DNA Storage Toolkit:
A Modular End-to-End DNA Data Storage Codec and Simulator

Puru Sharma

Gary Goh Yipeng

Bin Gao

Longshen Ou

Dehui Lin

Deepak Sharma

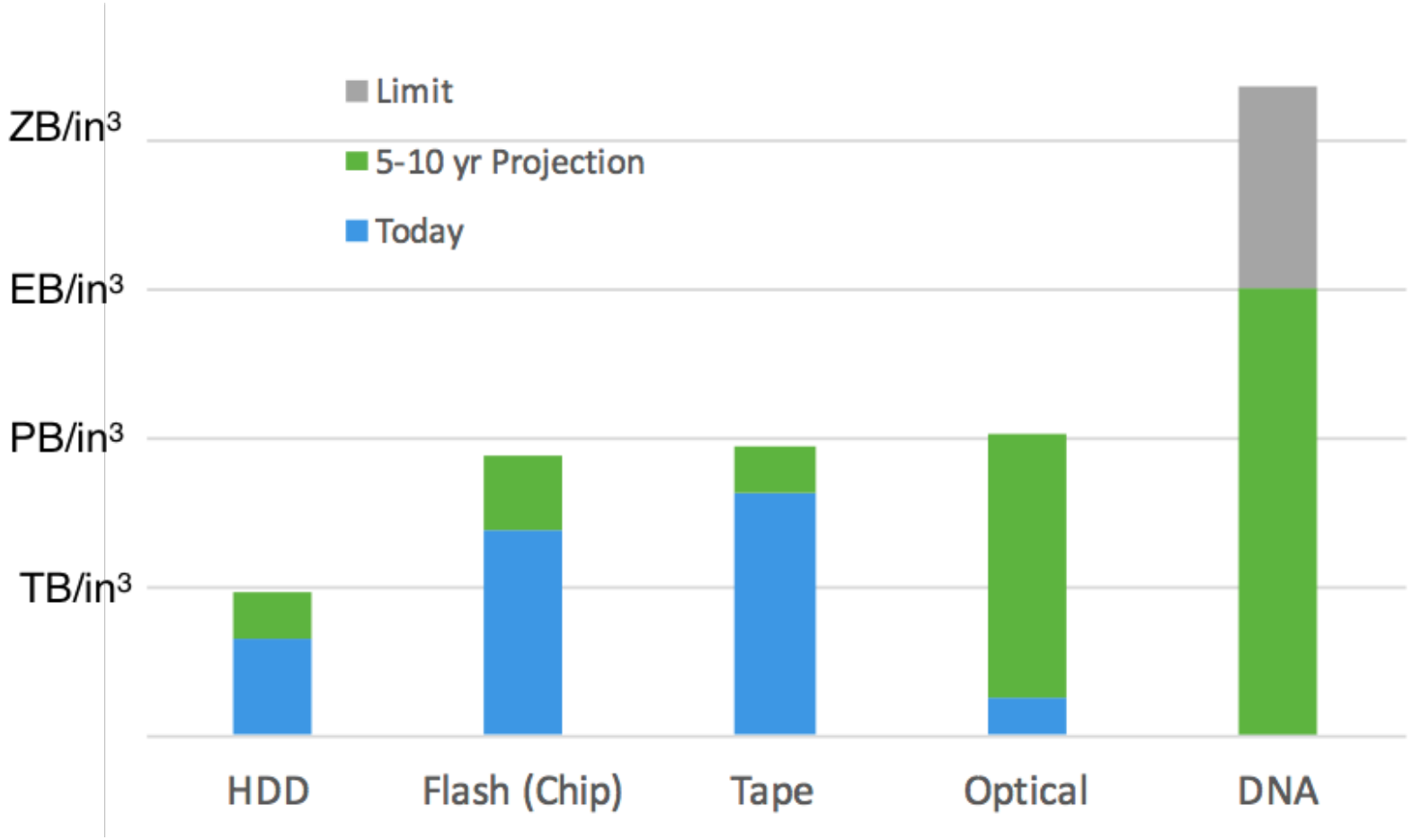
Djordje Jevdjic

NUS National University of Singapore

Why storing data in DNA molecules?

1. Incredible density
   - 6-7 orders of magnitude ahead of best alternatives!
2. Unmatched durability
   - Thousands/millions/billions of years (vs. 3-5 years for disks/flash)
3. Never obsolete: R/W interfaces will only improve with time
4. Efficient random access
5. Convenient for many data-parallel & near-data computations

