

Rapid, robust plasmid verification by *de novo* assembly of short sequencing reads

Jenna E. Gallegos^{1,†}, Mark F. Rogers^{2,†}, Charlotte A. Cialek^{2,3} and Jean Peccoud^{1,2,*}

¹Department of Chemical & Biological Engineering, Colorado State University, USA, ²GenoFAB, Inc. and

³Department of Biochemistry and Molecular Biology, Colorado State University, USA

Received July 02, 2020; Revised August 07, 2020; Editorial Decision August 13, 2020; Accepted August 24, 2020

ABSTRACT

Plasmids are a foundational tool for basic and applied research across all subfields of biology. Increasingly, researchers in synthetic biology are relying on and developing massive libraries of plasmids as vectors for directed evolution, combinatorial gene circuit tests, and for CRISPR multiplexing. Verification of plasmid sequences following synthesis is a crucial quality control step that creates a bottleneck in plasmid fabrication workflows. Crucially, researchers often elect to forego the cumbersome verification step, potentially leading to reproducibility and—depending on the application—security issues. In order to facilitate plasmid verification to improve the quality and reproducibility of life science research, we developed a fast, simple, and open source pipeline for assembly and verification of plasmid sequences from Illumina reads. We demonstrate that our pipeline, which relies on *de novo* assembly, can also be used to detect contaminating sequences in plasmid samples. In addition to presenting our pipeline, we discuss the role for verification and quality control in the increasingly complex life science workflows ushered in by synthetic biology.

INTRODUCTION

As synthetic biology programs increase in scale (1–3) workflows involving the high-throughput construction of dozens or even hundreds of plasmids are becoming increasingly common (4–6). DNA sequencing is an integral part of fabrication workflows involving the assembly of synthetic DNA fragments (7–10). Sequencing data can detect single point mutations (SNP) resulting from mistakes in the DNA synthesis processes (11–13) or from PCR. In addition, sequencing data can be used to detect structural issues such as inversion or duplication of genetic elements introduced during plasmid assembly.

Traditionally, plasmids have been verified using Sanger sequencing. This sequencing method requires a short, known sequence to initiate an up to ~1000 nucleotide read. Typically, Sanger sequencing is used to verify a sequence of interest that has been inserted into a plasmid backbone by sequencing from known universal primer binding-sites on either side of the multiple cloning site, providing 2× coverage of the insert. Many bioinformatics applications used to edit plasmid sequences have features that facilitate the visual inspection of the alignment of the sequence reads and the plasmid sequence. This approach is practical to verify the insert of a limited number of plasmids. The quality and length of Sanger sequencing reads also simplify sequence assembly.

However, Sanger sequencing is not a viable quality control option for verifying sequences of a large number of plasmid libraries. In order to sequence whole plasmids by Sanger sequencing, the user must first design and order primers, which adds to the time and cost involved in verification. Depending on the sequence of the plasmid, it may be difficult to design primers that will generate ample coverage, and structural features like hairpins can result in low quality reads (14). In addition, the analysis of sequencing reads is time consuming and error prone.

In order to streamline this process, we previously developed an application to automate assembly and sequence verification of plasmid sequences from Sanger sequencing reads (15). However, because this pipeline relies on reference-based assembly, it struggles to resolve duplications, inversions, and other rearrangements that result in structural issues.

Recently, synthetic biologists have demonstrated the possibility of sequencing multiple plasmids in a single Oxford Nanopore run (16). This multiplexed approach overcomes many of the limitations of Sanger sequencing. Oxford Nanopore sequencing is fast, inexpensive, and does not require primers. However, the analysis of data still relies on a reference-based assembly.

Reference-based assembly is disadvantageous for quality control workflows because it is biased towards a particular goal. If there are contaminating sequences present in a sam-

*To whom correspondence should be addressed. Tel: +1 970 491 2482; Email: jean.peccoud@colostate.edu

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

ple, for instance, the reads for those contaminants may be thrown out during the assembly process. Reference-based assembly also requires a full reference sequence by definition. However, researchers may sometimes wish to identify unknown plasmid samples, and even for known samples, a full reference sequence is not always available. It is very common for plasmids described in the literature to be accompanied only by a visual map or vague descriptions of how the plasmid was constructed. Without a full reference sequence, it is not possible to design primers for Sanger sequencing or conduct a reference-based assembly.

To address these limitations, we present an alternative approach based on short read sequencing and *de novo* assembly. The costs associated with short read sequencing are rapidly decreasing while speed and read length are increasing (17,18). The small amount of coverage needed for plasmids and various multiplexing options have also made short read sequencing a cost-effective option, as plasmid samples can easily be included within larger sequencing runs. Further, short read sequencing does not rely on primers and is less sensitive to secondary structures than long read sequencing. Though *de novo* pipelines exist for circular microbial genomes, these tools do not translate well to assembling plasmids, which are typically much smaller in size. DNA synthesis companies routinely rely on short read sequencing workflows for quality control in plasmid synthesis, but their assembly and verification tools are proprietary, and likely rely on reference-based assembly. Proprietary programs also exist in academic labs. However, our approach is the first open source, publicly available solution.

In this manuscript, we describe a plasmid verification pipeline that uses Illumina sequencing reads for *de novo* assembly. The pipeline was tested on a library of 96 plasmids designed to represent a broad range of variations of a common plasmid template. The tool produced correct assemblies for all 96 plasmid samples even when overall read quality was very low. Furthermore, we demonstrate that our script can determine whether or not a pool of reads is likely to contain contaminating sequences. We built a contamination test into our verification pipeline that informs the user of the likelihood of contamination. This tool can thus be used to verify known plasmid sequences, identify unknown plasmid sequences, and detect contamination in plasmid samples. The workflow described in this manuscript relies entirely on open source tools, making it, to our knowledge, the first non-proprietary tool for *de novo* assembly of plasmid sequences from short sequencing reads, and the first sequence assembly and verification tool that also predicts the likelihood of contamination.

MATERIALS AND METHODS

Plasmid preparation and sequencing

We built and tested our pipeline using Illumina sequencing reads from 96 plasmid DNA samples as previously described (19). Briefly, 6–24 individual transformants of eight different plasmids (ranging in size from 2521 to 3294 bp) were sequenced. Six of these plasmids are part of a family of synthetic plasmids that were synthesized, and sequence verified by Twist Biosciences (San Francisco, CA, USA). Five of these six differ from each other by one to four SNPs

which were intentionally introduced. These are referred to as ‘known-SNPs’. The sixth plasmid from Twist differs in that it lacks a 608 bp insert present in the other five. This variant was used to analyze the impact of contaminants containing insertions and deletions (INDELS). The remaining two plasmids are unrelated and were generated by Gibson assembly (20) of parts that have not been sequence verified. These were, therefore, expected to contain a small and variable number of unknown SNPs and INDELS.

Plasmids were isolated from *E. coli* cells, analyzed on an Agilent TapeStation, and submitted to seqWell (Beverly, MA, USA) for sequencing. seqWell uses a library prep technology called plexWell that enables the preparation of hundreds or thousands of multiplexed samples for Illumina sequencing in just 3 hours.

Sequencing resulted in 2.9 million read pairs for the 96 samples ($>1000 \times$ coverage), with read lengths of 35 to 251 bases and mean per-read PHRED quality scores (21) ranging from 2.0 to 38.8 (average of 35.3). Optimal filtering conditions removed 93% of reads, yielding 1682 to 11 938 filtered read pairs per plasmid with lengths of 125 to 251 bases and mean per-read quality from 36.8 to 38.8 (average of 38.0). In addition to raw reads in the form of FASTQ files, seqWell used their proprietary bioinformatics workflow to analyze the sequencing reads. They provided assembled FASTA files for 95 of the 96 samples.

