Hybrid Sequencing Facilitates Robust *De Novo* Plasmid Assembly

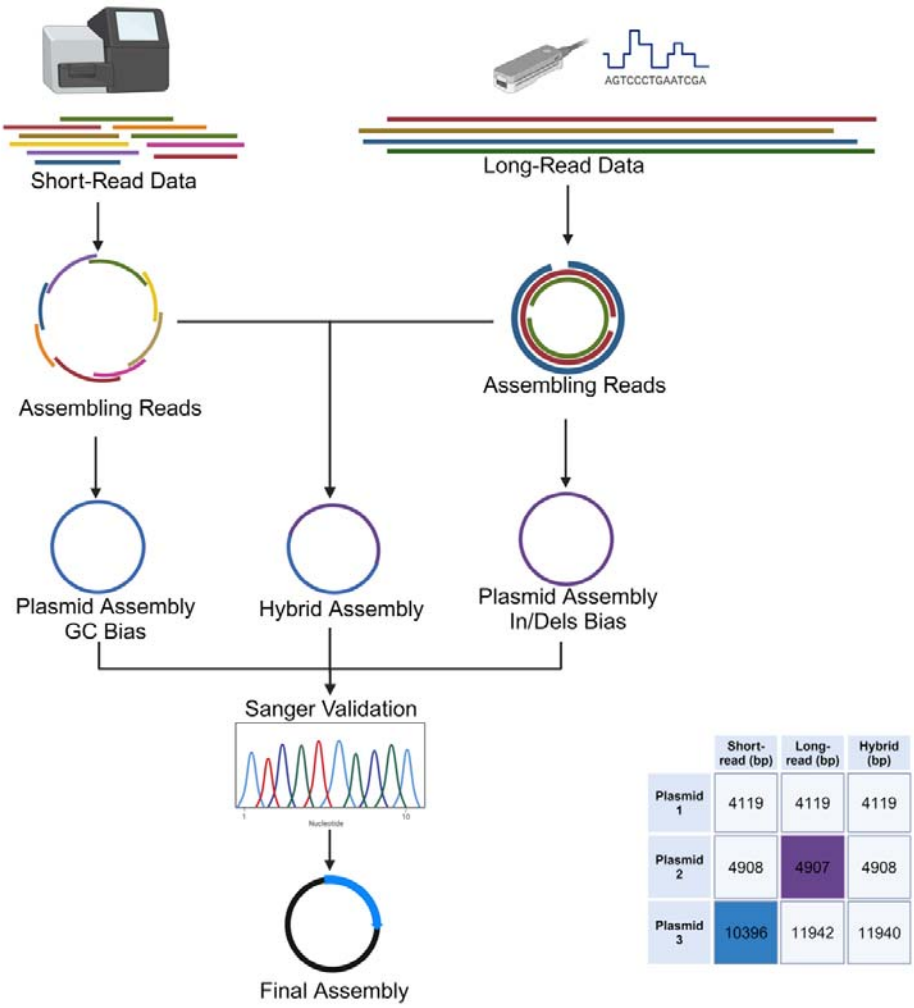
Sarah I. Hernandez[1], Casey-Tyler Berezin[1], Katie M. Miller[1], Samuel J. Peccoud[1], and Jean Peccoud[1,*]

[1] Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, Colorado, 80523, United States of America
* To whom correspondence should be addressed. Tel: [1-970-491-2482]; Email: jean.peccoud@colostate.edu

ABSTRACT

Despite the wide use of plasmids in research and clinical production, the verification of plasmid sequences is a bottleneck that is too often overlooked in the manufacturing process. Although sequencing platforms continue to improve, the method and assembly pipeline chosen still influence the final plasmid assembly sequence. Furthermore, few dedicated tools exist for plasmid assembly, especially for *de novo* assembly. Here, we evaluated short-read, long-read, and hybrid (both short and long reads) *de novo* assembly pipelines across three replicates of a 24-plasmid library. Consistent with previous characterizations of each sequencing technology, short-read assemblies had frequent issues resolving GC-rich regions, and long-read assemblies commonly had small insertions and deletions, especially in repetitive regions. The hybrid approach facilitated the most consistent assembly generation. Although Sanger sequencing can be used to verify specific regions, it requires a reference sequence to design primers, emphasizing the need for accurate *de novo* plasmid assembly tools. Some GC-rich and repetitive regions were difficult to resolve using any methods, suggesting that easily sequenced genetic parts should be prioritized in the design of new genetic constructs.

GRAPHICAL ABSTRACT



	Short-read (bp)	Long-read (bp)	Hybrid (bp)
Plasmid 1	4119	4119	4119
Plasmid 2	4908	4907	4908
Plasmid 3	10396	11942	11940

INTRODUCTION

Plasmids are critical tools used in research, industrial, and clinical settings for applications such as recombinant gene expression, designing genetic circuits, and the generation of clinical products like vaccines [1-8]. Sequence verification of these increasingly large plasmid libraries is critical to ensuring the expected product is made and to evaluate the biological effects of spontaneous and intentional mutations, including single nucleotide polymorphisms (SNPs). [2-5, 9]. Although DNA sequences are generally designed and documented digitally, and the necessity of openly providing DNA sequences and thorough gene annotations has been discussed many times, it is not uncommon for researchers to

have only a vague plasmid map and/or no reference sequence [10-13]. *De novo* sequence assembly is thus preferred to reference-based assembly as it can overcome reference bias and help identify unexpected mutations [1, 2, 11, 14-16]. Despite the importance of accurate plasmid sequence verification, it is often overlooked in favor of simpler "confirmation" methods, such as PCR amplification or restriction digests.

Compounding a reluctance to perform sequence verification is the lack of dedicated plasmid assembly tools. Amidst the many tools for genome and metagenome assembly, there are some designed to identify plasmids in the assembly of these larger datasets [17-25]. However, many of these methods can struggle to accurately reconstruct small (<25 kbp) plasmids or miss them altogether [26, 27]. To accelerate plasmid verification in high-throughput settings, assembly pipelines should aim to avoid the need for a reference sequence or manual intervention, as required by some methods [3, 28]. Here we used the two major end-to-end *de novo* assembly pipelines designed for plasmids: Epi2ME, a tool from Oxford Nanopore Technologies (ONT) for long-read assembly, and the open-source plasmid tool, PlasCAT, which can perform short-read, long-read, and hybrid assembly [2, 29-31].

Each of the available sequencing technologies have their own advantages and limitations. Sanger sequencing has long been the gold standard for sequencing, however, the Sanger method requires a reference sequence to design primers and is limited to short (~800 bp) sequences, necessitating many reactions to verify a whole plasmid [32-36]. The need to sequence unknown DNA templates led to the introduction of next-generation sequencing (NGS) technology, namely the Illumina fragmentation-based approach [5, 32-34, 37]. The resultant short sequence fragments (~250 bp) can be assembled to form the full sequence. While short-read sequencing generally provides good template coverage and high sequence accuracy, biases introduced in PCR steps result in the underrepresentation of GC-rich and repetitive regions [32, 38, 39]. Short-read sequencing can also underperform with low diversity libraries [38, 40]. Thus, third-generation sequencing methods that allow the sequencing of long-reads thousands of nucleotides long – as long as a plasmid itself – have been developed [3, 34]. Long-read sequencing can be faster and cheaper than fragmentation-based approaches, and better resolve long, complex sequences, but it is also more error prone than its predecessors and can struggle with smaller templates [1, 2, 27, 41-

43]. Recent advancements in genome assembly tools have indicated that a hybrid approach, using both short and long reads, can produce improved assemblies [23, 27, 42, 44]. Nevertheless, the ability to sequence growing libraries of DNA sequences, including genomes, with base-level precision is a continued pursuit [45].

Here, we evaluated the ability of the short- and long- read sequencing methods available from Illumina and ONT, respectively, to generate accurate *de novo* assemblies of plasmid sequences. We found that a hybrid assembly approach, using both short and long reads, produced the best assemblies. The short-read assemblies were limited by the quality and quantity of DNA used and struggled to assemble GC-rich regions, whereas the long-read assemblies had a higher incidence of insertions and deletions (collectively, indels) and mutations, as has been previously suggested [26, 42, 46]. We used Sanger sequencing to confirm several discrepancies between the assembly sequences and the plasmid reference sequences and found several cases where the assemblies consistently differed from the expected reference sequence, highlighting the importance of *de novo* assembly.

MATERIALS AND METHODS

Reagents

The Zyppy-96 Plasmid MagBead Miniprep Kit was purchased from Zymo Research (Irving, CA, USA, #D4102). The long-read library preparation kit and flow cell were purchased from Oxford Nanopore (UK, # SQK-RBK114.96) with the new chemistry flow cell (UK, # FLO-MIN114). For the short-reads, the library preparation kit (the ILMN DNA LP (M) Tagmentation 96), MiSeq cartridges, and iSeq cartridges were purchased from Illumina (San Diego, CA, USA. #20060059, #MS-103-1003, and #20031374 respectively). Sanger sequencing primers were designed and ordered from IDT (Coralville, IA). The BigDye Terminator™ V3.1 and BigDyeXTerminator™ purification kits were both purchased from Thermo Fisher (Waltham, MA. USA, #4337454 and #4376486 respectively).

Biological Resources

The plasmids used for library preparation, sequencing, and analysis were obtained from three vendors. Twelve plasmids were synthesized and sequence verified by Twist Biosciences (San Francisco, CA), and eleven were procured from Addgene (Watertown, MA, USA). One

plasmid solution was taken from a transfection kit where the mixture contains two plasmids with the same vector backbone, roughly 6000 bp, and two different inserts, around 1350 bp and 650 bp (Gibco #A14635) [47, 48]. The array of plasmids was given unique identifiers (e.g., Plasmid 1234) to anonymize the data set. This allowed for true *de novo* assembly, where the different methods could be compared on overall accuracy among the generated datasets.

Plasmid Isolation and Sequencing

Plasmid Isolation: The plasmid DNA was extracted from each of the 24 isolates on an epMotion 5075 TC liquid handler (Eppendorf, Hamburg, DE) using the Zyppy-96 Plasmid MagBead Miniprep Kit (Zymo Research, Irving CA, USA), according to manufacturer's instructions with the modification of pipet mixing during the lysis and neutralization steps and an extended elution time of 10 minutes. Samples were all quantified on a Synergy LX plate reader to determine the quality and quantity of samples after extraction via miniprep. All samples were required to have at least 35 ng/uL and an A260/280 purity reading of >1.8.

