srahunter: A User-Friendly Tool for Efficient Retrieval and Management of SRA Data

Abstract

Easy access and use of vast datasets are paramount for advancing scientific discovery in steadily expanding studies based on high-throughput sequencing (HTS). The Sequence Read Archive (SRA), now part of the International Nucleotide Sequence Database Collaboration (INSDC), is a public repository of reference for HTS reads and makes available an amazing sheer volume of information. Yet, researchers face significant challenges in efficiently accessing, downloading, and managing data and metadata. Here, we introduce *srahunter*, a tool designed to simplify data and metadata acquisition from SRA. Developed with Python and Bash, *srahunter* leverages the core functionalities of *SRA Toolkit* and *Entrez Direct*, to enable automated downloading, smart data management, and user-friendly metadata integration through an interactive HTML table. Compared with existing tools, *srahunter* increases the efficiency of metadata retrieval. *srahunter* supports HTS-based research by streamlined handling of SRA datasets, reducing the technical barriers to SRA data, *srahunter* can accelerate the development of genomics and multiple 'omics research.

Introduction

In the fast-evolving field of high-throughput sequencing (HTS) research, the availability of high-quality data is crucial for testing scientific hypotheses and driving new findings. The Sequence Read Archive (SRA), established by the National Center for Biotechnology Information (NCBI), is the largest public repository of high-throughput sequencing data (Leinonen et al., 2011; Kodama et al., 2012; Katz et al., 2022). Genomic and transcriptomic information is invaluable for researchers aiming to understand biological processes and interactions across the Tree of Life. However, efficient access, download, and management of sequencing data and metadata from the SRA poses significant challenges, primarily due to the overall amount of archived data (Leinonen et al., 2011). In 2023 the SRA database hosted more than 25.6 petabases from 14.8 million publicly available samples, a clear indication of the rapid and exponential growth of HTS-based research (Edgar et al., 2022; Katz et al., 2022).

NCBI *Entrez direc*t (Kans, 2024) and *Sra Toolkit* (https://github.com/ncbi/sra-tools) are command-line tools developed by the SRA staff that allow the download of data and metadata from SRA. Both tools contain a set of options that can be difficult to navigate. Indeed, setting all parameters and identifying the best way to process the data requires time and a medium level of coding capabilities, especially for an extensive list of accessions.

Recognizing these challenges, we have designed a tool to simplify data and metadata acquisition processes from the NCBI SRA. *srahunter* aims to address the critical bottlenecks in data retrieval and management, enabling researchers to streamline data handling. Features such as automated download, intelligent data management, and user-friendly metadata integration, including an interactive HTML table as standard output, can significantly simplify the workflow in HTS research.

Implementation

srahunter is written in Python, with the core functionalities based on *Sra toolkit* and *Entrez direct*. *srahunter* includes three main modules: the download module helps to get FASTQ sequencing data files from SRA, the metadata module allows the download of main run-associated metadata from SRA and also provides an interactive HTML output, whereas the full metadata module allows the download of all the run-associated metadata. Converting the SRA to FASTQ files can be parallelized across multiple cpu-cores. Storage availability is constantly checked during downloads, and all processes can be visually monitored through terminal messages and progress bars.

Input, flexible search, and output

srahunter takes as input an SRA accession list and can successfully download datasets produced by several sequencing platforms, including the widely used Illumina, PACBio, DNBSEQ, and ONT.

The download module automatically removes SRA files after successful conversion into FASTQ, data caching, download resuming, and error reporting in case of failed downloads (failed_list.csv file). Users can provide input parameters such as number of threads (default: 6), download directory (default: tmp_srahunter folder in the current directory), final output path (default: current directory), and maximum SRA file size (default 50 GB).

The metadata module features fast metadata retrieval, integrating the *Entrez direct* functionality and customizable output in CSV and HTML format. The tool supports an input list containing Run accession IDs; Bioproject accessions are not supported.

The fullmetadata module allows to obtain all the associated metadata from a list of Run accession numbers, exploiting the Entrez direct functionality, and parsing the XML output. In this case, the final output is represented by a CSV table.

The HTML table improves the output readability

In the case of the metadata module as default output, an interactive HTML table will be produced in addition to the classical CSV file. This interactive table is produced with the *datavzrd* tool, exploiting different features to facilitate the understanding of the downloaded metadata (Figure 1a). For each column, by clicking on the magnifying glass, it is possible to filter the data, whereas by clicking on the depicted chart, it is possible to produce a chart summarising the distribution of the column data, and this feature is available for both categorical and numeric variables (Figure 1b and Figure 1c).

