On-Ramp: a tool for rapid, multiplexed validation of plasmids using nanopore sequencing

Camille Mumm[1,3], Melissa L. Drexel[1,3], Torrin L. McDonald[1], Adam G. Diehl[2], Jessica A. Switzenberg[2], Alan P. Boyle[1,2,✉]

1. Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA 48109
2. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA 48109
3. These authors contributed equally to this work.

✉ Correspondence: apboyle@umich.edu

Abstract

The ability to generate vectors made of recombinant DNA has facilitated many achievements in molecular biology. As these vectors are assembled from many different parts and the enzymatic and bacterial processes used to create them can introduce errors, validation of the final product is an essential part of plasmid assembly. Sanger sequencing is the primary validation method available at a base-pair level, however it has many limitations. These include its inability to sequence through complex secondary structure, read length, and cost when validating the full sequence of plasmids or large numbers of plasmids. There is a need for a sequence validation method that can rapidly and affordably sequence the entirety of many plasmids simultaneously, using tools that are accessible to molecular cloning labs, without the need for the time and cost associated with next-generation sequencing. On-Ramp (Oxford-nanopore based Rapid Analysis of Multiplexed Plasmids) is a nanopore-based sequencing based approach that addresses this need by leveraging novel long-read technology. On-Ramp combines preparation protocols designed specifically for multi-plasmid pools and an analysis pipeline modified for accurate alignment of nanopore sequencing reads resulting from these plasmid pool methods. We demonstrate that through On-Ramp, large numbers of dissimilar or highly similar plasmids can be sequenced simultaneously and consensus sequences generated that capture single base-pair mutations, insertions or deletions, and highly repetitive region sequences can be resolved. Our tool allows for sequencing of multiple plasmids in their entirety at one-third the cost of Sanger sequencing, and through the use of our On-Ramp webapp, labs can obtain sequence results rapidly without the need for bioinformatic expertise. On-Ramp is available at http://OnRamp.BoyleLab.org.

Introduction

Cloning of recombinant DNA into plasmid vectors is a fundamental tool of molecular biology and central to many discoveries in genetics for decades, including the first sequencing of the human genome[1]. It continues to underpin modern-day research in genomics, protein expression and purification[2], transcriptional regulation[3], and gene therapies[4]. Consequently, the development of tools to improve efficiency, speed, and scale of recombinant plasmid assembly, multiplexing, and analysis is essential to advancing the pace of novel discovery across multiple fields. Developing tools for validation of cloned vectors is particularly crucial due to the error-prone nature of recombinant assembly. Examples of these errors include polymerases altering bases in a coding region that is subsequently amplified for cloning[5] or in off-target sites in the plasmid in the case of site-directed mutagenesis, and subtle structural rearrangements that can occur during bacterial transformation, which then propagate in downstream molecular plasmid manipulations**.** Illegitimate plasmid recombination during bacterial transformation is an established source of aberrant structures[6]. Incompletely circularized ligation products are substrates for bacterial recombination even in RecA- E.coli, often causing deletions while still retaining intact origins of replication and selection cassettes[6]. While advances have been made in a number of different enzyme-based techniques available for more efficient and flexible plasmid cloning *in vitro*, primary analysis methods of the resulting DNA construct are primarily limited to restriction digest analysis and subsequent Sanger sequencing for low-throughput analysis, or large-scale sequencing for high-throughput plasmid libraries.

While versatile and low-cost, restriction digestion coupled with gel electrophoresis is limited by its resolution to detecting plasmid sequence changes in the 100bp-1000bp range[7] and relies on unique restriction sites within the vector. This allows for simple verification of insert DNA presence or absence and to rule out highly aberrant structures, however it is unable to resolve single base-pair changes and small insertions or deletions. As a result, this approach is often only used as a first pass for low-throughput screening to prioritize clones for further analysis by Sanger sequencing. Sanger sequencing uses a PCR-amplification based approach to obtain base-pair resolution of DNA sequence in stretches of up to 1kb[8], to detect mutagenic changes and indels that occur during cloning and bacterial replication of recombinant plasmids.

Despite being an important tool for simple, low throughput sequence validations, Sanger sequencing also has a number of limitations. These include the need to synthesize target-specific primers for the sequenced region, inaccuracy in long mononucleotide stretches[9] or repetitive elements, and difficulty sequencing constructs with strong secondary structure. As Sanger sequencing has a maximum output of 1kb per sample, sequencing of multiple plasmid components or regions that exceed 1kb requires tiling over several runs, however this becomes expensive and laborious quickly when applied to multiple transformants. As a result, typical validation protocols using Sanger involve sequencing only a portion of the plasmid encoding transcription start sites or a modified functional region. However, this approach can miss errors on a plasmid that occur outside of the sequenced region. In plasmids where an inserted gene is larger than 1kb or contains multiple elements, only a portion of the functional regions are validated. As a result, recombinant manipulation- or bacterial recombination-induced errors in the vector that impact functional elements can propagate and impact downstream analysis. Even mutations in bacterial sequences, which are rarely checked by Sanger sequencing due to the assumption they are non-functional, can impact expression in reporter assays[10] and plasmid backbone context as well as element arrangement can impact expression[11,12]. This 1kb sequence limitation also makes Sanger unsuitable for cost-effective analysis of large-scale plasmid libraries.

High-throughput sequencing (HTS) addresses some Sanger sequencing limitations through facilitating sequencing of entire plasmids and scalability for large plasmid libraries, such as deep sequencing of multiplexed reporter assay (MPRA) libraries. However, cloning projects are rarely done at this scale. Thus cost, sample pooling coordination across research groups, indexing chemistry compatibility issues, and turnaround time are major barriers to widespread adoption of HTS outside of large-scale approaches. HTS, in the frequent case where plasmid libraries are identical outside of a unique modified region, cannot provide full-plasmid coverage to detect variation in individual plasmid backbones due to the inability to uniquely map short reads outside the unique region.

