11TH ANNUAL NON-VOLATILE MEMORIES WORKSHOP

20
NVMW

# SOLQC

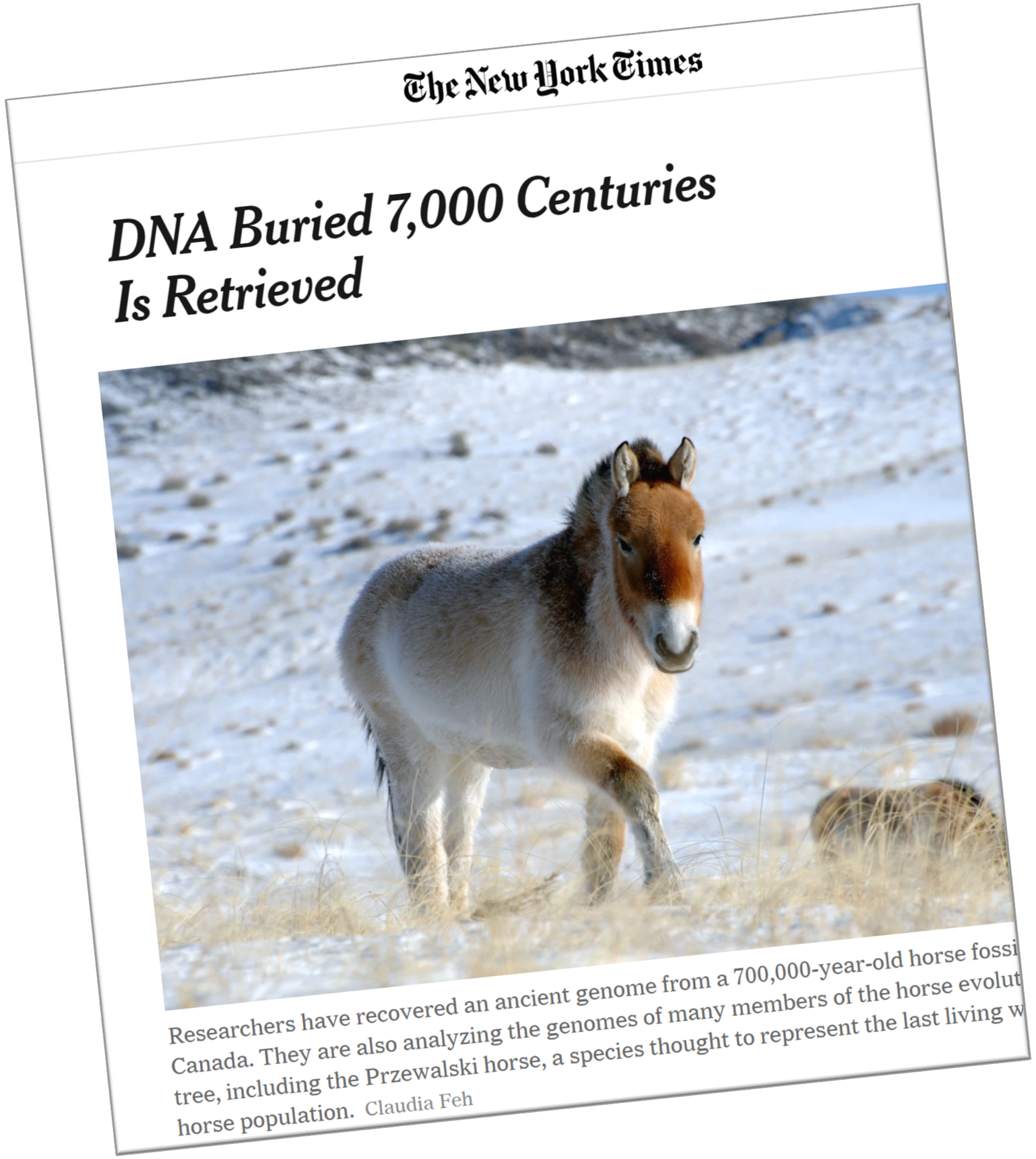
## Synthetic Oligo Library Quality Control Tool

*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!**



The New York Times

## DNA Buried 7,000 Centuries Is Retrieved

Researchers have recovered an ancient genome from a 700,000-year-old horse fossi
Canada. They are also analyzing the genomes of many members of the horse evolut
tree, including the Przewalski horse, a species thought to represent the last living w
horse population. Claudia Feh

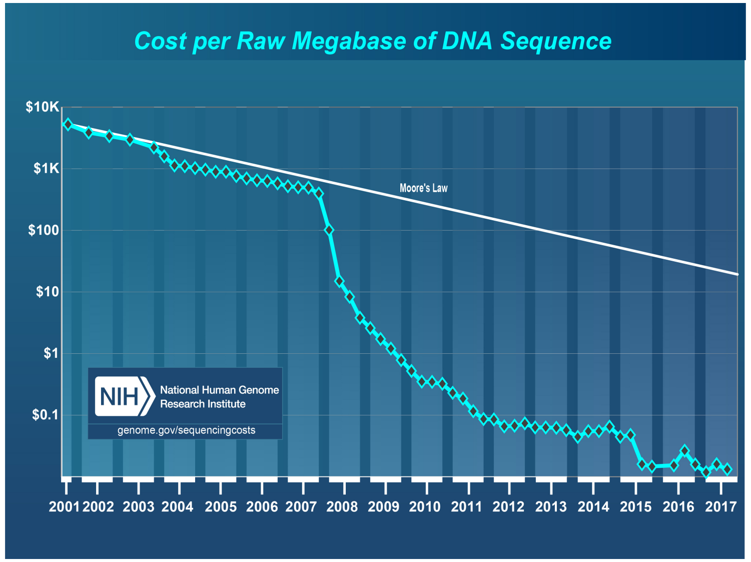
# DNA-Storage – Why?

## Capacity –

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

## Cost decreasing –

DNA write (synthesis) and read (sequencing) **costs are decreasing daily**



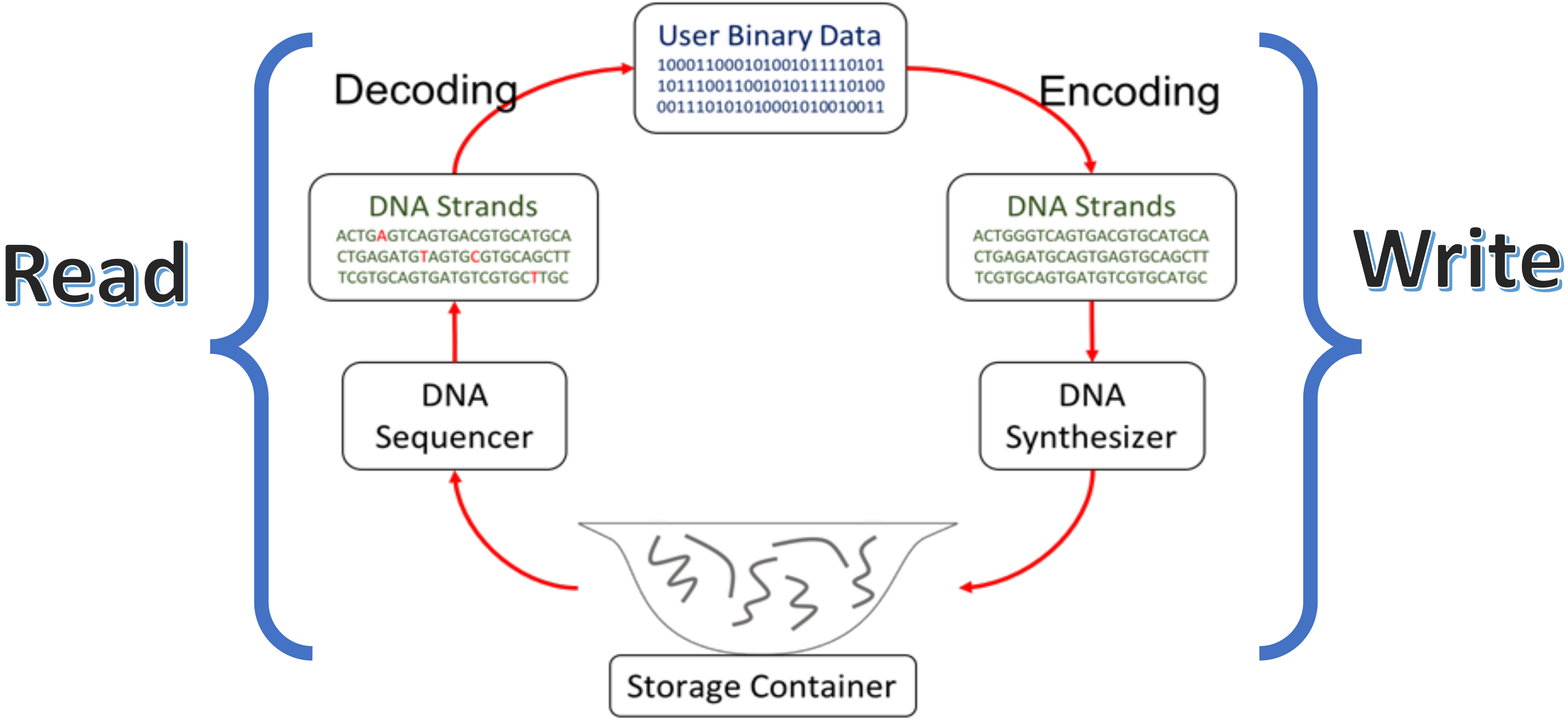
# DNA Storage Systems

- Feynman, **There's plenty of room at the bottom**. Engineering and Science, California Institute of Technology, **1960**.
- Church, Gao, and Kosuri, **Next-generation digital information storage in DNA**. Science, **2012**.
- Goldman, Bertone, Chen, Dessimoz, LeProust, Sipos, and Birney, **Towards practical, high-capacity, low-maintenance information storage in synthesized DNA**. Nature, **2013**.
- Grass, Heckel, Puddu, Paunescu, and Stark, **Robust chemical preservation of digital information on DNA in silica with error-correcting codes.** Angewandte Chemie International Edition, **2015**.
- Yazdi, Kiah, Garcia-Ruiz, Ma, Zhao and Milenkovic**, DNA-based storage: Trends and methods**. IEEE Trans. on Molecular, Biological and Multi-Scale Communications, **2015**.
- Bornholt, Lopez, Carmean, Ceze, Seelig, and Strauss, **A DNA-based archival storage system**. ASPLOS, **2016**.
- 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.**
- Erlich and Zielinski, **DNA fountain enables a robust and efficient storage architecture**. Science, **2017**.

# DNA Storage Systems

- Yazdi, Gabrys, and Milenkovic**. Portable and error-free DNA-based data storage**. Scientific Reports, **2017**.
- Heckel, Mikutis, and Grass. **A characterization of the DNA data storage channel.** *arXiv preprint*, **2018**.
- Organick, Ang, Chen, Lopez, Yekhanin, Makarychev, Racz, Kamath, Gopalan, Nguyen, Takahashi, Newman, Parker, Rashtchian, Stewart, Gupta, Carlson, Mulligan, Carmean, Seelig, Ceze, and Strauss. **Random access in large-scale DNA data storage.** Nature Biotechnology, **2018**.
- Gopalan, Yekhanin, Ang, Jojic, Racz, Strauss, and Ceze**. Trace reconstruction from noisy polynucleotide sequencer reads**, US Patent App **2018**.
- Takahashi, Nguyen, Strauss, and Ceze, **Demonstration of end-to-end automation of DNA data storage**. Scientific Reports, **2019**.
- Tabatabaei, Wang, Athreya, Enghiad, Hernandez, Leburton, Soloveichik, Zhao, and Milenkovic, **DNA punch cards: Encoding data on native DNA sequences via topological modifications**. BioRxiv, **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



Read

Decoding

User Binary Data

1000110001010010111110101
1011100110010101111110100
001110101010001010010011

Encoding

Write

DNA Strands

ACTGAGTCAGTGACGTGCATGCA
CTGAGATGTAGTGCGTGCAGCTT
TCGTGCAGTGATGTCGTGCTTGC

DNA Strands

ACTGGGTCAGTGACGTGCATGCA
CTGAGATGCAGTGAGTGCAGCTT
TCGTGCAGTGATGTCGTGCATGC

DNA Sequencer

DNA Synthesizer

Storage Container

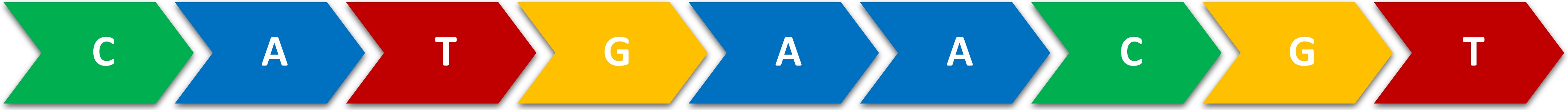
# DNA Intro

- DNA consists of 4 bases, aka nucleotides:

Adenine Cytosine Guanine Thymine

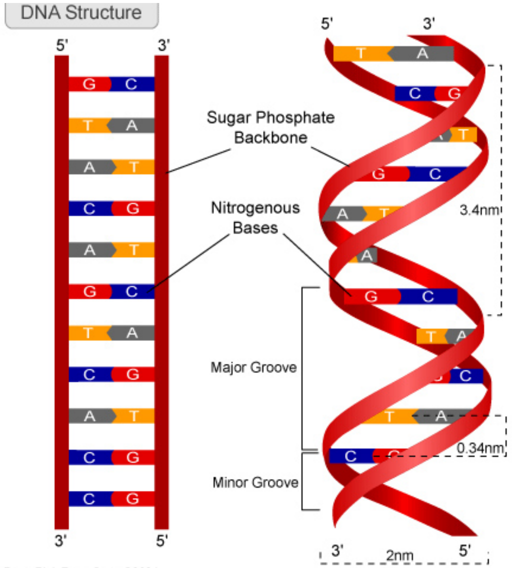
A C G T

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



C A T G A A C G T

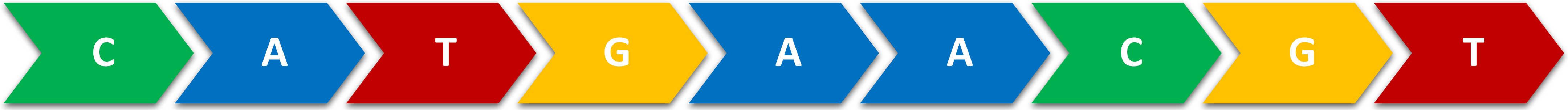
- C&G are complementary and A&T
  - Each strand can bond its complementary strand
  - Two strands can bind if they are complementary



Dept. Biol. Penn State ©2004

# 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

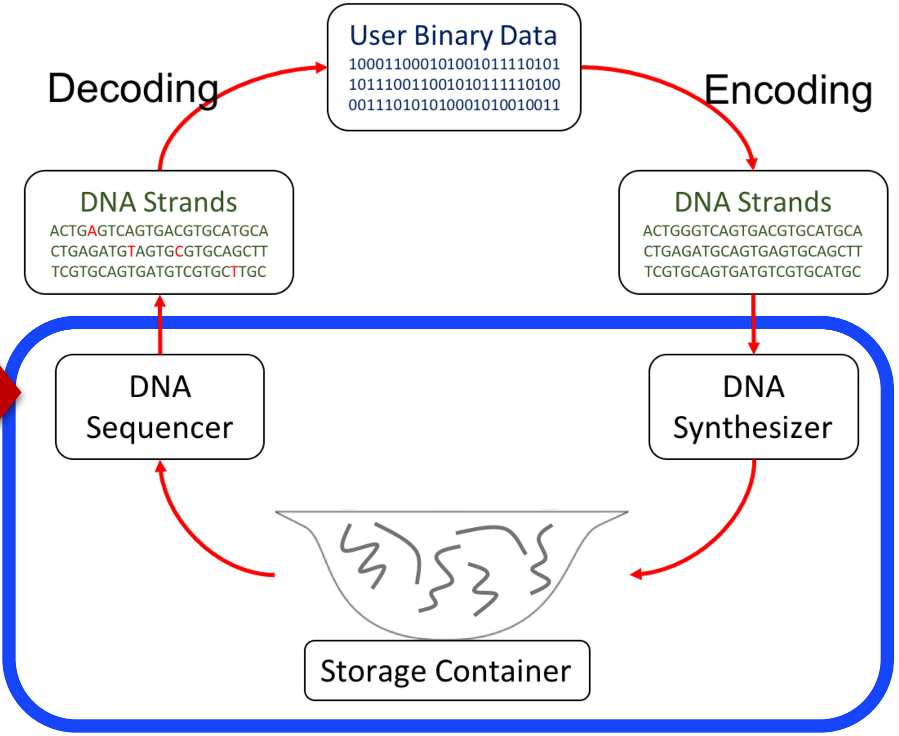


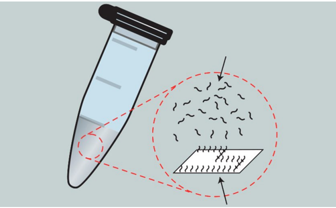
- However…
  - Strands are limited in their size (**~200** bases)
  - Strands are not ordered (a soup with many strands)

# 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

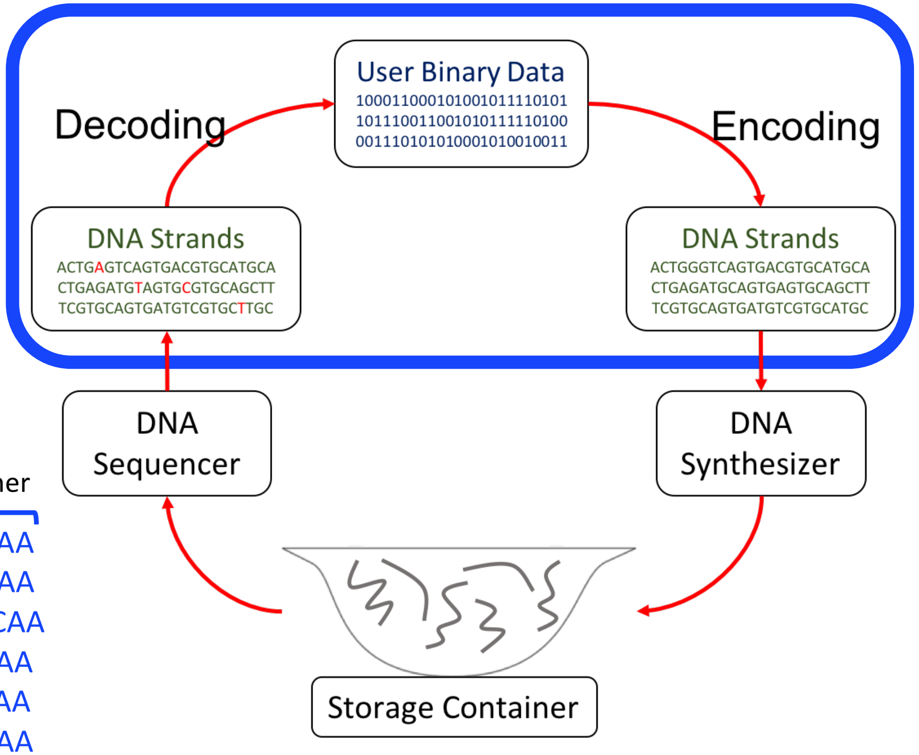
- **DNA Sequencing**: reading DNA strands
  - Generating many reads of each strand
  - Less expensive and faster than synthesis (per base)

