Exact- and Near-Deduplication in the Compression of Whole Human Genomes

Vinicius Cogo
Alysson Bessani

Main Goal

Combine **compression**, **deduplication**, and **similarity search** to **reduce** the **size** of **human genomes** in the **FASTQ** format

Ideas to Explore

FASTQ entry

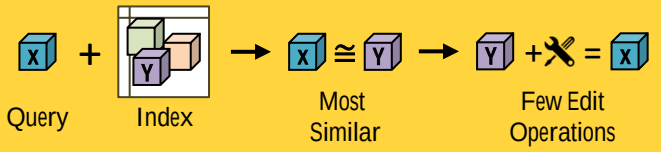
```
@SRR027520.280 length=25
GGTAGATAGGGTAAAGAAAATGTGG
+SRR027520.280 length=25
BBBBB=?:=DDAB@CBB>CABBBBB
```

Convert **quality scores** to **delta values**

```
BBBBB=?:=DDAB@CBB>CABBBBB

0,0,0,0,-5,2,-5,3,7,0,-3,1,…
-2,3,-1,0,-4,5,-2,1,0,0,0,0
```

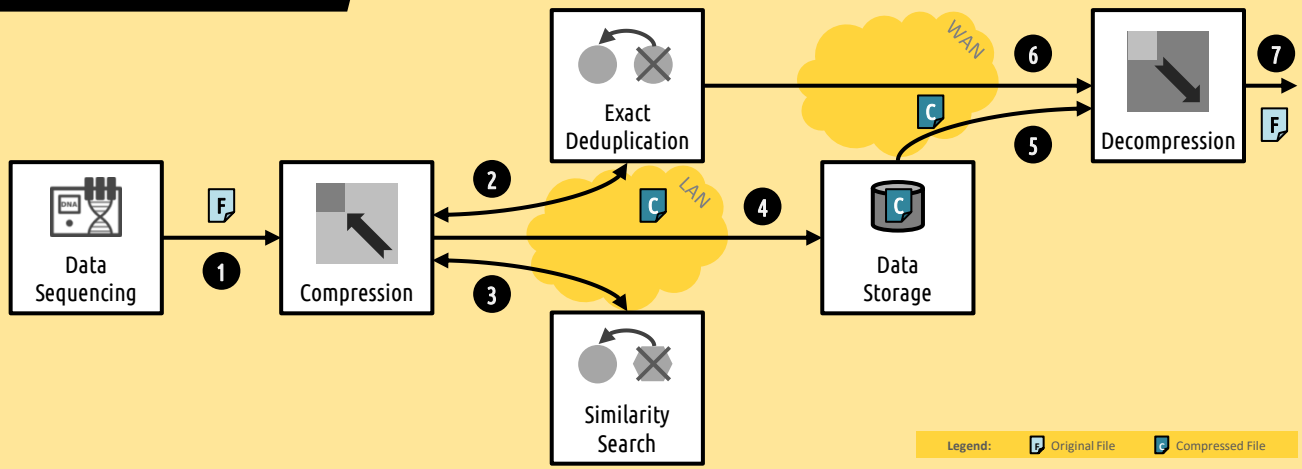
Similarity search using **LSH** (Locality Sensitive Hashing)



Explore **patterns** in **delta values**

Distribution of delta values	Sum of delta values	First character	What else?
35% $\Delta q = 0$	**30%** $\Sigma\Delta q = 0$	**50%** $1^{st}\ q = H$	
94% $\Delta q \in [-10,10]$	**98%** $\Sigma\Delta q \in [-20,20]$	**87%** $1^{st}\ q \in [A, H]$	

Overview



Expected Results

- **Feasible** index **sizes**
- **Adapt** to different systems
- **Reduce** data at least **5x**
- **Reduce more** in larger systems