Simulating Noisy Channels in DNA Storage

Mayank Keoliya, Puru Sharma, Djordje Jevdjic
School of Computing
National University of Singapore

write process:

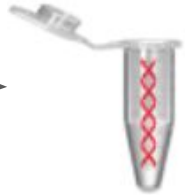
1011101...

encoding

ACGTCAG...

synthesis

Twist,
IDT



read process:

selecting

sequencing

Nanopore,
Illumina

ACGTCAG...

decoding

1011101...

DNA Storage Pipeline

Naive Method

Model described using P % error rate

$P_{insertion} = P / 3$

$P_{substitution} = P / 3$

$P_{deletion} = P / 3$

Ref: *Clustering Billions of Reads for DNA Data Storage*, Rashtchian et al. (2017)

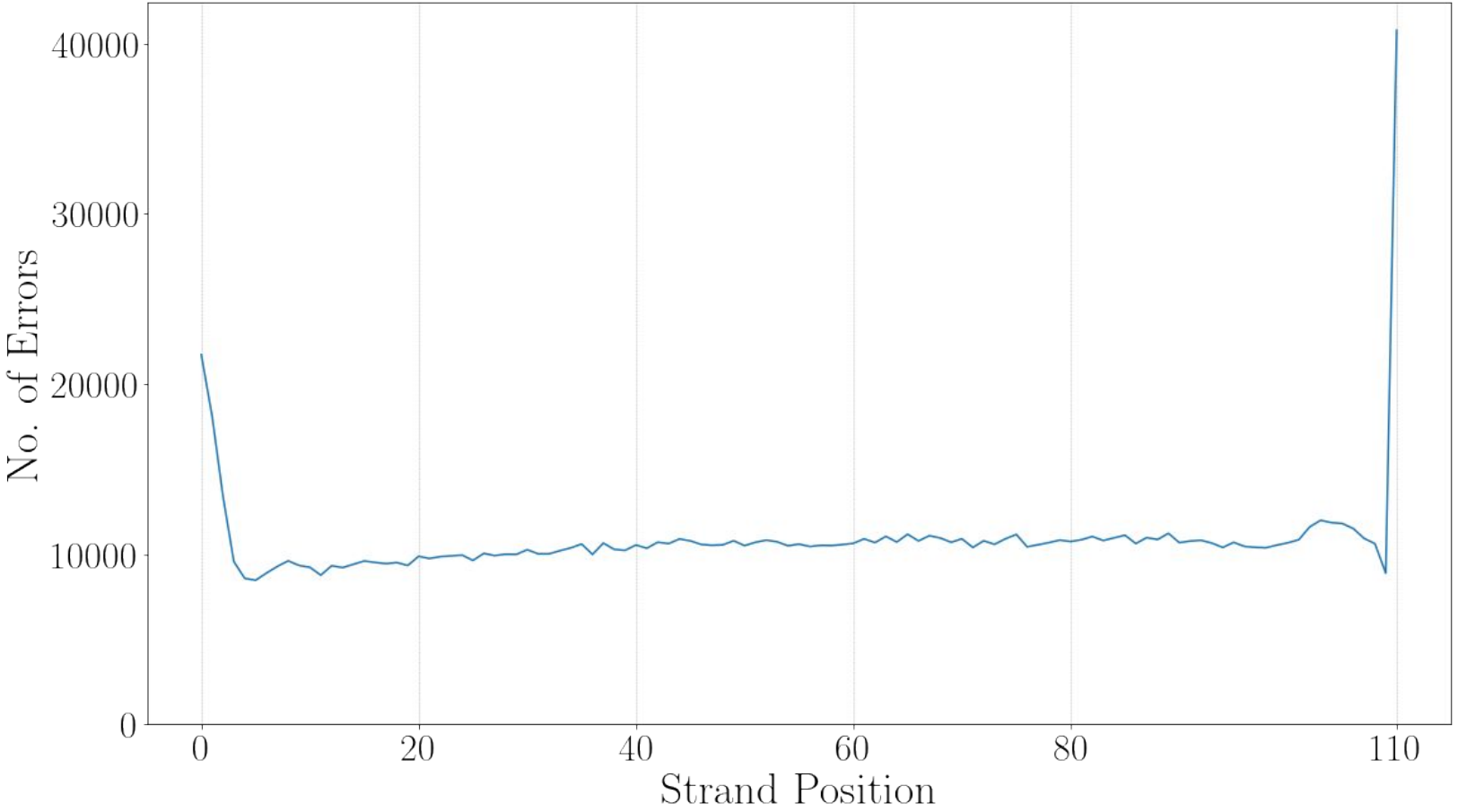
DNASimulator

Model uses prior sequenced datasets to calculate:

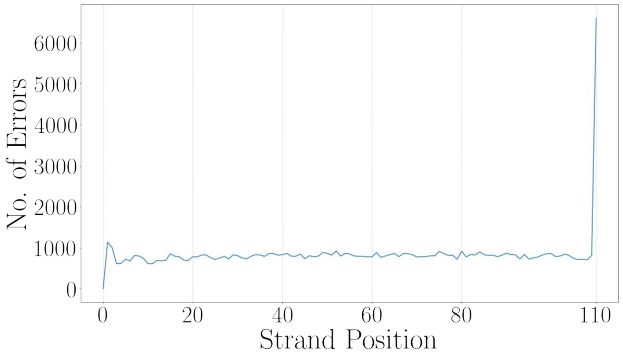
for each base *B* (= A, G, C or T),

$P_{ins \mid B}$, $P_{subs \mid B}$ and $P_{dels \mid B}$

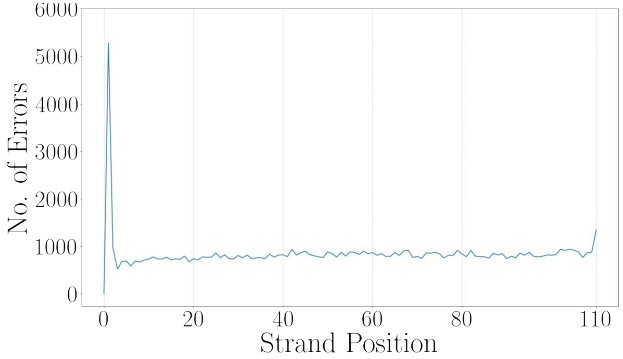
Ref: *DNA Storage Simulator*, Chaykin et al (2020).



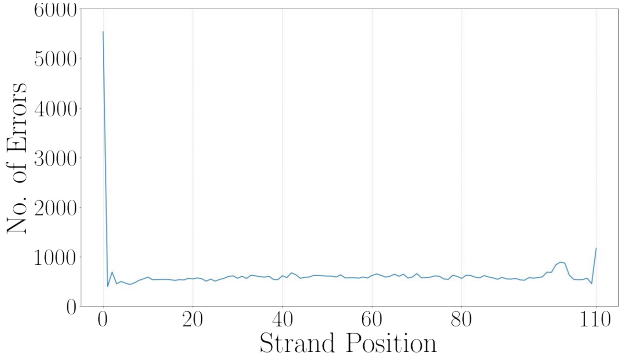
Per-Strand distribution of errors for Microsoft Nanopore dataset



