SRA Analysis Submission Guide

National Center for Biotechnology Information (NCBI), National Library of Medicine

Version 1.0 Draft E 20 Oct 2010

1 Contents

2 Overview

This document reviews submission procedures and guidelines for SRA analysis objects, including

- De novo assemblies (to be specified in a future version of this document)
- Reference alignments
- Sequence annotations (to be specified in a future version of this document)
- Abundance measurements (to be specified in a future version of this document)

In keeping with developing NIH policy, this document also shows how to submit primary sequencing data as a part of the analysis object.

2.1 History

Guidelines for SRA analysis submission were developed in conjunction with two NIH roadmap initiatives: The Cancer Genome Atlas (TCGA), and the Human Microbiome Project (HMP). The TCGA established early requirements to allow submission of all needed primary data through the BAM file format. The HMP pioneered requirements for annotation of raw sequencing data from metagenome projects where assembly into higher constructs is difficult.

2.2 Goals

1. Meet the needs of users by providing a home somewhere in the data model for all desired properties.
2. Distinguish where in the data model each desired property should reside.
3. Define processing directives that might be important to interpreting the sequencing/alignment data and loading it into an archive database.
4. Eliminate dependence on spreadsheets and filenames to convey metadata.
5. Provide searchable metadata that can be used by query writers in the public database.
6. Provide query source for programmatic construction of component descriptions that users of protected data will see inside the dbGaP authorized access download interface.

2.3 Scope

In its current revision, this document describes metadata needs for BAM file submission. It does not describe the submission modalities. Higher level analysis types and other analysis types are not described. Some BAM files are submitted using preexisting SRA data, other BAM files will be submitted containing de novo sequencing data as part of its payload. This document does not describe archive requirements for the BAM file read placement records, which may have additional requirements in order to be loaded into the NCBI alignment database. These requirements need further development.

2.4 Revision History

Drafts A-E created 2010-09-14 to 2010-10-08. Document released with draft status 20 Oct 2010.

2.5 Related Documents

Elements of TCGA project requirements have been incorporated into this document [Tim Fennell. *BAM File Format for TCGA Submissions.* Draft v2, July 9, 2009.]

Submitters should also consult the established SRA submission documentation :
http://trace.ncbi.nlm.nih.gov/Traces/sra/static/SRA_Submission_Guidelines.pdf
http://trace.ncbi.nlm.nih.gov/Traces/sra/static/SRA_Quick_Start_Guide.pdf
http://dbgap.ncbi.nlm.nih.gov/aa/aspera_transfer_guide.pdf

Here is the released SRA 1.2 XML Schema:
http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA_1-2/SRA.analysis.xsd?view=log

Here is the change notice for transition from SRA 1.1 (production) to SRA 1.2 (current draft):
http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA_1-2/SRA_XML_Schema_Change_Notice_SRA_1-2.pdf?view=co

For details on the SAM/BAM specification please reference:
http://samtools.sourceforge.net/SAM1.pdf

A BAM file validator utility is available here:
http://picard.sourceforge.net/command-line-overview.shtml – ValidateSamFile

3 Data Model

NCBI Object	Accession	Sequencer Production Unit	BAM Component
Submission envelope	SRA	n/a	n/a
Analysis	SRZ	n/a	BAM file
Study	SRP	n/a	n/a
Experiment	SRX	n/a	Library (LB)
Sample	SRS	n/a	Sample (SM)
Run	SRR	Lane/slide/plate	Read Group (RG)
Reference Sequence	NC_ and others	n/a	Sequence Dictionary (SQ)
Probe set	Pr	capture array	n/a

3.1 Submission Metadata

The submission metadata pertains the submission "package" or "envelope" conveying the data to the archive.
Submitter id/alias – Submitter's name or alias for the submission.
Submission date – ISO 8601 date for the date of transmission of the file to NCBI.
Submitter contact – name and email address of the submitter contact(s).
Center name – NCBI short name for the submitting center.