Oxford Nanopore Sequencing: Post isolation, 50 ng of each isolate was used for sequencing with the MinION Sequencer (Oxford Nanopore, UK). These sequencing libraries were prepared using the Rapid Barcoding Kit (#SQK-RBK114.96) with the Flow Cell (#FLO-MIN114) according to the manufacturer's instructions. Samples were run on the MinION with a maximum read length kept of 25 kbp. FASTQ files were generated from the super-high accuracy method of the Dorado basecaller within the MinKNOW software and were used for sequence validation, comparison, and evaluation.

Illumina Sequencing: After isolation, 200 ng of each isolate was used for sequencing on the MiSeq and iSeq (Illumina, CA). Both sequencing libraries were prepared on an epMotion 5075 TC liquid handler (Eppendorf, Hamburg, DE) using the ILMN DNA LP (M) Tagmentation 96 IPB kit protocol as described by the manufacturer. The pooled libraries were spiked with 1% v/v PhiX Control V3 (Illumina, San Diego, CA) and were diluted to a final loading concentration of 10 pM and 100 pM for the MiSeq and iSeq, respectively. The diluted libraries (600 µL and 20 µL) were loaded onto a MiSeq Reagent Nano Kit v2 (500-cycles) and iSeq 100 i1 Reagent v2 (300-cycle). FASTQ files generated were used for sequence validation, comparison, and evaluation.

Sanger Sequencing: Sequence validation was performed as needed for templates with generated sequence discrepancies. Primers were designed and ordered through IDT between 18 to 25 bps long to satisfy the following requirements: a GC content of 50% or higher, a melting temperature around 70 C, and no secondary structure (Supplementary Table S1). Fragments were prepared following the BigDye Terminator™ V3.1 kit as described, and the 10 µL reactions were diluted to 0.5X using the BigDyeXTerminator™ purification kit where described. Samples were sequenced using the LongFrag_BDX protocol. Generated results were immediately uploaded to SnapGene and compared to generated data.

De Novo Sequence Assembly: De novo sequence assembly was primarily performed using PlasCAT, an open-source plasmid assembly pipeline that was recently adapted to a web application (sequencing.genofab.com) [2, 29, 31]. In brief, the pipeline generates assemblies from short-reads, from long-reads, or from a hybrid approach (i.e. both short and long reads) via the gold-standard genome assembly tool, Unicycler [23]. The pipeline also performs some pre-processing of the data, either through Trimmomatic for short-reads [49] or Filtlong for long reads (https://github.com/rrwick/Filtlong), and subsets the long reads to a particular coverage using Rasusa [50, 51]. The short reads were filtered to a minimum length of 50 bp and a minimum quality score of 35. Filtlong was used to keep the best 80% of long reads (based on quality and length) and remove reads above the maximum read length of 20,000 bp [52]. Subsetting with rasusa was done using the default estimated size of 5,000 bp and 500X coverage, which gave better long-read assemblies than the default 1000X coverage. Long-read and hybrid assemblies are polished with racon [53]. Long-read assemblies were also generated using Oxford Nanopore's EpiPl2ME platform to serve as a method comparison (https://labs.epi2me.io/). This pipeline uses Flye for long-read assembly [54], the Medaka polisher (https://github.com/nanoporetech/medaka) [30, 55], and Trycycler to generate a consensus assembly [28]. For Epi2ME, we used the default estimated size of 7000 bp, 60X coverage, and end trimming of 150 bp.

RESULTS

Plasmid Sequencing and Assembly Pipelines

We sequenced a set of 24 plasmids and performed *de novo* assembly using short-reads from the Illumina MiSeq and long-reads from the ONT MinION. Three technical replicates were

sequenced from each plasmid, all pulling from the same initial purified plasmid solution. The open-source tool PlasCAT was used to generate assemblies in three ways: short-read only, long-read only, and a hybrid approach using short and long reads (Figure 1). This design allowed us to compare both the reproducibility of each sequencing technology across repeated library preparations and evaluate different plasmid assembly approaches.

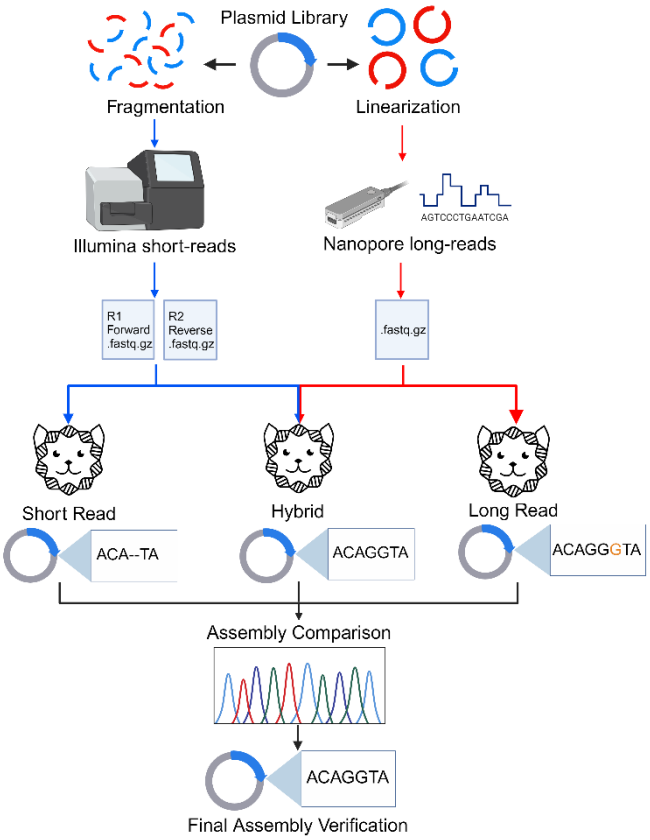


Figure 1: Overview of sequencing workflow. Multiple sequencing runs were performed on 24 plasmid samples. For short-read sequencing, plasmids were fragmented and chemically indexed before being loaded onto the Illumina MiSeq. Forward and reverse reads were generated and used with PlasCAT to generate both short-read and hybrid assemblies. For long-read sequencing, the plasmids were linearized, chemically indexed, and loaded onto the Oxford Nanopore MinION. Fastq files were generated and used for both long-read and hybrid assemblies. Sanger sequencing was used to confirm regions with discrepancies between assemblies.

Of the 72 total samples prepared, all generated data on the Nanopore sequencer, while only 69 samples generated fastq files with data on the MiSeq. Of the three failures, two were replicates of the same sample. Issues with sample dropouts prohibited us from generating assemblies for these three samples and were attributed to the library preparation procedure and not the assembly process. The short reads were trimmed to maintain a per-base quality

score of at least 35 and reads shorter than 50 bp were removed (Supplemental Figure S1). Before filtering, there were roughly 50,000 reads (combined forward and reverse) for each sample, representing at least 500 million bases per run. After filtering, only about 5,000 reads were retained per sample (Supplemental Figure S1). The long reads were filtered with Filtlong. There were initially about 50,000 reads per sample, representing over 3 billion bases per run. Filtlong retains the best 80% of the data, based on both quality and length. This is evidenced by an increase in average read length (from ~4,000 to ~6,000) and minimum read length (from 500 to ~1,000 bp) in post-filtered samples, despite the decrease in maximum read length to 20,000 bp. After filtering, there were about 25,000 reads per sample (Supplemental Figure S1).

Hybrid *De Novo* Plasmid Assembly Outperforms Short-Read or Long-Read Assembly

To quantify the robustness of each assembly method, we devised two scoring methods: an assembly score to represent the success of a particular assembly, and a sequence agreement score to assess the reproducibility of assemblies across multiple runs (Figure 2). An assembly score of 1 indicates a single contig was returned (a success), while a 0 was given if no assembly or if multiple contigs were returned (a failure). An overall assembly score was obtained by summing the scores of each of the three runs. A sequence agreement score of 1 indicates that the sequences of all successful (non-fragmented) assemblies were the same, or a 0 if not. Plasmids that only had one successful assembly were excluded from this scoring. The overall assembly and sequence agreement scores were converted into percentages based on the number of included runs and samples, respectively. We included one sample that was a mixture of two plasmids (Plasmid 3589) to see whether the plasmid assembly pipelines would generally return one contig, or if it could resolve mixtures of similar plasmids, but it was excluded from our formal data analysis.

We generated short-read assemblies for all 69 of the samples that produced sufficient data. Most of the assemblies were single contigs, however the pipeline returned 13 assemblies with multiple contigs (Figure 2A). Given that we expected all samples to result in a single contig representing the plasmid, these 13 assemblies were considered failures and given an assembly score of 0. There was only one sample that received an overall assembly score of 0, for which all three assemblies were fragmented (Plasmid 3135). Nevertheless, nearly 77% of

runs resulted in a successful assembly. Furthermore, 76% of samples had good sequence agreement scores, indicating a high level of support of those short-read assembly sequences. However, there

A

	Short-Read - PlasCAT										
	Run 1			Run 2			Run 3				
Plasmid ID	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Overall Assembly Score	Sequence Agreement Score
3125	**4119**	1	1	**4119**	1	1	**4119**	1	1	3	1
3137	**4875**	1	1	**4875**	1	1	**4875**	1	1	3	1
3139	**4908**	1	1	**4908**	1	1	4908	2	0	2	1
3113	**4908**	1	1	2547	3	0	**4908**	1	1	2	1
3144	4911	3	0	**4911**	1	1	**4911**	1	1	2	1
3148	**4947**	1	1	**4947**	1	1	4947	2	0	2	1
3122	**4966**	1	1	**4966**	1	1	**4966**	1	1	3	1
3118	5041	1	1	n/a	n/a	0	n/a	n/a	0	1	-
3117	**5050**	1	1	2547	3	0	**5050**	1	1	2	1
3115	**6815**	1	1	**6815**	1	1	**6815**	1	1	3	1
2101	**6824**	1	1	**6824**	1	1	**6824**	1	1	3	1
3123	**6832**	1	1	**6832**	1	1	**6832**	1	1	3	1
3124	**6819**	1	1	6819	3	0	**6819**	1	1	2	1
3127	6652	1	1	6586	1	1	n/a	0	0	2	0
3121	**7765**	1	1	**7765**	1	1	**7765**	1	1	3	1
3116	4454	3	0	**8756**	1	1	**8756**	1	1	2	1
3126	**4611**	1	1	**4611**	1	1	**4611**	1	1	3	1
3131	5295	1	1	5179	1	1	5170	3	0	2	0
3132	5467	1	1	5550	1	1	5383	1	1	3	0
3130	10229	1	1	10270	1	1	10329	1	1	3	0
3135	10462	2	0	10231	2	0	10462	3	0	0	0
3133	**14191**	1	1	**14191**	1	1	**14191**	1	1	3	1
3134	12843	2	0	15044	1	1	10462	2	0	1	-
3589	*5976*	*3*	-	*5976*	*3*	-	*6628*	*1*	-	-	-
	Total Score		19			18			16	53	16
	Percent		82.61%			78.26%			69.57%	76.81%	76.19%

