

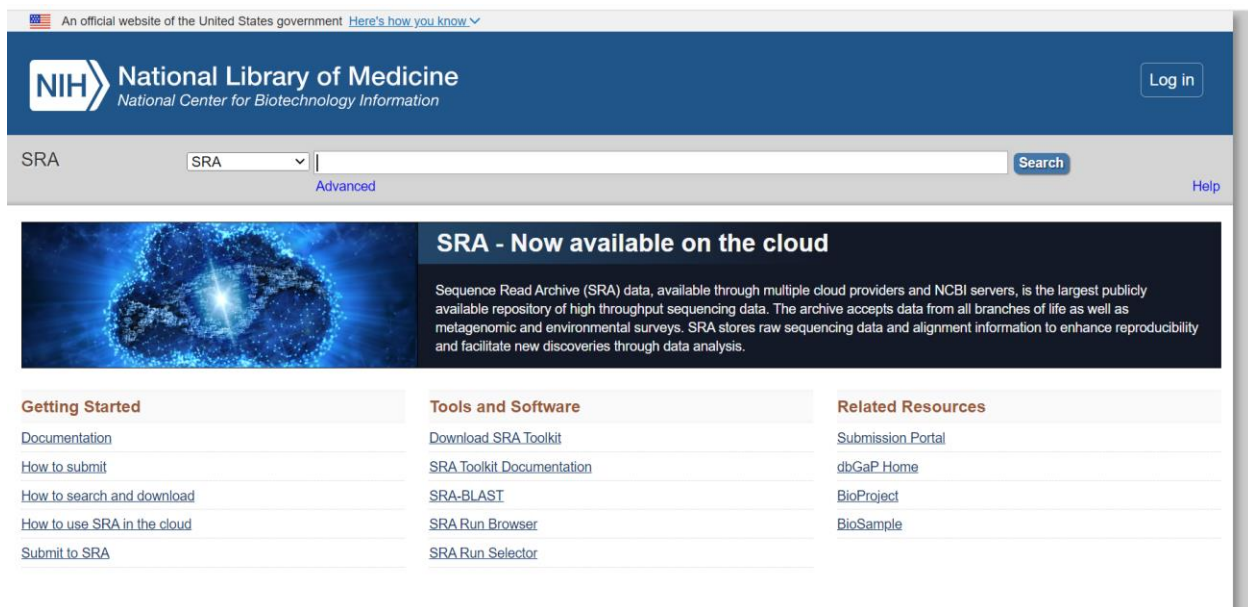
## Tutorial: How to Find Sequence Read Datasets for Read Cleaning

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.

An official website of the United States government [Here's how you know](#)

**NIH** National Library of Medicine  
National Center for Biotechnology Information

Log in

SRA

Create alert Advanced Help

Access  
Public (2,238)

Source  
DNA (858)  
RNA (580)

Type  
exome (3)  
genome (225)

Library Layout  
paired (2,174)  
single (64)

Platform  
Illumina (2,234)  
LS454 (1)

Strategy  
EpiGenomics (24)  
Exome (67)  
Genome (225)

Summary 20 per page

Send to: Filters: [Manage Filters](#)

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

**Search results**  
Items: 1 to 20 of 2238

1. ☐ [Adx\\_MCMV\\_2](#)  
1 ILLUMINA (Illumina NovaSeq 6000) run: 26.2M spots, 7.8G bases, 2.3Gb downloads  
Accession: SRX24782486

2. ☐ [Adx\\_MCMV\\_1](#)  
1 ILLUMINA (Illumina NovaSeq 6000) run: 25.8M spots, 7.7G bases, 2.3Gb downloads  
Accession: SRX24782485

3. ☐ [SHAM\\_MCMV\\_4](#)  
1 ILLUMINA (Illumina NovaSeq 6000) run: 25.4M spots, 7.6G bases, 2.3Gb downloads

Results by taxon

Top Organisms [Tree](#)

Homo sapiens (970)  
soil metagenome (209)  
rhizosphere metagenome (209)  
human gut metagenome (164)  
Mus musculus (147)  
All other taxa (539)  
[More...](#)

Search in related databases

Database	Access		all
	public	controlled	
BioSample	12		12
BioProject	51		51
dbGaP		4	4

