ACS Synthetic Biology

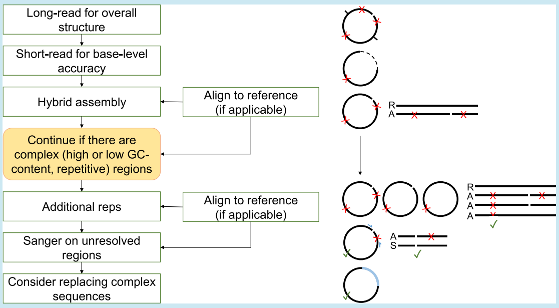
# Sequencing Strategy to Ensure Accurate Plasmid Assembly

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

Cite This: https://doi.org/10.1021/acssynbio.4c00539 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Despite the wide use of plasmids in research and clinical production, the need to verify plasmid sequences is a bottleneck that is too often underestimated 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 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 accurate, consistent assembly generation and identified mutations relative to the reference sequence. Although Sanger sequencing can be used to verify specific regions, some GC-rich and repetitive regions were difficult to resolve using any method, suggesting that easily sequenced genetic parts should be prioritized in the design of new genetic constructs.


Long-read for overall structure
Short-read for base-level accuracy
Hybrid assembly
Align to reference (if applicable)
Continue if there are complex (high or low GC-content, repetitive) regions
Additional reps
Align to reference (if applicable)
Sanger on unresolved regions
Consider replacing complex sequences
R
A
R
A
A
A
A
S

**KEYWORDS:** *whole-plasmid sequencing, NGS, nanopore, assembly, workflows, reproducibility*

## 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.[1−6] 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).[1−3,7,8] 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.[9−12] Many plasmids are generated by inserting a gene of interest into a plasmid backbone, and sequence verification is often overlooked in favor of simpler "confirmation" methods, such as PCR amplification or restriction digests. Yet, without confirming a plasmid's sequence, unrecognized deviations from an expected sequence threaten the accuracy of biological insights gained using such a plasmid.

We have proposed cryptographic algorithms to secure the exchange of plasmids in the life science community.[13−15] This cyberbiosecurity solution[16,17] makes it possible to retrieve the plasmid's and its developer's identities, retrieve the plasmid documentation, and detect the possible presence of mutations in the plasmid sequence from sequencing data without prior knowledge of the plasmid sequenced. The value of this technology depends on the availability of a streamlined and robust sequencing workflow. First, we developed a *de novo* assembly pipeline to produce the plasmid's physical sequence from short sequencing reads.[18] To streamline the execution of this bioinformatics pipeline, we deployed it on Amazon Web Services. We also developed a user interface allowing laboratory personnel to analyze plasmid sequencing data without installing software or requiring a bioinformatician's assistance. This tool, called PlasCAT,[19] is available at sequencing.genofab.com. Its modular architecture makes it possible to improve the underlying bioinformatics pipeline without significant modifications to the user interface. Here, we present an improved *de novo* assembly pipeline, analyze its reproducibility, and compare its performance to Epi2ME, a tool from Oxford Nanopore Technologies (ONT) for long-read assembly.[20]

The lack of dedicated plasmid assembly tools can explain people's reluctance to perform sequence verification of their plasmids. Amidst the many tools for genome and metagenome assembly, some are designed to identify plasmids in the assembly of these larger data sets.[21−29] However, many of these methods can struggle to accurately reconstruct small (<25 kbp) plasmids or miss them altogether.[30,31] To accelerate

**Received:** August 8, 2024
**Revised:** September 26, 2024
**Accepted:** October 24, 2024

plasmid verification in high-throughput settings, assembly pipelines should avoid the need for a reference sequence or manual intervention, as required by some methods.[1,32] *De novo* sequence assembly is preferred to reference-based assembly as it can also help overcome reference bias and identify unexpected mutations.[7,10,33−36]

Although Sanger sequencing has long been the gold standard for sequencing, it requires a reference sequence to design primers and is limited to short (∼800 bp) sequences, necessitating many reactions to verify a whole plasmid.[38−42] The need to sequence unknown DNA templates led to the introduction of next-generation sequencing (NGS) technology, namely the Illumina fragmentation-based approach.[3,38−40,43] While the short-read sequencing fragments (∼250 bp) generally provide good template coverage and high sequence accuracy, biases introduced in PCR steps result in the underrepresentation of GC-rich, GC-poor, and repetitive regions.[38,39,44−47] Short-read sequencing can also underperform with low diversity libraries.[44,46] Thus, third-generation sequencing methods that allow the sequencing of reads that are thousands of nucleotides long—as long as a plasmid itself—have been developed.[1,40] These long-read sequencing methods can be faster and cheaper than fragmentation-based approaches, and better resolve long, complex sequences, but have historically had lower accuracy than their predecessors and can struggle with smaller templates.[7,31,36,48,49] Recent advancements in genome assembly tools have indicated that a hybrid approach, using both short and long reads, can produce improved assemblies.[27,31,49−51] Nevertheless, the ability to sequence growing libraries of DNA sequences, including genomes, with base-level precision is a continued pursuit.[52] To our knowledge, a hybrid approach to *de novo* plasmid assembly using PlasCAT or other tools has not yet been systematically interrogated.

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.[30,49,53] 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. Importantly, *de novo* assembly outperformed reference-based assembly, which frequently showed reference bias and often did not match Sanger data. Thus, *de novo* hybrid assembly is the preferred method for high-throughput plasmid sequencing.

## ■ 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 R10.4 flow cell were purchased from Oxford Nanopore (UK, #SQK-RBK114.96 and #FLO-MIN114). For the short-reads, the ILMN DNA LP (M) Tagmentation 96 library preparation kit, 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 ThermoFisher (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 11 were procured from Addgene (Watertown, MA, USA). One plasmid solution was taken from a transfection kit where the mixture contained two plasmids with the same vector backbone, roughly 6000 bp, and two different inserts, around 1350 bp and 650 bp (Gibco #A14635).[54,55] 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 data sets.

