Mining Microbial Genomes from Datasets on the Sequence Read Archive

Haley Leffler[1], Sruthi Ganapaneni[1], Bhavya Papudeshi[2], Carrie Ganote[2], Sheri A. Sanders[2], Thomas G. Doak[2]

[1] Department of Human Biology, Indiana University, Bloomington, IN, [2] National Center for Genome Analysis Support, Pervasive Technology Institute, Indiana University, Bloomington, IN

NSF Awards
ABI-1759906
2018

Background

- **Sequence Read Archive (SRA)** hosts more than 14PB of raw sequencing data with metadata. Most published genomics projects around the world have deposited their sequences here, which is available for further downstream analysis.
- Searching through this database however requires lots of compute resources.

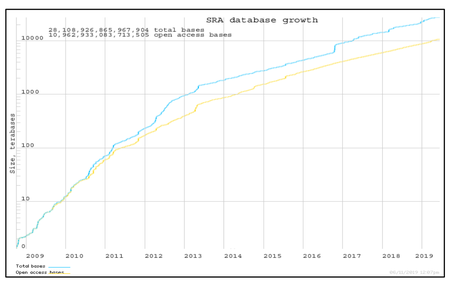


Figure 1: SRA growth diagram showing the exponentially increasing bases housed within the database from SRA website

- **SearchSRA gateway** was developed by the Edwards lab, San Diego State University, allowing researchers to identify other datasets in the SRA that contain a reference genome, taking only a couple hours to search through the entire database.
- **Jetstream** is an NSF cloud computing infrastructure that provides higher performance than a desktop or workstation and is easy for inexperienced researchers to use with its straightforward user interface. For this reason, this workflow was developed here.
- **Our objective is to develop a workflow that allows researchers to mine the SRA—using the SearchSRA gateway—then filter and visualize the identified datasets.**

Methods

- The workflow begins with uploading a reference genome to the Search SRA gateway, which aligns the genome against the SRA database and returns bam files for each alignment.
- The resulting bam files are filtered to include only those datasets that have good coverage of the reference genome.
- The filtered bam files are uploaded to Anvi'o, to visualize the coverage of the reference genome in the samples identified. This step helps further filter out false positives.
- We selected two reference genomes — bacteriophages, (Fig 2) to test the workflow's flexibility

1. **CrAssphage**, a highly abundant phage in the human gut microbiome.
   - 90% of humans have crAssphage in their gut microbiomes, but the significance of the phage is unknown.
   - CrAssphage has been previously found in termite gut metagenomes, humans, and some water samples. CrAssphage-like genomes have also been identified in Old-World and New-World primates, but have not been identified in other species.
   - Since 8.2% of the bacterial species in the human microbiome overlap with pig samples, we applied this workflow to identify crAss-like phage in pig microbiome samples.
2. **Pseudomonas phage PAK-P1**, a phage that infects the *Pseudomonas* class of bacteria, a leading cause of healthcare-associated infections in immunocompromised patients.
   - Pseudomonas phage PAK-P1 has been used to treat *Pseudomonas aeruginosa* infections.
   - This bacteriophage has relevant clinical applications, so through this workflow we study this phage's distribution across different environments and its genetic variation.

A) B)