De Novo assembly and sequence verification pipeline

Our sequence verification pipeline performs three key steps: quality filtering, plasmid assembly and assembly evaluation. It accepts as input, Illumina paired-end sequencing FASTQ files containing sequencing reads and produces as output a FASTA assembly file representing the predicted plasmid reference sequence, along with an estimated likelihood of contamination (Figure 1). Reference sequences were used to build and evaluate the pipeline; however, it is important to emphasize that the pipeline itself does not use a reference.

The first step is to filter the input sequences to eliminate all but the highest-quality reads. We tested two filtering tools, Trimmomatic, a commonly used open source software tool (22), and fastp, a more recent, high-throughput method, for comparison (23). Illumina sequencing can produce extremely high coverage for short plasmid sequences. Our experimental samples provided anywhere from 36 000 to 166 000 reads per plasmid, yielding coverage from $2800 \times$ to over $12 000 \times$. These values are orders of magnitude higher than typical recommendations for *de novo* genome assembly, which range from $13 \times$ to $60 \times$ (24,25). Higher coverage does not improve performance, and often hurts performance, sometimes yielding assemblies with multiple sequences (putative contigs) (26). For this reason, we used stringent criteria for filtering reads, retaining only the top 3% to 11% highest quality reads.

The second step is sequence assembly, in which two FASTQ files containing trimmed, forward and reverse, Illumina paired-end sequencing reads are accepted as input, and a FASTA file containing one (or more) circular contig(s) is produced as output. For this, we considered five different assembly tools for our process: Unicycler (27),

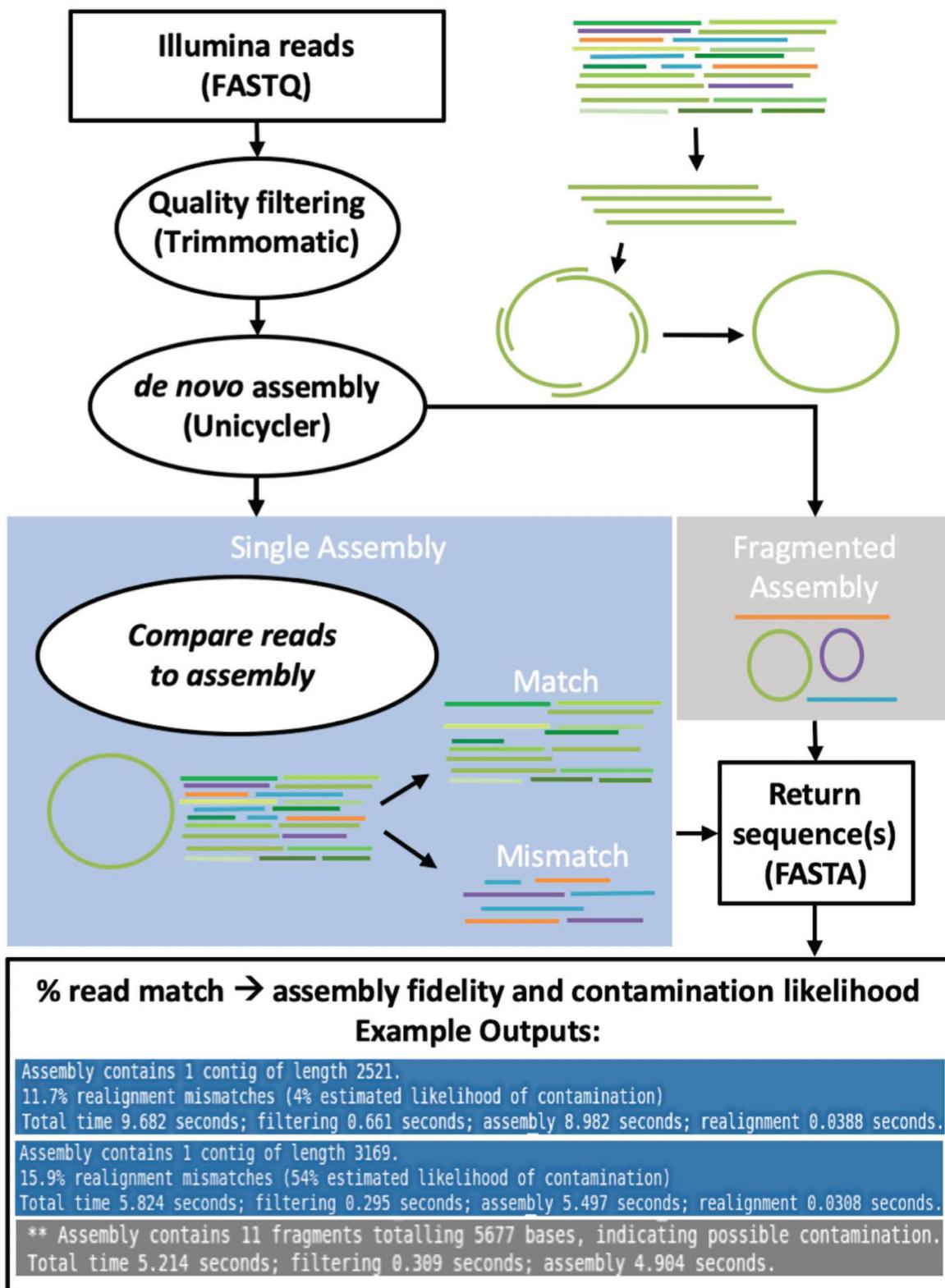


Figure 1. Schematic of pipeline. Illumina reads in the form of FASTQ files are filtered using Trimmomatic and assembled using Unicycler with parameters that have been specifically optimized for plasmid assembly. If a single assembly results, the unfiltered reads are compared with the assembled sequence and sorted according to whether or not they match 100%. Assembled sequences are returned as FASTA files. The percentage of matched reads is used to calculate the likelihood of contamination.

MIRA (28), Velvet (29), plasmidSPADEs (30) and Circlator (31). MIRA and Velvet are popular, general purpose assemblers that generate linear assemblies. Circlator is a post-processing tool that can convert linear assemblies to circular assemblies, while plasmidSPADEs and Unicycler are both designed to resolve chromosomes and plasmids from whole bacterial genome sequencing, yielding circular plasmid assemblies. We used MIRA and Velvet on their own and with Circlator, and we used Unicycler and plasmidSPADEs by themselves.

When assessing the quality of assembly of putative contigs, we imposed strict criteria: (i) the final assembly must contain only one contig that matches the correct plasmid sequence exactly and (ii) the pipeline should take no more than a few minutes to run.

Assessing assembly quality and detecting contamination

Because our pipeline is intended for use in the absence of a reference sequence, once our pipeline was established, it was necessary to develop additional reference-agnostic methods for assessing plasmid assembly. For this, we built an additional quality control step into the pipeline in which the filtered reads are mapped back to the assembly. The proportion of reads that matches the assembled contig exactly is used to assess the quality of the assembly.

The proportion of reads that matches the assembled contig was further used to assess the likelihood of contamination in the original sample. There are several scenarios by which a plasmid sample might become contaminated. For instance, if plasmid DNA is extracted from a mixed population of transformants, depending on the method of plasmid construction and screening, there could be sequences present that have SNPs or that lack or contain an extra insert. Additionally, if a contamination event occurs between different laboratory strains, sequences from completely unrelated plasmids may be present in the mixed sample. In each case, when the contaminated sample is sequenced, the resulting reads will include a combination of reads from the primary sample and reads from the contaminant.

To simulate each of these possible contaminations (SNPs, INDELs and unrelated sequences), we conducted a series of contamination simulation experiments by randomly sampling the filtered reads from distinct plasmids to create artificial contamination, and then built a procedure for detecting the contamination.

In each of these contamination simulation experiments, we select one of our plasmids and label it the ‘primary’ plasmid, and its filtered reads, the ‘primary’ read set. We then select a second, ‘contaminant’ plasmid at random that has a different reference sequence from that of the primary plasmid. In each contamination simulation experiment, a subset of reads from the contaminant library is artificially combined with a subset of reads from the primary library, the reads are assembled, and the resulting assemblies are assessed. Different contamination levels (up to 50%) were assessed.

