

Kmer2Reads, an associative index for Third Generation Sequencing data

Léa Vandamme, Bastien Cazaux and Antoine Limasset

Univ. Lille, CNRS, UMR 9189 - CRISTAL, F-59000 Lille

SeqBIM

November 21, 2023

Genome assembly is complex

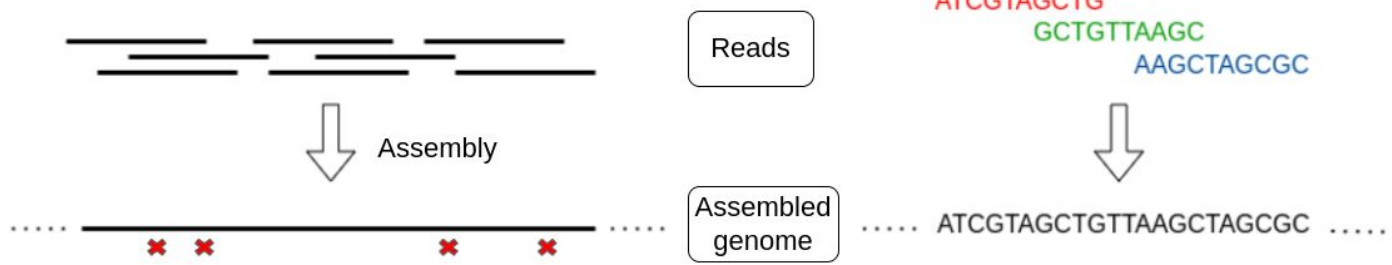


Long reads properties :

- Error rate : from ~0.1% to ~10%
- Length : 10 - 100 kilo bases

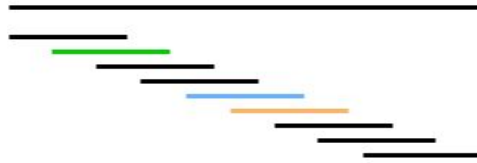


Hard task



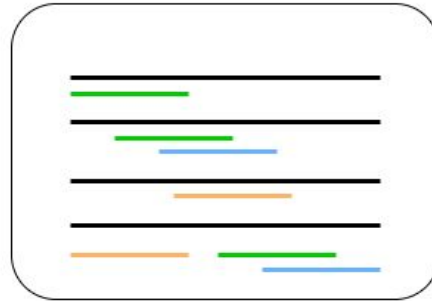
K-mer to reads index

Sequence of interest

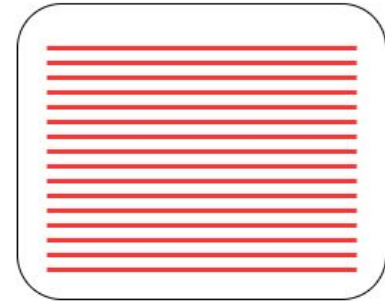


Cutted into k-mers

Reads of interest



Other reads



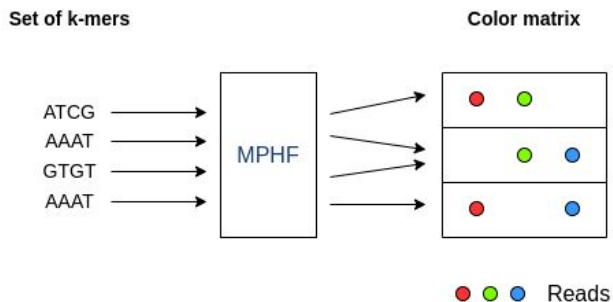
Select **reads** similar to our query **sequence**

↳ Select reads **sharing k-mers** with our query sequence

↳ Use **k-mer to reads** index

State of the art

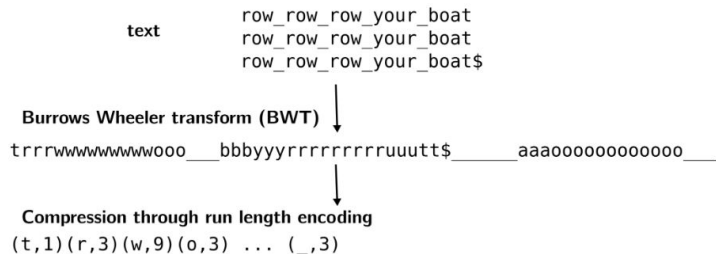
- Hashing methods (based on hash tables / MPMF)



