

SRA Submission Quick Start Guide

National Center for Biotechnology Information (NCBI)

National Library of Medicine

9 Feb 2010 Version 1.0 Draft A

The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Life Technologies AB SOLiD System®, Helicos Biosciences Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®. This Guide is intended for new and low volume submitters to the SRA.

Table of Contents

SRA Submission Quick Start Guide	1
1 Steps for SRA Submission	2
2 Login	3
3 Creating a New Submission	3
3.1 Submission Alias and Comment	3
4 Study.....	4
4.1 Creating a New Study.....	4
4.2 Describing a Study	4
5 Setting a Submission Release Date	5
6 Status	6
7 Sample	6
7.1 Creating Samples	6
7.2 Describing a Sample	6
8 Experiment.....	7
8.1 Creating Experiments.....	7
8.2 Describing an Experiment	7
• Meta Information.....	7
• Library.....	8
• Processing	8
9 Run	9
9.1 Creating Runs.....	9
9.2 Describing a Run	9
10 Data Transfer	10

1 Steps for SRA Submission

1. Gather Sequence Data Files
2. [Generate md5 checksums for the files](#)
3. Enter Metadata on SRA website
 - a. [Submission](#)
 - b. [Study](#)
 - c. [Sample](#)
 - d. [Experiment](#)
 - e. [Run](#)
4. [Transfer Data files to SRA](#)
5. Update Submission with [PubMed links](#), [Release Date](#), or Metadata Changes

Submissions Tracking Preferences							
Submission: SRA008281.25/foxa2							
Accession #	Submission #	Submitter	Updated	State	Status	Comments	
SRA008281.25	BCCAGSC : foxa2	BCCAGSC	2009-11-10 13:51	public		<ul style="list-style-type: none"> • SRP000660.1 : FoxA2 epigenetics • 10 samples • 17 experiments • 31 runs 	
Files							
Type	Accession #	Alias	Uploaded	Links	Files	Released	
STUDY	SRP000660.1	FoxA2 epigenetics	10 M	ok	done	2009-03-26 18:37:04	
SAMPLE	SRS002374.1	MM0325	10 M	ok	done	2009-11-10 13:51:56	
EXPERIMENT <input type="button" value="New Run"/>	SRX003293.1	MM0325_200GMAAXX	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014516.2	MM03251.	9 M	ok	done	2009-11-10 13:51:56	
EXPERIMENT <input type="button" value="New Run"/>	SRX003294.1	MM0325_2049EAAXX	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014515.2	MM03251..1	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014517.2	MM03251..3	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014518.2	MM03251..2	9 M	ok	done	2009-11-10 13:51:56	
SAMPLE	SRS002375.1	MM0261	10 M	ok	done	2009-11-10 13:51:56	
EXPERIMENT <input type="button" value="New Run"/>	SRX003295.1	MM0261_FC4315	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014501.2	MM02611.	9 M	ok	done	2009-11-10 13:51:56	
EXPERIMENT <input type="button" value="New Run"/>	SRX003296.1	MM0261_FC6331	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014502.2	MM02611..5	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014503.2	MM02611..4	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014504.2	MM02611..3	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014505.2	MM02611..2	9 M	ok	done	2009-11-10 13:51:56	
RUN	SRR014506.2	MM02611..1	9 M	ok	done	2009-11-10 13:51:56	

Figure 1 Example of a finished SRA Submission viewed from the Interactive Tool.

2 Login

From the [SRA Homepage](#):

Click the **Submit** tab.

Then click **NCBI PDA** (If you do not have a PDA account already, one will need to be created.)



Figure 2 From the 'Submit' tab, click NCBI PDA to login for Submission.

