Short Read Archive (SRA) Requirements Meeting

Created: July 29, 2007; Updated: November 20, 2010.

<table>
<thead>
<tr>
<th>Status</th>
<th>Historical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Date</td>
<td>2007</td>
</tr>
<tr>
<td>Inactive Date</td>
<td></td>
</tr>
<tr>
<td>Scope</td>
<td>INSDC SRA</td>
</tr>
</tbody>
</table>

Date: July 27, 2007

Place NISC Conf Room Rockville MD USA

1. Attendees

1. Mike Attili | Helicos | mattili at helicosbio.com
2. Vladimir Alekseyev | NCBI | aleksey at ncbi.nlm.nih.gov
3. Inna Belaia | NCBI | belaia at mail.nih.gov
4. Toby Bloom | BI | tbloom at broad.mit.edu
5. Vivian Bonazzi | NHGRI | bonazziv at mail.nih.gov
6. James Bonfield | Sanger | jkb at sanger.ac.uk
7. Kevin Bradtke | Genecodes | kbradtke at genecodes.com
8. Deanna Church | NCBI | church at ncbi.nlm.nih.gov
9. Guy Cochrane | EBI | cochrane at ebi.ac.uk
10. Anthony Cox | Illumina | anthony.cox at solexa.co.uk
11. David Dooling | Wash U | ddooling at watson.wustl.edu
12. Adam Felsenfeld | NHGRI | felsenfa at exchange.nih.gov
13. Paul Flicek | EBI | flicek at ebi.ac.uk
14. Tim Hunkapiler | ABI | tim at discoverybio.com
15. Steve Leonard | Sanger | srl at sanger.ac.uk
16. Elaine Mardis | Wash U | emardis at wustl.edu

Table continues on next page...
2. Data Production and Archive Requirements (led by David Dooling)

One approach is to show only the base coverage and coverage depth at each base of the reference.

A table of data requirements for each of the vendors was presented.

2.1. What do the centers currently do about data retention?

WUGC – We store SFF and metrics from 454 runs. We store R and D directories for two months, then purge them. For Illumina, we store everthing, but hope to store only the resulting .prb files.

WIBR – We keep all raw data for 1-2 months. Then we keep only a sub-sampling of images for subsequent quality analysis. We do our own alignments so the secondary analysis is not retained. For SOLiD, we do not pull images off the machine as this would be impractical. Just pull off processed intensity files (about 600 GB for 2 slides).

BCM – We are keeping everything for now but would like to get to a 3 month retention period. We would like to keep post-image analysis files.

SC – We keep images for a few weeks. We keep SFF from 454. We would like to keep SRF from Illumina once that is ready, and alignments are kept in a local format.

2.2. What should be publicly archived?

- For 454, SFF plus meta data should be archived. The new SRF format should subsume SFF.
• Everybody agreed archiving image data is infeasible.
• Metrics from the instrument runs should not be archived. Even though these are minimal in size, they are interesting only to production QC at the Centers and not to users of the archive.
• Everybody agreed it is important to accept experimental meta data in a form which can be readily archived (as opposed to now, where it is embedded in each trace record).
• Everybody agreed that Sanger data should remain in the current Trace Archive. Existing deposits of 454 data could be duplicated in the SRA, but this would entail additional migration work.
• Everybody expressed the desire that SRF be the transaction medium for run data. It was felt that once adopted, the vendors would augment or replace existing delivery formats with SRF.
• There was a question about logging four intensities (one per channel per cycle) rather than one (for the principal base call). Many Sanger-type sequencing analysis tools used only the intensity value of the principal base call at any consensus position. However, with new technologies having multiple high intensities per cycle would be an error.
• There was consensus that one should not archive images even for limited rescoring R+D purposes. This is because such research will be taking place at the Centers, and for those who want to engage in it, they should go buy an instrument, rather than relying on the SRA.
• Only a subset of quality values in the old phred range are actually used by these technologies and it’s not clear what they represent. Can fewer bits be used to encode them (other than the current 7 needed to encode 1-100)?
• here was a question about whether it would be necessary to store negative values for processed intensities, or simply truncate them to zero. This could increase the SFF file size by as much as 10%.

