



Compression of Human Genomic Data

Dominique Lavenier



GenScale, Univ. Rennes, IRISA/CNRS, INRIA - France



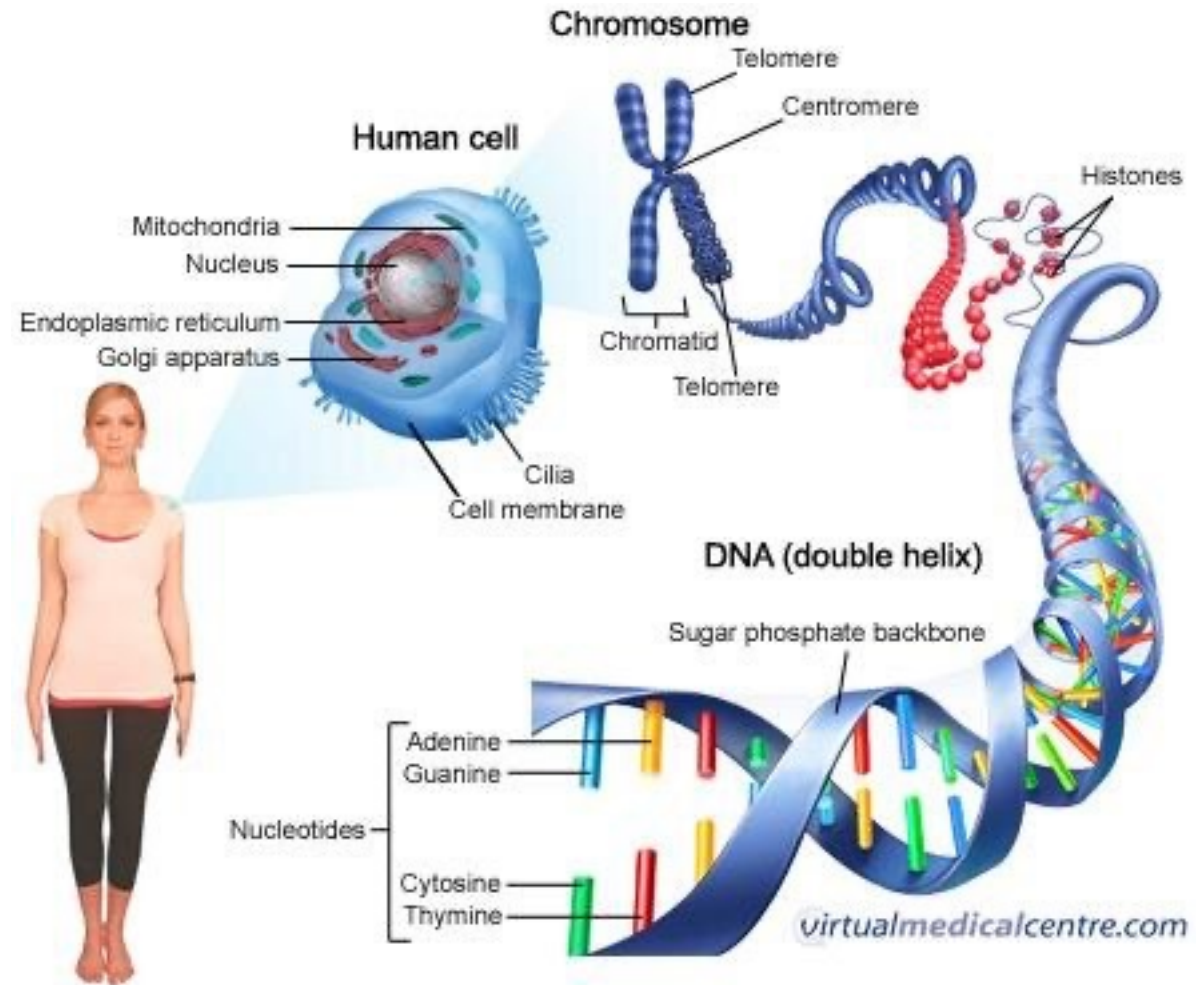
Human Genome

Human Genome

- 23 pairs of chromosomes / cell
- 3.2×10^9 bp

Each genome is different

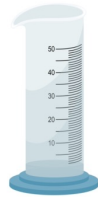
- Many small variations
 - SNPs (Single Nucleotide Polymorphism)
 - Micro-satellites
 - ...
- A few Large variations