3.2 Analysis Metadata

Analysis alias – Submitter's name or alias for the analysis object.
Analysis title – The title string that will be presented to users of the public archive when this record is retrieved in a search result. Please limit this string to 80 characters.
Analysis type – DE_NOVO_ASSEMBLY | REFERENCE_ALIGNMENT | SEQUENCE_ANNOTATION | ABUNDANCE_MEASUREMENT
Analysis Description – A free form description of the analysis product and the process by which it was produced.
Analysis date – ISO 8601 date when the analysis was completed and the BAM file written.
Analysis center – NCBI short name for center that performed the analysis
Analysis Files and Checksums – Each analysis file and its MD5 checksum.

3.2.1 Reference Alignment Metadata

This section enumerates metadata components that are specific to reference alignment analysis objects.
Standard Assembly – Controlled name for the reference assembly or set of reference sequences used in the alignment. The following table shows a catalog of standard assemblies that are supported by NCBI. Other SRAs may define and support different assemblies. A set of cross referenced sequences may also be specified as the reference assembly.

short_name	Description	source
GRCh37	GRCh37 is the Genome Reference Consortium Human Reference 37 released 24-FEB-2009, and includes haploid and alternative loci sequences. This reference can also be specified in the NAME field (db="gencoll", accession="GCA_000001405.1")	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.shtml
GRCh37-lite	GRCh37-lite is a subset of the full GRCh37 human genome assembly plus the human mitochondrial genome reference sequence (the "rCRS") from Mitomap.org. This set of sequences excludes all the alternate loci scaffolds of the full GRCh37 assembly, and has the pseudo-autosomal regions (PARs) on chromosome Y masked with Ns. This haploid representation of the genome is provided as a convenience for use in alignment pipelines that cannot handle the multiple placements expected in the PARs	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.shtml http://www.mitomap.org/MITOMAP

	and in regions of the genome that are represented by the alternate loci.	
HG18	The March 2006 human reference sequence (NCBI Build 36.1) was produced by the International Human Genome Sequencing Consortium and is distributed by UCSC.	http://genome.ucsc.edu/cgi-bin/hgGateway?db=hg18
NCBI36	NCBI Build 36.3 released 24 March 2008. This build consists of a reference assembly for the whole genome, alternate assemblies for the whole genome produced by Celera and by JCVI, plus alternate assemblies for some parts of the genome.	ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/
NCBI36-HG18_Broad_variant	Broad Institute variant of Build 36/HG 18.	ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/special_requests/assembly_variants/NCBI36-HG18_Broad_variant.README
NCBI36_BCCAGSC_variant	British Columbia Cancer Agency Genome Sequencing Center variant of Build 36/HG 18.	ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/special_requests/assembly_variants/NCBI36_BCCAGSC_variant.README
NCBI36_BCM_variant	Baylor College of Medicine variant of Build 36/HG 18.	ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/special_requests/assembly_variants/NCBI36_BCM_variant.README
NCBI36_WUGSC_variant	Washington University variant of Build 36/HG 18.	ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/special_requests/assembly_variants/NCBI36_WUGSC_variant.README

Custom assembly – It is possible to specify a list of contigs including de novo assemblies of unmapped reads that together comprises the reference sequence. More development is needed to define the business rules that would apply to this kind of reference specification.

Processing pipeline – The sequence of processes/tools/operations and their versions can be specified for the alignment process.
Processing directives – certain specific instructions to the data loading software, or properties that users of the data should be aware of:

- **alignment_includes_unaligned_reads** - Whether unaligned reads are provided in the alignment, and what to do with them
- **alignment_marks_duplicate_reads** - Whether duplicates are removed from the alignment
- **alignment_includes_failed_reads** - Whether non-PF filtered reads have been included in the read groups

3.3 Study Metadata

For open SRA submissions, the submitter must create or reference a SRA data producing study (SRP).