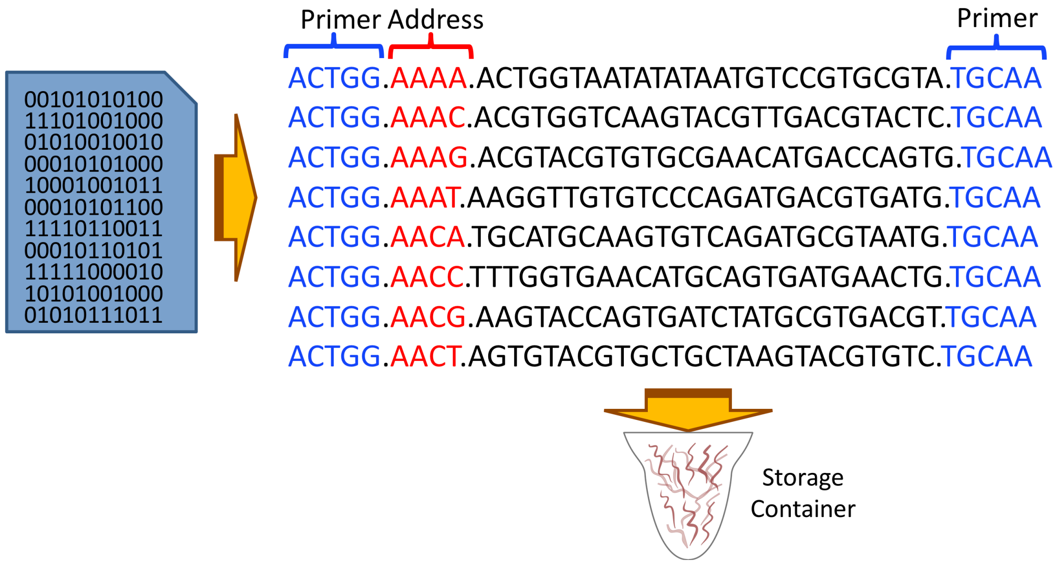




# How to Write Data into DNA?

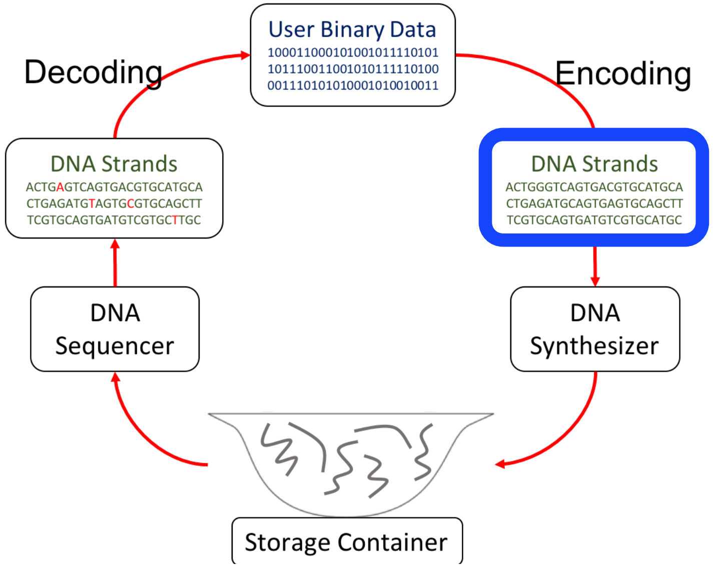
- Parse the file to strings of bits
- Each string is converted to a DNA strand with index and primer

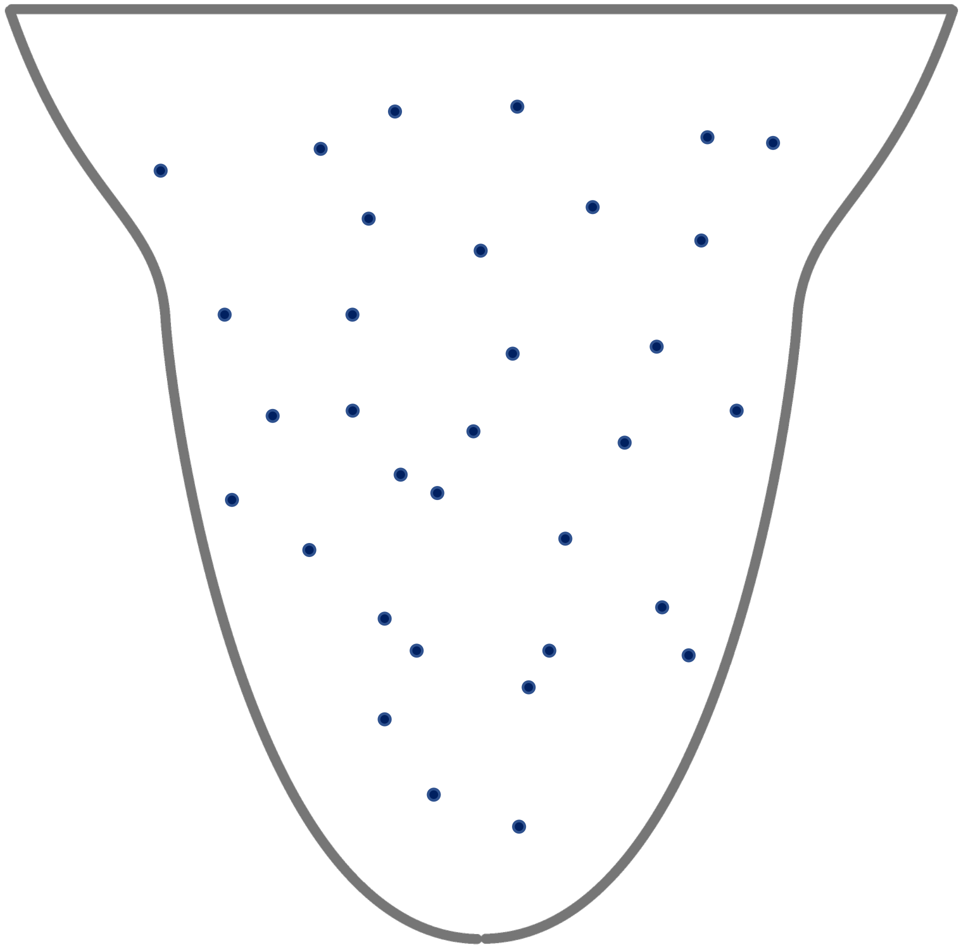


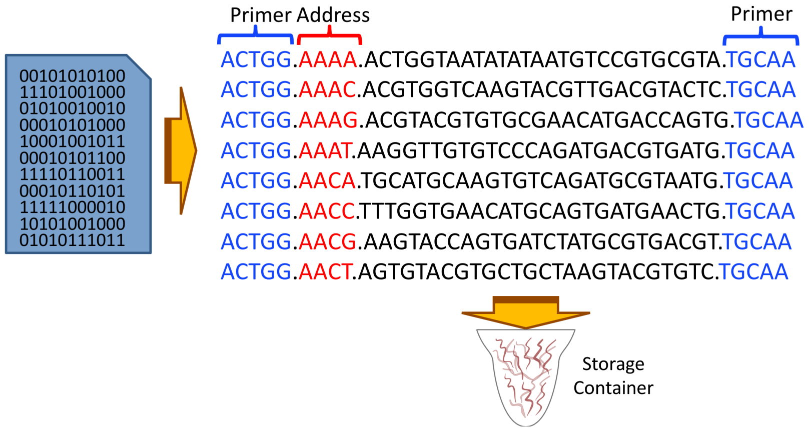


```
ACTGG.AAAA.ACTGGTAATATATAATGTCCGTGCGTA.TGCAA
ACTGG.AAAC.ACGTGGTCAAGTACGTTGACGTACTC.TGCAA
ACTGG.AAAG.ACGTACGTGTGCGAACATGACCAGTG.TGCAA
ACTGG.AAAT.AAGGTTGTGTCCCAGATGACGTGATG.TGCAA
ACTGG.AACA.TGCATGCAAGTGTCAGATGCGTAATG.TGCAA
ACTGG.AACC.TTTGGTGAACATGCAGTGATGAACTG.TGCAA
ACTGG.AACG.AAGTACCAGTGATCTATGCGTGACGT.TGCAA
ACTGG.AACT.AGTGTACGTGCTGCTAAGTACGTGTC.TGCAA
```

# DNA Storage Channel Model

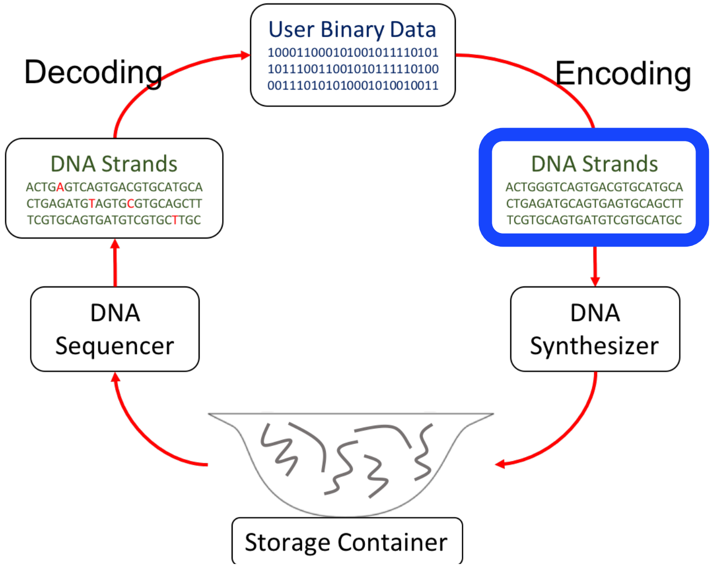






ACTGG.AAAA.ACTGGTAATATATAATGTCCGTGCGTA.TGCAA
ACTGG.AAAC.ACGTGGTCAAGTACGTTGACGTACTC.TGCAA
ACTGG.AAAG.ACGTACGTGTGCGAACATGACCAGTG.TGCAA
ACTGG.AAAT.AAGGTTGTGTCCCAGATGACGTGATG.TGCAA
ACTGG.AACA.TGCATGCAAGTGTCAGATGCGTAATG.TGCAA
ACTGG.AACC.TTTGGTGAACATGCAGTGATGAACTG.TGCAA
ACTGG.AACG.AAGTACCAGTGATCTATGCGTGACGT.TGCAA
ACTGG.AACT.AGTGTACGTGCTGCTAAGTACGTGTC.TGCAA

# DNA Storage Channel Model



User Binary Data
10001100010100101111010 1
10111001100101011111010 0
00111010101000101001001 1

Decoding

Encoding

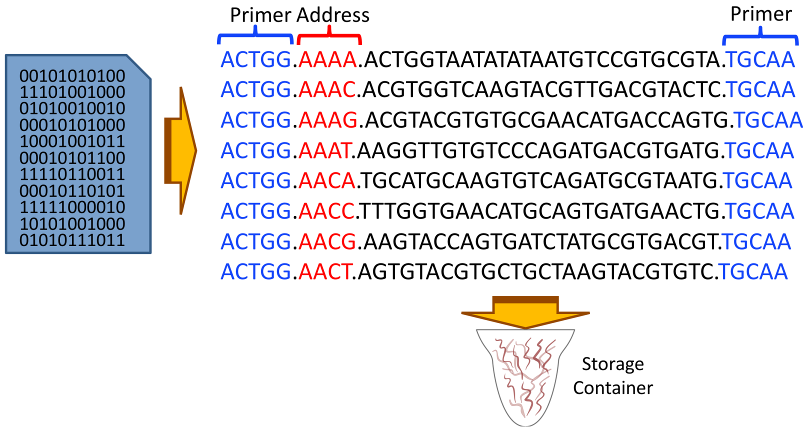
DNA Strands
ACTGAGTCAGTGACGTGCATGCA
CTGAGATGTAGTGCGTGCAGCTT
TCGTGCAGTGATGTCGTGCTTGC

DNA Strands
ACTGGGTCAGTGACGTGCATGCA
CTGAGATGCAGTGAGTGCAGCTT
TCGTGCAGTGATGTCGTGCATGC

DNA Sequencer

DNA Synthesizer

Storage Container

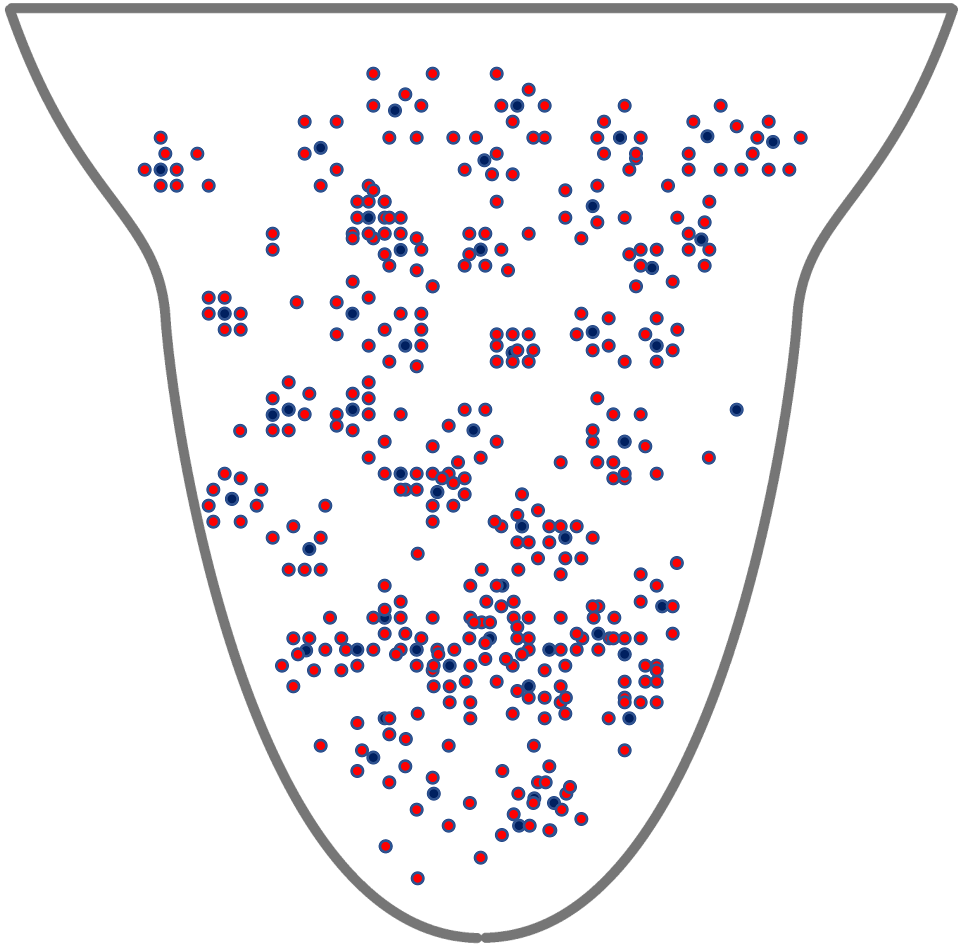


00101010100
11101001000
01010010010
00010101000
10001001011
00010101100
11110110011
00010110101
11111000010
10101001000
01010111011

Primer Address Primer

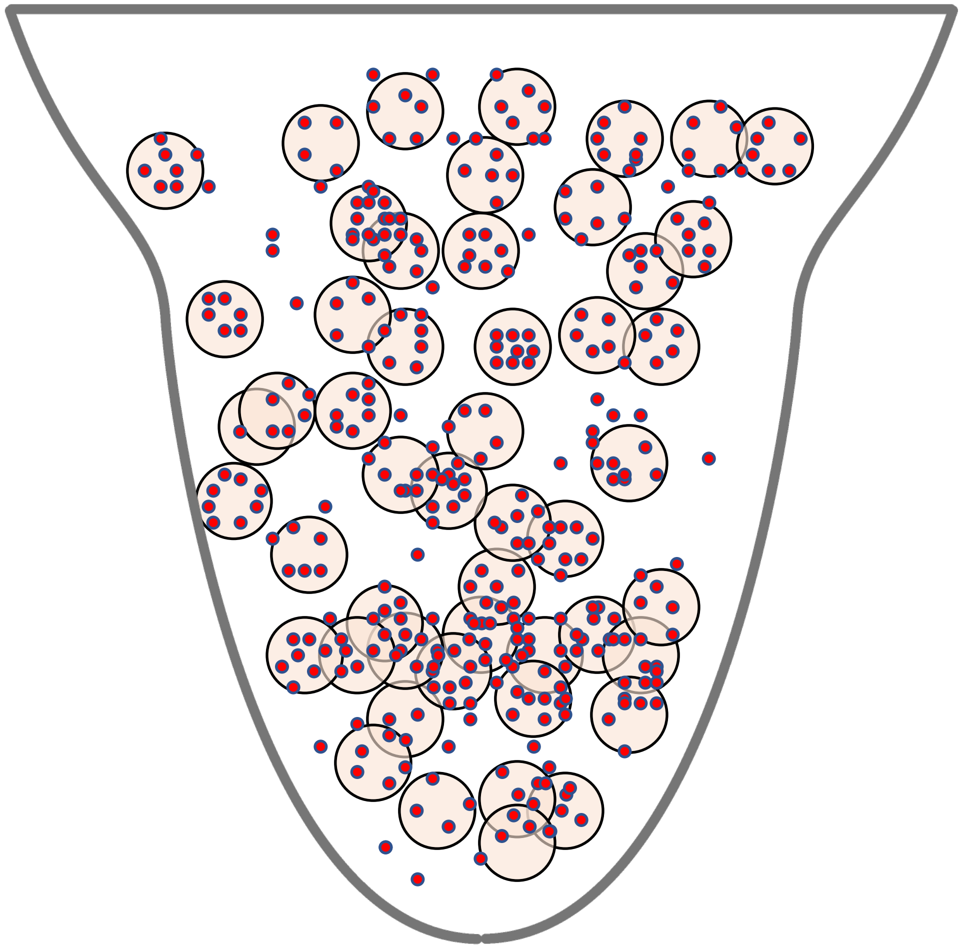
ACTGG.AAAA.ACTGGTAATATATAATGTCCGTGCGTA.TGCAA
ACTGG.AAAC.ACGTGGTCAAGTACGTTGACGTACTC.TGCAA
ACTGG.AAAG.ACGTACGTGTGCGAACATGACCAGTG.TGCAA
ACTGG.AAAT.AAGGTTGTGTCCCAGATGACGTGATG.TGCAA
ACTGG.AACA.TGCATGCAAGTGTCAGATGCGTAATG.TGCAA
ACTGG.AACC.TTTGGTGAACATGCAGTGATGAACTG.TGCAA
ACTGG.AACG.AAGTACCAGTGATCTATGCGTGACGT.TGCAA
ACTGG.AACT.AGTGTACGTGCTGCTAAGTACGTGTC.TGCAA