Use example: Pufferfish (Almodaresi, 2018), SRC (Marchet, 2020), BLight (Marchet, 2021),






- Full text indexing (locate occurrences of pattern in a text)

Locate occurrences of patterns in a text



Variant of FM-index : r-index (Mun, 2020)

State of the art

	Construction	Memory	Debit
Hashing methods (SRC)			
Full-text indexing (r-index)			~

Challenge :

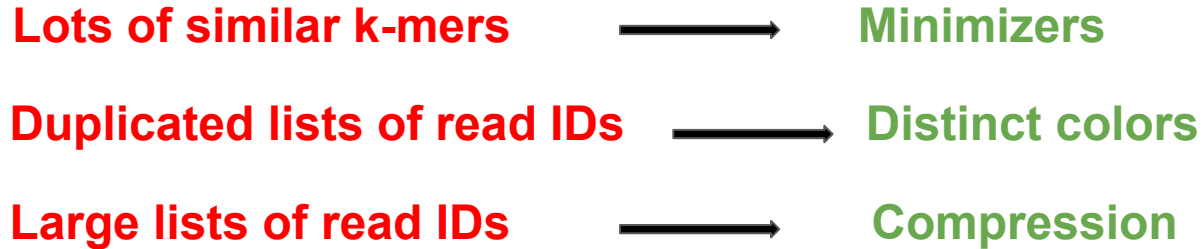
Scale to large genomes (Human genome = 100 million bp, 10 million reads)

- Limit memory and time cost

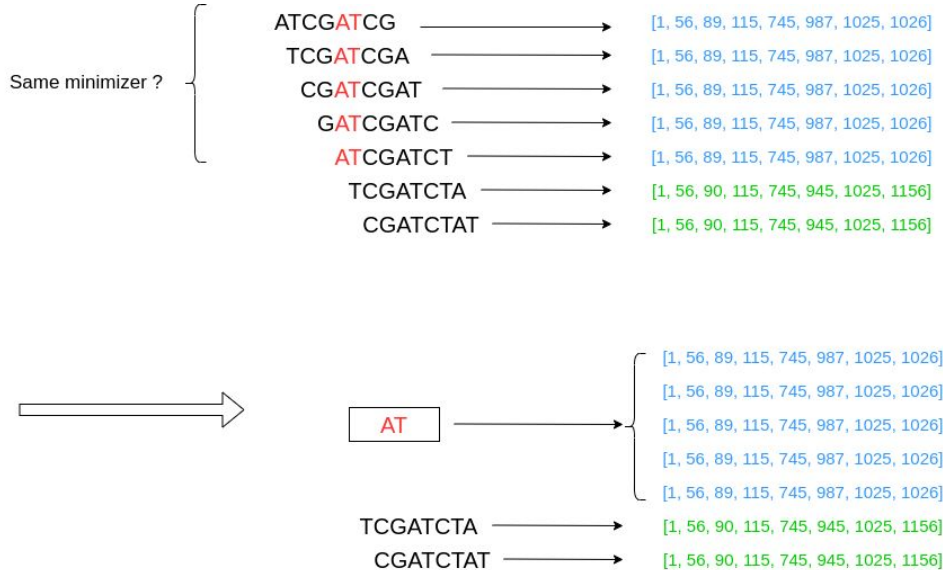
A naive version ?



Our contributions



Similar k-mers ?



Ecoli (error rate 1%, coverage = 100X, read length = 10000) :

- K-mers : 4,553,982
- M-mers : 443,123

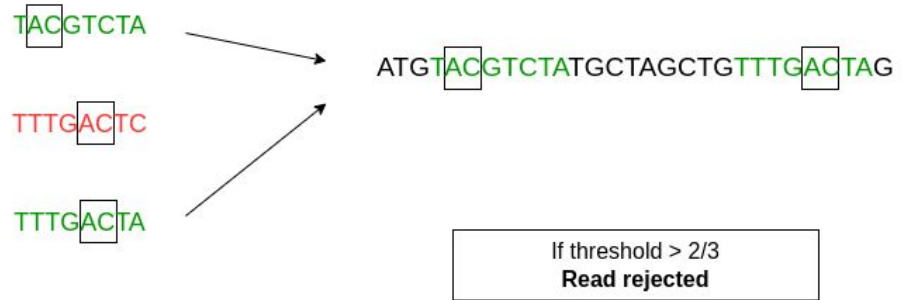
1 order of magnitude

False positives ?