For protected SRA submissions, the submitter must reference an existing dbGaP authorized access study (phs). Reference can be made to the study handle with refcenter="NCBI". Submitters should NOT create these records.

3.4 Sample Metadata

For open SRA submissions, the submitter must create or reference a SRA sample or BioSample (SRS).

For protected SRA submissions, the submitter must reference an existing BioSample record (SRS). Reference can be made to the submitted sample name with refcenter set to the original repository short name. Submitters should NOT create these records.

Open SRA samples or Biosamples have diverse attributes and information content.

Protected SRA samples are exported from dbGaP and make visible a standard subset of attributes, including at the time of this writing:
Title – Brief yet unique headline returned with the record as part of a search result.
Identifiers – SRS accession, dbGaP sample accession
Organism – Target organism {human}
Original_repository – Namespace for sample set {TCGA}
Submitted_sample_id – Sample name {TCGA aliquot id}
Submitted_subject_id – Subject name {TCGA subject id, substring of the aliquot id}
Sex – {male, female, unknown}
Sample_type – Project specific sample type {TCGA: normal, primary tumor, etc}
Is_tumor – {0,1}
Histological_type – Sample diagnosis {TCGA: Serous Cystadenocarcinoma, etc}

Analyte_type – {DNA, RNA, etc}
Study_name – Short name for the parent study {TCGA}
Description – Free form text describing the sample.
Links – Includes link to parent dbGaP authorized access study homepage

An example of a TCGA record that has this information:
http://www.ncbi.nlm.nih.gov/biosample/limits?term=TCGA-13-0725-01A-01D-0359-05

3.5 Library Metadata

Each library mentioned in the BAM will map to a new or existing SRA experiment. The SRA experiment contains the following data:

Experiment title – The title string that will be presented to users of the public archive when this record is retrieved in a search result. Please limit this string to 80 characters.
Experiment description – Description of the library and its sequencing.
Library Name – Controlled vocabulary of terms describing overall strategy of the library.
Library Strategy – Controlled vocabulary of terms describing overall strategy of the library. Terms used by TCGA include {WGS, WXS*, RNA-Seq*}.[i]
Library Source – Controlled vocabulary of terms describing starting material from the sample. Terms used by TCGA include {GENOMIC, TRANSCRIPTOMIC*}.
Library Selection method – Controlled vocabulary of terms describing selection or reduction method use in library construction. Terms used by TCGA include {Random, Hybrid Selection}.
Library Layout – Specification of the layout: fragment/paired, and if paired, the nominal insert size and standard deviation.
Library Protocol description – Description of the library construction protocol, or reference to a standard protocol.
Targeted loci* - Set of loci to be selected for sequencing {16S RNA, exome} and associated probes.
Platform – Controlled vocabulary of platform type {Illumina, LS454, AB_SOLID, CompleteGenomics}
Instrument model – Controlled vocabulary of instrument models {Illumina Genome Analyzer II, etc}
Expected sequence length – Number of raw bases or color space calls expected for the read (includes both mate pairs and all technical portions).
Sequence processing software and version – Name and version of sequencing processing software used.

3.6 Run Metadata

Each read group will map to exactly one new or existing SRA run.
Run name – Production flowcell/slide/plate name

Run date – ISO 8601 date the run was produced
Run center – NCBI center short name where the run was produced (useful if different from the submitter).
Run file info – Information about the run data file(s). If BAM, then this is the BAM file name and its checksum.
Processing directives – certain specific instructions to the data loading software, encoded as tag-value attributes, including:

- Actual raw sequence length, including both mate pairs and all technical portions.
- Quality scoring system {phred, log-odds}
- Quality basis character {! or @}

4 Reference Alignment

This section reviews preparation of the reference alignment submission.

4.1 Reference Alignment Analysis Object

Each reference alignment must have an associated SRA analysis metadata object. Generally there is a one-to-one relationship between the BAM file and the analysis object.