B

	Long-Read - PlasCAT										
	Run 1			Run 2			Run 3				
Plasmid ID	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Overall Assembly Score	Sequence Agreement Score
3125	4117	1	1	4119	1	1	4116	1	1	3	0
3137	4874	1	1	4875	1	1	4875	1	1	3	0
3139	4908	1	1	4909	1	1	4908	1	1	3	0
3113	**4908**	1	1	**4908**	1	1	**4908**	1	1	3	1
3144	**4911**	1	1	**4911**	1	1	**4911**	1	1	3	1
3148	**4947**	1	1	**4947**	1	1	**4947**	1	1	3	1
3122	4966	1	1	*4966*	1	1	4965	1	1	3	0
3118	**5041**	1	1	**5041**	1	1	**5041**	1	1	3	1
3117	**5050**	1	1	**5050**	1	1	**5050**	1	1	3	1
3115	6815	1	1	*6815*	1	1	6815	1	1	3	0
2101	6824	1	1	6820	1	1	6823	1	1	3	0
3123	6832	1	1	6832	1	1	6919	1	1	3	0
3124	**6819**	1	1	**6819**	1	1	**6819**	1	1	3	1
3127	7464	1	1	7457	1	1	7463	1	1	3	0
3121	7765	1	1	20878	1	1	7765	1	1	3	0
3116	8757	1	1	8756	1	1	8756	1	1	3	0
3126	4611	1	1	4611	1	1	4614	1	1	3	0
3131	6077	1	1	6080	1	1	6077	1	1	3	0
3132	6345	1	1	18620	1	1	6346	1	1	3	0
3130	11136	1	1	11137	1	1	11136	1	1	3	0
3135	12156	2	0	13059	1	1	12067	1	1	2	0
3133	**14191**	1	1	**14191**	1	1	**14191**	1	1	3	1
3134	15041	1	1	15044	1	1	15044	1	1	3	0
3589	*10410*	*8*	-	*12047*	*4*	-	*7456*	*2*	-	-	-
	Total Score		22			23			23	68	7
	Percent		95.65%			100.00%			100.00%	98.55%	30.43%

C

	Hybrid - PlasCAT										
	Run 1			Run 2			Run 3				
Plasmid ID	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Overall Assembly Score	Sequence Agreement Score
3125	**4119**	1	1	**4119**	1	1	9492	2	0	2	1
3137	**4875**	1	1	**4875**	1	1	**4875**	1	1	3	1
3139	**4908**	1	1	**4908**	1	1	4908	3	0	2	1
3113	**4908**	1	1	**4908**	1	1	**4908**	1	1	3	1
3144	12988	9	0	**4911**	1	1	**4911**	1	1	2	1
3148	**4947**	1	1	**4947**	1	1	14996	3	0	2	1
3122	**4966**	1	1	**4966**	1	1	**4966**	1	1	3	1
3118	5041	1	1	n/a	n/a	0	n/a	n/a	0	1	-
3117	**5050**	1	1	**5050**	1	1	**5050**	1	1	3	1
3115	**6815**	1	1	**6815**	1	1	**6815**	1	1	3	1
2101	**6824**	1	1	**6824**	1	1	**6824**	1	1	3	1
3123	**6832**	1	1	**6832**	1	1	**6832**	1	1	3	1
3124	**6819**	1	1	6819	8	0	**6819**	1	1	2	1
3127	7462	1	1	7549	1	1	n/a	n/a	0	2	0
3121	**7765**	1	1	**7765**	1	1	**7765**	1	1	3	1
3116	**8756**	1	1	**8756**	1	1	**8756**	1	1	3	1
3126	**4611**	1	1	**4611**	1	1	9559	12	0	2	1
3131	6078	1	1	6075	3	0	15236	16	0	1	-
3132	6342	1	1	6350	1	1	12116	10	0	2	0
3130	11139	1	1	11139	1	1	11140	1	1	3	0
3135	**13467**	1	1	13467	2	0	**13467**	1	1	2	1
3133	**14191**	1	1	**14191**	1	1	**14191**	1	1	3	1
3134	**15044**	1	1	**15044**	1	1	**15044**	1	1	3	1
3589	*6754;1486*	*2*	-	*7462*	*1*	-	*6749*	*1*	-	-	-
	Total Score		22			19			15	56	18
	Percent		95.65%			82.61%			65.22%	81.16%	85.71%

Figure 2: Hybrid assemblies outperform short- and long-read assemblies. *De novo* assemblies were generated from short-reads (A), long-reads (B), or from both (hybrid, C). All assembly pipelines

produced some fragmented assemblies (>1 contig) which were considered failures (red). Only the length of the longest contig is reported in these cases. Some short-read sequencing preparations did not produce sufficient data for assembly (n/a, red). An assembly score of 1 indicates a successful assembly (non-fragmented), and these are summed across the three replicates to generate an overall assembly score (maximum of 3). A sequence agreement score of 1 indicates that all successful assemblies were exact matches for one another, and the corresponding assembly lengths are bolded. Samples with only one successful assembly were not given a sequence agreement score and were excluded from the percentage calculation. The sample containing a mixture of two plasmids (Plasmid 3589) was also excluded from this analysis, as it was not expected to return only one contig. The hybrid assemblies had the highest sequence agreement scores, followed by short-read and then long-read assemblies. The long-read assemblies had the highest overall assembly score but failed to produce high sequence agreement scores, indicating lower reproducibility of assembly results.

were 5 samples which received a sequence agreement score of 0 for their short-read assemblies and for which no consensus could be reached (Figure 2A).

We compared the assemblies generated from one sequencing run using the iSeq, which produces 151 bp reads, to the MiSeq, which produces 251 bp reads. Overall, the assemblies generated by the iSeq were similar to the assemblies from the MiSeq (Supplemental Table S2). Four of the iSeq assemblies contained multiple fragments, which was comparable to the failure rate of the MiSeq assemblies. Although the assemblies generated by the iSeq reads were not considerably better or worse, the shorter length of the reads may result in worse resolution of repetitive regions. Thus, we continued only with MiSeq for short-read data for further analysis.

Compared to the short-read assemblies, the long-read assemblies had a higher overall assembly score but a lower sequence agreement score (Figure 2). While nearly 99% of long-read assemblies contained a single contig, they are less consistent across runs, with only 30% of samples having a sequence agreement score of 1. Most of the samples with sequence agreement scores of 0 appeared to vary only by small (⩽5 bp) insertions and deletions (collectively, indels), but in a few cases, there appeared to be a multiplicity issue where the length of one assembly was 2-3x longer than the assemblies from other runs (Plasmids 3121 and 3132).

The hybrid assembly pipeline produced the best assemblies overall. The overall assembly score (81%) was slightly lower than the long-read assemblies, due to the three failed short-read library preparations and the presence of more fragmented assemblies. This similarity to the short-read assemblies is consistent with the hybrid approach relying on the short-reads

to establish an initial scaffold onto which the long-reads are assembled. Nevertheless, nearly 87% of samples had a good sequence agreement score, significantly higher than both short-read and long-read assemblies, highlighting the ability of the hybrid approach to reproducibly generate high quality assemblies that leverage the strengths of both technologies. There were only 3 samples that received a sequence agreement score of 0, and these had not been resolved by short-read or long-read methods either (Plasmids 3127, 3130, and 3132).

Of note, short-read assembly of the two-plasmid mixture (Plasmid 3589) resulted in either a singular plasmid of about 6600 bp or the assembly contained three contigs, seemingly representing a backbone and two inserts (Figure 2A). Several of the long-read and hybrid assemblies returned plasmids around 7400 bp and 6700 bp. Each method seemed to struggle with resolving two highly similar structures.

Given that the long-read assemblies were less robust than short-read and hybrid assemblies, we compared the results we obtained from PlasCAT to assemblies generated from Epi2ME, a long-read assembly tool recommended by Oxford Nanopore. Epi2ME failed to produce an assembly in two cases (Supplemental Table S3), while PlasCAT always returned an assembly, albeit sometimes fragmented. Aside from these runs, 97% assemblies were successful; however, it appears that Epi2ME is restricted to returning only a single contig, which could have inflated this score. With almost 55% of samples having sequence assembly scores of 1, Epi2ME performed slightly better than PlasCAT for long-read assemblies but did not perform as well as the short-read or hybrid assemblies.

Both PlasCAT and Epi2ME subset the long-read data to a particular coverage level, since studies have shown improved outcomes utilizing smaller, higher quality datasets [25, 42, 51, 56, 57]. Indeed, running the PlasCAT pipeline on the first MinION run without subsampling the data did not produce any successful assemblies; they either failed or were highly fragmented, consisting of anywhere from 3 to 80 different contigs (Supplemental Table S4). The one exception was the two-plasmid mixture (Plasmid 3589) which produced two reasonable contigs sized 7447 bp and 6532 bp. Subsetting the data also led to more practical return times, cutting down the average time per assembly from about 30 minutes to 2 minutes. While PlasCAT takes a single subset of the data to produce an assembly, Epi2ME

uses Trycycler to generate a consensus assembly from 3 assemblies generated from 3 separate subsample sets, which likely improved its sequence agreement scores.