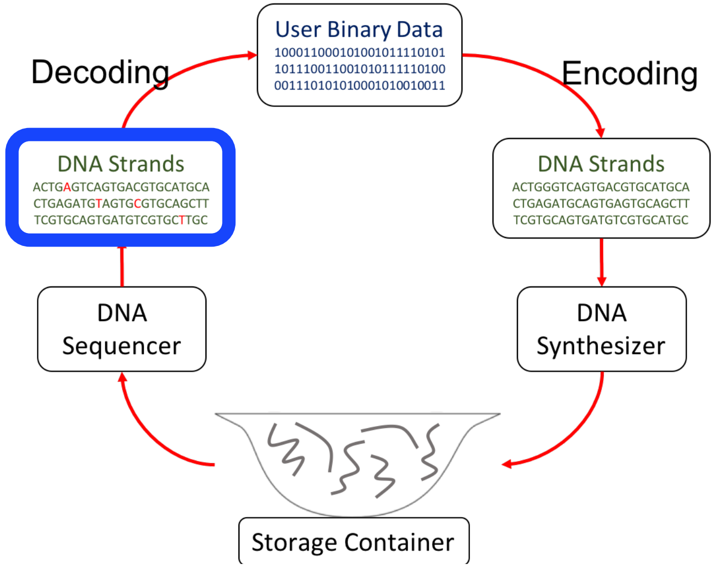
Storage
Container

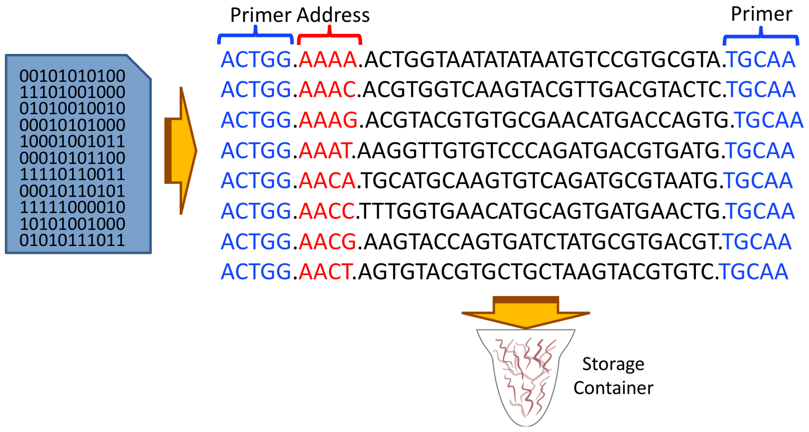


# DNA Storage Channel Model

## Clustering





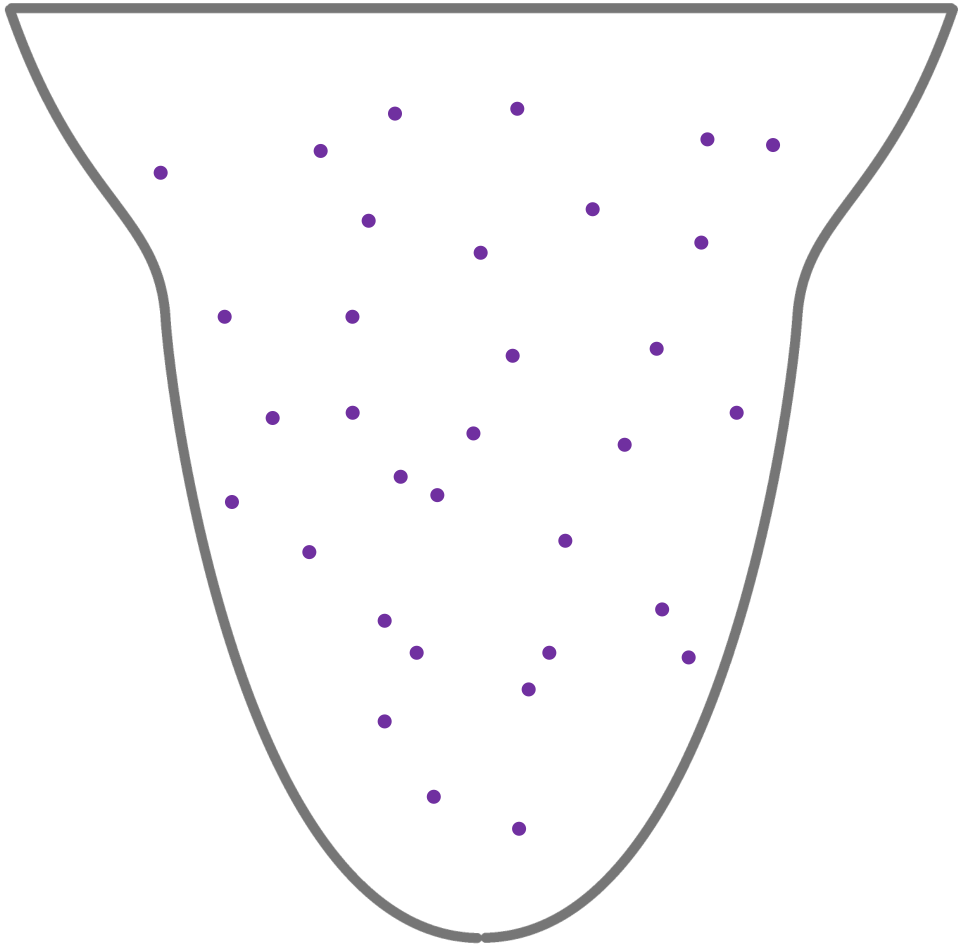


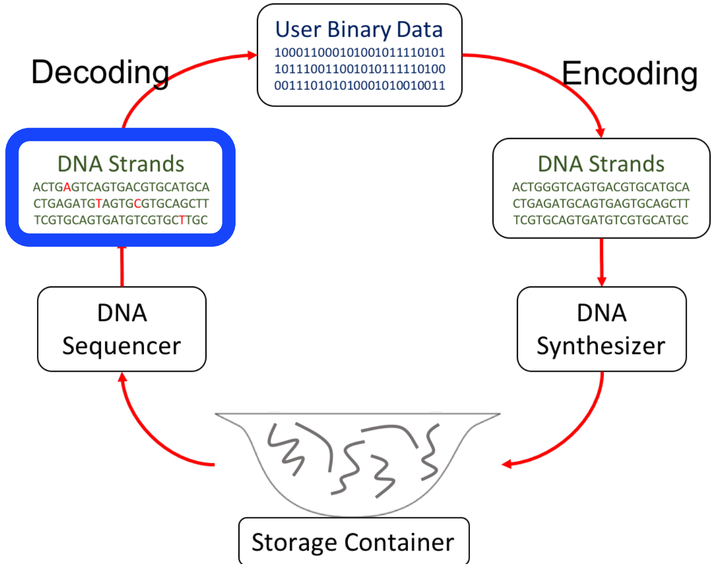
Primer Address Primer

ACTGG.AAAA.ACTGGTAATATATAATGTCCGTGCGTA.TGCAA
ACTGG.AAAC.ACGTGGTCAAGTACGTTGACGTACTC.TGCAA
ACTGG.AAAG.ACGTACGTGTGCGAACATGACCAGTG.TGCAA
ACTGG.AAAT.AAGGTTGTGTCCCAGATGACGTGATG.TGCAA
ACTGG.AACA.TGCATGCAAGTGTCAGATGCGTAATG.TGCAA
ACTGG.AACC.TTTGGTGAACATGCAGTGATGAACTG.TGCAA
ACTGG.AACG.AAGTACCAGTGATCTATGCGTGACGT.TGCAA
ACTGG.AACT.AGTGTACGTGCTGCTAAGTACGTGTC.TGCAA

# DNA Storage Channel Model

## Reconstruction

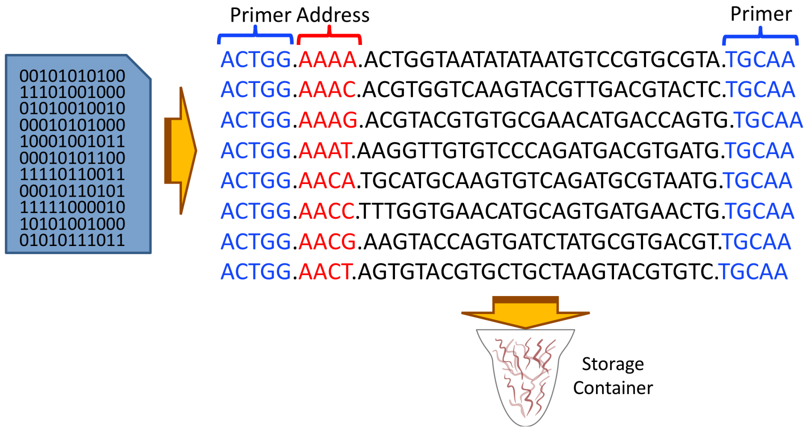




User Binary Data
10001100010100101111010 1
1011100110010101111101 00
0011101010100010100100 11

DNA Strands
ACTGAGTCAGTGACGTGCATGCA
CTGAGATGTAGTGCGTGCAGCTT
TCGTGCAGTGATGTCGTGCTTGC

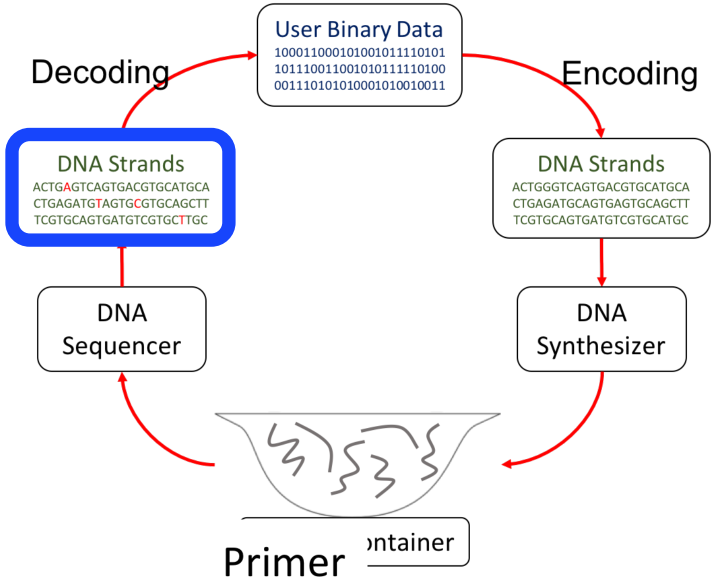
DNA Strands
ACTGGGTCAGTGACGTGCATGCA
CTGAGATGCAGTGAGTGCAGCTT
TCGTGCAGTGATGTCGTGCATGC



00101010100
11101001000
01010010010
00010101000
10001001011
00010101100
11110110011
00010110101
11111000010
10101001000
01010111011

ACTGG.AAAA.ACTGGTAATATATAATGTCCGTGCGTA.TGCAA
ACTGG.AAAC.ACGTGGTCAAGTACGTTGACGTACTC.TGCAA
ACTGG.AAAG.ACGTACGTGTGCGAACATGACCAGTG.TGCAA
ACTGG.AAAT.AAGGTTGTGTCCCAGATGACGTGATG.TGCAA
ACTGG.AACA.TGCATGCAAGTGTCAGATGCGTAATG.TGCAA
ACTGG.AACC.TTTGGTGAACATGCAGTGATGAACTG.TGCAA
ACTGG.AACG.AAGTACCAGTGATCTATGCGTGACGT.TGCAA
ACTGG.AACT.AGTGTACGTGCTGCTAAGTACGTGTC.TGCAA

# DNA Storage Channel Model



Decoding

User Binary Data
100011000101001011110101
101110011001010111110100
001110101010001010010011

Encoding

DNA Strands
ACTGAGTCAGTGACGTGCATGCA
CTGAGATGTAGTGCGTGCAGCTT
TCGTGCAGTGATGTCGTGCTTGC

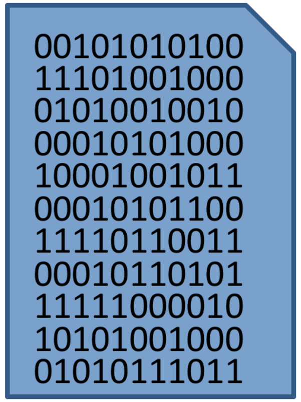
DNA Strands
ACTGGGTCAGTGACGTGCATGCA
CTGAGATGCAGTGAGTGCAGCTT
TCGTGCAGTGATGTCGTGCATGC

DNA Sequencer

DNA Synthesizer

ontainer

## ECC



00101010100
11101001000
01010010010
00010101000
10001001011
00010101100
11110110011
00010110101
11111000010
10101001000
01010111011

Primer Address

Primer

ACTGG.AAAA.ACTGGTAATATATAATGTCCGTGCGTA.TGCAA
ACTGG.AAAC.ACGTGGTCAAGTACGTTGACGTACTC.TGCAA
ACTGG.AAAG.ACGTACGTGTGCGAACATGACCAGTG.TGCAA
ACTGG.AAAT.AAGGTTGTGTCCCAGATGACGTGATG.TGCAA
ACTGG.AACA.TGCATGCAAGTGTCAGATGCGTAATG.TGCAA
ACTGG.AACC.TTTGGTGAACATGCAGTGATGAACTG.TGCAA
ACTGG.AACG.AAGTACCAGTGATCTATGCGTGACGT.TGCAA
ACTGG.AACT.AGTGTACGTGCTGCTAAGTACGTGTC.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

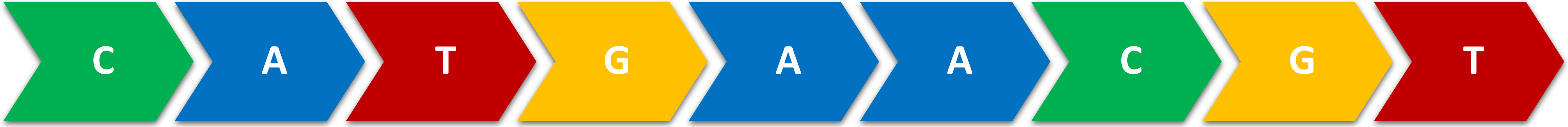
## Sequencing

Higher GC Content affects sequencing error

Presence of **Homopolymers** increases the error rate

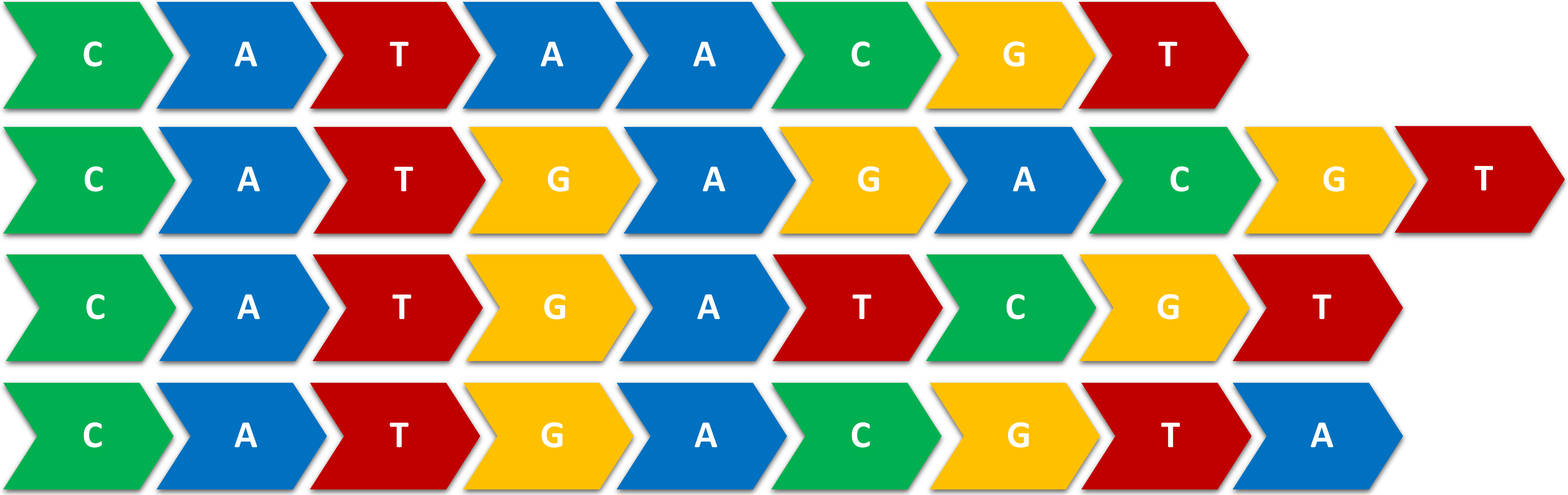
# Errors in DNA

Design:



C A T G A A C G T

Copies:



C A T A A C G T

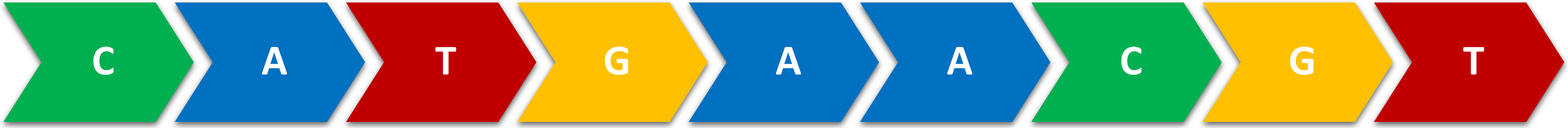
C A T G A G A C G T

C A T G A T C G T

C A T G A C G T A

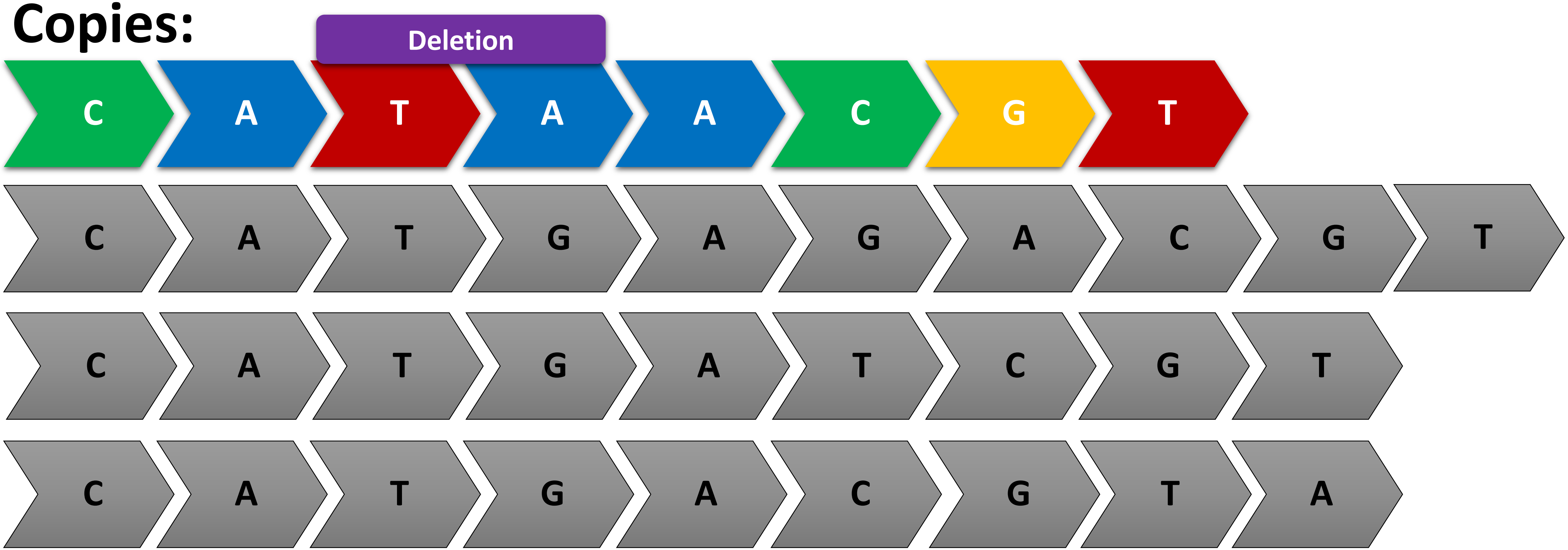
# Errors in DNA

Design:



C A T G A A C G T

Copies:



Deletion

C A T A A C G T

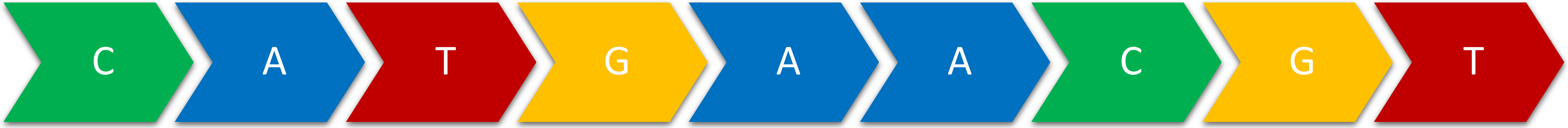
C A T G A G A C G T

C A T G A T C G T

C A T G A C G T A

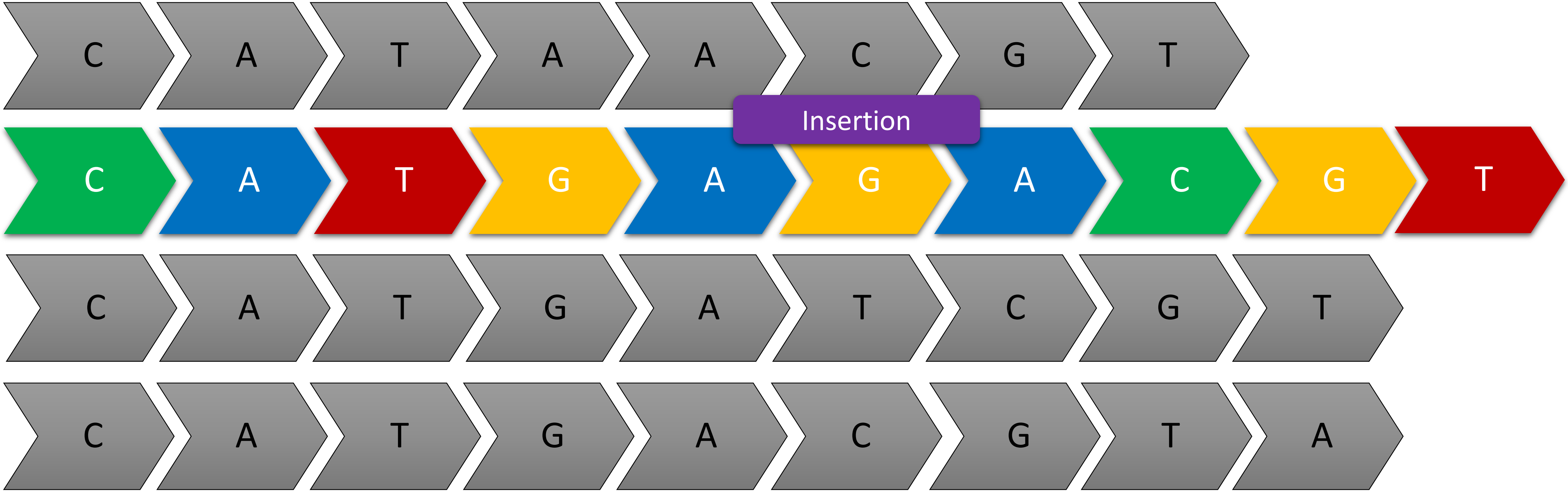
# Errors in DNA

## Design:



C A T G A A C G T

## Copies:



C A T A A C G T

Insertion

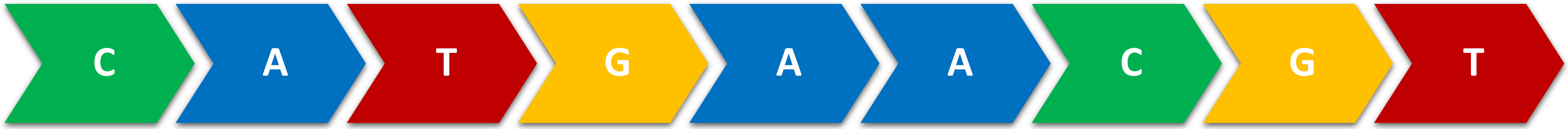
C A T G A G A C G T

C A T G A T C G T

C A T G A C G T A

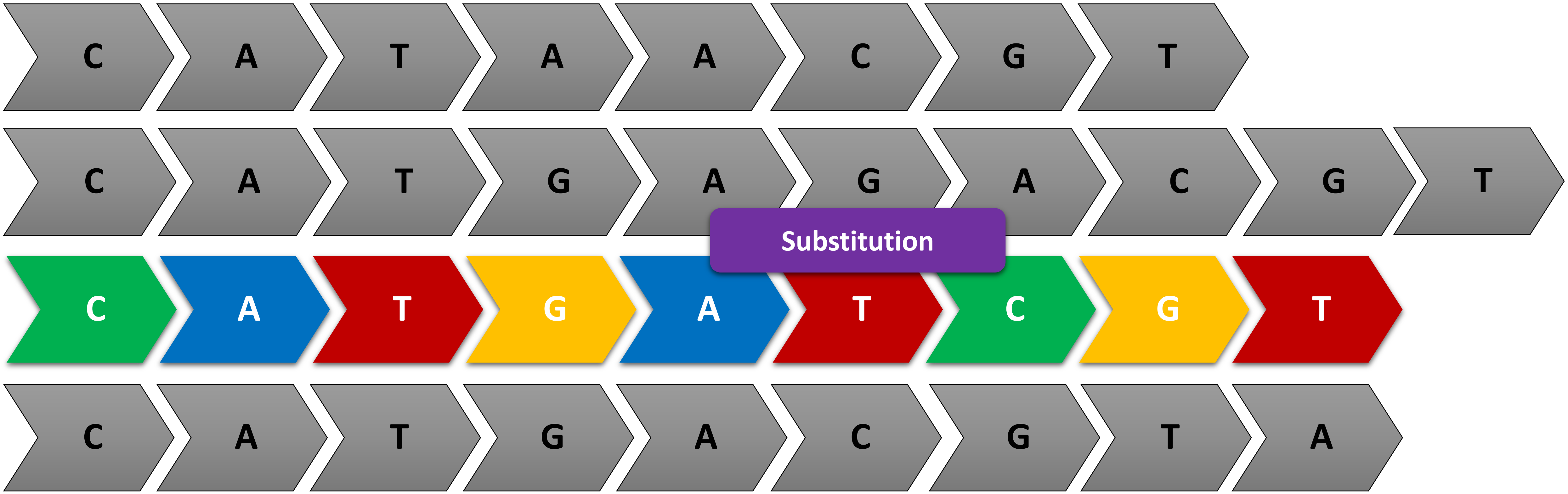
# Errors in DNA

**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

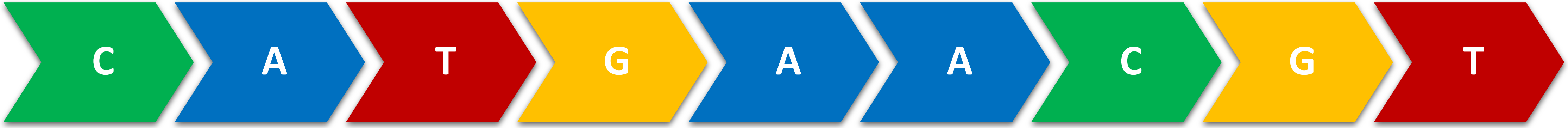
Substitution

C A T G A T C G T

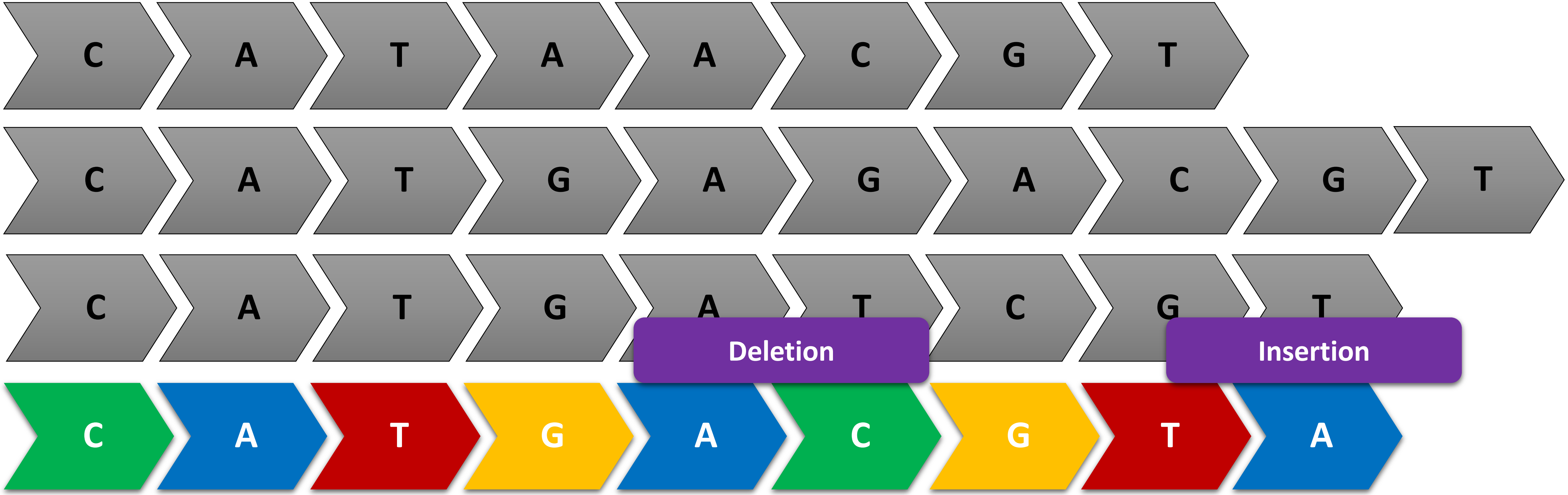
C A T G A C G T A

# Errors in DNA

## Design:

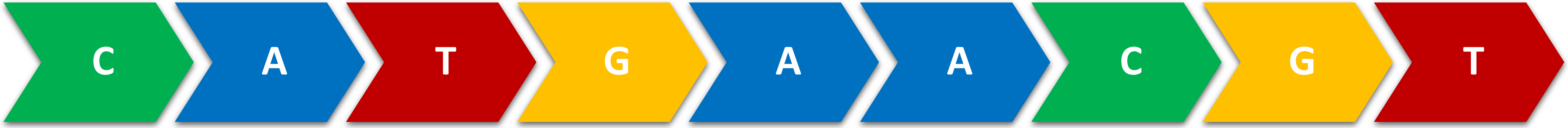


## Copies:

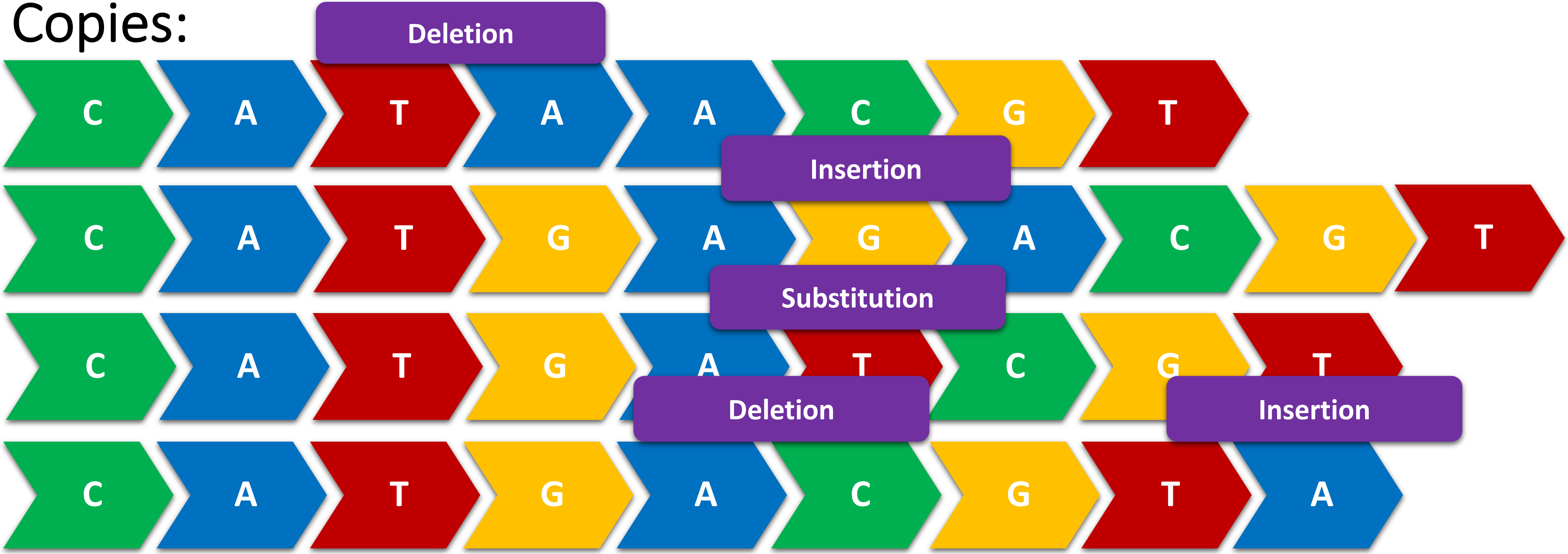


# Errors in DNA

Design:


C A T G A A C G T

Copies:


Deletion

C A T A A C G T

Insertion

C A T G A G A C G T

Substitution

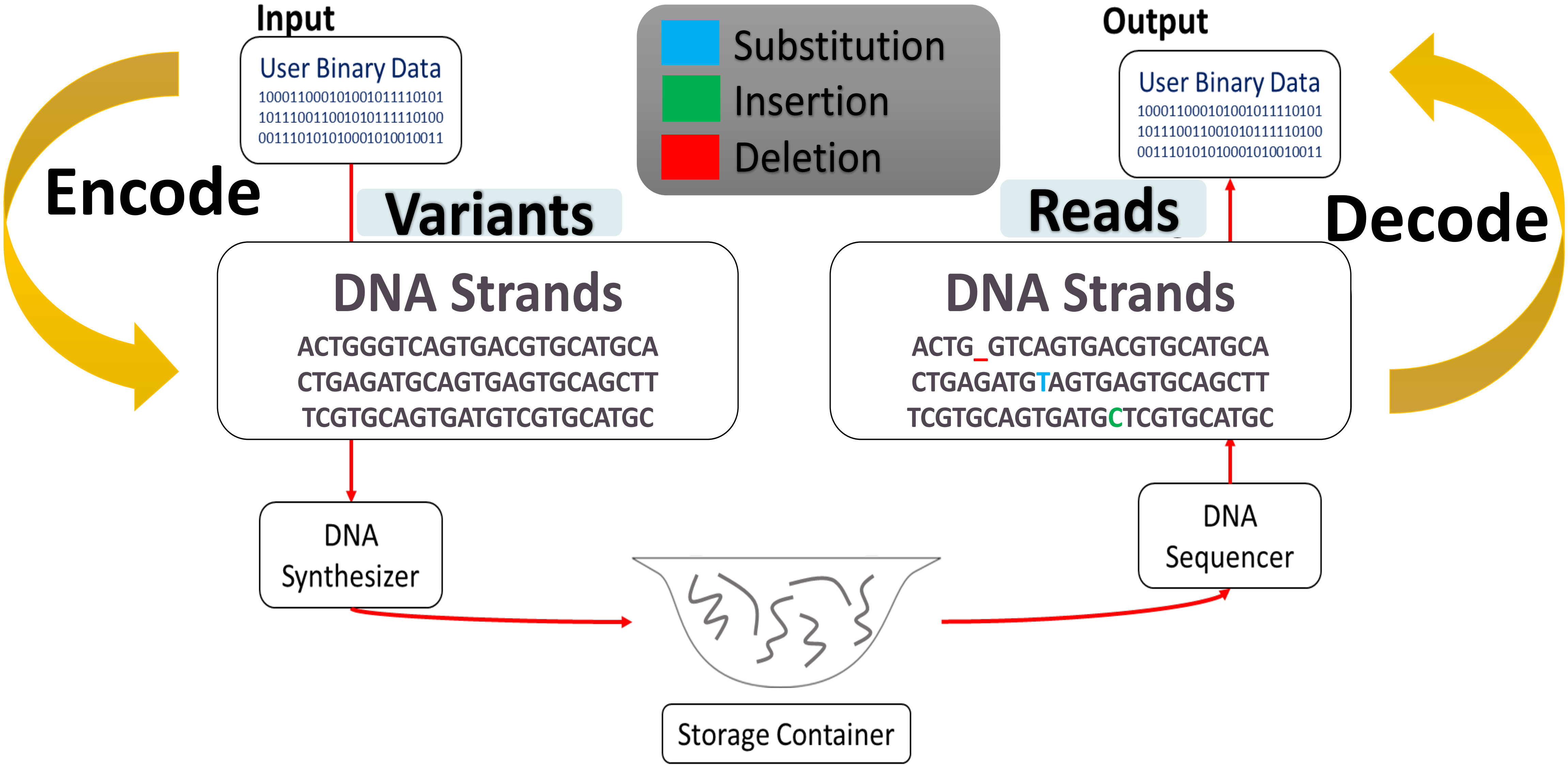
C A T G A T C G T

Deletion

Insertion

C A T G A C G T A

# Error Characterization



🟦 Substitution

🟩 Insertion

🟥 Deletion

Input

User Binary Data
100011000101001011110101
101110011001010111110100
001110101010001010010011