**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 min. 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 superhigh 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 and 20 $\mu$L) 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 (Table S1). Fragments were prepared following the BigDye Terminator V3.1 kit as described, and the 10 $\mu$L reactions were diluted to 0.5× 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).[37] Short-read assembly through this pipeline has been previously validated,[7] while the experimental data used in developing the long-read and hybrid pipelines has not been published. As such, this work represents the first comparative assessment of the accuracy of each pipeline implemented in PlasCAT. 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.[27] The pipeline also performs some preprocessing of the data, either through Trimmomatic for short reads[56] or Filtlong for long reads (https://github.com/rrwick/Filtlong), and subsets the long reads to a particular coverage using Rasusa.[57,58] 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.[59] Subsetting with Rasusa was done using the default estimated size of 5,000 bp and 500× coverage, which gave better long-read assemblies than the default 1000× coverage. Long-read and hybrid assemblies are polished Racon.[60] Long-read assemblies were also generated using Oxford Nanopore's EpiPI2ME platform to serve as a method comparison (https://labs.epi2me.io/). This pipeline uses Flye for long-read assembly,[61] the Medaka polisher (https://github.com/nanoporetech/medaka),[20,62] and Trycycler to generate a consensus assembly.[32] For Epi2ME, we used the default estimated size of 7000 bp, 60× 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.

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 (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 (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 ∼4000 to ∼6000) and minimum read length (from 500 to ∼1000 bp) in postfiltered samples, despite the decrease in maximum read length to 20,000 bp. After filtering, there were about 25,000 reads per sample (Figure S1).

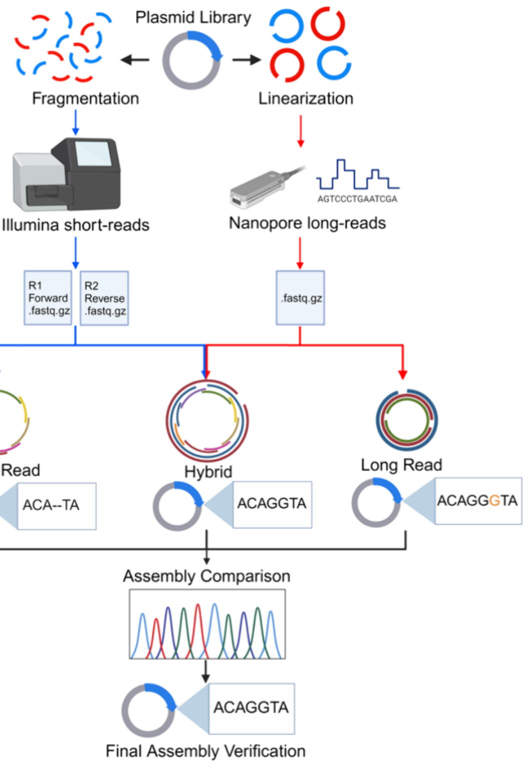
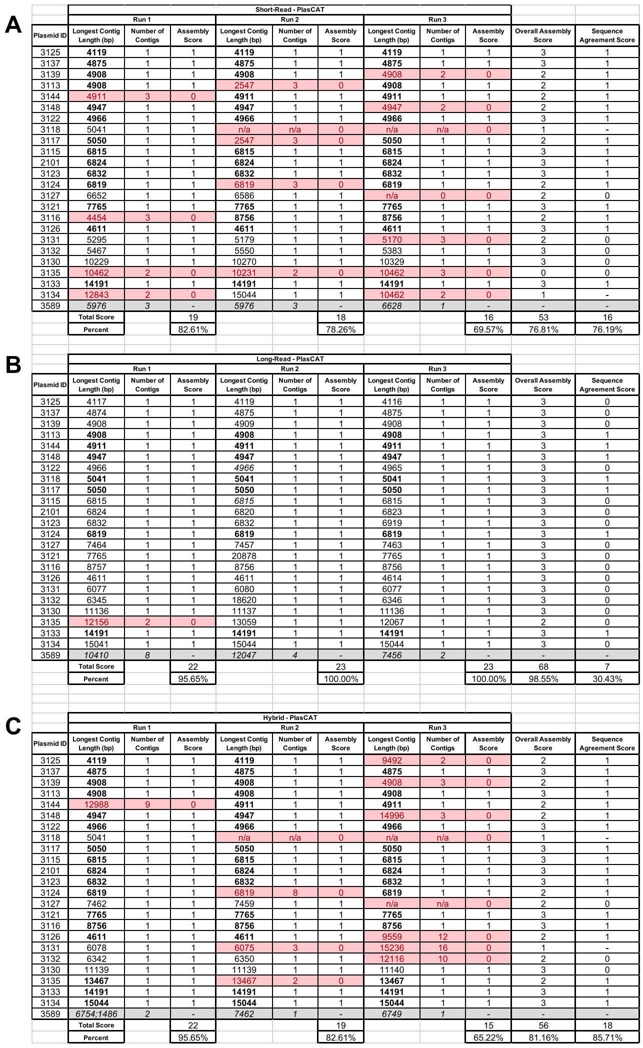


**Figure 1.** Overview of sequencing workflow. Multiple sequencing runs were performed on 24 plasmid samples. For short-read sequencing, plasmids were fragmented and chemically indexed before being loaded onto the Illumina MiSeq. Forward and reverse reads were generated and used with PlasCAT to generate 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.

**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 (nonfragmented) 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,



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 | 7459 | 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 (nonfragmented), 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.

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 that the assembly of most plasmids by short-read sequencing is reproducible across repeated library preparations. The sequence agreement score is ultimately more important than the assembly score, since the consistency provides researchers with a higher level of confidence that their assemblies are correct. Notably, there were 5 samples, including most of the plasmids larger than 10 kb, 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 (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). Nearly 99% of long-read assemblies contained a single contig, likely due, at least in part, to the length of the reads generated and the absence of any fragmentation steps. However, the long-read assemblies were not 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 2−3× longer than the assemblies from other runs (Plasmids 3121 and 3132). If a researcher's goal is only a high-level structural confirmation of the plasmid (i.e., was my gene of interest inserted?), then such assemblies may be sufficient. However, the inconsistency of the runs is a matter of concern both in terms of overall trust in the accuracy of long-read assemblies and when detailed sequence verification is required.

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 three 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).

