

Mining Microbial Genomes from Datasets on the Sequence Read Archive

Bhavya Papudeshi, Haley Leffler, Sruthi Ganapaneni, Carrie Ganote, Sheri A. Sanders,
Thomas G. Doak



Sequence Read Archive (SRA)

- The Sequence Read Archive (SRA) is a public repository maintained by National Center for Biotechnology (NCBI) to make raw sequence data from diverse projects available to the research community. SRA currently hosts around 11 PB of sequencing data.
- Reusing or adding these datasets to a research project can reduce sequencing costs of a project, generate enough evidence for statistical and computational methods to be applied, and greatly broaden the scope of studies.
- Accessing and retrieving specific datasets from SRA can be challenging, given its size and heterogeneous content.

Mining the SRA

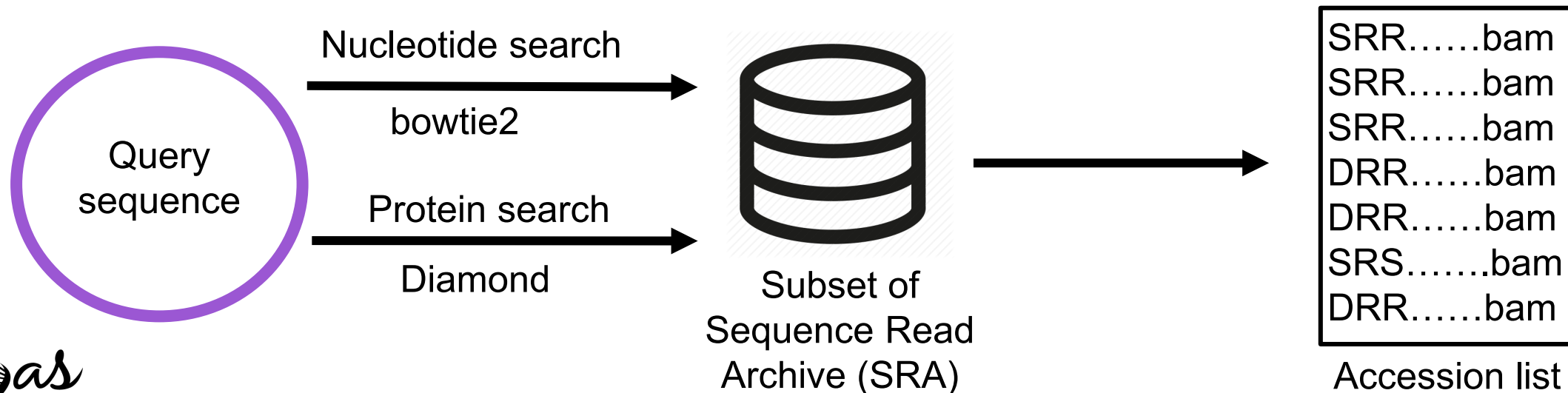
- Using the NCBI SRA website (<https://www.ncbi.nlm.nih.gov/sra>) allows users to lookup specific datasets using the accession, project name, location, or other metadata.
- Using filters, the users can also generate a list of metagenomics/metatranscriptomic datasets that contain a specific taxa of interest.
- In the case the user is looking for a novel microbial genome/sequences, the search becomes a little more challenging as this would require,
 1. Downloading all datasets (complete or subset) to a local storage space to search the query against.
 2. Require lots of compute resources for a long time, to align the query sequence against the large number of query datasets.