Solution : count shared k-mers between potential output reads and query sequence.

3 k-mers
Same minimizer

Only 2 found in the read



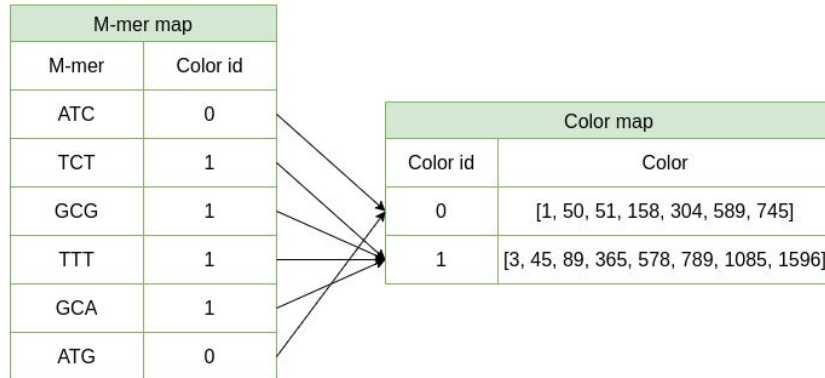
Dataset : Ecoli (error rate 1%, coverage = 100X, read length = 10000), seuil : 0.6
Query : sequence of length 50bp, extracted from a read

- Nombre reads total : 46.396
- Potential hits : 91
- Reads deleted : 1

Duplicated colors ?

Color = List of read IDs

Index **distinct** colors



Dataset : Ecoli (error rate 1%, coverage = 100X,
read length = 10000), seuil : 0.6
Query : sequence of length 50bp, extracted from a
read

- M-mer number : 443,123
- Color number : 88,345

1 order of magnitude

Large ID lists ?

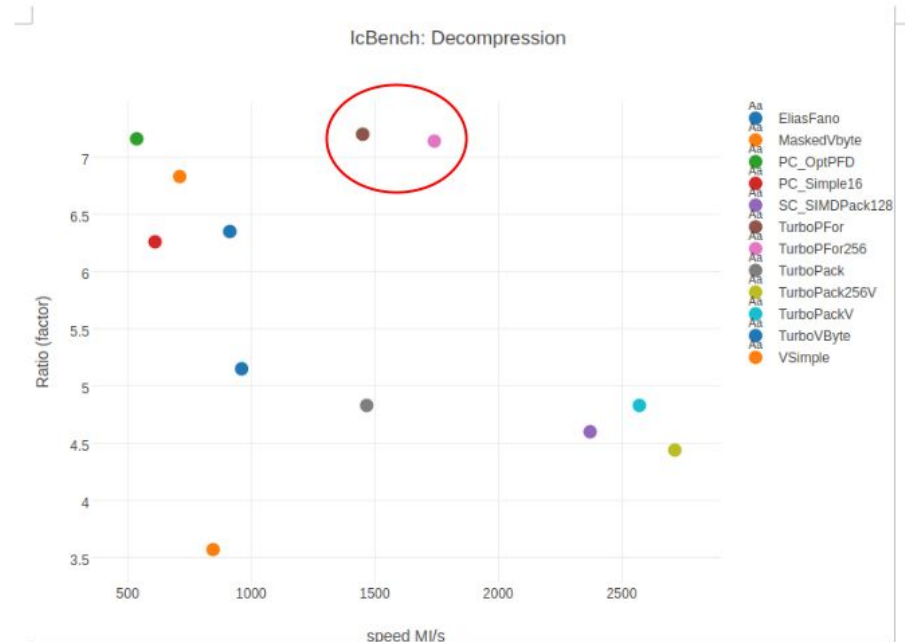
TurboPFor

Delta encoding :

[1, 1000, 1005, 1024, 1025, 1034]

-> [1, 999, 5, 19, 1, 9]

1 order of magnitude



Our tool : K2R

Error rate 1%, coverage = 100X, read length = 10.000

Dataset	K-mer number	M-mer number	Color number
E.Coli	4,553,982	443,123	88,345
C.Elegans	94,006,409	6,338,436	4,475,066