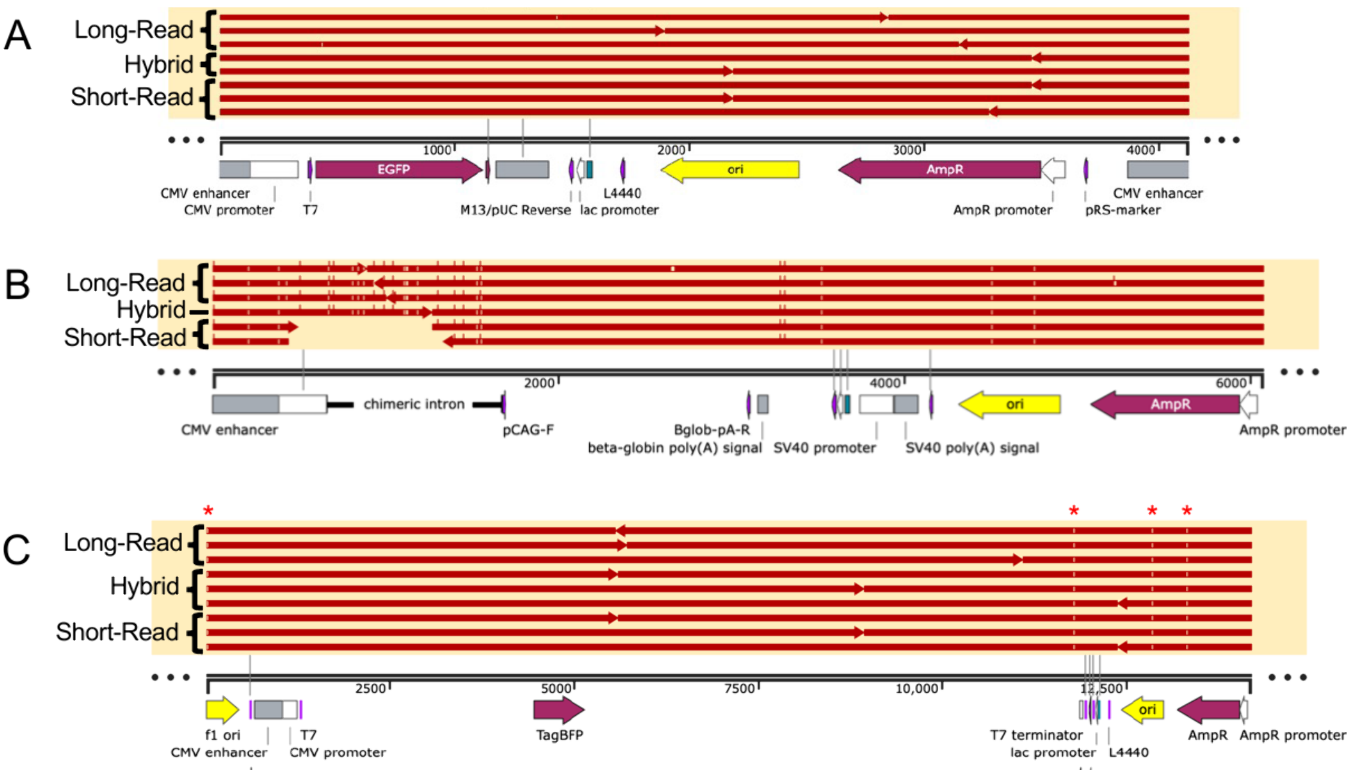


**Figure 3.** Alignment of assembly sequences to a reference sequence. (A) Representative example of a plasmid that is correctly assembled by any method (Plasmid 3125). (B) Short-read assemblies consistently miss GC-rich regions that are resolved by long-read or hybrid assemblies (Plasmid 3131). (C) Some deviations between assembly sequences and the reference sequence are consistent across methods (Plasmid 3133), indicating true mutations (red asterisks). Only successful assemblies were aligned to the reference sequence (1−3 per method). Each red line represents 1 assembly. Insertions, relative to the reference, appear as lines above the red line, while deletions and mismatches appear as gaps in the red line.

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 containing 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 *de novo* assembly tool recommended by Oxford Nanopore. Epi2ME failed to produce an assembly in two cases (Table S3), while PlasCAT always returned an assembly, albeit sometimes fragmented. Aside from these runs, 97% of 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.

Increasing sequencing depth improves assembly quality until it begins to plateau around 50× depth.[63] None of our samples had such low long-read sequencing depth that increasing the depth would have improved the results; almost all samples had initial coverage depth over 20,000× (assuming plasmid size of 5000 bp) with the lowest at roughly 1500×. Yet, a depth that is too high also impedes assembly quality, thus both PlasCAT and Epi2ME subset the long-read data to a particular coverage level to improve assembly results.[29,49,58,64−66] 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 (Supplementary Table S4). The one exception was the two-plasmid mixture (Plasmid 3589) which produced two reasonable contigs sized 7447 bp and 6532 bp. To subset the data, an estimated size for the plasmid must be provided to establish the number of reads needed to achieve a particular coverage level, presenting a potential barrier to *de novo* assembly. However, the default size parameters of PlasCAT and Epi2ME seemed to work well for samples of all sizes. While PlasCAT takes a single subset of the data to produce an assembly, Epi2ME uses Trycycler to generate a consensus assembly from three assemblies generated from three separate subsample sets, which likely improved its sequence agreement scores. Interestingly, reanalysis of the same data with the same parameters on Epi2ME occasionally returned different assemblies, suggesting a truly random seed used for subset generation. Although Trycycler was likely meant to accommodate this randomness, it was not intended to be a fully automated platform and will fail to produce a consensus if any assemblies are too different from one another.[32] 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.

Subsetting the data also led to more practical return times, cutting down the average time per PlasCAT assembly from about 30 to 2 min. All samples from the same PlasCAT run are executed in parallel, resulting in a fast turnaround time for

https://doi.org/10.1021/acssynbio.4c00539

processing large data sets. On the other hand, each Epi2ME assembly took a few minutes per sample, and they were run in succession, so each set of 24 assemblies took 2 to 3 h to complete. Within PlasCAT, long-read assembly was the fastest at only about 2 min per assembly, while each short-read assembly took, on average, 11 min to complete, ranging from 90 s to nearly 25 min. The hybrid assemblies took significantly longer: at least 5 min and up to 48 min.

***De Novo* Assembly Reveals 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;[38,44] 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).[53,67] 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 (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 (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.

