Efficient Storage of Whole Human Genomes

Vinicius Vielmo Cogo[*] and Alysson Bessani
LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
{vvcogo,anbessani}@ciencias.ulisboa.pt

Whole genome sequencing (WGS) digitises the complete DNA from an organism's cell. The standard raw data format from WGS is the FASTQ, which is composed of unique unrelated blocks with four lines each (i.e., comments, DNA sequences, and quality scores) [1]. Whole human genomes are big (up to 300GB per individual) and critical (contain identity- and health-related information).

The number of to-be-stored genomes is increasing exponentially, which is motivated by the decreasing cost for sequencing a human genome—$1000 nowadays [1]. Pressure is on hospitals and biobanks to store millions of genomes, and researchers would like to analyse thousands samples at time. Storing genomes efficiently may accelerate medical breakthroughs, but augments the risks for donors' privacy. The million-scale size and criticality of sets of genomes require systematic solutions to store and share this data in efficient, scalable, and secure ways.

Privacy-awareness. Some portions of human genomes are more important to our privacy than others. Based on published attacks, we analysed these privacy-sensitive portions and proposed a method that systematically detects them when they are sequenced [2]. Segregating this portion allows systems to explore adaptive access control based on donors' preferences for each portion, and support diverse dependability, security, and privacy premises for them. The detection method and subsequent steps must scale to large data sets and must support with parallelism the high throughput of WGS.

Cost-efficiency. Data reduction techniques decrease the data size and increase data density—resulting in higher cost-efficiency. However, the specificities of genomes and their formats render standard compression [3] and deduplication algorithms [4] ineffective. Specialised compression algorithms for FASTQ files partially solve the problem with compression ratios up to 5× [3]. Based on the values from Table 1, reducing data 5× still incur in several million dollars for a million genomes annually. Archiving it on a tape-based solution would also cost dozen million dollars, even if it was optimised to fit in rackscale [5]. These values consider a private infrastructure using hard disks [6] or public clouds—Amazon S3 provides standard storage and Amazon Glacier archival.

Table 1: Annual storage cost per genome (in $).

Reduction	**Gains**	**Disk**	**S3**	**Glacier**
0× Uncompressed	0%	195	99	25.2
5×	80%	39	19.8	5.04
200×	99.5%	0.975	0.495	0.126

Considering human genomes have more than 99.5% of similarity to each other, why does not compression achieve higher data reductions (e.g., 200×)? We observed that all compression algorithms for FASTQ consider only the to-be-compressed file or a single reference genome. There is a research opportunity in using the global knowledge on human genomes and inter-genome deduplication to increase the reduction gains towards annual storage costs near $1 per genome. We are exploring several possibilities: (1) using statistics about common DNA sequences, (2) grouping similar sequences (e.g., with Locality-Sensitive Hashing), (3) detecting and storing their differences, and (4) developing application-specific deduplication algorithms. Besides systems' performance and scalability, we are investigating also how application-specific knowledge can be used to adapt storage systems to become more efficient.

In conclusion, our contributions on privacy-awareness and cost-efficiency enable the efficient storage of genomic data and make security and dependability more affordable in this scenario.

References

[1] V. Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

[2] V. Cogo et al.A high-throughput method to detect privacy-sensitive human genomic data. In *14^{th} WPES*, 2015.

[3] S. Deorowicz and S. Grabowski. Data compression for sequencing data. *Algorithm Mol. Biol.*, 8(1):1, 2013.

[4] J. Paulo et al. A survey and classification of storage deduplication systems. *ACM Comput. Surv.*, 47(1):11, 2014.

[5] S. Balakrishnan et al. Pelican: A building block for exascale cold data storage. In *11^{th} OSDI*, 2014.

[6] D. Haussler et al. A million cancer genome warehouse. Technical report, UCB/EECS, 2012.

[*] VVC is a PhD student and will present the poster.

[†] Work supported by FCT-LaSIGE (UID-CEC-00408-2013), and by EC-SUPERCLOUD (H2020-ICT-643964).

Efficient Storage of Whole Human Genomes

Vinicius Vielmo Cogo, Alysson Neves Bessani
{vvcogo,anbessani}@fc.ul.pt
LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

LaSIGE FCT-UID-CEC-00408-2013 | Ciências ULisboa | SUPERCLOUD EC-H2020-ICT-643964

Motivation

Identity- and health-related information

000,000 Data sets with millions genomes

GB 300GB per sample

The **criticality** and **million-scale size** of sets of human genomes require **systematic solutions** to store and share this data **efficiently**