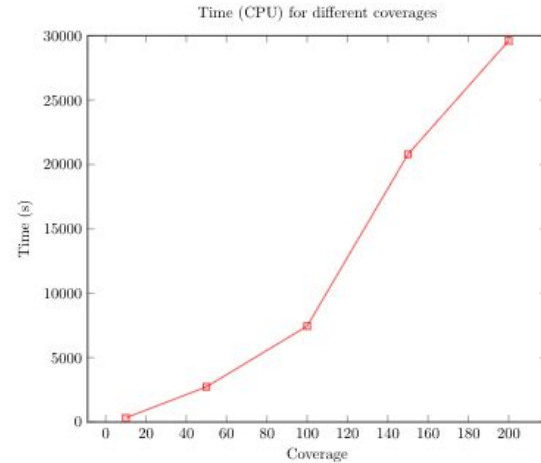
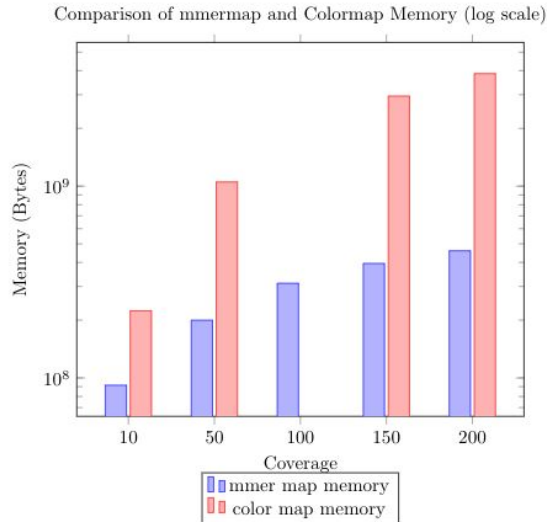
Results : memory & time

Index creation

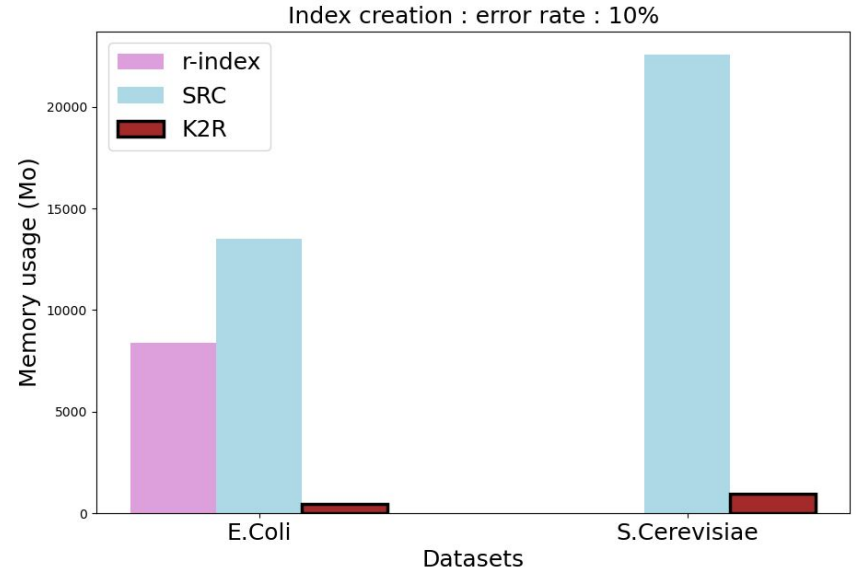
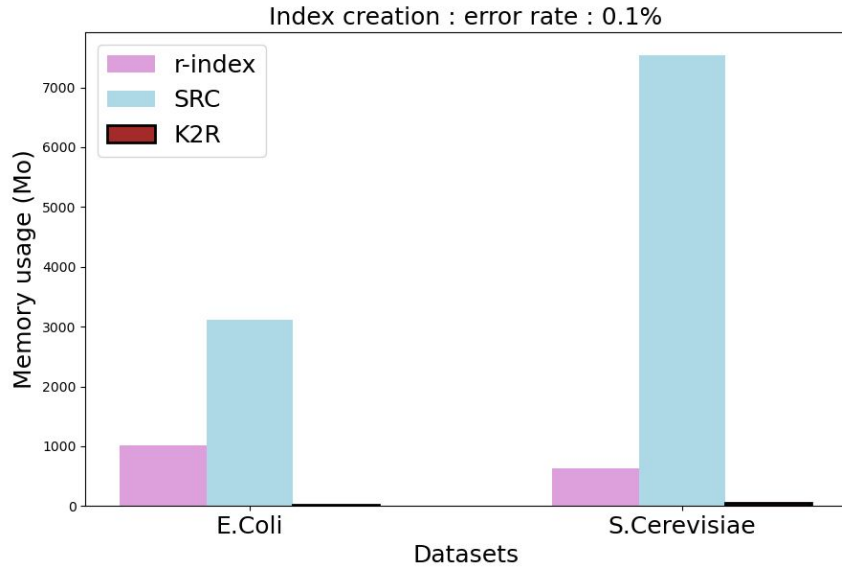
Parameters : error rate 1%, read length = 10.000bp

