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 | SUPER CLOUD 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

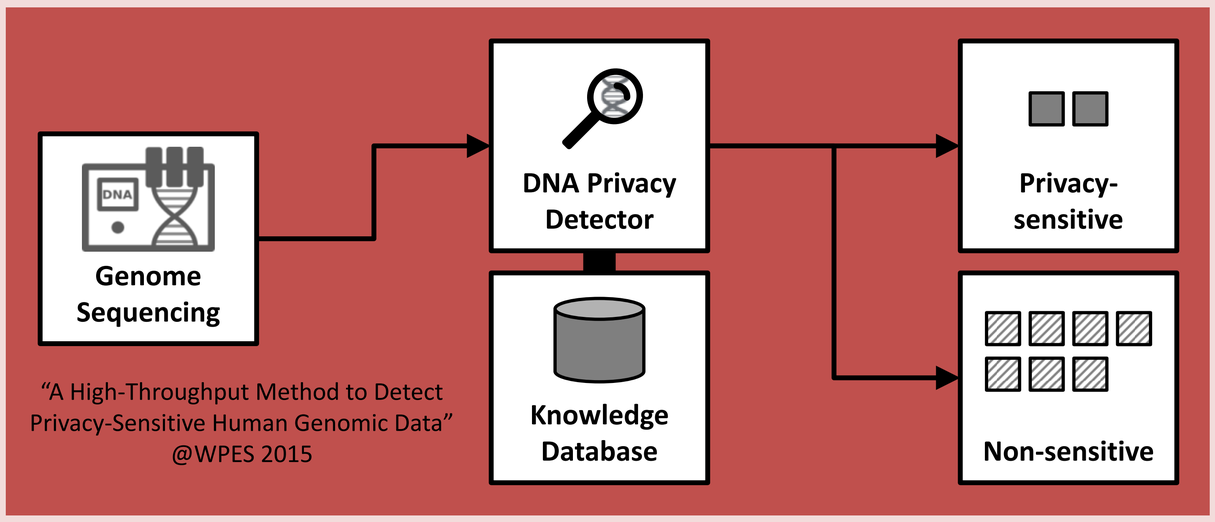
Enhance the **storage** ecosystem with **privacy-awareness** and **data reduction** to **enable** the **efficient storage** of **human genomes**

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**



Challenges:

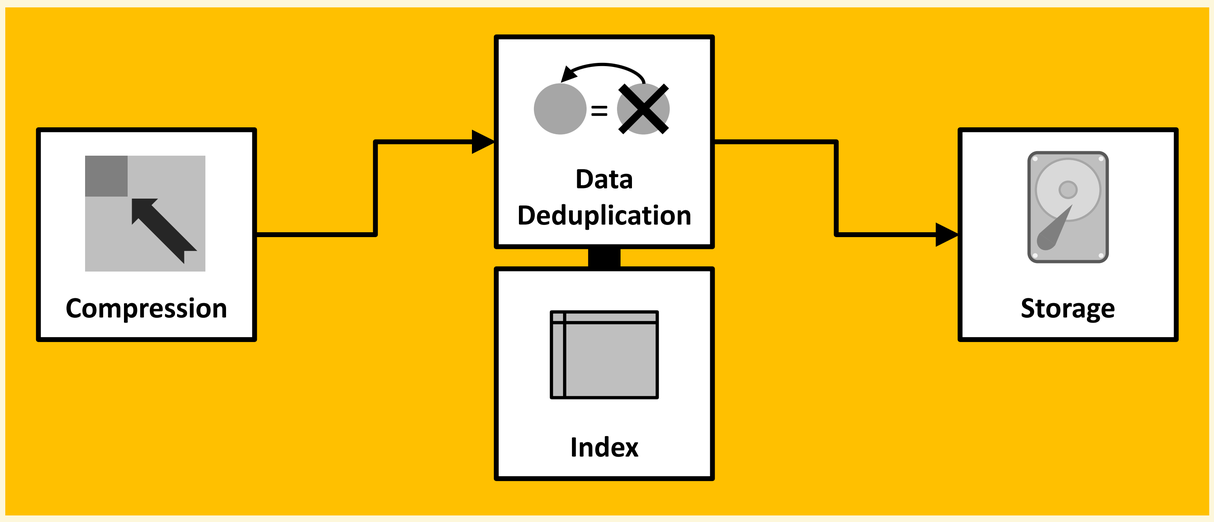
- ❑ High-throughput detection
- ❑ Completeness of the knowledge database
- ❑ Adapting subsequent components

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

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`
	`...`