Sequencing



DNA
extraction →



Illumina sequencer

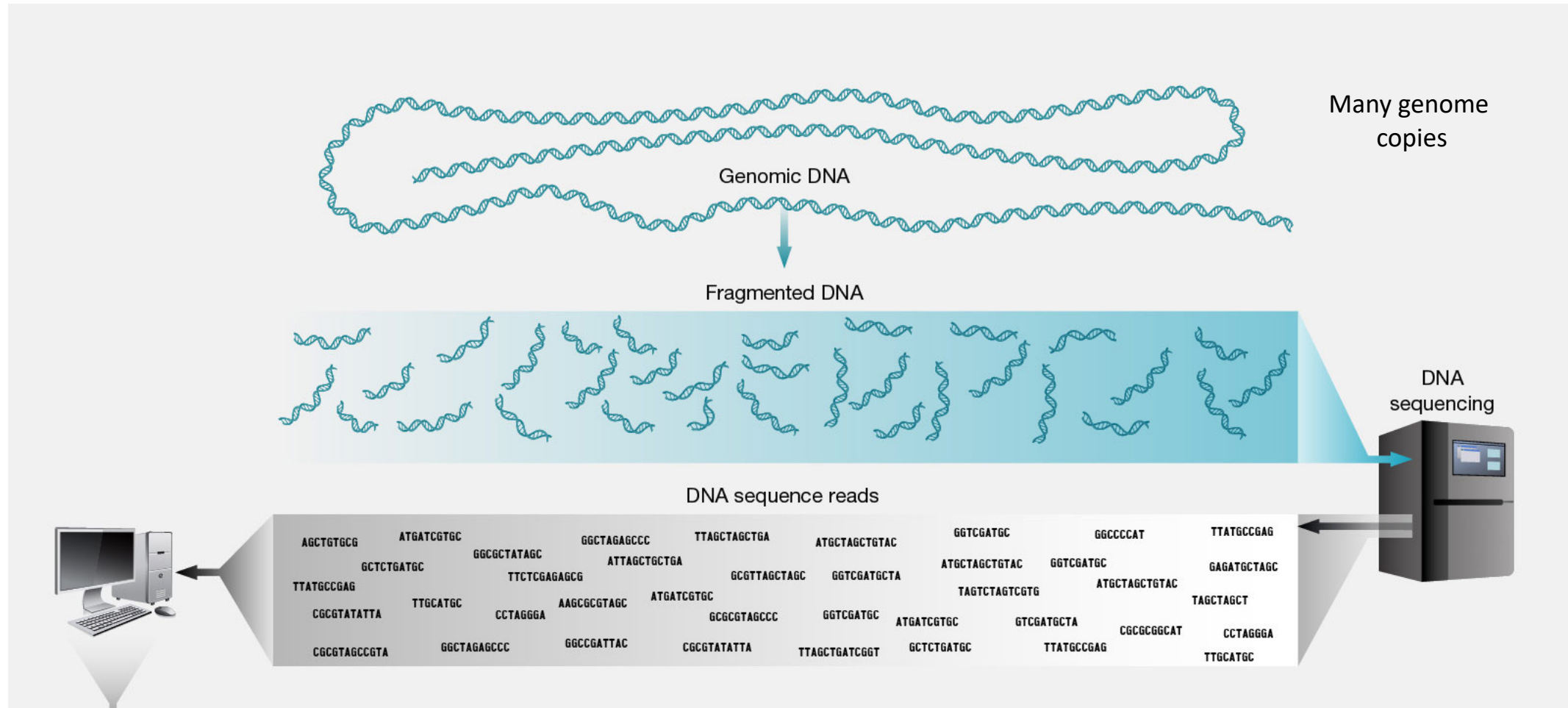


Raw data



300 – 500 GBytes

Sequencing principle



Genomic Data

Short reads

- Length: ~300 bp
- Error rate: ~0.1%
- Technology: Illumina

Long reads

- Length: 10Kbp to 100Kbp
- Error rate: 1%-3%
- Technology: PacBio, ONT

FASTA / FASTQ format → text file

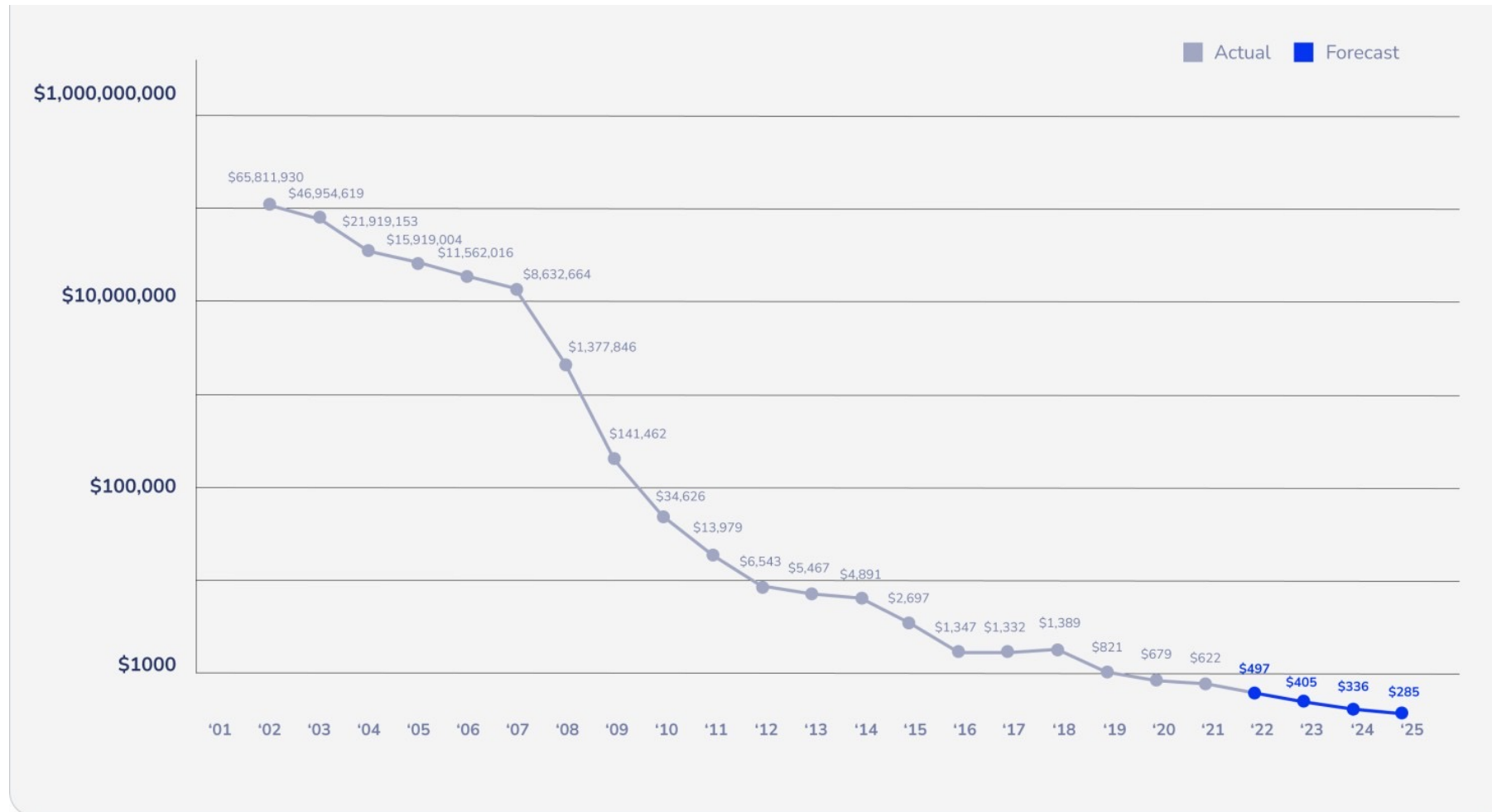
```
Identifier ● @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence ● TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign ● +
Quality scores ● hhhhhhhhhghghghhhhhfhhhhfffffe'ee['X]b[d[ed'[Y[~Y
Identifier ● @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence ● GATTTGTATGAAAGTATACAACATAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign ● +
Quality scores ● hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd
```

FASTQ