Dataset

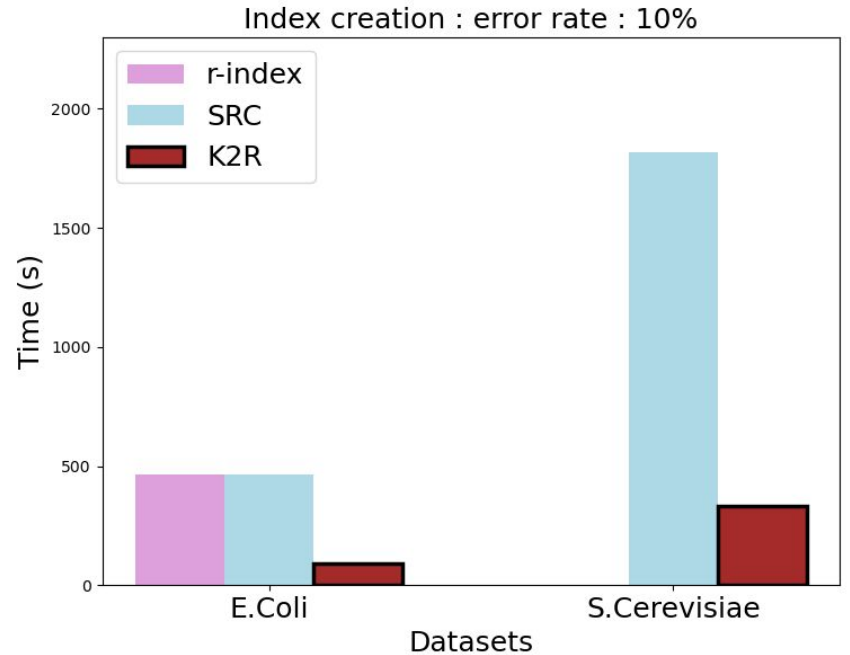
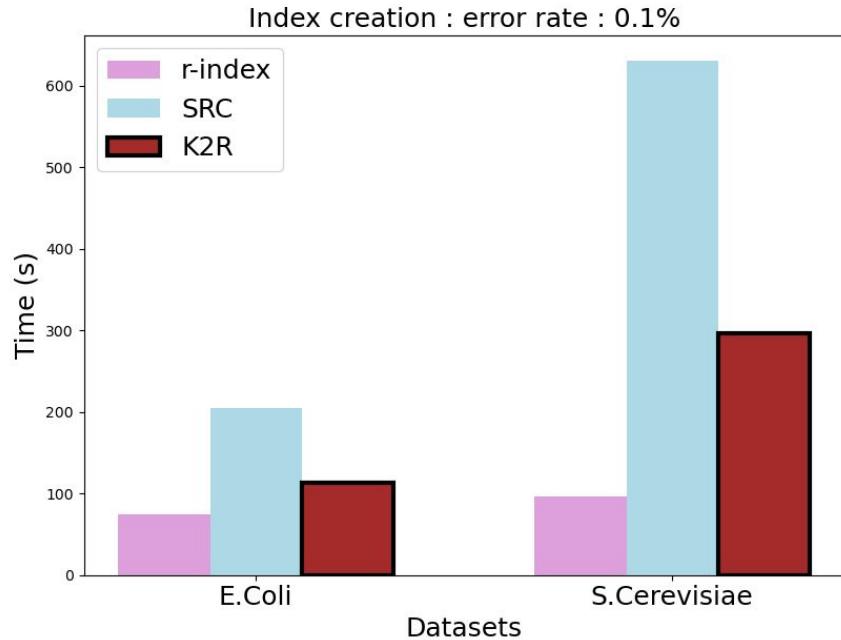
C.Elegans : 100X



Results : comparison



Results : comparison



Index files on disk

Index creation :

- C.Elegans (100.3MB) : error rate 1%, read length = 10.000bp.