3 Creating a New Submission

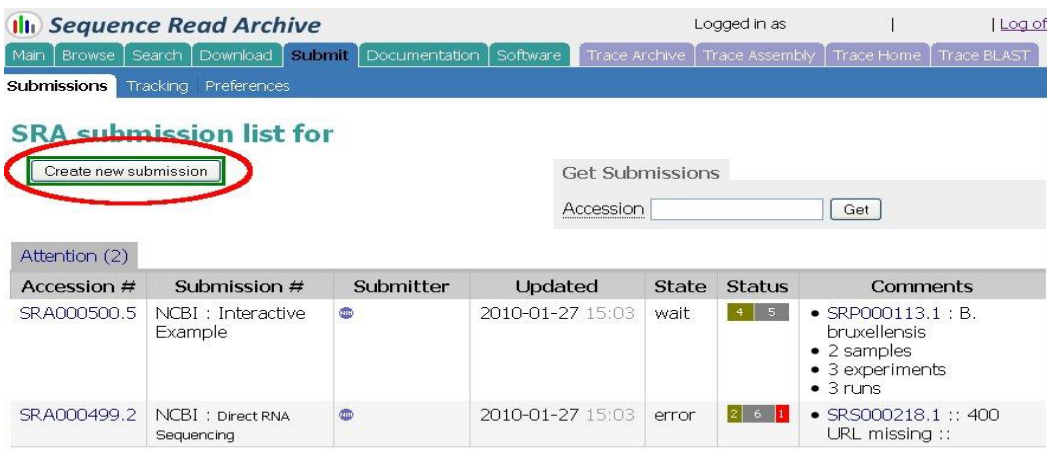


Figure 3 To start a Submission, click the 'Create new submission' button.

3.1 Submission Alias and Comment

Submission ID also known as '**Alias**' - will be used for tracking within the archive and for the submitter. This field should be something that makes sense to the submitter.

Example: *C. elegans resequencing project* (this field is NOT indexed in Entrez).

Submission Comment – area for submitter to enter a comment about the submission.

Example: *prepared with assistance by John Smith* (this field is NOT indexed in Entrez).

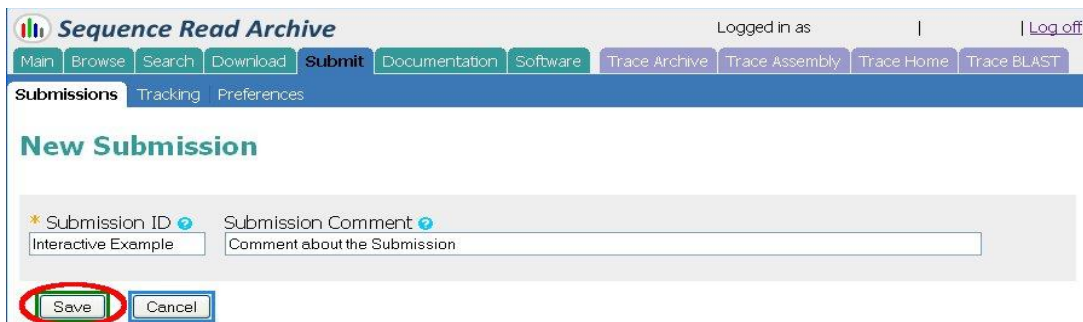


Figure 4 The Submission is not created until the 'Save' button is clicked.

4 Study

4.1 Creating a New Study

A Study identifies the sequencing study or project and may contain multiple experiments. Study accessions are a good choice as the reference in a scientific journal article because the published Study provides a good landing page for users seeking to download data.



Figure 5 Create the Study by clicking the button for 'New Study'.

4.2 Describing a Study

Alias- Used as a reference for the submitter and archive. (NOT an indexed field)

Title- Publicly viewable title. A title from a journal article or other descriptive title should be used.

Study Type- Drop-down menu providing a selection of different categories for sequencing projects. Used as a method for users to find general types of studies. Pick the closest category and avoid 'other' if possible.

Abstract- Describes the goals, purpose, and scope of the study.

Description- Allows for more extensive and free-form description of the study.

Project Name- Name used by submitter for the project, if different from the Study Title.

Project ID- [Genome Project ID](#). New Projects can be created [here](#).

Links and Attributes- Used to add URLs, Entrez Links, or other Attributes in a key-value pair configuration. If linking to other databases, please use the correct [database abbreviation](#).

- If the Study accompanies a journal article, enter the PubMed ID (pmid) as an 'entrez link' with "pubmed" as the 'DB' and the pmid as the 'ID'.

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed
18045790[uid] Search Clear

Display Settings: Summary Send to:

Database resources of the National Center for Biotechnology Information.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chebvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E.

Nucleic Acids Res. 2008 Jan;36(Database issue):D13-21. Epub 2007 Nov 27.

PMID: 18045790 [PubMed - indexed for MEDLINE]

Relevant articles: Free article

Figure 6 The PMID for an article is listed at the bottom of the article's PubMed Summary.