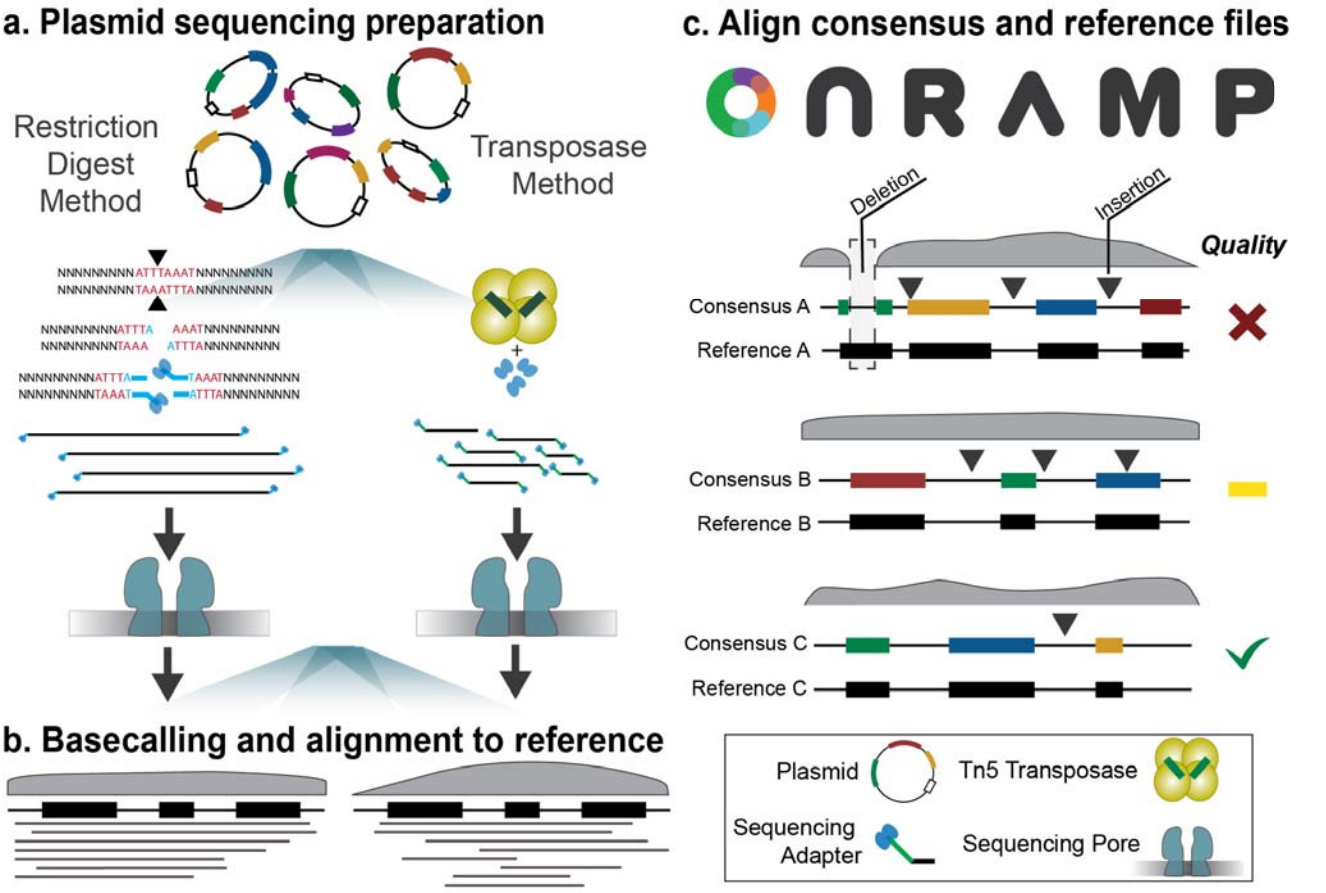


Figure 1. Plasmid library preparation and consensus sequence generation using On-Ramp.
a. Plasmids are prepared through either restriction digestion followed by adapter ligation, or transposase-based fragmentation and simultaneous adapter ligation. b. Prepared plasmids are run on a Nanopore sequencing platform and reads are basecalled then aligned to generate consensus files. c. Consensus files are aligned back to references to identify variation.

Here we present On-Ramp (Oxford Nanopore-based Rapid Analysis of Multiplexed Plasmids), a tool for plasmid validation that utilizes long-read sequencing technology, allowing for sequencing of multiple entire plasmids in a pooled format **(Figure 1)**. On-Ramp leverages the ONT long-read sequencing platform, which has recently gained significant momentum in genomics research due to its ability to resolve previously intractable complex structural variation[13]. ONT produces compact, benchtop sequencing platforms that generate single continuous reads on the order of megabases, and has been employed in whole genome sequencing (WGS) and targeted enrichment sequencing[14,15,16]. On-Ramp addresses the need for a medium-throughput approach that is rapid, simple, and more cost-effective than HTS, and provides flexible pooled plasmid sequencing to meet various sequencing needs within the lab. We created an all-in-one pooled plasmid sequencing protocol and analysis pipeline using ONT's Flongle platform and kits. Our preparation protocol is straightforward, rapid, and barcode-free, and the plasmids are deeply sequenced to base-pair resolution with single long reads spanning entire plasmids. Using existing long read processing software[17], consensus plasmid sequences are generated from sequencing reads, revealing detailed structural and single nucleotide variation. We provide three alternate methods developed for pooled plasmid preparation, allowing for flexibility in choosing either a transposase-based preparation, a restriction-based approach, or a modified restriction-based approach for barcode-free pooling of clonal plasmid copies. A streamlined web application is available (https://onramp.boylelab.org/) that allows labs to use On-Ramp's plasmid sequencing pipeline to analyze results, making interpretation accessible and simple.

Results

A. On-Ramp detects plasmid sequence variation

We first used simulated data to assess the ability of our On-Ramp analysis pipeline to accurately detect sequence variation within plasmids in a mixed plasmid pool. On-Ramp takes as input reads from Nanopore sequencing runs that have been basecalled, aligns these reads to user-provided reference files, and then compiles them into a single consensus sequence for each plasmid[17,18] to identify mutations, insertions or deletions against the reference files. We constructed simulated read data using NanoSim[19], a tool designed to simulate Nanopore reads of whole genome sequencing libraries. The library simulates plasmids prepared using the ONT Tn5 transposase with randomly distributed read start sites. This allowed for generation of a read library with a known plasmid of origin. Libraries were constructed for 30 dissimilar plasmids (average length 4.4kb) with known reference sequences. An average of 967 reads were generated for each plasmid and subsequently pooled. On-Ramp was first run in medaka mode on the 30 plasmid library to map pooled reads back to the plasmid references. Of this library's 29,984 reads, an average of 614 reads were assigned to each plasmid (of the 967 generated read pool). Alignments between the 30 references and the consensus sequences produced from this mapping contained three total gaps (2 missing single bases at the start of one consensus due to lack of depth, and a 1bp gap at a homopolymer run in another plasmid) with no mismatches detected in the alignments.

Given that this is a known plasmid set, we wanted to test the level of coverage needed to eliminate these gaps and produce a perfect consensus. To accomplish this, consensus sequences constructed using 500%, 100%, 50% and 10% of the 29,984 read pool (3085, 614, 306 and 60 reads per plasmid, respectively) were generated and gaps in the resulting consensus measured **(Figure 2a).** Decreasing coverage led to incomplete assembly, missing sequence at the consensus ends, and an increase in gaps at homopolymers using the simulated reads. However, many of these gaps may be specific to this Tn5 simulated data set, as restriction enzyme digested plasmid pools had fewer gaps at consensus ends as a majority of the reads span the full length of the plasmid.

Having constructed and characterized our simulated library, we next simulated the presence of sequence variation. Modified references were constructed containing a single insertion or deletion of 100bp, 10bp, or 1bp randomly generated throughout the plasmid. We produced six simulated reads sets using these modified references, each with an average of 970 reads and combined one of these insertion or deletion sets with the 30-plasmid pooled reads for each experiment. Each plasmid read pool, along with its corresponding references, was then analyzed using On-Ramp in medaka mode to generate polished consensus sequences. For each of the variant plasmid sets, we obtained a similar average number of plasmid-mapped simulated reads to the non-variant set. Using the pools containing reads generated from modified references, we accurately detected each one of the generated modifications at 970 reads per plasmid **(Fig 2b, 2c)**. Next, we tested the impact of read count per plasmid on On-Ramp's ability to detect each indel in the variant plasmid using simulations with 500 and 100 reads per plasmid **(Fig 2d, e).** Insertions and deletions of 100bp, 10bp and

1bp were all correctly identified even at 100 reads per plasmid. Read count did not impact ability to detect mutations, but rather affected whether additional variation occurred elsewhere in the consensus (points above the dotted lines in Fig 2d, e) as a result of lack of coverage, especially at map ends and homopolymers as discussed above.

