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]Computer Science Department, Technion, Haifa, 3200003, Israel.
[2]School of Computer Science, Herzliya Interdisciplinary Center, Herzliya, 4610101, Israel.
[*]The two first authors contributed equally to this work

Abstract—**DNA-based storage has attracted significant attention due to recent demonstrations of the viability of storing information in macromolecules using synthetic oligo libraries. As DNA storage experiments, as well as other experiments of synthetic oligo libraries are growing by numbers and complexity, analysis tools can facilitate quality control and help in assessment and inference. We present a novel analysis tool, called *SOLQC*, which enables fast and comprehensive analysis of synthetic oligo libraries, based on next generation sequencing (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 descriptions of the analysis results. The results are reported in a flexible report format. We demonstrate SOLQC by analyzing literature libraries. We also discuss the potential benefits and relevance of the different components of the analysis.**

I. Introduction

The recent progress in DNA synthesis and sequencing technologies has paved the way for the development of data storage technology based upon DNA molecules. A DNA storage system consists of three important components. The first is the DNA synthesis which produces the *oligonucleotides*, also called *strands* or *variants*, that encode the data. In order to produce strands with acceptable error rates, in a high throughput manner, the length of the strands is typically limited to no more than 250 nucleotides [1]. The second part is a storage container with compartments which stores the DNA strands, however without order. Finally, sequencing is performed to read back a representation of the strands, which are called *reads*.

The encoding and decoding stages are two processes, external to the storage system, that convert the users binary data into strands of DNA such that, even in the presence of errors, it will be possible to revert back and reconstruct the original binary data. The processes of synthesizing, storing, and sequencing the strands are all error prone. Each step in the process can independently introduce a significant number of errors, mostly of three types: deletion, insertion, and substitution. Since DNA storage, as well as any other storage channel, is a noisy information system, a mandatory step is to conduct a comprehensive characterization and analysis of those errors.

Most of the research on characterizing errors in synthetic DNA libraries has been done in the context of individual studies using synthetic DNA. Recently, in [4], Heckel, Mikutis, and Grass, studied the errors in a DNA storage channel based upon three different data sets from the experiments in [2], [3], [4]. In their work they studied the deletion/insertion/substitution rates and how it is affected by filtering reads with incorrect length (compared to the designed length). In particular, when they considered only reads with the correct length, they showed, as expected, that the deletion rate has been significantly decreased in all of the data sets.

A comprehensive summary of the previous works in DNA-storage, as well as the work regarding errors in DNA oligo libraries can be found in our full paper [5]. The tool is open software and available for the community. Installation instructions can be found in SOLQC website

II. SOLQC Tool

In this work we present our software tool, called **SOLQC - Synthetic Oligo Library Quality Control**. The tool is designed to enable and to facilitate individual labs obtaining information about DNA libraries and performing error analysis before or during experiments. In [5] we describe our methods and demonstrate the results of analyzing several libraries.

SOLQC generates a customized report consisting of several statistics and plots for a given input synthetic DNA library. The input to the SOLQC tool is the result of a sequencing reaction run on the library. It consists of the design variants and of all the sequenced reads and is operated in the following order.

1) **Preprocessing**: The reads can be filtered such that only valid reads will be processed by the tool.
2) **Matching**: Each read is matched to its corresponding variant. The matching step can be done by different strategies and approximations as described in [5].
3) **Alignment**: Every read is aligned according to its matched variant.
4) **Analysis**: The matched reads and their alignments are used in order to create error characterization and data statistics for the library.
5) **Report generation**: The output of our tool is a report which consists of the analysis results.

III. QC Analysis for Synthetic DNA Libraries

In this section we describe and discuss the statistical analysis performed and supported by the SOLQC tool. These statistics are explained on actual data from the experiment in [2] by Erlich and Zielinski. The details of this experiment are summarized in Table I. These statistical results 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 sequencing reads (error statistics). We sampled 1,689,319 out of the 15,787,115 reads of the library, and analyzed only reads with length at most 4 bases shorter or longer than the design's length, which is 152 (i.e., their base-length was between 148 and 156). Those reads were matched with their closest design variants using an approximation of the edit distance which calculated the edit distance between all reads and variants based upon the first 80 bases.

Table I
Experiment by Erlich and Zielinski [2]

Data size	2.11 MB
Design length	152 bases
Number of variants	72000
Number of reads	15,787,115
Number of sampled reads	1,689,315
Number of filtered reads	1,427,781
Synthesis Technology	Twist Bioscience
Sequencing Technology	Ilumina miSeq V4

1) **Total error rates** (Fig. 1). This plot presents the insertion, substitution, and deletion error rates as inferred from the reads in the library. Each bar in the X-axis presents the error type. The Y-axis presents the error rate (in log scale), calculated as the ratio between the total number of errors of each type and the total number number of read bases.