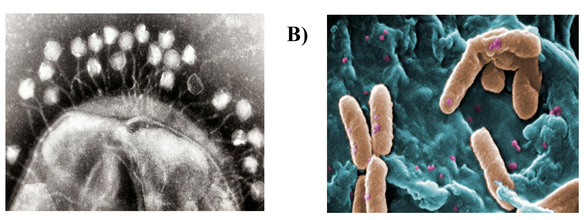
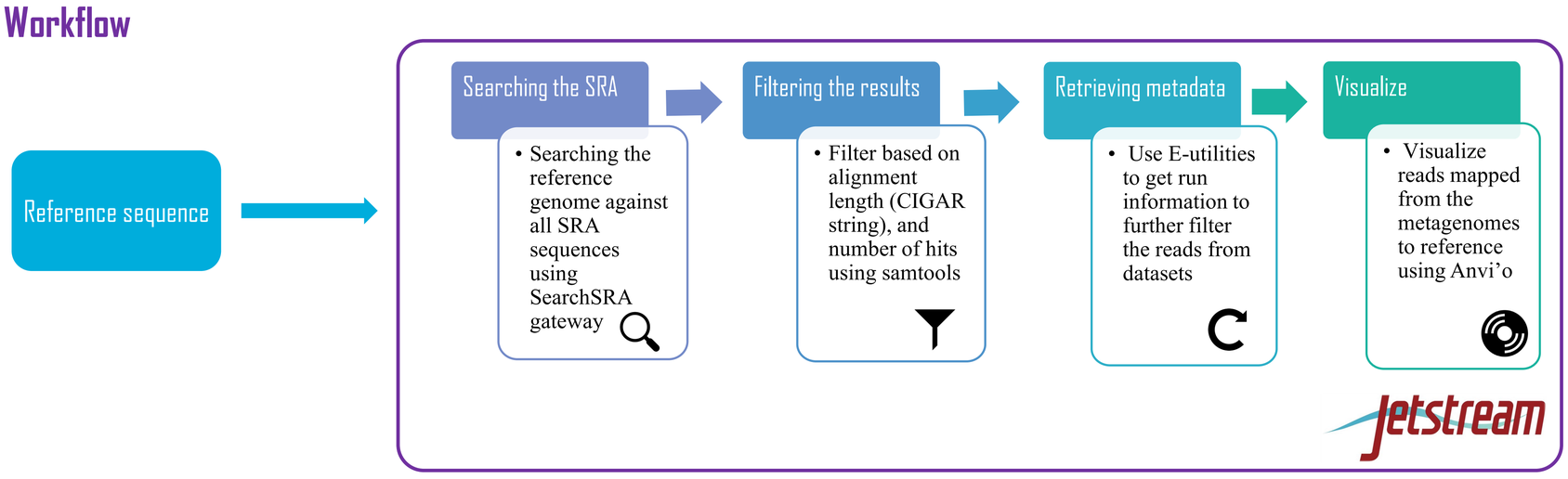


Figure 2: A) Bacteriophages infecting a bacterium
B) *Pseudomonas aeruginosa* bacteria

Workflow



Results

CrAssphage

- 10 pig (in pink) and human microbiome samples (blue) were mapped to the crAssphage genome, using Anvi'o in Fig 3.
- Of the 10 pig samples identified after SearchSRA, further filtering identified them as false positives (highlighted in black), and only two pig samples (highlighted in pink) had high coverage of the crAssphage genome.
- The coverage of the crAss- like sequences in pig samples (pink), when compared to human microbiome samples (blue), further confirms their presence.
- Cluster 1 (red) had the highest coverage in pig samples, this sequences belonged to conserved bacteriophages genes and not specific to crAssphage.