Our contamination detection algorithm uses simple string matching to compare an output assembly with the filtered reads used to create it. The algorithm iterates over all of the distinct high-quality reads, attempting to match

them to a location in the assembly sequence. It assesses the percentage of distinct reads that map exactly to a location in the assembly and compares it to distributions we have established experimentally.

RESULTS

Plasmid Identification and Verification

For quality filtering, we found that Trimmomatic (22) yielded reliably good data sets when we applied strict quality filtering that retained only the highest-quality 3% to 11% of read pairs (Figure 2). Users of our script may choose instead to use *fastp* (23) for quality filtering, since we found that both methods yielded similar results.

For assembly, the only tools that yielded single-contig assemblies in any of our tests were plasmidSPADEs and Unicycler. For our filtered data, plasmidSPADEs yielded 63/96 correct assemblies (66%) while Unicycler resolved all 96 correctly. Thus, we settled on the Unicycler program because we were able to derive quality assemblies with fewer parameter adjustments. It is worth noting that Unicycler is really a sophisticated wrapper for SPADES (32).

Interestingly, none of the assembly tools we tried worked well with whole sets of Illumina reads even though the data for most of the samples were relatively high quality (average PHRED score of 35.3, Figure 2). The most likely explanation is an overabundance of input data, a paradox that poses a significant challenge for *de novo* assembly when sequencing coverage exceeds 500x (26).

Thus, we tested whether high-quality reads or the number of reads were necessary to produce accurate assemblies. We ran Unicycler on random samples of read pairs such that the resulting input data contained the same number of pairs as the quality-filtered data sets. We found that the randomly sampled read pairs yielded assemblies with the same fidelity as those built from high-quality reads (Supplemental Figure TBD or data not shown). However, we persisted in using quality filtering as it reduced read quantity to a manageable level and the high-quality reads became important when predicting contamination. Using read filtering, optimal assemblies were achieved using coverage between 113× and 1030× (data not shown).

For plasmids ranging from 2.5 to 3.3 kb, complete assemblies by Unicycler with our plasmid-specific parameters required on average less than eight seconds (single processor on a Linux laptop with 16GB RAM), including Trimmomatic filtering. The Unicycler assembly process itself dominated this time, which increased approximately linearly with the number of input reads (Figure 3).

For 95 of the 96 experimental samples, our pipeline’s assembled contig matched the sequence generated by seqWell using their proprietary plasmid assembly workflow. For the one exception, seqWell was unable to provide an assembly, reporting that it was due to poor read quality. In this case, our pipeline was able to generate an assembly successfully after the quality filtering step eliminated more than 95% of the total read pairs.

72 of our 96 samples are clones derived from plasmids that were previously sequence-verified by Twist. It is worth noting that 48 of these 72 plasmids had the capacity to form substantial hairpins resulting in practically uninterpretable

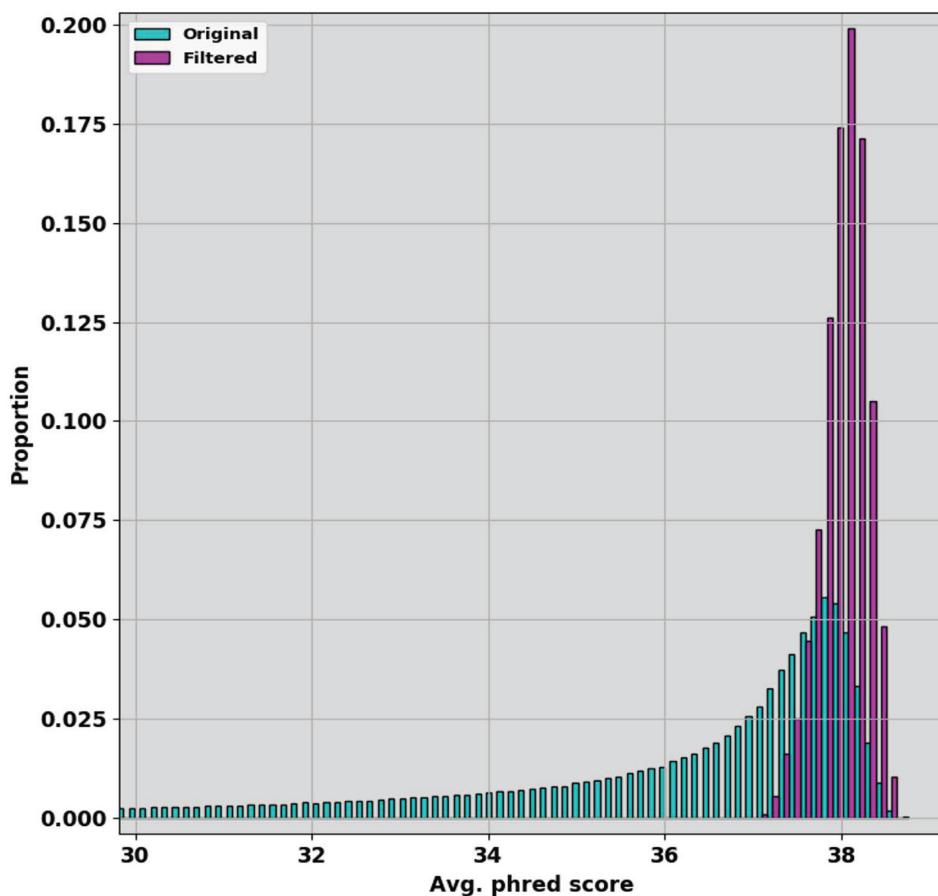


Figure 2. Filtering for high quality reads. Distributions of read qualities for original raw reads (cyan) and reads after strict filtering (7.1% of the original reads, magenta) show the dramatic increase in quality consistency in the reads provided to the assembler.

Sanger sequencing reads. For these previously sequence-verified samples, our pipeline correctly assembled all 72 plasmids, and succeeded in identifying all of the known SNPs that were present in 36 of these 72 samples.

The remaining 24 samples sequenced contained plasmids that were constructed from non-sequence-verified parts by Gibson assembly (20). Among the 24 samples, our pipeline detected 15 that contained unknown SNPs and/or INDELS. We validated two of the 15 variants by Sanger sequencing and confirmed that all 15 agreed with the assemblies provided by seqWell.

In conclusion, our workflow resulted in the correct assembly of all 96 samples, including the identification of several known and unknown variants, even when overall read quality was very low.

Assessing assembly quality without a reference and detecting contamination

In the previous section, we described developing and validating our pipeline by comparing the resulting assemblies to a known reference sequence. However, we built our pipeline using *de novo* assembly, so that it can also be used to assemble plasmid sequences when a reference is not available. One reference-agnostic way to assess an assembly's potential fidelity is to map the input reads to the assembly. To de-

termine if assembly quality could be reliably assessed in this manner, we challenged our assemblies by artificially spiking in 'contaminant' reads prior to assembly. We then compared the percent match between the assembled contig and the input reads for contaminated and uncontaminated reads to establish a baseline percent match for assessing assembly quality.

We ran this procedure using different proportions of contaminant reads, from 50% down to 5%, and evaluated 500 assemblies at each proportion. Each contamination simulation experiment yielded four possible outcomes (Figure 4A): 1) the assembly matches the primary reference, 2) the assembly matches the contaminant reference, 3) the assembly matches neither or 4) the pipeline yields a fragmented assembly containing more than one contig. By varying the relative proportions of primary and contaminant reads in the combined data sets, we assess the impact of SNPs, INDELS and other forms of contamination and evaluate the assembly pipeline's sensitivity to contaminated data as shown in Table 1.