Storage Density Projections*



*Credit: Luis Ceze & Karin Strauss, Microsoft

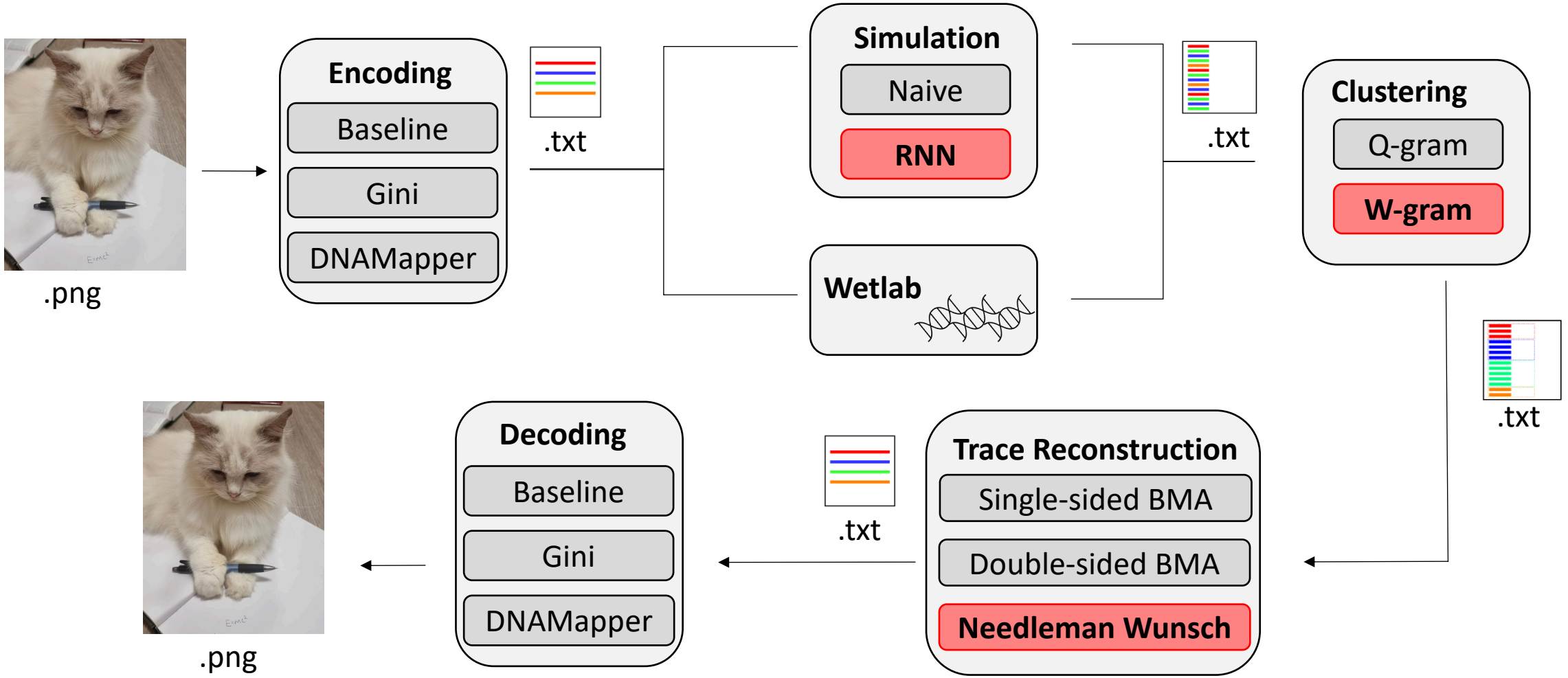
Key Problems with DNA Storage

1. Expensive R/W interfaces
   - Writing cost: $1K - $10K/MiB, reading cost: $10 - $10K/MiB

   → High cost also makes research very challenging
2. High read/write latency
   - Takes days to write DNA, hours (to days) to read (OK for archival storage)

   → But this delay slows down research
3. Extremely error-prone interfaces
   - Errors are very peculiar and hard to simulate
4. Requires expensive equipment and wetlab expertise
5. No complete open-source codec available

Our DNA Storage Pipeline

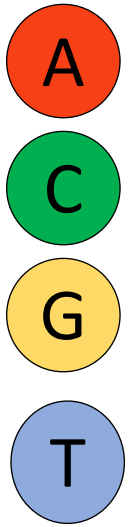


Outline

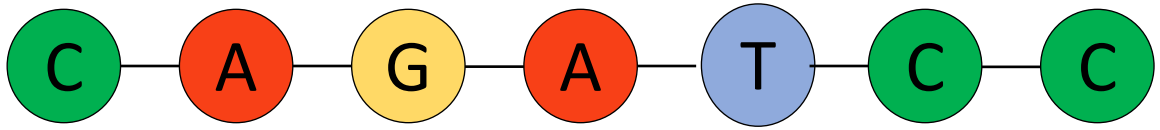
- Introduction
- **DNA Storage Basics**
- Our Toolkit
  - Encoding
  - Simulation
  - Clustering
  - Trace Reconstruction
  - Decoding
- Conclusion

DNA Molecules

4 nucleotides



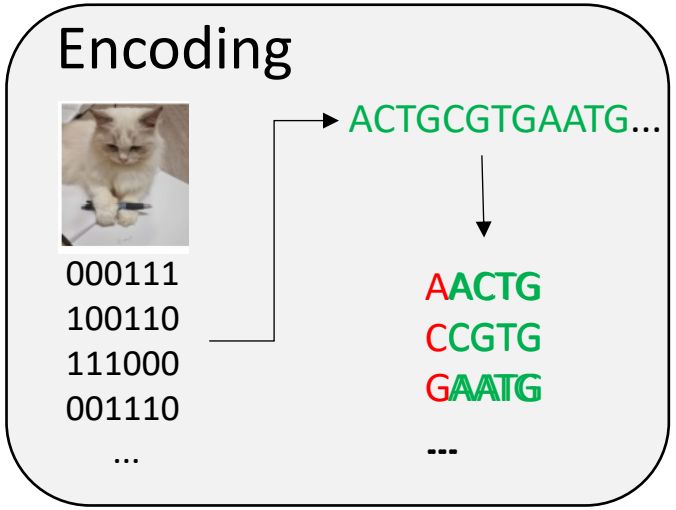
Synthetic DNA molecule



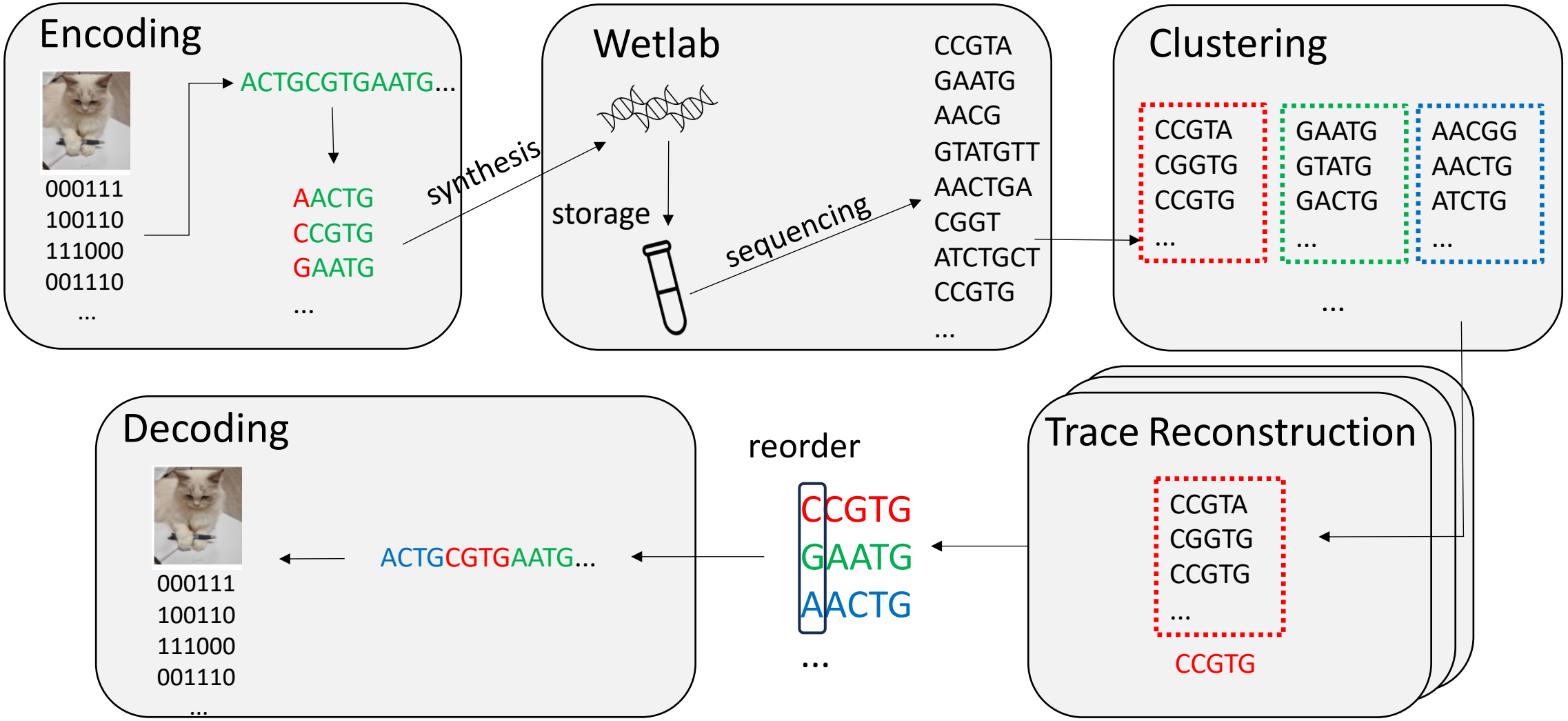
- Artificially created string of nucleotides
- No biological meaning