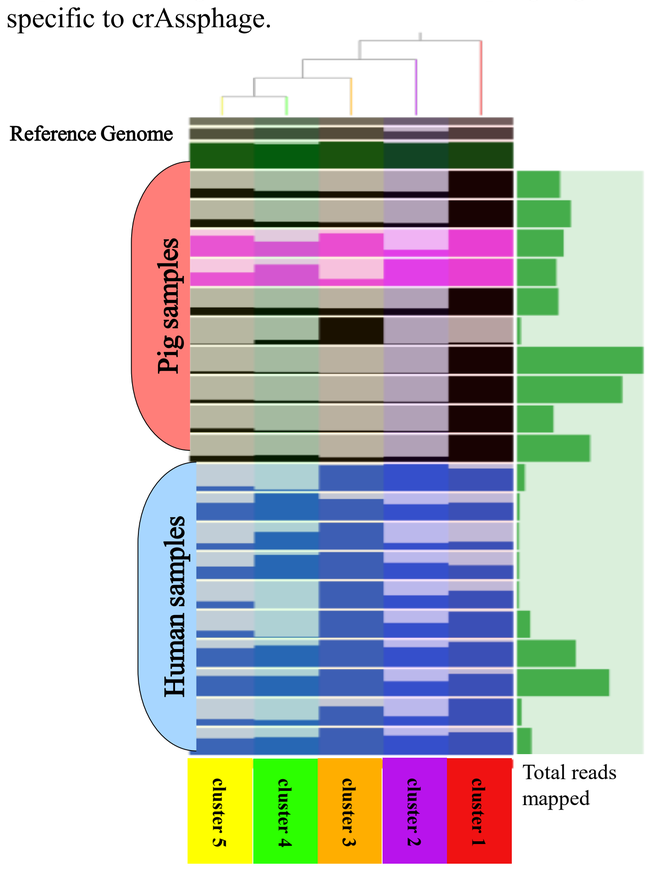


Figure 3: Vertical representation showing the genomic distribution of the 10 pig and human samples aligning to the crAssphage genome

Pseudomonas phage PAKP1

- Only four datasets from SRA were identified to contain the reference genome, after filtering.
- Three datasets (in black) had only partial coverage to the reference genome, containing sequences in only one of the clusters (green or blue).
- BLAST analysis of the green and blue clusters identified these sequences as head and/or tail genes (green) or regions of the phage genome containing host bacteria sequences (blue).
- The dataset (in purple) that had a high coverage to the reference genome was previously classified as an unidentified genome (SRR1518980) in SRA, but now **we can classify this genome as Pseudomonas phage PAKP1.**

Reading the Anvi'o figures:

- The aligned reads for each dataset (each row in Fig 3, concentric circle in Fig 4) are mapped to the reference genome (grey).
- Plotted clusters are determined based on sequence similarity.
  - Both Fig 3 and 4 show the mean coverage of each cluster to the reference genome.
  - The fuller each bar is on the figures, the greater the number of metagenomic reads for that section of the reference genome.

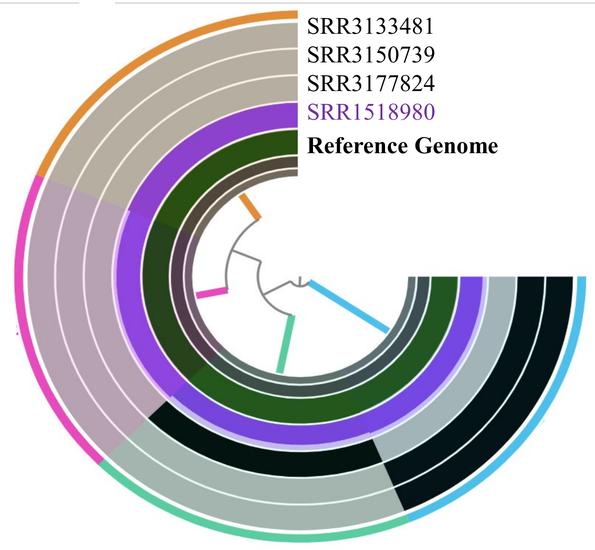


Figure 4: Circular representation showing the genomic distribution of samples containing hits to the Pseudomonas phage PAK-P1 reference genome

Conclusions

- Testing our workflow with the two reference genomes provided valuable feedback for improving the workflow.
  - **CrAssphage**

  Two pig samples were identified to contain crAss-like sequences after filtering. Visualization of these sequences show that they have a similar genomic distribution of crAssphage sequences as the human microbiome datasets.
  - **Pseudomonas Phage PAK-P1**

  There were only four datasets in SRA that were identified to contain Pseudomonas phage PAK P1 genes, possibly because the parameters for filtering were too strict or due to low coverage of this genome in the database.
- We developed a workflow to mine the SRA, identify, and visualize other datasets containing a genome of interest.
- The workflow and visualization tools are installed and available as a pre-configured image on the Jetstream cloud computing system. Contact NCGAS for access to this resource.
  - Jetstream: https://use.jetstream-cloud.org/application/images/831
  - GitHub repository with documentation: https://github.com/NCGAS/CEWiT-REU-Identifying-datasets-in-SRA-using-Jetstream
- Future directions will include downstream analysis steps to analyze the clusters further to predict functions of the cluster, and testing different filtering parameters.

Acknowledgements

Special thanks to, Robert A. Edwards from San Diego State University, CEW&T REU-W program for the opportunity to work with NCGAS, where the pipeline was developed, NCGAS and PTI for the funds to support this research.