data, which will become the primary object of study, exchange, and archiving.

PF: As with any archiving project, one needs to consider the cost of storage relative to the potential future reuse. In the course of the 1000 Genomes project, for example, the raw sequencing reads have been realigned and reanalysed many times. Different short read types will have different requirements for storage. For instance, RNA-seq and ChIP-seq datasets probably require less information to be stored than genome sequences where the goal is identifying variants.

SS: It's very hard to get researchers to agree to not keep their raw reads! Even if it's not a case of deleting the data, but just moving it to some sort of less accessible back-up storage. We definitely need to store the short reads in the short term. For instance, if you're comparing differences between two genome assemblies you need to be able to go back to the raw reads to check if the differences are real or if it's just an assembly problem. It's also important for other groups to be able to verify or replicate reported results. Perhaps data will be stored for a few years in a readily available format, then moved to back-up storage, before finally being deleted.

RK: If there is a good chance that the data are going to be used by others, then yes, there is a good case for storing them. Just storing the raw data for the sake of it, however, is probably not worth it. Higher-level data can sometimes be more useful.

MG: I strongly feel that the data should be archived. A lot of the genomics community would feel that generating data just to be thrown away would be anathema.

5. What is going to happen to the back catalog of data currently stored in the NCBI SRA?

DL: NCBI believes that it has the resources to support a static, unmonitored public archive for 12 months. After that, NCBI will re-evaluate. We can also transfer existing data to new providers by tape or disk. All publicly available data are accessible through EBI and DDBJ.

PF: It will continue to be available from the EBI. The data are currently mirrored.

SS: I don't know. We've downloaded some datasets so we'll be able to access them in future.

RK: I don't know in general, but the subset of the data useful for metagenomics is rapidly finding its way into other resources (for example, we have already deposited all our SRA data into MG-RAST).

MG: My impression is that they're certainly not going to delete all of that.

6. How will other data repositories fare in the future given the data deluge that is occurring? Will central repositories become a thing of the past?

DL: For most of the assembled sequence entries in GenBank, including the reference human genome sequence, the underlying sequence reads are not readily available. The phasing out of the SRA, while somewhat accelerated because of budgetary constraints, should not be unexpected given the evolution of next-generation sequencing applications. While the growth in the volume of data derived from next-generation sequencing will be steep, this can certainly be accommodated by the approaches the centralized databases have taken for several decades. So we believe we'll continue to see a mixture of distributed and centralized repositories in the biomedical and life sciences.

PF: I think the community wants something relatively simple. They want somewhere to store their data, and to be able to access it easily. There will always be a role for central repositories. The storage costs are similar whether data are stored centrally or in dispersed locations, but there are economies of scale involved in handling the data associated with a central repository. It is also more convenient for users such as journals and other researchers to have a central point to access the data.

SS: Central repositories are far more efficient. It's not clear if governments will be prepared to fund them, but they should do because it's cheaper. It would be a mess to have different data in different places, perhaps with duplications, with each database having different formats or policies for data access, different reliabilities and so on. GenBank, ENA and DDBJ have been hugely valuable for the community, not least because they have had strict policies for free availability of the data.

RK: In some ways a central repository doesn't make sense. It is hard to envisage a situation where a user will want to access both cancer genomics data and metagenomics data, for instance. The economies of scale with centralized databases are in some sense false. It is cheaper to have user-friendly resources tailored to the needs of the community they're serving. It costs money for users to spend time trying to work out how to access the data. It is likely that any central repository will run into similar problems to the NCBI SRA.

MG: Due to the huge size of the files, uploading and downloading data from a central repository is not easy. The model of a central archive may need to be revisited and we may see in the future an increased use of cloud computing resources.

Published: 22 March 2011

doi:10.1186/gb-2010-12-3-402
Cite this article as: GB Editorial Team: Closure of the NCBI SRA and implications for the long-term future of genomics data storage. *Genome Biology* 2011, **12**:402.