Human Genome

- Coverage: 30X to 50X
- 10⁹ Reads: 2x150 bp
- 300 GB to 500 GB

Human genome sequencing cost



<https://www.linkedin.com/pulse/cost-sequencing-human-genome-fall-over-50-2025-3billionpairs/>

Genomic data compression

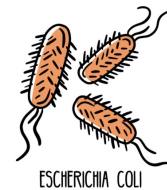
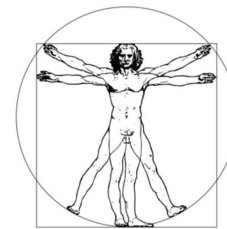
Lossless

Without reference

- Standard tools (gzip,bzip2,...)
- Genomic dedicated tools

With reference

- Genozip
 - currently the best tool (fast, good ratio)
 - used as reference



Compression with reference

Human Genome Reference: 3.2×10^9 nt

TTGAGACCAGATTAGAGACCAGGGGAGAGACCCAGGGAGTTA**CATACCATGACCAGGGGATT**TAGGGATTCCGATATAGGACGCCATTACCAGTGACCAGGATAGCCAGGTTAGA...

CATACCATGACCAGGGGATTTAGGGAA**A**T
GGTAGGACCAGGGATTGGGATTAGGAGA
TGGACCAGGATAGGACCAGATAACAGTA
GTCCACATACCATAGGACCACTTACCAT
...



42 (S, 27, A)
10988
1970 (D, 13)
7892
...