In Table 1, we report the average proportions of correct, fragmented and mismatching assemblies. Not surprisingly, as the proportion of contaminant reads increases, the proportion of assemblies correctly matching the primary reference sequences decreases. When the difference between the contaminant differs from the primary plasmid by one to

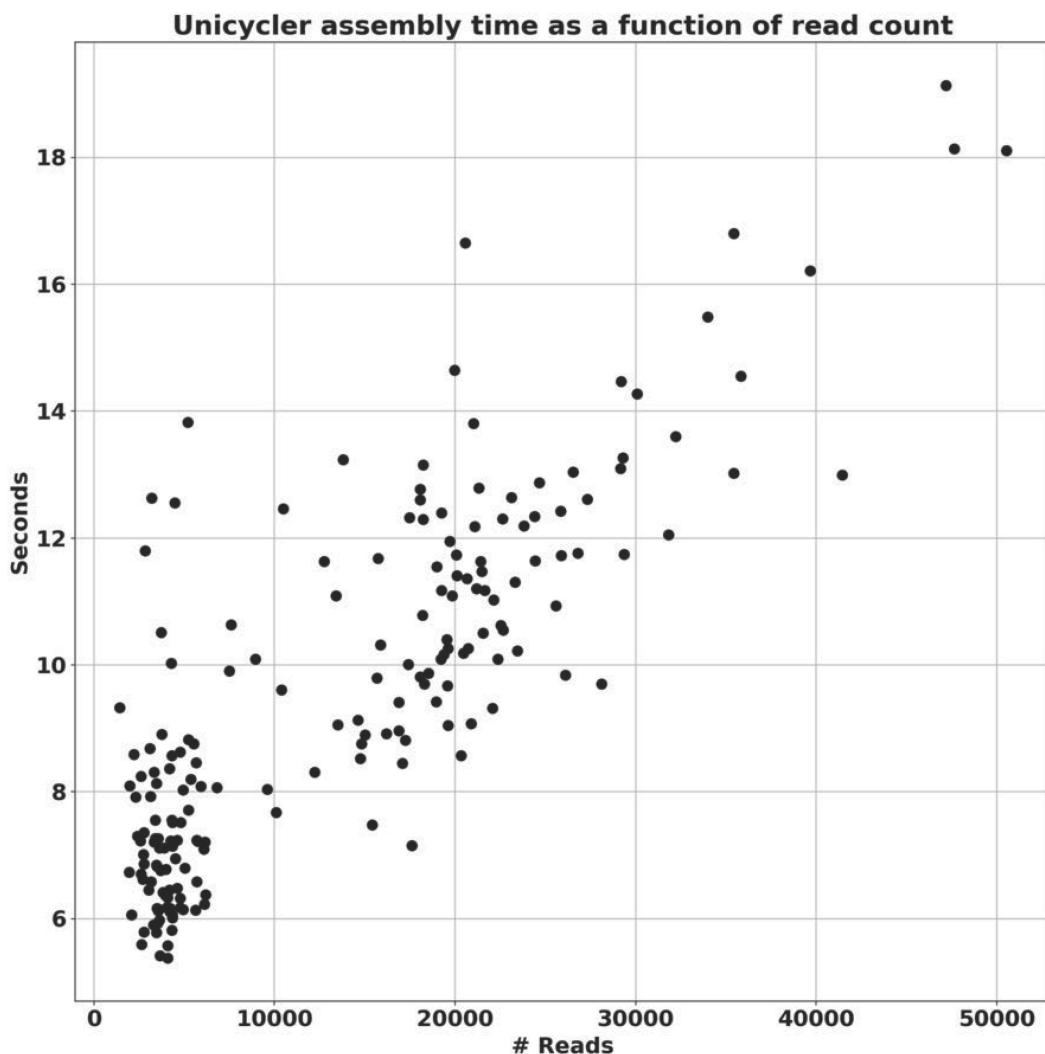


Figure 3. Unicycler assembly time as a function of read count. Experiments using up to 50 000 input reads show that Unicycler runtime increases approximately linearly with the number of input reads.

four SNPs, fully 90% of assemblies match the correct primary sequence at 35% contamination or less (Table 1, Figure 4B).

When the contaminating sequence contains an INDEL, we observe a sizeable increase in the number of fragmented assemblies (Table 1). It is worth noting that in some cases the fragmented assembly consists of just two sequences: the correct sequence for the smaller of the primary and contaminant source sequences, and a second sequence that matches the insertion. When the primary and contaminant sequences differed by a 608 bp INDEL, the pipeline failed to yield correct assemblies consistently when over 20% of reads were contaminating (Table 1, Figure 4C).

When the primary and contaminant populations come from plasmids with very little sequence similarity, the resulting assemblies are more severely affected than with SNPs or INDELs. As with INDELs, the high assembly fidelity was achieved only when contamination levels were below 20% of total reads (Figure 4D). At that ratio, the assembly may

be correct up to 86% of the time (Table 1). Similarly, we find that fragmented assemblies nearly disappear.

These results suggest that our pipeline produces robust results when contaminating reads are <20–35%, depending on the type of contaminant and its similarity to the primary sequence. However, even when the contamination is limited enough to yield exactly one contig, it is still important for the user to know whether any contamination is present in the sample. To predict contamination likelihood in a reference agnostic way, we mapped the input reads back to the assembly for contaminated and uncontaminated samples. We predicted that, because contaminating reads will not match, this approach would provide a potential measure of contamination likelihood.

We first note that, even after filtering, some reads may not map exactly even to a perfect assembly derived from an uncontaminated sample (33–35). However, the mismatch percentages were, as we predicted, higher for read data with contamination (Figure 5A, above). For the 96 assemblies