**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 nine plasmids that had deviations from the reference sequence in all three types of assemblies (Table S5). Of these, five plasmids consistently deviated from the reference sequence (Figure 3C). For example, the reference for Plasmid 3133 was 14194 bp, but all nine assemblies had six identical mismatches and two small deletions resulting in a 14191 bp plasmid. In addition, eight of the nine assemblies for Plasmid 3121 showed a 19 bp deletion corresponding to the T7 promoter. For all five samples, any successful assemblies that did not exactly match the others were long-read assemblies that supported the deviations but had additional indels.

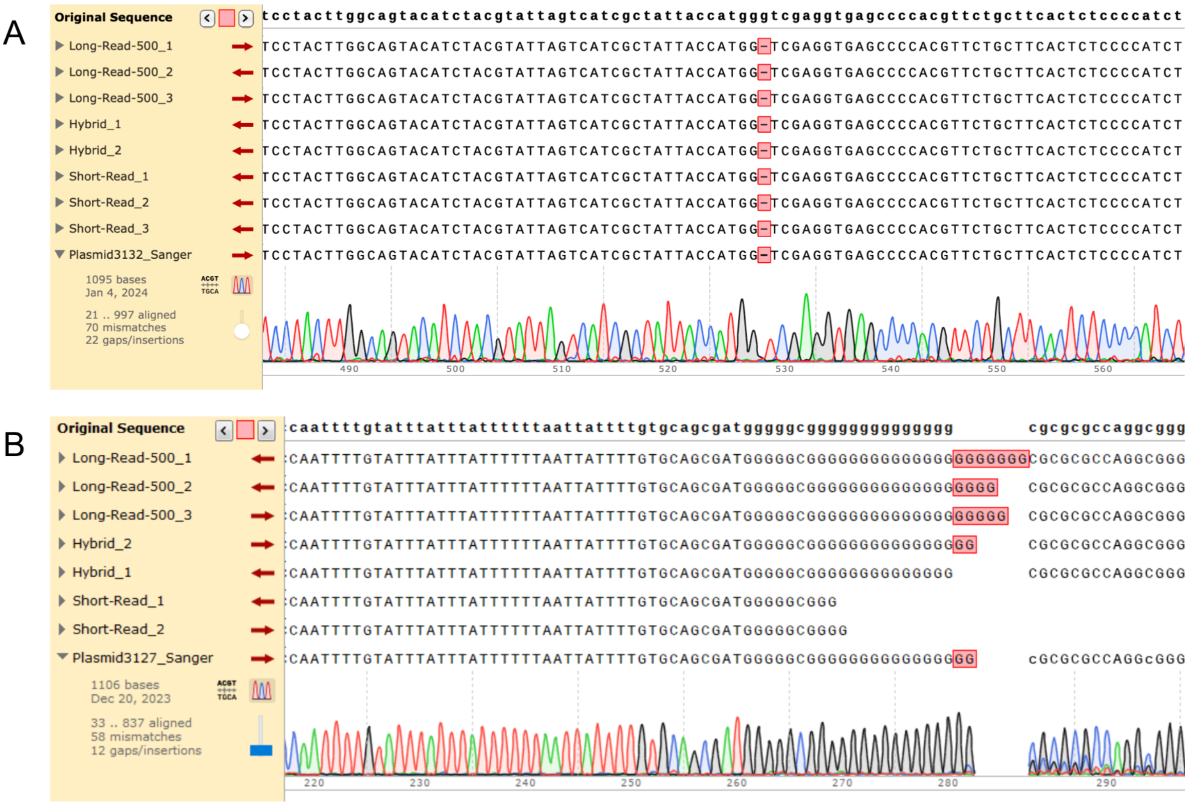
In the four other cases, there were deviations from the reference sequence, but it was not clear what the true sequence was. For example, Plasmid 3131 usually returned a 6077 or 6078 bp from long-read or hybrid assemblies; some errors were consistent, resulting in a plasmid roughly 13 bp larger than the reference, but a consensus could not be reached (Table S5). All four plasmids showed large (>750 bp) gaps in the short-read assemblies corresponding to the chicken $\beta$-actin (CAG) promoter[68] and adjacent chimeric intron (Figure 3B, Table S5). This region is highly repetitive and has a GC content of 73%. This region also contained 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 (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 $\beta$-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). Given the high level of support across methods for certain discrepancies along with our Sanger data, we excluded what seemed to be true deviations, recalculated the number of assemblies with errors, and found that hybrid assemblies had the fewest remaining errors of all types (Table 2).

One may assume that if a reference sequence is available, performing a reference-based assembly would lead to improved results over a *de novo* assembly. We used MIRA[42,69−71] to perform reference-based assembly of short reads for the 9 plasmids where the *de novo* assemblies deviated from the reference. A detailed discussion of these results is provided in the Supporting Information (Supplementary Discussion). Briefly, the *de novo* assemblies were much closer to the expected size for all samples (Table S6). Some of the deviations found in *de novo* assemblies were supported by the reference-based assemblies, however, there were several cases where the reference-based assembly showed evidence of reference bias: the assembly matched the reference sequence even when Sanger data supported the deviations found in the *de novo* assemblies (Figure S2). In addition, reference-based assemblies frequently contained nonstandard nucleotides, even when Sanger data showed clean peaks. These findings suggest that *de novo* assemblies are more accurate than reference-based assemblies, which is especially powerful since accurate reference sequences may not always be available.