Short reads

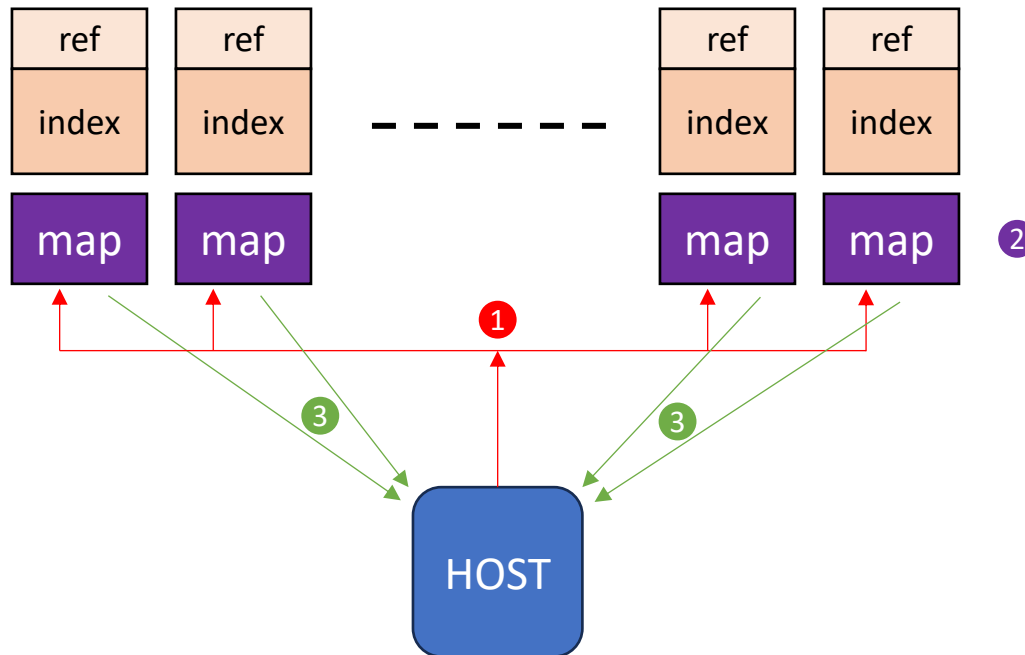


Mapping problem: find the best match as fast as possible

UPMEM PiM implementation

MRAM
Reference genome
+ index

DPU



The reference is distributed over the DPUs

- Each DPU as a fraction of the Human Genome in its MRAM
- An index is built to speed-up the mapping

Broadcast reads to DPUs

1

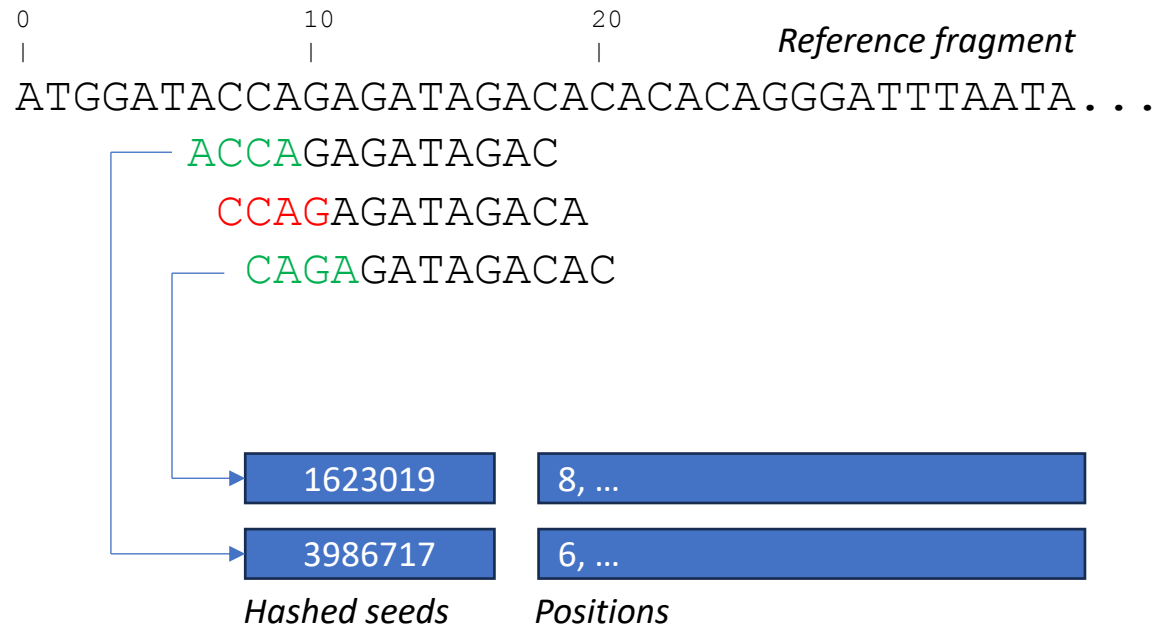
Mapping on DPUs

2

DPUs sent back match (if any)

3

Reference index



Each DPU indexes its reference fragment by iterating 13-mers (called seeds) and inserting their positions in a table

