Searching through the SRA (https://www.searchsra.org/)
A focus on the ECSS work

Mats Rynge (ECSS)
Eroma Abeysinghe (ECSS)
Kyle Levi
Robert Edwards (PI)
Searching the Sequence Read Archive using Jetstream and Wrangler

Kyle Levi
San Diego State University - Biological and Medical Informatics Program
San Diego, California, United States of America
levy@sdstate.edu

Erema Abyegunde
Indiana University – Science Gateway Research Center
Bloomington, Indiana, United States of America
erema@indiana.edu

Abstract

The Sequence Read Archive (SRA), the world’s largest database of sequences, hosts approximately 1.5 petabytes (1.5E15) of fragmentary data and is growing at the alarming rate of 18 TB per day. Yet
this vast trove of data is inaccessible to most researchers wanting
to analyze the data. The SRA requires large storage and computing facilities that are beyond the capacity of most laboratories. Enabling
terminals to explore the SRA and other large sequence archives will provide insight into ecology, medicine, and industrial applications. In this project we specify
how terminal-based streaming and save-prepare-view will improve insight into ecology, medicine, and industrial applications. In this project we specify
a novel terminal-based system that will enable insight to new ecosystems from different environments. We are developing a set of tools to enable
biologists to mine the metadata associated with the SRA using the
Jetstream and Wrangler APIs, which we will publish on GitHub as
an open-source tool. The system for terminal-based streaming and save-prepare-view
will improve insight into ecology, medicine, and industrial applications.

CCS Concepts

• applied computing → biocomputing; molecular sequence
  analysis; computational genomics

Keywords

Sequence Read Archive, SRA, Metadata, Jetstream, Wrangler, Bioinformatics, Apache Airflow, Dask, Central storage, Snow
Grid, SRA Gateway, Metadata Discovery Challenge

ACM Reference Format:
Searching the Sequence Read Archive using Jetstream and Wrangler. In
MIPS ’18: Pacific-Asia Workshop on Bioinformatics and Genomics (PABG)
https://doi.org/10.1145/3229278.3229287

1 Introduction

A rapid drop in the cost of sequencing DNA—from roughly $10,000
dollars per sample in October of 2008 to just $1,000 today—has
facilitated the rapid growth of the SRA from 106 to 1010 samples in
10 years and has opened new avenues of research. By 2030, experts
estimate that the SRA data will reach 1.5E15. This increase in data
will lead to new opportunities for bioinformatics research and the
need for tools to explore and analyze these data. The SRA is a
terminal-based system that provides insight into ecology, medicine,
and industrial applications. In this project, we specify a novel
terminal-based system that will enable insight into new ecosystems
from different environments.

In the SRA, data is organized into studies, each of which contains
one or more sample files. Each sample has one or more experiments,
and each experiment has one or more lanes. At the time of writing
there were 1,482,135 experiments and 127,709,986 runs in the SRA
(approximately 90% of experiments have one and a single lane). The
primary community established this database to store the metadata
data, but the computational barrier to exploring this data leaves it
behind the reach of many researchers.

Through these many types of genomic data within the SRA, this
project focuses on genomic metadata, because these contain
many different organisms and can attract a wider interest of
researchers compared to single-organism runs. There are few popu-
lar metadata approaches that address the need to explore the
metadata. For instance, pipelines that are designed to extract
meaningful information from the metadata (e.g., Bambusa, or the
dino-meta-bam pipeline) are fragmented and can miss information
that is useful to the researcher. These pipelines provide a
taxonomic profile of an environment and access computationally
demanding, but only provide information about which organisms
are present and are incapable of detecting any viruses that may
be in the sample. The second type of study, whole-genome (WGS)
metagenomics, is where random samples of the genome in the
environment are sequenced, usually without amplification, and
retrieved into order of all the DNA in the sample [6]. The analysis
of these samples is computationally intensive but provides a detailed

https://dl.acm.org/citation.cfm?id=3229278
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®.
Shotgun Sequenced Metagenomes
• What does my organism look like across multiple samples?
• How many samples have my organism?
• What percentage of the sample is my organism?
• What organisms are present in each community?
• What functional genes are present in each community?
• Does my organism look different between these two groups?
How will searchSRA.org help?

Less time spent…
• Setting up SSH keys
• Learning bash commands
• Waiting for downloads and extractions
• Troubleshooting program errors

More time spent…
• Learning about different sequencing technologies
• Analyzing real data
• Designing creative experiments
• Troubleshooting experimental design
Search SRA Gateway

• Search SRA gateway ([https://www.searchsra.org/](https://www.searchsra.org/)) is hosted by SciGaP service ([https://scigap.org/](https://scigap.org/)).


• SciGaP provided storage facility for user data files; inputs and outputs.

• Gateway maintenance, software updates are done by SciGaP team.
Search SRA Gateway Users

- Gateway has two main user groups; Admins and general users.
- For general users, the main services are authentication and authorization, their job monitoring, manages metadata and data generated by the jobs.
- Sharing service allows users to share their jobs with other gateway users.
- Gateway has ‘Cloning’ feature for users, this makes submitting a new job quicker and easier.
- Search SRA provides instructions on using the gateway, examples and citing methods for users.
Search SRA Gateway Admins

• Gateway admins take care of the gateway for its users.

• ‘Admin Dashboard’ is the workspace for gateway admins.

• Gateway admins can manage user access to the gateway; decide whom to grant access and at which level.

• ‘One stop view’ of current state of jobs running through the gateway. e.g.: how many failed last 24 hours.

• Configuring application tools, compute resources and storage resource registration, are few admin functions available to the gateway admins.
Backend: Jetstream(compute) + Wrangler(data)

- One master VM
  - Job scheduling, data management, …
  - HTCondor (and Slurm as an interface!)
  - Ganglia

- Autoscaled workers

- VM configuration managed by SaltStack

https://github.com/linsalrob/jetstream-sra-cluster-setup
https://github.com/linsalrob/jetstream-search-sra
https://github.com/linsalrob/jetstream-search-sra-proteins
Networking / Wrangler Mounts

Virtual machines (master shown)

Virtual switch to the Internet
Shade - OpenStack API Access

Scaling up: Cron scheduled script which cleans up old workers, determine demand based on the HTCondor queue, and start new VMs if needed.

Scaling down: VMs will shut down if now work scheduled to them.
Want to search the sequence read archive but don't know where to start? Check out our new website that allows you to run DNA or protein searches against all the metagenomes in the SRA. Search through $10^{14}$ bp of DNA!

edwards.sdsu.edu/research/simil... searchsra.org

Rob Edwards
@linsalrob

6:53 AM · Dec 6, 2018 · Twitter Web Client

143 Retweets 261 Likes

A. Murat Eren (Meren) @merenbey · 13h
Replying to @linsalrob
This is an excellent service to the community, Rob.

We will look into this with the intention to make it accessible to #anvio users as seamlessly as possible (i.e., looking at a pan-genome and feel the urge to turn it into a metapangenome? Edwards Lab to the rescue.)
Usage

CPU Hours

Number of Jobs
CPU Utilization
Network Utilization (~5 Gb/s total)
Our reach will forever exceed our grasp, but, in stretching our horizon, we forever improve our world.