Each Epi2ME assembly took a few minutes per sample, and they are run in succession, so each set of 24 assemblies took two to three hours to complete. On the other hand, all samples from the same PlasCAT run are executed in parallel, resulting in a fast turnaround time for processing large datasets. Long-read assembly was the fastest at only about 2 minutes per assembly, while each short-read assembly took, on average, 11 minutes to complete, ranging from 90 seconds to nearly 25 minutes. The hybrid assemblies took significantly longer: at least five minutes and up to 48 minutes.

De Novo Assembly Can Reveal Deviations from Reference Sequence

Some plasmids were easily assembled by any method and matched exactly to the reference sequence (Figure 3A). However, short-read sequencers are known to have biases associated with highly repetitive and/or GC-rich regions [32, 38], thus we expected that some short-read assemblies may not be representative of the sample. Indeed, aligning the assemblies generated from the short-reads to the reference sequence revealed significant gaps in some assembly sequences (Figure 3B). While long-read sequencers are better able to resolve repetitive and GC-rich regions, they have historically been marred by high error rates, can introduce indels, and appear to depend greatly on how the data is processed (i.e., subsetting, choice of tool) [46, 58]. Thus, hybrid assemblies are expected to resolve the gaps seen in short-read assemblies by using long-read data, while also leveraging the high accuracy of short reads to prevent indels and mismatches.

We found that all assembly methods resulted in some assemblies that had indels or mismatches compared to the reference sequence (Supplemental Table S5, Table 1). The short-read assemblies, and therefore the hybrid assemblies, were more likely to fail or be fragmented than the long-read assemblies. The short-read assemblies were also the most likely to have large indels, typically entire fragments missing (Table 1). However, long-read assemblies had more small indels (≤ 5 bp) and mismatches than the other methods. There were 10 plasmids where at least one long-read assembly had a deviation from the reference sequence, although the short-read and hybrid assemblies matched the reference perfectly. In all cases except one, these assemblies contained indels, mostly small (Supplemental Table

S5). There were also two samples that had assemblies that matched the reference length exactly but had several mismatches, which we did not encounter in the short-read or hybrid assemblies.

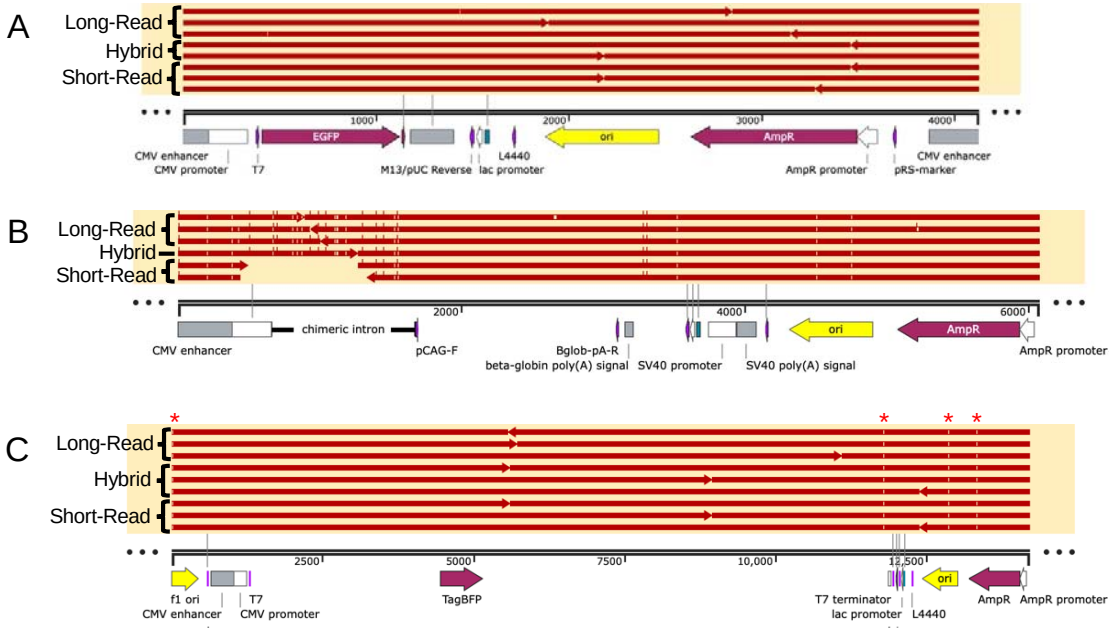


Figure 3: Alignment of assembly sequences to a reference sequence. (A) Representative example of a plasmid that is correctly assembled by any method (Plasmid 3123). **(B)** Short-read assemblies consistently miss GC-rich regions that are resolved by long-read or hybrid assemblies (Plasmid 3131). **(C)** Some deviations between assembly sequences and the reference sequence are consistent across methods, indicating true mutations (red asterisks). Only successful assemblies were aligned to the reference sequence (1-3 per method).

Table 1. Summary of errors in plasmid assemblies compared to the reference sequence.

Assembly Type	# Successful Assemblies	# Assemblies with Indels ≤ 5 bp	# Assemblies with Indels > 5 bp	# Assemblies with Mismatches
Short-Read	53	18	13	17
Long-Read	68	35	8	29
Hybrid	56	19	5	19

There were 9 plasmids that had deviations from the reference sequence in all three types of assemblies (Supplemental Table S5). Of these, 5 plasmids consistently deviated from the reference sequence. For example, the reference for Plasmid 3133 was 14194 bp, but all 9 assemblies had 6 identical mismatches and 2 small deletions resulting in a 14191 bp plasmid. In addition, 8 of the 9 assemblies for Plasmid 3121 showed a 19 bp deletion corresponding

to the T7 promoter. For all 5 samples, any successful assemblies that did not exactly match the others were long-read assemblies that supported the deviations but had additional indels.

In the 4 other cases, there were deviations from the reference sequence, but it was not clear what the true sequence was. For example, Plasmid 3131 usually returned a 6077 or 6078 bp from long-read or hybrid assemblies; some errors were consistent, resulting in a plasmid roughly 13 bp larger than the reference, but a consensus could not be reached (Supplemental Table S5). All 4 plasmids showed large (>750 bp) gaps in the short-read assemblies corresponding to the chicken β-actin (CAG) promoter [59] and adjacent chimeric intron (Figure 3B, Supplemental Table S5). This region is highly repetitive and has a GC content of 73%. This region also contained the discrepancies that could not be resolved between the long-read and hybrid assemblies.

Sanger sequencing primers were designed to target a few regions containing discrepancies (Supplemental Table S1). The Sanger reactions allowed us to confirm several deviations from the reference sequence (Figure 4). For example, it confirmed a 1 bp deletion near the chicken β-actin promoter in Plasmid 3132 (Figure 4A). It also confirmed a 2 bp insertion in a section of repeated Gs in the same promoter in Plasmid 3127, which was found in one hybrid assembly; the other assemblies had a variable number of Gs and the short-read assemblies missed the region completely (Figure 4B).

Although Sanger sequencing continues to be the gold standard for accurately resolving known short (~800 bp) sequences [60], it requires a reference sequence in order to generate primers and increasing the number of reactions to cover larger sections becomes costly and time-consuming (Supplemental Table S6). Thus, it is not a practical approach for *de novo* plasmid assembly. Although short-read assemblies outperform long-read assemblies, it is worth noting that short-read sequencing on the Illumina MiSeq is more expensive per sample than long-read sequencing on the ONT MinION (Supplemental Table S6). Long-read plasmid assembly may be best-suited for obtaining a complete structural overview of a plasmid, with the caveat that base-level alterations may be missed. On the other hand, short-read assembly may be sufficient for plasmids that do not have any highly repetitive or GC-rich regions. Given the increased robustness of hybrid assembly over either individual

method, it is clear that long-read sequencing should follow short-read sequencing to achieve the best results.

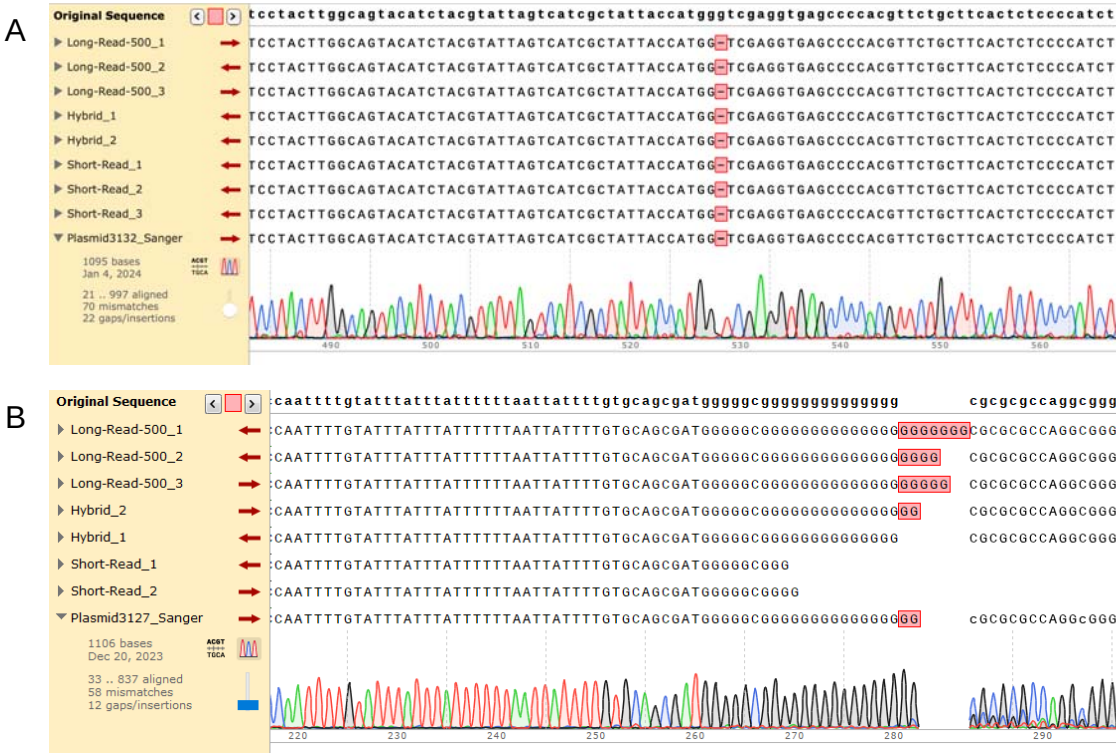


Figure 4: Sanger confirmation of discrepancies compared to reference sequence. All successful assemblies were aligned to the reference sequence. **(A)** Sanger sequencing confirmed a 1 bp deletion in the sequence of Plasmid 3132 that was found in all assemblies. **(B)** Sanger sequencing confirmed a 2 bp insertion in the sequence of Plasmid 3127, which was found in only one hybrid assembly. The 2 short-read assemblies missed this region entirely.

