Spring 6-1-2021

Detection of Antibiotic Resistance Genes in the Wastewater Microbial Metagenome

Alan Caparaz Le

Detection of Antibiotic Resistance Genes in the Wastewater Microbial Metagenome

A Research Project
Submitted in Partial Fulfillment of the
Requirements for the
Master's Degree in

BIOINFORMATICS

Presented to
The Faculty of the Department of Computer Science
San José State University

By
Alan Caparaz Le
May 2021

ABSTRACT

The existential threat of emerging antibiotic resistance in microbial communities poses significant risks to public health. In particular, wastewater can serve as a point of confluence for pharmaceuticals and antibiotic-resistant bacteria from urban and agricultural settings. While this is a prime environment for genetic drift and horizontal transfer of antibiotic resistance genes (ARGs) and mobile genetic elements, it also presents an opportunity for resistome monitoring via shotgun metagenomic sequencing and downstream analysis. This project reports the application of a hybrid assembly approach for the detection of ARGs within DNA derived from a wastewater sample collected from the San José-Santa Clara Regional Wastewater Facility, which serves a significant portion of the San Francisco Bay Area. Hybrid assembly (with polishing) of Nanopore-derived long reads and Illumina-derived short reads resulted in detection of additional ARGs compared to a previously-performed short-read-based approach.

ACKNOWLEDGMENTS

Heartfelt thanks go out to the following people for their expertise, guidance, and support in bringing this project to fruition.

- Dr. Wendy Lee for relentless support, advising, and for introducing me to this metagenomic research project.
- Dr. Cleber Ouverney and Michael Abrams for taking the initial steps in attempting to identify antibiotic resistance genes using their DNA library.
- Dr. Andreopoulos for serving on the committee for this project and providing feedback.
- Josh Garcia, Randy Kirchner, and Tim Andriese for enabling me to perform wet lab research in the Ouverney Lab in Duncan Hall.
- Steven Boring for maintaining the College of Science HPC.
- My peers in the MS Bioinformatics program for all the learning, sharing, and growing we’ve done together.
- My friends and family, the metagenomic community I am so proud to be part of.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 THE RESISTOME

Wastewater from agricultural activities and sewer systems presents a notable environment for the confluence of multiple types of antibiotics and antibiotic-resistant bacteria. The environmental prevalence of antibiotics is by no means limited to wastewater. Antibiotics have been detected in rivers, lakes, groundwater wells, and the soil microbiome [1], applying human-induced selective pressures that impact the frequencies of antibiotic-resistance genes (ARGs) in the community.

In 2018, antibiotic resistance in the United States was estimated to account for an additional national cost of treatment of $2.2 billion per year [2], or an average extra cost of $1,383 per bacterial infection treated. However, this additional treatment expense is not represented equally across demographics, as black and Hispanic patients, prisoners, the elderly, and patients with comorbid conditions are most likely to be affected by antibiotic-resistant infections. The consequences of antibiotic resistance may therefore include long-term inequities that affect society in more far-reaching aspects than healthcare alone.

Although humans have been exploiting modern antibiotics for nearly a century, the microbial capacity for antibiotic resistance is not a novel phenomenon. Rather, microbial DNA sequences from microbiomes such as 30,000-year-old permafrost samples and long-isolated cave structures indicate a baseline frequency of ARGs existed in bacterial genomes prior to widespread human adoption and application of antibiotics [3], [4]. As such, ARGs are concerning not because of evolutionary novelty, but because their distribution may rapidly change as humans continue to transform our local environments. One solution to such a dynamic resistome is a frequent and robust monitoring pipeline for the identification of emerging pathogen-associated ARGs.

1.2 Metagenomic Analysis of Environmental Samples

1.2.1 Detection of Uncultivable Species

Environmental samples containing microbial communities present a possible means of assessing a habitat's members and their frequencies. However, due to the high microbial diversity inherent to environmental samples, cultivable bacterial species (those which can be isolated in pure culture and represent less than 1% of mixed microbial communities) are unlikely to accurately represent the resistome [5]. One solution to this problem is cloning the extracted DNA into vectors, producing a library that can be clonally expanded using a competent prokaryotic host. However, this does not guarantee full coverage of the DNA from the environmental sample and inherently introduces additional fragmentation into the metagenome. Thus, one strategy for achieving broad detection of ARGs in a functional context is shotgun metagenomics. By immediately isolating DNA and sequencing the entire genomic or transcriptomic content of an environmental sample, cultivation-independent characterization of these community members is made feasible. One issue that may rise from this approach is poor alignment of the sequencing data to existing reference sequences, though this is gradually becoming less of an issue over time as more genomes of non-cultivable strains are made available [6].

1.2.2 Next-Generation Sequencing Strategies

The continued proliferation and improvement of high-throughput sequencing methodologies has enabled metagenomic analyses to be conducted at ever-increasing scale. Short reads generated by the Illumina sequencing-by-synthesis platform feature high per-base read accuracy and depth of coverage; however, due to their limited length, assembly from such reads may be susceptible to errors due to repetitive elements if the reads are unable to span the entire length of the element. Because metagenomic samples contain many species by definition, the number of contigs generated by this process may reflect either the diversity of the environmental sample or the inability to resolve reads into contiguous segments. The former is a natural property of the sample of interest, while the latter is frequently a consequence of poor coverage or inadequate tools and techniques. Long-read sequencing technologies such as the Oxford Nanopore Technologies platform aim to improve genome reconstruction by enabling assembly to be performed with larger starting reads, facilitating greater contiguity and localizing genes in relation to each other with greater accuracy [7]–[9]. However, low per-base accuracy and low throughput make it difficult to reliably apply Nanopore sequencing to applications that require high resolution, such as characterizing mobile genetic elements (MGEs), ARGs, or highly polymorphic genes [10].

1.2.3 Hybrid Assembly

Because of the relatively high error rate of the Nanopore long-read sequencing platform, it is useful to combine these sequences with short reads generated by the Illumina platform, allowing for both high contiguity and per-base accuracy. These two strengths are highly valuable in metagenomic analyses, since per-species coverage can vary greatly depending on the frequency of a given species and the genome sizes of other species within the sample. Approaches to hybrid assembly differ, but one approach is to generate a long-read assembly (with high expected contiguity) and use a polisher with short reads to correct incorrectly-called bases in the assembly [11]. In some cases, a short-read assembly is first generated, and a gap-filling stage using long reads is performed to improve contiguity [8]. A hybrid assembly strategy facilitates and enhances the recovery of multiple distinct genomes from metagenomic samples [12]–[14] allowing for the detection of not only ARGs or other sequences of interest, but also the species from which they are derived.

1.2.4 Taxonomic Classification and Detection of Antibiotic-Resistance Genes

Because of the presence of uncultivated microbial species, computational approaches for characterizing rich metagenomic datasets will continue to be necessary. For example, because microbial diversity is frequently a metric of interest in environmental samples, contigs generated from such a dataset need to be grouped taxonomically, a process known as "binning" [15]. Binning accuracy plays a critical role in the conclusions drawn from a dataset, particularly in the context of detecting horizontal transfer of MGEs and ARGs. These types of genetic structures in particular pose a daunting task for both public health and metagenomic analysis [16]. Indeed, metagenomic analysis performed using a simulated short-read-based metagenome shows very poor recovery of genomic islands and plasmid sequences [10]. While binning tools such as MetaBAT 2 allow for significant parameter tuning to improve performance in a short-read dataset [15], the upstream application of long-read sequencing technology enables the assembly of larger contigs, thus enabling more accurate binning and more complete genome assembly [17]. Sufficiently high depth of coverage using long-read sequencing can even preclude the need for a binning step entirely [18], though such deep sequencing may not be feasible for every research group. In comparison, read-based classifiers like Kraken 2 match k-mers against a reference database to identify the likely lowest common ancestor for a sequence [19]. This approach can reduce computational overhead and be used earlier in the analysis pipeline.

2 METHODS

2.1 Sample and Data Provenance

2.1.1 Fosmid Library

Vector-based storage and amplification of environmental genomic samples enable long-term preservation and study [5]. Construction of such libraries could be considered tantamount to taking a genomic snapshot of the environment at that time. The fosmid library for this study was prepared from an environmental wastewater sample that was collected from the San José-Santa Clara Regional Wastewater Facility, as described in a previous manuscript [20]. In brief, DNA from 1.0 g pellets of sediment from wastewater were processed using the Epicentre Meta-G-Nome DNA Isolation Kit, yielding high molecular weight genomic DNA. This DNA was then used to construct the fosmid library using the CopyControl™ Fosmid Library Production Kit with pCC1FOS vector cloning system and E. coli EPI300™-T1R Plating Strain, resulting in 4012 clones. Depending on the research application, DNA samples from each of these clones can be combined into pools. This allows for high-throughput metagenomic shotgun assembly to be performed.