$log_2|\{A, C, G, T\}| = 2$ bits of data per nucleotide

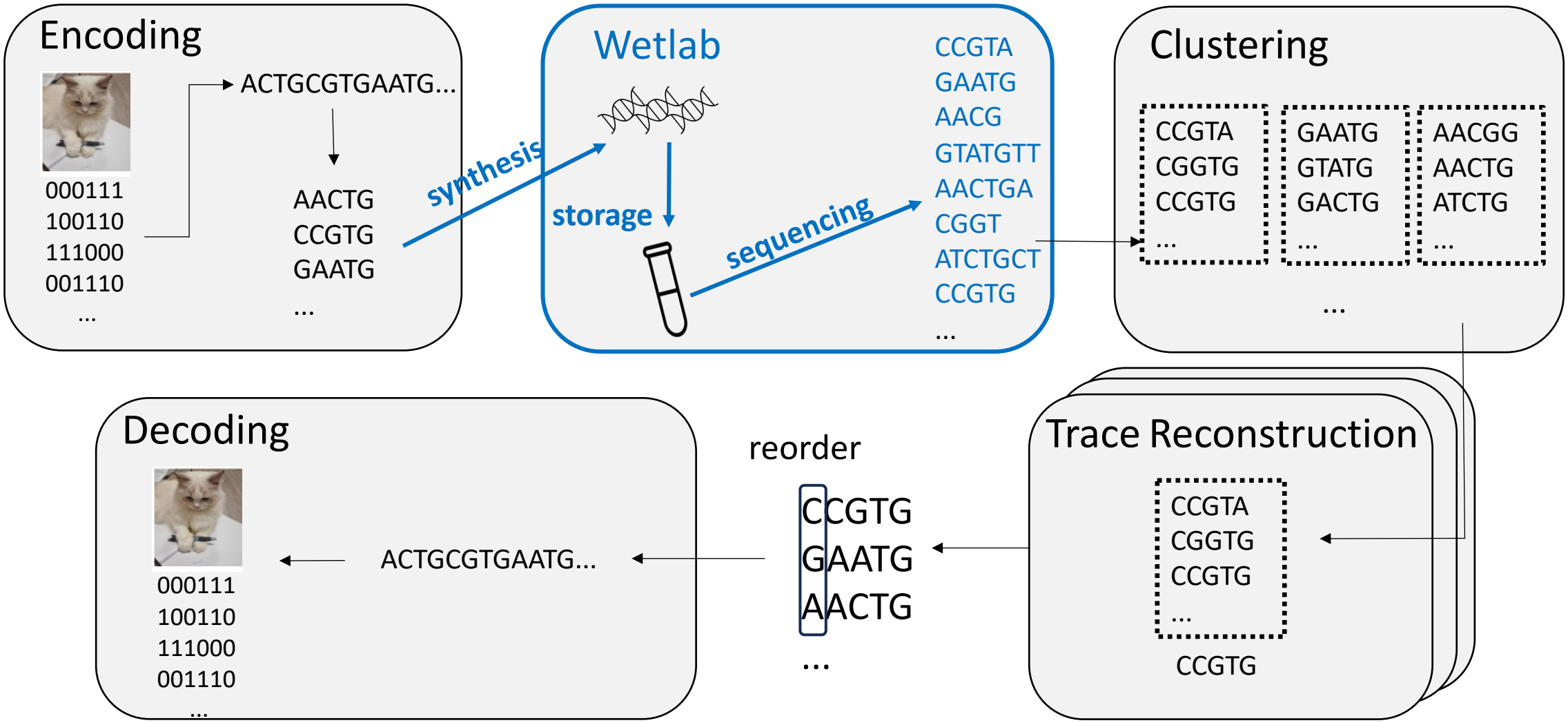
DNA Storage Pipeline



DNA Storage Pipeline



DNA Storage Pipeline – wetlab steps



Outline

- Introduction
- DNA Storage Basics
- **Our Toolkit**
  - Encoding
  - Simulation
  - Clustering
  - Trace Reconstruction
  - Decoding
- Conclusion