Encode

Variants

DNA Strands
ACTGGGTCAGTGACGTGCATGCA
CTGAGATGCAGTGAGTGCAGCTT
TCGTGCAGTGATGTCGTGCATGC

DNA Synthesizer

Storage Container

DNA Sequencer

Reads

DNA Strands
ACTG_GTCAGTGACGTGCATGCA
CTGAGATGTAGTGAGTGCAGCTT
TCGTGCAGTGATGCTCGTGCATGC

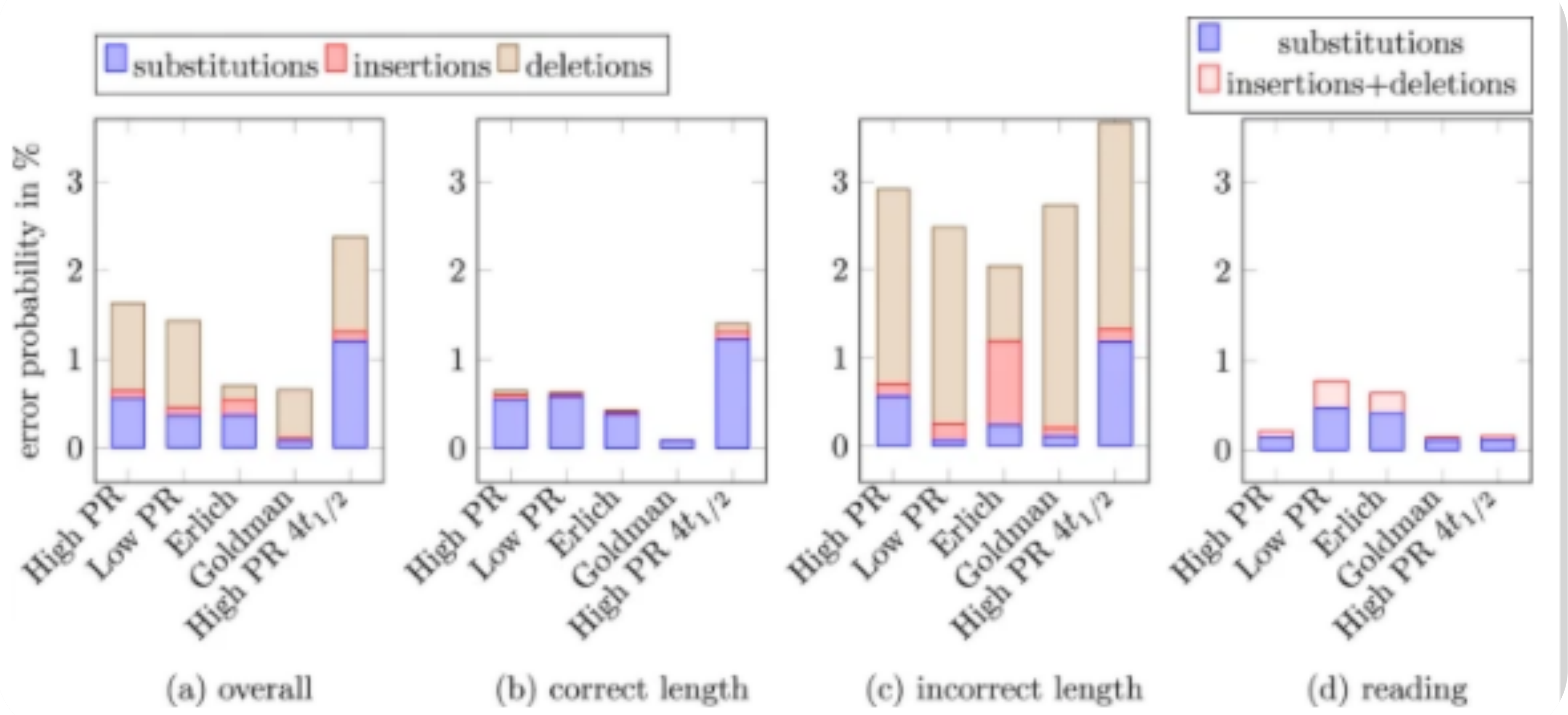
Output

User Binary Data
100011000101001011110101
101110011001010111110100
001110101010001010010011

Decode

# Error Characterization

Heckel, Mikutis, and Grass. **A characterization of the DNA data storage channel.** *Scientific Report*, **2019**.



(a) overall (b) correct length (c) incorrect length (d) reading

# SOLQC Pipeline

python™

edlib 1.3.8.post1

pandas

## 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

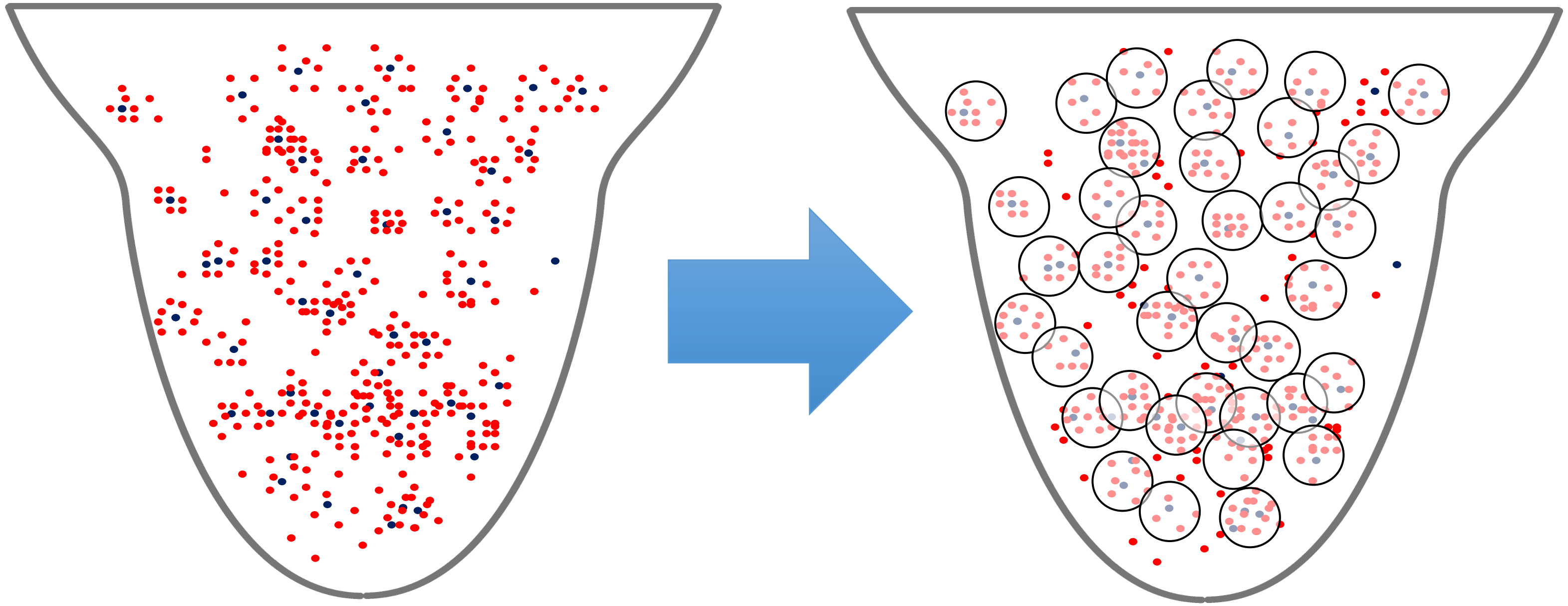
Quality report consisting of plots and statistical values

# SOLQC Pipeline

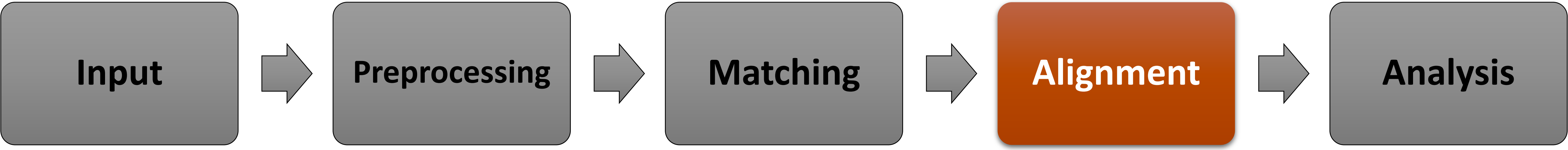
Input → Preprocessing → Matching → Alignment → Analysis

## Matching - Clustering

The set of reads which are matched to the same variant forms a **variant cluster.**



# SOLQC Pipeline



Input

Preprocessing

Matching

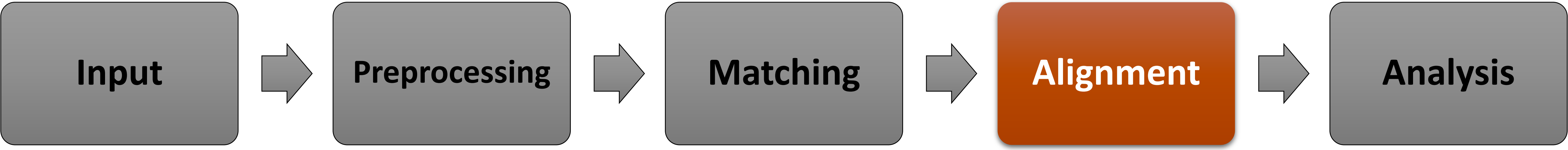
Alignment

Analysis

## 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

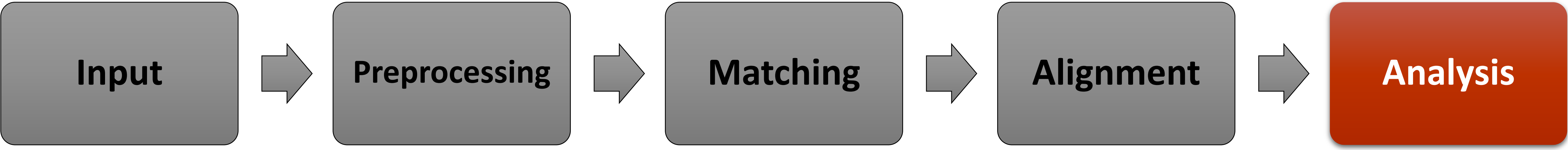


Input → Preprocessing → Matching → Alignment → Analysis

| Variant | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Read #1 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Read #2 | | | D | D | D | | | | | | | | | S | | | | | | | | | | | | | |
| Read #3 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Read #4 | | | | | | | | | D | | | | | | | | | | D | | | | | | | I | |
| Read #5 | | | | | | | | | | | | | | | S | | | | | | | | | | | | D |
| Read #6 | | | D | | | | | | | | | | | | | | | | | | | | | | S | | |
| Read #7 | | | | | | | | | | | | | | | | | | I | | | | | | | | | |
| Read #8 | | | | | | | S | | | | | | | | | | | | | | | | | | | | |
| Read #9 | | | | | I | | | D | D | | | | | | | | | | | | | | | | | | |
| Read #10 | | | I | | | | | | | D | D | D | | | | | | S | | | | | | | | | |
| Read #11 | | | | | | | | | | | | | | | | | | | | | | | D | D | D | D | D |
| Read #12 | D | D | D | | | | | | | | S | | | | | | | | | | | | | | | | |
| Read #13 | | | | | D | | | | | | D | | | | | | | | | | | | | | | | |
| Read #14 | | | | | | | | | | | S | | | | | | | | | | | | | | | | |
| Read #15 | S | I | | | | | | | | | | | | | | | | | | | | | | | | | |

D - Deletion S - Substitution I - Insertion

# SOLQC Pipeline



Input → Preprocessing → Matching → Alignment → Analysis

## Analysis

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

| | Grass et al. | Erlich & Zielinski | Organick et al. | Yazdi et al. |
|---|---|---|---|---|
| Storage size | 81KB | 2.11 MB | 200 MB (9.5 MB) | 3.633 KB |
| Design length | 158 | 152 | 150 | 880-1,060 |
| # variants | 5,000 | 72,000 | 607,150 | 17 |
| # reads | 3,312,235 | 15,787,115 | 62,879,612 | 6,660 |
| # filtered reads | 1,945,744 | 1,427,781 | 91,898 | 6,660 |
| Synthesis | CustomArray \| GenScript Make Research Easy | TWIST BIOSCIENCE | TWIST BIOSCIENCE | IDT INTEGRATED DNA TECHNOLOGIES |
| Sequencing | illumina® | illumina® | illumina® | Oxford NANOPORE Technologies |

Grass, Heckel, Puddu, Paunescu, and Stark, **Robust chemical preservation of digital information on DNA in silica with error-correcting codes**. Angewandte Chemie International Edition, 2015.

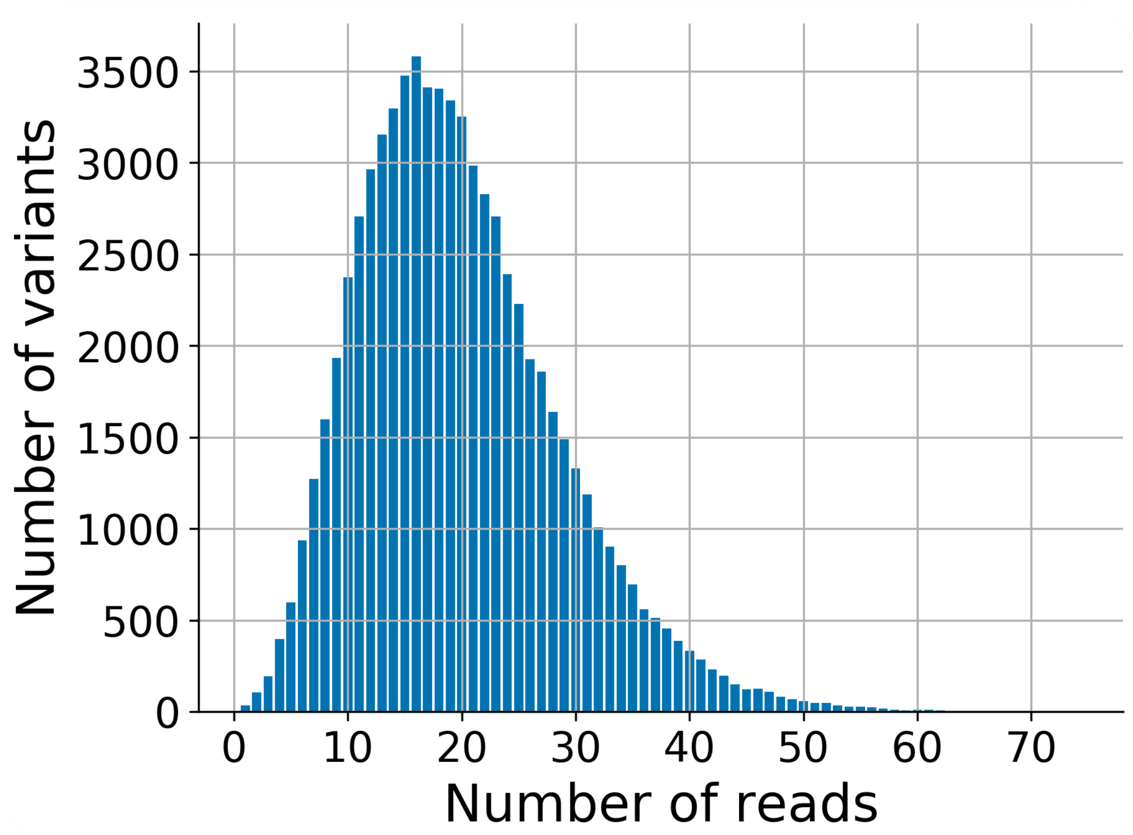
Erlich and Zielinski, **DNA fountain enables a robust and efficient storage architecture**. Science, 2017.

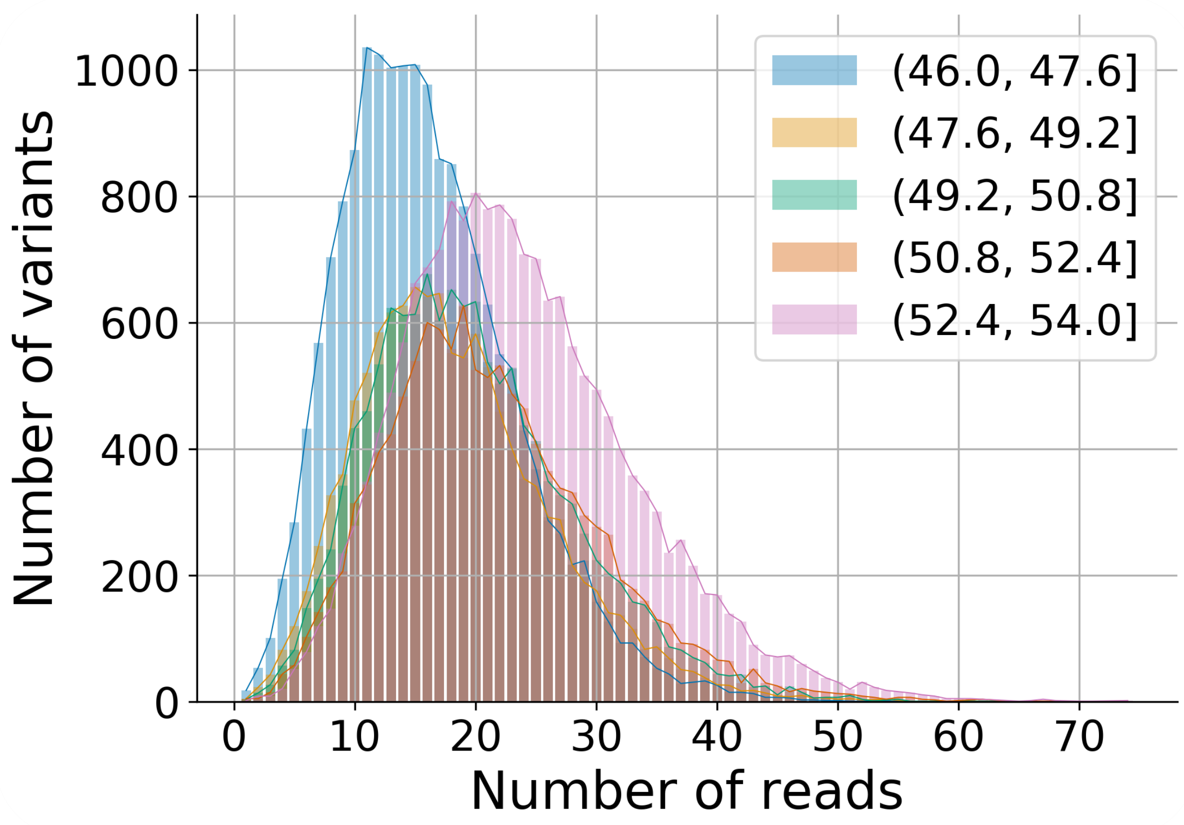
Organick, Ang, Chen, Lopez, Yekhanin, Makarychev, Racz, Kamath, Gopalan, Nguyen, Takahashi, Newman, Parker, Rashtchian, Stewart, Gupta, Carlson, Mulligan, Carmean, Seelig, Ceze, and Strauss. **Random access in large-scale DNA data storage.** Nature Biotechnology, 2018.

Yazdi, Gabrys, and Milenkovic. **Portable and error-free DNA-based data storage**. Scientific Reports, 2017.