2.1.2 Short-Read Sequencing Data Provenance

Short-read sequencing was performed previously using the Illumina HiSeq 1000 platform [20]. Short reads used in this current project have been previously made available under the NCBI SRA Accession ID `SRX286069` as part of the study "Activated Wastewater Metagenome". For the Activated Wastewater Metagenome sequencing, the 4012 fosmid clones were grouped into 12 pools containing approximately 334 clones each. Furthermore, the MG-RAST analysis of this dataset is available under the ID `mgm4521514.3`, offering quality control metrics, taxonomic classifications, and functional hits for genes and gene families.

Assuming all fosmid inserts are ~40kb in length and 3343 fosmid clones are represented, this dataset provides approximately 270X depth of coverage. Limiting the size of the input data can substantially improve processing time. In order to subsample the paired-end reads to 10% of the original sample (and thus a coverage of ~27X), the `sample` command of seqtk v1.3 was used

[21]. This subset can then be used with polishing tools, since lowering the coverage is not expected to significantly impact the final result polishing result.

2.1.3 Reference Sequences

Because the fosmid library clones were generated in the *E. coli* EPI300-T1^R Plating Strain, there exists the possibility of host sequences within the sequencing data. The canonical reference genome for *E. coli* DH10B was downloaded from NCBI entry NC_010473.1 [22] using the *esearch* and *efetch* utilities available in NCBI's Entrez toolkit. The pCC1FOS vector sequence was downloaded from NCBI entry EU140751.1. Resistance gene identification was performed on both references to determine the possible contribution of these sequences to the ARGs detected in each metagenomic assembly.

2.2 Library Preparation and Long-Read Sequencing

2.2.1 Fosmid Pool Preparation

Because DNA extracted from these clones has been used for multiple projects, fosmid pools had to be reconstructed from available aliquots of DNA, which had previously been used to create fosmid pools for an earlier sequencing experiment (see *Short-Read Sequencing Data Provenance*). Analysis was performed on the Agilent TapeStation platform to obtain concentrations for each pool (Figure 1). Out of 12 clonal pools from this previous dataset, 10 were recovered in sufficient quantity to represent each pool equally (insufficient DNA was available to regenerate Pool1 and Pool10). The new pool, Pool13, was created by combining 300ng of DNA from each available clonal pool. The resulting pool contained 3000ng of DNA in 19.97uL, enabling the ten pools to be sequenced in a single run. Pool13 was used for the long-read sequencing and downstream analysis described in Section 2.2 and Section 2.3.

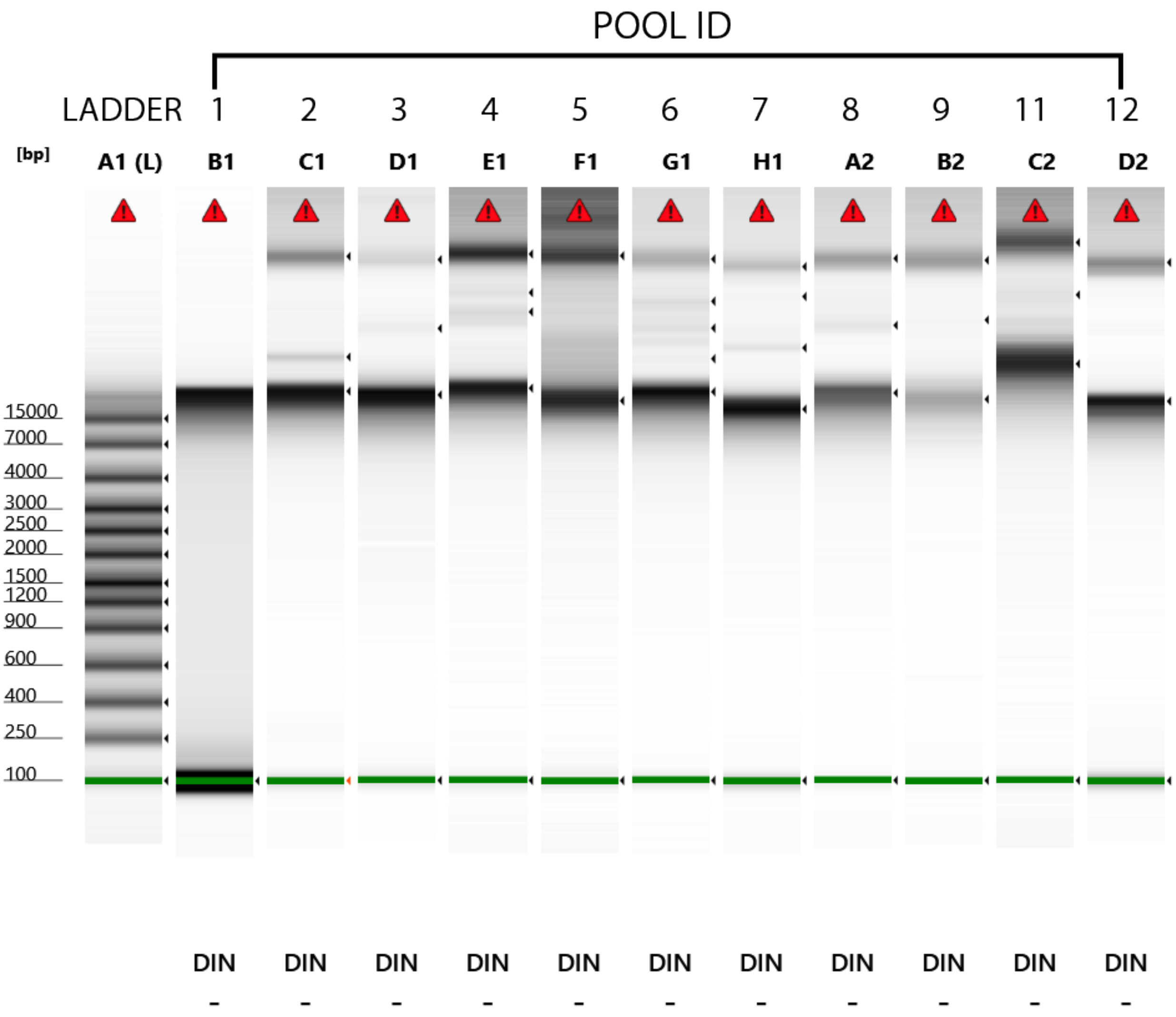


Figure 1. Fosmid pool fragment sizes via Agilent TapeStation.

2.2.2 Library Preparation

The Oxford Nanopore Technology MinION workflow consists of library preparation protocols, the MinION sequencing device, MinKNOW sequencing software v20.06.05, and downstream analysis tools. Library preparation was performed using the Ligation Sequencing Kit SQK-LSK109 and the Genomic DNA by Ligation protocol (Oxford Nanopore Technologies, vGDE_9063_v109_revV_14Aug2019). Preparation also required the NEBNext® Companion

Module for Oxford Nanopore Technologies® Ligation Sequencing (New England BioLabs Inc, E7180S) and Agencourt Ampure XP Beads (Beckman Coulter Life Sciences, A63880).

After preparation of the fosmid pools as Pool13, 6.66µL of Pool13 fosmid DNA containing 1µg of DNA was diluted with nuclease-free water for a final volume of 49µL. For DNA repair and end-prep, 47µL of diluted DNA was combined with library preparation reagents in a 0.2mL thin-walled PCR tube according to Table 1, and thermal cycling and magnetic bead purification was performed.

Table 1. Volumes for DNA repair and end-prep for Genomic DNA by Ligation

Reagent	**Volume**
DNA CS (positive control)	1µL
DNA (from fosmid clones)	47µL
NEBNext FFPE DNA Repair Buffer	3.5µL
NEBNext FFPE DNA Repair Mix	2µL
Ultra II End-prep reaction buffer	3.5µL
Ultra II End-prep enzyme mix	3µL
Total	60µL

Adapter ligation and further magnetic bead purification were performed with Long Fragment Buffer to complete library preparation and enrich for DNA fragments of 3kb and longer, per Table 2.

Table 2. Volumes for adapter ligation and cleanup

Reagent	Volume
Repaired and end-prepped DNA	60µL
Ligation Buffer (LNB)	25µL
NEBNext Quick T4 DNA Ligase	10µL
Adapter Mix (AMX)	5µL
Total	100µL

2.2.3 Long-Read Sequencing with MinKNOW and the MinION Mk1B

After library preparation, sequencing was performed using MinION sequencer with the R9.4.1 flow cell (Oxford Nanopore Technologies, FLO-MIN106). 12µL of the prepared library was combined with 37.5µL Sequencing Buffer (SQB) and 25.5µL Loading Beads (LB) in a new 0.2mL PCR tube. The MinION SpotON Flow Cell was loaded into the MinION Mk1B and loaded with 800µL of flow cell priming mix via the priming port. The prepared library was mixed by pipetting, then 75µL of the library was added to the sample port dropwise. MinKNOW software was used to monitor and operate the MinION sequencer during the course of the sequencing run. Although the MinION flow cell is capable of longer runs, the sequencing run was manually terminated after approximately 24 hours of runtime because the library was depleted by that time.