SearchSRA gateway

SearchSRA gateway

- Developed by Rob Edwards lab, San Diego State University on XSEDE resources, Jetstream and Wrangler, <https://www.searchsra.org/>.
- User uploads an input sequence, and waits ... while the website spins up multiple virtual machines and searches subsets of the SRA database stored on Wrangler, in parallel. The results are all concatenated and given back to the user as one output.
- For more information- <https://edwards.sdsu.edu/research/similarity-searching-the-sra/>



Create an account





The screenshot shows the SRA website interface. At the top, the browser address bar displays <https://www.searchsra.org>. The main navigation bar features the 'Searching SRA' logo on the left and links for 'Documentation', 'About', 'Contact', 'Examples', and 'Cite Us' on the right. A dark blue bar below the navigation contains the 'Create account' link, which is highlighted with a white mouse cursor, and a 'Log in' link to its right. Below the navigation bar, a banner image shows DNA sequence data (A, T, C, G) on a dark background. To the right of the banner, the 'SRA' section title is followed by a paragraph: '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®.' Below the banner, the 'Searching SRA Gateway' section is visible, featuring a 'Welcome!' message and a paragraph: 'We are generating sequence data at an alarming rate! The [Sequence Read Archive](#) currently contains ~10 petabytes of sequence data (~10¹⁶ pb) and through 2016 has grown at approximately 10 TB of sequence data per day. We're here to help you search through that data and find meaning to'. In the bottom-left corner of the page, there is a 'Login' link and a 'Register' link, with the 'Register' link highlighted by a red rectangular box. The Windows taskbar at the bottom shows various application icons and the system clock indicating 9:27 AM on 5/20/2019.

Upload a query sequence

The screenshot shows a web browser window with the URL <https://www.searchsra.org/experiment/create>. The browser's address bar and tabs are visible at the top. The main content area is titled "Application configuration" and contains several sections:

- A message box at the top says "This has not been shared".
- The "Application Inputs" section includes:
 - A "Fasta-Reference-File" input field with a "Choose File" button highlighted in red. The text next to it says "No file chosen" and "Max Upload Size: 10M". Below it is the instruction "Upload your input Fasta-Reference-File." and a "view file" button.
 - A dropdown menu labeled "Select existing Search IDs File OR Upload your own below" with "HMP" selected.
 - An "Optional Input Files" section with a "Choose Files" button and "No file chosen" text, also with a "Max Upload Size: 10M" limit.
- The "Notifications" section at the bottom has three buttons: "Save", "Save and launch" (highlighted in red), and "Start over".

Results from SearchSRA gateway

Experiment Status	COMPLETED								
Job	<table><thead><tr><th>Name</th><th>ID</th><th>Status</th><th>Creation Time</th></tr></thead><tbody><tr><td>A1524838725</td><td>439</td><td>COMPLETE</td><td>02/21/2019, 1:32 PM - GMT-0400 (Eastern Daylight Time)</td></tr></tbody></table>	Name	ID	Status	Creation Time	A1524838725	439	COMPLETE	02/21/2019, 1:32 PM - GMT-0400 (Eastern Daylight Time)
Name	ID	Status	Creation Time						
A1524838725	439	COMPLETE	02/21/2019, 1:32 PM - GMT-0400 (Eastern Daylight Time)						
Notifications To:									
Creation Time	02/21/2019, 1:32 PM - GMT-0400 (Eastern Daylight Time)								
Last Modified Time	02/21/2019, 9:11 PM - GMT-0400 (Eastern Daylight Time)								
Inputs	Fasta-Reference-File: reference  Select existing Search IDs File OR Upload your own below: All-SRA-metagenomes								
Outputs	Downloading-Details: report.txt  Search-SRA-Standard-Error: search-SRA.stderr  Search-SRA-Standard-Out: search-SRA.stdout 								
Storage Directory	Open								
Errors									

Report.txt

```
results_url =  
http://149.165.169.158/results/51  
24a181-38ef-47fe-98be-  
e2c5a7941b65/results.zip
```

