Website Integration of DNA Sequencing Facility

Sidharth Bansal, Krishan Kanji, Alex Li, Shreya Mantripragada, Pranav Sarathy

UC Berkeley DNA Sequencing Facility

Introduction

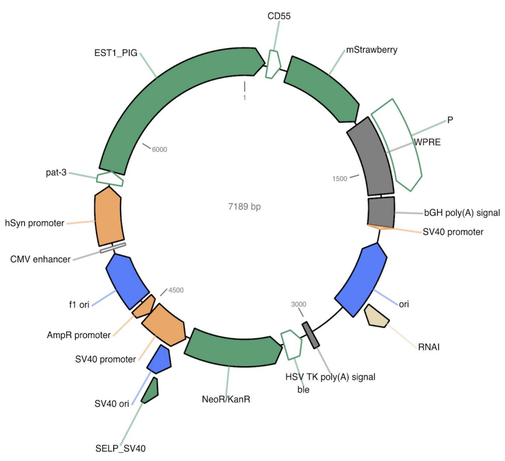
- DNA sequencing is essential modern biology
- DNA sequencing facilities turn biological samples into genomic data that can be used for diagnostics, research, and medicine.
- We aimed to analyze and improve the data pipeline for Oxford Nanopore sequencing to help with scalability
- This automation aims to enhance scalability, reduce manual errors, and improve turnaround time
- The pipeline must:
  - Process raw sequencing data
  - Identify successful and failed reads
  - Prepare and deliver results to customers

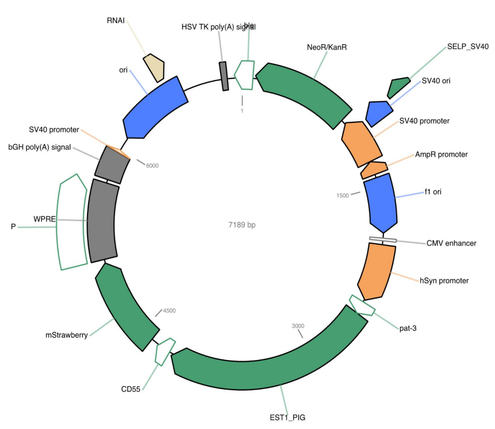
Objectives

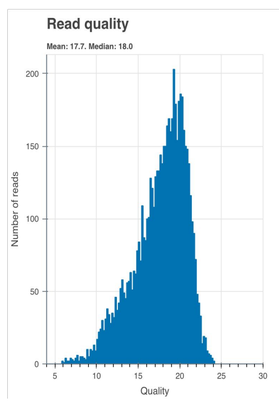
- Create a program to create a plasmid map from the DNA sequence
- Create an automated backup report for when the sequencing was not clear from the machine
- Merge preprocessing, spreadsheet creation, and clone validation jobs into one end-to-end script
- Build a supervisor module for fault/status reporting per barcode
- Enable automated, faster, and more reliable runs on Savio cluster

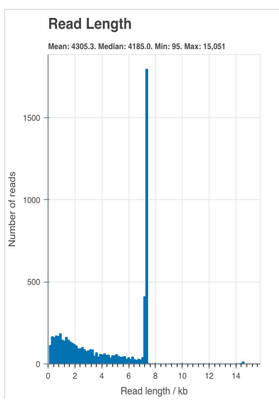
Results

- Helped automate the DNA sequencing pipeline using three previously independent scripts.
- Improved accuracy of clone validation and enhanced output quality using automated plasmid map generation alongside spreadsheet creation.









Materials

- Programming Libraries
  - Pandas
  - NumPy
  - Plannotate
  - Bokeh
  - geckodriver
  - NextFlow
- Environments
  - Savio
- Programming Languages
  - Python
  - Bash
- External Methods
  - EPI2ME

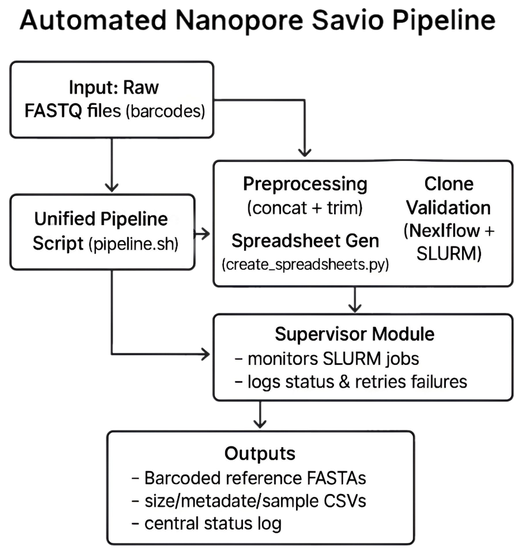
Methods

- Learned the current pipeline to understand the processes and where the work needs to be done
- Created a local script that took in a fasta file and outputted the desired PNG via the plannotate library
- Uploaded python file to Savio and figured out how to change the code to accommodate for the lack of a browser through Savio due to original libraries requiring browser
- Combined three independent scripts into a single end-to-end pipeline for preprocessing, spreadsheet generation, and clone validation
- Wrote a local supervisor script to monitor SLURM job status for each barcode and log failures
- Adapted the workflow for Savio by adjusting file paths, permissions, and container execution logic to ensure compatibility with the cluster environment

Conclusions

- Integration with the Savio cluster enables scalability and high-throughput processing, making it suitable for large-scale sequencing facilities
- Integrate the processing steps with automated merging of the future website

Feature	Database	Identity	Match Length	Description	Start	End	Length
WPRE	Snapgene	100.0%	100.0%	woodchuck hepatitis virus posttranscriptional regulatory element	1106	1695	589
SV40 promoter	Snapgene	100.0%	100.0%	SV40 enhancer and early promoter	4103	4461	358
SV40 ori	Snapgene	100.0%	100.8%	SV40 origin of replication	4117	4253	136



Acknowledgements/Contact

We thank Dr. Scott Geller, Nicolas Sapountzis, and Brian McCarthy for their guidance and support.

For inquiries, please contact:

- dnaseq@berkeley.edu
- (510) 642-6383