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.

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.

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

<table>
<thead>
<tr>
<th>Reduction</th>
<th>Gains</th>
<th>Disk</th>
<th>S3</th>
<th>Glacier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncompressed</td>
<td>0%</td>
<td>195</td>
<td>99</td>
<td>25.2</td>
</tr>
<tr>
<td>5×</td>
<td>80%</td>
<td>39</td>
<td>19.8</td>
<td>5.04</td>
</tr>
<tr>
<td>200×</td>
<td>99.5%</td>
<td>0.975</td>
<td>0.495</td>
<td>0.126</td>
</tr>
</tbody>
</table>

References
Efficient Storage of Whole Human Genomes

Motivation

Identity- and health-related information

![Image](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAoAAAAAICAQABJRh3AAAABGdBTUEAALGPC/xhB/R0kABX就近.jpeg)

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

Sequencing Genomes

Unaligned (FASTQ)

Aligned (SAM/BAM)

Assembled (FASTA)

<table>
<thead>
<tr>
<th>Reduction</th>
<th>Gains</th>
<th>Hard Disk</th>
<th>Amazon S3</th>
<th>Amazon Glacier</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x Uncompressed</td>
<td>0%</td>
<td>195</td>
<td>99</td>
<td>25.2</td>
</tr>
<tr>
<td>5x</td>
<td>80%</td>
<td>39</td>
<td>19.8</td>
<td>5.04</td>
</tr>
<tr>
<td>200x</td>
<td>99.5%</td>
<td>0.975</td>
<td>0.495</td>
<td>0.126</td>
</tr>
</tbody>
</table>

Annual storage cost per genome (in $)

- Adapt storage systems with application-specific knowledge
- Adapt data management to differentiated portions