2.3 Processing of Long-Read Sequencing Data

A high-level depiction of the data analysis workflow is illustrated in Figure 2. After acquiring raw long-read sequencing data in the FAST5 format, guppy v4.4.1 (Oxford Nanopore Technology) with the High-Accuracy configuration for the R9.4.1 platform (dna_r9.4.1_450bps_hac.cfg) was used for basecalling. Basecalling was performed locally in Ubuntu 16.04 with CUDA acceleration on an Nvidia RTX 3080 GPU, providing a substantial performance uplift compared to basecalling with a local AMD 3700X CPU. All FASTQ files generated via this process were concatenated into a single file using the `cat` command.

2.3.1 Adapter Removal and Filtering

Porechop v0.2.4 was used to remove adapter sequences from the reads. Porechop first aligns a subset of the reads to a library of known adapters. After known adapter sequences are detected in this subset, the rest of the reads are aligned to the known adapter sequences; if a matching adapter sequence is found in the read, then the sequence is trimmed [24]. Filtlong v0.2.0 was then used to remove the worst reads, using a minimum length of 1kb [25]. Filtlong assigns an internal score to each read based on read length, mean quality score, and a sliding window quality score. For this long-read sequencing run, the default Filtlong score weighting was used, and the 10% worst-scoring reads were discarded.

2.3.2 Assembly

Three separate approaches were used to generate long-read, short-read, and hybrid assemblies. First, MEGAHIT, an assembler designed to accommodate metagenomic data [26], was used to generate an assembly from the entire set of short reads. Hybrid assembly was then performed using the MEGAHIT short-read assembly, the processed long and short reads, and the hybrid metagenomic assembler OPERA-MS v0.8.3, which uses Pilon v1.22 to polish the MEGAHIT short-read assembly [8], [27]. OPERA-MS assembly was performed with and without polishing, and the default recommended references from the Genome Taxonomy Database were used for reference-based clustering.

In the second approach, long reads were assembled using Flye v2.8.3 and the `--meta` flag for metagenomic assembly [11]. This assembly was then polished in two iterations using Flye's built-in polisher and the original long reads. This long-read assembly was finally polished using the short reads and Racon v1.4.20 [28] to generate the hybrid assembly.

In the final approach, Unicycler v0.4.4 was used to generate short-read, long-read, and hybrid metagenomic assemblies [14]. Note that Unicycler is designed to assemble sequences from bacterial isolates rather than metagenomic samples, and this approach is primarily exploratory.

2.3.3 Quality Control of Long and Short Reads

Quality control metrics were generated for both the basecalled long reads and the trimmed/filtered long reads using NanoPlot v1.32.1 and NanoQC v0.9.4 [29]. For the existing short reads, fastp v0.20.1 was used [30].

QUAST v5.02 was used to generate assembly metrics for the metagenomic assemblies [31]. MultiQC v1.10.1 was used to simultaneously and interactively visualize assembly metrics for multiple assemblies [32].

2.3.4 Taxonomic Classification and Annotation

Kraken 2 is a rapid and accurate taxonomic classifier that was used to assign taxonomy to sequences generated in this workflow [19]. The 12/2/2020 release of the pre-built k2_standard database containing archaea, bacteria, viral, plasmid, and human sequences was downloaded through the Kraken 2 project webpage. Classification by Kraken 2 v2.1.1 was performed on the short reads and short read assembly, the long reads and long-read assembly, and the hybrid assembly. The output from Kraken 2 was then parsed and visualized using Krona, which generates interactive taxonomic visualizations in HTML [33].

The Resistance Gene Identifier (RGI) was used to predict the resistome of the metagenomic contigs using reference data downloaded from the Comprehensive Antibiotic Resistance Database (CARD), a curated database providing reference sequences and tools for resistome monitoring and analysis [34]. In order to identify allelic variants and gene homologs of antibiotic resistance genes, RGI relies on the WildCARD dataset, which is comprised of CARD's "Resistomes & Variants" and "Prevalence Data" data. WildCARD v3.1.1 was used for this project, but different versions of WildCARD could result in different results as annotations change or grow over time.

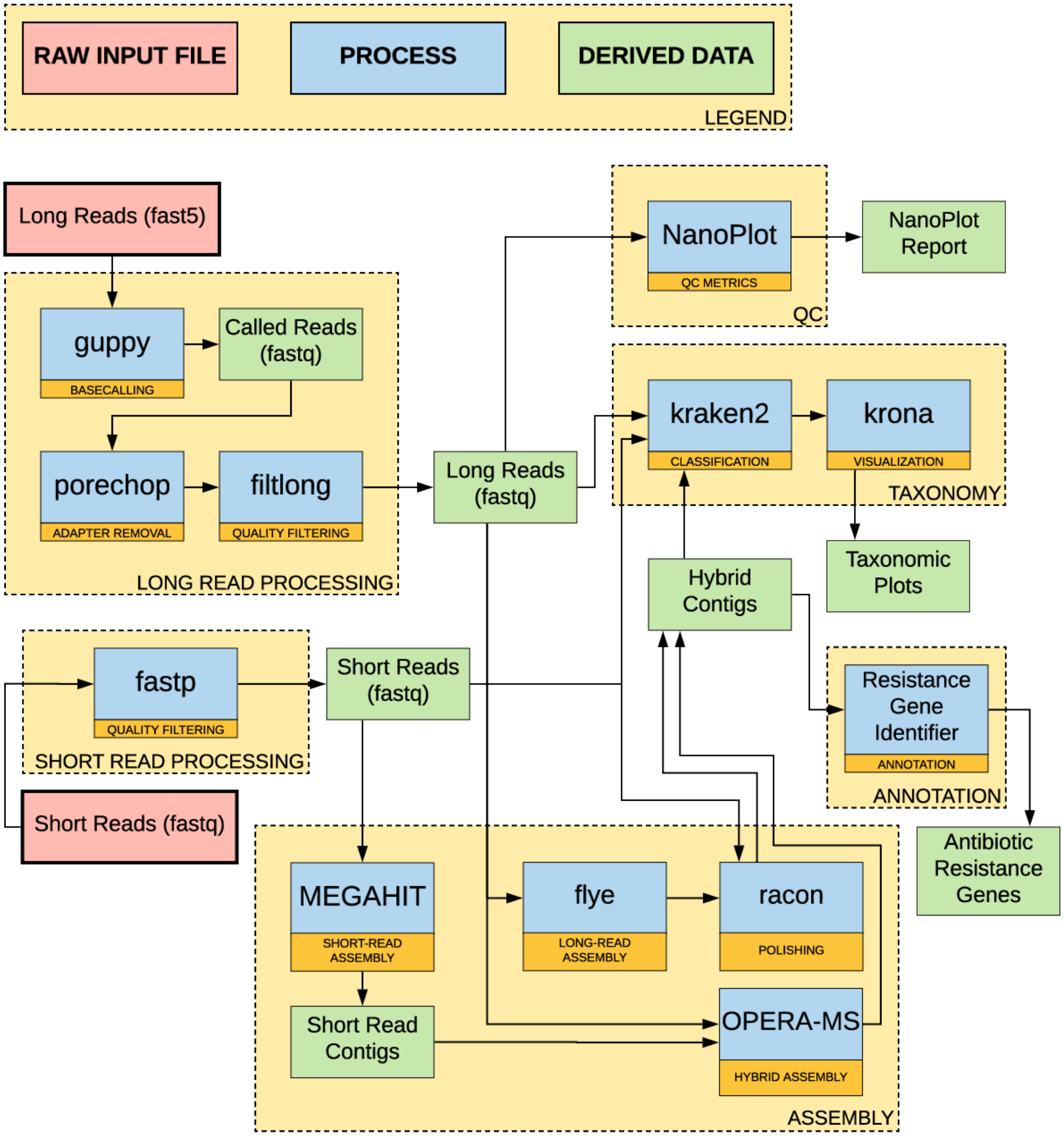


Figure 2. Overview of the sequence data analysis workflow.

2.4 Computing

2.4.1 Computing Resources

Computing was primarily performed on the San José State University College of Science High Performance Computing cluster (COS-HPC), a computing cluster available to University students, faculty, and staff. As a shared utility, the COS-HPC (funded by a $900,798 grant from the National Science Foundation, award ID #1626645) allocates requested resources to jobs according to current demand and availability. Jobs were submitted to the COS-HPC using the Simple Linux Utility for Resource Management (SLURM) Workload Manager [35], typically using the `srun` or `sbatch` commands. For jobs submitted using `sbatch`, shell scripts were written on the server using the `nano` command line text editor or remotely using Microsoft Visual Studio Code and the Remote-SSH extension. An example of a job submission script for `sbatch` is shown in Figure 3.

```
#!/bin/bash

#SBATCH --partition=nodes
#SBATCH --job-name=opera-ms
#SBATCH --output=./logs/opera_%j.out
#SBATCH --error=./logs/opera_%j.err
#SBATCH --ntasks=24
#SBATCH --ntasks-per-node=24
#SBATCH --mem=96G
#SBATCH --time=48:00:00
#SBATCH --verbose

# BEGIN SCRIPT
perl ~/repos/OPERA-MS/OPERA-MS.pl \
    --contig-file 03_megahit/results_short/short_contigs.fa \
    --short-read1 ./02_fastp_short/ww_1.fastq \
    --short-read2 ./02_fastp_short/ww_2.fastq \
    --long-read ./02_filtered/pool13_2_filt.fastq \
    --out-dir 04_hybrid \
    --num-processors 24 \
    --no-gap-filling
```

Figure 3. Example job submission script for SLURM.

