Summary of the meeting on Deep Sequencing Data Management
Keystone Conference, Whistler, March 21
Fitzsimmons Room, 21:00-22:30 PST

Participants

Todd Allen, PARTNERS AIDS Research Center, tallen2@partners.org
Rick Bushman, University of Pennsylvania, bushman@mail.med.upenn.edu
Angela Ciuffi, University of Lausanne, Angela.Ciuffi@chuv.ch
Danny Douek, VRC, NIAID, ddouek@mail.nih.gov
Will Fisher, LANL, wfischer@lanl.gov
Brian Foley, LANL, btf@lanl.gov
Brandon Keele, NCI, SAIC-Fredreick, keelebf@mail.nih.gov
Johnson Mak, Burnet Institute, mak@burnet.edu.au
Darrin Martin, University of Cape Town, Darrin.Martin@uct.ac.za
Doug Richman, UCSD, drichman@ucsd.edu
Redmond Smyth, Burnet Institute, Redmond@burnet.edu.au
Amalio Telenti, Centre Hospitalier Universitaire Vaudois, Amalio.Telenti@chuv.ch
Yegor Voronin, Global HIV Vaccine Enterprise, yvoronin@vaccineenterprise.org

Joining by phone:
Dave O'Connor, University of Wisconsin, david.h.oconnor@gmail.com
Robert Reinhart, AVAC, rjreinhard@gmail.com

Unable to attend:
Paul de Bakker, Broad Institute, debakker@broadinstitute.org
Kyle Bittinger, University of Pennsylvania, kylebittinger@gmail.com
Matthew Henn, Broad Institute, mhenn@broadinstitute.org
Bette Korber, LANL, btk@lanl.gov
Athe Tsibris, PARTNERS AIDS Research Center, atsibris@partner.org

Meeting Description

The meeting was co-organized by LANL HIV database and the Enterprise Secretariat to discuss the issue of management of the large data sets generated in the deep sequencing experiments. The news of the closure of the GenBank's Short Read Archive (SRA), which surfaced when meeting planning was underway, made this discussion particularly timely. While the majority of discussion focused on deep sequencing of viral population, many of the discussed issues will be relevant to deep sequencing of B-cell and T-cell repertoires, as well as to transcriptome analyses.

Discussion Summary

- Closing of SRA is a lesson in inappropriate approach to deep sequencing (DS) data management. The general perception is that its failure was caused primarily by a very rigid data structure, which did not match the experimental design and provided no flexibility to adapt it to the diverse needs of the research community. Several people also commented that SRA was plagued by technical issues, with unreliable and difficult to use tools for data upload and retrieval. However, it is important to hear from GenBank on what exactly went wrong and how they are planning to address these issues in the future. It is also important to let them know the field-specific needs of the HIV research community to make sure that the next iteration of SRA addresses them.

- Importance of data management experts was highlighted. They are needed to curate databases, develop tools, and assist data producers with annotation and deposition of data. It was agreed that these experts would be most effective if placed directly into laboratories of data producers, even if for limited periods of time (although some data-rich labs have 3-5 full-time positions to manage and analyze the data).
- DS data requires multiple levels of processing, with each level allowing multiple approaches with no clear winner at this point and constantly appearing novel and improved methods ("everyone is doing it differently"). Thus, it was generally agreed that depositing raw "off-the-machine" data is essential. However, final processed data set should also be deposited in order to support the conclusions made in the publication, inform researchers how the data has been analyzed previously and what conclusions were drawn. There was no clear agreement on what intermediate processing steps should be routinely included as this may depend on the specific experimental design.
    - In the opinion of some participants, "processed data will not be of interest to those who are qualified to make use of DS datasets". It is critical that scientific interests of data consumers be a major factor when designing the DS database.
- It was suggested that several well-described properly-annotated quality-controlled reference datasets should be provided to the community. These data sets can be used to validate and compare novel tools for data processing and analysis and serve as examples for the field. LANL HIV database seems to be well positioned to host these datasets.
- The strategic approach to DS data management should acknowledge the fact that the field is rapidly changing right now. It is quite likely that sequencing technology and analysis will be radically different in 5 years from now. Thus, the efforts are better spent on solving immediate challenges in the field and providing flexible solutions to accommodate the latest developments in technologies and tools.
    - Closure of SRA is one such immediate challenge, which has to be solved quickly. Currently there is no place to deposit data in support of a publication.
- The mantra for the DS database should be "Make easy things easy and hard things possible".
- It was suggested that commercial data-storing solutions could be explored. With the vast amounts of generated data, data transfer is quickly becoming a rate-limiting step. One approach is to store data in the cloud and provide tools that would process the data remotely without download to user's computer and provide only the results. Some commercial solutions allow free data storage and charge for CPU usage.
- It was suggested that physics and astronomy fields should be consulted as they have been facing similar issues (of managing extremely large data sets) for some time already and some of their approaches may be applicable in the HIV field.

  - Within virology field, the issues faced by the HIV research community will be shared by many other viral fields, such as HCV or Dengue. We should explore possibilities for collaboration. HIV field may become the leader in this area.

- Proper controls are very important for DS datasets and have to be included. The field needs to agree on a minimal set of controls that have to be provided with each dataset.

- It appears that multiple tools are currently being developed for the same functionality by multiple parties, which are often not aware of each other's efforts. A repository or a comprehensive list of available tools and of their functionality would be very helpful.

  - One possibility is to adopt an open-source modular approach to software development, similar to R project, where modules can be produced by the community and provided for everyone to download.

- Viral sequences are currently not viewed as infringing on patient privacy (such as, for example, SNP or genome data). However, with development of cheaper and very comprehensive sequencing technologies, viral sequences can be used as fingerprints to uniquely identify a person. The issue of patient privacy needs to be addressed and common guidelines need to be developed and adopted.

- Data integration needs to be considered as well, as in the near future DS will be used to generate linked data sets of patient's genome, transcriptome, virus populations, etc… Data sets are expected to become orders of magnitude larger and more complex. Integration of sequence data with immunological data presents a separate set of challenges.

- There was an agreement that there is an urgent need in guidelines for metadata to be included in annotation and on appropriate ontologies and vocabulary for that metadata.

Action items

1. Brian Foley and Will Fisher: to explore whether LANL HIV DB would be able to host reference DS datasets and tools.

   Status: LANL HIV DB is planning to make available several datasets from CHAVI and will consider hosting relevant software (or links) for data analysis.
2. Volunteers needed: to contact GenBank about closure of SRA, discuss their future plans and communicate the needs and perspective of HIV research community.
3. Volunteers needed: to draft guidelines for metadata to be included in annotation and on appropriate ontologies and vocabulary for that metadata.
4. Volunteers needed: to contact people responsible for R project to hear their advice and "lessons learned" from establishing that platform.
5. Enterprise Secretariat: to organize a follow-up meeting to continue discussion and update on progress since last meeting.

   Status: A meeting is being planned for July-August to be held in New York (TBC).