## ■ DISCUSSION

Calls for consistent, accurate sequence verification have been left unanswered for too long.[10,12] Even when researchers may not expect single base pair accuracy to be important, it is important to remember that even single base pair changes can have unintended biological effects, exemplified by SNPs, various diseases, and the sequence similarity between certain fluorescent proteins.[72,73] We evaluated several approaches for 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. In several cases, the *de novo* assemblies for a plasmid consistently differed from the reference sequence, regardless of sequencing method, and several of these



A

```
Original Sequence          tcctacttggcagtacatctacgtattagtcatcgctattaccatgggtcgaggtgagccccacgttctgcttcactctccccatct
Long-Read-500_1        →   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
Long-Read-500_2        ←   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
Long-Read-500_3        →   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
Hybrid_1               ←   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
Hybrid_2               ←   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
Short-Read_1           ←   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
Short-Read_2           ←   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
Short-Read_3           ←   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
Plasmid3132_Sanger     →   TCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGG-TCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
   1095 bases
   Jan 4, 2024
   21 .. 997 aligned
   70 mismatches
   22 gaps/insertions
```

B

```
Original Sequence          caatttgtatttatttattttttaattattttgtgcagcgatgggggcggggggggggggg             cgcgcgccaggcggg
Long-Read-500_1        ←   CAATTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGGGGGCGCGCGCCAGGCGGG
Long-Read-500_2        ←   CAATTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGGG      CGCGCGCCAGGCGGG
Long-Read-500_3        →   CAATTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGGGGG     CGCGCGCCAGGCGGG
Hybrid_2               →   CAATTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGG        CGCGCGCCAGGCGGG
Hybrid_1               ←   CAATTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGG          CGCGCGCCAGGCGGG
Short-Read_1           ←   CAATTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGG
Short-Read_2           →   CAATTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGG
Plasmid3127_Sanger     →   CAATTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCGATGGGGGCGGGGGGGGGGGGGGG        cGCGCGCCAGGcGGG
   1106 bases
   Dec 20, 2023
   33 .. 837 aligned
   58 mismatches
   12 gaps/insertions
```

**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 two short-read assemblies missed this region entirely.

### Table 2. Summary of Errors in Plasmid Assemblies Compared to the Reference Sequence, Excluding True Deviations

| assembly type | # successful assemblies | # assemblies with indels ≤ 5 bp | # assemblies with indels > 5 bp | # assemblies with mismatches |
|---|---|---|---|---|
| short read | 53 | 9 | 10 | 8 |
| long read | 68 | 25 | 6 | 19 |
| hybrid | 56 | 8 | 2 | 8 |

deviations were confirmed by Sanger sequencing. Importantly, we found *de novo* assembly to be more accurate than traditional reference-based assembly. Providing a reference sequence sometimes led to reference bias wherein the reference-based assembly preferentially matched the reference sequence, even when Sanger sequencing confirmed a true deviation. Among *de novo* assemblies, the hybrid approach, leveraging the high accuracy of short reads with the ability of long reads to resolve complex regions, led to the best, most reproducible assemblies.

This workflow was developed with the large-scale production and verification of high-throughput plasmid libraries in mind. With the increasing ease and availability of DNA synthesis and sequencing technologies,[74] there is increasing demand for high-throughput methods to produce hundreds to thousands of plasmids. The major bottleneck in the production of these libraries is the sequence verification step.[7,9,12] While one may wish to argue that short- or long-read sequencing on its own can provide sufficient sequence verification, the hybrid approach not only leverages the high accuracy of short reads but also 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. When deviations from the expected sequence arise during hybrid assembly, confidence can be gained by sequencing additional independent replicates of your plasmid. Given the propensity of DNA to mutate, errors could arise in one bacterial colony and not another, and frequent sequencing can help detect which mutations arose and when. Alternatively, sequencing multiple replicates from the outset as we did here, assuming there will be some failure, may save time in the long run despite incurring higher costs. Sanger sequencing can provide additional confirmation of potential errors.

This work aimed to combine different sequencing technologies and bioinformatics tools to develop a high-performance plasmid sequencing pipeline that minimizes the risk of sequencing errors. This risk should always be considered when reviewing plasmid sequencing data. The decision to collect additional data using different technologies and replicate the sequencing experiment should be motivated by an economic analysis considering two factors: the consequences of sequencing errors and the cost of sequencing. The consequences of sequencing errors can be evaluated as the potential economic loss resulting from working with an incorrect plasmid. For example, a company using a plasmid in a regulated biomanufacturing process can justify higher quality control costs than a graduate student using plasmids in relatively cheap phenotyping assays. The cost structure of sequencing data can hinder additional data collection. It is easier to order additional data when paying on a per-sample basis, as when working with specialized service providers. Users operating their own instruments may find it more difficult to justify the cost of a sequencing run if they only have a few plasmids to sequence.

The analysis of the risks of sequencing error and the cost structure of NGS may even lead an investigator to use Sanger sequencing. An increasing number of Sanger reactions is required to cover large sequences, which becomes costly and time-consuming for whole-plasmid sequencing (Table S7). However, Sanger sequencing may still make sense for users who need to accurately verify the inserts of only a limited number of plasmids at a time. While Sanger sequencing is the gold standard for accurately resolving known short sequences,[75] it requires a reference sequence to generate primers, making it impractical for *de novo* assembly. Thus, a Sanger-based approach may be compatible with using long-read sequencing to obtain a high-level overview of the plasmid's structure in order to design primers.

Similarly, long-read plasmid assembly can be sufficient for quickly and cheaply eliminating constructs or colonies that will not produce the desired genotype or phenotype from high-throughput screening experiments. For example, when we perform ligation reactions, we expect that some plasmids will not take up the inserts, thus we perform long-read sequencing to identify samples where ligation failed and continue onto hybrid sequencing with only the promising candidates. These initial long-read assemblies can have issues like indels that may make them less trustworthy but provide a reasonable starting point for further analysis. On the other hand, short-read plasmid assembly may be sufficient for plasmids that are not highly repetitive and with an overall and parts-level GC content of around 50%; otherwise, large fragments may be missing or the assemblies may be fragmented. Although short-read assemblies can 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 (Table S7).

It should be emphasized that certain highly repetitive, GC-rich sequences remained difficult to resolve to single base pair accuracy by any of the bioinformatics pipelines, with the short-read data performing the worst in these areas. Specifically, there were four plasmids containing a GC-rich region encompassing a CAG promoter and adjacent chimeric intron, which could not be resolved by any assembly method. Although targeted Sanger sequencing may resolve these sequences with good confidence (Figure 4), regions with GC content outside the typical (40−60%) range as well as repetitive regions that can form hairpins can present challenges for Sanger, short-read, and long-read sequencing alike.[7,47,75] Optimization of sequencing library preparation kits that can overcome the GC-bias would make short-read sequencing a more desirable approach. For example, the ExpressPlex kit from SeqWell is more resilient to a range of GC values and to lower inputs of DNA which mitigates some of the issues we found with short-read sequencing.[76] If some genetic parts are unreliable to sequence regardless of the method, especially to single base pair accuracy, these parts must be cataloged and flagged as difficult sequences. Ideally, these parts would be replaced with others that can perform the same function but are more readily sequenced. This is especially important for developing new plasmid backbones that can be used to produce a wide variety of plasmids that are easily sequenced.