DISCUSSION

We evaluated several sequencing analysis pipelines for *de novo* plasmid assembly in terms of their ability to produce a successful (single contig) assembly as well as to reproducibly assemble a plasmid across three technical replicates of a 24-plasmid library. We also compared the obtained assembly sequences to a reference sequence and found several plasmids that consistently differed from the reference, regardless of sequencing or assembly method. A hybrid approach to sequencing, leveraging the high accuracy of short reads with

the ability of long reads to resolve complex regions, led to the best, most reproducible assemblies.

Long-read assemblies appeared to vary the most based on the tools used to process the data. PlasCAT, which uses miniasm for long-read assembly (through Unicycler), and Epi2ME, which uses Flye for long-read assembly, implement some of the best available assembly tools [61]. Each pipeline utilizes different tools for filtering and subsetting the data and for polishing the assemblies, but both are into easy-to-use full-service workflows suitable for high-throughput *de novo* plasmid sequence verification. Given that a hybrid approach is superior for *de novo* plasmid assembly, it is critical that assembly tools can accommodate data from multiple sources, making a vendor-independent solution like PlasCAT appealing.

Subsetting long-read sequencing data is now commonly implemented to produce better, more accurate assemblies [25, 42, 50, 51, 56, 57]. To subset the data, an estimated size for the plasmid must typically be provided to establish the number of reads needed to achieve a particular coverage level, presenting a potential barrier to *de novo* assembly. Two avenues of estimating plasmid size are to look at the most common read length in a dataset, or to generate an initial assembly using the default settings, then rerun the same data using the obtained assembly size as the new input size. However, this approach seemed to lead to worse assemblies from Epi2ME, while the default parameter seemed to work well for most samples. Similarly, of all the plasmid size options on PlasCAT, the default size of 5000 bp worked the best across all our samples. If the default size parameter for a tool results in generally good assemblies for plasmids of all sizes, this is advantageous for *de novo* assembly.

Interestingly, re-analysis of the same data with the same default parameters on Epi2ME returned different assemblies for some samples, suggesting a truly random seed used for subset generation. This proposed an interesting problem, if random subsetting can change the results, then how can the true sequence be identified? It seems that Trycycler was meant to accommodate this randomness, but it was not intended to be a fully automated platform, as it will fail to produce a consensus if any assemblies are too different from one another, leading to some instances of failure by Epi2ME [28]. The implementation of Trycycler in Epi2ME may have allowed Epi2ME to produce more similar assemblies across the three

replicates than PlasCAT, with the trade-off of potential run failure. On the other hand, PlasCAT's long-read analysis returned the same assembly each time, suggesting a systematic randomness in how data is subset, and an increased ability to reliably reproduce assemblies. There is therefore a balance that must be met between generating independent, random subsets and the ability to generate assemblies in a reproducible way.

All the bioinformatics pipelines struggled to resolve certain highly repetitive, GC rich regions to single base pair accuracy, with the short-read data performing the worst in these areas. Specifically, there were 4 plasmids containing a GC-rich region encompassing a CAG promoter and adjacent chimeric intron, which could not be resolved by any assembly method. Known PCR-induced biases in short-read sequencing can lead to an underrepresentation of GC rich regions, explaining the poor assembly of this genetic part using short-reads [33, 40]. On the other hand, the persistence of repetitive G's in the presented difficulties for the long-read sequencing reads (Figure 3). Although targeted Sanger sequencing may resolve these sequences with good confidence (Figure 4), GC-rich genetic parts such as these can have a propensity to form structural hairpins which can inhibit Sanger sequencing.

While one may wish to argue that short- or long-read sequencing on their own can provide sufficient sequence verification, calls for consistent, accurate sequence verification have been left unanswered for too long [11, 13]. If a high-level structural analysis of a genetic construct is all that is needed, long-read sequencing may be sufficient, but the base-level accuracy of the assembly may not be trustworthy in the absence of additional replicates. For a plasmid with an overall and parts-level GC-content around 50%, short-read sequencing may be sufficient, but individuals should expect that large fragments may be missing or that the assemblies may be fragmented. Even when researchers may not expect single base pair accuracy to be important, it is important to remember that single base pair changes can have unintended effects, exemplified by SNPs, various diseases, and even the sequence similarity between certain fluorescent proteins [62, 63]. Thus, robust hybrid assembly, possibly followed with Sanger sequencing for confirmation, is the ideal choice for verifying plasmid sequences. Performing several sequencing runs also allows researchers to develop confidence in a *de novo* assembly. The hybrid approach not only leverages the high accuracy of short-reads and Sanger reactions, but the ability of long-reads to resolve complex regions.

The cost associated with this choice will be the highest, but in return, reproducibility and confidence will skyrocket.

Certain plasmids, or at least some genetic parts, remain difficult to resolve by any method. Regions with GC-content outside the typical (40-60%) range as well as repetitive regions that can form hairpins present challenges for Sanger, short-read, and long-read sequencing alike [2, 60]. The prevalence of difficult-to-sequence parts pose an interesting dilemma. If some parts will be unreliable to sequence regardless of the method, especially to single base pair accuracy, how can sequence verification be enforced? These parts must be cataloged and flagged for verification with Sanger sequencing. Ideally, these parts would be replaced with others that can perform the same function but are more readily sequenced.

This work was performed with the large-scale production and verification of high-throughput plasmid libraries in mind. With increasing ease and availability of DNA synthesis and sequencing technology [64], there is increasing demand for high-throughput production methods to produce hundreds to thousands of plasmids. The major bottleneck in the production of these libraries is the sequence verification step [2, 10, 13]. Tools such as Trycycler that require manual intervention for users to reliably generate good assemblies may be useful for difficult-to-assemble plasmids but become impractical to scale up [3, 65]. Indeed, even perusing the multiple contigs in the fragmented assemblies generated here revealed contigs that could reasonably represent the plasmid of interest. Nevertheless, the pipelines described in this manuscript address a critical need for better plasmid library validation, generating reliable data faster, and they make it easier for non-technical users to carry out complex bioinformatics analyses. As sequencing technologies and bioinformatics tools will continue to improve, further work is needed to optimize analysis pipelines for *de novo* plasmid assembly. Recent work has led to the development of DNA signatures that can verify plasmid sequences using only *de novo* assembly [1, 66]. By incorporating a compressed version of a sequence into a signature and inserting this into a plasmid, plasmids can be instantly verified against the original reference sequence even with no prior knowledge of the sequence, emphasizing the need for and power of *de novo* assembly [1, 66].

DATA AVAILABILITY

All data generated for this manuscript are publicly available in a repository and can be accessed at https://figshare.com/s/cb61b237859049e68e52.

Supplementary Data are available at NAR online.

AUTHOR CONTRIBUTIONS

Sarah Hernandez: Conceptualization, Methodology, Formal Analysis, Visualization, Writing—original draft & editing. Casey-Tyler Berezin: Conceptualization, Formal analysis, Software, Writing—review & editing. Katie Miller: Formal analysis, Writing—review. Samuel Peccoud: Software. Jean Peccoud: Conceptualization, Methodology, Writing—review & editing.

FUNDING

This work was supported by the National Institutes of Health [R01GM147816 and R21AI168482], the National Science Foundation [2123367], and the Suzanne and Walter Scott Foundation. Funding for open access charge: National Institutes of Health.

ACKNOWLEDGEMENTS

Graphical abstract and Figure 1 were created with BioRender.com.

CONFLICT OF INTEREST

J.P., K.M., and S.P. have financial interests in GenoFAB, Inc., a company which may benefit or be perceived as benefiting from this publication.