State-of-the-Art

No automatic detection of privacy-sensitive sequences

High (**lossy**) compression ratio, **slow** (de)compression

No deduplication for genomic data

Objective

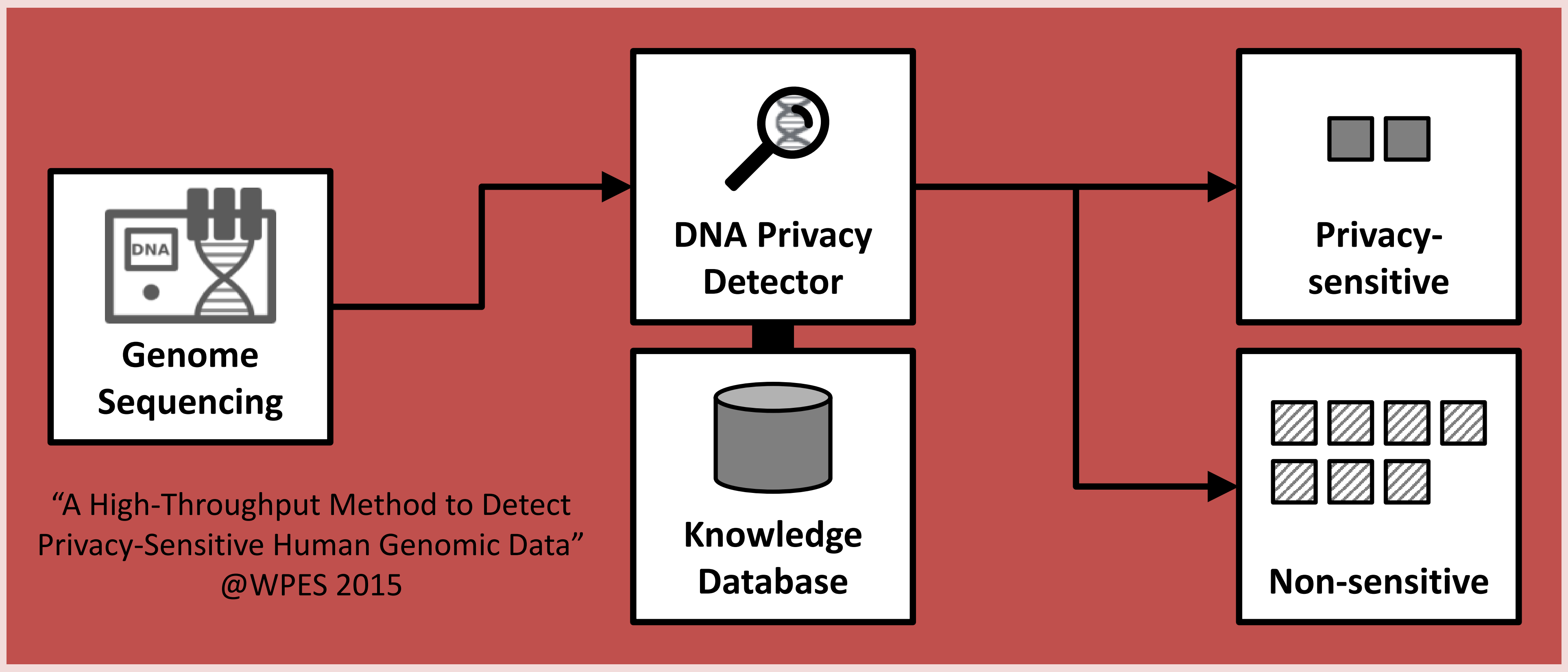
Enhance the **storage** ecosystem with **privacy-awareness** and **data reduction** to **enable** the **efficient storage** of **human genomes**

Privacy-Awareness

Some **data** are **more important** than others to our **privacy**

Goals:

- **Systematically detect** the **privacy-sensitive portions**
- **Create a database** of known privacy-sensitive sequences
- **Differentiate** them according to their **relevance to privacy**



"A High-Throughput Method to Detect Privacy-Sensitive Human Genomic Data" @WPES 2015

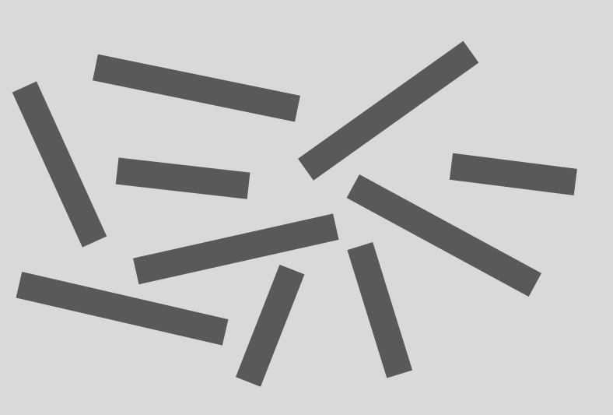
Challenges:

❑ High-throughput detection

❑ Completeness of the knowledge database

❑ Adapting subsequent components

Sequencing Genomes



Unaligned (FASTQ)