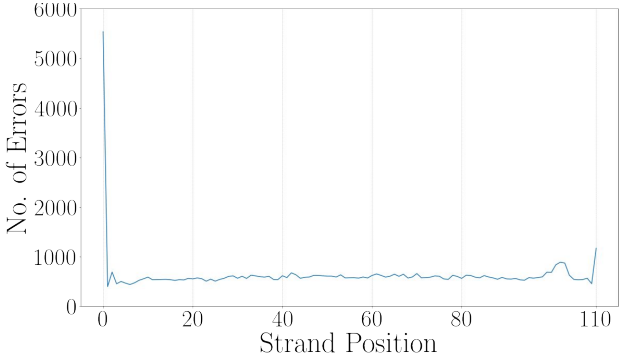
Deletion of G



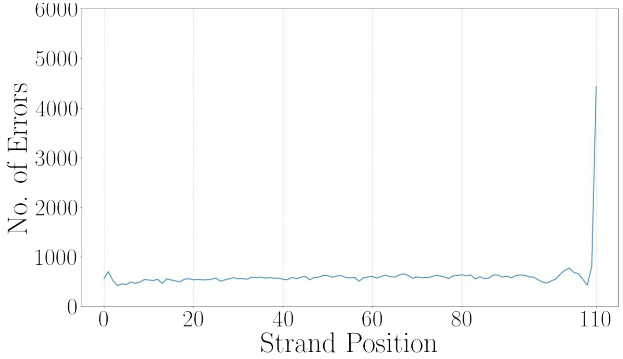
Deletion of C



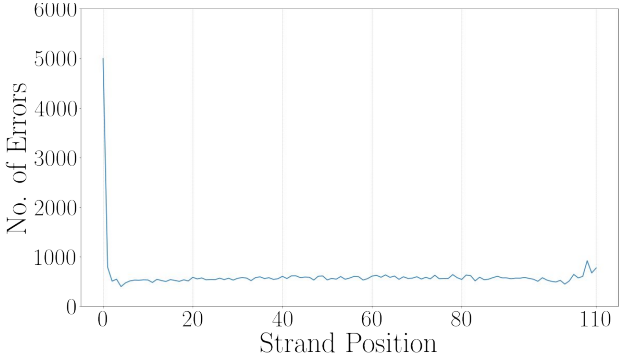
Deletion of T



Insertion of T



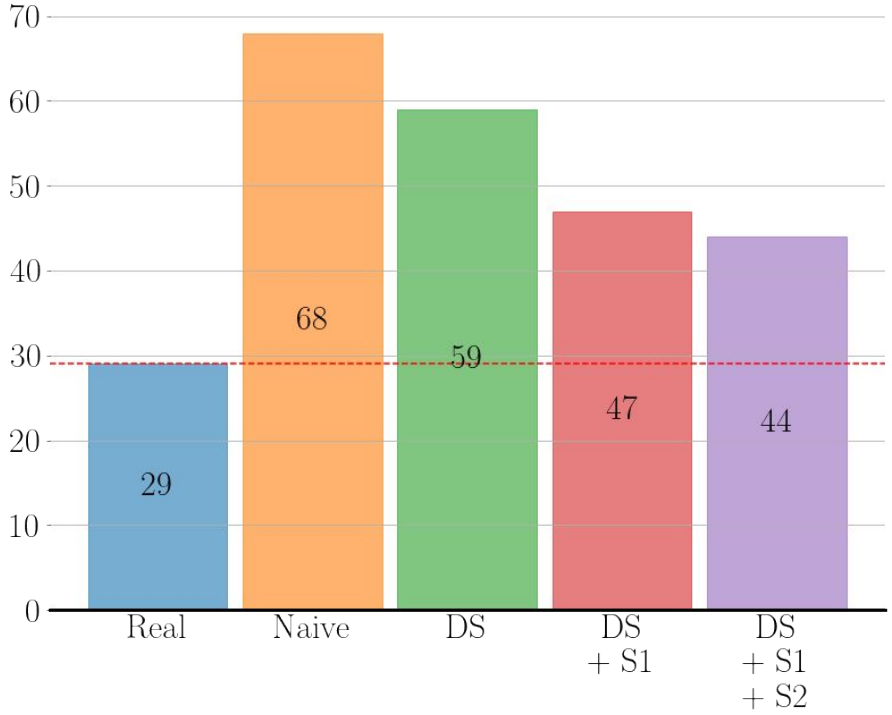
Insertion of G



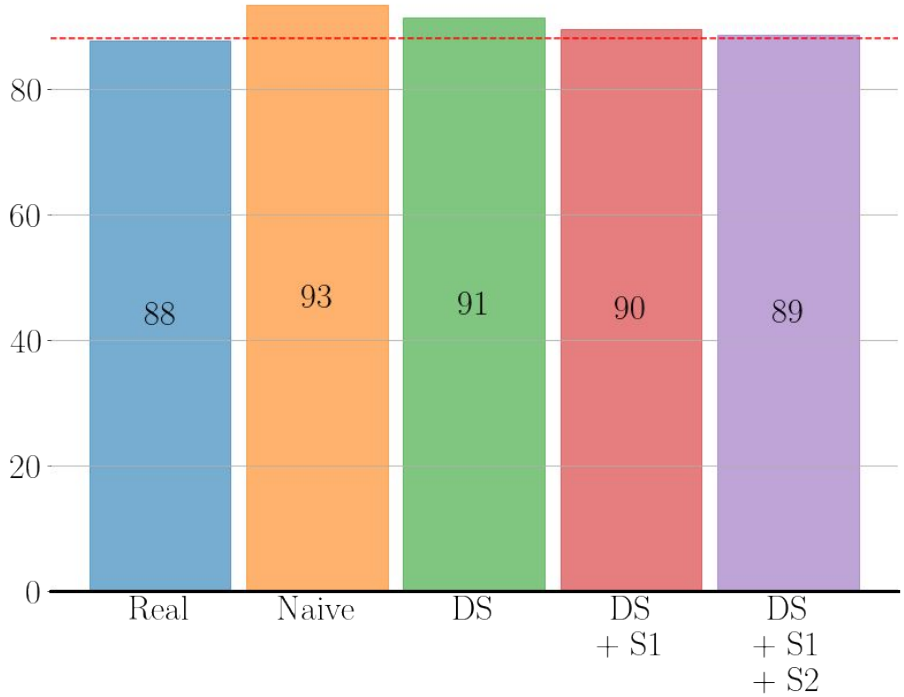
Insertion of C

Per-Strand distribution of errors **by error type** for Microsoft Nanopore dataset

Results



Per-Strand TR accuracy for datasets



Per-Char TR accuracy for datasets