# Histogram of cluster size per variant

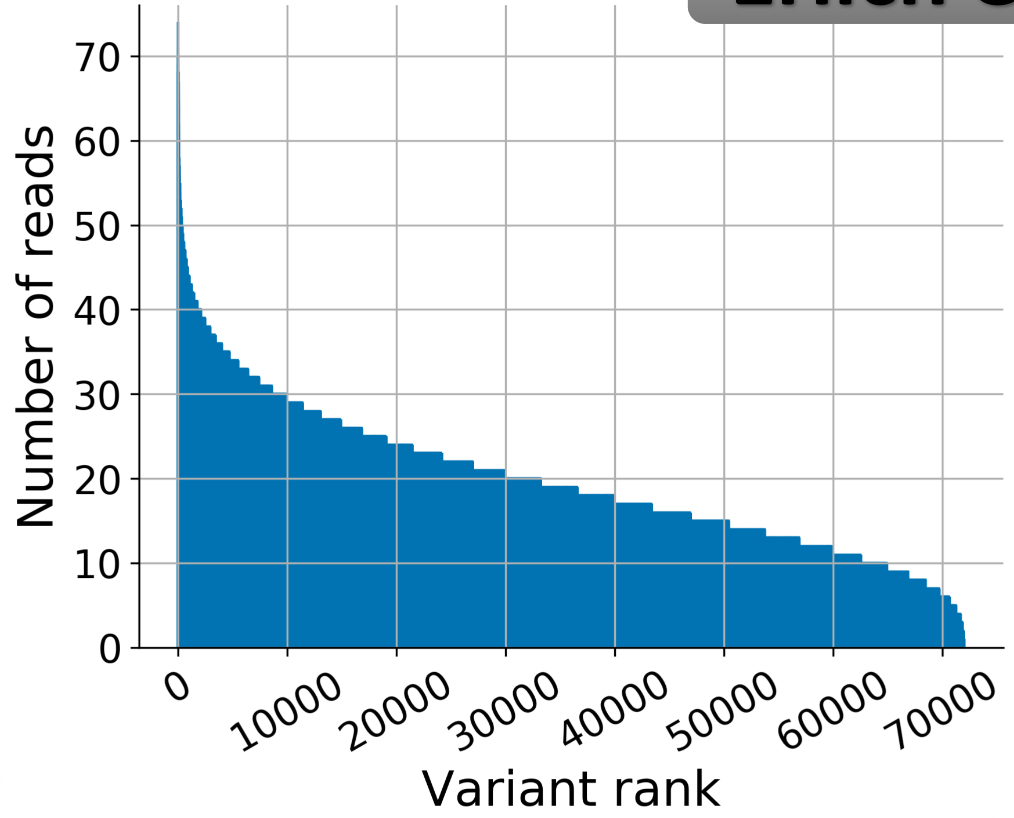
## Erlich & Zielinski

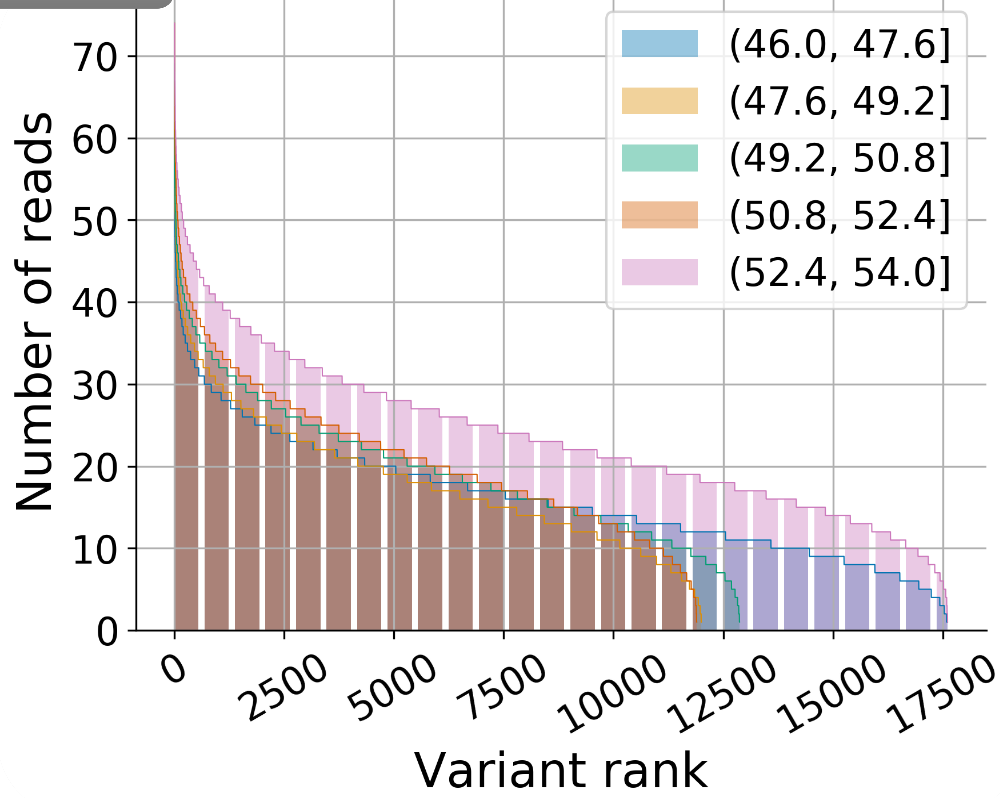




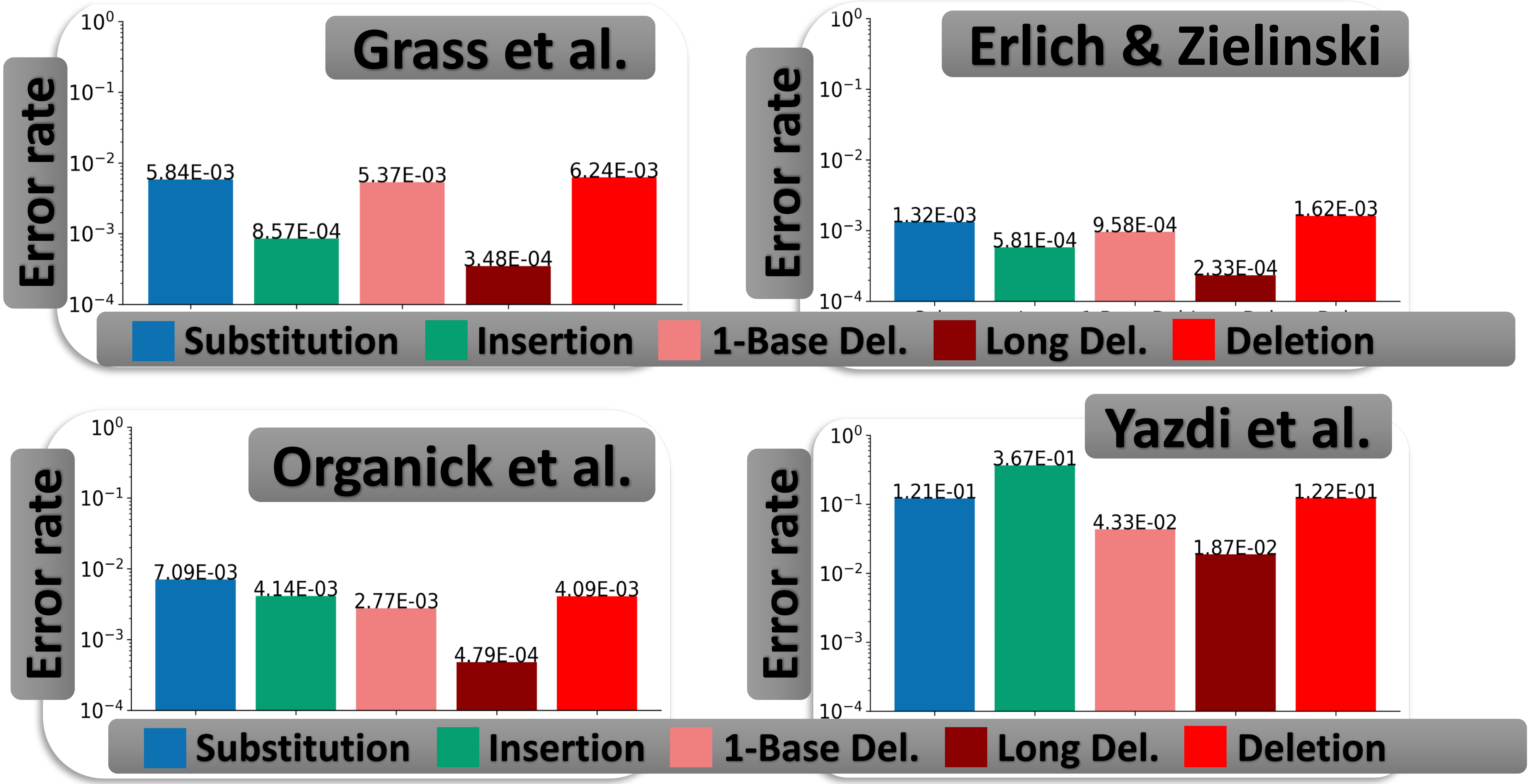
# Sorted bar plot of the number of filtered reads per variant

## Erlich & Zielinski



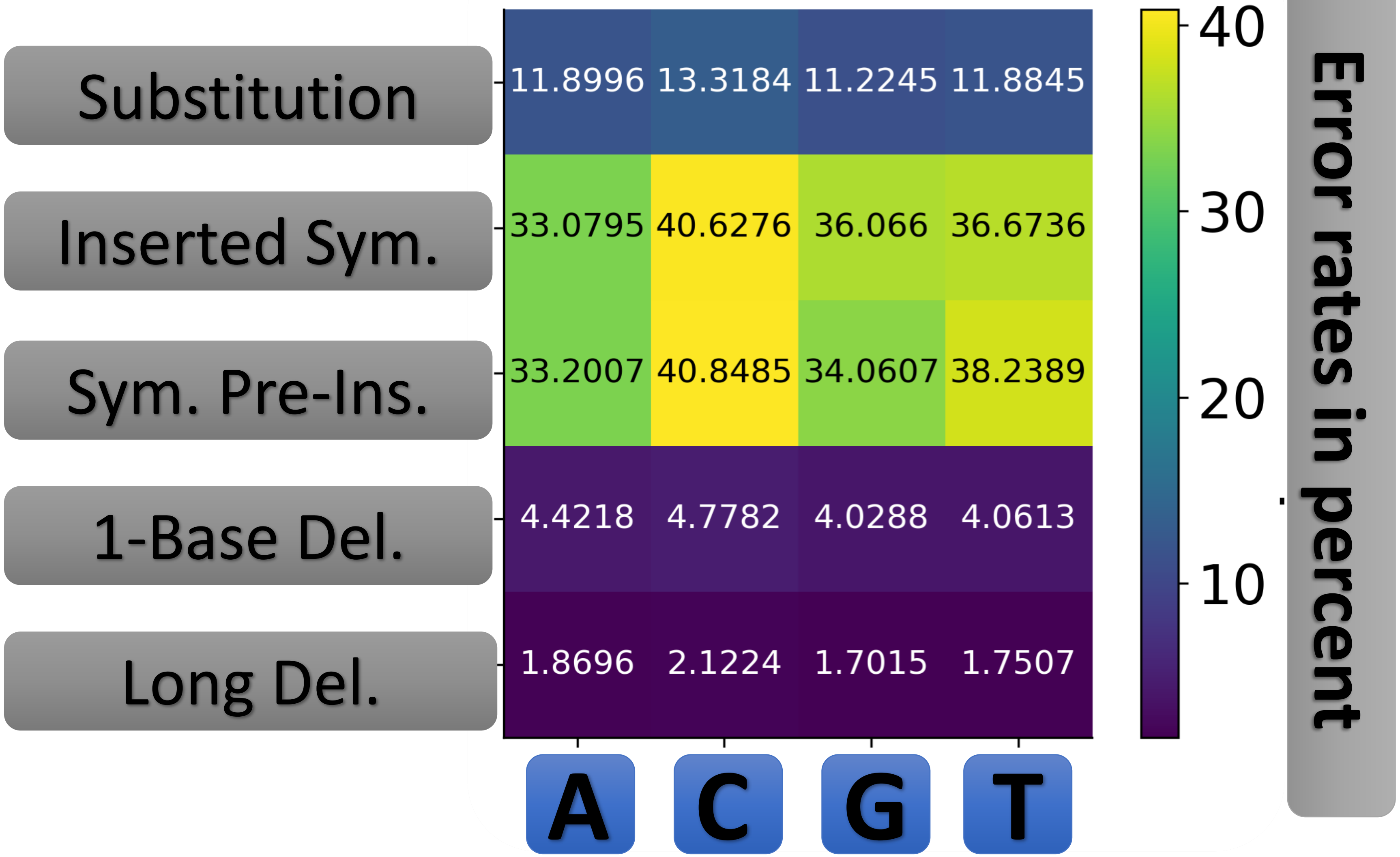


# Total error rates



# Error rates, stratified by symbol

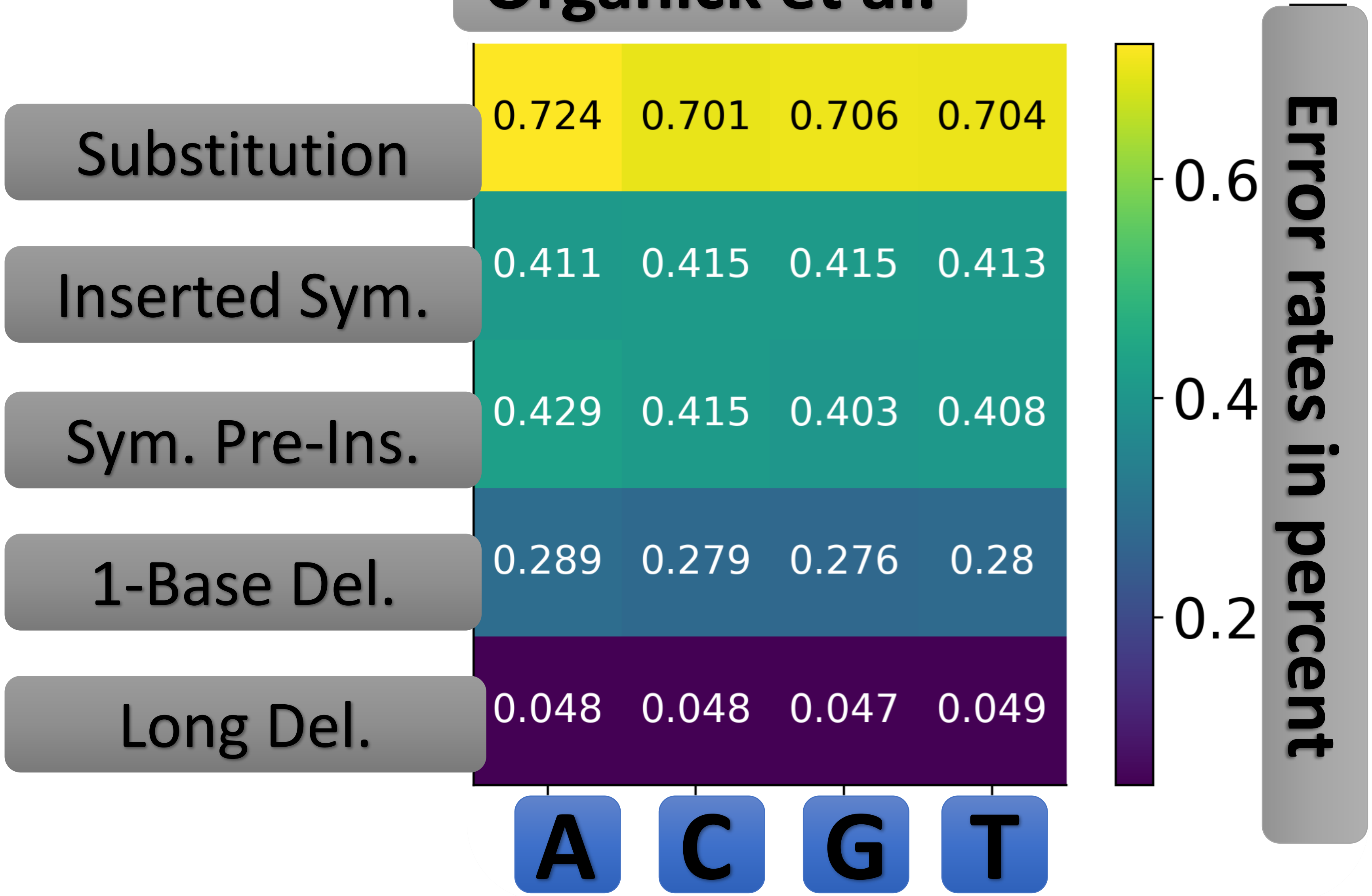
## Yazdi et al.



| | A | C | G | T |
|---|---|---|---|---|
| Substitution | 11.8996 | 13.3184 | 11.2245 | 11.8845 |
| Inserted Sym. | 33.0795 | 40.6276 | 36.066 | 36.6736 |
| Sym. Pre-Ins. | 33.2007 | 40.8485 | 34.0607 | 38.2389 |
| 1-Base Del. | 4.4218 | 4.7782 | 4.0288 | 4.0613 |
| Long Del. | 1.8696 | 2.1224 | 1.7015 | 1.7507 |

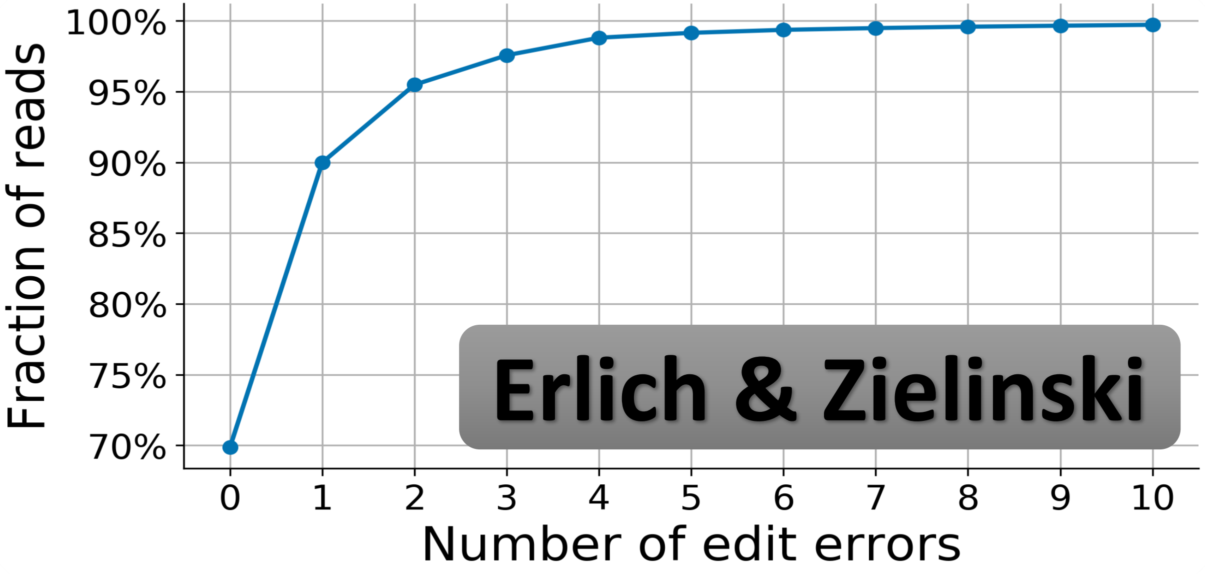
# Error rates, stratified by symbol

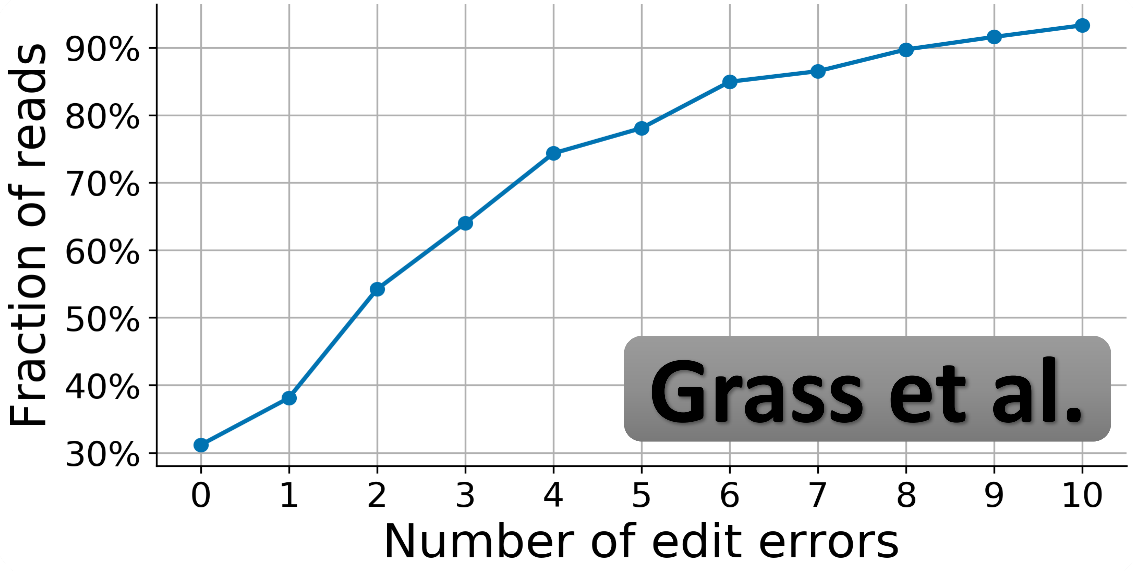
## Organick et al.