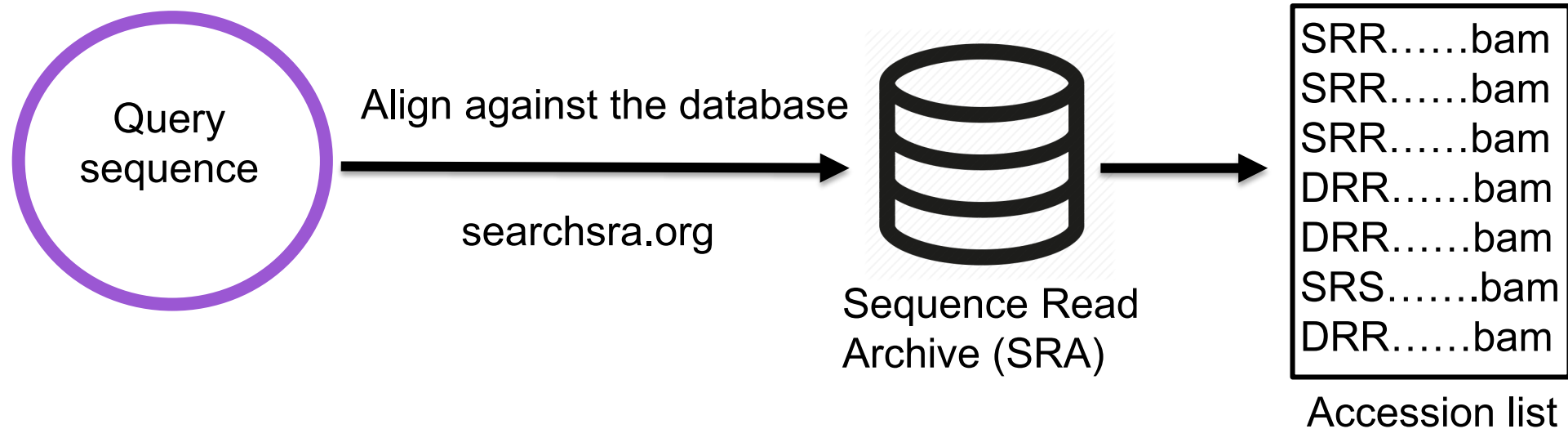
```
results_size = 496M  
total_compute_hours = 178.64
```




Downloading and filtering the results

Downloading the results

- Download SearchSRA gateway results— a list of bam files that contain information of only the aligned reads from each SRA sample aligned against the query sequence.
- Generates bam files for even those SRA samples that had no hits as well, in this case the bam files are empty