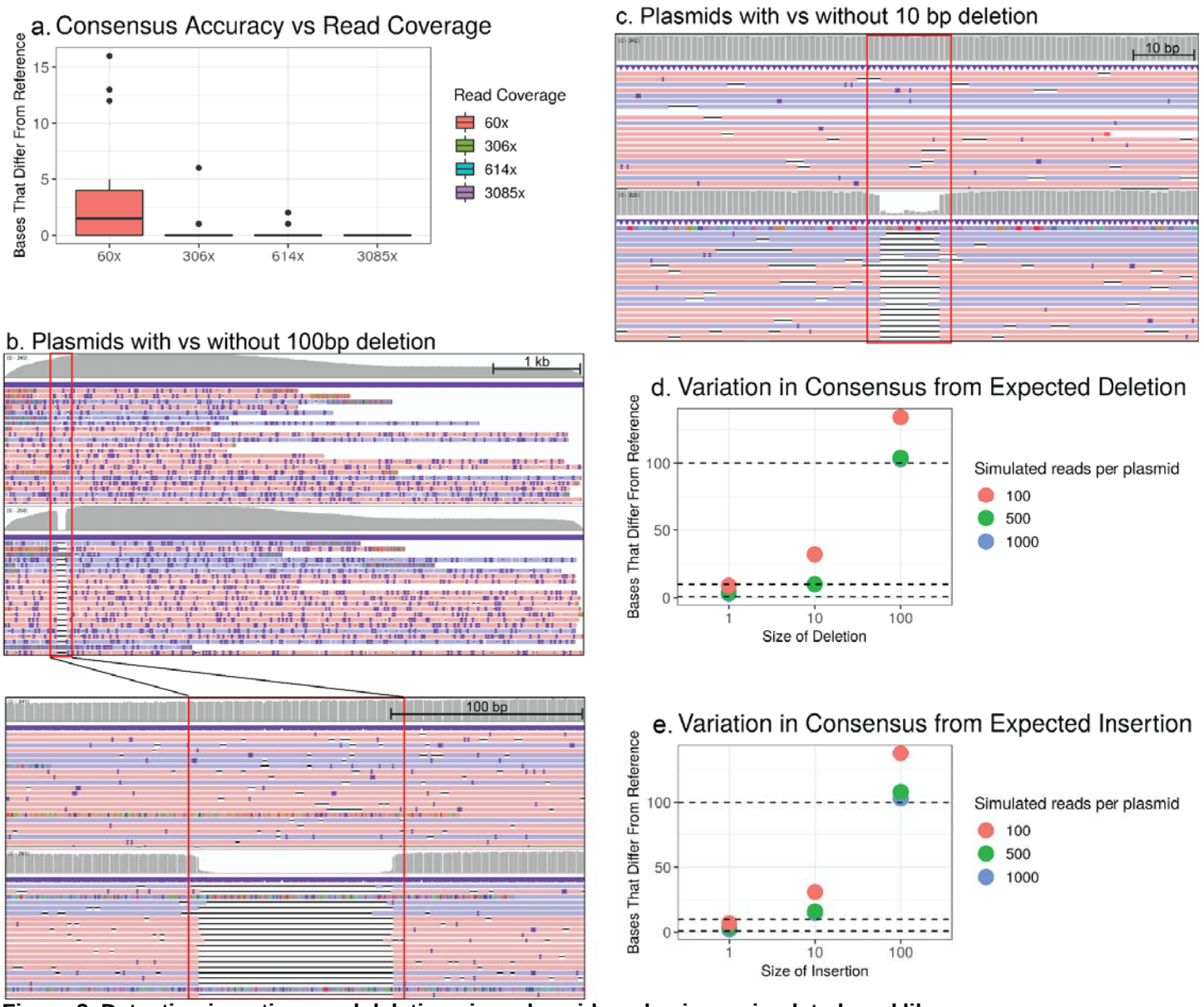


Figure 2. Detecting insertions and deletions in a plasmid pool using a simulated read library
a. Number of gaps in consensus sequence (before indels) for simulated read depth experiments at various read coverage. **b and c.** IGV view of read pileups for reads with vs reads without a 100 bp deletion (b) and a 10 bp deletion (c). Deletions are highlighted by red boxes. Gray top row shows read depth at each position. Purple lines are minus-strand reads, red lines represent plus-strand reads. Dark purple marks are bases that differ from reference in each read. **d and e**. Number of base-pair differences between reference and consensus files for each simulation condition at different read depths. Dotted lines indicate expected number of differences due to simulated deletion (d) or insertion (e).

B. On-Ramp correctly assigns reads to highly similar plasmids

We next tested On-Ramp's ability to correctly assign reads originating from a pool of plasmid sequences with high similarity to each other without barcoding. We created plasmid references that differed only in unique regions of 24bp, 12bp, or 6bp in length (16 unique references at each length) and using NanoSim, constructed simulated plasmid read data. An average of 972 simulated reads were generated for each of the 16 reference plasmids in the pool. Reads and references were then provided to On-Ramp in either biobin or medaka mode. In biobin mode, On-Ramp scans the provided reference sequences for any unique sequence to use for

distinguishing each reference. It then aligns each read to these unique regions to obtain an alignment score, and the read is assigned to the reference where it meets the scoring criteria. Once reads are assigned to reference sequences, On-Ramp applies medaka's consensus tool to each bin of reads individually to generate a consensus for each plasmid. Using On-Ramp in biobin mode with default scoring, less than 6% of reads were assigned to the incorrect reference. In medaka mode of On-Ramp, the medaka consensus tool generates polished consensus sequences from alignments to user-provided reference sequences in place of a draft genome assembly. Unlike biobin, it simultaneously maps all reads against all references to generate the consensus.

For both biobin and medaka modes, the number of simulated reads that were uniquely assigned decreases as the length of the unique sequence decreases (**Extended Fig 1a-d**), and errors in the consensus at map ends and homopolymers increase with decreasing read coverage. Despite these complications, On-Ramp generated accurate consensus sequences for plasmids that differed by only 24bp or 12bp for 3.4kb plasmids. Errors were primarily limited to homopolymers and gaps at the start and end of the reference sequences where there was low read coverage in consensus sequences, which is a limitation of using NanoSim for library simulation. This drop off in end coverage is significantly reduced in real data by utilizing the restriction digest preparation where the majority of reads span the length of the plasmid, increasing end read coverage and the number of uniquely mappable reads containing the unique sequence marker **(Extended Fig 2).**

More reads were uniquely assigned to each reference for the 6bp marker pool in biobin mode, but the consensus sequences contained more gaps compared to medaka mode. In medaka mode the number of uniquely assigned reads is much lower for the 6bp marker, however the consensus had fewer gaps as a result of medaka using non-uniquely assigned reads. While these consensus sequences are more complete, we do not recommend using medaka mode for highly similar plasmid pools (<24bp unique sequence) as the consensus may rely too heavily on non-uniquely assigned reads. Biobin and medaka modes are both reliable for correctly mapping similar plasmid pools where there is at least 24bp of unique sequence to differentiate the plasmids. For highly similar plasmid pools with less unique sequence, an alternative plasmid preparation protocol, described below (restriction-based clonal preparation), bypasses these issues. These results indicate that On-Ramp can correctly assign unique reads originating from a pool of highly similar plasmids sufficiently to build an accurate consensus even when the plasmids differ by only 12bp within a 3.4kb plasmid.

C. Plasmid library preparation and sequencing

We next evaluated the performance of On-Ramp with sequencing of real plasmids. Plasmid preparation protocols were adapted from Nanopore which require ligation of DNA ends with specialized adapters used to facilitate sequencing. Here we describe two methods for preparing plasmid pools based on a fragmentation and adapter attachment methodology **(Figure 1a).** For the first method, ONT's Tn5 transposase randomly fragments pooled plasmid DNA and simultaneously ligates adapters onto fragment ends. In the second method, plasmids are linearized by single restriction enzyme digestion and pooled. The pooled plasmids are mono-adenylated and adapters are ligated to plasmid ends using ONT sequencing ligation kits. Following adapter ligation, flow cells are loaded onto the MinION sequencer with an attached Flongle adapter and primed using ONT's protocol (see Methods). Pooled plasmid libraries are then loaded onto the primed flow cells and sequenced for 24-48 hours. Basecalled reads and plasmid reference files are provided to On-Ramp for analysis. A consensus file is generated for each plasmid using the medaka mode of On-Ramp and then aligned to its reference using Emboss Needle[20] to generate its optimal global pairwise sequence alignment.