Submissions Tracking Preferences

Submission: SRA009008.10/Sewage Signatures

Study: SRP000905

Meta information

Alias: Sewage indicator signatures Title: Diversity and population structure of sewage derived microorganisms in wastewater treatment plant influent

Study type: Metagenomics

Abstract: The release of untreated sewage introduces non-indigenous microbial populations of uncertain composition into surface waters. We used massively parallel 454 sequencing of hypervariable regions in rRNA genes to profile microbial communities from eight untreated sewage influent samples of two wastewater treatment plants (WWTP) in metropolitan Milwaukee. The sewage profiles included a discernible human fecal signature made up of several taxonomic groups. Samples from Jones Island and South Shore WWTPs had very consistent community

Description: [Empty text box]

Project name: Sewage-derived bacteria in wastewater treatment plant influent Project ID: 0

Links and Attributes

URL link: Visualization and Analyt http://vamps.mbi.edu

URL link: Great Lakes WATER Ins http://www.gliwi.uwm.edu

NCBI Site map All databases PubMed Search

Sequence Read Archive

SRP000905 Diversity and population structure of sewage derived microorganisms in wastewater treatment plant influent

Study Type: Metagenomics

Submission: SRA009008 by Marine Biological La on 2009-06-28T21:39:31Z

Abstract: The release of untreated sewage introduces non-indigenous microbial populations of uncertain composition into surface waters. We used massively parallel 454 sequencing of hypervariable regions in rRNA genes to profile microbial communities from eight untreated sewage influent samples of two wastewater treatment plants (WWTP) in metropolitan Milwaukee. The sewage profiles included a discernible human fecal signature made up of several taxonomic groups. Samples from Jones Island and South Shore WWTPs had very consistent community profiles. Overall, the sewer system appears to be a defined environment with both infiltration of rainwater and stormwater inputs modulating community composition. Microbial sewage communities represent a combination of inputs from human fecal microbes and enrichment of specific microbes from the environment to form a unique population structure.

Description: [Empty text box]

Properties: External Links: Visualization and Analysis of Microbial Population Structures (VANES) Great Lakes WATER Institute

Download fastq for entire study

Experiments

Accession	Spots	Bases
Total: 9	279,757	34.7M
SRX005900	31,665	4.0M
SRX005901	36,879	4.7M
SRX005902	27,226	3.4M
SRX005903	34,017	4.3M
SRX005904	29,248	3.7M
SRX005905	35,434	4.5M
SRX005906	19,368	2.5M
SRX005907	23,722	3.0M
SRX005908	42,198	4.6M

Write to the Help Desk | Contact Us | Disclaimer | Accessibility
National Center for Biotechnology Information | U.S. National Library of Medicine

Last update: Tue, 09 Feb 2010 Rev. 103054

Figure 7 Example showing the SRA Interactive Tool (left) compared to the view of the same Study in Entrez (right). Above Study in Entrez

5 Setting a Submission Release Date

A release date is required for all submissions. It is advisable to enter a release date before loading any data into a Submission. This will prevent accidental early release of data. Dates may be set for up to one year in the future in anticipation of a publication release date.

Submissions Tracking Preferences

Submission: SRA000500.2/Interactive Example

Accession #	Submission #	Submitter	Updated	State	Status	Comments
SRA000500.2 (as Admin)	NCBI : Interactive Example	Adam Stine	2009-10-09 10:24	completed	i	• SRP000113.1 : B. bruxellensis

Files

Type	Accession #	Alias	Uploaded	Links	Files	Released
STUDY	SRP000113.1	B. bruxellensis	1 m	ok	done	

New Sample New Experiment

Release hold release to: (YYYY-MM-DD)

Figure 8 To save the date on which the submission is scheduled to be published/released to the public, enter a date in the box using a YYYY-MM-DD format, then click 'Release' The release date can be changed as long as the submission has not yet been made public

6 Status



Figure 9 The Status Bar

The status bar provides a visual representation of the current state of the submission and files in Tracking. **Done** (Dark Green) indicates the number of completed objects. **Wait** (Gray) further information or files uploads are needed. **Processing** (Light Blue- not shown) an object is currently being processed, if an object/file is processing for more than 48 hours, [contact SRA](#). **Queue** (Dark Blue) the object will be processing when the pipeline is available. **Replaced** (Bright Green) an object/file was replaced by another. **Error** (Red) intervention is required, please [contact SRA](#).

7 Sample

7.1 Creating Samples

Each unique sample used in a study needs to have its own sample object within the submission. The exception to that rule is when samples were intentionally pooled. A pooled sample is one sample but should explicitly describe as much as is known about what was in the pool. Please [contact SRA](#) for help with barcoded or indexed samples. A Study may contain multiple Samples.