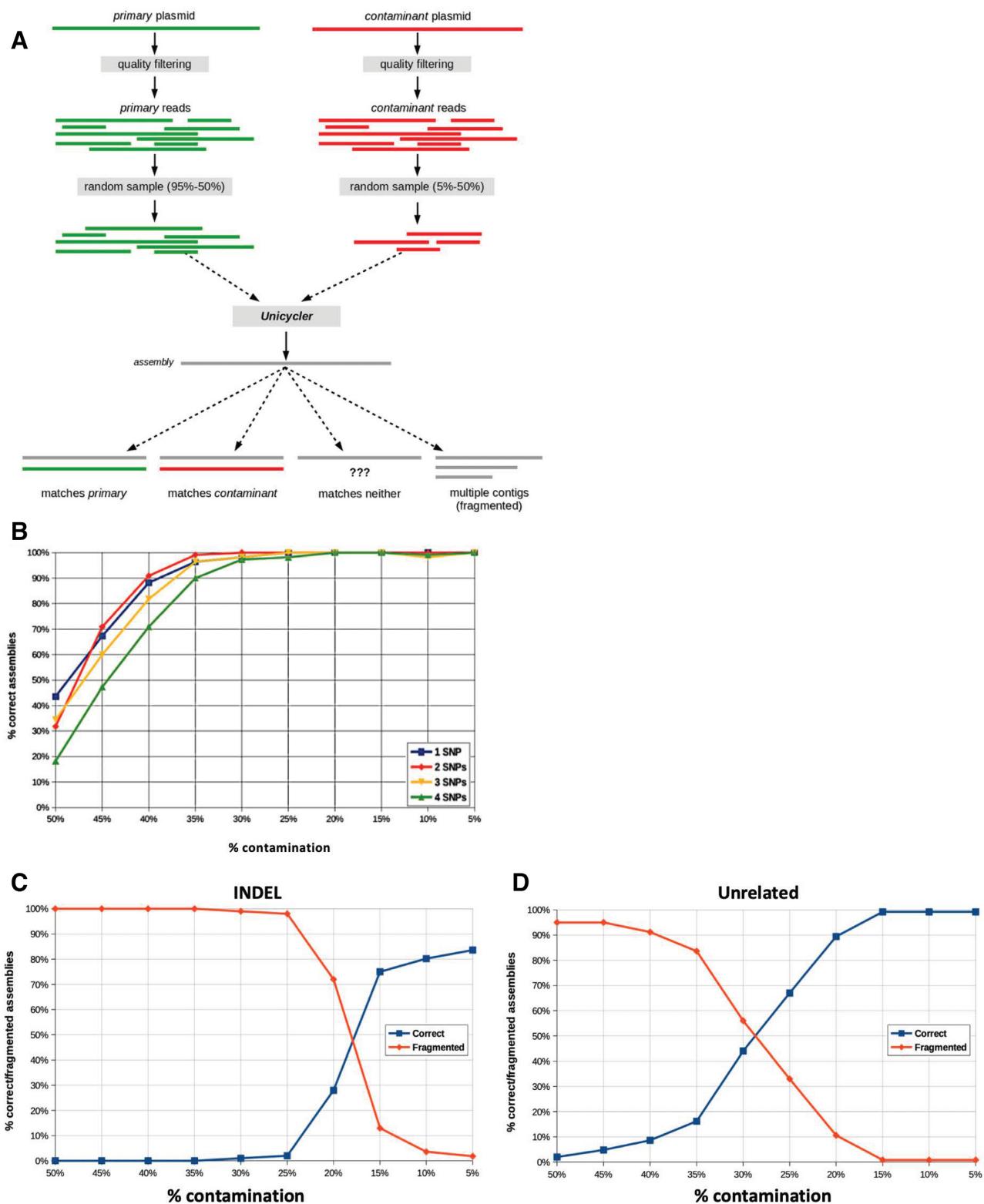


Figure 4. Determining the reliability of assemblies in the presence of contaminating reads. (A) Workflow for creating a set of reads ('primary') with a specific proportion of contaminating reads to simulate how the pipeline is affected by contamination. Datasets of reads with 5–50% contaminating reads were created, assembled, and verified against the reference sequence. (B) Percentage of correct assemblies when contaminating reads contained 1–4 SNPs. Percentage of correct and fragmented assemblies when (C) a 608 bp INDEL was introduced into contaminating reads, or (D) contaminating read sequences are completely unrelated to primary sequence.

Table 1. Contamination test summary

Variant type	%Contamination	%Correct	%Fragmented	%Mismatches
None	0	100%	0%	12.29%
1 SNP	20%	100%	0%	13.38%
	35%	96%	0%	14.41%
	50%	44%	0%	14.68%
	20%	100%	0%	14.35%
2 SNPs	35%	96%	0%	14.82%
	50%	32%	0%	15.96%
	20%	100%	0%	14.04%
3 SNPs	35%	96%	0%	15.21%
	50%	35%	0%	16.58%
	20%	100%	0%	14.85%
4 SNPs	35%	90%	0%	16.83%
	50%	18%	0%	17.99%
	10%	99%	1%	13.49%
608 bp indel	20%	89%	11%	14.78%
	35%	16%	84%	17.28%
	50%	2%	95%	21.00%
	10%	86%	4%	18.24%
Unrelated	20%	28%	72%	23.96%
	35%	0%	100%	NA
	50%	0%	100%	NA

Each row of this table represents the average result from a single simulation experiment (repeated 500 times). Each experiment differed by the amount and type of contamination simulated. Variant Type denotes by what metric the ‘contaminating’ sequence (reads from a different sample that were artificially spiked) differs from the ‘primary’ sequence (the reference sample). %Contamination refers to the percentage of the read library derived from contaminating reads. %Correct denotes the percentage of correct assemblies yielded. %Fragmented denotes the percentage of fragmented assemblies yielded. %Mismatches denotes the percentage of assemblies that did not exactly match the reference.

described above, we find that on average 11.7% of filtered input reads fail to map exactly to their associated assembly (Figure 5A, above). With contaminated data we find that overall 15.9% of input reads fail to map to an assembly. Interestingly, without the filtering step, the same percent of contaminated and uncontaminated reads mapped to the assembly (Figure 5A), suggesting that a filtering step is needed to detect contamination. Further, we find that the percentage of misaligned reads is highly predictive of contaminated data, even when the contamination is relatively subtle, as with SNPs (AUC 0.86–0.96, Figure 5B). We use this metric in our pipeline as a measure of the quality of the assembly and the likelihood of contamination.

Analysis of a library of 96 plasmids

The motivation for developing this *de novo* assembly pipeline was the implementation of new a method to verify the authenticity and integrity of plasmids (36,37). In this situation, it is not possible to assume the plasmid identity and therefore *de novo* assembly is a requirement. Supplementary Tables S1 to S3 summarize the results obtained on the data produced as described in the Methods section.

We sequenced 24 clones of a control plasmid with a size of 2521 bp (Supplementary Table S1). All clones led to a complete and perfect assembly but the probability of contamination varied from 1% to 93%. We observed similar results when sequencing 48 plasmids with a sequence length of 3129 bp (Supplementary Table S2). These plasmids were

provided as sequence-verified plasmids by a gene synthesis company. They had five different inserts different from each other by a single point mutation. Finally, we sequenced 24 plasmids that had been produced by assembling synthetic DNA fragments (Supplementary Table S3). In this case, it was expected that the plasmids would differ from each other as the sequence of the synthetic DNA fragments had not been verified prior to their assembly in a complete plasmid. The outcome of the sequencing read assembly are consistent with this assumption. In one case, 4 out 12 plasmids led to an assembly that perfectly matched the expected plasmid sequence. In the other case, 5 of the 12 clones led to a perfect match.

Analysis of additional data sets

We also assessed the tool performance on four additional datasets collected by other investigators. Three of these also provided reference sequences we could use to assess the fidelity of our assemblies. The remaining data set had no reference sequences, so we relied entirely on the pipeline’s contamination estimates to assess the likely quality of complete assemblies.

The first data set came from a recent study by Chiniquy *et al.* (38) that provided sets of 24 replicates from each of four plasmids identified as pXMJ19, pms6126, pGEN-292 and pskb3-CopR1598 (Supplementary Tables S4–S7). The plasmid lengths ranged from 4 to 10.3 kb. For pXMJ19 and pms6126, the pipeline generated a high-quality assembly for every replicate. With one exception, these assemblies were consistently within 2% of the correct sequence length, with sequence identity between 99.3% and 100% for those regions covered. In addition, 18 of the 24 replicates of pGEN-292 led to fragmented assemblies. The six successful assemblies resulted in partial coverages of the plasmid sequence. We noticed that this plasmid sequence contains several repeats up to 38 bases long that may impede assembly. Finally, the assembly of pskb3-CopR1598 data resulted in a number of fragmented assemblies and poor sequence coverage.

Our second data set was provided by the Jonathan Schlebach lab at Indiana University (personal communication). It consisted of 24 closely related plasmids that differed only in adjacent nucleotides. Since, the reference sequence of the individual plasmids was not available, it was not possible to determine the accuracy of the assembly over the entire sequence. However, the pipeline generated high-quality assemblies for all 24 plasmids, consistently within 0.1% of the given sequence length, and having sequence identity over 99.6% over the regions covered (see Supplementary Table S8).

