UNIVERSIDADE DE LISBOA • FACULDADE DE CIÊNCIAS • DEPARTAMENTO DE INFORMÁTICA

Efficient Storage of Whole Human Genomes

PhD. Student: Vinicius Vielmo Cogo

Advisor: Prof. Dr. Alysson Neves Bessani

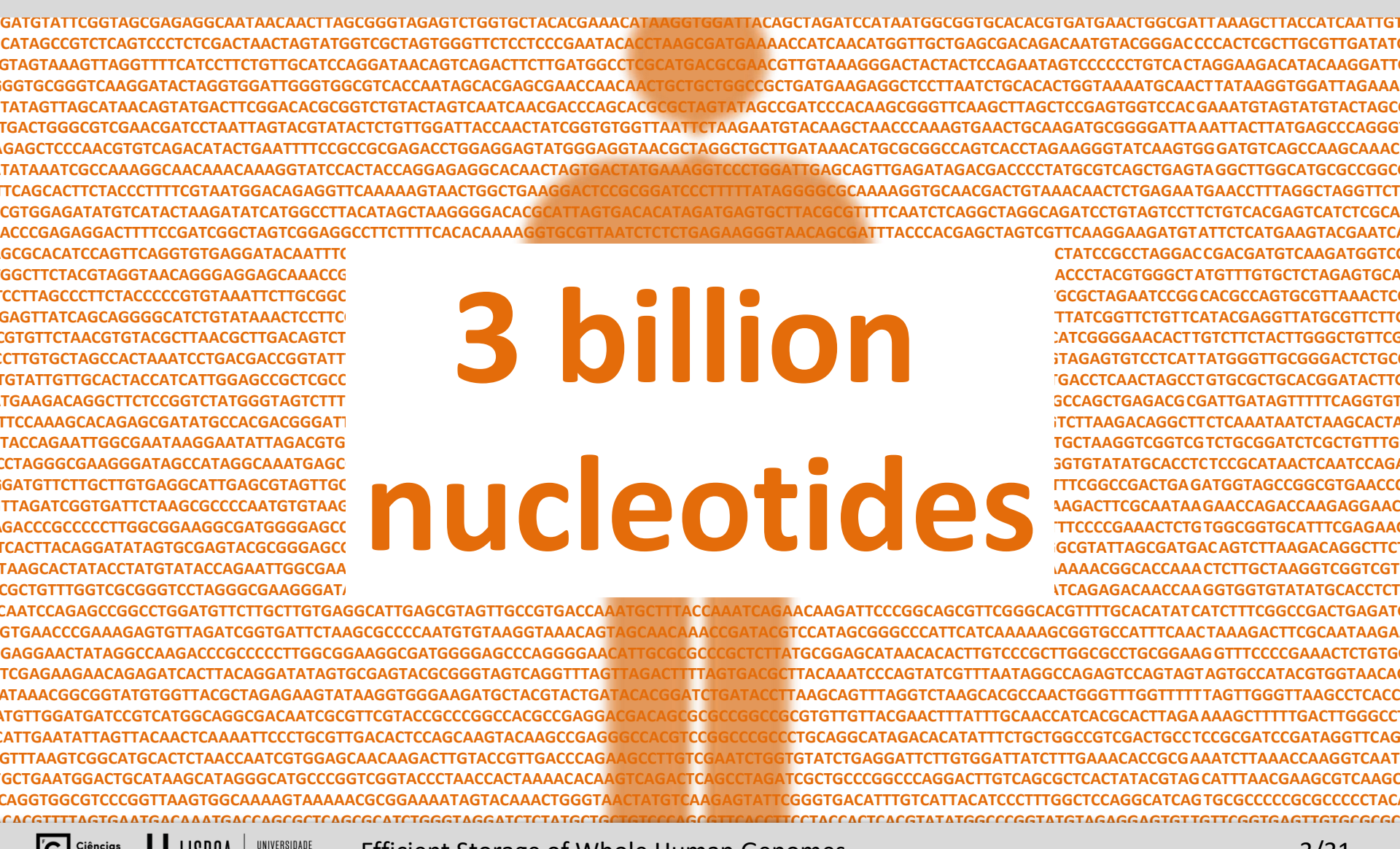
2016

1. Introduction

DNA

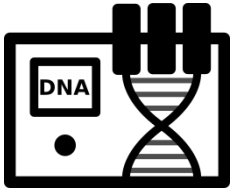
1. Introduction

Human Genome

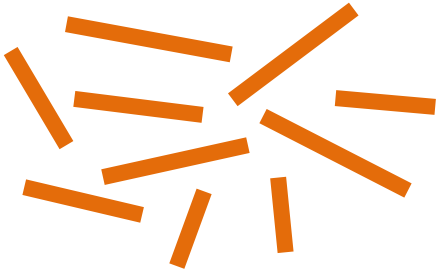


1. Introduction

Genomic data



Sequencing



FASTQ
(raw)
(unaligned)

Alignment

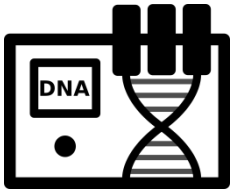
SAM
(aligned)

Assembly

