SRA Concepts

National Center for Biotechnology Information – National Library of Medicine
Draft B 13 Mar 2008

1 Overview

The current Trace Archive for capillary-based sequencing platforms tracks sequencing data as individual traces. The arrival of new massively parallel sequencing technologies has complicated representation of such experimental data. In addition, investigators are conducting increasingly diverse experiments with greater throughput. More data producers (particularly those from smaller labs) are coming on line. Finally, we continue to see greater reliance of the community on public resources like those at NCBI to accession experimental data for archival, retrieval, and publication.

The Sequence Read Archive (SRA) is an entirely new resource at NCBI. It is being designed specifically meet the challenges presented by massively parallel sequencing technologies.

This document defines entities and relations that make up the Sequence Read Archive (SRA). .

1.1 Goals

- Provide a central repository for next generation sequencing data.
- Provide links to other resources referencing or using this data.
- Provide users with retrieval based on ancillary information and sequence comparison.
- Track studies and experiments (project metadata).
- Allow flexible submission and retrieval of ancillary data.
- Improve database efficiency through normalization of data structures.

- Separate submission from content.
- Establish basis for user-interactive submission and retrieval.

1.2 Related Documents

- NCBI Trace Archive Documentation

2 Concepts

A fundamental departure from the current Trace Archive design separates the experimental data from its metadata. The metadata are now organized as follows:

Study – A study is a set of experiments and has an overall goal.
Experiment – An experiment is a consistent set of laboratory operations on input material with an expected result.
Sample – An experiment targets one or more samples. Results are expressed in terms of individual samples or bundles of samples as defined by the experiment.
Run – Results are called runs. Runs comprise the data gathered for a sample or sample bundle and refer to a defining experiment.
Submission – A submission is a package of metadata and/or data objects and a directive for what to do with those objects.

2.1 Differences with the Trace Archive

The separation of metadata from data allows for separable submissions. Thus information about the study or experiment can be posted when it becomes decided (typically early in the project life cycle), and can await the experimental results (typically late in the project life cycle). This separation of concerns also allows for new features including hold-until-publish and user-controlled modification.

Metadata disassociation also permits gathering of richer information about studies. The old method of collecting study specific attributes (for example, the *salinity* attribute for seawater metagenomics studies) has been dropped in favor of a more flexible system of Center-controlled vocabulary. This policy change allows the SRA to serve as a repository of important ancillary data while avoiding the pitfall of ontology development.

All next generation sequencing technologies perform image processing in order to reduce sequencing data to base, quality, and intensity calls. These will be accepted in either the incipient Sequence Read Format (SRF) data file or manufacturer specific data files. Currently, the Trace Archive accepts only ZTR files.

The SRA will from its inception accept any secondary analyses typically performed on the next generation data. These might include alignments, small-scale assemblies, oligo profiles. As many of these analyses are currently being developed, the SRA will accept their reports and data in “blob” form with virtually no internal structure. This will establish a new “one-stop” submission paradigm appropriate for small projects, projects executed by automatic pipelines, and projects submitted from newer Centers with little experience interacting with NCBI. Future releases of the SRA will improve the routing of analysis data to downstream repositories. The Trace Archive accepts only traces.

The vast increase in the amount of data offered by the next generation technologies has required modification of the capillary-based sequencing notions of reads and read accessions. The SRA will be offering reads for retrieval based on type and container relationships specified in the retrieval query. Accessions may be assigned as specified by the user. However, there will be no tracking of individual read accessions even if these are designated in the original submission because of the excessive amount of storage space needed just to accommodate accessions.

2.2 Submission

A **submission** is a wrapper around study metadata and run data, plus directives for the SRA operators.

Properties include:

center_name
submission_name
submission_date
contact_info
study accession (and possibly one or more experiment accessions)
directive

Here are some directives (operations) concerning submissions:

- New – This is a new submission for this **study.**
- Modify – This submission entirely replaces a previously referenced submission. The goal is usually to repair a flawed submission, or to complete an interrupted submission.
- Release – Release to the public an embargoed submission.
- Suppress – The submission should be entirely suppressed.

Submissions are tracked using a publicly available web interface. Submissions can be referenced in two ways: by a public moniker assigned by the Center, and by a private key that is returned to the submitting Center once its submission has been logged at NCBI. A submitting Center can update or replace only its own submissions. Both keys must be used in all subsequent transactions concerning a submission.

