CyVerse Documentation

CyVerse

Feb 18, 2021

Contents

|CyVerse_logo|_

 Learning Center Home

Goal

Import data from the NCBI Sequence Read Archive into your data store (SRA) via the Discovery Environment

The NCBI Sequence Read Archive (SRA) is a repository for high-throughput sequencing data. You can import data from the SRA into your Data Store using the Discovery Environment SRA-Import App.

Tip: According to the SRA homepage: "Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®."

Prerequisites

2.1 Downloads, access, and services

In order to complete this tutorial you will need access to the following services/software

Prerequisite	Preparation/Notes	Link/Download
CyVerse account	You will need a CyVerse account to complete this exercise	Register

2.2 Platform(s)

We will use the following CyVerse platform(s):

Platform	Interface	Link	Platform Documentation	Quick Start
Discovery Environment	Web/Point-and-click	Discovery Environment	DE Manual	Guide

2.3 Input and example data

In order to complete this quickstart you will need to have the following inputs prepared

Input File(s)	Format	Preparation/Notes	Example Data
SRA Accession number	N/A	We will cover how to search for an accession	In this example, we will download accession SRR1761506

Get started

Note: **Searching the SRA:** Searching the SRA can be complicated. Often a paper or reference will specify the accession number(s) connected to a dataset. You can search flexibly using a number of terms (such as the organism name) or the filters (e.g. DNA vs. RNA). The SRA Help Manual provides several useful explanations. It is important to know is that projects are organized and related at several levels, and some important terms include:

- **Bioproject**: A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium of coordinating organizations; see for example Bio Project 272719

- **Bio Sample**: A description of the source materials for a project

- **Run**: These are the actual sequencing runs (usually starting with SRR); see for example SRR1761506

1. Obtain an SRA accession number (starting SRR***); If you do not have an accession, you can go to the SRA homepage and search using a variety of search terms and filters (e.g. DNA vs. RNA, exome vs. genome, etc.)

Note: On the SRA homepage for each accession, you may wish to record some useful information about the run, including the sequencing format and the file size.

2. Log in to the Discovery Environment and click on NCBI-SRA-Fastq-dump-2.8.1 App or clcik on **Apps** to search for an launch this App.

3. Name your analysis and enter any desired comments

4. Under "Inputs" enter the SRA accession run number (if you have already downloaded an SRA file you can use this App to decompress it into a fastq file - search for the file using the 'Browse' button)

Tip: Depending on the file size, this will take several minutes

5. (optional) Under "optional parameters" check 'Split files' if your data are paried-end

---

Tip: The SRA page for your run should indicate 'SINGLE' or 'PAIRED' under Library Layout; https://www.ncbi.nlm.nih.gov/sra/?term=SRR1761506

---

6. (optional) Under "Output" enter a custom name for 'Sra output folder name' or leave the default

7. Click **Launch Analysis**

8. To view the status of the import and obtain results click on the **Analysis** icon

9. When the job status is marked 'Completed' in the Analysis window (you may have to refresh), click on the job name (e.g. 'SRA-Import-0.1.0_analysis1') to view the result in your data store

---

3.1 Summary

In addition to a folder of logs you should have the following files: - A compressed file (including sequence data and metadata) in the NCBI ".sra" format - An output folder (default:'sra_out') containing your fastq file (sequence data). If paired-end, and the 'Split files' option was checked, you will have two .fastq files (_1 for left-reads, _2 for right reads).

Next Steps:

Some common next steps include

1. Using FastQC to check the quality of the sequence reads

2. Using Trimmomatic to filter and trim reads for quality control

Both of these applications are available for use in the Discovery Environment. See DE Apps catalog

---

3.2 Additional information, help

Fix or improve this documentation

- Search for an answer: **|CyVerse Learning Center|**

- Ask us for help: click **|Intercom|** on the lower right-hand side of the page

- Report an issue or submit a change: **|Github Repo Link|**

- Send feedback: Tutorials@CyVerse.org

Learning Center Home