SOLQC
Synthetic Oligo Library Quality Control Tool

Omer Sabary, Yoav Orlev, Roy Shafir, Leon Anavy, Eitan Yaakobi, Zohar Yakhini
DNA-Storage – Why?

Stability –
DNA can still recovered from 700,000 years old horse!
DNA-Storage – Why?

Capacity –
DNA is extremely dense.
$10^9$ GB /mm$^3$

Cost decreasing –
DNA write (synthesis) and read (sequencing) costs are decreasing daily
DNA Storage Systems

- Blawat, Gaedke, Hutter, Chen, Turczyk, Inverso, Pruitt, and Church, *Forward error correction for DNA data storage*. Int. Conf. on Computational Science, 2016.
- Helixworks: 2016, first commercially available DNA storage medium.
DNA Storage Systems

- Takahashi, Nguyen, Strauss, and Ceze, **Demonstration of end-to-end automation of DNA data storage.** Scientific Reports, 2019.
- Anavy, Vaknin, Atar, Amit, and Yakhini, **Improved DNA based storage capacity and fidelity using composite DNA letters.** Nature Biotechnology, 2019.
- DNA Catalog: **2019, the first to store 16GB of data.**
- Iridia: **2019, complete DNA storage system on a chip.**
DNA Storage

Decoding

User Binary Data
100011000101001011110101
101110011001010111110100
001110101010001010010011

Encoding

DNA Strands
ACTGAGTCAGTGACGTGATGCA
CTGAGATGAGTGCAGTGCAGTCTT
TCGTGCAGTCATGTCGTGCCTT

DNA Strands
ACTGGGTCACTGACGTGACGTGCA
CTGAGATGAGTGCAGTGCAGTCTT
TCGTGCAGTCATGTCGTGCCTT

DNA Sequencer

DNA Synthesizer

Storage Container

Read

Write
DNA Intro

- DNA consists of 4 bases, aka nucleotides:
  - Adenine
  - Cytosine
  - Guanine
  - Thymine

- DNA strand, aka oligonucleotide, is a string of the nucleotides

- C&G are complementary and A&T
  - Each strand can bond its complementary strand
  - Two strands can bind if they are complementary
How to Write Data into DNA?

- Convert a binary sequence into a quaternary sequence
  - \(A = 00\), \(C = 01\), \(G = 10\), \(T = 11\)
  - 01.00.11.10.00.00.01.10.11

- However...
  - Strands are limited in their size (~200 bases)
  - Strands are not ordered (a soup with many strands)
How to Write Data into DNA?

• **DNA Synthesis**: artificially generating DNA strands
  • Strands are generated by appending one base at a time
  • Typical lengths are ~200 bases (due to technology limitations)
  • Each strand has thousands copies

• **DNA Sequencing**: reading DNA strands
  • Generating many reads of each strand
  • Less expensive and faster than synthesis (per base)
How to Write Data into DNA?

• Parse the file to strings of bits
• Each string is converted to a DNA strand with index and primer
DNA Storage Channel Model
DNA Storage Channel Model

Decoding
- User Binary Data
  10010110001110011011010110011010011010010000000000001
- DNA Strands
  ACTAGTAATATAATGATCGTGTA
  ACTGGAAACACTCGTCAGCTACCTCTC
  DNA Synthesizer
- DNA Sequencer
  GAGTAGACGTACCTGCAGCTACCTCTC

Encoding
- DNA Strands
  ACTAGTAATATAATGATCGTGTA
  ACTGGAAACACTCGTCAGCTACCTCTC
- DNA Synthesizer
  DNA Sequencer

Primer Address
- ACTGG.AAAA
- ACTGG.AAAC
- ACTGG.AAAC
- ACTGG.AAGA
- ACTGG.AAAAT
- ACTGG.AACA
- ACTGG.AACA
- ACTGG.AACCT
- ACTGG.AACCT
- ACTGG.AAAG
- ACTGG.AAAT
- ACTGG.AAAC