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 nontechnical users to carry out complex bioinformatics analyses. 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.[1,77] Both PlasCAT and Epi2ME are 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. As sequencing technologies and bioinformatics tools continue to improve, further work is needed to optimize analysis pipelines for *de novo* plasmid assembly as well as improve verification and documentation practices surrounding plasmids. Recent work has led to the development of DNA signatures that can embed identifying information directly into plasmid sequences to facilitate simpler plasmid verification using only *de novo* assembly.[36,78] 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 accurate *de novo* assembly methods.[36,78]

## ■ ASSOCIATED CONTENT

### Data Availability Statement

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

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.4c00539.

All Supporting Information, including additional methods, data, and references on reference-based assembly; Sanger primer sequences; iSeq short-read assemblies; Epi2ME assemblies; long-read assemblies without subsampling; summary of errors found in assemblies; cost analysis; summary of read statistics before and after filtering (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

**Jean Peccoud** − *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, Colorado 80523, United States of America;* orcid.org/0000-0001-7649-6127; Phone: 1-970-491-2482; Email: jean.peccoud@colostate.edu

### Authors

**Sarah I. Hernandez** − *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, Colorado 80523, United States of America*

**Casey-Tyler Berezin** − *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, Colorado 80523, United States of America*

**Katie M. Miller** − *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, Colorado 80523, United States of America*

**Samuel J. Peccoud** − *Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, Colorado 80523, United States of America*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acssynbio.4c00539

### Author Contributions

S.H. performed conceptualization, methodology, formal analysis, visualization, and writing—original draft and editing. C.-T.B. performed conceptualization, formal analysis, software, and writing—review and editing. K.M. performed formal analysis and writing—review. S.P. developed software. J.P. performed conceptualization, methodology, and writing—review and 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.

### Notes

The authors declare the following competing financial interest(s): J.P., K.M., and S.P. have financial interests in GenoFAB, Inc., a company that may benefit or be perceived as benefiting from this publication.

## ■ ACKNOWLEDGMENTS

We thank Dr. Stephen Coleman for access to the Illumina iSeq sequencing machine. Figure [1] was created with BioRender.com.

## ■ REFERENCES

(1) Brown, S. D.; Dreolini, L.; Wilson, J. F.; Balasundaram, M.; Holt, R. A. Complete sequence verifcation of plasmid DNA using the Oxford Nanopore Technologies' MinION device. *BMC Bioinf.* **2023**, *24* (116), 166.

(2) Rozwandowicz, M.; Brouwer, M. S. M.; Fischer, J.; Wagenaar, J. A.; Gonzalez-Zorn, B.; Guerra, B.; Mevius, D. J.; Hordijk, J. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J. Antimicrob. Chemother.* **2018**, *73*, 1121−1137.

(3) Cameron, D. E.; Bashor, C. J.; Collins, J. J. A brief history of synthetic biology. *Nature Reviews Microbiology* **2014**, *12* (5), 381−390.

(4) Munnelly, K. Engineering for the 21st Century: Synthetic Biology. *ACS Synth. Biol.* **2013**, *2*, 213−215.

(5) Peccoud, J. Synthetic Biology: fostering the cyber-biological revolution. *Synth. Biol.* **2016**, *1* (1), No. ysw001.

(6) Ghaffarifar, F. Plasmid DNA vaccines: where are we now? *Drugs Today* **2018**, *54* (5), 315−333.

(7) Gallegos, J. E.; Rogers, M. F.; Cialek, C. A.; Peccoud, J. Rapid, robust plasmid verification by de novo assembly of short sequencing reads. *Nucleic Acids Res.* **2020**, *48* (18), No. e106.

(8) Shapland, E. B.; Holmes, V.; Reeves, C. D.; Sorokin, E.; Durot, M.; Platt, D.; Allen, C.; Dean, J.; Serber, Z.; Newman, J.; Chandran, S. Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process. *ACS Synth. Biol.* **2015**, *4* (7), 860−866.

(9) Peccoud, J. Data sharing policies: share well and you shall be rewarded. *Synth. Biol.* **2021**, *6* (1), No. ysab028.

(10) Peccoud, J.; Anderson, J. C.; Chandran, D.; Densmore, D.; Galdzicki, M.; Lux, M. W.; Rodriguez, C. A.; Stan, G.-B.; Sauro, H. M. Essential information for synthetic DNA sequences. *Nat. Biotechnol.* **2011**, *29* (1), 22−22.

(11) Peccoud, J.; Gallegos, J. E.; Murch, R.; Buchholz, W. G.; Raman, S. Cyberbiosecurity: from naive trust to risk awareness. *Trends Biotechnol.* **2018**, *36* (1), 4−7.

(12) Thuronyi, B. W.; DeBenedictis, E. A.; Barrick, J. E. No assembly required: Time for stronger, simpler publishing standards for DNA sequences. *Plos Biology* **2023**, *21* (11), No. e3002376.

(13) Kar, D. M.; Ray, I.; Gallegos, J.; Peccoud, J.; Digital Signatures to Ensure the Authenticity and Integrity of Synthetic DNA Molecules. In *Nspw '18: Proceedings of the New Security Paradigms Workshop* Association for Computing Machinery 2018, 110 122 DOI: .

(14) Kar, D. M.; Ray, I.; Gallegos, J.; Peccoud, J.; Ray, I. Synthesizing DNA molecules with identity-based digital signatures to prevent malicious tampering and enabling source attribution. *Journal of Computer Security* **2020**, *28* (4), 437−467.

(15) Gallegos, J. E.; Kar, D. M.; Ray, I.; Ray, I.; Peccoud, J. Securing the Exchange of Synthetic Genetic Constructs Using Digital Signatures. *ACS Synth. Biol.* **2020**, *9* (10), 2656−2664.

(16) Peccoud, J.; Gallegos, J. E.; Murch, R.; Buchholz, W. G.; Raman, S. Cyberbiosecurity: From Naive Trust to Risk Awareness. *Trends Biotechnol* **2018**, *36* (1), 4−7.