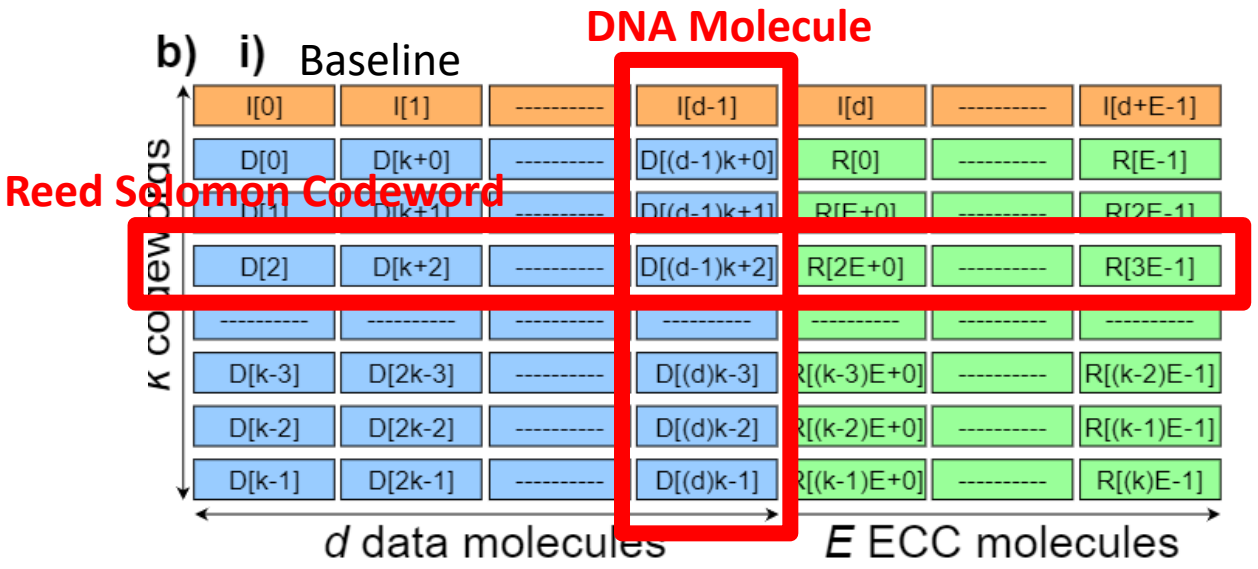
Encoding

3 encoding schemes provided.

i) Baseline: Introduced by Organick et al. [1]

ii) Gini: Introduced by Lin et al. [2]

iii) DNAMapper: Introduced by Lin et al. [2]



ii) Gini

I[0]	I[1]	----------	I[d-1]	I[d]	----------	I[d+E-1]
D[0]	D[k+0]	----------	D[(d-1)k+0]	R[0]	----------	R[E-1]
D[1]	D[k+1]	----------	D[(d-1)k+1]	R[E+0]	----------	R[2E-1]
D[2]	D[k+2]	----------	D[(d-1)k+2]	R[2E+0]	----------	R[3E-1]
----------	----------	----------	----------	----------	----------	----------
D[k-3]	D[2k-3]	----------	D[(d)k-3]	R[(k-3)E+0]	----------	R[(k-2)E-1]
D[k-2]	D[2k-2]	----------	D[(d)k-2]	R[(k-2)E+0]	----------	R[(k-1)E-1]
D[k-1]	D[2k-1]	----------	D[(d)k-1]	R[(k-1)E+0]	----------	R[(k)E-1]

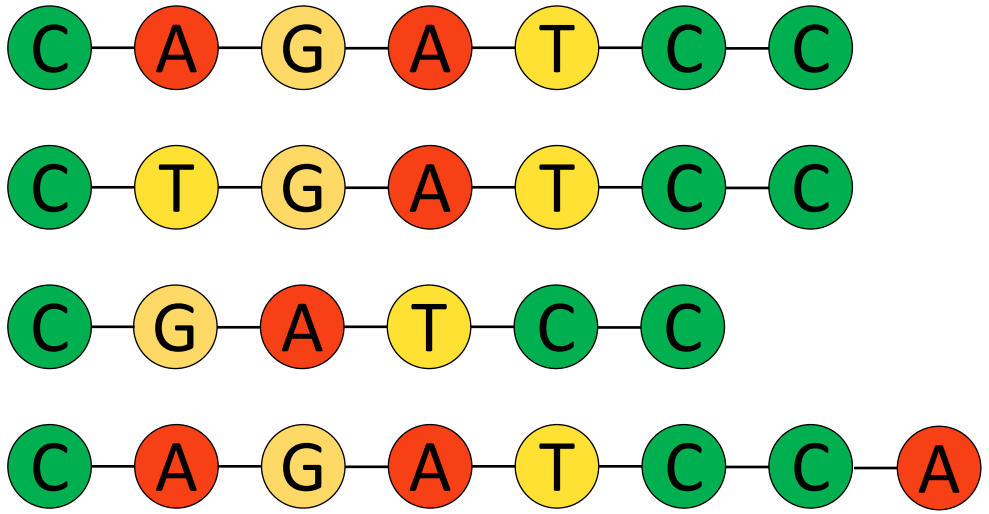
iii) DNA Mapper

I[0]	I[1]	----------	I[d-1]	I[d]	----------	I[d+E-1]
P[d+0]	P[d+1]	----------	P[2d-1]	R[0]	----------	R[E-1]
P[3d+0]	P[3d+1]	----------	P[4d-1]	R[E+0]	----------	R[2E-1]
P[5d+0]	P[5d+1]	----------	P[6d-1]	R[2E+0]	----------	R[3E-1]
----------	----------	----------	----------	----------	----------	----------
P[4d+0]	P[4d+1]	----------	P[5d-1]	R[(k-3)E+0]	----------	R[(k-2)E-1]
P[2d+0]	P[2d+1]	----------	P[3d-1]	R[(k-2)E+0]	----------	R[(k-1)E-1]
P[0]	P[1]	----------	P[d-1]	R[(k-1)E+0]	----------	R[(k)E-1]

[1] Organick et al., "Random Access in large-scale DNA data storage" Nature Biotechnology, 2018

[2] Lin et al., "Managing Reliability Skew in DNA Storage" ISCA, 2022

Wetlab



Original

Substitution (A→T)

Deletion (of A)

Insertion (of A)

Sequencing produces many (buggy) copies of each molecule:

synthesized

CAGATCC

Simulator?