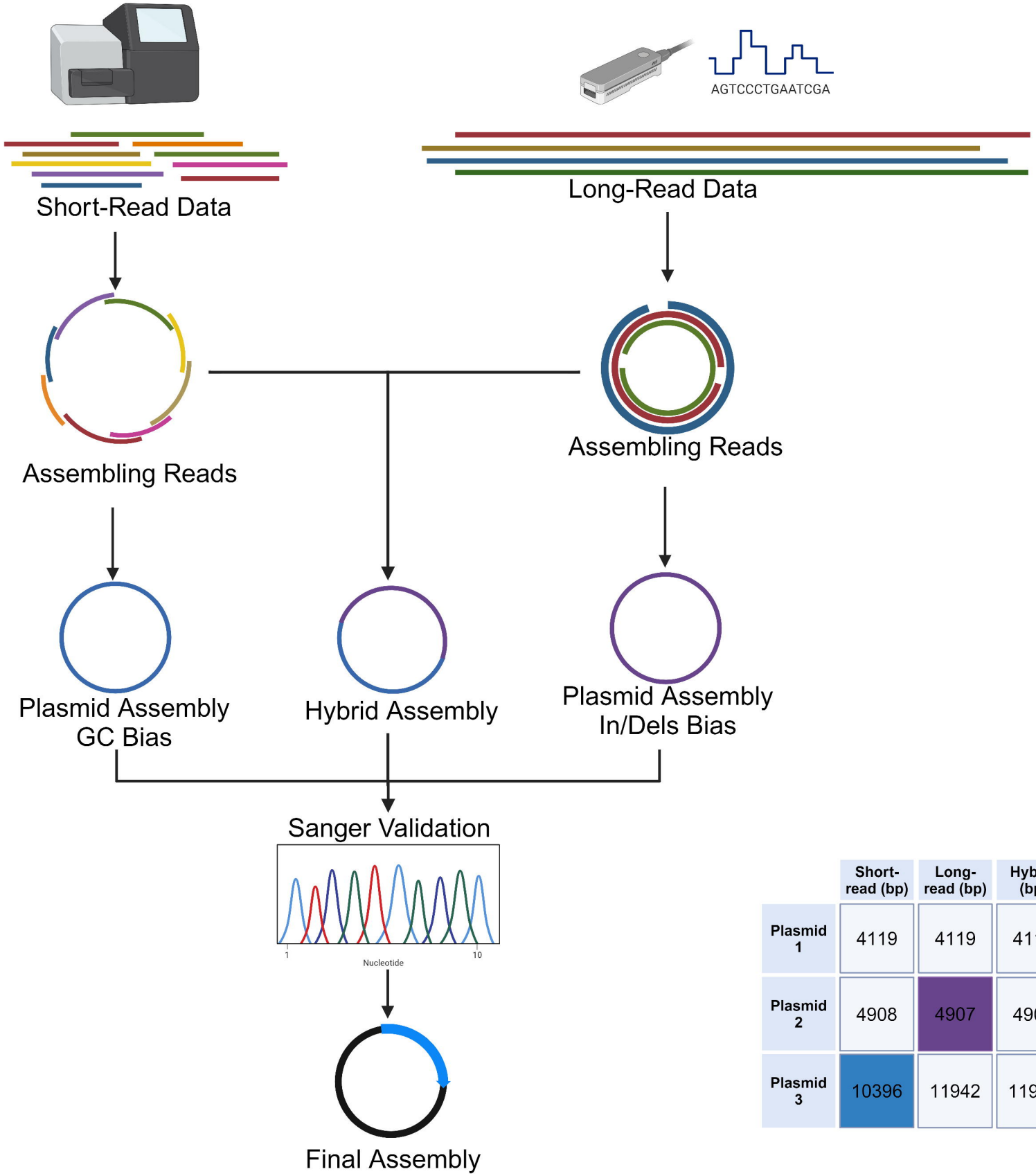
REFERENCES

1. Gallegos, J.E., et al., *Securing the Exchange of Synthetic Genetic Constructs Using Digital Signatures.* ACS Synth. Biol., 2020. **9**(10): p. 2656-2664.
2. Gallegos, J.E., et al., *Rapid, robust plasmid verification by de novo assembly of short sequencing reads.* Nucleic Acids Research, 2020. **48**(18).
3. Brown, S.D., et al., *Complete sequence verifcation of plasmid DNA using the Oxford Nanopore Technologies' MinION device.* BMC Bioinformatics, 2023. **24**(116).

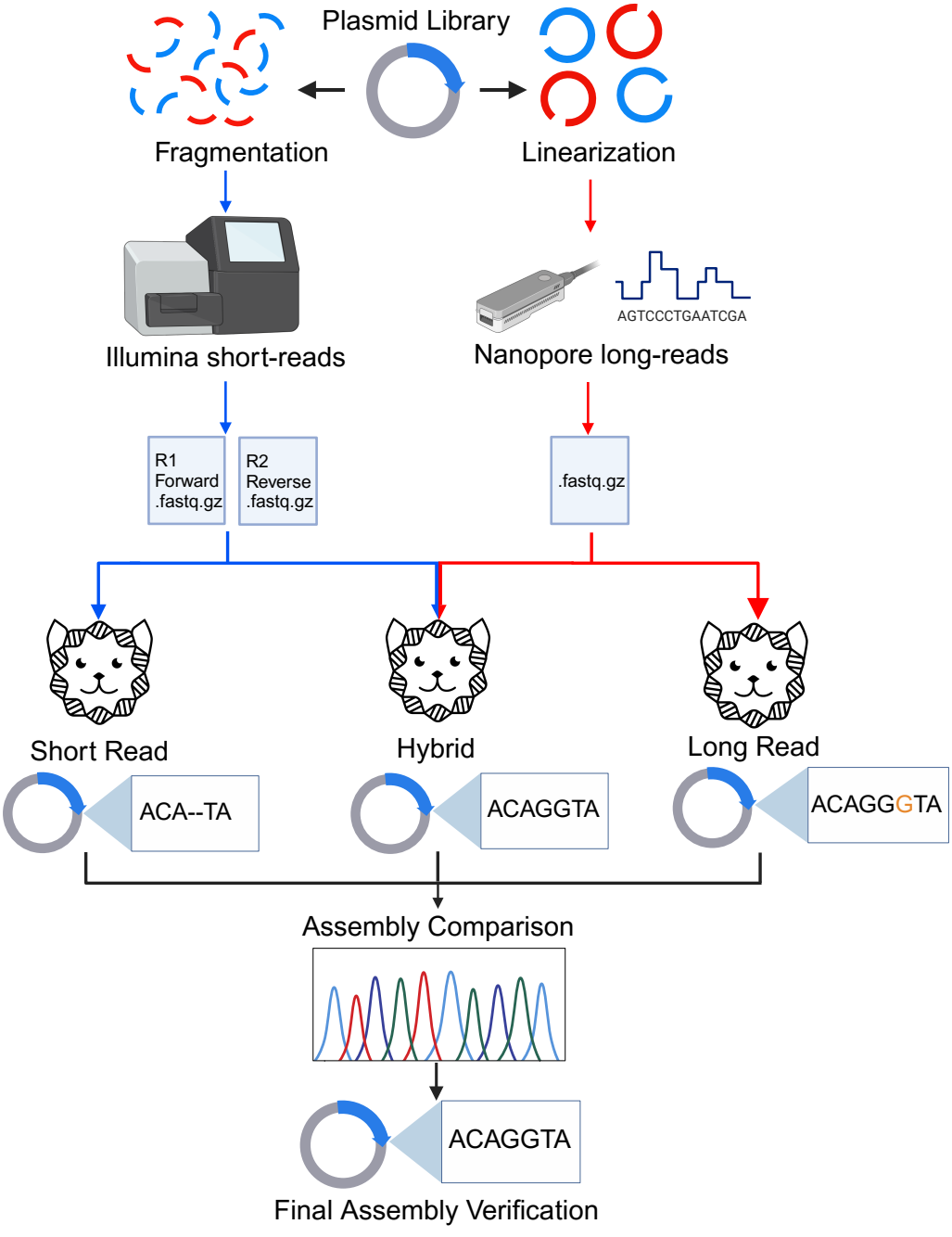
4. Rozwandowicz, M., et al., *Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae.* Journal of Antimicrobial Chemotherapy, 2018. **73**: p. 1121-1137.
5. Cameron, D.E., C.J. Bashor, and J.J. Collins, *A brief history of synthetic biology.* Nature Reviews Microbiology, 2014. **12**(5): p. 381-390.
6. Munnelly, K., *Engineering for the 21st Century: Synthetic Biology.* ACS Synth. Biol., 2013(2): p. 213-215.
7. Peccoud, J., *Synthetic Biology: fostering the cyber-biological revolution.* Synthetic Biology, 2016. **1**(1).
8. Ghaffarifar, F., *Plasmid DNA vaccines: where are we now.* Drugs Today, 2018. **54**(5): p. 315-33.
9. Shapland, E.B., V. Holmes, and C.D. Reeves, *Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process.* ACS Synthetic Biology, 2015. **4**(7): p. 860-866.
10. Peccoud, J., *Data sharing policies: share well and you shall be rewarded.* Synthetic Biology, 2021. **6**(1): p. ysab028.
11. Peccoud, J., et al., *Essential information for synthetic DNA sequences.* Nature Biotechnology, 2011. **29**(1): p. 22-22.
12. Peccoud, J., et al., *Cyberbiosecurity: from naive trust to risk awareness.* Trends in biotechnology, 2018. **36**(1): p. 4-7.
13. Thuronyi, B.W., E.A. DeBenedictis, and J.E. Barrick, *No assembly required: Time for stronger, simpler publishing standards for DNA sequences.* Plos Biology, 2023. **21**(11): p. e3002376.
14. Chen, N.-C., et al., *Reference flow: reducing reference bias using multiple population genomes.* Genome Biology, 2021. **22**(1).
15. Valiente-Mullor, C., et al., *One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads.* PLOS Computational Biology, 2021. **17**(1): p. e1008678.
16. Lau, J. *Reference bias: Challenges and solutions*. SevenBridges Blog, 2017.
17. Antipov, D., et al., *plasmidSPAdes: assembling plasmids from whole genome sequencing data.* Bioinformatics, 2016. **32**(22): p. 3380-3387.
18. Rozov, R., et al., *Recycler: an algorithm for detecting plasmids from de novo assembly graphs.* Bioinformatics, 2017. **33**(4): p. 475-482.
19. Antipov, D., et al., *Plasmid detection and assembly in genomic and metagenomic data sets.* Genome Res, 2019. **29**(6): p. 961-968.
20. Gomi, R., K.L. Wyres, and K.E. Holt, *Detection of plasmid contigs in draft genome assemblies using customized Kraken databases.* Microb Genom, 2021. **7**(4).
21. Pellow, D., et al., *SCAPP: an algorithm for improved plasmid assembly in metagenomes.* Microbiome, 2021. **9**(1): p. 144.
22. Gupta, S.K., S. Raza, and T. Unno, *Comparison of de-novo assembly tools for plasmid metagenome analysis.* Genes & Genomics, 2019. **41**(9): p. 1077-1083.
23. Wick, R.R., et al., *Unicycler: resolving bacterial genome assemblies from short and long sequencing reads.* PLoS computational biology, 2017. **13**(6): p. e1005595.
24. Tang, X., et al., *PLASMe: a tool to identify PLASMid contigs from short-read assemblies using transformer.* Nucleic Acids Research, 2023. **51**(15): p. e83-e83.
25. Bouras, G., et al., *Plassembler: an automated bacterial plasmid assembly tool.* Bioinformatics, 2023. **39**(7).
26. Berbers, B., et al., *Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modifed Bacillus.* Nautre Research, 2020. **10**.
27. Johnson, J., M. Soehnlen, and H.M. Blankenship, *Long read genome assemblers struggle with small plasmids.* Microbial Genomics, 2023. **9**(5).

