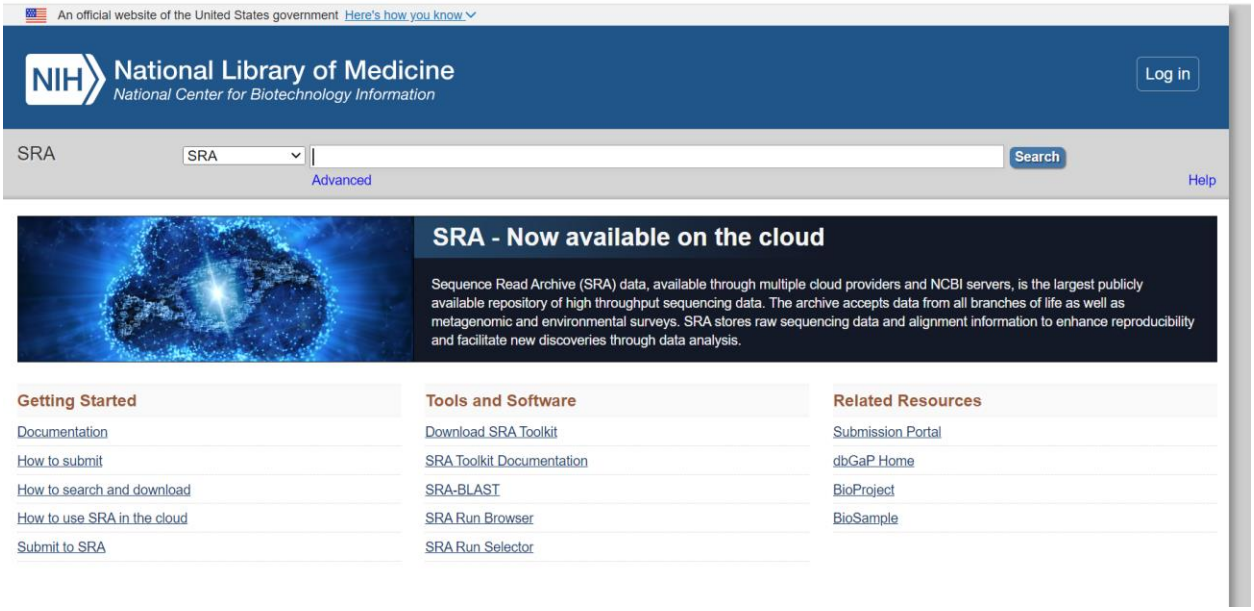
Tutorial: How to Find Sequence Read Datasets for Read Cleaning

To perform read cleaning on sequence data, you'll need to obtain appropriate datasets. Follow these steps to find and download sequencing reads from the NCBI Sequence Read Archive (SRA):

- **Visit the NCBI SRA Website**

Navigate to the NCBI SRA homepage:

https://www.ncbi.nlm.nih.gov/sra/



- **Access the Search Interface**

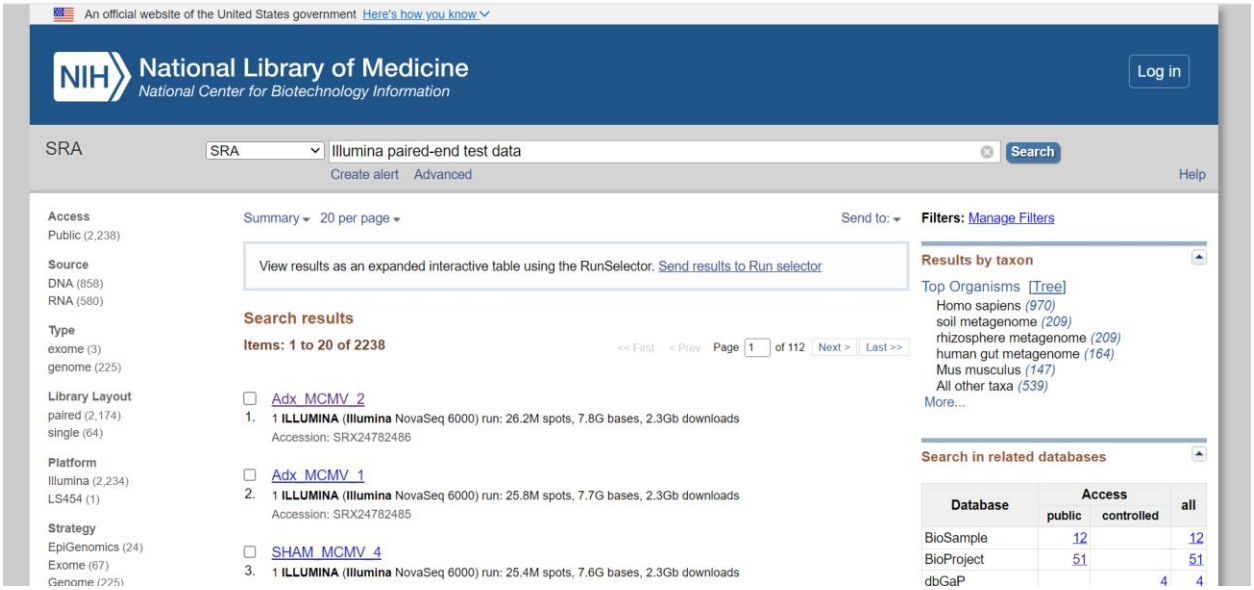
On the homepage, you will see the main search bar where you can input your query.