Primer
- TGGTGAAATAAGCTAGCTACCTGCAGCTACCTCTC
- TGGTGAAATAAGCTAGCTACCTGCAGCTACCTCTC
- TGGTGAAATAAGCTAGCTACCTGCAGCTACCTCTC
- TGGTGAAATAAGCTAGCTACCTGCAGCTACCTCTC
- TGGTGAAATAAGCTAGCTACCTGCAGCTACCTCTC
- TGGTGAAATAAGCTAGCTACCTGCAGCTACCTCTC
- TGGTGAAATAAGCTAGCTACCTGCAGCTACCTCTC

Storage Container

14
DNA Storage Channel Model

Clustering
DNA Storage Channel Model

Reconstruction
DNA Storage Channel Model

ECC

Primer Address

- ACTGG.AAAA.ACTGGTAATATATAATGTCCGTGCCGTA.TGCAA
- ACTGG.AAAC.ACGTGGTCAAGTACGTTGACGTACTC.TGCAA
- ACTGG.AAAG.ACGTACGTGTGCAGAACATGACCAGTG.TGCAA
- ACTGG.AAAT.AAGGTTGTGTCCAGATGACGTGATG.TGCAA
- ACTGG.AACA.TGCATGCAAGTGCAGATGCTAATG.TGCAA
- ACTGG.AACC.TTTGTTGAAACATGCACTGATGAACTG.TGCAA
- ACTGG.AACG.AAGTACCAGTACTATGCGTGACGT.TGCAA
- ACTGG.AACT.AGTTACGTGCTGCTAAGTACGTGT.C.TGCAA

Storage Container
Errors in DNA

**Synthesis**

- Mostly for chemical reasons
- Each copy of a certain sequence has different errors

**PCR**

- Creates a bias - prefers one sequence over another

**Sequencing**

- Higher GC Content affects sequencing error
- Presence of Homopolymers increases the error rate
Errors in DNA

Design:

Copies:
Errors in DNA

Design:

| C | A | T | G | A | A | C | G | T |

Copies:

| C | A | T | A | A | C | G | T |

Deletion
Errors in DNA

Design:

Copies: Insertion
Errors in DNA

Design:

C A T G A A C G T

Copies:

C A T A A C G T
C A T G A G A C G T
C A T G A T C G T
C A T G A T C G T

Substitution
Errors in DNA

Design:

Copies:

Deletion

Insertion
Errors in DNA

Design:

Copies:

- Deletion
- Insertion
- Substitution
- Deletion
- Insertion
Error Characterization

Input

User Binary Data
100110001010011110101
1011100110010111110100
00111010101000101001011

Variants

DNA Strands
ACTGGGTCAGTGACGTGCATGCA
CTGAGATGCAGTGAGTGCAGCTT
TCGTGCAGTGATGTCGTGCATGC

DNA Synthesizer

Substitution
Insertion
Deletion

Reads

DNA Strands
ACTG\_GTCAGTGACGTGCATGCA
CTGAGATG\_TAGTGAGTGCGACGCTT
TCGTGCAGTGATGCTCGTGCATGC

DNA Sequencer

Output

User Binary Data
100110001010011110101
1011100110010111110100
00111010101000101001011

Decode

Storage Container
Error Characterization

**Input**
Synthetic DNA library:
- Design variants
- NGS results.

**Step 0 – Preprocessing**
Filtering invalid sequences by their length.

**Step 1 – Clustering**
Matching each read with its design variant.

**Step 2 – Alignment**
Calculation the alignment path of each read vs. variant.

**Step 3 – Analysis**
Characterization and analysis of the errors in the library.

**Output**
Quality report consisting of plots and statistical values.
SOLQC Pipeline

Matching - Clustering
The set of reads which are matched to the same variant forms a variant cluster.
SOLQC Pipeline