Low-interest seeds are filtered out

Mapping

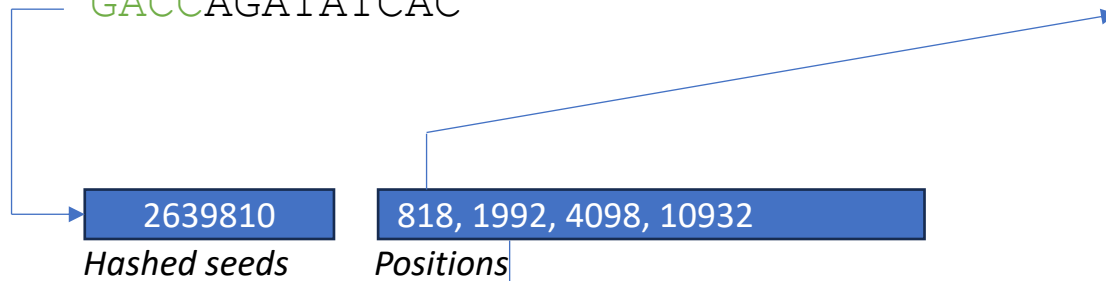
Read

GGGACCAGATATCACACGATATTAAGGACCAGGTA

GGGACCAGATATC

GGACCAGATATCA

GACCAGATATCAC



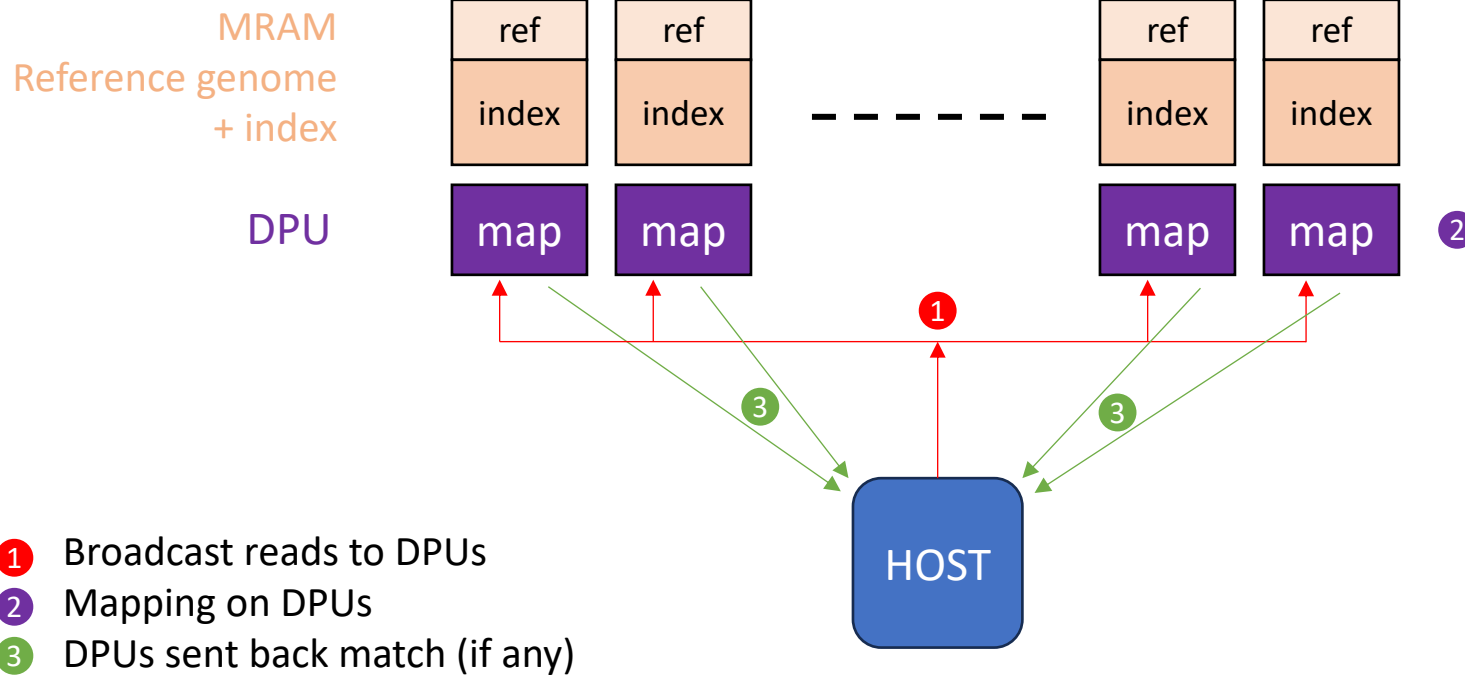
Size seed = 13

818
|
...TAG**GACCAGATATC**GTCCAGG**ACTG**ACCCTTTAGATCTC...
GG**GACCAGATATC**CACACGAT**ATTA**AGGACCAG**G**TA
reference

1992
|
...GG**GACCAGATATC**CACACGATATCA**A**TGACCAGGTAGTT...
GG**GACCAGATATC**CACACGATATTA**A**GGACCAG**G**TA
reference

 Mapping: position 1990

Implementation limitation



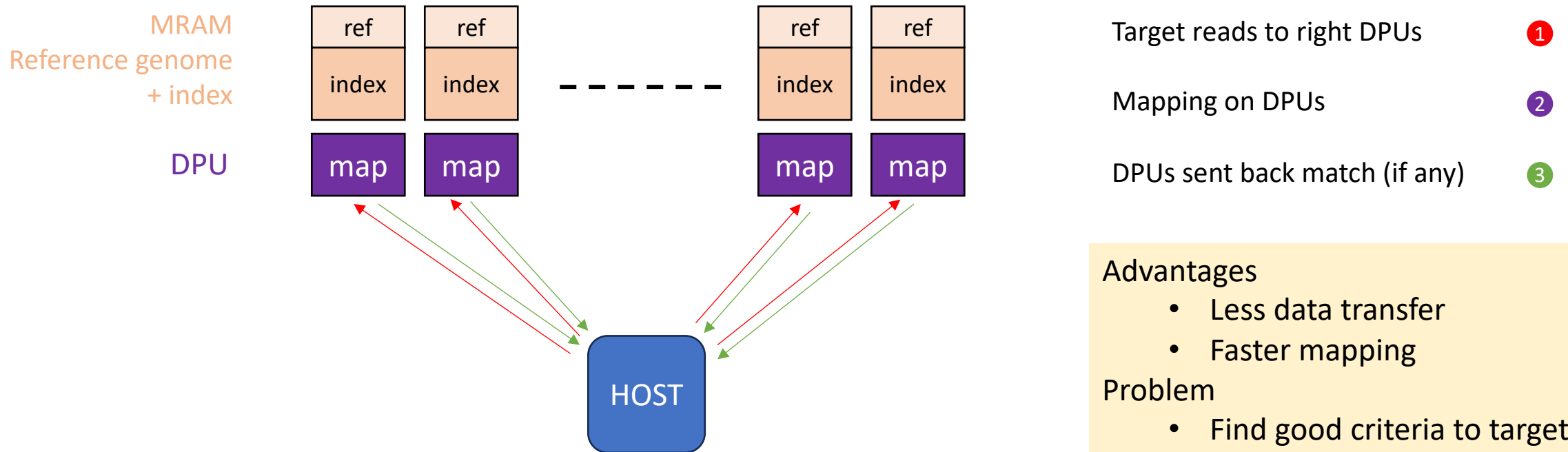