28. Wick, R.R., et al., *Trycycler: consensus long-read assemblies for bacterial genomes.* Genome Biology, 2021. **22**.
29. Berezin, C.-T., et al. *PlasCAT: Plasmid Cloud Assembly Tool*. 2023.
30. Epi2ME Labs. *Epi2ME Labs Blog*. 2023 12/06/2023.
31. Peccoud, S., et al., *PlasCAT: Plasmid Cloud Assembly Tool.* Bioinformatics, 2024.
32. Mardis, E.R., *Next-Generation Sequencing Platforms.* Annu. Rev. Anal. Chem., 2013. **6**: p. 287-303.
33. Hu, T., et al., *Next-generation sequencing technologies: An overview.* Human Immunology, 2021. **82**(11): p. 801-811.
34. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.
35. Peccoud, J., et al., *Targeted development of registries of biological parts.* Plos one, 2008. **3**(7): p. e2671.
36. Wilson, M.L., et al., *Sequence verification of synthetic DNA by assembly of sequencing reads.* Nucleic Acids Research, 2012. **41**(1): p. e25-e25.
37. Buermans, H.P.J. and J.T.d. Dunnen, *Next generation sequencing technology: Advances and applications.* Biochimica et Biophysica Acta, 2014. **1842**(10): p. 1932-1941.
38. Tilak, M.-K., et al., *Illumina Library Preparation for Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA.* Genome Biology and Evolution, 2018. **10**(2): p. 616-622.
39. Liao, X., et al., *Current challenges and solutions of de novo assembly.* Quantitative Biology, 2019. **7**: p. 90-109.
40. Aird, D., et al., *Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.* Genome Biology, 2011. **12**.
41. Zhao, W., et al., *Oxford nanopore long-read sequencing enables the generation of complete bacterial and plasmid genomes without short-read sequencing.* Front. Microbiol. Sec. Evolutionary and Genomic Microbiology, 2023. **14**.
42. De Maio, N., et al., *Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes.* Microbial genomics, 2019. **5**(9): p. e000294.
43. Xia, Y., et al., *Strategies and tools in illumina and nanopore‐integrated metagenomic analysis of microbiome data.* iMeta, 2023. **2**(1).
44. Khrenova, M.G., et al., *Nanopore sequencing for de novo bacterial genome assembly and search for single-nucleotide polymorphism.* International Journal of Molecular Sciences, 2022. **23**(15): p. 8569.
45. Gallegos, J.E., et al., *Challenges and opportunities for strain verification by whole-genome sequencing.* Scientific Reports, 2020. **10**(1): p. 5873.
46. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis.* Genome Biology, 2020. **21**(30).
47. ThermoFisher Scientific, *Expi293™ Expression System USER GUIDE*. 2020, ThermoFisher Scientific. p. 1-32.
48. Janeway CA Jr, T.P., Walport M, et al., *The structure of a typical antibody molecule*, in *Immunobiology: The Immune System in Health and Disease.* 2001, New York: Garland Science.
49. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.
50. Hall, M.B., *Rasusa: Randomly subsample sequencing reads to a specified coverage.* Journal of Open Source Software, 2022. **7**(69): p. 3941.
51. Lonardi, S., et al., *When less is more: 'slicing' sequencing data improves read decoding accuracy and de novo assembly quality.* Bioinformatics, 2015. **31**(18): p. 2972-2980.
52. Wick, R.R. and P. Menzel. *Filtlong*. Filtlong 2021 [cited 2024.
53. Vaser, R., et al., *Fast and accurate de novo genome assembly from long uncorrected reads.* Genome Res., 2017. **27**(5): p. 737-746.

54. Kolmogorov, M., et al., *Assembly of long, error-prone reads using repeat graphs.* Nature Biotechnology, 2019. **37**: p. 540-546.
55. Wright, C., et al. *epi2me-labs/wf-denovo-assembly*. epi2me-labs 2022 11/16/2023
56. Murigneux, V., et al., *Comparison of long-read methods for sequencing and assembly of a plant genome.* GigaScience, 2020. **9**(12).
57. Wick, R.R. and K.E. Holt, *Benchmarking of long-read assemblers for prokaryote whole genome sequencing.* F1000Research, 2021. **8**: p. 2138.
58. Boostrom, I., et al., *Comparing long-read assemblers to explore the potential of a sustainable low-cost, low-infrastructure approach to sequence antimicrobial resistant bacteria with oxford nanopore sequencing.* Frontiers in Microbiology, 2022. **13**: p. 796465.
59. Alexopoulou, A.N., J.R. Couchman, and J.R. Whiteford, *The CMV early enhancer/chicken β actin (CAG) promoter can be used to drive transgene expression during the differentiation of murine embryonic stem cells into vascular progenitors.* BMC Cell Biology, 2008. **9**(1): p. 2.
60. Crossley, B.M., et al., *Guidelines for Sanger sequencing and molecular assay monitoring.* 2020. **32**(6): p. 767-775.
61. Wick, R. and K. Holt, *Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 4; peer review: 4 approved].* F1000Research, 2021. **8**(2138).
62. Morise, H., et al., *Intermolecular Energy Transfer in the Bioluminescent System of Aequorea.* Biochemistry, 1974. **13**(12): p. 2656-2662.
63. Weiner, M.P. and T.J. Hudson, *Introduction to SNPs: discovery of markers for disease.* Biotechniques, 2002. **32**(Sup): p. S4-S13.
64. Hughes, R.A. and A.D. Ellington, *Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology.* Cold Spring Harbor Perspectives in Biology, 2017. **9**(1).
65. Wick, R.R., L.M. Judd, and K.E. Holt, *Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing.* PLOS Computational Biology, 2023. **19**(3): p. e1010905.
66. Berezin, C.-T., et al., *Cryptographic approaches to authenticating synthetic DNA sequences.* Trends in Biotechnology, 2024.



	Short-read (bp)	Long-read (bp)	Hybrid (bp)
Plasmid 1	4119	4119	4119
Plasmid 2	4908	4907	4908
Plasmid 3	10396	11942	11940


Plasmid Library
Fragmentation
Linearization
Illumina short-reads
AGTCCCTGAATCGA
Nanopore long-reads
R1
Forward
.fastq.gz
R2
Reverse
.fastq.gz
.fastq.gz
Short Read
Hybrid
Long Read
ACA--TA
ACAGGTA
ACAGGGTA
Assembly Comparison
ACAGGTA
Final Assembly Verification

A

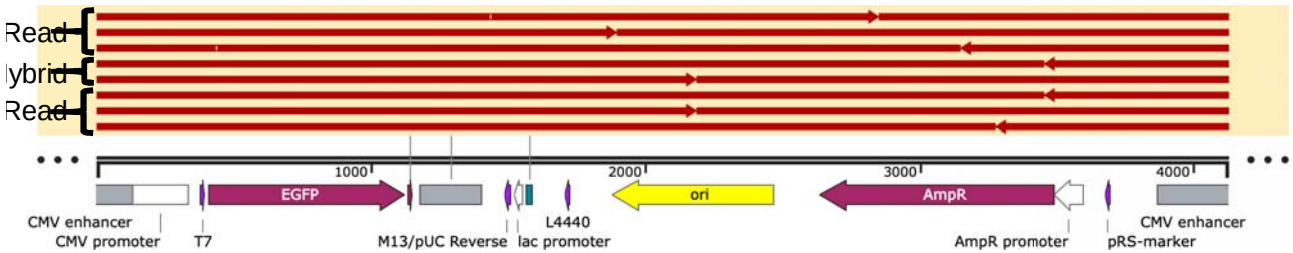
	Short-Read - PlasCAT										
	Run 1			Run 2			Run 3				
Plasmid ID	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Overall Assembly Score	Sequence Agreement Score
3125	**4119**	1	1	**4119**	1	1	**4119**	1	1	3	1
3137	**4875**	1	1	**4875**	1	1	**4875**	1	1	3	1
3139	**4908**	1	1	**4908**	1	1	4908	2	0	2	1
3113	**4908**	1	1	2547	3	0	**4908**	1	1	2	1
3144	4911	3	0	**4911**	1	1	**4911**	1	1	2	1
3148	**4947**	1	1	**4947**	1	1	4947	2	0	2	1
3122	**4966**	1	1	**4966**	1	1	**4966**	1	1	3	1
3118	5041	1	1	n/a	n/a	0	n/a	n/a	0	1	-
3117	**5050**	1	1	2547	3	0	**5050**	1	1	2	1
3115	**6815**	1	1	**6815**	1	1	**6815**	1	1	3	1
2101	**6824**	1	1	**6824**	1	1	**6824**	1	1	3	1
3123	**6832**	1	1	**6832**	1	1	**6832**	1	1	3	1
3124	**6819**	1	1	6819	3	0	**6819**	1	1	2	1
3127	6652	1	1	6586	1	1	n/a	0	0	2	0
3121	**7765**	1	1	**7765**	1	1	**7765**	1	1	3	1
3116	4454	3	0	**8756**	1	1	**8756**	1	1	2	1
3126	**4611**	1	1	**4611**	1	1	**4611**	1	1	3	1
3131	5295	1	1	5179	1	1	5170	3	0	2	0
3132	5467	1	1	5550	1	1	5383	1	1	3	0
3130	10229	1	1	10270	1	1	10329	1	1	3	0
3135	10462	2	0	10231	2	0	10462	3	0	0	0
3133	**14191**	1	1	**14191**	1	1	**14191**	1	1	3	1
3134	12843	2	0	15044	1	1	10462	2	0	1	-
3589	*5976*	*3*	-	*5976*	*3*	-	*6628*	*1*	-	-	-
	Total Score		19			18			16	53	16
	Percent		82.61%			78.26%			69.57%	76.81%	76.19%