The Tn5-based method for plasmid pool preparation was used as a test case to generate actual, non-simulated reads. The Rapid Sequencing Kit from ONT was used, which employs a Tn5 transposase to fragment plasmids, and equimolar pools for two runs were prepared, one 7-plasmid pool and one 16-plasmid pool. Sequencing of the 7-plasmid pool using a Flongle flow cell generated 14,464 reads with a read length N50 of 5.78kb, 13,599 passed with a Q >= 7 quality score and were used in subsequent analysis. The 16-plasmid pool had a read length N50 of 4.74kb and generated 10,265 reads, 8,354 of which passed. On average, 934 reads mapped to each plasmid in the 7-plasmid pool, and 172 reads mapped in the 16-plasmid pool **(Fig 3a)**. Consensus accuracy averaged 4.4 gaps per plasmid for the 7-plasmid run, as measured by per-base differences in consensus vs reference (gaps), not including two 21bp deletions resulting from modifications in the actual plasmids **(Fig 3b).** In the 16 plasmid experiment, 12 plasmids had low (<100x) read depth, leading to an increased number of errors in the consensus and were unlikely to be true mutations **(Fig 3a, b)**. Despite this, 7 of the 16 plasmids had fewer than 10 errors in consensus vs reference alignments.

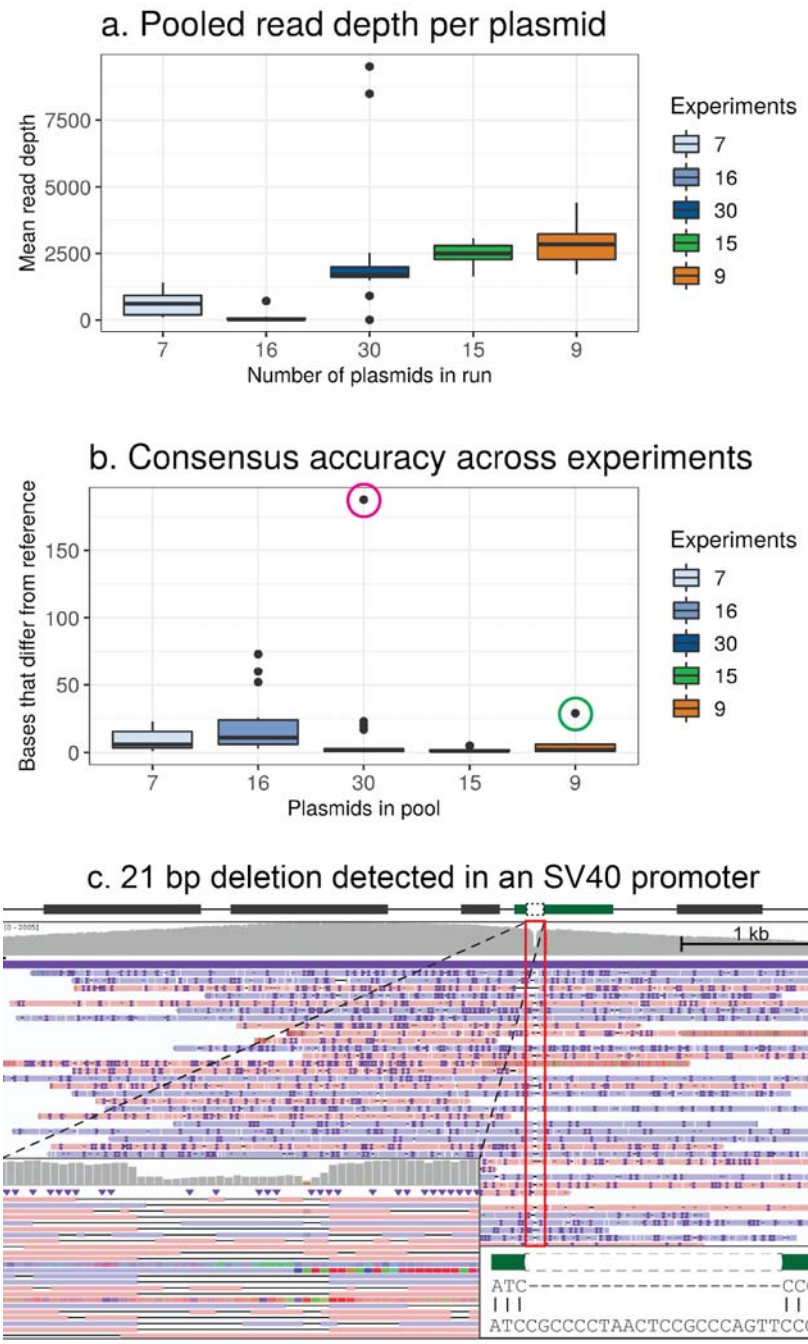


Figure 3. Characteristics of plasmid sequencing experiments and an example of variant detection
a. Per-plasmid read depth across pooled sequencing runs. b. Per-plasmid count of bases in consensus sequence that differ from reference. Circled in fuchsia is a plasmid with a 185bp and circled in green is an example of differences from the reference due to a 29bp deletion in a repetitive element in one plasmid in the pool c. IGV browser view of read alignments that reveal a 21bp deletion (red box) in an SV40 promoter (green) of a plasmid backbone. Left inset - IGV view zoomed in on the deletion region, black lines indicate deletions. Right inset - reference sequence matched to span of deletion in consensus at base-pair level. Grey row indicates per-base read depth.

The next test case of a 30-plasmid pool was prepared using the restriction enzyme method. Plasmids were linearized using restriction digestion, pooled, A-tailed, and then adapters were ligated and samples prepared for sequencing on a Flongle flow cell using the ONT Ligation sequencing kit. A total of 105,680 passing reads with a N50 length of 5.86kb was obtained and aligned to the reference plasmids. Even with the reported 10% error rate for the flongle flow cells, coverage for plasmids across all experiments was high enough (at least 900 reads per each plasmid) to allow for base-pair resolution for all but one of the sequenced plasmids **(Fig 3a)**. The single exception was expected as it was a sample that was known to have failed restriction digest check indicating it likely would not match any of the provided references**.** Plasmids had 3.3

gaps per plasmid on average, with a median gap count of 2 and excluding a 185bp deletion in one plasmid **(Fig 3b)**.

D. Nanopore plasmid sequencing reveals regulatory structural mutations

The high sequencing coverage from even Nanopore's lowest-capacity flow cell (Flongle) allowed for highly similar plasmids to be distinguishable from each other. Three plasmids within the 7-plasmid experiment were similar in sequence except for small 4bp connector sequences and corresponding reads were correctly assigned to their respective reference. Additionally, real sequence variation was detectable at this read level - an unexpected 21bp deletion was identified in two of the plasmids that contain the SV40 promoter **(Fig 3c)** that had not been detectable by diagnostic restriction digest. This deletion was visualized using Integrated Genome Viewer (IGV)[21] included in the On-Ramp website, and read coverage was assessed to determine if this was a true deletion. When compared to plasmids of similar length and read coverage containing the same promoter, this plasmid was the only case of such a deletion, suggesting it was not due to error prone sequencing or low read coverage.

In a separate run, 9 pooled plasmids were sequenced using restriction-based sample preparation, including 6 containing small 40bp repetitive regions in 4 or 6 copies within different plasmids **(Fig 3a-b)**. These regions were previously intractable to Sanger sequencing with no sequence data available through the region, likely due to high secondary structure. Not only was nanopore technology able to sequence completely through the repetitive region and On-Ramp able to assign reads accurately, but sufficient reads were obtained to reveal a single deletion of one of the 6 repetitive elements in one plasmid **(Fig 4d).**

E. Validating plasmid sequences from pools of plasmid clones

A common use-case for validating many plasmids at once on a scale suitable for On-Ramp is simultaneous screening of multiple clonal copies of a few plasmids in order to obtain a positive clone on a first pass validation. On-Ramp's pooled, whole-plasmid protocol and analysis is well-suited for this type of validation. However, as a result of the pooling process, reads originating from different clones of the same plasmid would all map to a single reference. This makes it impossible to differentiate clones with the desired sequence from those with indels or mutations using the resulting sequence data. To address this problem we have developed a simple modified restriction digestion method leveraging the long-read nature of Nanopore sequencing which allows for differentiation of highly similar or clonal plasmids within the same pool.

