Bioinformatics, 37(5), 2021, 720–722
doi: 10.1093/bioinformatics/btaa740
Advance Access Publication Date: 25 August 2020
Applications Note

Downloaded from https://academic.oup.com/bioinformatics/article/37/5/720/5896982 by Department of mathematics user on 03 October 2021

Sequence analysis

SOLQC: Synthetic Oligo Library Quality Control tool

Omer Sabary[1,*,†], Yoav Orlev[2,†], Roy Shafir[1,2], Leon Anavy[1], Eitan Yaakobi[1] and Zohar Yakhini[1,2]

[1]The Henry and Marilyn Taub Faculty of Computer Science, Technion, Haifa, 3200003, Israel and [2]School of Computer Science, Herzliya Interdisciplinary Center, Herzliya 4610101, Israel

*To whom correspondence should be addressed.

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

Associate Editor: Inanc Birol

Received on November 26, 2019; revised on July 27, 2020; editorial decision on August 11, 2020; accepted on August 19, 2020

Abstract

Motivation: Recent years have seen a growing number and an expanding scope of studies using synthetic oligo libraries for a range of applications in synthetic biology. As experiments are growing by numbers and complexity, analysis tools can facilitate quality control and support better assessment and inference.

Results: We present a novel analysis tool, called *SOLQC*, which enables fast and comprehensive analysis of synthetic oligo libraries, based on NGS analysis performed by the user. SOLQC provides statistical information such as the distribution of variant representation, different error rates and their dependence on sequence or library properties. SOLQC produces graphical reports from the analysis, in a flexible format. We demonstrate SOLQC by analyzing literature libraries. We also discuss the potential benefits and relevance of the different components of the analysis.

Availability and implementation: SOLQC is a free software for non-commercial use, available at https://app.gitbook.com/@yoav-orlev/s/solqc/. For commercial use please contact the authors.

Contact: omersabary@cs.technion.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DNA synthesis technology has greatly developed over recent years and is holding a promise to enable a leap in using natural systems for various applications. Synthetic oligonucleotide libraries (OLs) consisting of thousands of DNA sequences, often (and herein) referred to as **variants**, have become a common tool in molecular biology. Studies using OLs performed systematic, unbiased investigations of gene regulation mechanisms and genetic variations (Kotler *et al.*, 2018; Levy *et al.*, 2017; Sharon *et al.*, 2012). Design and optimization of protein engineering and CRISPR-Cas9-based tools for genome editing also used OLs (Miles *et al.*, 2016). Synthetic DNA is also an attractive alternative for data storage media. With an information density orders of magnitude better than that of magnetic media and due to its highly robust chemical properties DNA can potentially efficiently store data for centuries. This was demonstrated in a series of studies covering a variety of encoding schemes, sequencing technologies and data access capabilities (Anavy *et al.*, 2019; Blawat *et al.*, 2016; Church *et al.*, 2012; Erlich and Zielinski, 2017; Goldman *et al.*, 2013; Grass *et al.*, 2015; Organick *et al.*, 2018; Yazdi *et al.*, 2017).

The process of using OLs in such studies usually starts with a design file containing the DNA variants, which will be synthesized as millions of physical oligonucleotides (oligos). These oligos will typically be sequenced in one or more steps of the experimental process. It is important to control the quality of the OL throughout the process to ensure that the results stem from the biology and not from technical noise and other biases related to the DNA synthesis and sequencing. The processes of synthesizing, storing, sequencing and handling oligonucleotides are all error prone. These errors include sequence alterations of specific molecules in the form of base substitutions, insertions and deletions as well as frequency variation between the different variants that can result from non-uniform synthesis or biases in amplification steps (Heckel *et al.*, 2019; Pan *et al.*, 2014; Ruijter *et al.*, 2009). While quality assessment of NGS data is common in many experimental pipelines, it is usually done on natural DNA and focuses on the technical quality assessment reported by the NGS platform and on assessing possible contamination in the samples without using any information about the expected sequences (Andrews *et al.*, 2010). Characterizing errors in OLs, based on the library design, has only been done in the context of individual studies with no standard assessment protocols and tools (Heckel *et al.*, 2019; Kosuri and Church, 2014; Organick *et al.*, 2018; Tian *et al.*, 2004).

In this work, we present SOLQC, a software tool that supports and potentially standardize the statistical analysis and quality control of OLs. The tool is designed to facilitate analysis by individual labs to obtain information about DNA libraries and to perform error analysis before or during experiments. Supplementary Information reports results from analyzing several literature libraries.