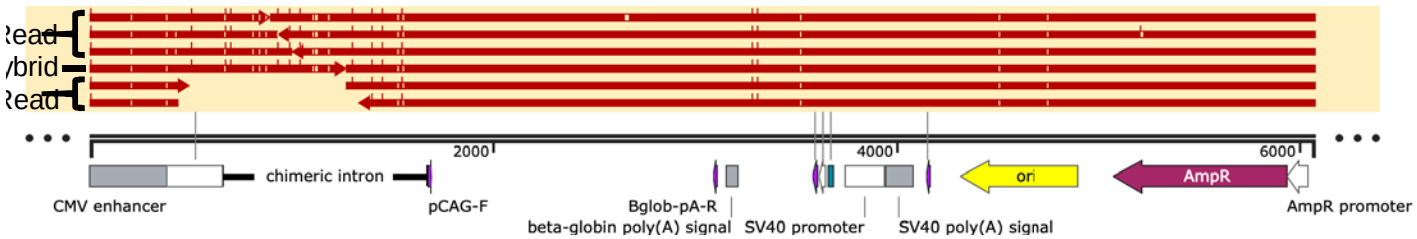
B

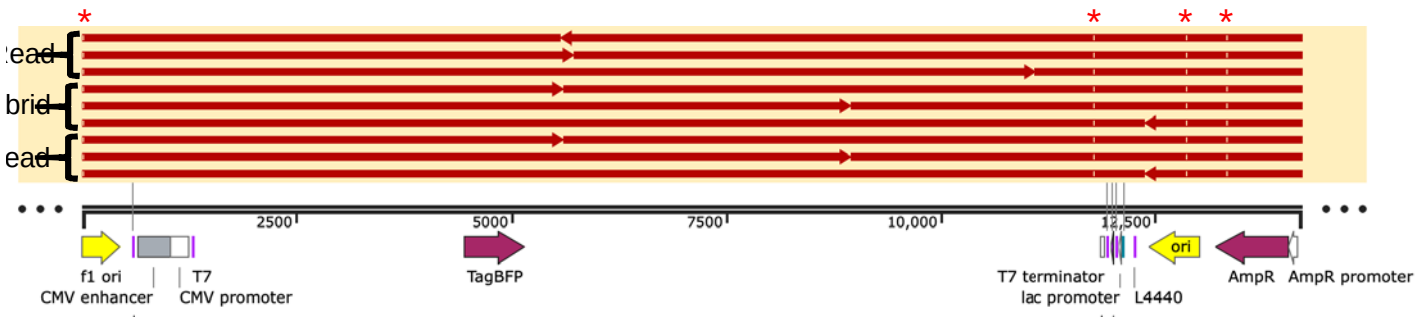
	Long-Read - PlasCAT										
	Run 1			Run 2			Run 3				
Plasmid ID	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Overall Assembly Score	Sequence Agreement Score
3125	4117	1	1	4119	1	1	4116	1	1	3	0
3137	4874	1	1	4875	1	1	4875	1	1	3	0
3139	4908	1	1	4909	1	1	4908	1	1	3	0
3113	**4908**	1	1	**4908**	1	1	**4908**	1	1	3	1
3144	**4911**	1	1	**4911**	1	1	**4911**	1	1	3	1
3148	**4947**	1	1	**4947**	1	1	**4947**	1	1	3	1
3122	4966	1	1	*4966*	1	1	4965	1	1	3	0
3118	**5041**	1	1	**5041**	1	1	**5041**	1	1	3	1
3117	**5050**	1	1	**5050**	1	1	**5050**	1	1	3	1
3115	6815	1	1	*6815*	1	1	6815	1	1	3	0
2101	6824	1	1	6820	1	1	6823	1	1	3	0
3123	6832	1	1	6832	1	1	6919	1	1	3	0
3124	**6819**	1	1	**6819**	1	1	**6819**	1	1	3	1
3127	7464	1	1	7457	1	1	7463	1	1	3	0
3121	7765	1	1	20878	1	1	7765	1	1	3	0
3116	8757	1	1	8756	1	1	8756	1	1	3	0
3126	4611	1	1	4611	1	1	4614	1	1	3	0
3131	6077	1	1	6080	1	1	6077	1	1	3	0
3132	6345	1	1	18620	1	1	6346	1	1	3	0
3130	11136	1	1	11137	1	1	11136	1	1	3	0
3135	12156	2	0	13059	1	1	12067	1	1	2	0
3133	**14191**	1	1	**14191**	1	1	**14191**	1	1	3	1
3134	15041	1	1	15044	1	1	15044	1	1	3	0
3589	*10410*	*8*	-	*12047*	*4*	-	*7456*	*2*	-	-	-
	Total Score		22			23			23	68	7
	Percent		95.65%			100.00%			100.00%	98.55%	30.43%

C

	Hybrid - PlasCAT										
	Run 1			Run 2			Run 3				
Plasmid ID	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Longest Contig Length (bp)	Number of Contigs	Assembly Score	Overall Assembly Score	Sequence Agreement Score
3125	**4119**	1	1	**4119**	1	1	9492	2	0	2	1
3137	**4875**	1	1	**4875**	1	1	**4875**	1	1	3	1
3139	**4908**	1	1	**4908**	1	1	4908	3	0	2	1
3113	**4908**	1	1	**4908**	1	1	**4908**	1	1	3	1
3144	12988	9	0	**4911**	1	1	**4911**	1	1	2	1
3148	**4947**	1	1	**4947**	1	1	14996	3	0	2	1
3122	**4966**	1	1	**4966**	1	1	**4966**	1	1	3	1
3118	5041	1	1	n/a	n/a	0	n/a	n/a	0	1	-
3117	**5050**	1	1	**5050**	1	1	**5050**	1	1	3	1
3115	**6815**	1	1	**6815**	1	1	**6815**	1	1	3	1
2101	**6824**	1	1	**6824**	1	1	**6824**	1	1	3	1
3123	**6832**	1	1	**6832**	1	1	**6832**	1	1	3	1
3124	**6819**	1	1	6819	8	0	**6819**	1	1	2	1
3127	7462	1	1	7549	1	1	n/a	n/a	0	2	0
3121	**7765**	1	1	**7765**	1	1	**7765**	1	1	3	1
3116	**8756**	1	1	**8756**	1	1	**8756**	1	1	3	1
3126	**4611**	1	1	**4611**	1	1	9559	12	0	2	1
3131	6078	1	1	6075	3	0	15236	16	0	1	-
3132	6342	1	1	6350	1	1	12116	10	0	2	0
3130	11139	1	1	11139	1	1	11140	1	1	3	0
3135	**13467**	1	1	13467	2	0	**13467**	1	1	2	1
3133	**14191**	1	1	**14191**	1	1	**14191**	1	1	3	1
3134	**15044**	1	1	**15044**	1	1	**15044**	1	1	3	1
3589	*6754;1486*	*2*	-	*7462*	*1*	-	*6749*	*1*	-	-	-
	Total Score		22			19			15	56	18
	Percent		95.65%			82.61%			65.22%	81.16%	85.71%


Read
ybrid
Read
1000
2000
3000
4000
EGFP
ori
AmpR
CMV enhancer
CMV promoter
T7
M13/pUC Reverse
L4440
lac promoter
AmpR promoter
CMV enhancer
pRS-marker

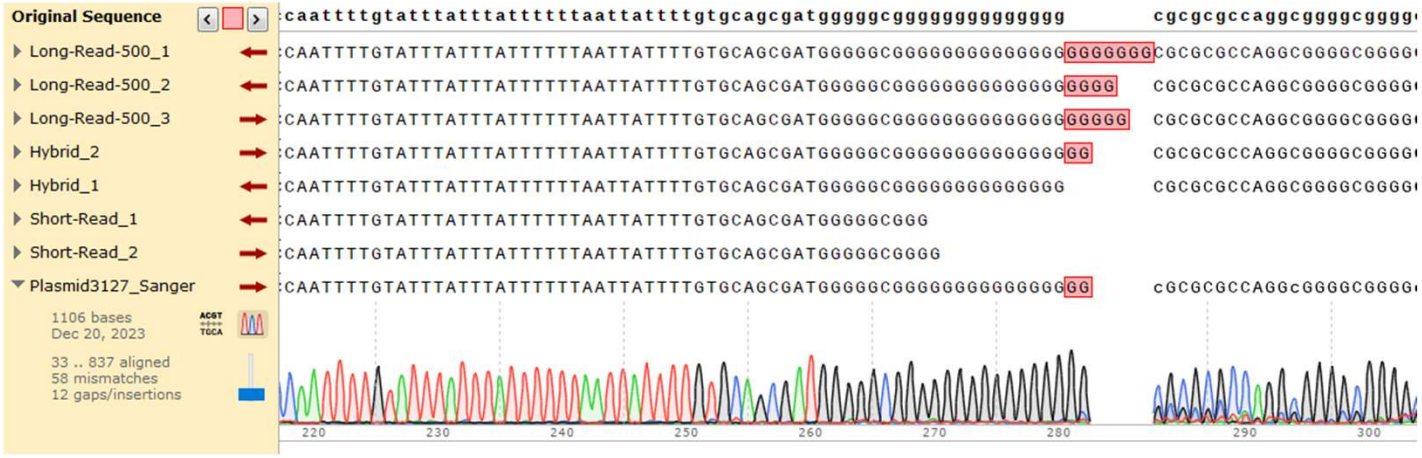

Read
ybrid
Read
2000
4000
6000
chimeric intron
ori
AmpR
CMV enhancer
pCAG-F
Bglob-pA-R
beta-globin poly(A) signal
SV40 promoter
SV40 poly(A) signal
AmpR promoter


*
*
*
*
ead
bri
ead
2500
5000
7500
10,000
12,500
ori
f1 ori
CMV enhancer
T7
CMV promoter
TagBFP
T7 terminator
lac promoter
L4440
AmpR
AmpR promoter

A

Original Sequence
gttccgcgcacatttccccgaaaagtgccacctgg tcgacattgattattgactagttattaatagtaatcaattacggggtcatta
Long-Read-500_1
Long-Read-500_2
Long-Read-500_3
Long-Read-1000_1
Long-Read-1000_2
Long-Read-1000_3
Hybrid_1
Hybrid_2
Short-Read_1
Short-Read_2
Short-Read_3
Plasmid3132_Sanger
GTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGGGTCGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTA

B


Original Sequence
caattttgtatttatttatttttaattattttgtgcagcgatgggggcgggggggggggggg cgcgcgccaggcggggcgggg
Long-Read-500_1
Long-Read-500_2
Long-Read-500_3
Hybrid_2
Hybrid_1
Short-Read_1
Short-Read_2
Plasmid3127_Sanger
CAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGGGGGGGGGCGCGCGCCAGGCGGGGCGGGG
CAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGGGGGCGCGCGCCAGGCGGGGCGGGG
CAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGGGCGCGCGCCAGGCGGGGCGGGG
CAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGGCGCGCGCCAGGCGGGGCGGGG
CAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGCGCGCGCCAGGCGGGGCGGGG
CAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGG
CAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGG
CAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGGCGCGCGCCAGGcGGGGCGGGG
1106 bases
Dec 20, 2023
33 .. 837 aligned
58 mismatches
12 gaps/insertions
220
230
240
250
260
270
280
290
300