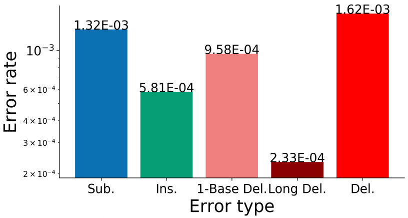

Figure 1. Total error rates.

2) **Histogram of the cluster size per variant** (Fig. 2). The plot in Fig. 2 presents the histogram of the variant cluster size, which is the number of filtered reads, per design variant. The X-axis presents the size of a variant cluster, starting from the size of the smallest variant cluster among all the variants in the library and up to the largest variant cluster value. The Y-axis presents the number of variants in the library that have a cluster of size x. According to the matching step, the cluster size for each of the design variants is calculated and the histogram is generated by counting the number of variants with a given cluster size. Note that the sum of the y values in this histogram is the number of variants in the experiment, which is 72,000 in [2].

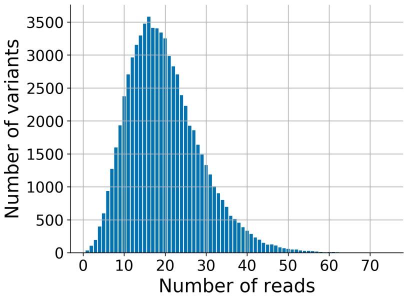

Figure 2. Histogram of the number of filtered reads per variant.

3) **Error rate stratified by symbol** (Fig. 3). This plot presents by a heat map the symbol dependent, error distribution. Each square presents for each type of error, its error rate for the specific symbol. For insertions we address both the inserted symbol, and the symbol before the insertion. The x, y entry in the heat map is calculated to be the ratio between the number of type y errors of base x and the expected number of base x in the reads[1].

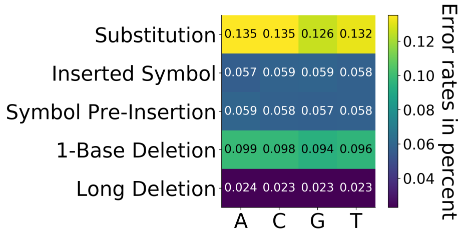

Figure 3. Error rates stratified by symbol. Note that the numbers are in percents, e.g. 0.024 for "A" long deletion, means that 0.024 percents of the occurrences of base A in the library creates long deletion error.

4) **Error rate per position** (Fig. 4). This plot presents the error rate for every error type as it is reflected in a specific position of the strand. The X-axis presents the position in the strand, from 5' to 3'; note that the phosphoramidite synthesis direction is 3' to 5'. It is important to emphasize that we report rates as calculated from the alignment results. These rates reflect both synthesis as well as sequencing errors. The Y-axis presents the error rates per position in all reads in log scale. For every position between 0 (the first position, from 5' to 3') and 151 (the last position in [2]) and for each error type as described in Fig. 1, the tool calculates the error rate as the ratio between the number of errors of each type and the number of filtered reads.

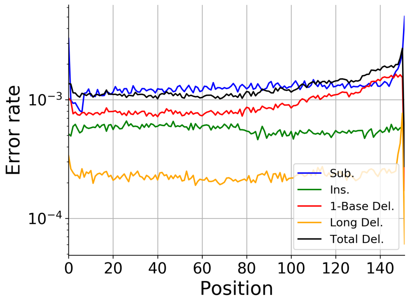

Figure 4. Error rates by position. X-axis represents position counted from the 5' end of the designed variant. The Y-axis is log-scale.

IV. Use-Case Examples

Two use-case examples for SOLQC (more examples can be found in the paper):

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. The user can characterize the DNA channel using data from previous experiments of various technologies and design parameters. Then, using this information, the user can design appropriate error-correcting codes and coding techniques to improve the error rates.

Binning of synthetic DNA-libraries. The result of a sequencing reaction on a given library does not include the matching of each read to its design variant. SOLQC provides several methods to bin the reads according to their corresponding design variants. The matching/clustering methods can be performed on libraries with or without the barcode. In addition, users can get coverage depth statistics from SOLQC as well as quality related statistics, which can be different for different variants or set of variants. Lastly, in applications like data storage, the set of reads that is binned to any given variant can be used in order to decode the stored variant.

References

[1] S. L. Beaucage and R. P. Iyer. Advances in the synthesis of oligonucleotides by the phosphoramidite approach. *Tetrahedron*, 48(12):2223 – 2311, 1992.

[2] Y. Erlich and D. Zielinski. DNA fountain enables a robust and efficient storage architecture. *Science*, 355(6328):950–954, 2017.

[3] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494(7435):77, 2013.

[4] R. Heckel, G. Mikutis, and R. N. Grass. A characterization of the DNA data storage channel. *arXiv:1803.03322*, 2018.

[5] O. Sabary, Y. Orlev, R. Shafir, L. Anavy, E. Yaakobi, and Z. Yakhini. SOLQC : Synthetic oligo library quality control tool. *bioRxiv*, 2019.

[1] The expected number of base x in the reads is calculated as the sum of the products of the number of base x in each of the design variants, and the number of reads matched to it.