All reads send to all DPUs

- High data transfer
- High DPU workload

Only a few DPUs send back mapping positions

Read dispatcher

Instead of broadcasting all reads to all DPU, send reads selectively to some DPUs



- Advantages**
- Less data transfer
 - Faster mapping
- Problem**
- Find good criteria to target the right DPUs

Bloom filter

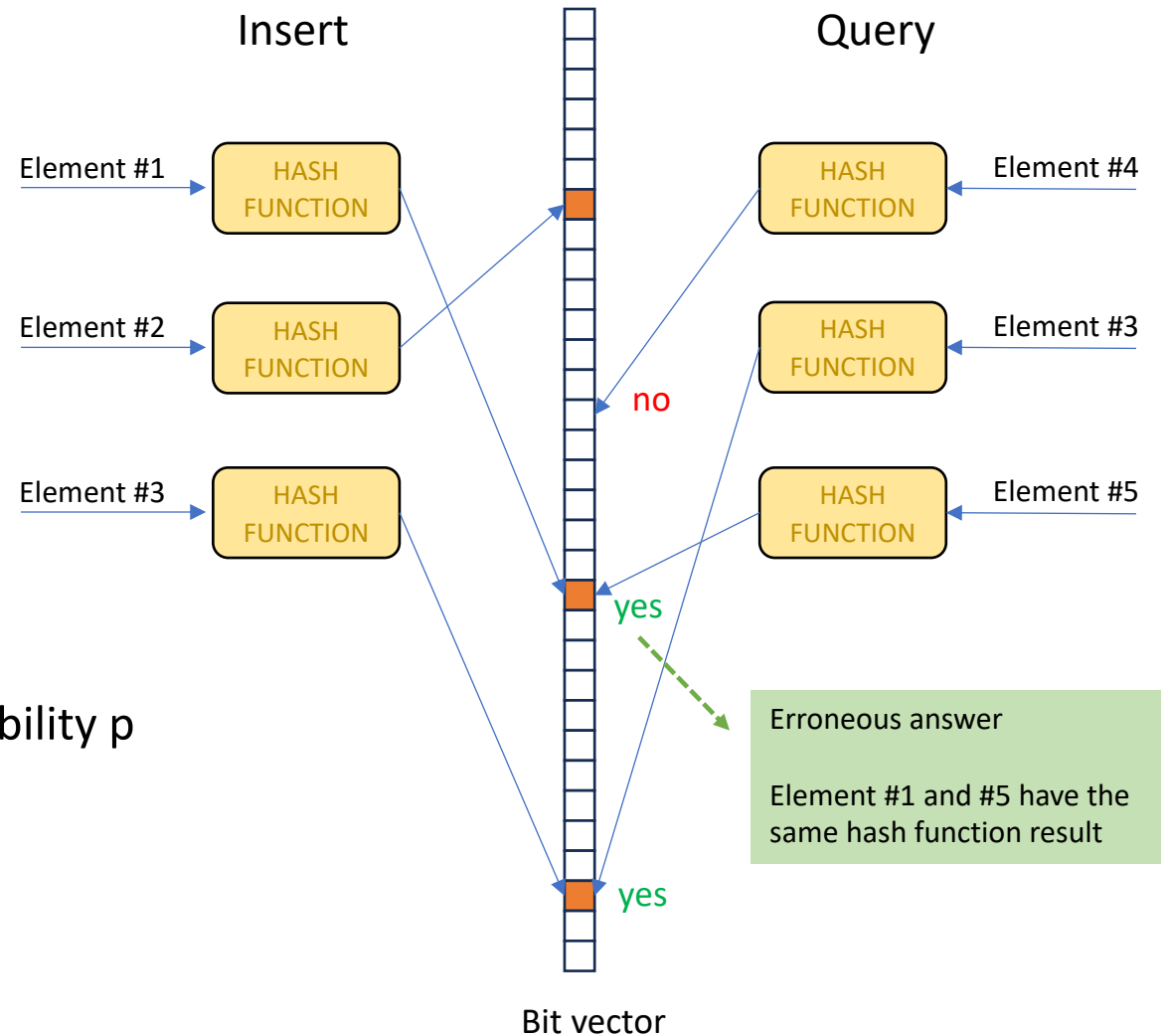
Probabilistic data structure

Two operations

- Insert an element
- Query the presence of an element
 - no → the element is absent
 - yes → the element is present with a probability p