Submissions Tracking Preferences

Submission: SRA000500.2/Interactive Example

Accession #	Submission #	Submitter	Updated	State	Status	Comments
SRA000500.2 (as Admin)	NCBI : Interactive Example	Adam Stine	2009-10-09 10:24	completed	1	• SRP000113.1 : B. bruxellensis

Files

Type	Accession #	Alias	Uploaded	Links	Files	Released
STUDY	SRP000113.1	B. bruxellensis	1 m	ok	done	

hold release to: (YYYY-MM-DD)

Figure 10 Click the 'New Sample' button to create a new Sample.

7.2 Describing a Sample

Alias- Used as a reference for the user and archive. (NOT an indexed field)

Title-Publicly viewable title. A formal title used to describe the Sample. If the submission goes along with a journal publication, the title should distinguish samples within the article.

Anonymized Name- Anonymous public name of the sample. For example, *HapMap human isolate NA12878*.

NCBI Taxon ID- ID number from the [NCBI Taxonomy database](#). For samples that do not have an appropriate Taxonomy entry, the submitter will need to apply for a new or provisional Taxon ID. [Email SRA](#) for assistance to establish new or provisional Taxon IDs.

Description- Allows for more extensive and detailed description of the sample.

Links and Attributes- Used to add URLs, Entrez Links, or other Attributes in a key-value pair configuration. If linking to other databases, please use the correct [database abbreviation](#).

Submissions Tracking Preferences

Submission: SRA000500.2/Interactive Example

Sample : SRS000220

Name and description

*Alias: England Barrel Title: Brettanomyces bruxellensis- England Barrel Anonymized name:

*NCBI Taxon ID: 13366 Look up Search via Entrez

Description: Culture taken from a 'Burton Union' in England. Culture grown in a lab to acquire a clone of Brettanomyces bruxellensis.

Links and Attributes

Add Save Back

Figure 11 Samples should be fully described such that a user does not need to find an accompanying publication. The information is not stored until the 'Save' button is clicked. Saved samples can be updated as necessary.

8 Experiment

8.1 Creating Experiments

Experiment describes the library, platform selection, and processing parameters. Each change to the library or sequencer parameters requires the creation of a new experiment. A Sample can contain multiple experiments but each experiment contains only one library.

Submission: SRA000500.2/Interactive Example

Accession #	Submission #	Submitter	Updated	State	Status	Comments
SRA000500.2 (as Admin)	NCBI : Interactive Example	Adam Stine	2009-11-03 11:09	queue	1 2 1	<ul style="list-style-type: none"> SRP000113.1 : B. bruxellensis SRS000219.1 : Belgium SRS000220.1 : England Barrel SRX000199.1 : English Barrel

Files

Type	Accession #	Alias	Uploaded	Links	Files	Released
SAMPLE	SRS000219.1	Belgium	3 W	queue	done	
STUDY	SRP000113.1	B. bruxellensis	3 W	ok	done	2010-04-11 00:00:00
SAMPLE	SRS000220.1	England Barrel	3 W	queue	done	
EXPERIMENT <input type="button" value="New Run"/>	SRX000199.1	English Barrel	0 m	ok	wait	

hold release to: (YYYY-MM-DD)

Figure 12 Click the 'New Experiment' button to begin creating an Experiment.

8.2 Describing an Experiment

- **Meta Information**

Platform- This describes the sequencing platform used in the experiment.

Alias- Used as a reference for the user and archive. (NOT an indexed field)

Title- A publicly viewable and formal title used to describe the Experiment.

Study Accession- Links this Experiment to a previously created Study

Sample Accession- Links this Experiment to a previously created Sample
Design Description- Describes the setup and goals of this Experiment

- **Library**

Name- Name of the Library that was sequenced

Strategy- Sequencing strategy used in the experiment

Source- Type of genetic source material sequenced

Selection- Method of selection or enrichment used in the Experiment

Layout- Configuration of the read layout. Paired, Fragment, etc.

Nominal Size (paired)- Size of the insert for Paired reads. (Required)

Nominal Standard Deviation (paired)- Standard deviation of insert size (typically ~10% of Nominal Size)

Library Construction Protocol- An area to give a description on the library construction techniques and reagents used.(Required)

- **Processing**

This section varies with the sequencer selected. Please pay close attention to the answers provided in this section, as they may affect proper loading of data.

Links and Attributes- Used to add URLs, Entrez Links, or other Attributes in a key-value pair configuration. If linking to other databases, please use the correct [database abbreviation](#).