sequenced

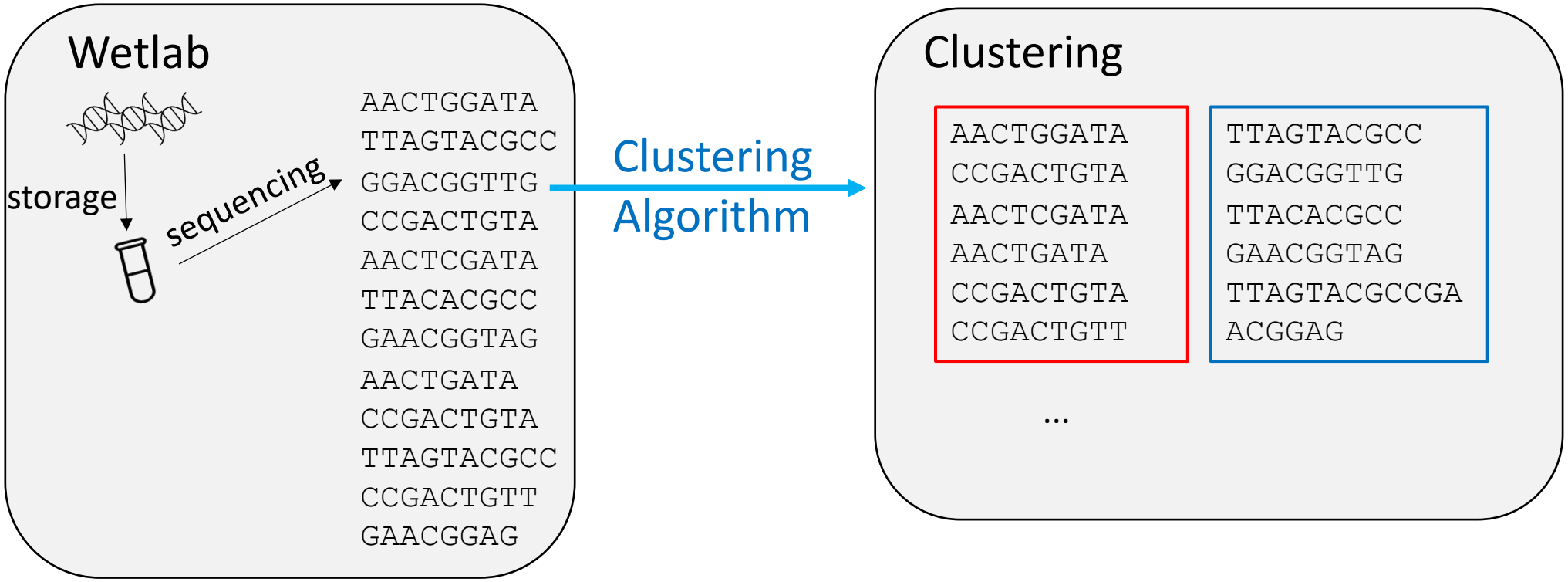
CAGATCC

CAGATC

AAGATCCA

AGATTCC

Clustering


Wetlab
storage
sequencing
AACTGGATA
TTAGTACGCC
GGACGGTTG
CCGACTGTA
AACTCGATA
TTACACGCC
GAACGGTAG
AACTGATA
CCGACTGTA
TTAGTACGCC
CCGACTGTT
GAACGGAG
Clustering
Algorithm
Clustering
AACTGGATA
CCGACTGTA
AACTCGATA
AACTGATA
CCGACTGTA
CCGACTGTT
TTAGTACGCC
GGACGGTTG
TTACACGCC
GAACGGTAG
TTAGTACGCCGA
ACGGAG
...

Clustering

- Using edit distance for clustering is too slow.
- We can approximate using q-gram binary signatures and Hamming distance.

	CT	GA	GT	
ACGTCGAAC	0	1	1	} Low Hamming distance
ACGTCCGAAC	0	1	1	

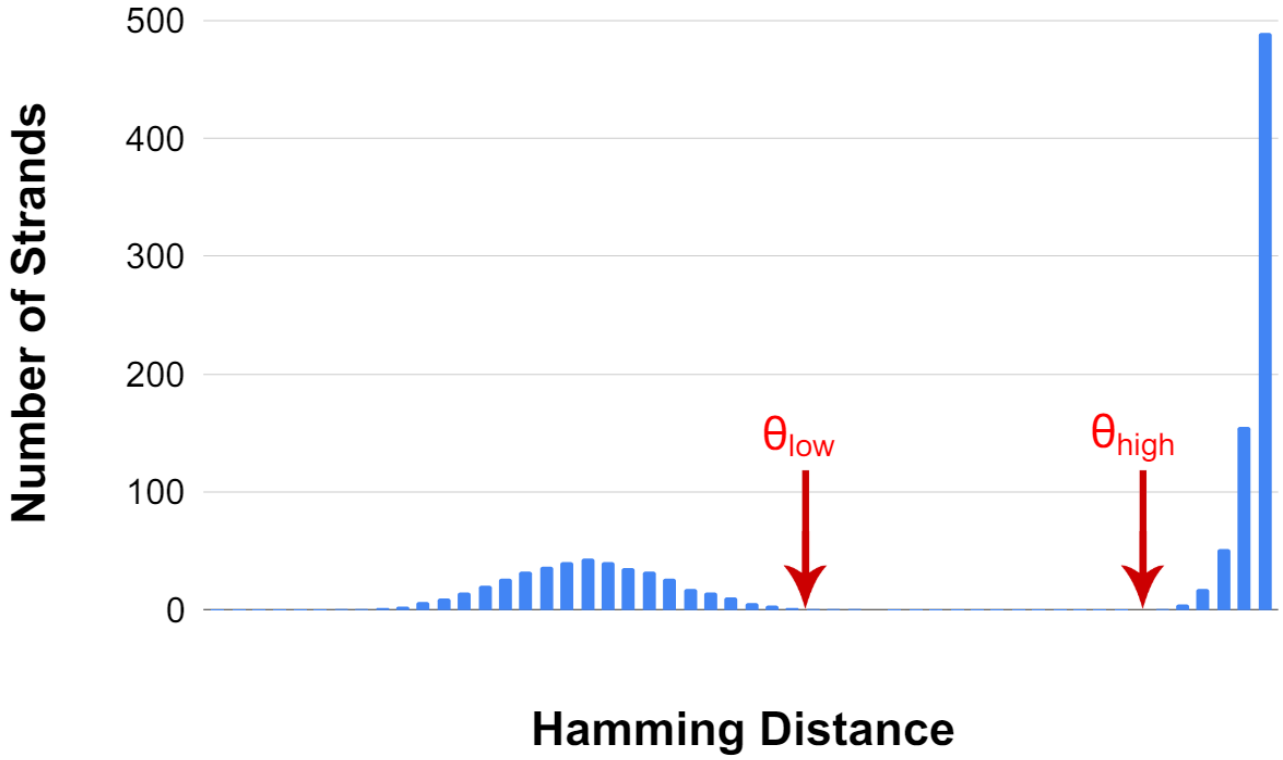
ACGTCGAAC	0	1	1	} High Hamming distance
TACCTATTCC	1	0	0	

[3] Rashtchian et al., “Clustering Billions of Reads for DNA Data Storage” Neurips, 2017

Autotuning the threshold parameters

The threshold parameters for the binary signature Hamming distance need to be tuned based on dataset.

