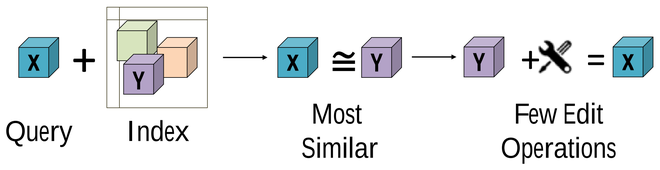
On the Challenges of Deduplicating Human Genomic Sequencing Data

Vinicius Cogo
Alysson Bessani

Overview

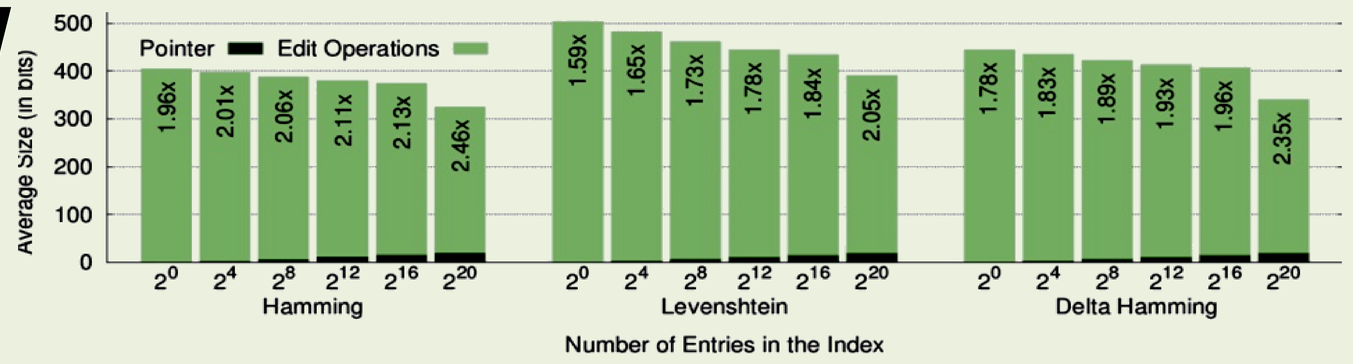
Reduce human genomes in the **FASTQ** format

Similarity-based deduplication + Delta encoding

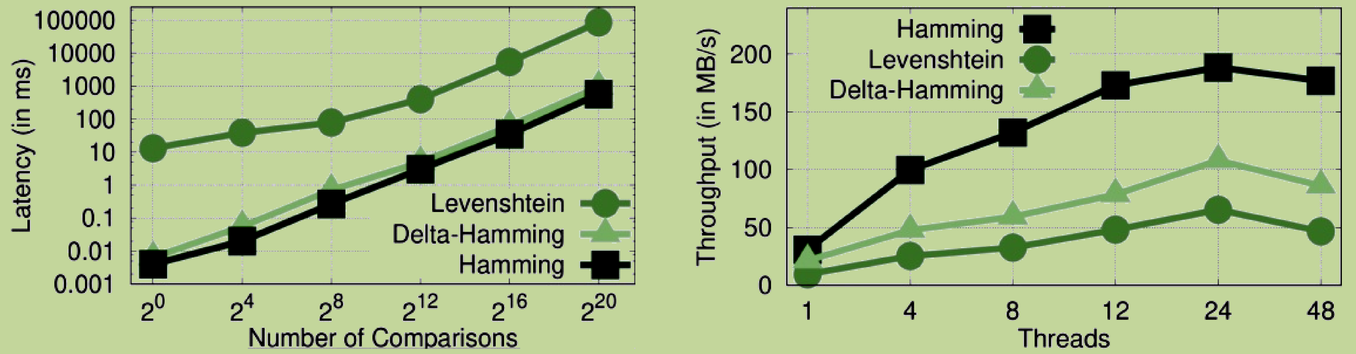


Results

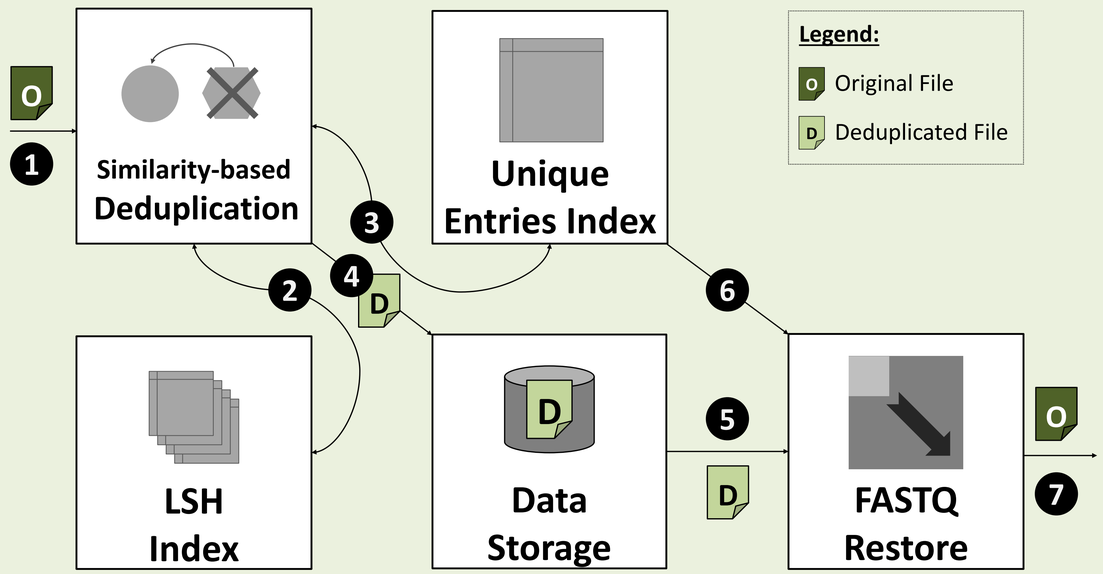
Compression ratio



Throughput



Architecture



Comparison

- **Compression ratio:**

≈75%
of results from the best

2.46x vs. **3.14x** ZPAQ
(QS-only)

6.10x vs. **8.20x** LFQC
(FASTQ)

- **Decompression throughput:**

200 MB/s

4x faster than GZIP

80x faster than ZPAQ&LFQC

U LISBOA | Ciências ULisboa | LASIGE | FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
(Project: UID/CEC/00408/2013)