This specific script is used to call perl and run the OPERA-MS tool for hybrid metagenomic assembly. Parameters specific to the sbatch command are specified in the header with the prefix #SBATCH.

Resource usage in bioinformatic workloads may scale in several key areas. For example, in genomic assembly, read length, library size, and reference genome size (if available) all directly contribute to the space and time complexity of a given operation [36]. Though there are many ways in which performance or resource usage can be improved, such as compression [37], many bioinformatic workloads can specifically benefit from parallelization, or the use of multiple computing threads. This is dependent on algorithms and approaches that can subdivide the primary objective into discrete tasks that are not immediately dependent on the output of previous tasks [38]. Given appropriate software tools, the 72 compute nodes of the COS-HPC allow for orders of magnitude improvements in performance compared to a single-threaded application. In addition, a high-capacity scratch space and 128GB of RAM per compute node enable analyses that would otherwise encounter prohibitive storage and memory limits on a personal computer. After completing the most demanding computational workloads on the COS-HPC, data was downloaded from the COS-HPC to a personal computer for local analysis as necessary using Windows Subsystem for Linux 2 (WSL2) and `rsync`.

2.4.2 Environments and Package Management

Manual installation of packages and dependencies can present a major hindrance to workflows that incorporate many packages. To this end, the Conda package manager was used to create and manage environments, collect the appropriate package versions, and resolve dependency trees [39]. Because of the wide variety of tools used in this bioinformatic workflow, it may also be preferable to execute different stages of the workflow in separate environments. This reduces the likelihood of different packages experiencing conflicting dependencies. To this end, separate environments were created for each analysis step using Conda. Another benefit of this practice is that it enables later versions of packages to be used, since each environment's dependency tree is smaller and less likely to generate conflicts. This has been particularly important in the use of long-read sequencing software, as this has been an area of very active development, resulting in constant updates to functionality, version numbers, and dependencies. In total, 13 separate Conda environments were generated and used for these analyses.

For tools that were not available via Conda package manager, such as OPERA-MS, the source code was downloaded and compiled on the COS-HPC. The resulting binaries were saved to and executed from the `bin` folder located in the user's home directory.

3 RESULTS

3.1 Quality Control of Long and Short Reads

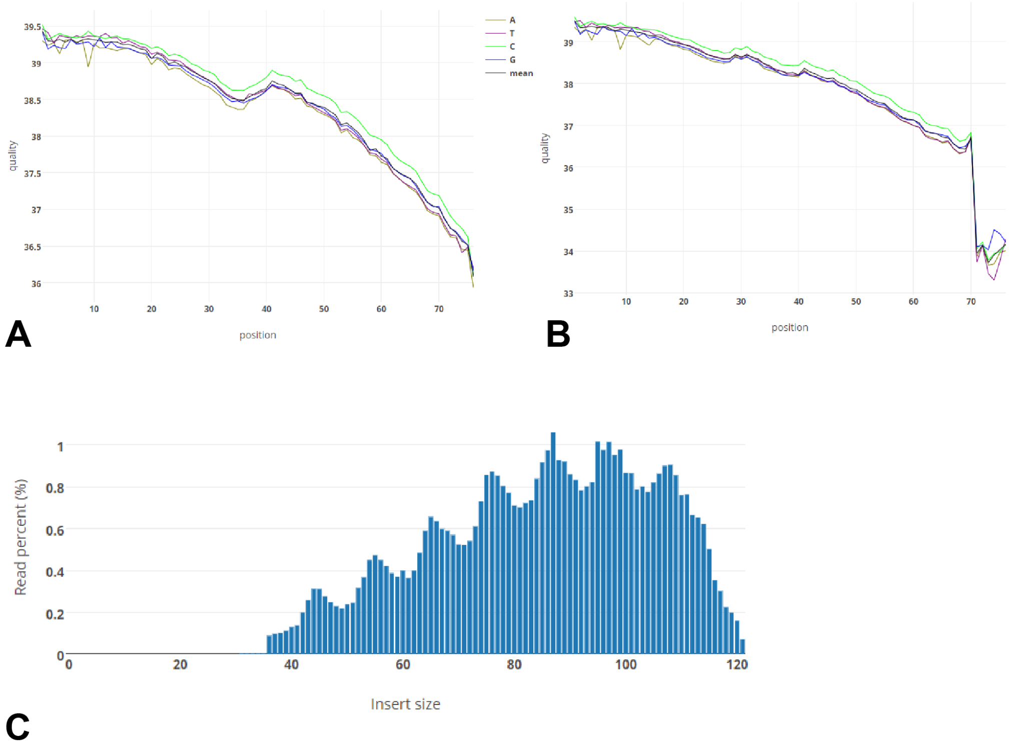


Figure 4. fastp report for Illumina HiSeq 1000 short reads.

Mean per-base quality scores for forward (A) and reverse reads (B). Insert size distribution (C).

The sequencing dataset from the Illumina HiSeq 1000 contained a total of 36Gb of short reads. Of these, fastp reported that 97.47% of the bases had a quality score of at least 30. Except for the tail end of the reverse reads, each base position had a mean quality score of at least 36, indicating high confidence in the base accuracy (Figure 4A and B). Insert sizes ranged from 36bp to 122bp, with a peak at 87bp (Figure 4C).

A 24-hour sequencing run with the MinION generated 6.35Gb of long reads with a mean read length of 7.4kb. For comparison, reads generated on the Illumina platform are commonly 150bp pairs (2 x 150bp). Basecalling with guppy yielded a total of 854,745 reads. After adapter trimming with Porechop, the nucleotide frequencies of the heads and tails of the reads improved significantly, though it does appear that some adapter sequences remain, visible in Figure 5B and C in approximately the first and last 40bp of each read.

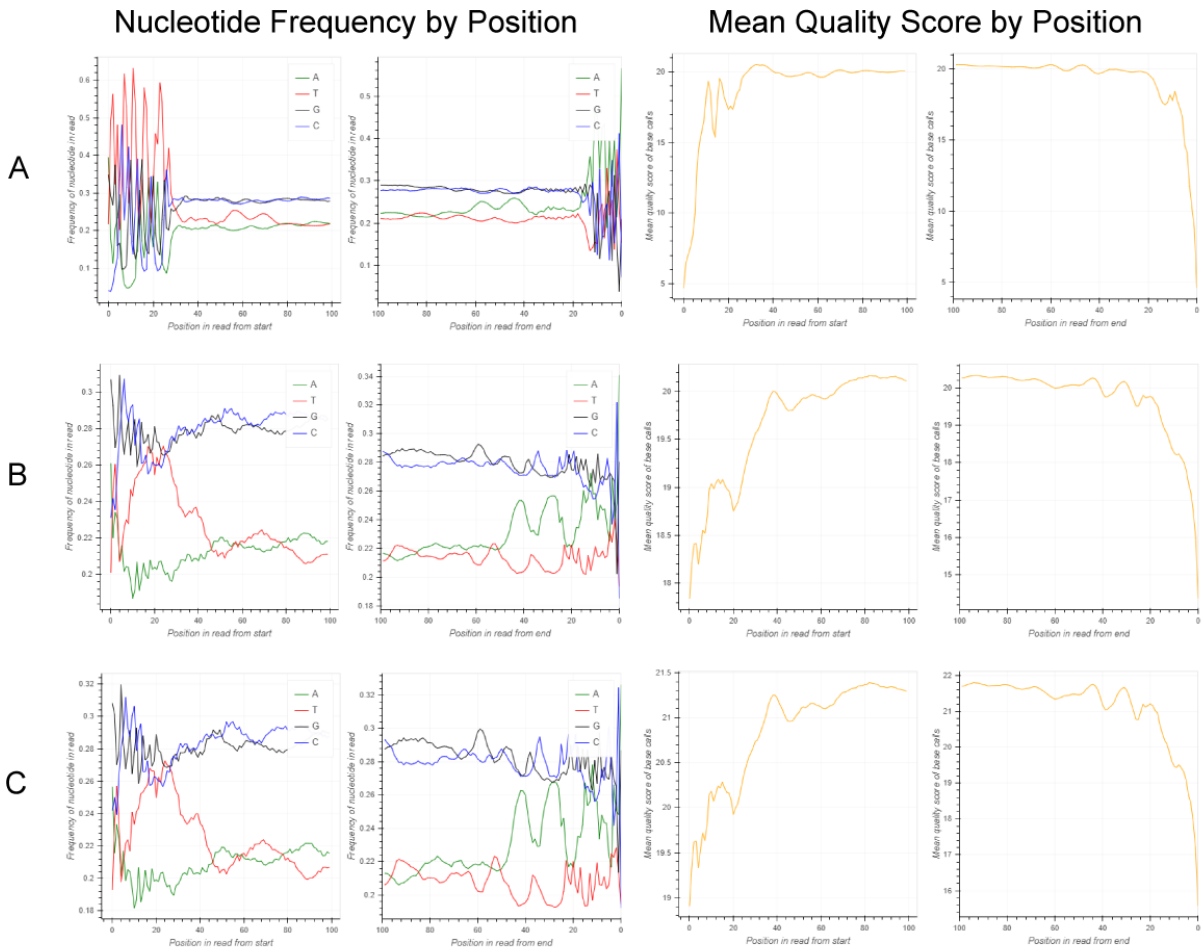


Figure 5. NanoQC plots for adapter removal and trimming results.