Figure 1. *HTML output provided by srahunter. a) An example of the srahunter HTML table is provided, close to each column header a magnifying glass and a schematic chart allow to filter and to produce a bar plot of the data distribution respectively. The bar plot can be produced for both categorical variables (e.g., platform used to sequence the sample; panel b) and numeric variables (e.g., average length of the produced reads; panel c)*

Comparison with existing tools

For the metadata and fullmetadata modules, we compared the speed of metadata retrieving with two popular tools, namely *Kingfisher* and *ffq*. Runtime depends on software, environment, connection speed, and input. The *ffq*, *Kingfisher*, and *srahunter* tools were compared by recording the time needed for retrieving metadata of 1, 10, 100,1000, and 10000 samples using the bash time function. The results show that *Kingfisher* and *srahunter* have a significantly lower data retrieval time than ffq (Figure 2a) with no significant differences between the *srahunter* metadata and fullmetadata modules. To better highlight the possible differences between the three tested tools, we compared the median time to compute one sample, which is lower for the *srahunter* metadata module than *Kingfisher*, with a value of 0.038s and 0.05s respectively, whereas in the case of fullmetadata module the time per sample is higher than *Kingfisher* (0.06s) (Figure 2b).

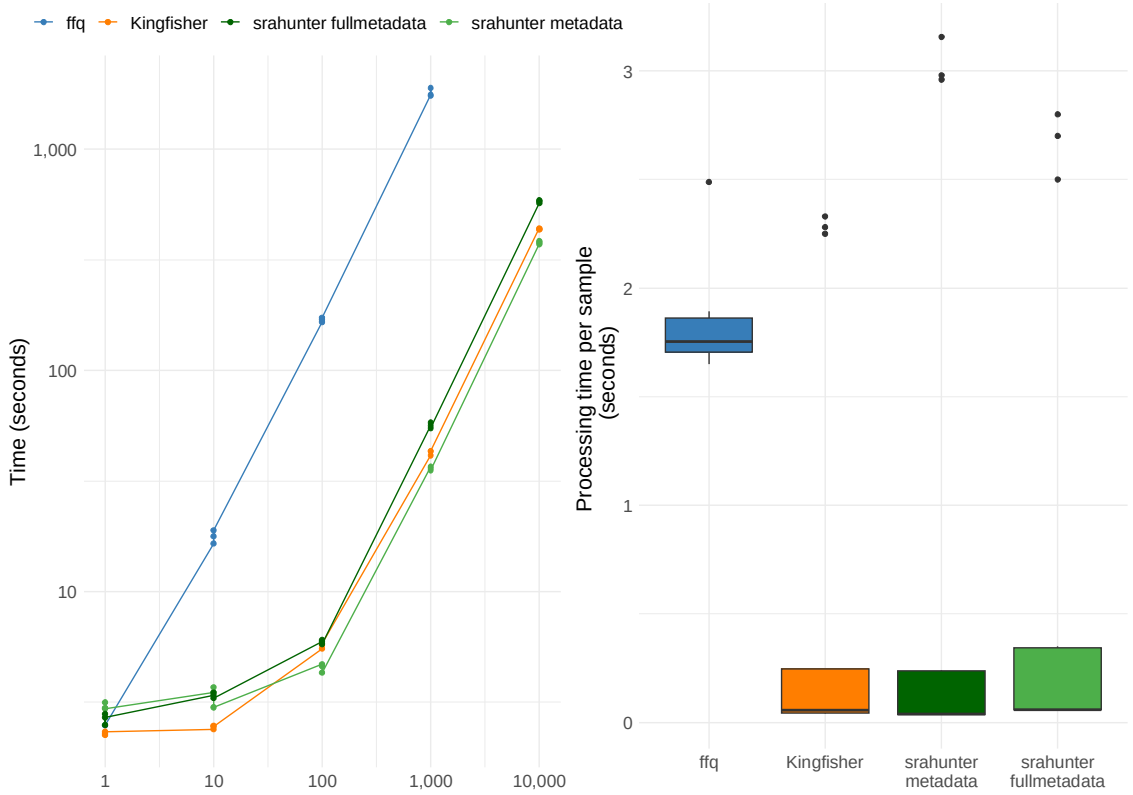

Figure 2. *Performance of three different tools in the retrieval of SRA metadata. In the figure panels, ffq, Kingfisher, and srahunter are depicted in blue, orange, and green, respectively. a) The line plot indicates the average time needed to obtain SRA metadata of 1, 10, 100, 1000, and 10000 samples with ffq, Kingfisher, and srahunter. b) The boxplot shows the average time required to analyze one sample between Kingfisher and srahunter*