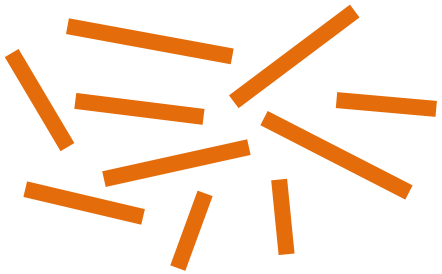
FASTA
(contiguous)

1. Introduction

Genomic data

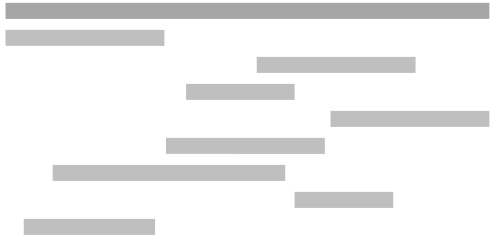


Sequencing



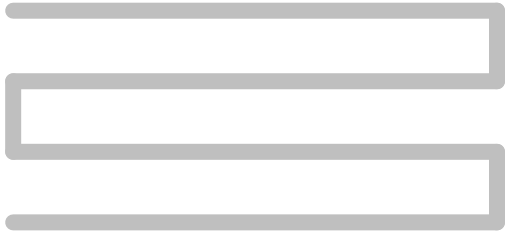
FASTQ (raw) (unaligned)

Alignment



SAM (aligned)

Assembly



FASTA (contiguous)

1. Introduction

Motivation

Human genomes are
critical and **difficult to store at scale**

Motivation

Critical:

- **Impact** several **areas**—medical, social, legal, etc.
- Large sample collections **accelerate breakthroughs**
- Hints on individuals' **health status and identity**