4.1.1 Example Reference Alignment XML

```
<?xml version="1.0" encoding="UTF-8"?>
<ANALYSIS_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <ANALYSIS accession="SRZ000303"
            alias="G2043.TCGA-13-0751-10A-01D-0446-08.bam"
            center_name="BI"
            broker_name="dbGaP"
            analysis_center="BI"
            analysis_date="2009-07-31T00:00:01-05:00"
  >
    <TITLE>TCGA reference alignment of sample TCGA-13-0751-10A-01D-0446-08</TITLE>
    <STUDY_REF accession="SRP000677"/>
    <DESCRIPTION>The Cancer Genome Atlas (TCGA) HG18 reference alignment of
                 sample TCGA-13-0751-10A-01D-0446-08</DESCRIPTION>
    <ANALYSIS_TYPE>
      <REFERENCE_ALIGNMENT>
        <ASSEMBLY>
           <STANDARD short_name="NCBI36-HG18_Broad_variant" />
        </ASSEMBLY>
        <RUN_LABELS>
          <RUN accession="SRR024521" read_group_label="A"/>
          <!-- etc -->
        </RUN_LABELS>
        <SEQ_LABELS>
          <SEQUENCE accession="NC_001807.4" seq_label="chrM"/>
          <!-- etc -->
        </SEQ_LABELS>
          <PROCESSING>
            <PIPELINE>
              <PIPE_SECTION section_name="Mapping">
                <STEP_INDEX>0</STEP_INDEX>
                <PREV_STEP_INDEX>NIL</PREV_STEP_INDEX>
                <PROGRAM>maq</PROGRAM>
```

```
                <VERSION>0.7.1-9</VERSION>
                <NOTES>/seq/software/picard/current/3rd_party/maq/maq map</NOTES>
              </PIPE_SECTION>
            </PIPELINE>
            <DIRECTIVES>
               <alignment_includes_failed_reads>true</alignment_includes_failed_reads>
               <alignment_includes_unaligned_reads>true</alignment_includes_unaligned_reads>
               <alignment_marks_duplicate_reads>true</alignment_marks_duplicate_reads>
            </DIRECTIVES>
          </PROCESSING>
      </REFERENCE_ALIGNMENT>
    </ANALYSIS_TYPE>
    <TARGETS>
      <TARGET sra_object_type="SAMPLE" accession="SRS004147"
              refcenter="NCBI" refname="TCGA-13-0751-10A-01D-0446-08" />
    </TARGETS>
          <DATA_BLOCK>
            <FILES>
              <FILE filename="TCGA-13-0751-10A-01D-0446-08_IlluminaGA-DNASeq_whole.bam"
                    filetype="bam" checksum_method="MD5"
                    checksum="94f9600fb59166d80da81020bf9d3a8f"/>
            </FILES>
          </DATA_BLOCK>
  </ANALYSIS>
</ANALYSIS_SET>
```

4.2 Reference Alignment of Existing Runs

This use case entails submission of a reference alignment using runs that have already been submitted to the SRA, and whose run accessions are already known. In this case the submission metadata includes ONLY the SRA analysis object (in addition to the SRA submission object).

4.3 Reference Alignment with de novo Runs

This use case covers submission of reference alignment where the payload of the alignment data implicitly contains the raw sequencing data that should be loaded into the SRA. This submission mode is not currently fully supported, but submitters can prepare submissions according to these instructions. In this case submission metadata must also include at least one SRA experiment and one or more SRA run objects (in addition to the SRA submission object), as well as SRA study and sample records if these do not already exist.

4.3.1 SRA Study and Samples

If the submitter is submitting BAM files against existing study and sample objects (the normal case for protected submissions), then the SRA study and sample ids need to be referenced. The reference can be by accession (SRP, SRS), or by refname and refcenter (see previous instructions). For new submissions to the open SRA the submitter will need to create study and sample objects as part of the submission.