2.3. Should the results of secondary analysis be archived?

A major part of the value provided by the new technologies is resequencing. Alignments to reference sequences, and analysis derived from these alignments, provide the primary starting point for investigators. There was a discussion about whether these should therefore be captured. In some cases Centers are performing their own secondary analysis. In order to control the scope of the SRA, only primary analysis results (the results of processing a sequencing run without respect to a reference) would be archived. The expectation is that other analyses would be archived by downstream resources even if these don’t exist at the moment.
3. Database Structure (led by Vladimir Alekseyev)

3.1. Should project registration be required for all SRA submissions?

The proposal is to ask submitters to describe their project in terms of an experimental design. Such a descriptor could be linked in as a first class object in the Entrez system.

- There is an effort underway to develop an international project id (ISNCD) that would represent project tracking information mirrored at NCBI, Ensembl, DDJB, and possibly other archives. So NCBI_PROJECT_ID will be subsumed by this new id.
- If we separate project metadata from data submissions, then there will be an asynchrony problem. Others suggested one should allow, for small projects at least, unified submission of meta data and data (“in-line” submission feature).
- The group wanted careful definition of fields in a RFC type document. There was a discussion about how deeply to represent the project meta data. The consensus seemed to be that some metadata will be useful to be able to query against when extracting data from the SRA, but these should not be required fields, nor should we attempt to design an ontology for the various experiments as these are being addressed by other efforts (for example CAMERA and Gemina). Also, the metadata acquisition should be flexible, allowing for center defined tag-value pairs.
- One observation is that metadata submitted and stored on the project level will be small so there is no need to make it efficient. Also, the SRA submission process may accept metadata as a proxy for other resources, some of which have yet to be designed. The SRA itself is not intended to track project metadata.

Another popular feature will be “hold until publish”. Now that submissions are tracked by experiment or project, this will be easy to implement.

3.2. Introduction to the “Spot” Abstraction

A common property of the new technologies is that they gather, through image processing, one intensity function for each reaction container (well/spot/bead), which we will call a “spot”. Adapters, paired end reads, linkers, barcodes, and other subsequences can be represented as annotations on the native read that partition it. In order to access the usable sequence itself, or the other component “technical reads”, the SRA would supply through its meta data directives for how to parse the native sequence to extract these objects.

All reads would receive an accession in the form run.spot.read. This accessioning scheme is indexed, rather than random access. Thus reads do not have a “name” as such. The scheme presented encodes enough information to locate a read individually while eliminating the need to store a tag for each read that would use almost as much space as the read itself. The accession is stable in that it is decided at the time of submission. Therefore, any downstream process will be able to refer to it so long as the order within the submission remains unchanged.
• There was quite a lot of discussion about whether reads need to be named. The consensus developed that doing so would be too expensive in terms of space requirements. The SRF format will be supporting read names, but this was done in order to address an application space beyond SRA. Therefore, SRA should not archive read names, although these can be used in submissions.
• There will be a need to call out in both SRF and SRA whether encoding of intensity values is in terms of base space, flow space, or color space.
• There was a discussion about whether padding the accession string is a good idea. EMBL is trying to migrate away from that. On the other hand, it is convenient for searching and sorting to have fixed length strings.
• There is a need to make the notion of an experiment flexible. It should be as big as a genome project and as small as a lane or region.
• There is a need to allow for incremental submissions, particularly when data sets are huge.
• There was a discussion about whether to allow for many-to-many experiment to project mapping. Right now the abstraction says that a project or study is composed of one or more experiments, each of which may generate a run of data.
• Do we need to know the total number of spots in the submission or expected total for the experiment?

3.3. Can we reasonably expect to submit, archive, and download all this data?

There was a lively and perhaps inconclusive discussion about whether the level of detail being proposed in the SRA will result in an unmanageable torrent of data. It was proposed that even under full compression, 1 GB of sequencing data will result in 10 GB of storage data, and that with 100-200 new technology instruments producing each week this could amount to deposits of 1 TB per day. Has there been any planning or modeling of what might happen if this situation were to materialize?

A related question is whether centralization of the archive makes any sense. Would it not be better to provide a central indexing service that leads back to the Centers, who will actually provision the data requested by users? Centers responded by saying that they do not want to be in the business of satisfying user requests for data, and that they were looking to NCBI to handle this.