Submission: SRA000500.2/Interactive Example

Experiment

The screenshot shows a web form for creating an experiment. It is divided into several sections:

- Meta information:** Includes fields for Platform (Illumina Genome Analyzer II), Alias (English Barrel), Title (Burton Union culture sample), Study accession (SRP001113), and Sample accession (SRS000220 (England Barrel)). There is a Design description field containing the text "De novo sequencing of B. bruxellensis".
- Library:** Includes fields for Name (UK-1), Strategy (WGS), Source (GENOMIC), Selection (RANDOM), Layout (PAIRED), Nominal size (bp), and Nominal standard deviation (bp) (0).
- Library Construction protocol:** A text area containing detailed instructions: "Illumina Paired-End DNA Sample Prep Kit (PE-102-1001) is used to build DNA libraries with insert sizes from 200-500 bp for paired-end sequencing. The kit provides reagents for repairing the ends of DNA that have been fragmented by sonication. Ends are repaired with a combination of fill-in reactions and exonuclease activity to produce blunt ends. An '&'- base is added to the blunt ends followed by ligation to Illumina Paired-End Sequencing adapters. These".
- Processing:** Includes fields for Planned read length (bp) for mate 1 (35) and mate 2 (35).
- Links and Attributes:** A section with an "Add" button and a "Save" button (circled in red) and a "Back" button.

Figure 13 Click the 'Save' button to store the Experiment information. Saved Experiments can be updated as necessary.

9 Run

9.1 Creating Runs

Runs describe the files that belong to the previously created Experiments. Runs are divided by production run of the sequencer. Experiments may contain many Runs depending on how many sequencer runs were involved in data acquisition.

Accession #	Submission #	Submitter	Updated	State	Status	Comments
SRA000500.2 (as Admin)	NCBI : Interactive Example	Adam Stine	2009-11-03 11:09	queue	1 2 1	<ul style="list-style-type: none">SRP000113.1 : B. bruxellensisSRS000219.1 : BelgiumSRS000220.1 : England BarrelSRX000199.1 : English Barrel

Type	Accession #	Alias	Uploaded	Links	Files	Released
SAMPLE	SRS000219.1	Belgium	2 M	queue	done	
STUDY	SRP000113.1	B. bruxellensis	2 M	ok	done	2010-04-11 00:00:00
SAMPLE	SRS000220.1	England Barrel	2 M	queue	done	
EXPERIMENT	SRX000199.1	English Barrel	1 M	ok	wait	

New Sample New Experiment

Release hold release to: (YYYY-MM-DD)

Figure 14 Click the 'New Run' button to the right of the Experiment for which a Run is needed. Each Experiment will have its own 'New Run' button.

9.2 Describing a Run

Alias- Used as a reference for the user and archive. (NOT an indexed field)

Run data file type- The storage format (srf, sff, fastq, etc) of the sequence data being submitted. More information about the file types currently accepted by the SRA can be found in the [SRA Submission Guidelines](#).

File Name- Name of the file transferred to the SRA including any file extensions. Some data types require multiple files to be combined into a tar archive and some require single files.

MD5 checksum- A checksum or hash sum generated for the file listed in 'File Name' that is used to detect errors introduced through storage or transfer. SRA uses the file name and md5 checksum to track and link files to their proper Runs.

Unix- md5sum <file>

OS X- md5 <file>

Windows- Application required. [Fsum Frontend](#) and [WinMD5Sum](#) are two possible options.

Plate and Region- Only seen on certain file types like FASTQ. Because some file types have limited addressing information, these fields allow the user to provide the address information for the sequencing media used.

Submission: SRA000500.2/ Interactive Example

Experiment: SRX000199/English Barrel

Run

General info

*Alias [?](#)
FC105

*Run data file type [?](#) srf

Data blocks [?](#)

*File name ?	*MD5 checksum ?	
FC502W9AAXX.8.srf	0cf9194c1bb4bf91371096585e536bfe	Delete

Add

Links [?](#) and Attributes [?](#)

Add

Save Back

Figure 15 Click the 'Save' button to store the Run information. Runs can only be updated until data has been loaded for the Run. Once there is data in a Run, it will be locked from further updates. Contact SRA for changes to be made to locked Runs.

10 Data Transfer

After the metadata is entered, you may upload data to the SRA.

Upload via FTP:

ftp://sra:61c6lwoE6!@ftp-private.ncbi.nih.gov/
(Windows Explorer or an FTP client may be used)

[FileZilla](#) is one of many free FTP clients.

Or from unix/linux

Address: ftp-private.ncbi.nih.gov

Login: sra

Password: 61c6lwoE6! ([contact SRA](#) for the current password)

If everything is correct files will be linked and loaded automatically.

Additional information on data transfer methods is available in the [SRA Submission Guidelines](#).

Please write to sra@ncbi.nlm.nih.gov for answers to submission questions.