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.
SOLQC Pipeline

Input → Preprocessing → Matching → Alignment → Analysis

Variant

<table>
<thead>
<tr>
<th>Read #1</th>
<th>Read #2</th>
<th>Read #3</th>
<th>Read #4</th>
<th>Read #5</th>
<th>Read #6</th>
<th>Read #7</th>
<th>Read #8</th>
<th>Read #9</th>
<th>Read #10</th>
<th>Read #11</th>
<th>Read #12</th>
<th>Read #13</th>
<th>Read #14</th>
<th>Read #15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td>D</td>
<td>D</td>
<td></td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td>D</td>
<td></td>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td></td>
<td></td>
<td>D</td>
<td></td>
<td>S</td>
<td></td>
<td></td>
<td>D</td>
<td></td>
<td></td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend:**
- **D** - Deletion
- **S** - Substitution
- **I** - Insertion
The matched reads and their error vectors are used in order to create error characterization and data statistics for the library, as will be described in the sequel.
## Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage size</td>
<td>81KB</td>
<td>2.11 MB</td>
<td>200 MB (9.5 MB)</td>
<td>3.633 KB</td>
</tr>
<tr>
<td>Design length</td>
<td>158</td>
<td>152</td>
<td>150</td>
<td>880-1,060</td>
</tr>
<tr>
<td># variants</td>
<td>5,000</td>
<td>72,000</td>
<td>607,150</td>
<td>17</td>
</tr>
<tr>
<td># reads</td>
<td>3,312,235</td>
<td>15,787,115</td>
<td>62,879,612</td>
<td>6,660</td>
</tr>
<tr>
<td># filtered reads</td>
<td>1,945,744</td>
<td>1,427,781</td>
<td>91,898</td>
<td>6,660</td>
</tr>
</tbody>
</table>

**Synthesis**  
CustomArray, GenScript  
**Sequencing**  
illumina, Oxford Nanopore Technologies


Histogram of cluster size per variant

Erlich & Zielinski
Sorted bar plot of the number of filtered reads per variant

Erlich & Zielinski
Total error rates

Grass et al.

Error rate

- Substitution
- Insertion
- 1-Base Del.
- Long Del.
- Deletion

Organick et al.

Error rate

- Substitution
- Insertion
- 1-Base Del.
- Long Del.
- Deletion

Erlich & Zielinski

Error rate

- Substitution
- Insertion
- 1-Base Del.
- Long Del.
- Deletion

Yazdi et al.

Error rate

- Substitution
- Insertion
- 1-Base Del.
- Long Del.
- Deletion
Error rates, stratified by symbol

Yazdi et al.

Error rates in percent

Inserted Sym.

Substitution

Sym. Pre-Ins.

1-Base Del.

Long Del.

A  C  G  T
Error rates, stratified by symbol

Organick et al.

<table>
<thead>
<tr>
<th>Error Rate (in percent)</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution</td>
<td>0.724</td>
<td>0.701</td>
<td>0.706</td>
<td>0.704</td>
</tr>
<tr>
<td>Inserted Sym.</td>
<td>0.411</td>
<td>0.415</td>
<td>0.415</td>
<td>0.413</td>
</tr>
<tr>
<td>Sym. Pre-Ins.</td>
<td>0.429</td>
<td>0.415</td>
<td>0.403</td>
<td>0.408</td>
</tr>
<tr>
<td>1-Base Del.</td>
<td>0.289</td>
<td>0.279</td>
<td>0.276</td>
<td>0.28</td>
</tr>
<tr>
<td>Long Del.</td>
<td>0.048</td>
<td>0.048</td>
<td>0.047</td>
<td>0.049</td>
</tr>
</tbody>
</table>
Cumulative distribution based upon the number of errors

- Erlich & Zielinski
- Grass et al.
- Organick et al.
- Yazdi et al.
Histogram of the length of the reads
Erlich & Zielinski

Unfiltered

Filtered

Fraction of reads

Length

10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1}

50 171 292

Fraction of reads

10^{-4} 10^{-3} 10^{-2} 10^{-1}

148 152 156

Length
Error rates per position

Erlich & Zielinski

Grass et al.

Organick et al.

Yazdi et al.
Error rates stratified by GC-content

Erlich & Zielinski

Organick et al.

Erlich & Zielinski

Organick et al.
Error rates stratified by GC-content

Grass et al.

Yazdi et al.
Thank You!

- Design
- NGS Files
- Library Configuration
- Analysis Configuration

If you use this tool please cite this paper.