For plasmid libraries containing single plasmid clones or highly similar plasmids (<24 bp difference), each clone is cut with a different unique restriction enzyme from its matched partners before pooling **(Fig 4a)**. This generates a different, unique linear order of sequence for each clone, despite having identical total sequence content. During analysis, a copy of the plasmid reference sequence is provided for each clone with the linear sequence set starting at the digest site that matches the enzyme used for that clone (termed 'rotated' reference). The functionality is provided by On-Ramp where available restriction sites are identified in a plasmid reference and a user is given the option to select appropriate sites for their experiment. While each cut clone contains the same sequence, the alternate digest sites create linear fragments (reads) that map precisely to their matched 'cut' reference sequence, but poorly to the same sequence reference 'cut' at any other site (**Fig 4a**). This method is uniquely feasible using long-read sequencing, as reads obtained from using nanopore easily encompass the entire length of the plasmid in a single read.

We first tested this approach using simulated reads representing 5 clones of a 2.6kb plasmid, with the closest modeled restriction sites 197bp apart (representing 7.6% of the total plasmid length). 18% of the simulated reads uniquely mapped to their matching reference **(Extended Fig 3)**. This allowed for differentiation of reads originating from different clones and provided an average of 738 reads assigned to each clonal plasmid. The real-world in-vitro functionality of this approach was then tested using five clones each of three similar ~6.5kb plasmids (15 total). Each plasmid clone was linearized with a different unique restriction enzyme, pooled, and adapters were ligated then subsequently sequenced. All three plasmids were identical except for a ~500bp insertion region, and clones were predicted to be identical based on previous restriction digest analysis. The closest cut sites were 579bp apart. We obtained 97.1k passing reads with a N50 length of 6462bp. An average of 2704 reads uniquely mapped to each rotated reference **(Fig 4b)** with a mean depth of 2534.9, compared to 7 reads uniquely mapping with a mean depth of 2.2 to non-rotated references. This indicates that a difference in restriction enzyme cut location is sufficient to create reads that align uniquely to their rotated reference, despite having identical total sequence content.

A second clonal plasmid pool experiment containing three clones with repetitive sequences demonstrated the ability of On-Ramp's protocol and analysis pipeline to detect and map mutations to a specific clone. We include nine plasmids in this pool, with a unique plasmid, and three sets of clones containing two (set 1), or three (sets 2 and 3) clones each **(Fig 4c).** 40,311 reads were generated for the 9 plasmids, and 29,304 reads passed QC, giving a mean depth of 2870 reads per plasmid. Following read assignment and consensus generation using On-Ramp, results were visualized using IGV. Using this restriction-barcoding approach, On-Ramp was able to detect a single 40bp deletion in a functionally crucial repetitive region of a 4.1kb plasmid on one of the three clones **(Fig 4d).**

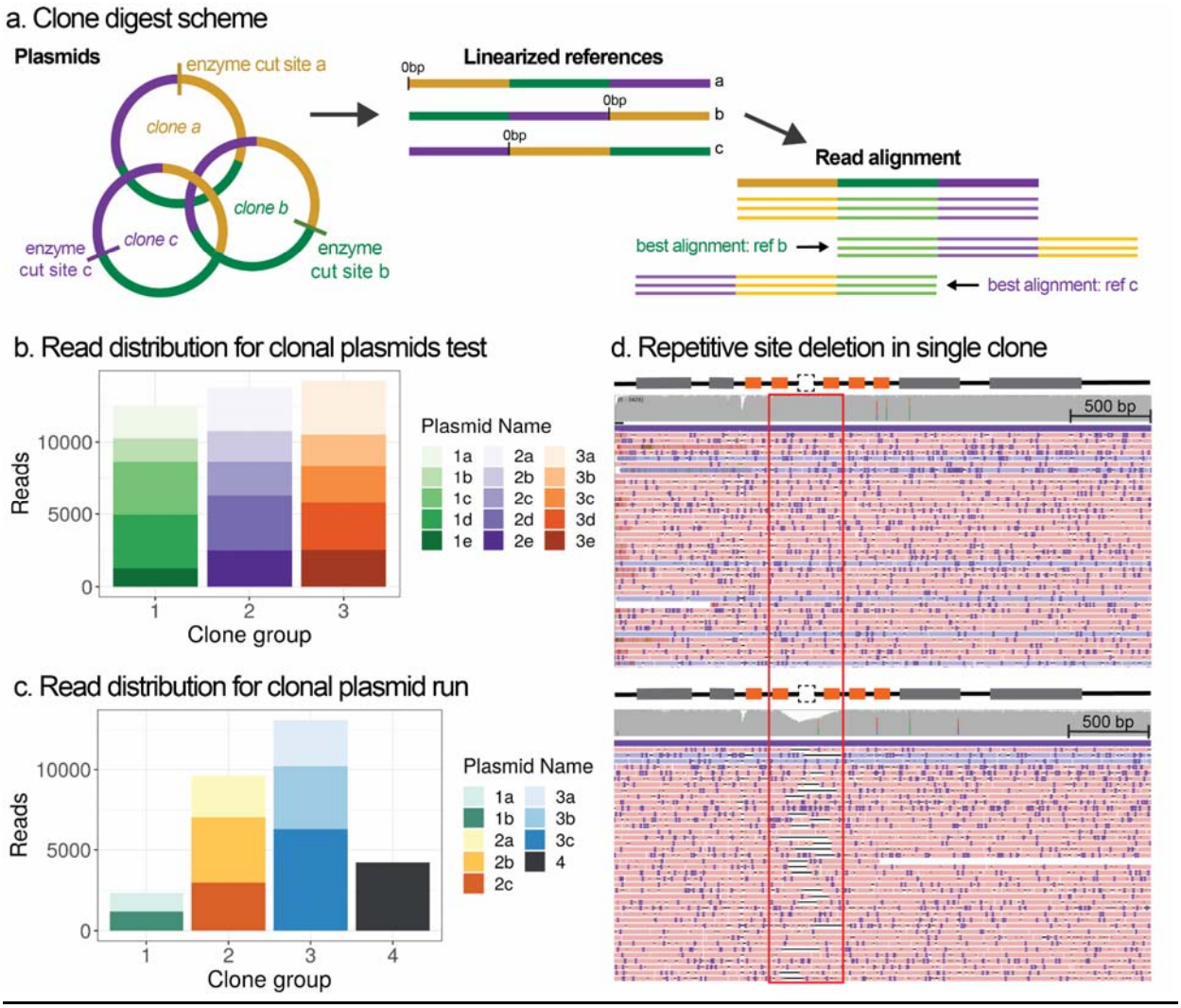


Figure 4. Restriction-digest barcoding for highly similar or clonal plasmids
a. Diagram of restriction cut-site method for unique read mapping of clonal plasmids. b. Number of reads mapping uniquely to each plasmid in a 15-plasmid clonal test pool. c. Reads mapping uniquely to each reference in a second clonal run. d. IGV view of reads mapping to clone without (top) or clone with (bottom) a 40bp deletion in a repetitive region (highlighted in yellow), and plasmid map diagrams above. Thick gray bar indicates per-position read depth. Pink bands are plus-strand, purple bands minus-strand individual reads. Black horizontal lines indicate deletions, purple marks represent base mismatches compared to reference.

Discussion

Assessing recombinant plasmid sequence fidelity is an integral part of any molecular cloning workflow. While Sanger sequencing is a cost-effective method for low-throughput plasmid validation, it can be inadequate for whole-plasmid sequencing, and handles regions with complex secondary structure poorly. As an alternative, high-throughput sequencing (HTS) quickly becomes costly, time consuming, and complex when dealing with multiplexed pools. Due to its short-read nature, HTS cannot identify and correctly assign mutations outside unique regions for highly similar plasmid pools. With the introduction of Oxford Nanopore Technologies' (ONT) sequencing platforms, sequencing of many plasmids in their entirety at high read depth is now possible.