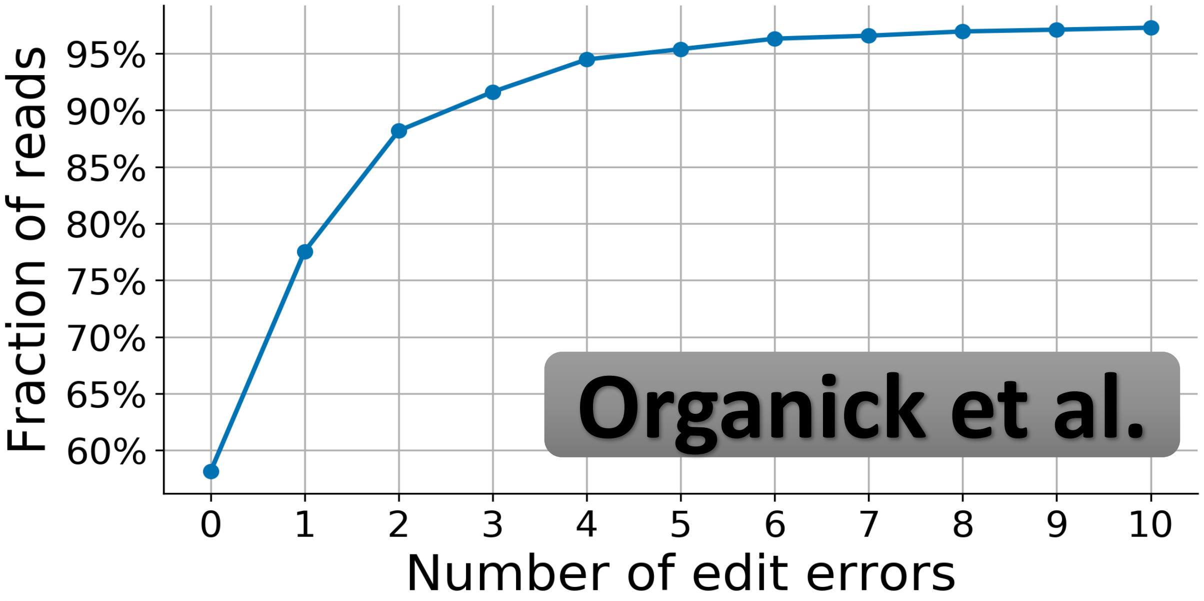


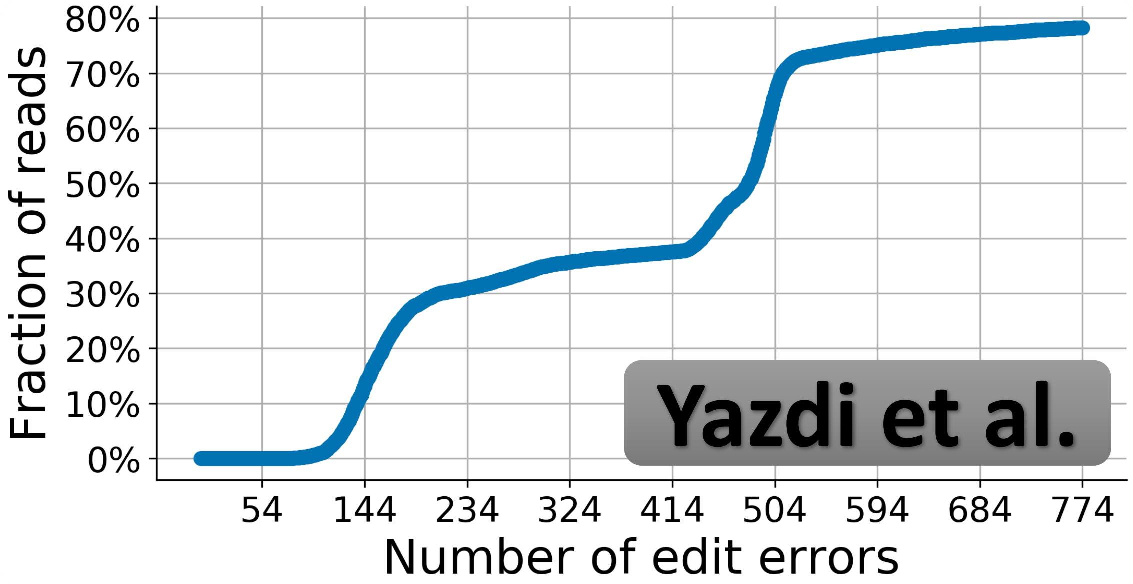
|  | A | C | G | T |
|---|---|---|---|---|
| Substitution | 0.724 | 0.701 | 0.706 | 0.704 |
| Inserted Sym. | 0.411 | 0.415 | 0.415 | 0.413 |
| Sym. Pre-Ins. | 0.429 | 0.415 | 0.403 | 0.408 |
| 1-Base Del. | 0.289 | 0.279 | 0.276 | 0.28 |
| Long Del. | 0.048 | 0.048 | 0.047 | 0.049 |