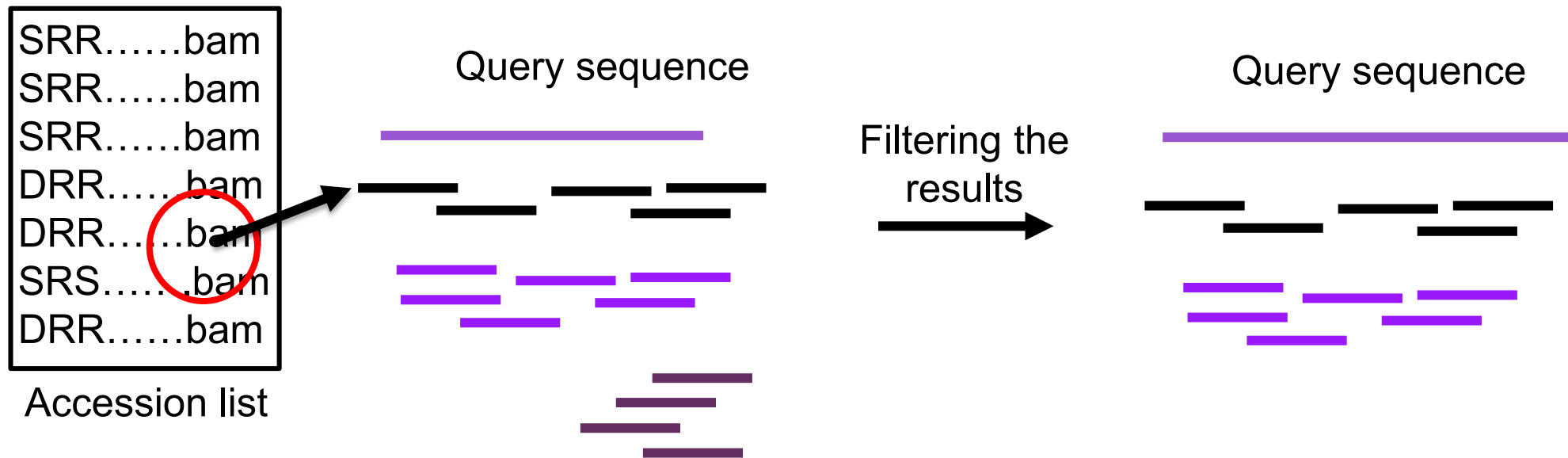


Filtering the results

We run the filtering script that generates a new directory labelled as subset with a list of only those bam files that have,

- alignment length more than 100bp, and
- more than 10 hits mapping to the query sequence per SRA sample

This script will remove both the empty bam files and potential false positives.

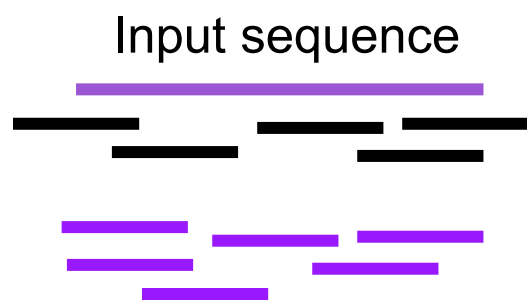




Collecting the metadata/run information

Further filtering using metadata

- Using E-utilities toolkit developed by NCBI, we can lookup the metadata associated with the previously filtered datasets to further narrow down the datasets.
- For example, if query sequence was identified in a human gut sample, and in this study we are looking for the prevalence of the genome in other human gut samples – we can narrow down the filtered datasets to only human gut samples.



Filtered list of bam files in subset directory

Get the metadata of the accessions



SRR..	human
SRR...	water
DRR...	human
DRR...	water
SRS...	human
DRR...	wastewater