The third data set included four previously published plasmids (39). Interestingly, these plasmids are part of reverse genetics systems to produce plant viruses from DNA molecules. Their sizes range from 7.6 to 14.4 kb. Their sequences are more complex than simple protein expression vectors as they include entire viral genomes. This yielded four distinct vectors, each of which had many more reads of higher quality per plasmid than those used to develop the pipeline. Using default parameters, our pipeline succeeded in creating an assembly for just one of these (Supplementary Table S9). Interestingly, the successful assembly

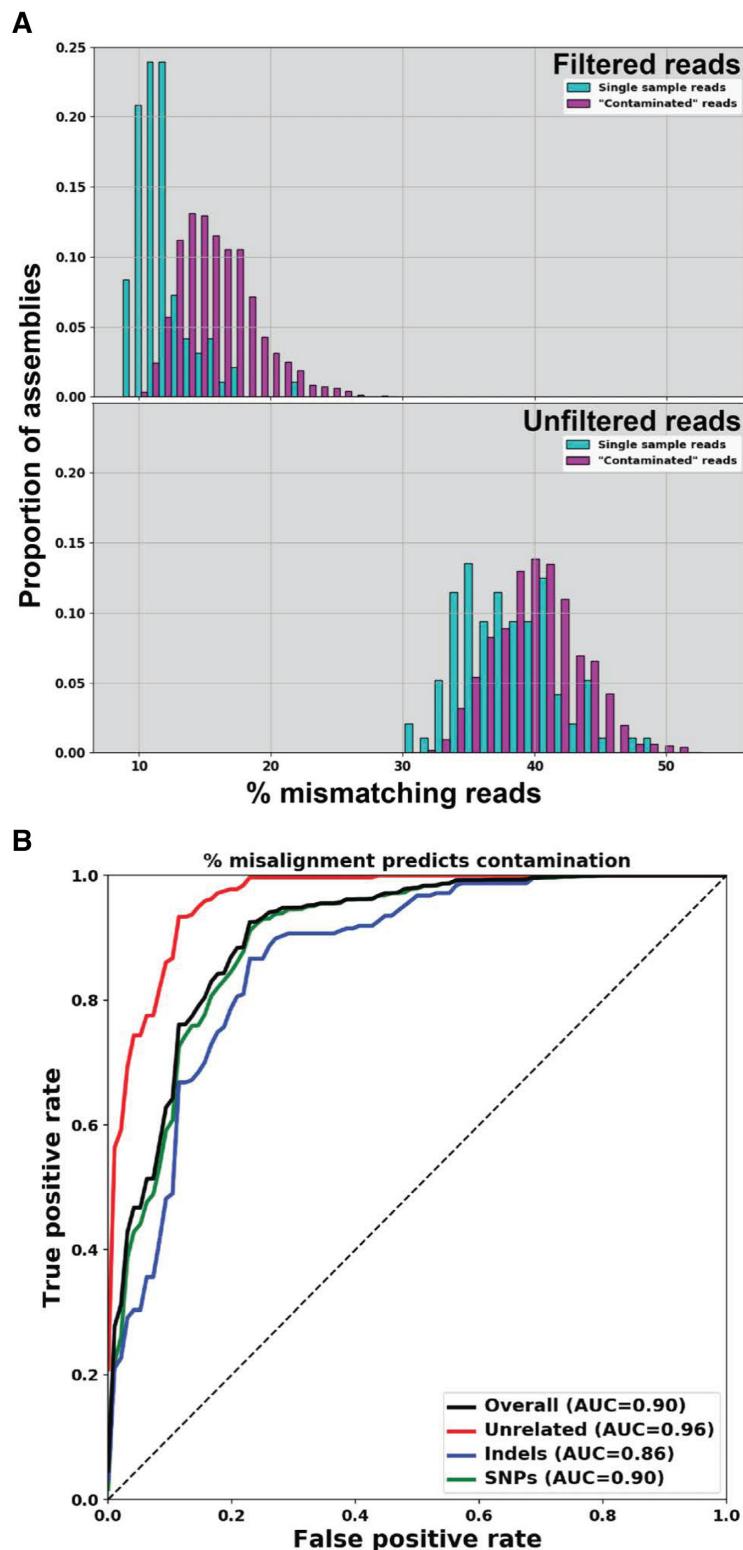


Figure 5. Misalignment of reads to the assembly can be used as a marker of contamination. (A) Input reads aligned back to their associated *de novo* assembly for contaminated and clean samples for filtered (top) and unfiltered (bottom) reads. (B) Receiver operating characteristic (ROC) curve demonstrating that, using the percent of mismatching reads alone, we can predict contamination with high accuracy. Here, true positives are assemblies correctly predicted as contaminated, and false positives are clean assemblies incorrectly predicted as contaminated. AUC: area under the curve.

was for the longest sequence (14 223 bases), and the only one that contains no repeat sequences of any length. The remaining three sequences contain repeats that are longer than the longest input reads. Given the high quality of the input reads, >60% of read pairs remained even after filtering. As we noted earlier, assemblies may suffer from a deficit of read coverage, which could explain why the longest sequence suffered the least degradation.

Finally, we tested a set of 91 plasmids provided by SeqWell. These plasmids did not come with reference sequences. For these tests, we thus relied on contamination predictions to assess the quality of our assemblies (Supplementary Tables S10 and S11). Out of the 91 samples, the pipeline yielded complete assemblies for 60 (66%) of them. Of these, 11 yielded strong evidence of contamination (estimated probabilities ranging from 80% to 100%), while the results obtained with three plasmids suggested the possibility of contaminations (estimated probabilities 51–66%). The remaining 31 assemblies were fragmented.

DISCUSSION

We have demonstrated the feasibility of rapidly producing highly accurate plasmid assemblies from short read sequencing data by *de novo* assembly. Our publicly available workflow relies on Unicycler (27), an open source tool that was developed for assembly of circular genomes. We identified Unicycler parameters and determined read filtering thresholds that are optimal for plasmid assemblies.

We incorporated into our pipeline a novel method for detecting contamination by analyzing the percent of input reads that map exactly to locations in the final assembly. In general, our pipeline can generate accurate assemblies with fewer than 1000 reads, regardless of quality; but if we focus on high-quality reads, we can begin to detect contamination. We find that only a small fraction of high-quality reads fails to map to an assembly, while increased levels of non-mapping reads correlate with contamination. We find that the resulting percent-matching metric can provide coarse information about whether the input data are likely to be contaminated. Despite our method's simplicity, our experiments show that it is surprisingly adept at discriminating between 'clean' and contaminated assemblies, demonstrating sensitivity even to contaminating plasmids with single-nucleotide mutations at relatively low levels.

The bioinformatics pipeline was developed using a dataset that included 96 clones corresponding to eight different plasmids. We assessed the tool performance by analyzing four additional datasets collected by other investigators. Together, these additional data sets included reads for 215 clones corresponding to 123 different plasmids representing a broad range of plasmid size and sequence complexity. These additional results suggest that the pipeline can yield high-quality assemblies in a variety of contexts. Importantly, these additional datasets were produced using different library preparation protocols. A number of factors such as sequence complexity, sequence length, read numbers, and read quality seem to influence the success of the assembly process. In many cases, the pipeline default parameters work well. In some instances, some minor parameter tuning resulted in more reliable assemblies. As we gather a

greater variety of data, we anticipate refining this pipeline to provide accurate assemblies for a broader spectrum of applications and sequencing data. Complex sequences with long repeats are expected to remain beyond the reach of a tool relying exclusively on short reads. We envision generalizing the pipeline to support a hybrid assembly strategy to combine the benefits of short and long reads (40,41).

The metric we use to detect contamination in single-contig assemblies is a rudimentary first step. We envision future improvements in several directions for this work. First, our rudimentary metric could be replaced with a more sophisticated machine learning approach that incorporates features of mapped and unmapped reads, such as the number of mismatching bases or their associated quality scores. Second, either this machine learning model or our realignment method could aid in selecting a contig that matches the correct plasmid sequence in cases where Unicycler produces more than one contig. Third, we wish to find ways for our pipeline to provide greater detail regarding the likely nature of any contamination it detects. We believe our likelihood score will be useful but reporting the nature of any contamination detected would make it easier for users to assess their samples more fully.