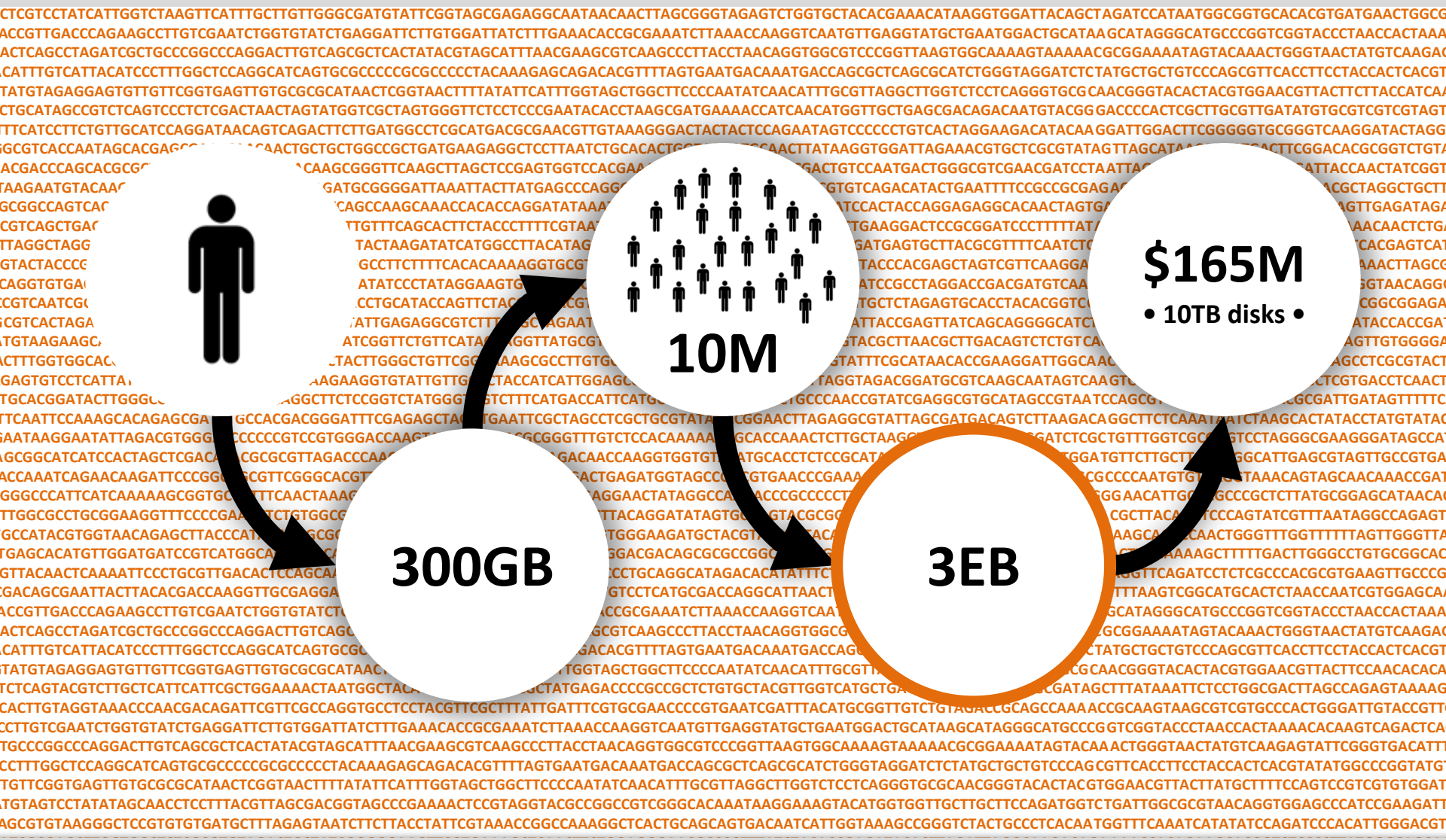
Motivation

Difficult to store at scale:

- Up to **300GB** per sample
- Sequenced **several times** during life
- **Whole populations** instead of a few individuals
- **Storing 1M** human **genomes**:
  $65M → 1st year, **$35M / year** → next years

1. Introduction

Problem size



Alternatives

Alternatives to the high costs:

- using cheaper infrastructures
- knowing what to protect
- reducing the data size

Outline

2. State of the art

3. Proposal
   - DETECTION
   - REDUCTION
   - STORAGE
   - AUDITABILITY

4. Plan

Outline

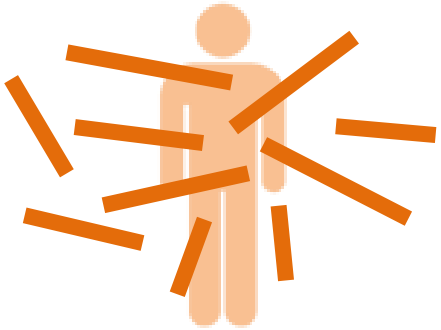
2. **State of the art**

3. Proposal
   - DETECTION
   - REDUCTION
   - STORAGE
   - AUDITABILITY

4. Plan

2. State of the art

The evolution of genomic data storage



	small	private public cohort	sequences	e.g., INSDC
		private public integrated	whole genomes	e.g., SRA

2. State of the art

The evolution of genomic data storage

- **Multi-tenant** public infrastructures (e.g., ELIXIR)
- **Cloud** storage (e.g., Illumina BaseSpace)
  - 👍 Infinite Scalability
  - 👍 Pay-as-you-go
  - 👎 Trust in a single entity
  - 👎 Locked in a single platform
- **Multi-cloud** solutions (e.g., DepSky)
  - We intend to explore alternatives
  - Tolerate organizational faults
  - Avoid lock-in

Identified open problems

Privacy-awareness:

- **Differential privacy** and **fully homomorphic encryption**
- **Detection** is a long-term ambition from the research and clinical communities
- **Manual classification** or orthogonal to published works