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
XR#\8)X)*B,eP]‘]_Ba_^_3_[Y?7+9IG:9IC
```

Aligned (SAM/BAM)

```
1:497:R:-272+13M17D 113 1 497 37 37M 15 100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>
XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0
```

Assembled (FASTA)

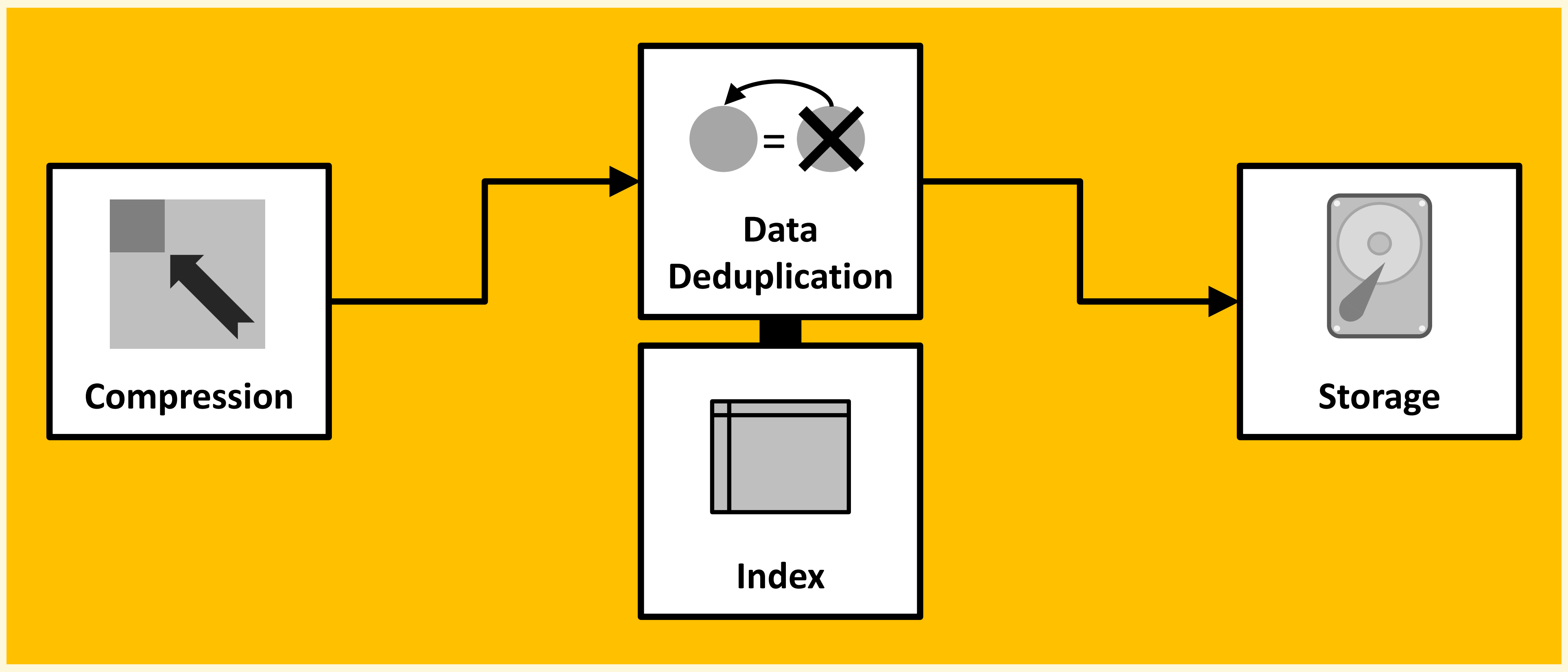
```
>SRR001666 chromosome 3
GCCGGTCCGCGCAGGCGCAGCGGGGTCGCAGGGGCGGGA
...
>SRR001666 chromosome 7
GATGGGTTCCTGGACCCTCCCCTCTCACCCTGGTCCCTCAGT
...
```

Data Reduction

Human genomes are more than **99.5%** similar to each other

Goals:

- Integrate **lossless compression** and **deduplication**
- Focus on **compression ratio** and **decompression time**
- **Deduplicate** similar portions



Challenges:

❑ High compression ratio in lossless algorithms

❑ Fast decompression with few resources

❑ Grouping similar sequences and storing their differences

❑ Developing application-specific deduplication

❑ Storing a human genome with less than $1 per year

Reduction	Gains	Hard Disk	Amazon S3	Amazon Glacier
0x Uncompressed	0%	195	99	25.2
5x	80%	39	19.8	5.04
200x	99.5%	0.975	0.495	0.126

Annual storage cost per genome (in $)

❑ Adapt storage systems with application-specific knowledge

❑ Adapt data management to differentiated portions