These results have important implications for basic and applied research. Contamination and sample misidentification are rampant problems in mammalian cell research (42–45). While comparably less attention has been paid to the potential for similar problems in model organisms and *in vitro*, it has been demonstrated that many plasmids currently in circulation do not match their supposed references (46,47). To ameliorate this problem, researchers can sequence their plasmids relatively inexpensively by coupling them to other sequencing runs at a local core facility and then using our pipeline to ensure that plasmid sequences are correct before running any experiments. SeqWell has made the chemistry they use for plasmid sequencing available as library prep kits, which should facilitate the process of plasmid verification by Illumina sequencing. This small up-front cost could save considerable time and resources, as well as limiting reproducibility issues.

The ability to rapidly and automatically validate plasmids by Illumina sequencing will be especially applicable to large-scale plasmid production projects associated with synthetic biology (48). The rapidly decreasing cost of DNA synthesis and sophisticated computer-aided design tools (49–51) have facilitated disciplined factorial experiments involving large libraries of plasmids with various genetic parts (52) for applications such as improved genetic circuit design (53), multiplexed CRISPR-Cas (54) or metabolic engineering (55). These workflows, which are often automated (56,57), necessitate an automated sequence verification pipeline like the one we have described.

Further, this pipeline will be an asset to another large-scale application: high-throughput plasmid libraries. Increasingly, high-throughput screens are being performed from libraries of hundreds to thousands of genes encoded on plasmids or lentiviral 'transfer plasmid' genomes. Panels of this sort are so robust that they have made it possible to specify each gene involved in a molecular pathway (58–60) in a single screen, work that would have previously taken years to decades. Notably, plasmid libraries have been

shown to perform sgRNA CRISPR/Cas9 knock-out (58) or shRNA knock-down (61).

Though these technologies were introduced in 2014 and 2011, respectively, they have been slow to be adopted, in part we believe from the difficulty of generating plasmid libraries. Though, the bottleneck of these types of screens is still generating a plasmid library, which includes the essential step of verifying plasmid sequences. The plasmid verification pipeline described here can solve a critical need to make generating high throughput screening libraries faster, cheaper, and therefore more feasible.

This work is also relevant for data security (62,63). There is increasing interest in using DNA as a medium for information storage (64,65). Recently, researchers began experimenting with storing information in plasmids, as they can be rapidly replicated (66,67). A technology for encrypting and securing digital information using cybersecurity techniques has also been developed for plasmids (36,37). Maintaining digital references for these sequences would defeat the purpose of storing the information in DNA, so these efforts will be greatly facilitated by the rapid *de novo* assembly pipeline we've described.

The described pipeline was built using open source tools and is available to the community through the supplement.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

NSF [1934573, 1832320]. The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The open access publication charge for this paper has been waived by Oxford University Press – NAR Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.

Conflict of interest statement. J.P. holds an equity stake in GenoFAB Inc., a company that may benefit or may be perceived to benefit from the publication of this article. C.C. and M.R. are employees of GenoFAB, Inc.

REFERENCES

1. Peccoud,J. (2016) Synthetic Biology: fostering the cyber-biological revolution. *Synth. Biol.*, **1**, ysw001.
2. Chao,R., Mishra,S., Si,T. and Zhao,H. (2017) Engineering biological systems using automated biofoundries. *Metab. Eng.*, **42**, 98–108.
3. Hillson,N., Caddick,M., Cai,Y., Carrasco,J.A., Chang,M.W., Curach,N.C., Bell,D.J., Le Feuvre,R., Friedman,D.C., Fu,X. *et al.* (2019) Building a global alliance of biofoundries. *Nat. Commun.*, **10**, 2040.
4. Ma,X., Liang,H., Cui,X., Liu,Y., Lu,H., Ning,W., Poon,N.Y., Ho,B. and Zhou,K. (2019) A standard for near-scarless plasmid construction using reusable DNA parts. *Nat. Commun.*, **10**, 3294.
5. Guo,X., Chavez,A., Tung,A., Chan,Y., Kaas,C., Yin,Y., Cecchi,R., Garnier,S.L., Kelsic,E.D., Schubert,M. *et al.* (2018) High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR–Cas9 in yeast. *Nat. Biotechnol.*, **36**, 540.
6. Pollak,B., Matute,T., Nuñez,I., Cerdá,A., Lopez,C., Vargas,V., Kan,A., Bielinski,V., von Dassow,P. and Dupont,C.L. (2020) Universal loop assembly: open, efficient and cross-kingdom DNA fabrication. *Synth. Biol.*, **5**, ysaa001.
7. Smanski,M.J., Bhatia,S., Zhao,D., Park,Y., Woodruff,L.B., Giannoukos,G., Ciulla,D., Busby,M., Calderon,J. and Nicol,R. (2014) Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol.*, **32**, 1241.
8. Cress,B.F., Toparlak,O.D., Guleria,S., Lebovich,M., Stieglitz,J.T., Englaender,J.A., Jones,J.A., Linhardt,R.J. and Koffas,M.A. (2015) CRISPRPathBrick: modular combinatorial assembly of type II-A CRISPR arrays for dCas9-mediated multiplex transcriptional repression in *E. coli*. *AcS Synth Biol.*, **4**, 987–1000.
9. Iverson,S.V., Haddock,T.L., Beal,J. and Densmore,D.M. (2016) CIDAR MoClo: Improved MoClo assembly standard and new *E. coli* part library enable rapid combinatorial design for synthetic and traditional biology. *AcS Synth Biol.*, **5**, 99–103.
10. Kuivanen,J., Holmstrom,S., Lehtinen,B., Penttila,M. and Jantti,J. (2018) A high-throughput workflow for CRISPR/Cas9 mediated combinatorial promoter replacements and phenotype characterization in yeast. *Biotechnol. J.*, **13**, e1700593.
11. Czar,M.J., Anderson,J.C., Bader,J.S. and Peccoud,J. (2009) Gene synthesis demystified. *Trends Biotechnol.*, **27**, 63–72.
12. Marchand,J.A. and Peccoud,J. (2012) Building block synthesis using the polymerase chain assembly method. *Methods Mol. Biol.*, **852**, 3–10.
13. Peccoud,J. (2012) In: *Gene Synthesis: Methods and Protocols*. Springer, NY.
14. Nelms,B.L. and Labosky,P.A. (2011) A predicted hairpin cluster correlates with barriers to PCR, sequencing and possibly BAC recombining. *Sci. Rep.*, **1**, 106.
15. Wilson,M.L., Cai,Y., Hanlon,R., Taylor,S., Chevreux,B., Setubal,J.C., Tyler,B.M. and Peccoud,J. (2013) Sequence verification of synthetic DNA by assembly of sequencing reads. *Nucleic Acids Res.*, **41**, e25.
16. Currin,A., Swainston,N., Dunstan,M.S., Jervis,A.J., Mulherin,P., Robinson,C.J., Taylor,S., Carbonell,P., Hollywood,K.A. and Yan,C. (2019) Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. *Synth. Biol.*, **4**, ysz025.
17. Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.
18. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135.
19. Gallegos,J.E., Kar,D.M., Ray,I., Ray,I. and Peccoud,J. (2019) Securing the exchange of synthetic genetic constructs using digital signatures. *bioRxiv* doi: <https://doi.org/10.1101/750927>, 26 September 2019, preprint: not peer reviewed.
20. Gibson,D.G., Young,L., Chuang,R.Y., Venter,J.C., Hutchison,C.A. 3rd and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–345.
21. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
22. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
23. Chen,S.F., Zhou,Y.Q., Chen,Y.R. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, 884–890.
24. Fang,H., Wu,Y., Narzisi,G., ORawe,J.A., Barrón,L.T.J., Rosenbaum,J., Ronenius,M., Iossifov,I., Schatz,M.C. and Lyon,G.J. (2014) Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med.*, **6**, 89.
25. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L. and Bignell,H.R. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
26. Lonardi,S., Mirebrahim,H., Wanamaker,S., Alpert,M., Ciardo,G., Duma,D. and Close,T.J. (2015) When less is more: 'slicing' sequencing data improves read decoding accuracy and *de novo* assembly quality. *Bioinformatics*, **31**, 2972–2980.
27. Wick,R.R., Judd,L.M., Gorrie,C.L. and Holt,K.E. (2017) Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.*, **13**, e1005595.
28. Chevreux,B., Pfisterer,T., Drescher,B., Driesel,A.J., Muller,W.E.G., Wetter,T. and Suhai,S. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.*, **14**, 1147–1159.