Take a tiny sample and plot the Hamming distance for them:



Trace Reconstruction

GTACCAGTCGAGTAAAGC

GCCGTGCGTAAGCT

GTACAATGTCGTGTAAC

GTCATGGTCAGTAAGC

GTACAGTCCGTAAAGC

TACGTGTATAGC

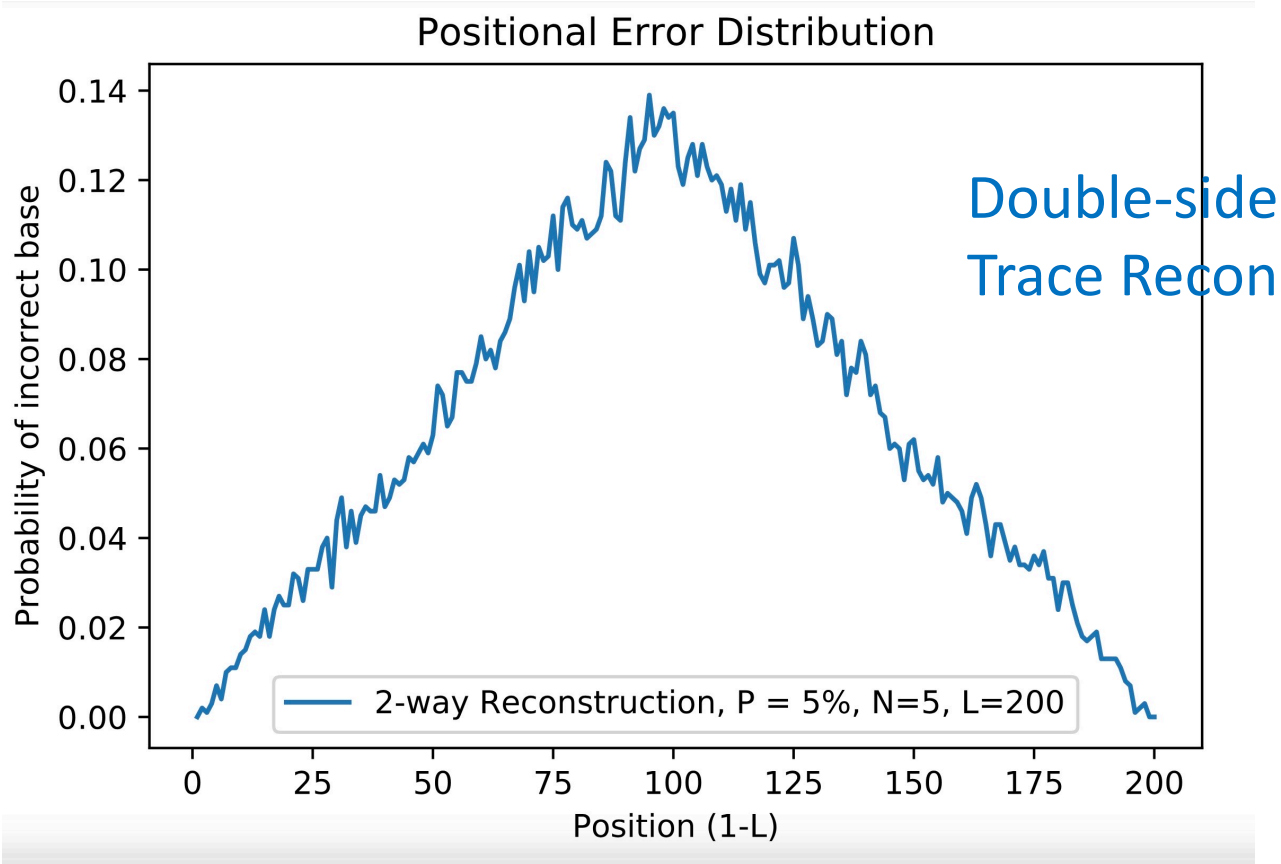
GATACAGCACGTGAAGC

What is the consensus?

Trace Reconstruction

Trace Reconstruction

Trace Reconstruction



Double-sided Bitwise Majority Trace Reconstruction

[2] Lin et al., “Managing Reliability Skew in DNA Storage” ISCA, 2022

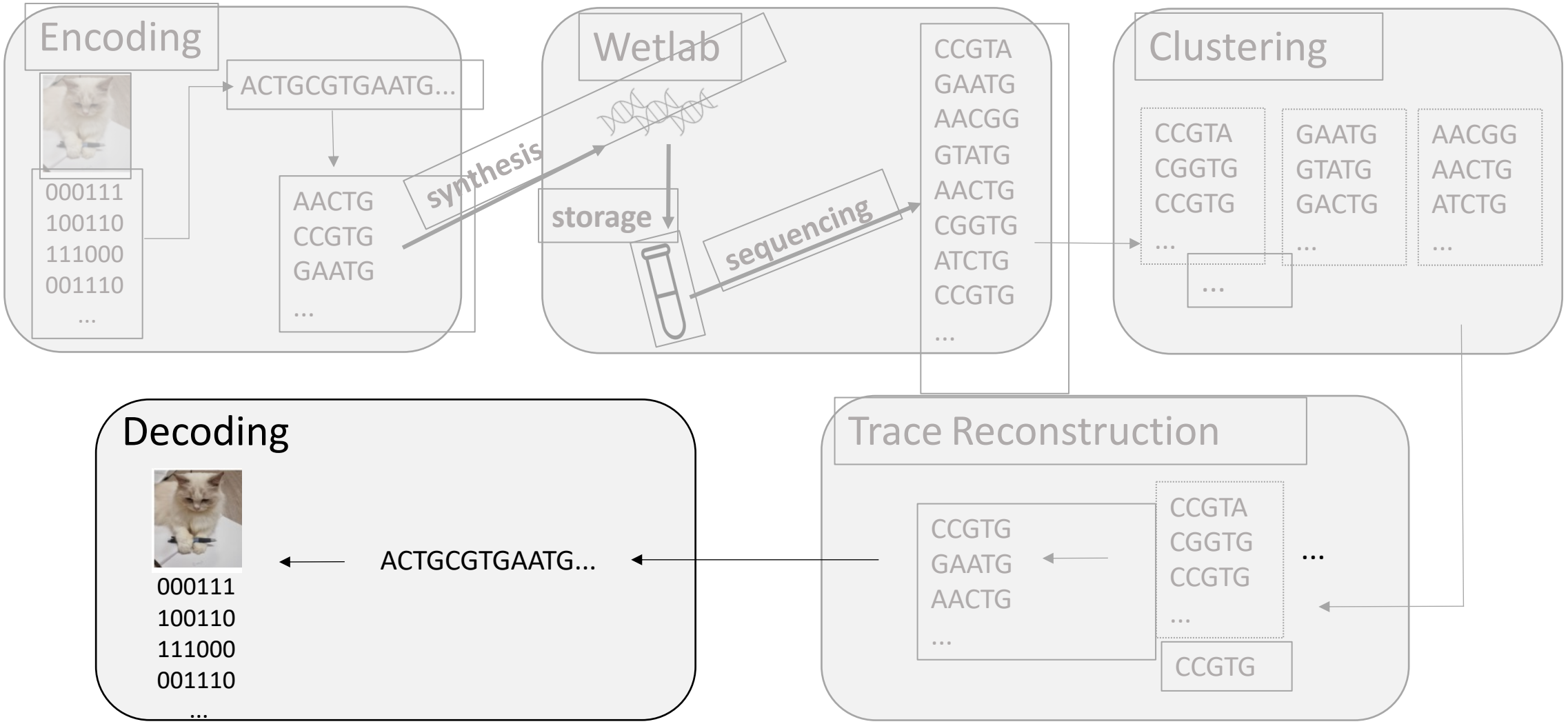
Needleman-Wunsch

```
Trace1        ACGTTGAA…
Trace2        AGTTGACA…
Trace3        AACGTTGAA…
Trace4        ACGTTCAA…
   …             …
```

```
Alignment { A-CGTTGA_A…
          { A--GTTGACA…
          { AACGTTGA_A…
          { A-CGTTCA_A…
```

```
Consensus   ACGTTGAA
```