4.3.2 Specify SRA Experiments

In addition to specifying SRA Analysis, the submitter must create metadata objects for SRA experiments. An SRA experiment corresponds to a library created for a platform. Generally there will be one SRA experiment per library (denoted by LB tag in the read group component

of the BAM file header). Also, there is generally one SRA experiment per sample (denoted by the SM tag in the read group component of the BAM file header).

In open SRA submissions the submitter should determine which existing SRA study and SRA sample records to reference, or create these as part of the submission. In protected SRA submissions, the parent SRA study and SRA sample records are prepared by dbGaP submission staff, and these must be referenced by the submitter rather than created.

4.3.3 Specify SRA Runs

Once SRA experiments have been created, the submitter must create SRA Run objects. These correspond to the run file or set of run files associated with a production unit of the sequencing instrument. For 454, the recommended production unit is plate or region. For Illumina, the recommended production unit is a lane. For SOLiD the recommended production unit is a slide. Other divisions are possible, and there is no strict rule that requires an SRA run to comprise a given sequencing production unit.

Where raw sequencing data are being delivered within the BAM file, it is necessary to create a mapping between the read group tag (**ID**) and the SRA run as referenced by its accession or namespace/alias (**CN**/**PU**).

Here is an example of how to encode the DATA_BLOCK descriptor for loading of a run from a BAM file. The same BAM file may be used as the source for multiple runs. This file is the same as that being specified in the analysis xml. Different data are being extracted from the BAM file for the respective archive resources.

```
<DATA_BLOCK name="30MWUAAXX081205" sector="3">
 <FILES>
  <FILE
   filetype="bam"
   filename=" TCGA-06-0188-01A-01D-0373-08_IlluminaGA-DNASeq_whole.bam"
   checksum="e6b8b7ee08e67434cf027c3255c1f633"
   checksum_method="MD5">
  </FILE>
 </FILES>
</DATA_BLOCK>
```

5 BAM File Format

The BAM file format is specified fairly loosely. Submitters need to comply with an additional policy layer for submissions to an archive to be possible.

5.1 BAM Read Inclusion Rules

- **Header included** – The optional BAM header should be included.

- **Center coherence** – Each BAM file should contain sequencing from at most one center.
- **Sample coherence** – Each BAM file should contain sequencing from at most one sample. Pooled sample runs that pertain to the same study should be demultiplexed into their respective sample fractions and delivered as separate BAM files.
- **Hard Clipping Disallowed** - Hard clipping would be disallowed for submissions that embed de novo sequencing data (rather than being submitted independently). The read's complete sequence must be supplied in order to recreate the original read.
- **Read group congruent to production run** - The read descriptor block must be used on submission to confer metadata on the collection of reads co-produced by the sequencing instrument (same half plate, same lane, same slide, etc).
- **Reconstructability** - The read's original name, base call vector, quality score vector must be reproducible. If the read has clip points computed by the instrument, these must be preserved.
- **Unmapped reads included** – The BAM file must contain all reads regardless of mapping status – i.e. all unmapped reads must be included in the file. When an unmapped read has a mate pair that is mapped then the unmapped read should be stored with the mate's reference sequence and position (with the read unmapped flag set to communicate its status). Unmapped reads with unmapped mate pairs must be stored after all mapped reads. If unmapped reads are not available, set the SRA analysis processing directive **alignment_includes_unaligned_reads = false.**
- **Duplicate reads marked** – The BAM file must contain all copies of duplicate reads. If duplicate reads have not been marked, set the SRA analysis processing directive **alignment_marks_duplicate_reads=false.**
- **Vendor quality filtered reads identified** - Files may contain both both passing and failing a vendor specific quality check. In the case that such "failing" reads are included they are to be stored using the same rules as any other read but **must** be marked by setting the "read fails platform/vendor quality checks" flag (flag bit 0x200). If reads failing vendor quality checks (for example non-PF-filtered reads), please set the SRA analysis processing directive **alignment_includes_failed_reads=false.**
- **Nonredundancy** - A single record per read should exist in the file – if multiple equally good alignments exist for a read one should be picked at random and a mapping quality of 0 assigned
- **Position order** – Reads must be contained in the BAM file by order of position. Specifically, sort order (**SO**) should be set to "coordindate".