However, current ONT technology and publications have primarily targeted whole-genome sequencing and assembly, variant detection and transcriptomics[22]. Here we present On-Ramp, a set of modified wet-lab protocols specific to pooled plasmid preparation using the Nanopore platform, and an associated computational pipeline designed to support analysis of either different or highly similar plasmid pools. With the addition of our On-Ramp webapp, we provide a rapid (2.5 hours for preparation, 16-24 hours for results) and cost-effective approach for medium-throughput plasmid sequencing and data interpretation that is accessible to labs without the need for extensive bioinformatics support.

We developed the On-Ramp sequencing pipeline using and modifying Nanopore long-read analysis tools to handle plasmid pool data, which were initially designed to handle mapping reads to whole genomes. Testing of On-Ramp using simulated read libraries demonstrated its ability to correctly assign sequencing reads to reference sequences and construct consensus sequences even with highly similar plasmids. On-Ramp was able to detect mutations in plasmids as small as 1bp. Three different barcode-free protocols were tested for pooled plasmid preparation; Tn5 fragmentation, restriction linearization, and restriction-enzyme tagging for clonal populations, providing flexibility in choosing a protocol based on plasmid characteristics. Using these protocols, we prepared and sequenced 7, 9, 15, 16, and 30-plasmid pools using Flongle flow cells. Analysis of read data from these experiments demonstrated that On-Ramp provides high sequence read depth across plasmid pools, generating consensus sequences spanning entire plasmid lengths at base pair resolution, allowing for classification of highly similar plasmids. A key concern with the Flongle is batch effects, wherein different flow cells can have variable active pore numbers for sequencing. Using On-Ramp, full plasmid sequences can be assembled, even with low read counts from few pores (here, Flongles with as few as 20 pores were used). A previously unknown mutation was identified within a single plasmid, showcasing the ability of the tool to determine uncharacterized structural and sequence variation. Additionally, our digest-based method for separation of clonal plasmid copies provides a barcode-free approach for plasmid pooling that maintains full-plasmid read length. This avoids loss of reads without a classifiable barcode signal.

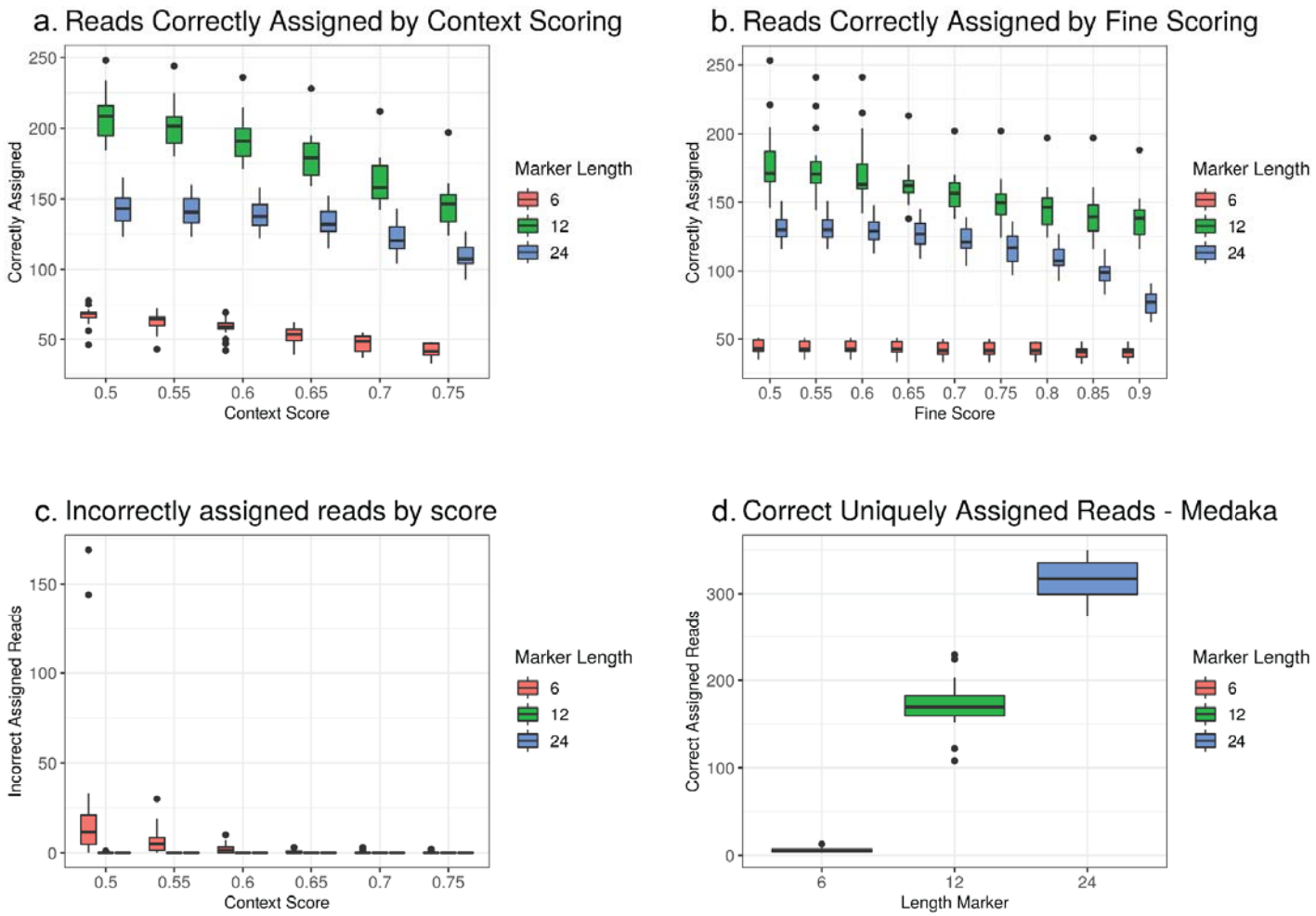
Using our webapp, users can upload reference plasmid sequences and basecalled read files from a Nanopore-platform based sequencing run to rapidly obtain consensus plasmid sequences for many plasmids at once in a matter of minutes. On-Ramp and Nanopore-based plasmid pooled sequencing have the potential to provide a flexible alternative to the current Sanger sequencing standard for plasmid validation for multiple sequenced plasmids, full-plasmid sequences, or sequencing through regions of high secondary structure. Using the protocols and Nanopore components described in our methods, full-plasmid sequences for a run of as few as eight 6kb plasmids can be obtained at the same cost as the equivalent data using Sanger sequencing, and cost-effectiveness increases with more or larger plasmids. The 30-plasmid pool was run at 25% of the cost of Sanger for the same data, and obtained sub-population level sensitivity. For a pool of 4 clones each of five 5kb plasmids using pooled plasmid sequencing on a Nanopore Flongle flow cell, the cost is 1/3 that of Sanger sequencing, including all reagents needed to prepare the plasmids. Additionally, detection of sub-population level variation of a deletion within a repetitive element in one of our plasmids demonstrated the frequency with which bacterial populations can mutate plasmid components, making full-plasmid sequencing for validation an important component of molecular cloning workflows.

Using medaka to generate consensus sequences, we were able to rapidly validate our plasmids based on alignments to reference sequences. Some limitations of this approach arise as a result of medaka being a reference-based approach, as opposed to an assembly-based method. For instance, while we were able to detect most variation in our constructs, consensus sequences for plasmids with very large indels (>1000bp) or where large portions of the plasmid have inserted backwards relative to the reference could not be generated. However, these large rearrangements should be easily detected by complementary diagnostic restriction digest tests. Using the alternate biobin mode, choosing unique regions in the reference is essential to binning reads. Indels in the unique portion of the reference can lead to incorrectly binned reads or failure to generate a consensus. An alternative method is to use the clonal restriction-based method presented above to separate reads from highly similar plasmids. Lastly, we found instances of mixed plasmid populations post-transformation (Fig 3d). This mixed population can be missed in the consensus file however, this issue can be addressed by viewing the read alignments in IGV or similar program. This problem also appears in Sanger sequencing results where sequence files will not show sub-population structure, but could be detectable in trace files.

