Biofoundry-Scale DNA Assembly Validation Using Cost-Effective High-Throughput Long-Read Sequencing

ACS Synthetic Biology

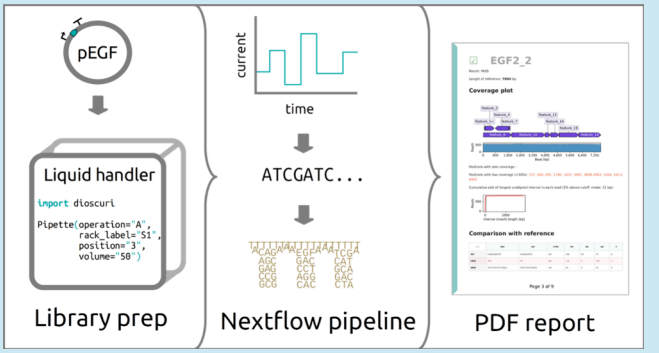
pubs.acs.org/synthbio
Technical Note

Biofoundry-Scale DNA Assembly Validation Using Cost-Effective High-Throughput Long-Read Sequencing

Peter Vegh,* Sophie Donovan, Susan Rosser, Giovanni Stracquadanio, and Rennos Fragkoudis*

Cite This: ACS Synth. Biol. 2024, 13, 683−686

ABSTRACT: Biofoundries are automated high-throughput facilities specializing in the design, construction, and testing of engineered/synthetic DNA constructs (plasmids), often from genetic parts. A critical step of this process is assessing the fidelity of the assembled DNA construct to the desired design. Current methods utilized for this purpose are restriction digest or PCR followed by fragment analysis and sequencing. The Edinburgh Genome Foundry (EGF) has recently established a single-molecule sequencing quality control step using the Oxford Nanopore sequencing technology, along with a companion Nextflow pipeline and a Python package, to perform in-depth analysis and generate a detailed report. Our software enables researchers working with plasmids, including biofoundry scientists, to rapidly analyze and interpret sequencing data. In conclusion, we have created a laboratory and software protocol that validates assembled, cloned, or edited plasmids, using Nanopore long-reads, which can serve as a useful resource for the genetics, synthetic biology, and sequencing communities.



KEYWORDS: *biofoundry, DNA assembly, plasmid validation, sequencing*

INTRODUCTION

A critical step of the Design Build Test Learn (DBTL) cycle widely adopted in synthetic biology[1] is verifying the fidelity of the DNA construct, obtained in the Build phase, to the designed DNA sequence. Several factors may lead to erroneous DNA constructs, including incorrect input DNA, problems during assembly or sample handling, misannealing overhangs, addition of point mutations, and homologous or other forms of recombination. Currently, the verification step relies mostly on a restriction digest or PCR followed by fragment analysis (FA) and sequencing.[2] FA provides a cost-efficient but indirect, low confidence confirmation of construct correctness by checking the presence of specific restriction enzyme recognition sites and the fragment size. Conversely, DNA sequencing approaches provide a nucleotide-level readout at the expense of substantially higher cost. Sanger sequencing, for example, is not feasible in most cases due to the cost and high number of reactions, but it may be useful for verifying targeted regions or in small batches of similar constructs assembled from shared genetic parts.[2] Second- and third-generation sequencing methods provide a solution to sequencing large batches of plasmid constructs due to their high-throughput and no requirement for using primers. In this Technical Note, we describe a single-molecule sequencing DNA assembly quality control solution at the Edinburgh Genome Foundry (EGF) that can be utilized by biologists and the sequencing community. EGF is an automated high-throughput facility (biofoundry) specializing in the modular assembly of DNA constructs (plasmids), using Golden Gate cloning. EGF's platform is species agnostic, and its outputs are used in projects as diverse as programming of stem cells for personalized medicine applications, vaccine development, gene therapy, and many more.