29. Zerbino,D.R. and Birney,E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.

30. Antipov,D., Hartwick,N., Shen,M., Raiko,M., Lapidus,A. and Pevzner,P.A. (2016) plasmidSPADEs: assembling plasmids from whole genome sequencing data. *Bioinformatics*, **32**, 3380–3387.

31. Hunt,M., De Silva,N., Otto,T.D., Parkhill,J., Keane,J.A. and Harris,S.R. (2015) Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.*, **16**, 294.

32. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S. and Prjibelski,A.D. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

33. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., Del Angel,G., Rivas,M.A. and Hanna,M. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491.

34. Manley,L.J., Ma,D. and Levine,S.S. (2016) Monitoring error rates in Illumina sequencing. *J. Biomol. Tech.*, **27**, 125–128.

35. Schirmer,M., D'Amore,R., Ijaz,U.Z., Hall,N. and Quince,C. (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, **17**, 125.

36. Kar,D.M., Ray,I., Gallegos,J., Peccoud,J. and Ray,I. (2020) Synthesizing DNA molecules with identity-based digital signatures to prevent malicious tampering and enabling source attribution. *J. Comput. Secur.*, **18**, 437–467.

37. Kar,D.M., Ray,I., Gallegos,J. and Peccoud,J. (2018) Digital signatures to ensure the authenticity and integrity of synthetic DNA molecules. *Nspw '18: Proceedings of the New Security Paradigms Workshop*, 110–122.

38. Chiniyup,J., Garber,M.E., Mukhopadhyay,A. and Hillson,N.J. (2020) Fluorescent amplification for next generation sequencing (FA-NGS) library preparation. *BMC Genomics*, **21**, 85.

39. Pasin,F., Tseng,X.-A., Bedoya,L.C., Heydarnejad,J., Deng,T.-C., Garcia,J.A. and Chen,Y.-R. (2018) Streamlined generation of plant virus infectious clones using the pLX mini binary vectors. *J. Virol. Methods*, **262**, 48–55.

40. Utturkar,S.M., Klingeman,D.M., Land,M.L., Schadt,C.W., Doktycz,M.J., Pelletier,D.A. and Brown,S.D. (2014) Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, **30**, 2709–2716.

41. Antipov,D., Korobeynikov,A., McLean,J.S. and Pevzner,P.A. (2016) hybridSPADEs: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, **32**, 1009–1015.

42. Freedman,L.P., Gibson,M.C., Ethier,S.P., Soule,H.R., Neve,R.M. and Reid,Y.A. (2015) Reproducibility: changing the policies and culture of cell line authentication. *Nat. Methods*, **12**, 493.

43. Lacroix,M. (2008) Persistent use of “false” cell lines. *Int. J. Cancer*, **122**, 1–4.

44. Capes-Davis,A., Theodosopoulos,G., Atkin,I., Drexler,H.G., Kohara,A., MacLeod,R.A., Masters,J.R., Nakamura,Y., Reid,Y.A. and Reddel,R.R. (2010) Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer*, **127**, 1–8.

45. Lorsch,J.R., Collins,F.S. and Lippincott-Schwartz,J. (2014) Fixing problems with cell lines. *Science*, **346**, 1452–1453.

46. 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. *et al.* (2008) Targeted development of registries of biological parts. *PLoS One*, **3**, e2671.

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

48. Casini,A., Storch,M., Baldwin,G.S. and Ellis,T. (2015) Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.*, **16**, 568–576.

49. Czar,M.J., Cai,Y. and Peccoud,J. (2009) Writing DNA with GenoCAD. *Nucleic Acids Res.*, **37**, W40–W47.

50. Wilson,M.L., Hertzberg,R., Adam,L. and Peccoud,J. (2011) A step-by-step introduction to rule-based design of synthetic genetic constructs using GenoCAD. *Methods Enzymol.*, **498**, 173–188.

51. Wilson,M.L., Okumoto,S., Adam,L. and Peccoud,J. (2014) Development of a domain-specific genetic language to design Chlamydomonas reinhardtii expression vectors. *Bioinformatics*, **30**, 251–257.

52. Adames,N.R., Wilson,M.L., Fang,G., Lux,M.W., Glick,B.S. and Peccoud,J. (2015) GenoLIB: a database of biological parts derived from a library of common plasmid features. *Nucleic Acids Res.*, **43**, 4823–4832.

53. Nielsen,A.A., Der,B.S., Shin,J., Vaidyanathan,P., Paralanov,V., Strychalski,E.A., Ross,D., Densmore,D. and Voigt,C.A. (2016) Genetic circuit design automation. *Science*, **352**, aac7341.

54. McCarty,N.S., Graham,A.E., Studená,L. and Ledesma-Amaro,R. (2020) Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nat. Commun.*, **11**, 1281.

55. Garst,A.D., Bassalo,M.C., Pines,G., Lynch,S.A., Halweg-Edwards,A.L., Liu,R., Liang,L., Wang,Z., Zeitoun,R., Alexander,W.G. *et al.* (2016) Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechnol.*, **35**, 48.

56. Linshiz,G., Jensen,E., Stawski,N., Bi,C., Elsbree,N., Jiao,H., Kim,J., Mathies,R., Keasling,J.D. and Hillson,N.J. (2016) End-to-end automated microfluidic platform for synthetic biology: from design to functional analysis. *J. Biol. Eng.*, **10**, 3.

57. Carbonell,P., Jervis,A.J., Robinson,C.J., Yan,C., Dunstan,M., Swainston,N., Vinaixa,M., Hollywood,K.A., Currin,A., Rattray,N.J.W. *et al.* (2018) An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Commun. Biol.*, **1**, 66.

58. Zhou,Y., Zhu,S., Cai,C., Yuan,P., Li,C., Huang,Y. and Wei,W. (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, **509**, 487–491.

59. Jaitin,D.A., Weiner,A., Yofe,I., Lara-Astiaso,D., Keren-Shaul,H., David,E., Salame,T.M., Tanay,A., van Oudenaarden,A. and Amit,I. (2016) Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, **167**, 1883–1896.

60. Golden,R.J., Chen,B., Li,T., Braun,J., Manjunath,H., Chen,X., Wu,J., Schmid,V., Chang,T.-C. and Kopp,F. (2017) An argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature*, **542**, 197–202.

61. Sims,D., Mendes-Pereira,A.M., Frankum,J., Burgess,D., Cerone,M.-A., Lombardelli,C., Mitsopoulos,C., Hakas,J., Murugaesu,N. and Isacke,C.M. (2011) High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol.*, **12**, R104.

62. Murch,R.S., So,W.K., Buchholz,W.G., Raman,S. and Peccoud,J. (2018) Cyberbiosecurity: an emerging new discipline to help safeguard the bioeconomy. *Front. Bioeng. Biotech.*, **6**, 39.

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

64. Church,G.M., Gao,Y. and Kosuri,S. (2012) Next-generation digital information storage in DNA. *Science*, 1226355.

65. Erlich,Y. and Zielinski,D. (2017) DNA fountain enables a robust and efficient storage architecture. *Science*, **355**, 950–954.

66. Nguyen,H.H., Park,J., Park,S.J., Lee,C.S., Hwang,S., Shin,Y.B., Ha,T.H. and Kim,M. (2018) Long-term stability and integrity of plasmid-based DNA data storage. *Polymers-Basel*, **10**, 28.

67. Akhmetov,A., Ellington,A.D. and Marcotte,E.M. (2018) A highly parallel strategy for storage of digital information in living cells. *BMC Biotech.*, **18**, 64.