4. Database Submission Format (led by Asim Siddiqui)

The SRF format was reviewed. Issues about read ids vs. read names were debated.

SRF is a separate effort that will hopefully conclude with a 1.0 specification sometime in August.

5. Data Retrieval (led by Gabor Marth)

This discussion tried to anticipate uses for the SRA. Some key points:
• here is a need to track the provenance of the source material (DNA). How was it isolated, was it methyl filtrated etc. These would have bearing on library construction and assembly. At the same time, one should not try to invent an ontology to describe this, just useful fields.
• Another need for data tracking is library stats: expected insert size etc.
• A discussion took place about whether to provision all the data from a run, or actively quality filter the data down to the “usable” subset. While this might be convenient for some applications, it is also fraught with issues. Historically, the Trace Archive accepted all data from a run regardless of quality level. Also, the issue of whether something is usable because of low quality of contamination is often not knowable until downstream processes are applied.
• A similar discussion ensued about accepting reads that did not align to the reference sequence used in the experiment. The observation was made that not aligning to a reference sequence is not a reason to not submit.

5.1. How will assemblies use this data?
There is a localization issue when referring to reads in the assembly or alignment context. If reads are accessed in storage order, then the time needed to perform random access retrieval will dominate any assembly download or display function. Therefore, reads will have to be reordered. The question arises whether the SRA will do this on retrieval. Various proposals include using a prefix on read ids in order to embed tracking information needed for localization. Then there would need to be a directive that would tell the output streamer to formulate the ids in a certain way.

5.2. What are the units of retrieval?
The use case for the short reads may determine the retrieval chunks. They could be run/region or plate/slide/lane order, or some other locality. There may be context-driven retrieval. This area will require further requirements development.

Everybody agreed that the user should be informed as to the expected size and time of the data download, and for the user to have the opportunity to cancel it. A web tool that would report download status similar to submission status might be warranted.

6. Data Submission Software (led by Guy Cochrane)
There will be three submission activities for a project:

1. Registration activity (email, web?)
2. Meta data submission to SRA (xml)
3. Experimental results submission to SRA (srf)

These activities might happen at different times, or the same time for small projects.
6.1. How does one mask off data that is not part of the experiment?

One of the issues is how to deal with contaminants. These are often not found until relatively late in the project life cycle, which may be well after the sequencing data have been submitted. There was a debate about whether tracking of contaminants should be the responsibility of the SRA, or downstream archives. Clearly it is convenient to be able to download a contaminant-free dataset.

Therefore, we may need facilities to:
  - Mask data within a run
  - Suppress data within a submission
  - Withdraw a submission

6.2. Should data that cohabit a run but otherwise are not related share a submission?

There was a consensus that Centers should endeavor to split up unrelated portions of a run so that each portion maps to an experiment. But this may require development of SRF slicing and dicing utilities. There was a suggestion to publish the specs and interfaces for such utilities, and let the vendors develop these.

7. Policy Issues: Recommendations to Funding Agencies (led by Elaine Mardis)

7.1. Will there be support for medical resequencing?

No. The policy development is underway, but we should assume this is not in scope at this time.

7.2. Is it ok to just submit bases and qualities?

There was a discussion about whether one should allow archival bases and quality data only. This would certainly be simpler and faster. But experience with Trace Archive was that in the long run requiring archival of the intensity files was very rewarding, and that migrating early submissions proved impossible. Another issue supporting full disclosure of intensity data is that vendors will want maximal representation of their data. Leaving out intensities will raise more questions than answers. Also, what is the value added by using four channels of quality scoring if intensity data is tracked? Finally, it is in general it is difficult to later change the rules to make them more stringent.

8. Summary (led by Deanna Church)

- It appears that the data model proposed is adequate.
- SRF is the submission and probably retrieval medium.
- SRF will be supported by all the vendors.
• Read names will be accepted but not tracked by the SRA.
• Read names might be auto generated by the SRA.
• Distinct project registration is important, but there should be an inline solution.
• SRA should be tracking meta data at the levels of experiment and run.
• NCBI will issue a straw man xml schema and gather further comments.
• NCBI aims for an October release of the SRA.
• SRF will be finalized in August.

9. Supplementary items

Vladimir Alekseyev's Presentation for the meeting