RESULTS AND DISCUSSION

The DNA verification by sequencing step at the end of the Build phase of a synthetic biology (engineering) project aims to compare the fidelity of an assembled DNA construct to its corresponding designed sequence. Here, we present a one-step software pipeline that facilitates and expedites analysis and interpretation of sequencing data (Figure 1). Specifically, we wanted to obtain an annotated comparison of the sequenced and the expected DNA, and a judgment call for each sample, based on various checks.

Received: September 20, 2023
Revised: January 16, 2024
Accepted: January 16, 2024
Published: February 8, 2024

https://doi.org/10.1021/acssynbio.3c00589
ACS Synth. Biol. 2024, 13, 683−686

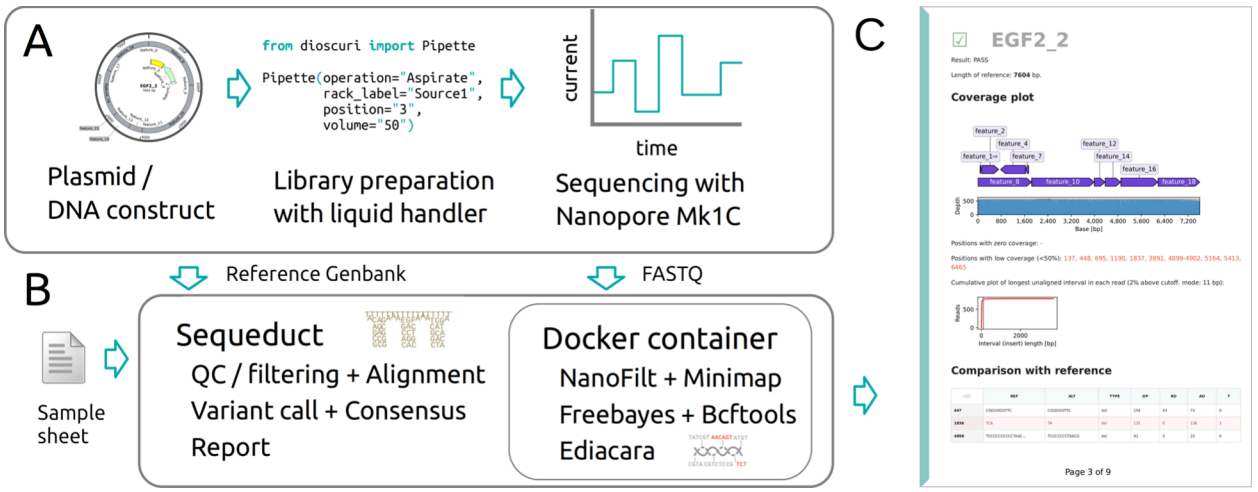


Figure 1. An overview of the sequencing pipeline. (A) The assembled or cloned plasmids are prepared into libraries using a liquid handling platform. Libraries are loaded onto a Flongle flow cell in an Oxford Nanopore Mk1C sequencer. (B) The FASTQ files are analyzed with a Nextflow pipeline (Sequeduct) that utilizes a Docker container with all required software. (C) The Ediacara Python package creates a PDF report for an easy overview and interpretation of the results.

We hereby refer to *errors* as differences (or mutations or variants) between the designed and assembled DNA sequence. These errors include single nucleotide variants (SNV), small insertions and deletions (indels), structural variations (SV), and sequencing errors.[3] Although random sequencing errors can be mitigated with increased sequencing depth, systematic errors are much harder to avoid; here, we focused on SNVs and SVs.