- **Search for Illumina Paired-End Test Data**

Then click on the **Search** button.

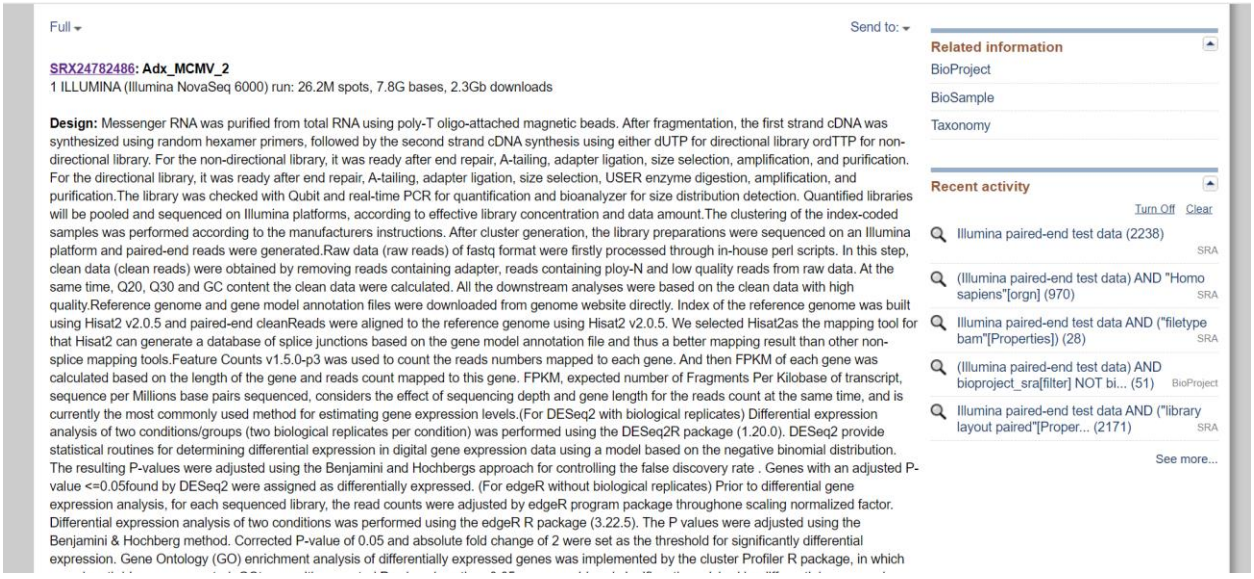
Review the Search Results

The search will return a list of datasets related to Illumina paired-end test data.



Select a Suitable Dataset

Browse through the available datasets and choose one that fits your project's requirements and your computational resource once you have selected a dataset. Scroll down on the page to find the SRR number so it can be used by fastqc to dump in the terminal:



Locate the SRR Number

Scroll down on the dataset's page until you find the **SRR (Sequence Read Archive Run) number**. This unique identifier is crucial for downloading the data.

1. **Use the SRR Number with FastQC**
2. With the SRR number, you can use tools like **`fastq-dump`** or **`prefetch`** to download the sequencing data via the terminal. Then, utilize **FastQC** for quality control and read cleaning.
3. By following these steps, you can efficiently locate and obtain the sequence reads necessary for your read cleaning tasks.