Decoding


Encoding
ACTGCGTGAATG...
000111
100110
111000
001110
...
AACTG
CCGTG
GAATG
...
synthesis
Wetlab
storage
sequencing
CCGTA
GAATG
AACGG
GTATG
AACTG
CGGTG
ATCTG
CCGTG
...
Clustering
CCGTA
CGGTG
CCGTG
...
GAATG
GTATG
GACTG
...
AACGG
AACTG
ATCTG
...
...
Trace Reconstruction
CCGTG
GAATG
AACTG
...
CCGTA
CGGTG
CCGTG
...
...
CCGTG
Decoding
ACTGCGTGAATG...
000111
100110
111000
001110
...

Wetlab Simulation

Sequencing produces many (buggy) copies of each molecule:

synthesized

CAGATCC

sequenced

CAGATCC

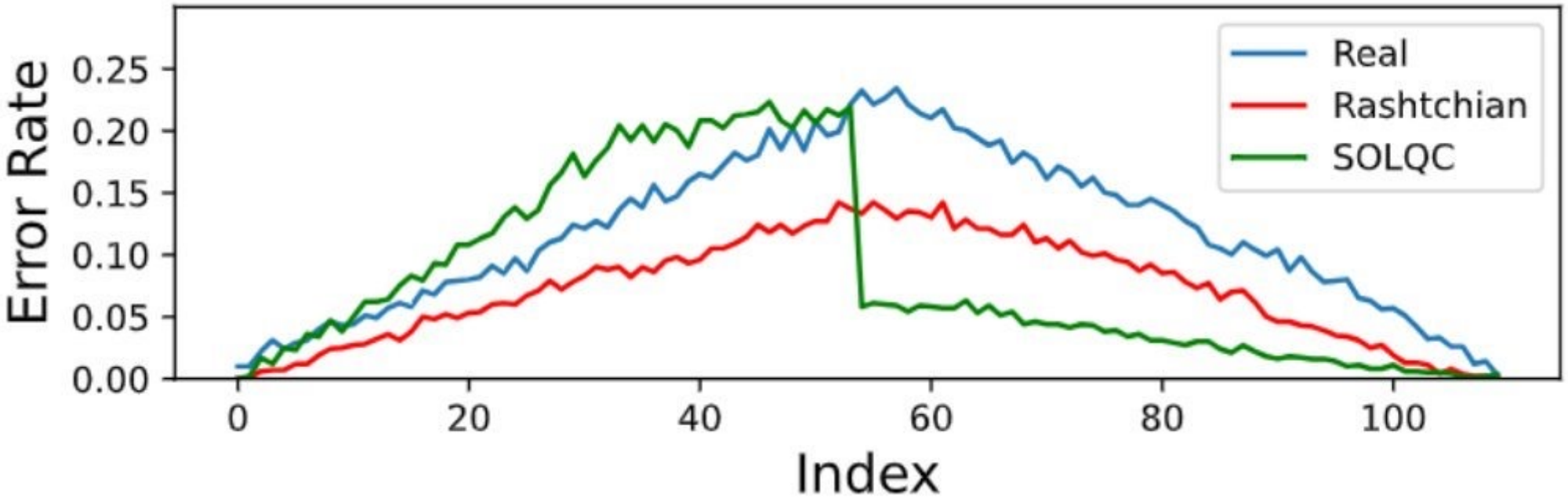
CAGATC

AAGATCCA

AGATTCC

Wetlab Simulation

We evaluate the simulation by comparing how the Trace Reconstruction module performs on the simulated data.