We applied automated protocols for the Oxford Nanopore Technologies (ONT) rapid barcoding kits to simultaneously fragment each sample (plasmid or DNA construct) and ligate individual barcodes. Up to 96 barcoded samples are pooled into a single library and loaded onto Flongle flow cells in a MinION Mk1C sequencer (see Methods section). A Nextflow[4] pipeline, named Sequeduct, has been created to perform alignment, variant detection, and reporting (https://github.com/Edinburgh-Genome-Foundry/Sequeduct/). This pipeline requires the FASTQ files of the partial and full length reads obtained from sequencing, the reference Genbank files of the designed sequences, and a sample sheet. The generated PDF report contains a chapter for each barcode: read statistics are followed by a histogram of the read lengths, which is a good indicator of the presence of structural errors. A displayed coverage chart with the reference sequence visualizes any deletions (Supporting Information S1). Visualizing insertions in a similar plot is not straightforward, therefore a cumulative plot of the longest unaligned interval is provided as an indicator of large insertions. A variant (error) table is also provided in a simplified variant call format (VCF),[5] which lists point and few nucleotide variants but not structural variants such as large deletions or insertions. Variants at homopolymer stretches are flagged as this is a known systemic sequencing error.[3] Variants are also annotated on the reference sequence on a second plot for an easy overview and navigation. Based on the above results, each of the plasmids with a sufficient number of reads is assigned a pass/fail outcome, as detailed in the Methods. A summary spreadsheet of the results, based on a review of the report, is also created. This can be revised by the user. A detailed guide on the interpretation of the report is published online (https://github.com/Edinburgh-Genome-Foundry/Sequeduct_demo). The pipeline also returns the *variant call* consensus FASTA sequence for each barcode.

If structural variants are found, then a subsequent task is to describe their nature and provide an explanation. This is largely beyond the scope of validation, but a second pipeline which aligns user-specified DNA part sequences against a *de novo* plasmid sequence assembled from the reads, using Canu,[6] is also provided. Alignments are reported and visualized in a PDF file (Supporting Information S2). This is useful for evaluating plasmids (or other DNA) that are assembled from parts using various Golden Gate toolkits, such as EMMA,[7] MoClo,[8] or Mobius,[9] or any other method, such as Gibson[10] assembly, and helps clarify whether part or sample mix-ups, recombination events, or overhang misannealing has occurred.

Multiple alternative approaches have been published by Oxford Nanopore Technologies and other research laboratories. The EPI2ME Clone validation workflow uses *de novo* assembly to produce a FASTA file for each sample.[11] The SequenceGenie workflow analyzes data from a multiplexed sequencing approach using a novel sample barcoding system.[12] The MinION Plasmid Sequence Verification Pipeline provides a cost-effective way of sequencing plasmids for clinical research applications.[13] Circuit-seq creates *de novo* assemblies from multiplexed samples,[14] while OnRamp is reference-based.[15] In comparison to these, Sequeduct performs an evaluation against an expected sequence and focuses on the produced report and downstream interpretation of results, which are more suitable for engineering biology and quality control purposes.

Several companies provide a Nanopore sequencing service of plasmids for a fee (~$15/plasmid). Performing the sequencing in-house is cost-competitive and fast, provided that at least 24 samples are sequenced at the same time. In any case, Sequeduct is free software and can be used with FASTQ data from sequencing providers in order to generate a more detailed and targeted report. This addresses a general need of quickly interpreting sequencing verification results and helps researchers routinely verify a received plasmid, an introduced mutation, the results of a cloning experiment, or DNA assembled from parts.

Ongoing development aims to incorporate additional functionalities and improvements. These include deconvolut-

https://doi.org/10.1021/acssynbio.3c00589
ACS Synth. Biol. 2024, 13, 683−686

ing mixed samples, where multiple plasmids are in the same sample or use the same barcode. This would allow sequencing polyclonal ("polyploid") samples or combinatorial assemblies or pooling multiple plasmids into the same barcode. Similarly, we plan to address more error scenarios in the pipeline as we accumulate sequencing results. We anticipate that analysis of synthetic design outcomes using full length sequence results will lead to the establishment of more robust DNA design rules. For example, analysis of a collection of plasmid sequencing data can point to sequence patterns that interfere with cloning or cause recombination. Alternatively, a review of the results can help find problematic components in a project so that these can be avoided in the next round of design.

In conclusion, we have set up a complete solution—consisting of laboratory and software protocols—for the validation of assembled, cloned, or edited plasmids and other DNA, using long-reads. The software is available under a free and open-source license (GPLv3) to encourage contributions and feedback from biofoundries and the sequencing community.