5.2 BAM Read Annotation Rules

The BAM file should contain (at a minimum) the following annotation for each read stored:

1. The full, unclipped, set of sequenced bases as produced by the sequencing instrument. While the specification allows the use of "=" for a base that matches the reference, the submitter should always insert the actual bases and never = signs.
2. A set of base quality scores. Each center should communicate clearly what quality scores are being used, e.g. quality scores from Bustard, from Gerald, scores calibrated

using a custom pipeline etc. This information should be communicated via the SRA Run metadata as there is no current way to provide this info within the BAM format itself.
3. All flag fields correctly and fully set based on knowledge of the read
4. The reference sequence name (or index), alignment start position and CIGAR string for all mapped reads
5. The mate sequence name (or index) and alignment start position set of the mate pair record (regardless of its mapping status)
6. The inferred insert size in the case that both ends of a mate pair are mapped
7. A phred scaled mapping score representing the probability that the alignment is incorrect if the read is mapped. If the aligner used does not provide such a score, the value 255 should be used.
8. @RG: a read group attribute denoting which read group described in the header each read comes from

5.3 BAM Reference Annotation Rules

Currently only standard assemblies are supported. Assemblies based on GRCh-37 are preferred, and are usually required by funders. Retrospective assemblies based on NCBI Build 36/HG-18 have been cataloged and are indexed by name of the assembly and the submitting center. Labels used in the BAM file to denote a reference sequence should be the same label used in one of these assemblies. Each contigs must map to a current or inactive accession at one of the partner INSDC sequence databases. Assemblies cannot be mixed or added to.

A more flexible method of determining the reference assembly as a set of contigs from various sources including de novo assembly is being designed and will be described in future revisions of this document.

5.4 BAM Header

The text header block described in the SAM specification is optional for BAM files, but is required for submission. The header should consist of:

1. A single line starting @HD that indicates the BAM specification version number and the sort order of the file (coordinate/read).
2. Many lines starting @SQ. These lines are one per sequence in the reference assembly to which these reads were aligned. Each sequence should be named as in the reference assembly (for example, *chr1*). Only sequences available in that assembly may be used.
3. One record per sequencing production unit starting with @RG, to be configured as follows:

   ID: an arbitrary ID used to link reads back to the read group header
   PL: the sequencing platform that generated the reads.
   PU: the "platform unit" - a unique identifier which tells you what run/experiment created the data. For Illumina, please follow this convention: Illumina flowcell barcode suffixed with a period and the lane number (and further suffixed with

period followed by sample member name for pooled runs). If referencing an existing already archived run, then please use the run alias in the SRA.
LB: the unique identifier of the sequencing library that was sequenced. This should correspond to the SRA library name for already-archived runs.
DT: the run start date of the instrument run. Please use ISO-8601 format.
SM: the sample identifier. This should be the sample alias loaded in the SRA or in the metadata being submitted to the SRA.
CN: the sequencing center that produced the data (This should be the INSDC short name for the Center.)

4. One or more records starting with @PG that records the program invocation that created the alignment product.

5.5 Working with BAM

BAM file format is described in http://samtools.sourceforge.net/SAM1.pdf .

A BAM validator called ValidateSamFiles tool is available at:
http://picard.sourceforge.net/command-line-overview.shtml - ValidateSamFile

Duplicate reads can be marked using the Picard MarkDuplicates tool:
http://picard.sourceforge.net/command-line-overview.shtml - MarkDuplicates

[i] * denotes choice that will be available in SRA 1.2 or later