(A) Raw basecalled reads. (B) Reads after adapter removal. (C) Reads after quality trimming.

Quality filtering performed with Filtlong eliminated the reads with the worst mean quality scores, resulting in 576,977 remaining long-read sequences (Figure 5). Although the quality scores in the heads and tails of the reads were still lower than the middles of the reads, they were still suitable for use in assembly.

After quality processing, was calculated for the long-read dataset. Each fosmid insert contains roughly 40kb of genomic DNA, and Pool13 contains roughly 3300 clones. The fosmid library size can therefore be approximated at 132Mb. After adapter removal, filtering, and quality trimming, the long-read sequencing run generated 5.7Gb, for an average of 43X coverage.

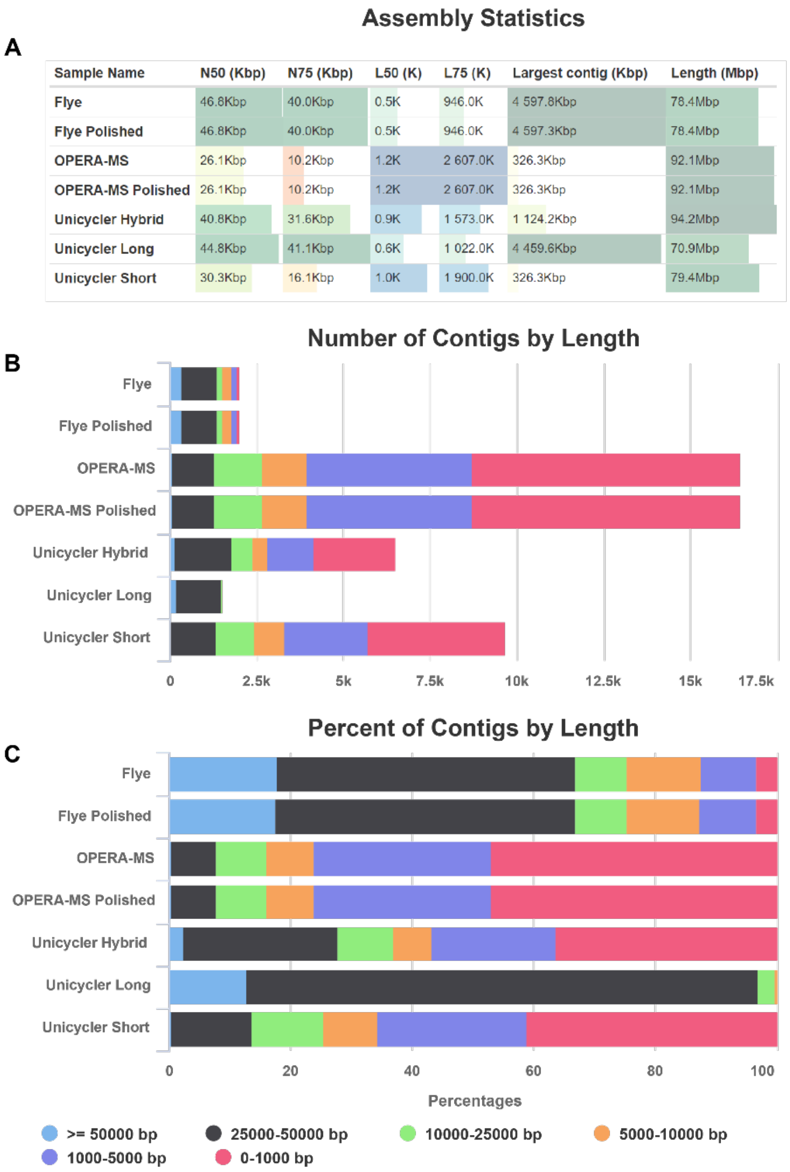
3.2 Assembly

Seven metagenomic assemblies were generated: two short-read assemblies, two long-read assemblies, and three hybrid assemblies. QUAST metrics for all seven assemblies are shown in Figure 6A. All of the short-read assemblies ("OPERA-MS", "OPERA-MS Polished", and "Unicycler Short") had more contigs and shorter contigs than the long-read and hybrid assemblies ("Flye", "Flye Polished", "Unicycler Long", "Unicycler Hybrid") Figure 6B. The Flye long-read assembly and Flye + Racon hybrid assembly exhibited high contiguity, with an N50 of 46.8Kb and L50 of 500 contigs (out of 2023 contigs). Furthermore, almost 20% of all contigs in these two assemblies were longer than 50Kb (Figure 6C). In contrast, the OPERA-MS-based assemblies had very few contigs longer than 50Kb, with N50 of 26Kb and L50 of 1244 (out of 16450 contigs).

3.3 Taxonomic Classification of Reads and Contigs

Taxonomic classification of both the Illumina short reads and the Nanopore long reads using Kraken 2 showed fewer unclassified reads in the latter (Figure 7). While 56% of short reads were unclassified, only 5% of long reads were unclassified. Assembly of short reads also reduced the frequency of non-hits. The OPERA-MS short-read-only assembly had a non-hit frequency of

29%, and the Flye long-read-only assembly had a non-hit frequency of 0.8% (a six-fold improvement). Polishing did not visibly affect taxonomic classification results.



Sample Name	N50 (Kbp)	N75 (Kbp)	L50 (K)	L75 (K)	Largest contig (Kbp)	Length (Mbp)
Flye	46.8Kbp	40.0Kbp	0.5K	946.0K	4 597.8Kbp	78.4Mbp
Flye Polished	46.8Kbp	40.0Kbp	0.5K	946.0K	4 597.3Kbp	78.4Mbp
OPERA-MS	26.1Kbp	10.2Kbp	1.2K	2 607.0K	326.3Kbp	92.1Mbp
OPERA-MS Polished	26.1Kbp	10.2Kbp	1.2K	2 607.0K	326.3Kbp	92.1Mbp
Unicycler Hybrid	40.8Kbp	31.6Kbp	0.9K	1 573.0K	1 124.2Kbp	94.2Mbp
Unicycler Long	44.8Kbp	41.1Kbp	0.6K	1 022.0K	4 459.6Kbp	70.9Mbp
Unicycler Short	30.3Kbp	16.1Kbp	1.0K	1 900.0K	326.3Kbp	79.4Mbp

Figure 6. QUAST-derived metrics for seven metagenomic assemblies.

(A) Assembly statistics. (B) Number of contigs by length. (C) Percent of contigs by length.