METHODS

Sequencing. Plasmid DNA is prepared using the Wizard SV 96 Plasmid DNA Purification System (Promega), and the concentration is measured by fluorescence-based quantification (Qubit dsDNA BR Assay Kit, Thermo Scientific). Samples are normalized to be within the 20−90 fmol/μL range, and 1 μL of sample is used for library preparation. The protocols for the ONT Rapid Barcoding Kit (SQK-RBK004) or Rapid Barcoding Kit 96 (SQK-RBK110.96) are performed following the manufacturer's instructions on an Opentrons OT-2 (SQK-RBK004) or Tecan Freedom EVO200 (SQK-RBK110.96) liquid handling robot. Libraries are loaded onto Flongle flow cells (R9.4.1) and run for up to 24 h on a MinION Mk1C device that performs basecalling using Guppy v4.3.4.

Analysis. The "pass" folder of the FASTQ sequencing data is used in the analysis. The pipeline is written in Nextflow and is available on GitHub with documentation and an example data set at https://github.com/Edinburgh-Genome-Foundry/Sequeduct_demo. The first workflow ("preview") generates summary plots of each barcode for an overview of the sequencing run, using NanoPlot, part of the NanoPack suite.[16] The second workflow ("analysis") filters FASTQ files using NanoFilt followed by read alignment to the reference sequence using minimap2.[17] The reference sequence files can be created with a sequence editor or batch simulated using DNA Cauldron.[18] SAMtools[19] is used to obtain coverage data, and variants are called with freebayes.[20] Consensus sequence files are created with BCFtools.[21] As part of the pipeline, a Python package, Ediacara, was also written to generate a report PDF that visualizes results for each barcode/plasmid (https://github.com/Edinburgh-Genome-Foundry/Ediacara). The package utilizes Biopython[22] and DNA Features Viewer.[23] Each plasmid is assigned an outcome: samples with an insufficient number of reads (sequencing problems) resulting below 30× coverage are marked as "low coverage". The remaining samples are marked "fail" if problems are detected: zero coverage sections, a majority of reads having an unaligned (insert) segment, or a consensus sequence length outside tolerance. The "warning" label is applied for cases where the errors are below the set threshold levels. All other samples are marked as "pass". An explanation of the report is also included in its Appendix section. In addition to the pipeline, a Dockerfile is also provided to generate a Docker image with all of the required software.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.3c00589.

Supporting Information S1: example PDF report of the "analysis" pipeline; Supporting Information S2: example PDF report of the "review" pipeline (PDF)

AUTHOR INFORMATION

Corresponding Authors

Rennos Fragkoudis − *Edinburgh Genome Foundry, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom; Department of Biochemistry and Biotechnology, University of Thessaly, 41500 Larissa, Greece;* orcid.org/0000-0002-8451-2665; Email: r.fragkoudis@ed.ac.uk

Peter Vegh − *Edinburgh Genome Foundry, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom;* orcid.org/0000-0002-0133-3240; Email: peter.vegh@ed.ac.uk

Authors

Sophie Donovan − *Edinburgh Genome Foundry, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom*

Susan Rosser − *Edinburgh Genome Foundry, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom;* orcid.org/0000-0002-2560-6485

Giovanni Stracquadanio − *Edinburgh Genome Foundry, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom;* orcid.org/0000-0001-9819-3645

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.3c00589

Author Contributions

All authors contributed to the design of the sequencing quality control step and pipeline, and the preparation of the manuscript. P.V. wrote the manuscript, designed and implemented the bioinformatics pipeline, and interpreted results. S.D. wrote the manuscript, implemented the laboratory protocol, and interpreted results. G.S. designed the bioinformatics pipeline. R.F. wrote the manuscript and contributed to the design of the laboratory protocol and pipeline.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The Edinburgh Genome Foundry is supported by the BBSRC (BB/M018040/1) and the BBSRC/MRC/EPSRC-funded UK Centre for Mammalian Synthetic Biology as part of the RCUK's Synthetic Biology for Growth program. This work was supported by the UKRI EPSRC Fellowship (EP/V033794/1) to G.S.