Compact data structure

- Vector of N bits
- Hash function returns a value in $[0, N-1]$



Bloom filter associated to a DPU

reference

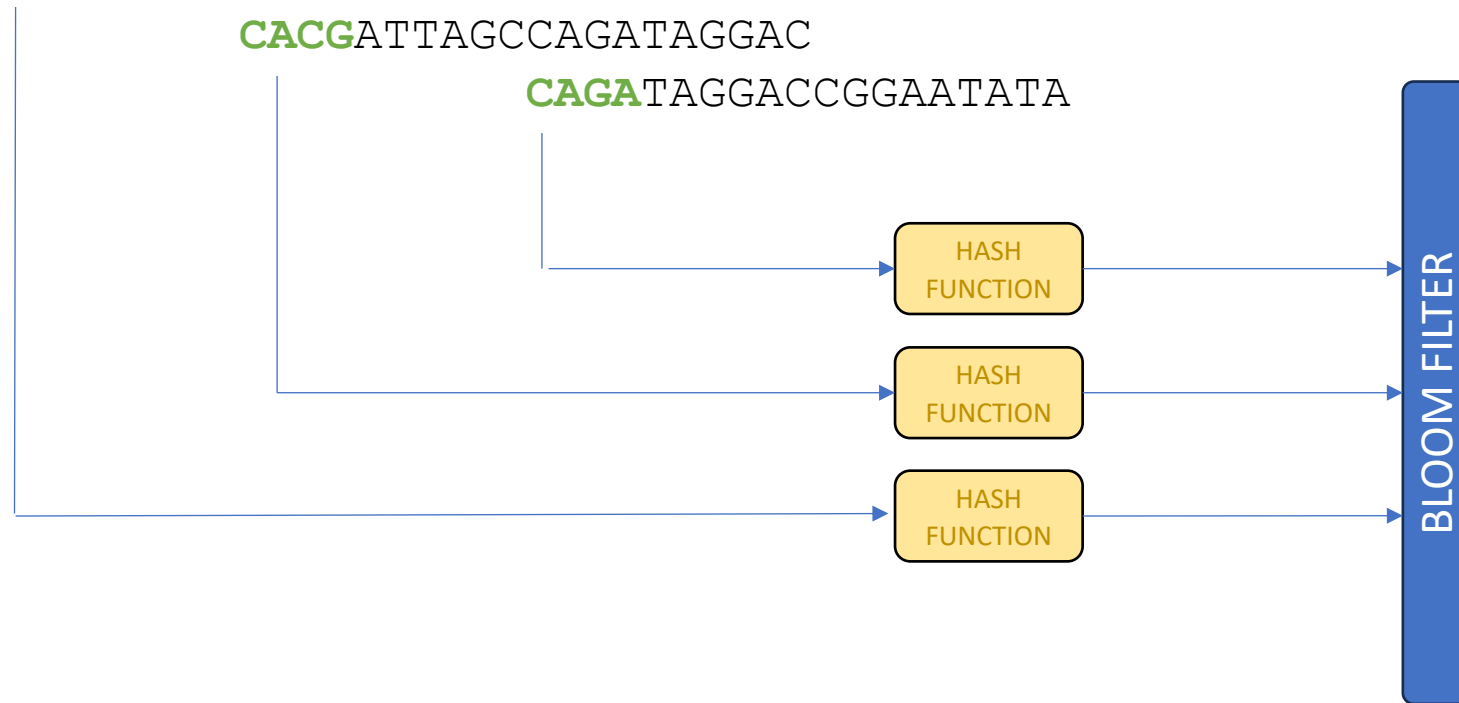
TTGACACCAGGATAGACACGATTAGCCAGATAGGACCGGAATATATCCAGATTAGCCGAGATGGATA

CAGGATAGACACGATTAGCC

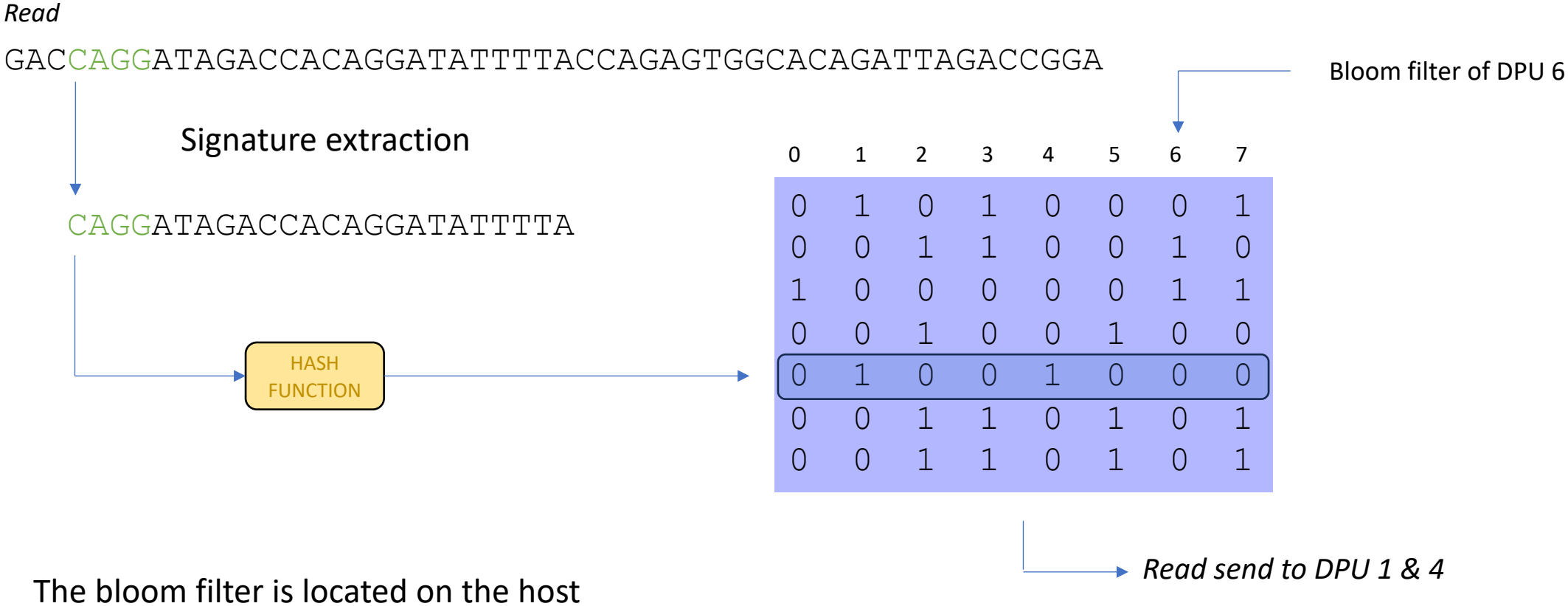
CACGATTAGCCAGATAGGAC

CAGATAGGACCGGAATATA

Set of signatures
Substrings of 70 nt



Bloom filter based dispatcher



Experiment

- Server:
 - Intel Xeon Silver 4215 @ 2.5 GHz
 - 256 GB of DDR4, 20 UPMEM DIMMS @ 350 MHz (160 GB, 2560 DPUs)
- Data sets

	Plain file size (GB)	Read count	Read size	Year
Genome GRCh38	3.2	1	3.1 Gbp	2013
Reads H1 (SRR14724532)*				
FASTQ	2x 121.1	356 M	151 bp	2021
FASTA	2x 66.3			
Reads H2 (SRR14724533)†				
FASTQ	2x 122.9	361 M	151 bp	2021
FASTA	2x 67.3			

 reference



Sequencing data



* <https://www.ebi.ac.uk/ena/browser/view/SRR14724532>

† <https://www.ebi.ac.uk/ena/browser/view/SRR14724533>

Bloom filter dispatcher efficiency

8 ranks (512 DPUs)

- Reference: 6.25×10^6 nt / DPU (MRAM full)

Size of the Bloom filter: ~ 10 GBytes

On average, each read is send to ~ 2 DPUs → High efficiency

Host bottleneck



- DPUs are idle $\sim 40\%$ of the time

Compression ratio

Software	% of original size		
	FASTQ	FASTA	
UPMEM PIM	5.52	5.05	