## 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:

Full

**SRX24782486: Adx\_MCMV\_2**  
1 ILLUMINA (Illumina NovaSeq 6000) run: 26.2M spots, 7.8G bases, 2.3Gb downloads

**Design:** Messenger RNA was purified from total RNA using poly-T oligo-attached magnetic beads. After fragmentation, the first strand cDNA was synthesized using random hexamer primers, followed by the second strand cDNA synthesis using either dUTP for directional library or dTTP for non-directional library. For the non-directional library, it was ready after end repair, A-tailing, adapter ligation, size selection, amplification, and purification. For the directional library, it was ready after end repair, A-tailing, adapter ligation, size selection, USER enzyme digestion, amplification, and purification. The library was checked with Qubit and real-time PCR for quantification and bioanalyzer for size distribution detection. Quantified libraries will be pooled and sequenced on Illumina platforms, according to effective library concentration and data amount. The clustering of the index-coded samples was performed according to the manufacturers instructions. After cluster generation, the library preparations were sequenced on an Illumina platform and paired-end reads were generated. Raw data (raw reads) of fastq format were firstly processed through in-house perl scripts. In this step, clean data (clean reads) were obtained by removing reads containing adapter, reads containing ploy-N and low quality reads from raw data. At the same time, Q20, Q30 and GC content the clean data were calculated. All the downstream analyses were based on the clean data with high quality. Reference genome and gene model annotation files were downloaded from genome website directly. Index of the reference genome was built using Hisat2 v2.0.5 and paired-end clean reads were aligned to the reference genome using Hisat2 v2.0.5. We selected Hisat2 as the mapping tool for that Hisat2 can generate a database of splice junctions based on the gene model annotation file and thus a better mapping result than other non-splice mapping tools. Feature Counts v1.5.0-p3 was used to count the reads numbers mapped to each gene. And then FPKM of each gene was calculated based on the length of the gene and reads count mapped to this gene. FPKM, expected number of Fragments Per Kilobase of transcript, sequence per Millions base pairs sequenced, considers the effect of sequencing depth and gene length for the reads count at the same time, and is currently the most commonly used method for estimating gene expression levels. (For DESeq2 with biological replicates) Differential expression analysis of two conditions/groups (two biological replicates per condition) was performed using the DESeq2R package (1.20.0). DESeq2 provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting P-values were adjusted using the Benjamini and Hochberg approach for controlling the false discovery rate. Genes with an adjusted P-value <= 0.05 found by DESeq2 were assigned as differentially expressed. (For edgeR without biological replicates) Prior to differential gene expression analysis, for each sequenced library, the read counts were adjusted by edgeR program package through scaling normalized factor. Differential expression analysis of two conditions was performed using the edgeR R package (3.22.5). The P values were adjusted using the Benjamini & Hochberg method. Corrected P-value of 0.05 and absolute fold change of 2 were set as the threshold for significantly differential expression. Gene Ontology (GO) enrichment analysis of differentially expressed genes was implemented by the cluster Profiler R package, in which

Related information

BioProject  
BioSample  
Taxonomy

Recent activity

Turn Off Clear

1 Illumina paired-end test data (2238) SRA

2 (Illumina paired-end test data) AND "Homo sapiens"[orgn] (970) SRA

3 Illumina paired-end test data AND ("filetype bam"[Properties]) (28) SRA

4 (Illumina paired-end test data) AND bioproject\_sra[filter] NOT bi... (51) BioProject

5 Illumina paired-end test data AND ("library layout paired"[Proper... (2171) SRA

[See more...](#)

## 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.

Submitted by: University of California, Davis

Study: Mouse RNAseq - Thymus, SHAM vs Adx Mice w/ MCMV  
[PRJNA1117947](#) • [SRP511375](#) • [All experiments](#) • [All runs](#)  
[show Abstract](#)

Sample: Adx\_MCMV 3  
[SAMN41592519](#) • [SRS21500369](#) • [All experiments](#) • [All runs](#)  
[Organism](#): [Mus musculus](#)

Library:  
[Name](#): CRRA210030415-1A  
[Instrument](#): Illumina NovaSeq 6000  
[Strategy](#): RNA-Seq  
[Source](#): TRANSCRIPTOMIC  
[Selection](#): Oligo-dT  
[Layout](#): PAIRED

Runs: 1 run, 26.2M spots, 7.8G bases, [2.3Gb](#)

Run	# of Spots	# of Bases	Size	Published
<a href="#">SRR29264829</a>	26,158,698	7.8G	2.3Gb	2024-06-03

ID: 33107824

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.