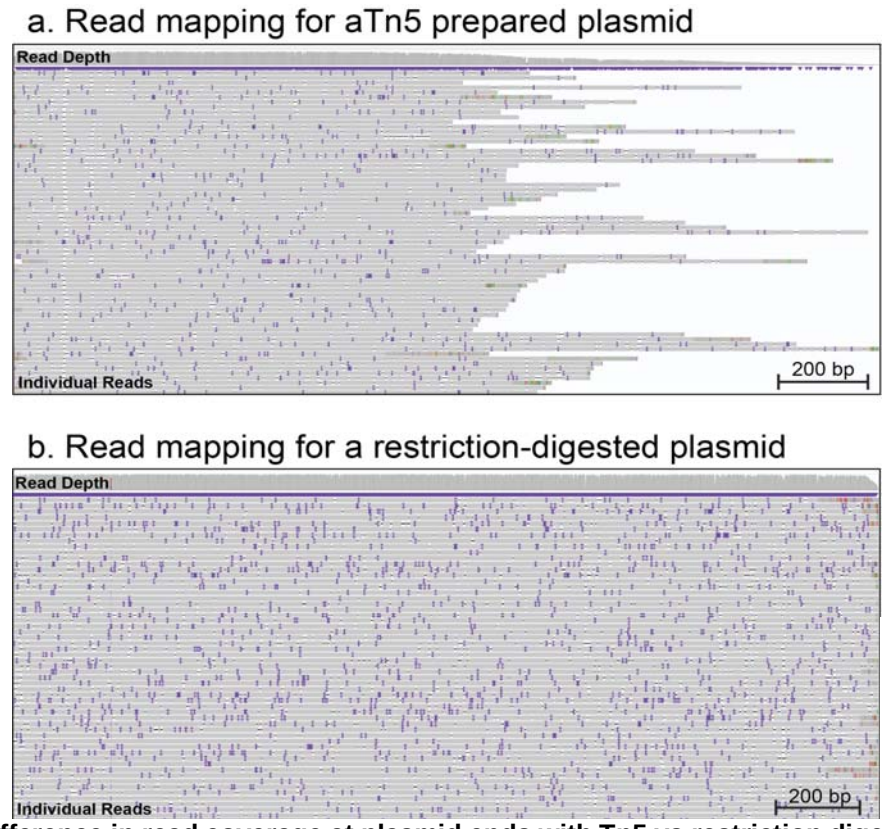
Through this protocol and website, On-Ramp presents an alternate medium-throughput approach to analysis in routine molecular cloning workflows. It provides for rapid sequencing of entire plasmids during

validation on a larger scale than Sanger sequencing, and more affordability and simplicity than high-throughput sequencing.

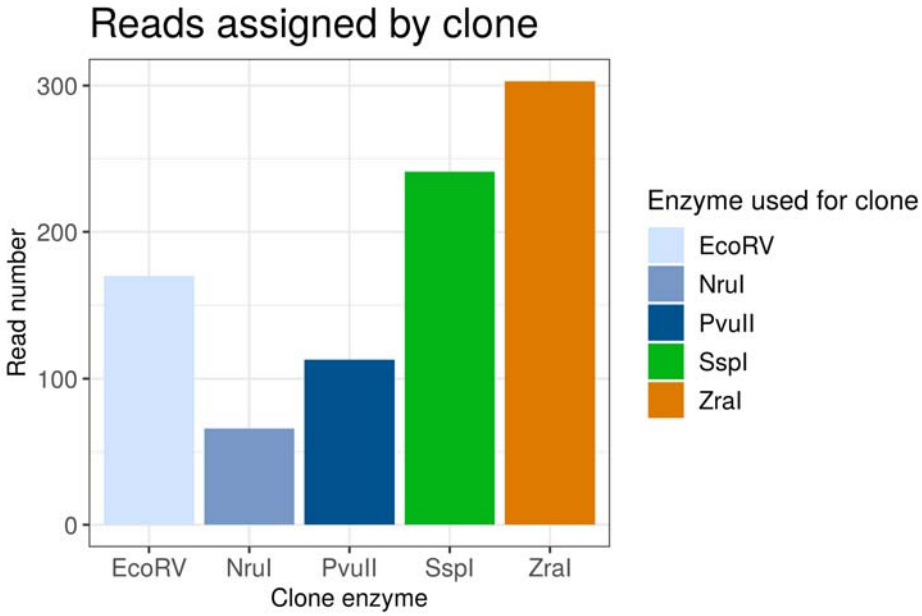
Extended Data Figures



Extended Figure 1. Number of correctly assigned reads for each of the 30 simulated plasmids containing a 6bp, 12bp, or 24bp unique region using different modes. Reads aligned using biobin mode (a,b) with a fixed fine score, comparing read counts at different context scores (a) or with a fixed context score, comparing read counts at different fine scores (b). The number of reads incorrectly assigned using different context scores (c). Count of uniquely mapping, correctly assigned reads using medaka mode (d).



Extended Figure 2. Difference in read coverage at plasmid ends with Tn5 vs restriction digest preparation
IGV view showing read coverage (uniquely mapped reads) for the end of a plasmid prepared using Tn5 (a) or a plasmid prepared using restriction digest (b) with read coverage indicated by height of gray panel at top.



Extended Figure 3. Number of uniquely assigned reads for each clone in a simulated clonal experiment containing 5 identical plasmids. Each clone was cut with a different single enzyme (indicated along x-axis) located in a different position along the plasmid.

Methods

Vector construction and maintenance

Plasmids were constructed using either EMMA[23] or gateway- or restriction-based cloning methods. The EMMA toolkit was a gift from Yizhi Cai (Addgene kit # 1000000119). Various parts from the toolkit were used for construction of the vectors, and mCherry was cloned from pHR-SFFV-KRAB-dCas9-P2A-mCherry to become a usable part. pHR-SFFV-KRAB-dCas9-P2A-mCherry was a gift from Jonathan Weissman (Addgene plasmid # 60954 ; http://n2t.net/addgene:60954 ; RRID:Addgene_60954)[24]. Expression vectors were grown in either Stbl3 or DH5 chemically competent E. coli strains.

Tn5-based plasmid preparation

For TN5-based preparation, plasmids were treated using the Rapid sequencing kit and following ONT's protocol (ONT, SQK-RAD004). Pooled plasmid DNA is brought to 7.5uL using H2O and combined with 2.5uL FRA, and incubated 30°C for 1 minute and then at 80°C for 1 minute then put on ice. 1uL of RAP is added and mixed by flicking, then spun down and incubated for 5 minutes at room temperature. DNA is loaded onto a primed flow cell.

Plasmid pool linearization by restriction enzyme and end-repair

Plasmid DNA was isolated using the QIAprep Spin Miniprep Kit following the manufacturer's protocol (QIAGEN, 27104) and eluted in water. Plasmids were linearized by restriction digest using a unique cut site, with times, temperatures and reaction volumes varied for other enzymes according to NEB recommendations. Example pooled restriction digest: NEB Buffer 3.1 (NEB,B7203S) was added to 1X and the final volume was adjusted with nuclease free water to 200uL. SwaI (NEB, R0604L) was added according to the total amount of DNA present for linearization (minimum 10 Units enzyme per 1 ug DNA), and the sample was digested at 25°C for 30 minutes. Plasmid pools were generated prior to digest if all contained the same unique restriction site, or after digest for plasmid pools where each plasmid required a different restriction enzyme. For plasmids where different restriction enzymes are used on each plasmid, heat-inactivation of each enzyme (following manufacturer instructions) or if not possible, column cleanup (Qiaquick PCR purification kit, QIAGEN, 28104) to remove enzyme was done and is a crucial step prior to pooling to prevent cross-cutting of other plasmids in the pool after combination by still-active enzymes.