(17) Murch, R. S.; So, W. K.; Buchholz, W. G.; Raman, S.; Peccoud, J. Cyberbiosecurity: An Emerging New Discipline to Help Safeguard the Bioeconomy. *Front Bioeng Biotechnol* **2018**, *6*, 39.

(18) Gallegos, J. E.; Rogers, M. F.; Cialek, C. A.; Peccoud, J. Rapid, robust plasmid verification by de novo assembly of short sequencing reads. *Nucleic Acids Res.* **2020**, *48* (18), No. e106.

(19) Peccoud, S.; Berezin, C. T.; Hernandez, S. I.; Peccoud, J.; Alkan, C. PlasCAT: Plasmid Cloud Assembly Tool. *Bioinformatics* **2024**, *40* (5), No. btae299.

(20) Epi2ME Labs. *Epi2ME Labs Blog*. Epi2ME, 2023.

(21) Antipov, D.; Hartwick, N.; Shen, M.; Raiko, M.; Lapidus, A.; Pevzner, P. A. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* **2016**, *32* (22), 3380−3387.

(22) Rozov, R.; Brown Kav, A.; Bogumil, D.; Shterzer, N.; Halperin, E.; Mizrahi, I.; Shamir, R. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* **2017**, *33* (4), 475−482.

(23) Antipov, D.; Raiko, M.; Lapidus, A.; Pevzner, P. A. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.* **2019**, *29* (6), 961−968.

(24) Gomi, R.; Wyres, K. L.; Holt, K. E. Detection of plasmid contigs in draft genome assemblies using customized Kraken databases. *Microb. Genomics* **2021**, *7* (4), No. 000550.

(25) Pellow, D.; Zorea, A.; Probst, M.; Furman, O.; Segal, A.; Mizrahi, I.; Shamir, R. SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome* **2021**, *9* (1), 144.

(26) Gupta, S. K.; Raza, S.; Unno, T. Comparison of de-novo assembly tools for plasmid metagenome analysis. *Genes & Genomics* **2019**, *41* (9), 1077−1083.

(27) Wick, R. R.; Judd, L. M.; Gorrie, C. L.; Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology* **2017**, *13* (6), No. e1005595.

(28) Tang, X.; Shang, J.; Ji, Y.; Sun, Y. PLASMe: a tool to identify PLASMid contigs from short-read assemblies using transformer. *Nucleic Acids Res.* **2023**, *51* (15), e83−e83.

(29) Bouras, G.; Sheppard, A. E.; Mallawaarachchi, V.; Vreugde, S.; Marschall, T. Plassembler: an automated bacterial plasmid assembly tool. *Bioinformatics* **2023**, *39* (7), No. btad409.

(30) Bas, B.; Saltykova, A.; Garcia-Graells, C.; Philipp, P.; Arella, F.; Marchal, K.; Winard, R.; Vanneste, K.; Roosens, N. H. C.; De Keersmaecker, S. C. J. Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modifed Bacillus. *Sci. Rep.* **2020**, *10*, 4310.

(31) Johnson, J.; Soehnlen, M.; Blankenship, H. M. Long read genome assemblers struggle with small plasmids. *Microb. Genomics* **2023**, *9* (5), No. mgen001024.

(32) Wick, R. R.; Judd, L. M.; Cerdeira, L. T.; Hawkey, J.; Méric, G.; Vezina, B.; Wyres, K. L.; Holt, K. E. Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol.* **2021**, *22*, 266.

(33) Chen, N.-C.; Solomon, B.; Mun, T.; Iyer, S.; Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* **2021**, *22* (1), 8.

(34) Valiente-Mullor, C.; Beamud, B.; Ansari, I.; Francés-Cuesta, C.; García-González, N.; Mejía, L.; Ruiz-Hueso, P.; González-Candelas, F. 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), No. e1008678.

(35) Lau, J. Reference bias: Challenges and solutions. In *SevenBridges Blog*, 2017.

(36) Gallegos, J. E.; Kar, D. M.; Ray, I.; Ray, I.; Peccoud, J. Securing the Exchange of Synthetic Genetic Constructs Using Digital Signatures. *ACS Synth. Biol.* **2020**, *9* (10), 2656−2664.

(37) Peccoud, S.; Berezin, C. T.; Hernandez, S. I.; Peccoud, J.; Alkan, C. PlasCAT: Plasmid Cloud Assembly Tool. *Bioinformatics* **2024**, *40*, No. btae299.

(38) Mardis, E. R. Next-Generation Sequencing Platforms. *Annu. Rev. Anal. Chem.* **2013**, *6*, 287−303.

(39) Hu, T.; Chitnis, N.; Monos, D.; Dinh, A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **2021**, *82* (11), 801−811.

(40) Heather, J. M.; Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **2016**, *107* (1), 1−8.

(41) Peccoud, J.; Blauvelt, M. F.; Cai, Y.; Cooper, K. L.; Crasta, O.; DeLalla, E. C.; Evans, C.; Folkerts, O.; Lyons, B. M.; Mane, S. P.; Shelton, R.; Sweede, M. A.; Waldon, S. A.; Thattai, M. Targeted development of registries of biological parts. *PLoS One* **2008**, *3* (7), No. e2671.

(42) Wilson, M. L.; Cai, Y.; Hanlon, R.; Taylor, S.; Chevreux, B.; Setubal, J. C.; Tyler, B. M.; Peccoud, J. Sequence verification of synthetic DNA by assembly of sequencing reads. *Nucleic Acids Res.* **2012**, *41* (1), e25.

(43) Buermans, H. P. J.; den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2014**, *1842* (10), 1932−1941.

(44) Tilak, M.-K.; Botero-Castro, F.; Galtier, N.; Nabholz, B. Illumina Library Preparation for Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA. *Genome Biology and Evolution* **2018**, *10* (2), 616−622.

(45) Liao, X.; Li, M.; Zou, Y.; Wu, F.; Yi-Pan; Wang, J. Current challenges and solutions of de novo assembly. *Quant. Biol.* **2019**, *7*, 90−109.

(46) Aird, D.; Ross, M. G.; Chen, W.-S.; Danielsson, M.; Fennell, T.; Russ, C.; Jaffe, D. B.; Nusbaum, C.; Gnirke, A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **2011**, *12*, R18.