Genozip -e	4.08	18.46	
Genozip -e fast	5.27	18.49	<i>fast mode</i>
Genozip -e @16	4.09	18.56	<i>16 threads</i>
Genozip -e fast @16	5.32	18.61	

Compression time

Current implementation uses 8 ranks (512 DPUs)

Software	Compression time (s)	
	Elapsed	CPU
UPMEM PiM	 188	 2251
genozip -e	310	9546
genozip -e -fast	227	7000
genozip -e -@16	510	8151
genozip -e -@16 -fast	362	5680

Energy consumption

Software	Energy consumption (kJoules)			
	DRAM*	PKG*	DPU [†]	Total
UPMEM PiM	9.12	20.37	7.16	🏆 36.65
genozip -e	15.38	41.66	N/A	57.04
genozip -e -fast	11.52	31.89	N/A	43.41
genozip -e -@16	25.26	61.70	N/A	86.96
genozip -e -@16 -fast	18.52	44.60	N/A	63.12

* Measured with Intel® RAPL

† Estimated with execution time

Conclusion

Sequencing the human genome is becoming inexpensive

- Use in personalized medicine
- Medical data that needs to be kept for several decades
- Efficient storage needed (space / time / energy)

PiM implementation of genomic data compression

- relieves server load
- reduces power consumption

Many other genomic algorithms benefit from PiM accelerators

- BioPiM project: <https://biopim.eu/>
- GenoPiM project: <https://genopim.irisa.fr/>

Thank you for your attention

Dominique Lavenier

lavenier@irisa.fr

GenScale, Univ. Rennes, IRISA/CNRS, INRIA - France