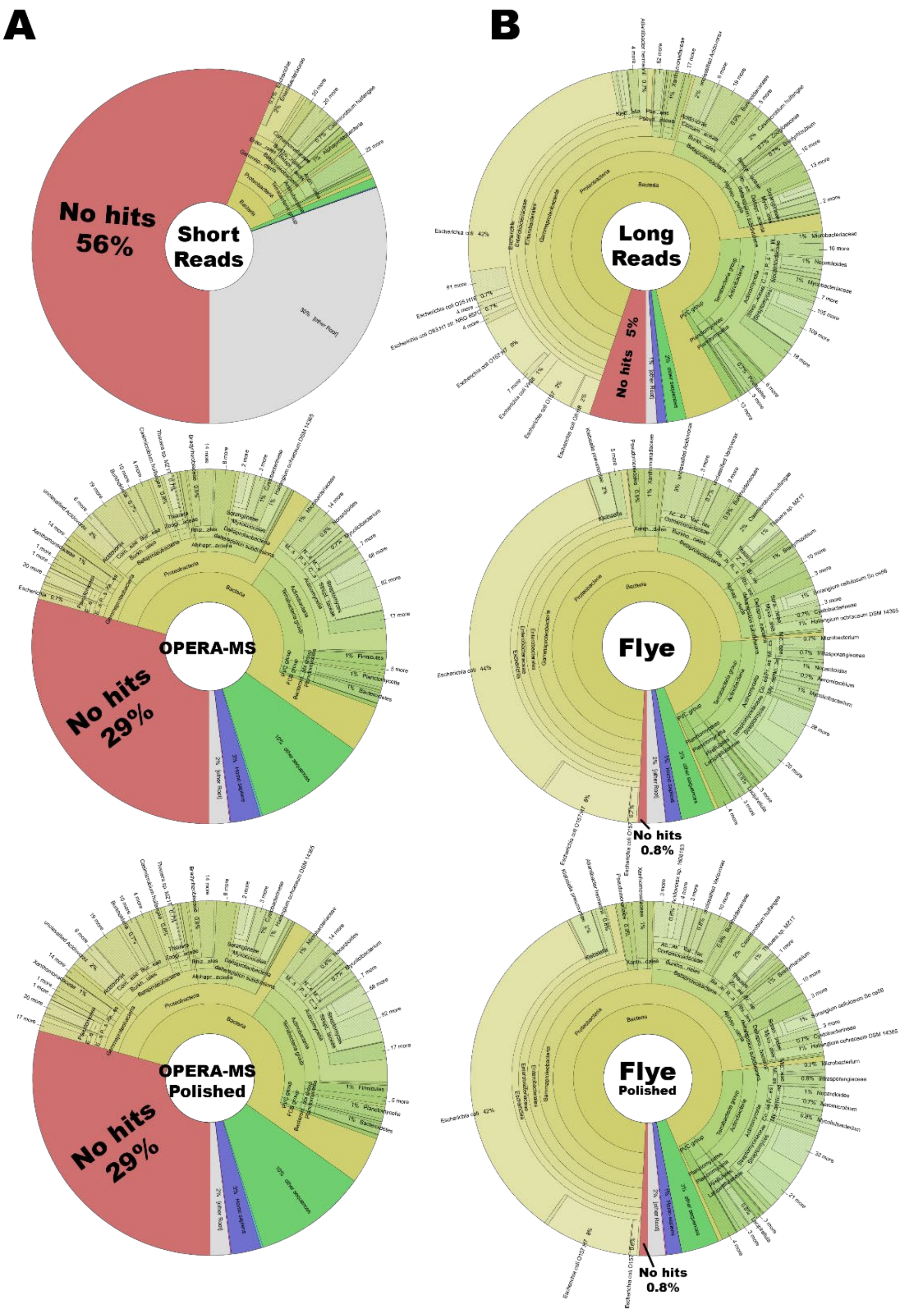


Figure 7. Taxonomic classification of long reads shows reduced frequency of non-hits compared to short reads. Assembly also reduces frequency of non-hits.

(A) From top to bottom: Short reads, the OPERA-MS short-read assembly, and polished OPERA-MS assembly. (B) Long reads, the Flye long-read assembly, and Racon-polished long-read assembly

3.4 Identification of Resistance Genes

The RGI aligned 83 ARGs to the polished contigs generated using the OPERA-MS hybrid assembler. A heat map was generated from the RGI results depicting the expression of the detected antibiotic resistance genes and their families (Figure 8). ARG hits are classified as either Strict matches (>95% identity) or Perfect matches (100% identity). A comparison of hit types and hit counts is shown in Table 3. After polishing with short reads, the Flye hybrid assembly showed 15 additional Perfect hits compared to the unpolished assembly, and 16 fewer Strict hits. On closer examination, 15 Strict hits were upgraded to Perfect hits, 4 Strict hits were downgraded to non-hits, and 3 new Strict hits were observed. In the OPERA-MS assemblies, one additional ARG (antibiotic resistant LpsB) was detected in the polished hybrid assembly compared to the unpolished short-read assembly. LpsB encodes a glycotransferase that is involved in lipopolysaccharide (LPS) synthesis, contributing to resistance by disrupting the interaction of LPS with the antibiotic colistin [40]. Furthermore, 18 ARGs were detected in the polished hybrid assembly that were not detected in the pCC1FOS vector or the E. coli DH10B reference genome (Table 4).

Table 3. Number of ARGs detected per assembly strategy

Assembly and Sequencing Type		**Total ARG Hits**	**Strict ARG Hits**	**Perfect ARG Hits**
Short-Read Assembly *(Abrams et al.)*		46	N/A	N/A
Pool13 Shotgun Metagenomic Sequencing Assemblies *(Total)*		94	45	49
Long-Read Assembly	**Flye**	74	45	29
	Flye Polished	73	29	44
Short-Read Assembly	**OPERA-MS**	82	33	49
	OPERA-MS Polished	83	34	49
Reference	*E coli* DH10B	64	19	45

Table 4. Non-fosmid-system-associated ARGs detected in OPERA-MS hybrid assembly

Gene Family	Gene Name
AAC(2')	AAC(2')-IIa
ANT(3")	aadA13
APH(3')	APH(3')-IIa
CTX-M beta-lactamase	CTX-M-130
Cfr 23S ribosomal RNA methyltransferase	cfrC
IDC beta-lactamase	IDC-2
Intrinsic peptide antibiotic resistant Lps	LpsB
OXA beta-lactamase	OXA-205 OXA-664
TEM beta-lactamase	TEM-116
major facilitator superfamily (MFS) antibiotic efflux pump	Rhodococcus fascians cmr pexA qacEdelta1
pmr phosphoethanolamine transferase	ArnT basS
resistance-nodulation-cell division (RND) antibiotic efflux pump	adeF opmE
trimethoprim resistant dihydrofolate reductase dfr	dfrA1

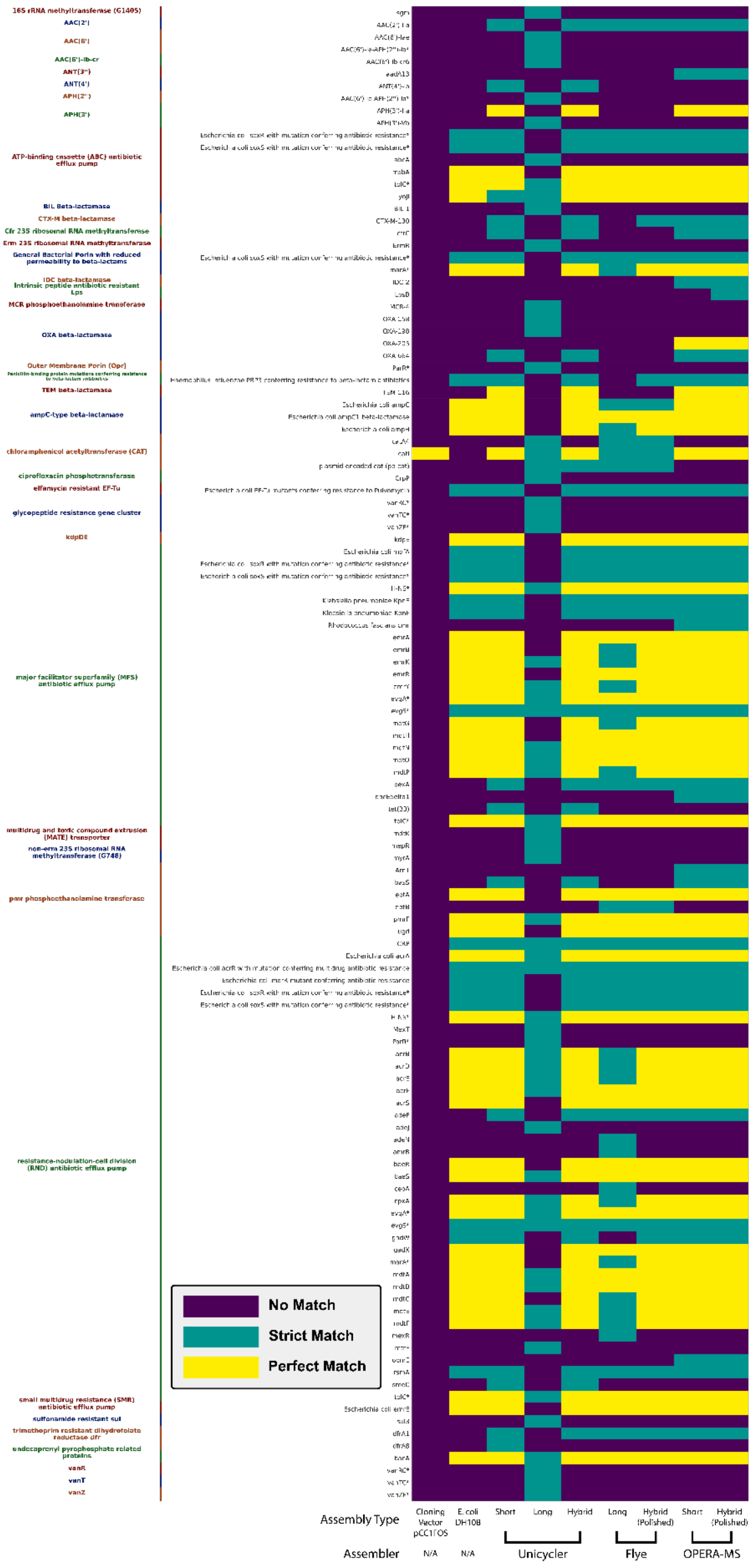


Figure 8. Heat map of detected ARGs.