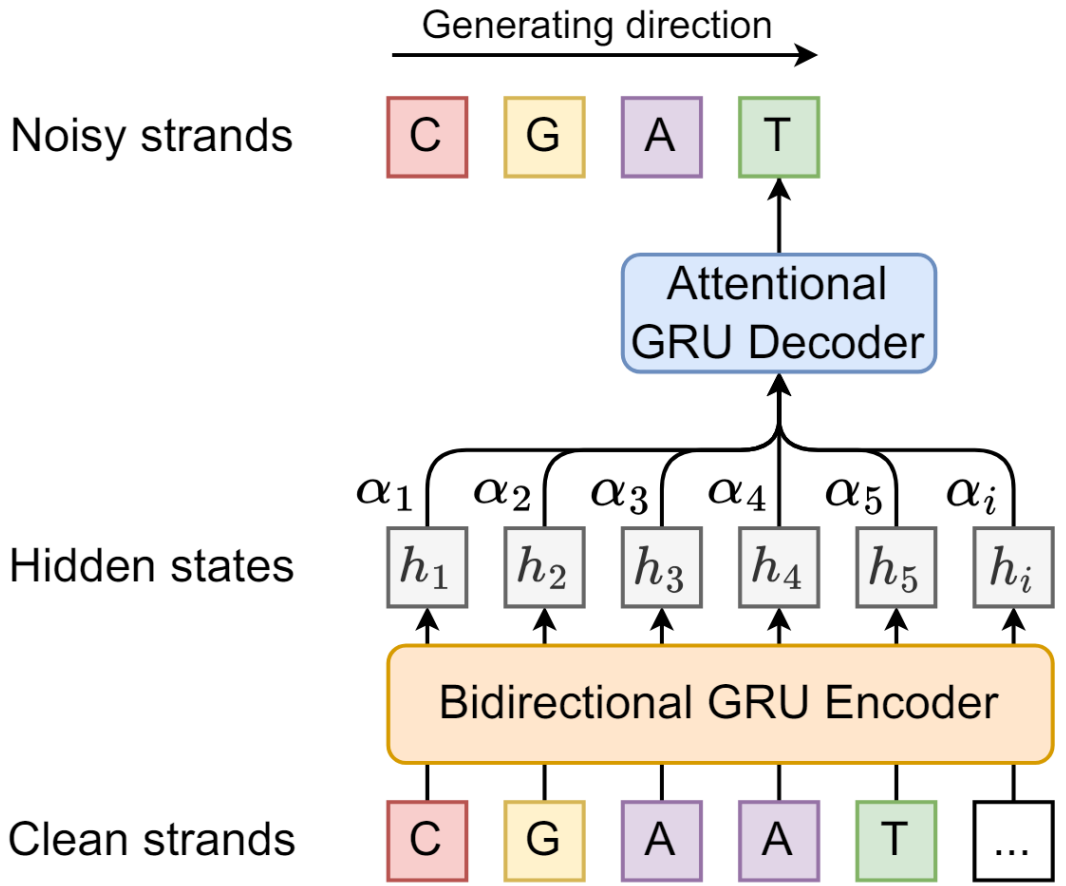
Our Simulation

Sequence-to-sequence problem.

Model based on basic NMT model [3].

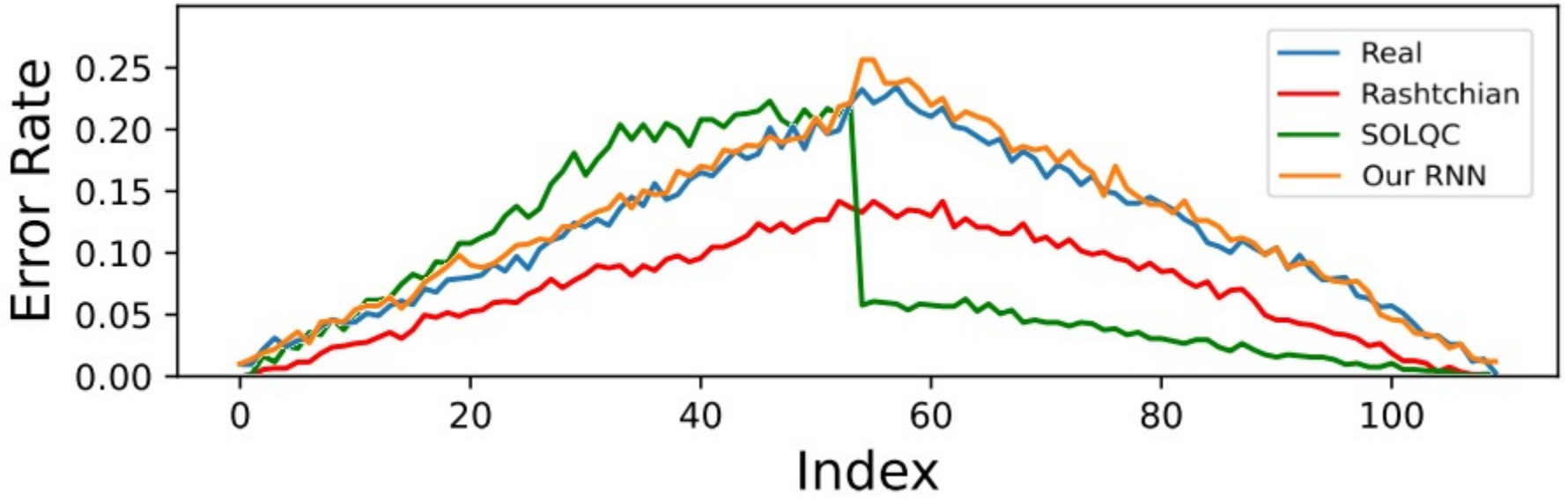
Attention based encoder-decoder.

[Hidden layer size = 128, Greedy sampling.]



[3] Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate" ICLR 2015

Our Simulation

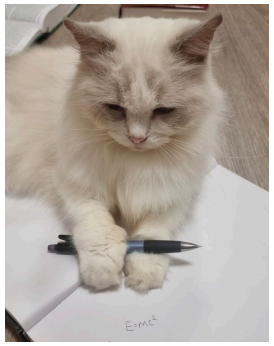


The performance of the reconstruction module on our simulated data is very similar to real data.

We can evaluate later modules of this pipeline using our simulated data!

Toolkit

https://github.com/prongs1996/DNAStorageToolkit/



.png

Encoding

- Baseline
- Gini
- DNAMapper

.txt

Simulation

- Naive
- **RNN**

Wetlab

.txt

Clustering

- Q-gram
- **W-gram**

.txt

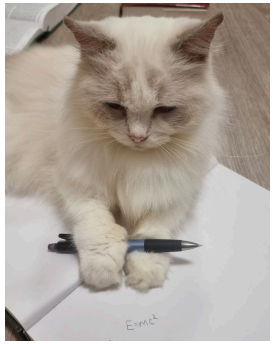
Trace Reconstruction

- Single-sided BMA
- Double-sided BMA
- **Needleman Wunsch**

.txt

Decoding

- Baseline
- Gini
- DNAMapper



.png

Conclusion

- First open end-to-end DNA storage toolkit
- Very accurate simulator for wetlab steps

Repository: https://github.com/prongs1996/DNAStorageToolkit/

Thank you! Questions?

Puru Sharma

Gary Goh Yipeng

Bin Gao

Longshen Ou

Dehui Lin

Deepak Sharma

Djordje Jevdjic

NUS National University of Singapore