Digested plasmids were diluted and pooled into a single 1.5mL microcentrifuge tube using the following rules to calculate desired amount of each plasmid: 1. using an equimolar amount of each plasmid, 2. a maximum of 1000ng total plasmid for the entire pool 3. using at least 10ng of each plasmid 4. a total 50uL volume. Amount of each plasmid in a pool ranged from 15ng-100ng across experiments in this paper. If any digests generated 3' or 5' overhanging bases, pooled plasmids were end-repaired using 1uL (5U) DNA Polymerase I Klenow Fragment (NEB M0210S) with 33 µM each dNTP and 1x NEB CutSmart buffer per 1000ng DNA pool, with incubation for 15 minutes at 25°C, and heat inactivation for 20 minutes at 75°C. Following digestion and end repair, A-tailing was completed using 1uL of 10mM dATP and Taq DNA polymerase (NEB, M0273S) per 50uL of sample with incubation at 75°C for 15 minutes.

ONT Adaptor Ligation

For restriction-prepared enzymes, following DNA linearization, end-repair and A-tailing, ONT's ligation sequencing kit was used (ONT, SQK-LSK109) to add adaptors. One half volume of ligation buffer (4X T4 ligase buffer, 60% PEG 8000), 5uL of T4 DNA ligase (NEB, M0202M), and 2.5uL of AMX (ONT, SQK-LSK109) was added to the plasmid mixture then incubated on a tube rotator at room temperature for 10 minutes. One volume of 1X Tris-EDTA buffer (pH 7.5; Invitrogen, 15567027) and 0.3X room temperature SPRI beads (Beckman Coulter, B23317) were added for selection of >2 kb fragments. The sample-SPRI bead mix was incubated on a tube rotator for 10 minutes on the bench at room temperature. The SPRI beads were washed twice with 100uL of Long Fragment Buffer (LFB; ONT, SQK-LSK109) and the sample was eluted in 9uL of Elution Buffer (EB; ONT, SQK-LSK109).

Nanopore sequencing

Flongle flow cells were primed for the sequencing runs following ONT's standard protocol, using flow cell priming buffers provided by ONT. Briefly, flow cells are QCd to check for a usable number of active pores (~ 0.5-1 pores per plasmid was used here as the minimum). Flow cell was washed with FB then SQB buffer

mixed 1:1 with water. DNA prepared from previous steps is mixed with SQB and LB immediately prior to loading following ONT's protocols.

Simulated reads

NanoSim was used to construct pooled plasmid read libraries. First, a model was created using 81,070 reads (N50=6,003bp) from a previous plasmid sequencing experiment, and the 30 plasmid sequences (average length = 4,318.7bp) were used as the reference genome and input in the characterization set. This model was then used to simulate reads from other plasmid references and from references constructed with 1, 10, 100, and 1000bp deletions and insertions of random sequence.

Bioinformatics pipeline

Basecalling was completed using Guppy (Oxford Nanopore Technologies, 4.5.2) using the dna_r9.4.1_450bps_hac.cfg configuration and passing reads (Q > 7) were filtered using Guppy or NanoFilt[25]. Adapters were trimmed using Porechop (https://github.com/rrwick/Porechop). On-Ramp allows users to use Porechop and NanoFilt to trim reads and filter by q score and read length. Reference sequences were generated using SnapGene (https://www.snapgene.com/). The reads and references were then used as input for On-Ramp during testing.

Medaka

The medaka consensus (https://github.com/nanoporetech/medaka) module was utilized to generate consensus sequences from read pileups using the '-g' flag to stop filling in gaps with draft/reference sequence during consensus stitching. The resulting alignments are then filtered (MAPQ >= 10) for visualization using the Integrative Genomics Viewer (IGV)[26]. Final pairwise alignments were constructed between the reference and consensus sequences generated by medaka using Emboss needle.

Binning

The biobin module mode of plasmid sequencing was used to bin reads based on unique sequences in the provided references. The biobin mode/module searches the reference sequences for unique sequences longer than 3bp and a set is constructed for each plasmid reference. Each input read was then aligned to these regions using Biopython pairwise aligner with alignment parameters match: 3, mismatch: -6, open_gap: -10, extend: -5. Reads were first aligned to an extended portion of the plasmid containing 20bp flanking the unique region and assessed using the 'context score'. For reads that passed this threshold, the aligned portion was then aligned and scored against the exact unique region and high scoring reads (fine score > 80) were assigned to the plasmids. Each of the resulting bins was then passed to medaka for consensus polishing.

Acknowledgements

M.D. and T.M. were supported by NIH Training Grant Michigan Predoctoral Training in Genetics (T32GM007544) C.M was supported by University of Michigan Genome Science Training Program (5T32HG000040-27).

Author Contributions

A.P.B., T.M., M.D. and C.M conceived the project. C.M and A.G.D. developed On-Ramp pipeline and webapp. C.M performed data analysis. All authors guided the experiment design and data analysis strategy (C.M conceived and performed simulated experiments, T.M conceived and performed Tn5 in-vitro experiments, M.D. conceived and performed in-vitro clonal plasmid sequencing experiment, and J.S and M.D ran the other restriction-based in-vitro sequencing experiments). A.P.B. supervised the experiments, analysis, and data interpretation. C.M, M.D. and T.M wrote the manuscript and all authors contributed edits and revisions. All authors read and approved the final manuscript.

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

2. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**, 172 (2014).

3. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

4. Mali, S. Delivery systems for gene therapy. *Indian J. Hum. Genet.* **19**, 3–8 (2013).

5. Potapov, V. & Ong, J. L. Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One* **12**, e0169774 (2017).

6. Conley, E. C., Saunders, V. A. & Saunders, J. R. Deletion and rearrangement of plasmid DNA during transformation of Escherichia coli with linear plasmid molecules. *Nucleic Acids Res.* **14**, 8905–8917 (1986).

7. Cutting, G. R., Antonarakis, S. E., Youssoufian, H. & Kazazian, H. H., Jr. Accuracy and limitations of pulsed field gel electrophoresis in sizing partial deletions of the factor VIII gene. *Mol. Biol. Med.* **5**, 173–184 (1988).

8. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).

9. Shinde, D., Lai, Y., Sun, F. & Arnheim, N. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.* **31**, 974–980 (2003).

10. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).

11. Kittleson, J. T., Wu, G. C. & Anderson, J. C. Successes and failures in modular genetic engineering. *Curr. Opin. Chem. Biol.* **16**, 329–336 (2012).

12. Williams, J. A., Carnes, A. E. & Hodgson, C. P. Plasmid DNA vaccine vector design: impact on efficacy, safety and upstream production. *Biotechnol. Adv.* **27**, 353–370 (2009).

13. Sanchis-Juan, A. *et al.* Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med.* **10**, 95 (2018).

14. McDonald, T. L. *et al.* Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat. Commun.* **12**, 3586 (2021).

15. Gilpatrick, T. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).

16. Bowden, R. *et al.* Sequencing of human genomes with nanopore technology. *Nat. Commun.* **10**, 1869 (2019).

17. *medaka: Sequence correction provided by ONT Research*. (Github).

18. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).

19. Yang, C., Chu, J., Warren, R. L. & Birol, I. NanoSim: nanopore sequence read simulator based on statistical characterization. *Gigascience* **6**, 1–6 (2017).

20. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

21. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

22. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).

23. Martella, A., Matjusaitis, M., Auxillos, J., Pollard, S. M. & Cai, Y. EMMA: An Extensible Mammalian Modular Assembly Toolkit for the Rapid Design and Production of Diverse Expression Vectors. *ACS Synth. Biol.* **6**, 1380–1392 (2017).

24. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

25. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

26. Robinson, J. T., Thorvaldsdóttir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *bioRxiv* 2020.05.03.075499 (2020) doi:10.1101/2020.05.03.075499.