# Cumulative distribution based upon the number of errors



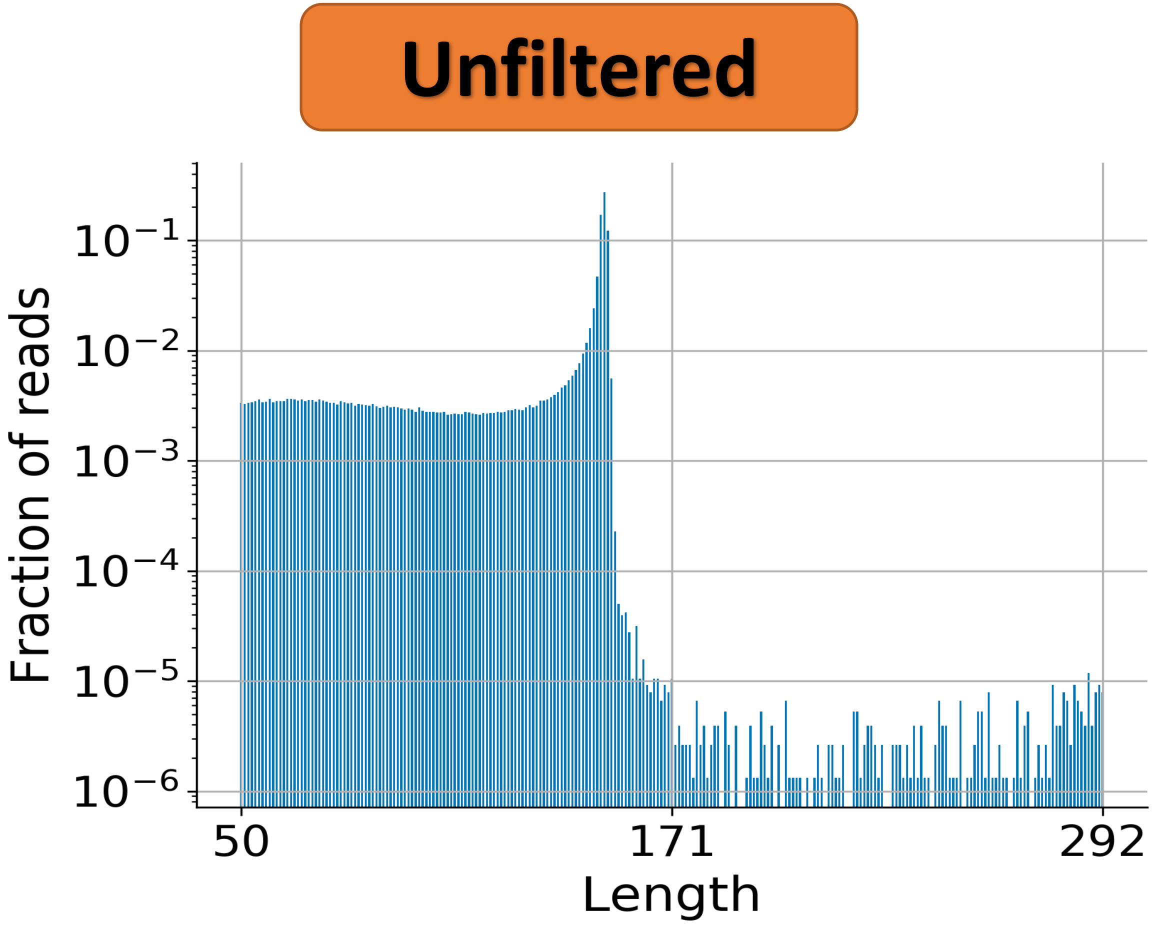


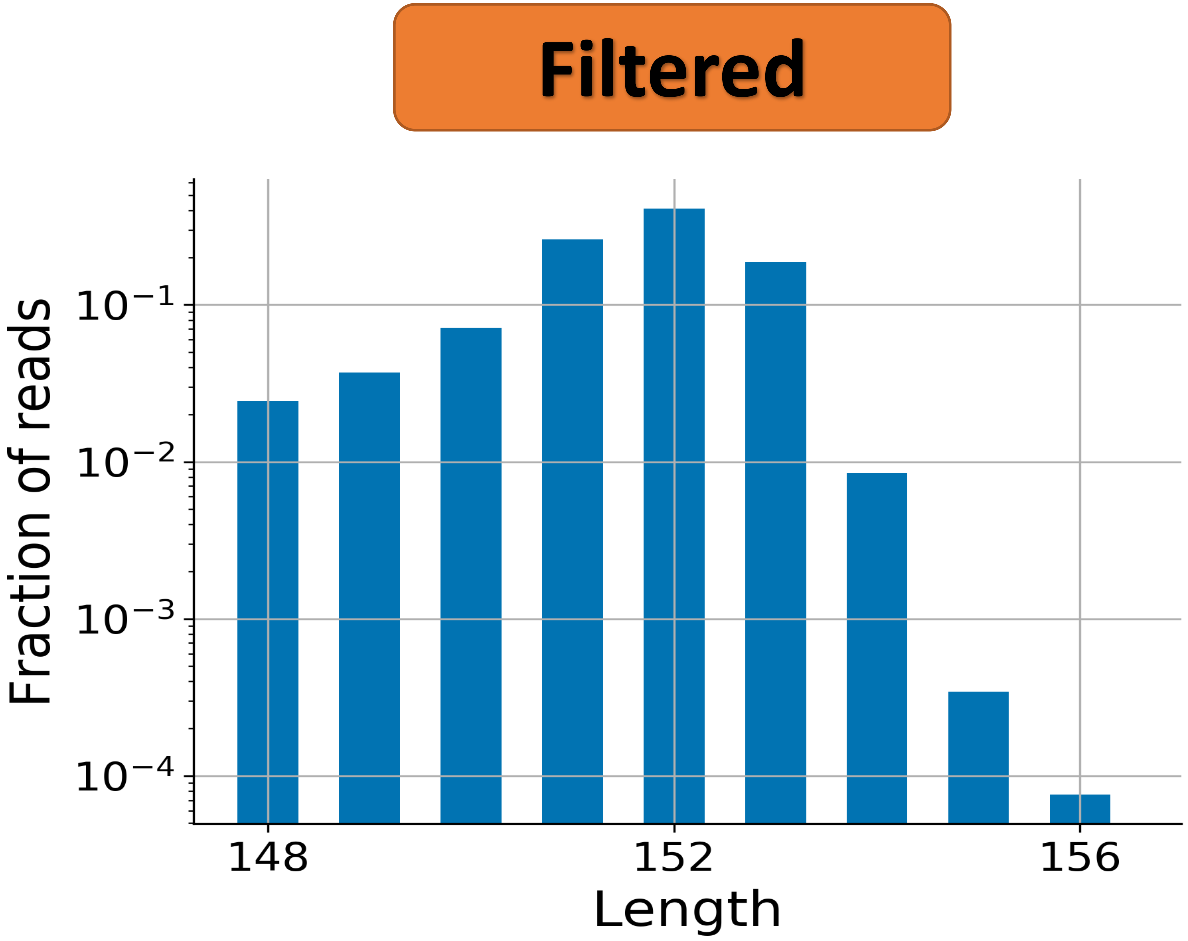




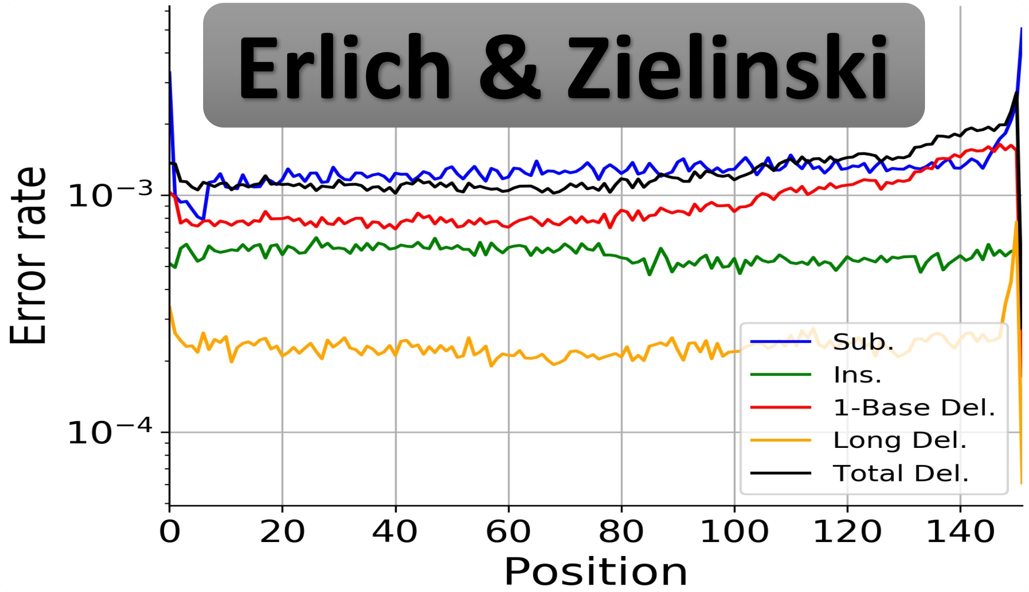
# Histogram of the length of the reads

## Erlich & Zielinski

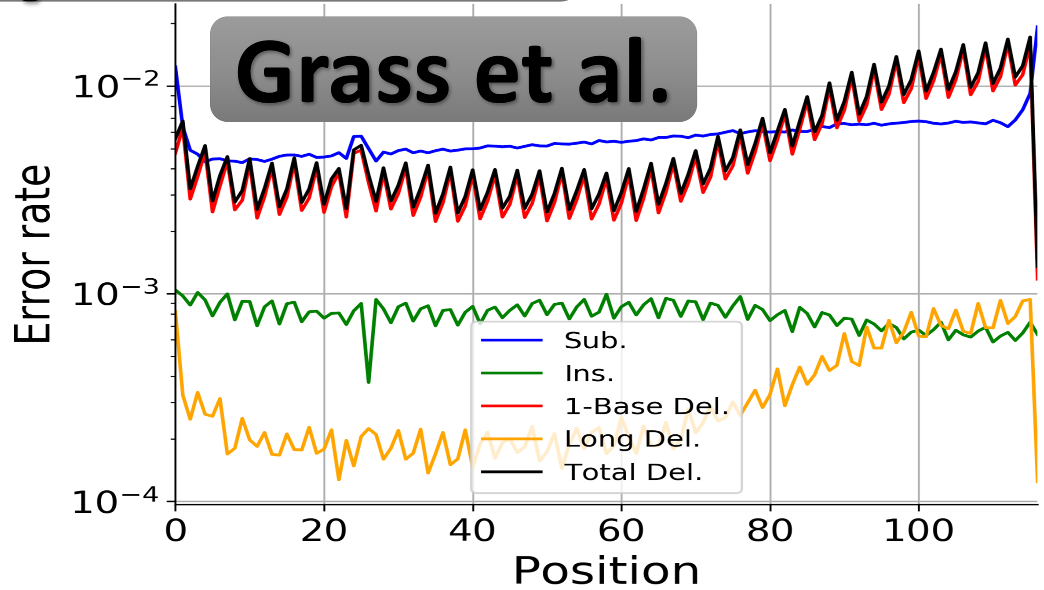

Unfiltered


Filtered

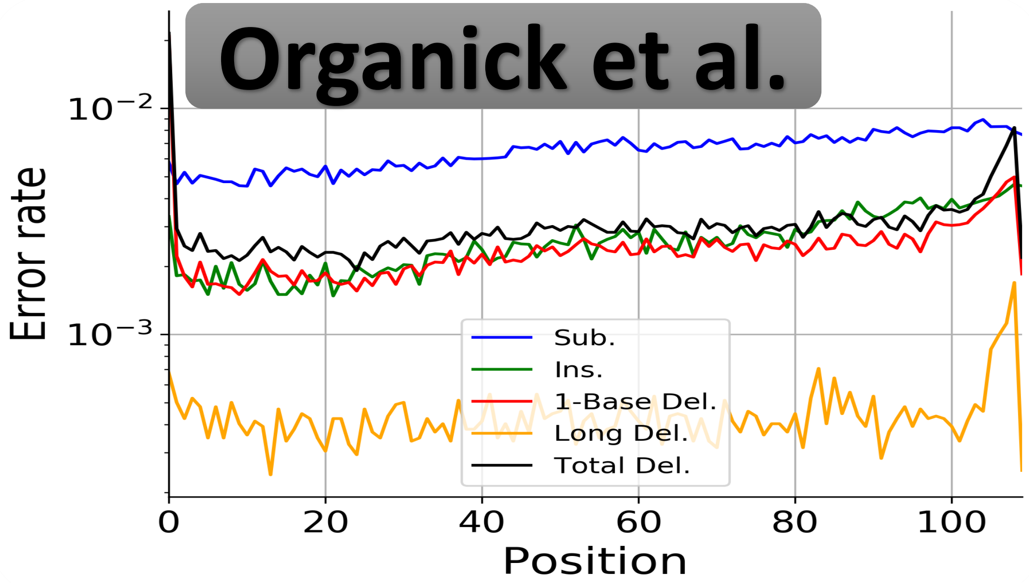
# Error rates per position



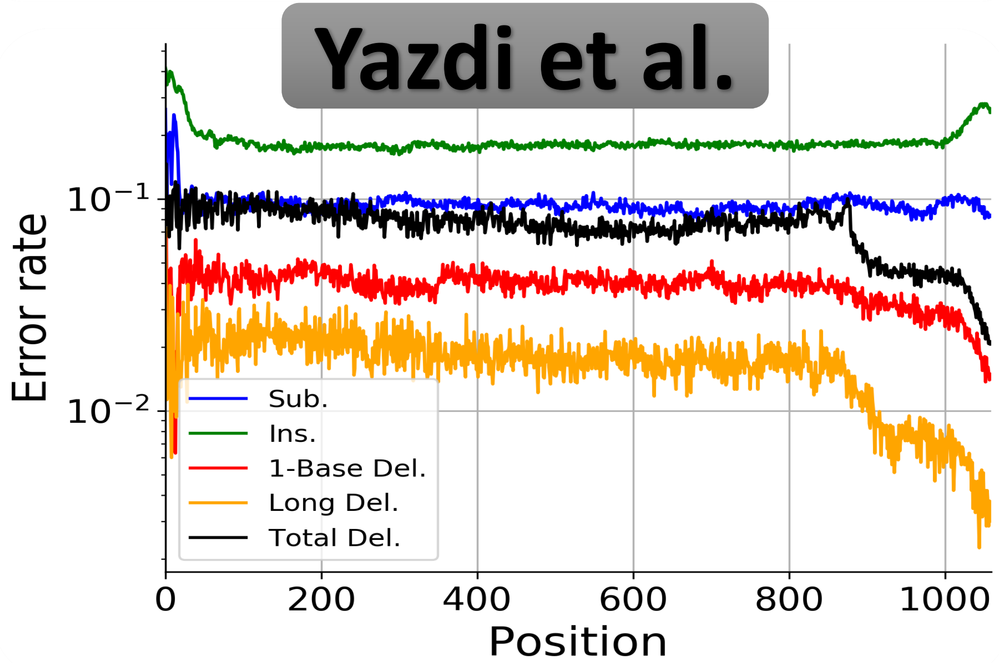
Erlich & Zielinski



Grass et al.



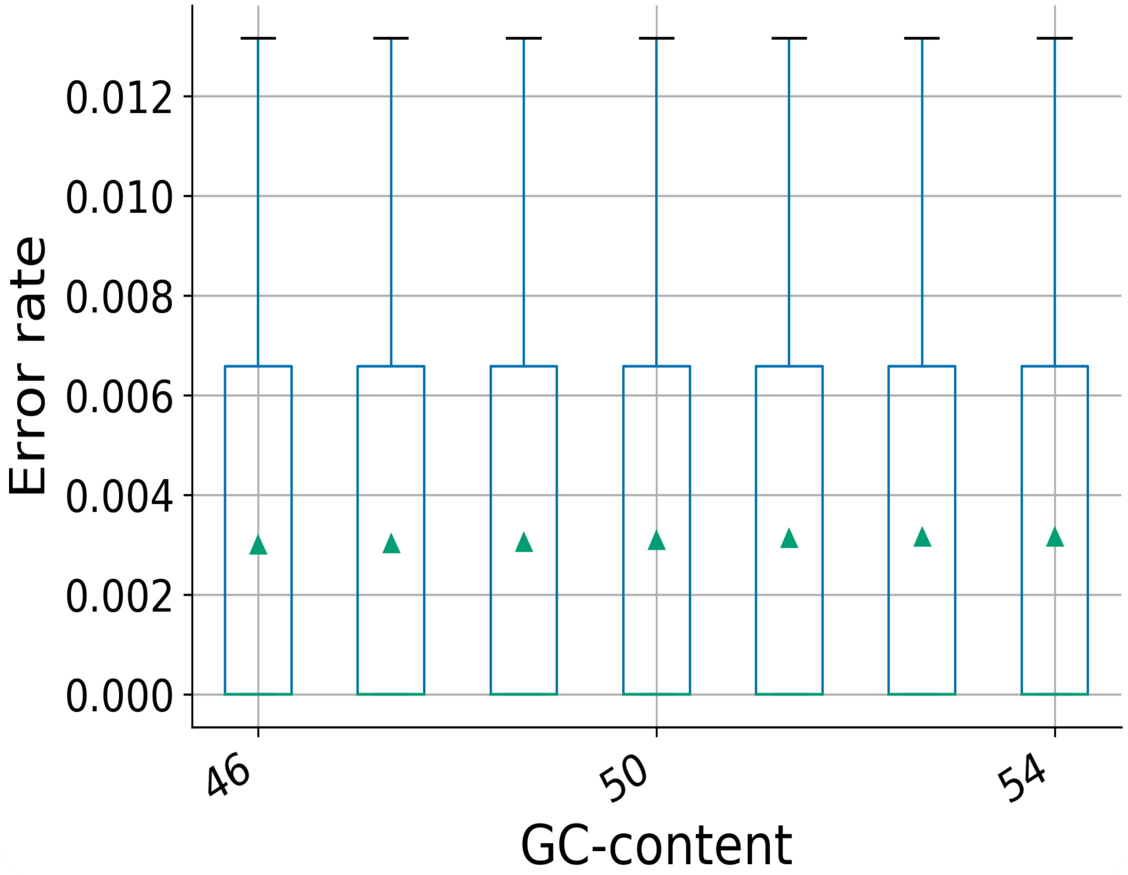
Organick et al.



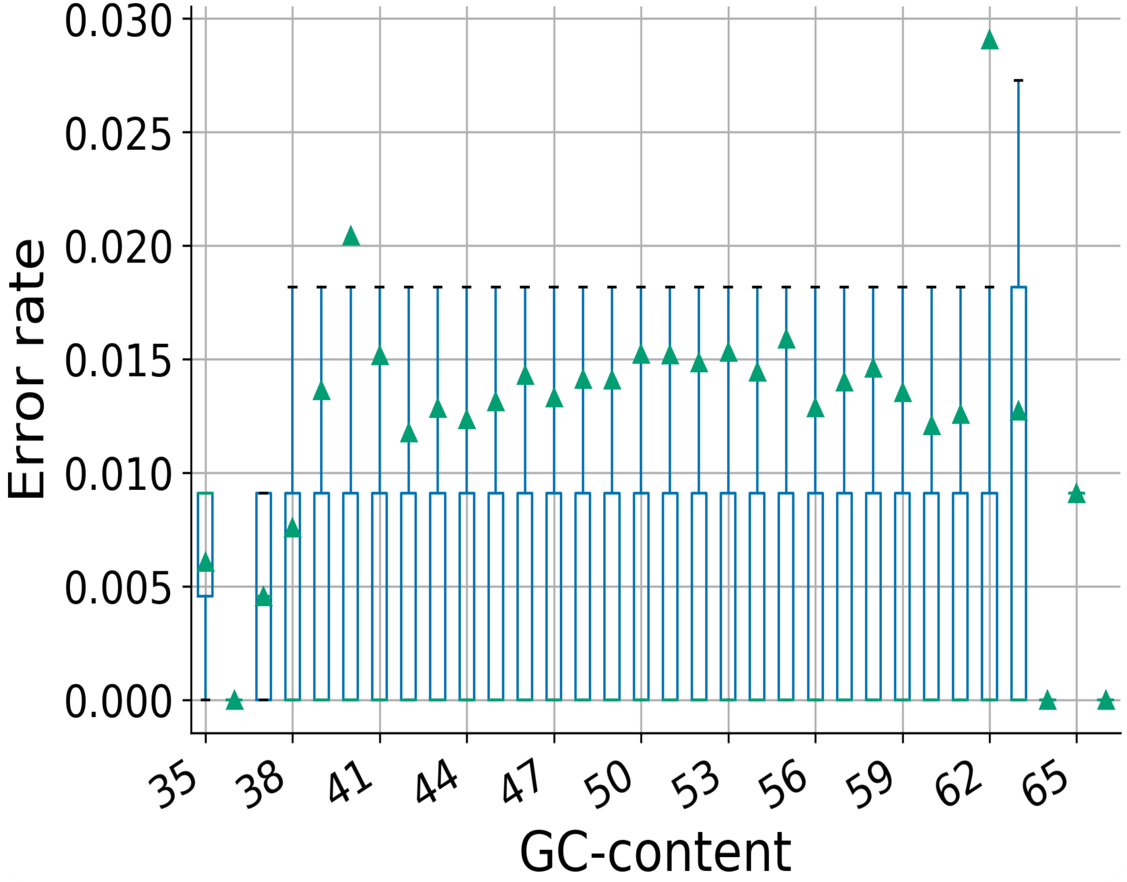
Yazdi et al.

# Error rates stratified by GC-content

## Erlich & Zielinski

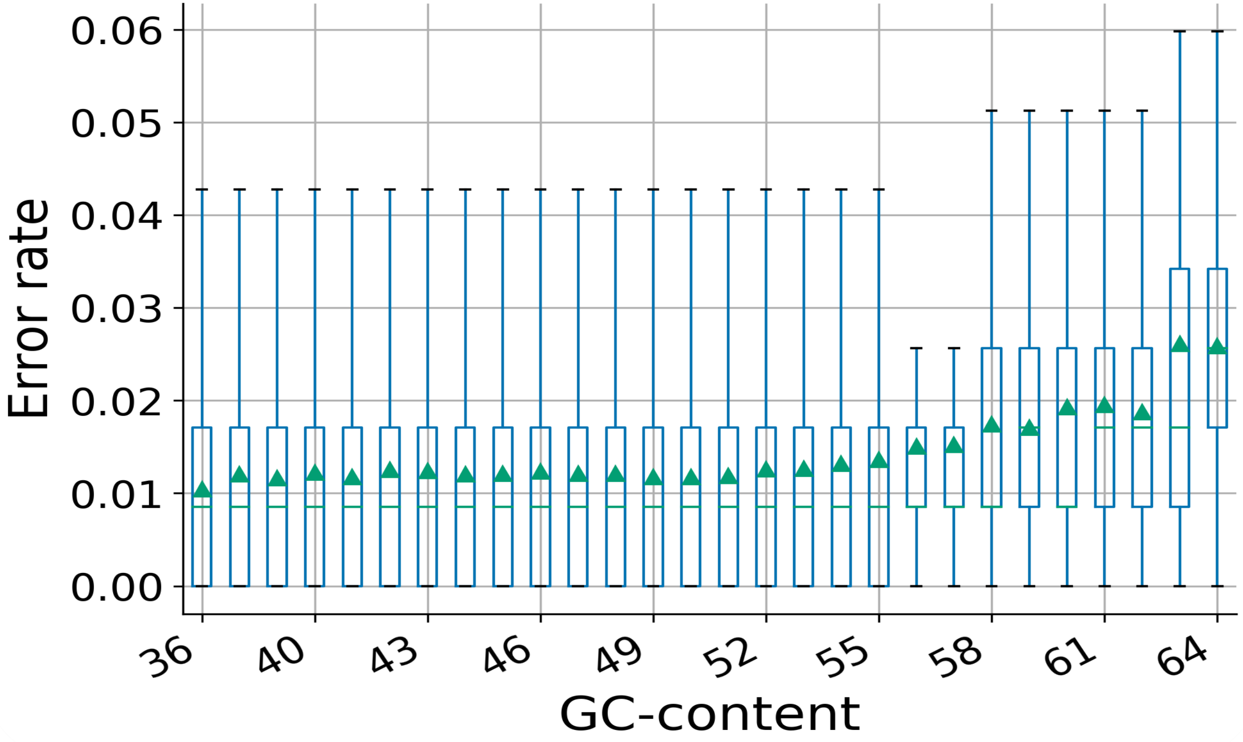


## Organick et al.

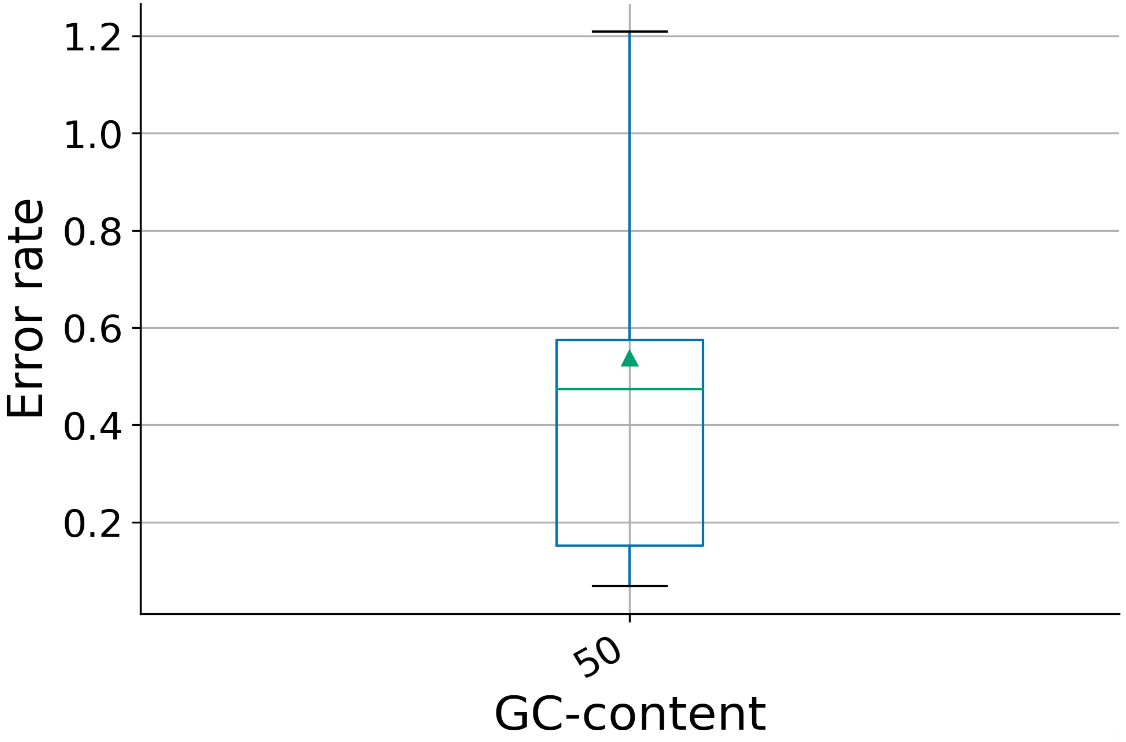


# Error rates stratified by GC-content

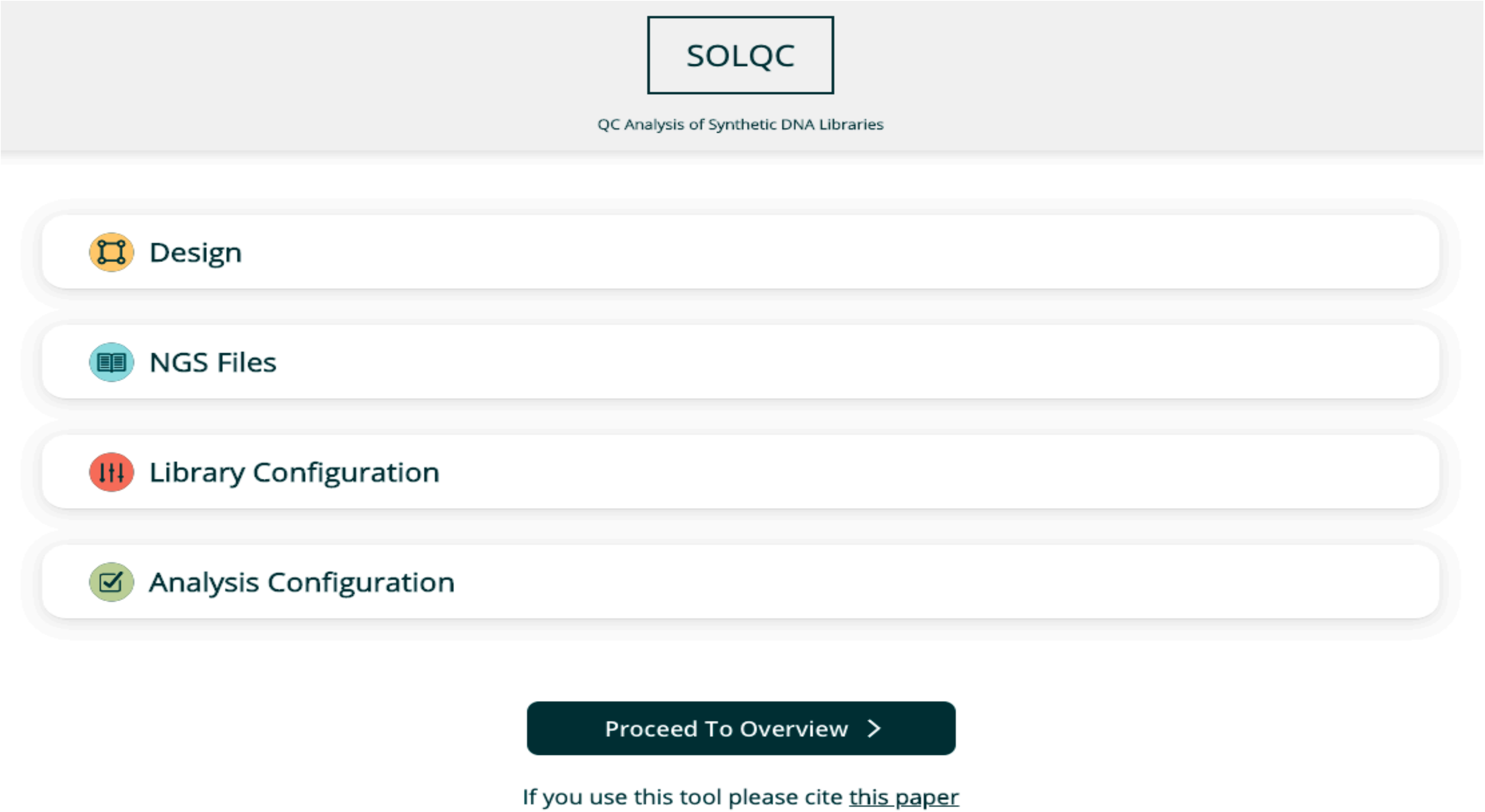
## Grass et al.



## Yazdi et al.



# Thank You!



SOLQC

QC Analysis of Synthetic DNA Libraries

If you use this tool please cite this paper