A submission can be held until publication (“HUP”, or embargo).   Such submissions do not appear in public status information. The submitting Center must send a directive to NCBI releasing the submission to the public (there are no decision rules).

2.3 Study

A study contains the project metadata. The universal project id replaces the currently used NCBI project id and will be valid at EMBL and other repositories. The SRA will require that such a project id exist prior to any submission, as this will be the only way to track related submissions including those from collaborating Centers.

A study is composed of a set of experiments. The experiment is a logical entity describing the target (one or more samples), and the sequencing method used on that sample. The three elements of an experiment are:
Design – Captures the layout of the experiment, including sample organization, library organization, and spot organization as determined by the investigator.
Platform – Specifies the next generation sequencing technology and parameters selected by the Center.
Processing – Specifies the next generation sequence processing parameters as determined by the vendor or manufacturer.

The study representation can be used gather together distinct aspects of a sequencing project. Consider for example a hybrid project where two next generation sequencing runs are performed in a random phase to cover a bacterial genome, followed by scaffolding provided with a paired-ends library. The study can refer to the overall sequencing effort directed at the genome and would consist of two experiments, one for the multiple next generation runs and one for the paired-end sequencing.

Much of the data contained in the study objects could be initialized in the SRA well before sequencing actually takes place. Subsequently, when the data become available, the Center may complete the submission by adding to the study, thus avoiding much correspondence concerning ancillary data that tends to take place immediately before publication.

2.4 Run

Each experiment in a study may receive one or more runs of sequencing. Sequencing runs identify a sample accession. Thus the study and its experiments must be defined before the submission of run data. The advantage of this approach is flexibility: multiple samples can be sequenced by a single run, or multiple runs can encompass a single sample. Study metadata can still be included in the submission so long as they define all tags used in the run data.

2.5 Sample Organization

An experiment will support the following sampling formats:

Individual – The sequencing effort targets one sample only.
Multiplexed – The sequencing effort targets a set of samples and each read can be mapped to a sample. This mapping would be resolved on data retrieval.

Pooled – The samples that are being sequenced are known and can be listed, but the reads cannot be mapped to them.
Population – The samples cannot be distinguished but their overall number may be known.

Where possible, a sample should be identified by its taxon id rather than a name. In some cases a taxon id will not be relevant. Where bar codes (known oligo sequences incorporated into the sequencing material) are known, these should be listed by the submitter.

2.6 Library Organization

Some concepts in library construction are borrowed from capillary-based sequencing. Although it is not always necessary to construct a library for next generation technologies (this is indeed one of the advantages), some aspects of the source material may be important to track, including:
Library Name – Center assigned library name often used by collaborators
Library Source – Type of DNA or RNA used in the experiment
Library Selection – Whether the source material was selected for certain properties (for example methyl filtrated)
Library Layout – Number, order, orientation, and distance of associated reads (for example, 2 Kbp paired ends).
Library Protocol – Free form text that documents the "library construction" steps.

2.7 Spot Organization

A spot is a new kind of abstraction that captures all the data associated with one intensity function in time. Thus reads related by mate pairing or bar coding can be tracked implicitly by virtue of sharing a "reaction container." The concept is roughly analogous to that of a *growth template* in capillary-based sequencing, in which mate pair reads are related by their sharing of an insert in the cloning vector.

A read is classified as to whether it is a technical read (primer, linker, adapter, bar code, etc) or an application read (single read, paired ends, etc). Reads are indexed by one of three methods: base-based coordinates, cycle-based coordinates, or by alignment to an expected oligo sequence (such as a linker or bar code).

Specification of spot decoding is done at the level of the experiment design, so that this information is bound once per experiment, and not once per read, as is currently done.

2.8 Read Organization

NCBI proposes the following accession format for each read:

```
SRA000000.ssss.rrr
```

where SRA000000 denotes the sample accession, ssss denotes the “spot” number in the run, and rrr denotes the read index within the spot. A motivation for using integer encoding of read names in this way is to eliminate the heavy burden of accessing small units of information with large randomly accessible keys. Integer encoding allows for implicitly indexed access which takes far less computational time to manage. In addition, range specifications can replace lists when referring to contiguous read sets.

For example, application mate pairs might be retrieved for a certain sample by the following regular expression: *SRA458693.*.00[24]* .