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

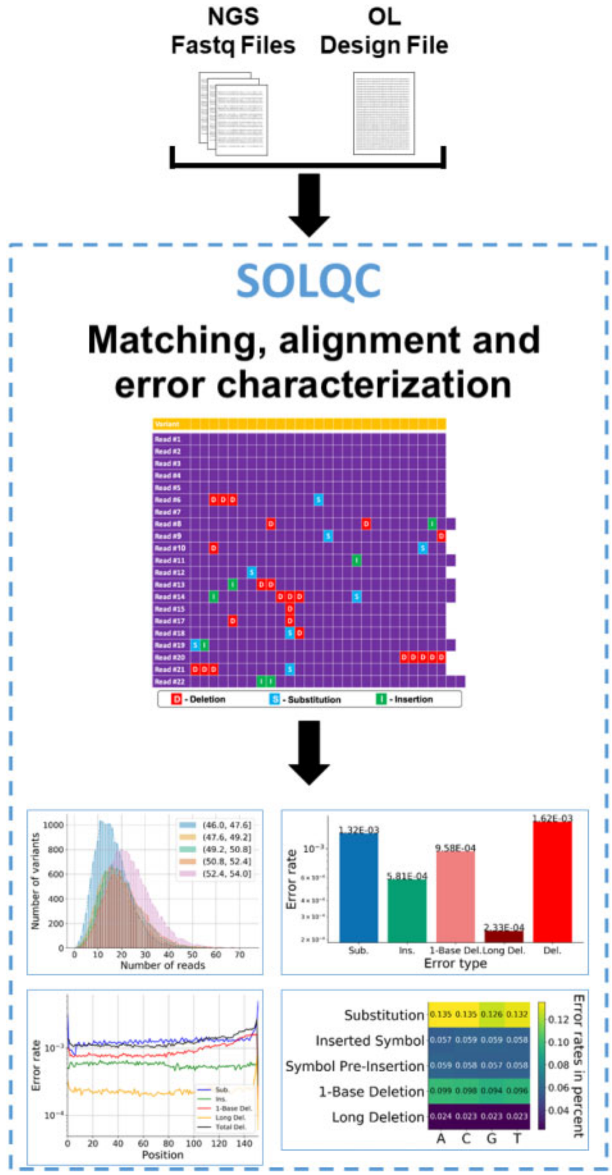

Fig. 1. SOLQC workflow. (Top) Input files. (Center) An example of 22 reads (purple) aligned to a variant of length 27 (yellow). Deletions, substitutions and insertions are marked in red, blue, green, respectively. (Bottom) Four example analyses based on data from Erlich and Zielinski (2017). (i) Histogram of the number of filtered reads per variant, stratified by the GC content. (ii) Total error rates. (iii) Error rates by position. *X*-axis represents position counted from the 5′ end of the designed variant. The *Y*-axis is log-scale. (iv) Error rates stratified by symbol. Note that the numbers are in percents. For example, the value of 0.024 for 'A' long deletion, means that 0.00024 of the occurrences in base A in the library creates long deletion error

2 SOLQC tool

We present a quality control software tool, called *SOLQC—Synthetic Oligo Library Quality Control*. SOLQC generates a report summarizing the analysis of the synthetic DNA library. Figure 1 depicts the overall workflow of SOLQC together with example analyses.

2.1 Inputs and outputs

Any input for SOLQC includes:

1. Design file: This file consists of the design variants that were synthesized, in a csv format. The tool also supports an IUPAC description (Johnson, 2010) of the design.
2. NGS results files: In fastq format and containing all resulting the NGS reads.
3. Library configuration: Auxiliary parameters which consist of other details about the design variants such as information regarding the barcodes etc.

Detailed description of SOLQC report, installation instructions, example data and an example report file are available as Supplementary Information and on SOLQC website: https://yoav-orlev.gitbook.io/solqc/.

2.2 SOLQC workflow

SOLQC includes a complete analysis pipeline where each step can be customized using the configuration file:

1. **Preprocessing:** The reads can be filtered so that only valid reads will be processed by the tool. The selection of valid reads can be configured by the user according to the sequence barcode and its length.
2. **Matching**: Each read is matched to its corresponding variant. The set of reads which are matched to the same variant form a *variant cluster*. The matching parameters are defined by the user. Matching can be based on barcode sequences included in the variants, full-length edit distance calculation or approximation. Alternatively, matching information can be given by the user.
3. **Alignment**: Every read is aligned according to its matched variant and an error vector is computed which represents the inferred error types at each position of the variant.
4. **Analysis**: The matched reads and their error vectors are used to characterize errors and produce data statistics for the library.
5. **Report generation:** The output of our tool is a report which consists of analysis results, as selected by the user, in a customizable format.

2.3 Statistical QC analysis for synthetic DNA libraries

SOLQC analyses are divided into two parts; the first one addresses the composition of the synthesized library (composition statistics), and the second one addresses the errors inferred from the NGS reads (error statistics). Detailed description of the analyses, as well as their figures, can be found in Supplementary Material Document.