(47) Browne, P. D.; Nielsen, T. K.; Kot, W.; Aggerholm, A.; Gilbert, M. T. P.; Puetz, L.; Rasmussen, M.; Zervas, A.; Hansen, L. H. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience* **2020**, *9* (2), No. giaa008.

(48) Zhao, W.; Zeng, W.; Pang, B.; Luo, M.; Peng, Y.; Xu, J.; Kan, B.; Li, Z.; Lu, X. Oxford nanopore long-read sequencing enables the generation of complete bacterial and plasmid genomes without short-read sequencing. *Front. Microbiol.* **2023**, *14*, No. 1179966.

(49) De Maio, N.; Shaw, L. P.; Hubbard, A.; George, S.; Sanderson, N. D.; Swann, J.; Wick, R.; AbuOun, M.; Stubberfield, E.; Hoosdally, S. J.; Crook, D. W.; Peto, T. E. A.; Sheppard, A. E.; Bailey, M. J.; Read, D. S.; Anjum, M. F.; Walker, A. S.; Stoesser, N. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genomics* **2019**, *5* (9), No. e000294.

(50) Xia, Y.; Li, X.; Wu, Z.; Nie, C.; Cheng, Z.; Sun, Y.; Liu, L.; Zhang, T. Strategies and tools in illumina and nanopore-integrated metagenomic analysis of microbiome data. *iMeta* **2023**, *2* (1), No. e72.

(51) Khrenova, M. G.; Panova, T. V.; Rodin, V. A.; Kryakvin, M. A.; Lukyanov, D. A.; Osterman, I. A.; Zvereva, M. I. Nanopore sequencing for de novo bacterial genome assembly and search for single-nucleotide polymorphism. *International Journal of Molecular Sciences* **2022**, *23* (15), 8569.

(52) Gallegos, J. E.; Hayrynen, S.; Adames, N. R.; Peccoud, J. Challenges and opportunities for strain verification by whole-genome sequencing. *Sci. Rep.* **2020**, *10* (1), 5873.

(53) Amarasinghe, S. L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M. E.; Gouil, Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **2020**, *21*, 30.

(54) ThermoFisher Scientific. *Expi293 Expression System USER GUIDE*. ThermoFisher Scientific: 2020; pp 1−32.

(55) Janeway, C. A., Jr.; T, P.; Walport, M.; et al.The structure of a typical antibody molecule. In *Immunobiology: The Immune System in Health and Disease*, Garland Science: New York, 2001.

(56) Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30* (15), 2114−2120.

(57) Hall, M. B. Rasusa: Randomly subsample sequencing reads to a specified coverage. *Journal of Open Source Software* **2022**, *7* (69), 3941.

(58) Lonardi, S.; Mirebrahim, H.; Wanamaker, S.; Alpert, M.; Ciardo, G.; Duma, D.; Close, T. J. When less is more: 'slicing' sequencing data improves read decoding accuracy and de novo assembly quality. *Bioinformatics* **2015**, *31* (18), 2972−2980.

(59) Wick, R. R.; Menzel, P.*Filtlong*. 2021.

(60) Vaser, R.; Sovic, I.; Nagarajan, N.; Šikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **2017**, *27* (5), 737−746.

(61) Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **2019**, *37*, 540−546.

(62) Wright, C.; Griffiths, S.; Nicholls, S.; Parker, M.; Horner, N. *epi2me-labs/wf-denovo-assembly*. 2022.

(63) Zhang, T.; Xing, W.; Wang, A.; Zhang, N.; Jia, L.; Ma, S.; Xia, Q. Comparison of long-read methods for sequencing and assembly of lepidopteran pest genomes. *International Journal of Molecular Sciences* **2023**, *24* (1), 649.

(64) Murigneux, V.; Rai, S. K.; Furtado, A.; Bruxner, T. J. C.; Tian, W.; Harliwong, I.; Wei, H.; Yang, B.; Ye, Q.; Anderson, E.; et al. Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience* **2020**, *9* (12), No. giaa146.

(65) Wick, R. R.; Holt, K. E. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research* **2019**, *8*, 2138.

(66) Hall, M. B. Rasusa: Randomly subsample sequencing reads to a specified coverage. *J. Open Source Software* **2022**, *7* (69), 3941.

(67) Boostrom, I.; Portal, E. A.; Spiller, O. B.; Walsh, T. R.; Sands, K. 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*, No. 796465.

(68) Alexopoulou, A. N.; Couchman, J. R.; Whiteford, J. R. The CMV early enhancer/chicken $\beta$ 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), 2.

(69) Chevreux, B.; Wetter, T.; Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *German Conf. Bioinf.* **1999**, *99*, 45−56.

(70) Cock, P. J.; Grüning, B. A.; Paszkiewicz, K.; Pritchard, L. Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. *PeerJ.* **2013**, *1*, No. e167.

(71) The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* **2022**, *50* (W1), W345−W351.

(72) Morise, H.; Shimomura, O.; Johnson, F. H.; Winant, J. Intermolecular Energy Transfer in the Bioluminescent System of Aequorea. *Biochemistry* **1974**, *13* (12), 2656−2662.

(73) Weiner, M. P.; Hudson, T. J. Introduction to SNPs: discovery of markers for disease. *Biotechniques* **2002**, *32*, S4−S13.

(74) Hughes, R. A.; Ellington, A. D. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harbor Perspect. Biol.* **2017**, *9* (1), No. a023812.

(75) Crossley, B. M.; Bai, J.; Glaser, A.; Maes, R.; Porter, E.; Killian, M. L.; Clement, T.; Toohey-Kurth, K. *Guidelines for Sanger sequencing and molecular assay monitoring.* **2020**, *32* (6), 767−775.

(76) SeqWell. *ExpressPlex Library Prep Kit*. 2024. https://seqwell.com/expressplex-library-prep-kit/

(77) Wick, R. R.; Judd, L. M.; Holt, K. E. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *PLOS Computational Biology* **2023**, *19* (3), No. e1010905.

(78) Berezin, C.-T.; Peccoud, S.; Kar, D. M.; Peccoud, J. Cryptographic approaches to authenticating synthetic DNA sequences. *Trends Biotechnol.* **2024**, *42*, 1002.