4 DISCUSSION

4.1 Current Practice of Metagenomic Hybrid Assembly Workflows

The laboratory user experience with the Nanopore platform was positive, especially in regard to cost, labor, and documentation. Nanopore's Genomic DNA by Ligation kit and protocol allowed the entire fosmid library to be sequenced at 43X coverage for under $200 worth of reagents (not including the MinION flow cell or the MinION Mk1B). Additionally, preparing the sequencing library and starting up the sequencing process took under 3 hours. Sequencing on the Nanopore platform will likely continue to improve in accessibility and ease of use, and further development in downstream long-read analysis tools is expected.

As demonstrated above, long-read sequencing with the Nanopore platform enables hybrid assembly to be performed, potentially enhancing recovery of sequences from metagenomic DNA samples by combining the high per-base accuracy of Illumina reads with contiguity provided by the Nanopore and SMRT platforms. Compared with the ease of sequencing on the Nanopore platform, the current ecosystem of tools enabling hybrid assembly in a metagenomic context remains an area of much-needed development. Many tools do not specifically accommodate metagenomic libraries, and instead focus on single-genome datasets. The Unicycler assembler is one example; because it was explicitly designed for handling reads from isolated bacterial species [14], it poorly scales when assembling large metagenomic datasets [12]. In addition, the bioinformatics community has yet to converge on standardized workflows for metagenomic hybrid assembly. Efforts to rectify this problem are reflected in the continued development of application-specific workflows and pipelines [41]–[43].

4.2 Fosmid Library Considerations

Because the DNA samples used for library preparation were derived from previously-generated fosmid inserts, the pooled sequencing results likely do not represent full coverage of the genomes present in the original DNA extraction. Note that due to the limited capture space of the fosmid system, recapitulation of entire genomes for less-common species is unlikely, presenting a technical limitation that cannot be surmounted by increasing sequencing coverage. Furthermore, there are two possible sources of DNA contamination: the fosmid vector (pCC1FOS) and the DH10B competent cells used to propagate the fosmids. The fosmid cloning system does present benefits in other aspects, particularly in the isolation and targeted cultivation of clones containing genes of interest, making it a powerful tool for molecular biology applications. It should not, however, be considered equivalent to a fresh DNA extraction from a wastewater sample.

4.3 Long Read Assembly and Hybrid Assembly

The Flye long-read assembly featured high contiguity and excellent taxonomic matching, per the QUAST (Figure 6) and Kraken 2/Krona results (Figure 7). The Flye assembly also generated a 4.6Mb contig, which is expected to correspond to the *E. coli* chromosome, though this binning step has not been performed. By the assembly metrics, the long-read sequencing and assembly process was successful in generating a high-contiguity assembly, especially compared to the short-read derived assemblies.

The polished hybrid assembly generated using OPERA-MS did not greatly differ from the short-read assembly with MEGAHIT. Contig metrics generated by QUAST for these two assemblies were very similar, indicating that contiguity was not significantly affected by the polishing process with short and long reads. OPERA-MS, though effective at recovering low-abundance bacterial genomes from metagenomic samples, has also been shown to generate less contiguity than assemblers like Flye and Canu [44]. This was replicated in the current study (Figure 6). Note that OPERA-MS was unable to be run with gap-filling enabled due to memory constraints on the COS-HPC; further investigation is required to determine if this could be remedied using the computing cluster's high-memory nodes.

One consideration with this dataset is that the nature of the fosmid library and the high coverage of the short-read sequencing dataset could mean that near-optimal assembly of the fosmid library has already been achieved by the short-read assembly. Because the fosmid inserts are already a subset of the wastewater metagenome, deeper sequencing of the inserts is unlikely to further improve contiguity; as a result, the addition of long reads did not appear to substantially modify the assembly metrics. Further analysis would be required to determine the extent to which the polishing process modified sequences within the contigs. To determine the contribution of the long-read and short-read datasets to the hybrid assemblies, it may be of value to repeat hybrid assembly using subsamples of the input reads. For example, by reducing the input read datasets to 10X coverage each for both the short and long reads, a larger difference between hybrid and short-read-only assembly might be observed. This would be an effective way of determining whether there exists a critical coverage level or ratio of short to long reads that strongly justifies a hybrid assembly strategy.

4.4 ARG DETECTION

Resistance gene identification was performed on the fosmid vector sequence as well as the *E. coli* DH10B strain in order to detect any ARGs associated with the fosmid DNA library system and host strain. One ARG, the chloroamphenicol acetyltransferase gene *catI,* was detected in the fosmid vector sequence. This is expected, as one of the advertised features of the vector is chloramphenicol resistance, which enables isolation of successfully-transformed cells in culture.

Many of the ARGs detected in the metagenomic assemblies are also found in the reference sequence for the DH10B host. Further processing could be performed to exclude *E. coli*-mapped reads from the assembly process, as well as use a tool such as bbsplit to remove any remaining vector sequences.

Fewer total ARG hits were observed in the long-read-based assemblies compared to the short-read-based assemblies. Even among the hybrid assemblies, the polished Flye assembly had 10 fewer hits than the OPERA-MS polished assembly. Because RGI detects hits based on percent sequence identity, shorter ARGs could be more easily lost due to sequencing artifacts compared to longer ARGs. The OPERA-MS polished hybrid assembly did result in detection of an additional ARG (LpsB) compared to the OPERA-MS short-read-only assembly. This suggests that greater accuracy derived from the polishing process enabled alignment of a contig with the LpsB sequence, though this would need to be confirmed with targeted molecular assays. Further analysis is required to compare the sets of Strict and Perfect hits for each assembly against each other to determine if there is a bias in which types of ARGs each assembly is able to recover.

Finally, compared to previous research conducted using this fosmid DNA library, the assemblies presented in this project identified 48 more ARGs (Table 3). Additional analysis is required to determine the overlap between the 46 ARGs identified by Abrams et al. and the 94 ARGs detected with the current short-read, long-read, and hybrid assemblies. Several factors could cause this difference. First, Abrams et al. sequenced 38 fosmid clones, while the current metagenomic study sequenced over 3000 clones. The second factor is that different workflows were used for analysis, with the previous resistome-characterization step performed by MG-RAST instead of RGI. In addition, reference sequences, databases, and tools could have significantly changed since the Abrams et al. research was conducted. Finally, long-read and hybrid assembly could have improved the resolution and recovery of ARG sequences, enabling them to be identified via RGI.

5 CONCLUSION

Hybrid assembly enables greater recovery of genomes and genes of interest from metagenomic environmental samples [12], [13], [23], [43]. The ability to successfully detect antibiotic resistance genes in wastewater samples using a shotgun metagenomic sequencing strategy marks a key milestone in developing information-driven approaches to epidemiology and public health. The workflow demonstrated here shows that long-read datasets can be quickly acquired, processed, assembled, and screened for ARGs. This ability to rapidly generate datasets at a low cost could facilitate longitudinal monitoring of dynamic environments, such as the wastewater resistome. As antibiotic selective pressure drives the emergence and transfer of ARGs, the ability to measure trends in urban and agricultural resistomes could help mitigate the over $2 billion annual cost of treating antibiotic-resistant infections, as well as the additional agricultural costs of lost or non-usable livestock. Antimicrobial resistance rates have also been shown to vary spatially and seasonally, furthering the need for broadly applicable monitoring systems and analysis workflows [45], [46]. As analysis pipelines, software, and sequencing technology continue to mature, hybrid assembly will surely continue facilitating advances in metagenomics and other bioinformatic applications.

6 REFERENCES

[1] K. Kumar, S. C. Gupta, Y. Chander, and A. K. Singh, "Antibiotic Use in Agriculture and Its Impact on the Terrestrial Environment," in *Advances in Agronomy*, vol. 87, Academic Press, 2005, pp. 1–54. doi: 10.1016/S0065-2113(05)87001-4.

[2] K. E. Thorpe, P. Joski, and K. J. Johnston, "Antibiotic-Resistant Infection Treatment Costs Have Doubled Since 2002, Now Exceeding $2 Billion Annually," *Health Aff. (Millwood)*, vol. 37, no. 4, pp. 662–669, Mar. 2018, doi: 10.1377/hlthaff.2017.1153.

[3] V. M. D'Costa *et al.*, "Antibiotic resistance is ancient," *Nature*, vol. 477, no. 7365, Art. no. 7365, Sep. 2011, doi: 10.1038/nature10388.

[4] K. Bhullar *et al.*, "Antibiotic resistance is prevalent in an isolated cave microbiome," *PloS One*, vol. 7, no. 4, p. e34953, 2012, doi: 10.1371/journal.pone.0034953.

[5] M. R. Rondon *et al.*, "Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms," *Appl. Environ. Microbiol.*, vol. 66, no. 6, pp. 2541–2547, Jun. 2000, doi: 10.1128/AEM.66.6.2541-2547.2000.