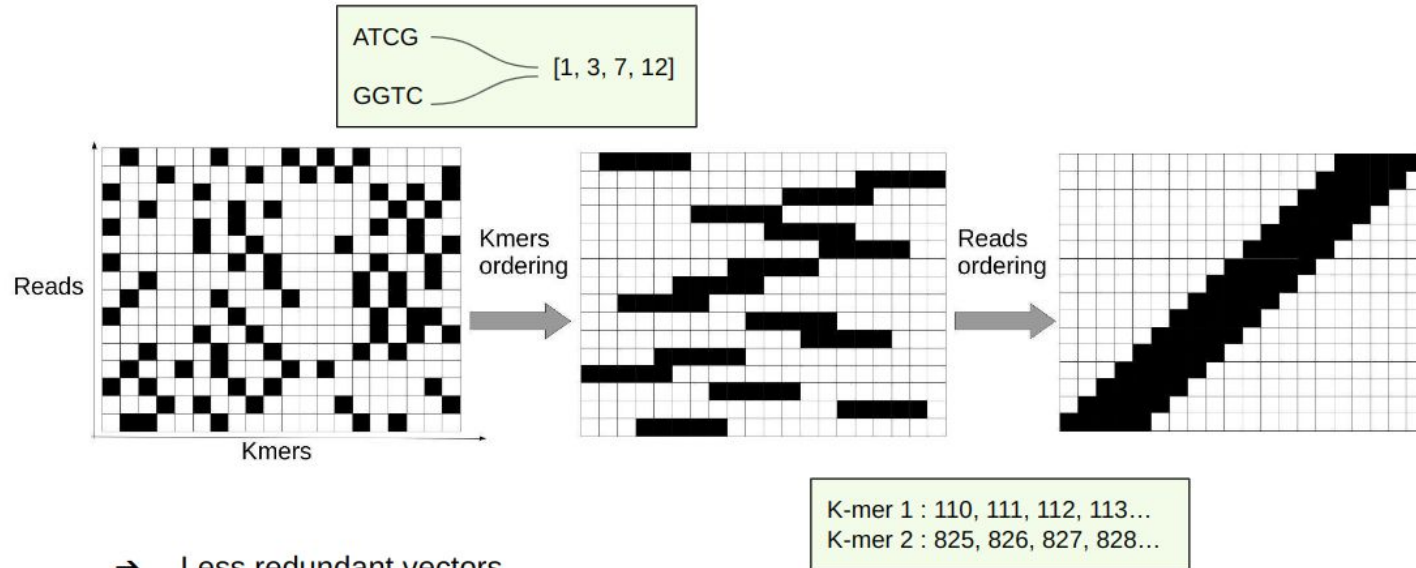
Coverage	Read FASTA file size (Go)	M-mer binary file size	Color binary file size	Total	Total / input
10	0.958	0.057	0.168	0.225	0.23
50	4.7	0.137	0.901	1.038	0.22
100	9.4	0.218	1.8	2.018	0.21
150	15	0.278	2.6	2.878	0.19
200	19	0.324	3.4	3.724	0.19

Conclusion

	Construction	Memory	Debit
Read Connector	✓	✗	✓
r-index	✗	✓	~
K2r	✓	~	✓

Perspectives

Sorting reads



- Less redundant vectors
- Improve delta encoding

Perspectives

- Bottleneck : memory (challenge r-index ?)
- Soon : article
- **Applications : open to collaborations :)**

Take home messages

Goals

- Associate k-mers to reads, in order to study genome sequences from raw reads

Etat de l'art

- Hashing methods (memory expensive)
- Full-text indexing (hard to construct)

Take home messages

Solutions

- Minimizers (False positives are managed)
- Colors
- Compression

K2R

- Fast
- Dynamic
- Scalable