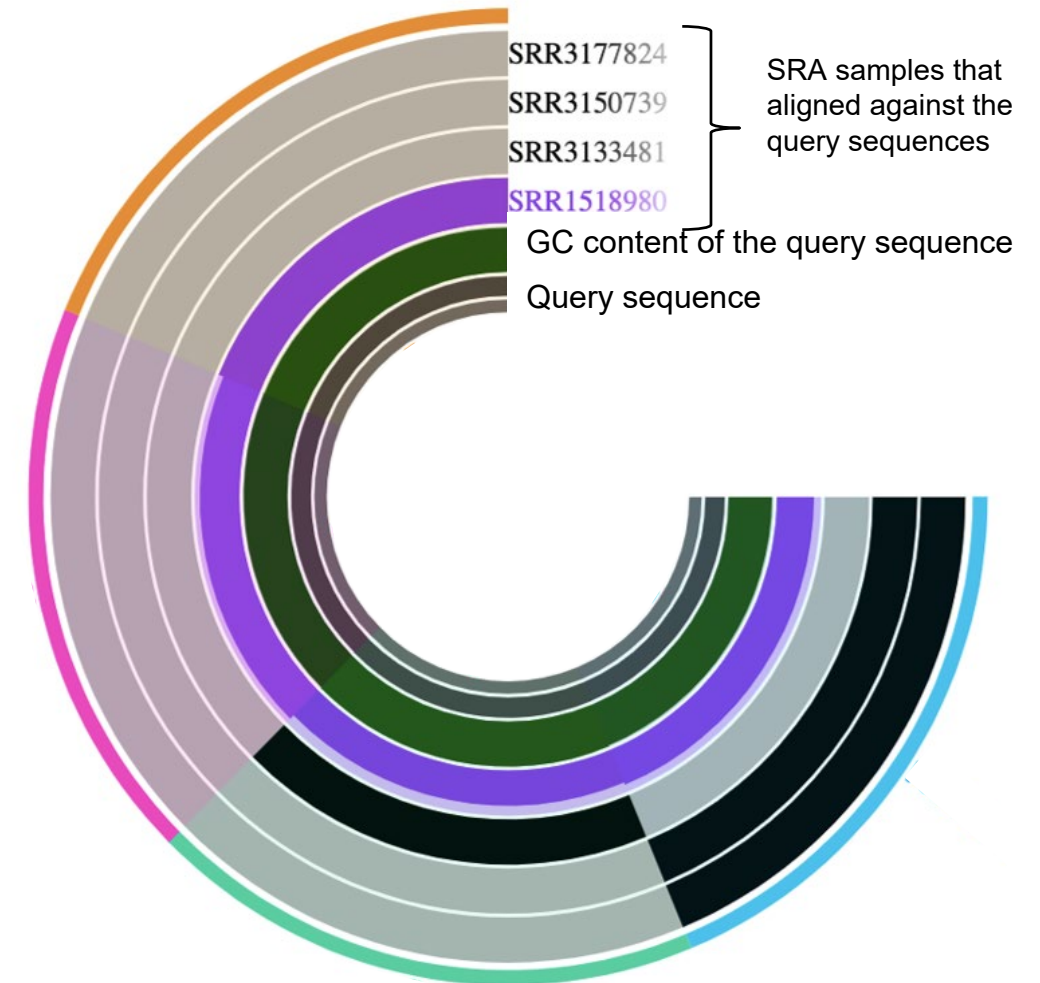
Metadata for the subset



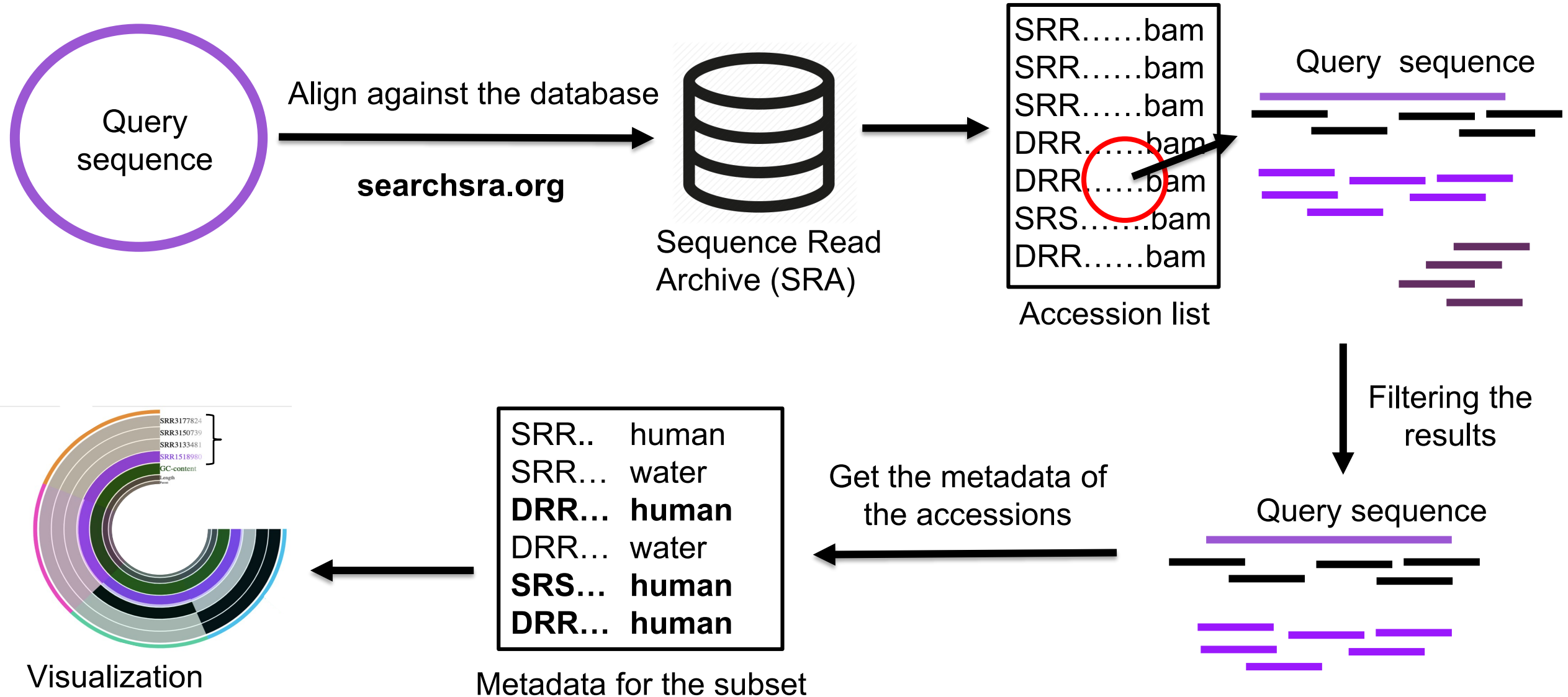
Visualization and further downstream analysis using Anvi'o