[6] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends Genet.*, vol. 30, no. 9, pp. 418–426, Sep. 2014, doi: 10.1016/j.tig.2014.07.001.

[7] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, "Continuous base identification for single-molecule nanopore DNA sequencing," *Nat. Nanotechnol.*, vol. 4, no. 4, Art. no. 4, Apr. 2009, doi: 10.1038/nnano.2009.12.

[8] D. Bertrand *et al.*, "Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes," *Nat. Biotechnol.*, vol. 37, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s41587-019-0191-2.

[9] R. M. Leidenfrost, D.-C. Pöther, U. Jäckel, and R. Wünschiers, "Benchmarking the MinION: Evaluating long reads for microbial profiling," *Sci. Rep.*, vol. 10, no. 1, p. 5125, Dec. 2020, doi: 10.1038/s41598-020-61989-x.

[10] F. Maguire, B. Jia, K. L. Gray, W. Y. V. Lau, R. G. Beiko, and F. S. L. Brinkman, "Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands," *Microb. Genomics*, vol. 6, no. 10, Oct. 2020, doi: 10.1099/mgen.0.000436.

[11] M. Kolmogorov *et al.*, "metaFlye: scalable long-read metagenome assembly using repeat graphs," *Nat. Methods*, vol. 17, no. 11, Art. no. 11, Nov. 2020, doi: 10.1038/s41592-020-00971-x.

[12] L. Liu *et al.*, "High-quality bacterial genomes of a partial-nitritation/anammox system by an iterative hybrid assembly method," *Microbiome*, vol. 8, no. 1, p. 155, Nov. 2020, doi: 10.1186/s40168-020-00937-3.

[13] H. Xie, C. Yang, Y. Sun, Y. Igarashi, T. Jin, and F. Luo, "PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning," *Front. Genet.*, vol. 11, 2020, doi: 10.3389/fgene.2020.516269.

[14] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, "Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads," *PLOS Comput. Biol.*, vol. 13, no. 6, p. e1005595, Jun. 2017, doi: 10.1371/journal.pcbi.1005595.

[15] D. D. Kang *et al.*, "MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, Jul. 2019, doi: 10.7717/peerj.7359.

[16] L. S. Frost, R. Leplae, A. O. Summers, and A. Toussaint, "Mobile genetic elements: the agents of open source evolution," *Nat. Rev. Microbiol.*, vol. 3, no. 9, pp. 722–732, Sep. 2005, doi: 10.1038/nrmicro1235.

[17] E. L. Moss, D. G. Maghini, and A. S. Bhatt, "Complete, closed bacterial genomes from microbiomes using nanopore sequencing," *Nat. Biotechnol.*, vol. 38, no. 6, Art. no. 6, Jun. 2020, doi: 10.1038/s41587-020-0422-6.

[18] S. M. Nicholls, J. C. Quick, S. Tang, and N. J. Loman, "Ultra-deep, long-read nanopore sequencing of mock microbial community standards," *GigaScience*, vol. 8, no. 5, May 2019, doi: 10.1093/gigascience/giz043.

[19] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2," *Genome Biol.*, vol. 20, no. 1, p. 257, Dec. 2019, doi: 10.1186/s13059-019-1891-0.

[20] M. E. Abrams *et al.*, "Metagenomic characterization of a microbial community in wastewater detects high levels of antibiotic resistance genes."

[21] H. Li, *lh3/seqtk*. 2021. Accessed: May 23, 2021. [Online]. Available: https://github.com/lh3/seqtk

[22] T. Durfee *et al.*, "The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse," *J. Bacteriol.*, vol. 190, no. 7, pp. 2597–2606, Apr. 2008, doi: 10.1128/JB.01695-07.

[23] E. Haghshenas, H. Asghari, J. Stoye, C. Chauve, and F. Hach, "HASLR: Fast Hybrid Assembly of Long Reads," *iScience*, vol. 23, no. 8, Aug. 2020, doi: 10.1016/j.isci.2020.101389.

[24] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, "Completing bacterial genome assemblies with multiplex MinION sequencing," *Microb. Genomics*, vol. 3, no. 10, Oct. 2017, doi: 10.1099/mgen.0.000132.

[25] R. Wick, *rrwick/Filtlong*. 2021. Accessed: May 03, 2021. [Online]. Available: https://github.com/rrwick/Filtlong

[26] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph," *Bioinforma. Oxf. Engl.*, vol. 31, no. 10, pp. 1674–1676, May 2015, doi: 10.1093/bioinformatics/btv033.

[27] B. J. Walker *et al.*, "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement," *PLOS ONE*, vol. 9, no. 11, p. e112963, Nov. 2014, doi: 10.1371/journal.pone.0112963.

[28] R. Vaser, I. Sovic, N. Nagarajan, and M. Sikic, "Fast and accurate de novo genome assembly from long uncorrected reads," *Genome Res.*, p. gr.214270.116, Jan. 2017, doi: 10.1101/gr.214270.116.

[29] W. De Coster, S. D'Hert, D. T. Schultz, M. Cruts, and C. Van Broeckhoven, "NanoPack: visualizing and processing long-read sequencing data," *Bioinformatics*, vol. 34, no. 15, pp. 2666–2669, Aug. 2018, doi: 10.1093/bioinformatics/bty149.

[30] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "fastp: an ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018, doi: 10.1093/bioinformatics/bty560.

[31] A. Mikheenko, V. Saveliev, and A. Gurevich, "MetaQUAST: evaluation of metagenome assemblies," *Bioinforma. Oxf. Engl.*, vol. 32, no. 7, pp. 1088–1090, Apr. 2016, doi: 10.1093/bioinformatics/btv697.

[32] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.

[33] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a Web browser," *BMC Bioinformatics*, vol. 12, p. 385, Sep. 2011, doi: 10.1186/1471-2105-12-385.

[34] B. P. Alcock *et al.*, "CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D517–D525, Jan. 2020, doi: 10.1093/nar/gkz935.

[35] A. B. Yoo, M. A. Jette, and M. Grondona, "SLURM: Simple Linux Utility for Resource Management," in *Job Scheduling Strategies for Parallel Processing*, Berlin, Heidelberg, 2003, pp. 44–60. doi: 10.1007/10968987_3.

[36] S. Baichoo and C. A. Ouzounis, "Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment," *Biosystems*, vol. 156–157, pp. 72–85, Jun. 2017, doi: 10.1016/j.biosystems.2017.03.003.

[37] G. Benoit *et al.*, "Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph," *BMC Bioinformatics*, vol. 16, no. 1, p. 288, Sep. 2015, doi: 10.1186/s12859-015-0709-7.

[38] F. E. Psomopoulos and P. A. Mitkas, "Bioinformatics algorithm development for Grid environments," *J. Syst. Softw.*, vol. 83, no. 7, pp. 1249–1257, Jul. 2010, doi: 10.1016/j.jss.2010.01.051.

[39] "Conda — Conda documentation." https://docs.conda.io/en/latest/ (accessed May 01, 2021).

[40] M. I. Hood, K. W. Becker, C. M. Roux, P. M. Dunman, and E. P. Skaar, "Genetic Determinants of Intrinsic Colistin Tolerance in Acinetobacter baumannii," *Infect. Immun.*, vol. 81, no. 2, pp. 542–551, Feb. 2013, doi: 10.1128/IAI.00704-12.

[41] D. H. Huson *et al.*, "MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs," *Biol. Direct*, vol. 13, Apr. 2018, doi: 10.1186/s13062-018-0208-7.

[42] S. Krakau *et al.*, *nf-core/mag: mag 1.2.0 - Yellow Squirrel*. Zenodo, 2021. doi: 10.5281/ZENODO.3589527.

[43] R. Van Damme, M. Hölzer, A. Viehweger, B. Müller, E. Bongcam-Rudloff, and C. Brandt, "Metagenomics workflow for hybrid assembly, differential coverage binning, transcriptomics and pathway analysis (MUFFIN)," Bioinformatics, preprint, Feb. 2020. doi: 10.1101/2020.02.08.939843.

[44] Y. Hu, L. Fang, C. Nicholson, and K. Wang, "Implications of Error-Prone Long-Read Whole-Genome Shotgun Sequencing on Characterizing Reference Microbiomes," *iScience*, vol. 23, no. 6, p. 101223, Jun. 2020, doi: 10.1016/j.isci.2020.101223.

[45] A. Erb, T. Stürmer, R. Marre, and H. Brenner, "Prevalence of antibiotic resistance in Escherichia coli: overview of geographical, temporal, and methodological variations," *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.*, vol. 26, no. 2, pp. 83–90, Feb. 2007, doi: 10.1007/s10096-006-0248-2.

[46] E. G. Ramsey *et al.*, "Seasonal variation in antimicrobial resistance rates of community-acquired Escherichia coli bloodstream isolates," *Int. J. Antimicrob. Agents*, vol. 54, no. 1, pp. 1–7, Jul. 2019, doi: 10.1016/j.ijantimicag.2019.03.010.