ABBREVIATIONS

VCF, variant call format; DBTL, design build test learn; FA, fragment analysis

REFERENCES

(1) Holowko, M. B.; Frow, E. K.; Reid, J. C.; Rourke, M.; Vickers, C. E. Building a biofoundry. *Synth. Biol.* **2021**, *6*, ysaa026.

(2) Zulkower, V. Computer-Aided Planning for the Verification of Large Batches of DNA Constructs. *Methods Mol. Biol. Clifton NJ.* **2021**, *2229*, 167−174.

(3) Delahaye, C.; Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. *PLoS One* **2021**, *16*, e0257521.

(4) Ewels, P. A.; Peltzer, A.; Fillinger, S.; Patel, H.; Alneberg, J.; Wilm, A.; Garcia, M. U.; Di Tommaso, P.; Nahnsen, S. The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **2020**, *38*, 276−278.

(5) Danecek, P.; Auton, A.; Abecasis, G.; Albers, C. A.; Banks, E.; DePristo, M. A.; Handsaker, R. E.; Lunter, G.; Marth, G. T.; Sherry, S. T.; McVean, G.; Durbin, R. The variant call format and VCFtools. *Bioinforma. Oxf. Engl.* **2011**, *27*, 2156−2158.

(6) Koren, S.; Walenz, B. P.; Berlin, K.; Miller, J. R.; Bergman, N. H.; Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722−736.

(7) 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.* **2017**, *6*, 1380−1392.

(8) Engler, C.; Youles, M.; Gruetzner, R.; Ehnert, T.-M.; Werner, S.; Jones, J. D. G.; Patron, N. J.; Marillonnet, S. A Golden Gate Modular Cloning Toolbox for Plants. *ACS Synth. Biol.* **2014**, *3*, 839−843.

(9) Andreou, A. I.; Nakayama, N. Mobius Assembly: A versatile Golden-Gate framework towards universal DNA assembly. *PLoS One* **2018**, *13*, No. e0189892.

(10) Gibson, D. G.; Young, L.; Chuang, R.-Y.; Venter, J. C.; Hutchison, C. A.; Smith, H. O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **2009**, *6*, 343−345.

(11) EPI2ME Labs Clone validation workflow (https://github.com/epi2me-labs/wf-clone-validation).

(12) Currin, A.; Swainston, N.; Dunstan, M. S.; Jervis, A. J.; Mulherin, P.; Robinson, C. J.; Taylor, S.; Carbonell, P.; Hollywood, K. A.; Yan, C.; Takano, E.; Scrutton, N. S.; Breitling, R. Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. *Synth. Biol.* **2019**, *4*, ysz025.

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

(14) Emiliani, F. E.; Hsu, I.; McKenna, A. Multiplexed Assembly and Annotation of Synthetic Biology Constructs Using Long-Read Nanopore Sequencing. *ACS Synth. Biol.* **2022**, *11*, 2238−2246.

(15) Mumm, C.; Drexel, M. L.; McDonald, T. L.; Diehl, A. G.; Switzenberg, J. A.; Boyle, A. P. Multiplexed long-read plasmid validation and analysis using OnRamp. *Genome Res.* **2023**, *33*, 741−749.

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

(17) Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **2021**, *37*, 4572−4574.

(18) Zulkower, V. Computer-Aided Design and Pre-validation of Large Batches of DNA Assemblies. *Methods Mol. Biol. Clifton NJ.* **2021**, *2229*, 157−166.

(19) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **2009**, *25*, 2078−2079.

(20) Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv*, **2012**.

(21) Danecek, P.; Bonfield, J. K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M. O.; Whitwham, A.; Keane, T.; McCarthy, S. A.; Davies, R. M.; Li, H. Twelve years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008.

(22) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.

(23) Zulkower, V.; Rosser, S. DNA Features Viewer: a sequence annotation formatting and plotting library for Python. *Bioinformatics* **2020**, *36*, 4350−4352.

https://doi.org/10.1021/acssynbio.3c00589
ACS Synth. Biol. 2024, 13, 683−686