Visualizing the results using Anvi'o

- Anvi'o is an open source tool that allows some really fun visualization and analysis of whole genome metagenomics data.
- In this workflow, we initially use the Anvi'o image for visualization. The figure represents how the reads from filtered SRA samples (four concentric circles labelled as SRR****) align against the query sequence (inner most ring).
- Can look around further using other analysis/workflows available through Anvi'o = [documentation available here](http://merenlab.org/2016/06/22/anvio-tutorial-v2/)



Overview of the workflow



Conclusions

- This workflow was developed and tested using the crAssphage genome that is found in the 90% of the human gut microbiome samples.
- This workflow is transferable to other datasets, and is actively being developed to test and include other features.
- This workflow can identify other environments where the query sequence can be found in, as well as widen the scope of the study by generating enough evidence for statistical and computational methods to be applied.
- The workflow and the scripts are available to the community through GitHub with documentation: <https://github.com/NCGAS/CEWiT-REU-Identifying-datasets-in-SRA-using-Jetstream>
- The workflow and visualization tools are installed and available as a pre-configured image on the Jetstream cloud computing system (<https://use.jetstream-cloud.org/application/images/831>). Contact NCGAS (help@ncgas.org) for access to this resource.

Acknowledgements

- National Center for Genome Analysis Support ([NCGAS](#)), NSF funded organization at Indiana University to support researchers with their genomic analysis.
- Our collaborations with [IU Pervasive Technology Institute](#) and the [Pittsburgh Supercomputing Center \(PSC\)](#) at Carnegie Mellon University.
- Within Indiana University [Pervasive Technology Institute](#) and a management unit of the [Research Technologies](#) Division of [University Information Technology Services](#).
- [Rob Edwards lab](#) at San Diego State University, who developed SearchSRA gateway and help throughout the project.
- [XSEDE Jetstream cloud](#) computing resource for developing this work and making this work available to the community.

ncgas

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

Affiliations:



**PERVASIVE
TECHNOLOGY INSTITUTE**



Collaborator: [ABI-1458689 2015](#), [ABI-1759914 2018](#)



RESEARCH TECHNOLOGIES
UNIVERSITY INFORMATION TECHNOLOGY SERVICES

NSF Awards:

[DBI-1062432 2011](#), [ABI-1458641 2015](#), [ABI-1759906 2018](#)



/ncgasiu



@ncgas



ncgas.org