2.3.1 Composition statistics

1. **Symbol composition.** Distribution of all bases by position in the sequence. This is calculated based on the design and the observed reads separately.
2. **Read length.** Distribution of the read lengths observed in the NGS data compared to the design variant lengths.
3. **Variant cluster size.** Distribution of the variant cluster sizes (i.e. the number of filtered reads matched for each design variant). Variants can be stratified by different parameters such as GC content (Fig. 1i).

2.3.2 Error statistics

1. **Total error rates.** Insertion, substitution, single base deletion and long deletion rates inferred from aligning the reads to the matching variants (Fig. 1ii).
2. **Error rates per position.** Insertion, substitution, single base deletion and long deletion rates as a function of the position in the variant (Fig. 1iii).
3. **Error rates stratified by symbol.** Symbol-dependent error rates (Fig. 1iv).

Downloaded from https://academic.oup.com/bioinformatics/article/37/5/720/5896982 by Department of mathematics user on 03 October 2021

4. **Deletion length distribution.** Distribution of the lengths of all deletions.
5. **Error rates stratified by GC content.** Distribution of the error rates stratified by the GC-content of the matched variant.

2.4 Use-case examples

SOLQC can be useful in the following use-case examples:

1. **Comparison of different OL designs.** Many design parameters may affect the quality of the synthesized OL. To asses these parameters, the user may try out different mini-libraries and compare their quality and error rates using SOLQC.
2. **Binning of OL NGS output.** SOLQC includes various methods for matching NGS reads to design variants. Users may incorporate SOLQC in their analysis pipeline and get the matching results together with coverage and quality statistics.
3. **Assessment of synthesis and sequencing technologies**. Users may choose to synthesize or sequence OL using different platforms to compare quality performance. SOLQC can serve as standard tool in such comparisons. A comparison between four OLs can be found in Supplementary Material.
4. **Design of error-correcting codes and coding techniques for DNA storage.** In data storage applications, SOLQC can be used as a characterization tool of the DNA channel. Using this information, the user can design appropriate error-correcting codes and coding techniques to improve the error rates.
5. **Standardization and reproducibility.** SOLQC helps detecting whether a library is behaving as previous libraries from the same vendor with similar preparation characteristics. Thus, SOLQC supports uniformity for OLs use in different labs, or in the same lab at different times or by different lab members.

Financial Support: This project has received funding from the European Union's Horizon 2020 Research And Innovation Programme under grant agreement No. 851018. L. Anavy is supported by the Adams Fellowships Program of the Israel Academy of Sciences and Humanities..

Conflict of Interest: none declared.

References

Anavy,L. *et al.* (2019) Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nat. Biotechnol.*, **37**, 1229–1236.

Andrews,S. *et al.* (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Blawat,M. *et al.* (2016) Forward error correction for DNA data storage. *Proc. Comput. Sci.*, **80**, 1011–1022.

Church,G.M. *et al.* (2012) Next-generation digital information storage in DNA. *Science*, **337**, 1628–1628.

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

Goldman,N. *et al.* (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, **494**, 77–80.

Grass,R.N. *et al.* (2015) Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem. Int. Ed.*, **54**, 2552–2555.

Heckel,R. *et al.* (2019) A characterization of the DNA data storage channel. *Scientific Reports, 9, 9663*.

Johnson,A.D. (2010) An extended IUPAC nomenclature code for polymorphic nucleic acids. *Bioinformatics*, **26**, 1386–1389.

Kosuri,S. and Church,G.M. (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods*, **11**, 499–507.

Kotler,E. *et al.* (2018) A systematic p53 mutation library links differential functional impact to cancer mutation pattern and evolutionary conservation. *Mol. Cell*, **71**, 178–190.

Levy,L. *et al.* (2017) A synthetic oligo library and sequencing approach reveals an insulation mechanism encoded within bacterial σ 54 promoters. *Cell Rep.*, **21**, 845–858.

Miles,L.A. *et al.* (2016) Design, execution, and analysis of pooled in vitro CRISPR/Cas9 screens. *FEBS J.*, **283**, 3170–3180.

Organick,L. *et al.* (2018) Random access in large-scale DNA data storage. *Nat. Biotechnol.*, **36**, 242–248.

Pan,W. *et al.* (2014) DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnol.*, **14**, 10.

Ruijter,J. *et al.* (2009) Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.*, **37**, e45–e45.

Sharon,E. *et al.* (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, **30**, 521–530.

Tian,J. *et al.* (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.

Yazdi,S.H.T. *et al.* (2017) Portable and error-free DNA-based data storage. *Sci. Rep.*, **7**, 5011.

Downloaded from https://academic.oup.com/bioinformatics/article/37/5/720/5896982 by Department of mathematics user on 03 October 2021