Data reduction:

- **Referential compression** for FASTA (reduces more than **700x**)
- Quality scores have higher entropy and no reference (most use lossy)
- Sample and read IDs make block deduplication inefficient
- No study on genomic specificities for deduplication

Storage auditability:

- Auditing the **integrity** of stored data in untrusted repositories
- Opportunity to audit **data access** in untrusted repositories

Outline

2. State of the art

3. **Proposal**
   - DETECTION
   - REDUCTION
   - STORAGE
   - AUDITABILITY

4. Plan

3. Proposal

Main goal

Design and develop **a set of tools** to advance the **efficiency** and **privacy-awareness** of storing human genomic data

3. Proposal

Thesis hypothesis

Privacy-aware and **data-reduction**
techniques are key elements
to **enable** the efficient affordable **storage**
of whole human **genomes**

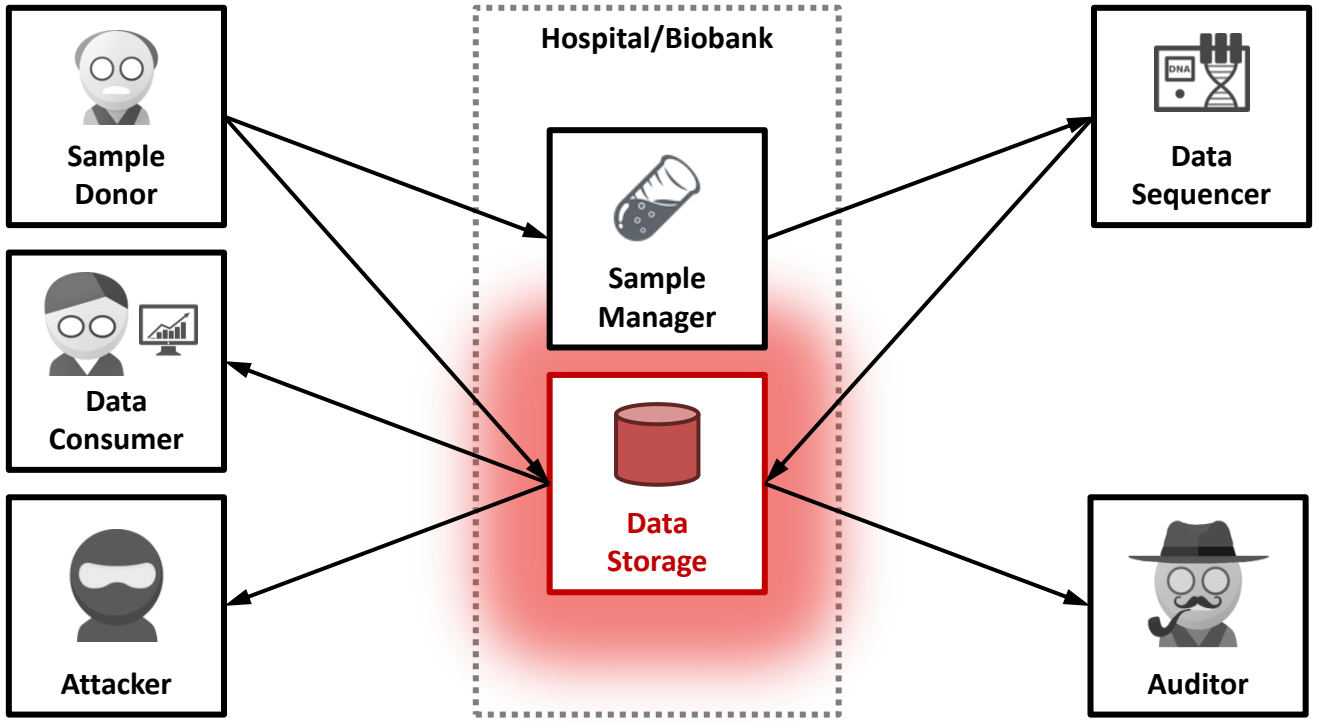
3. Proposal

Contributions

1. a study on the **privacy-sensitiveness** of human genomes
2. a method to **detect** the privacy-sensitive genomic data
3. **data reduction** algorithms specialized for genomic data
4. an efficient **storage system** that integrates several storage infrastructures
5. an independent **data auditability** scheme

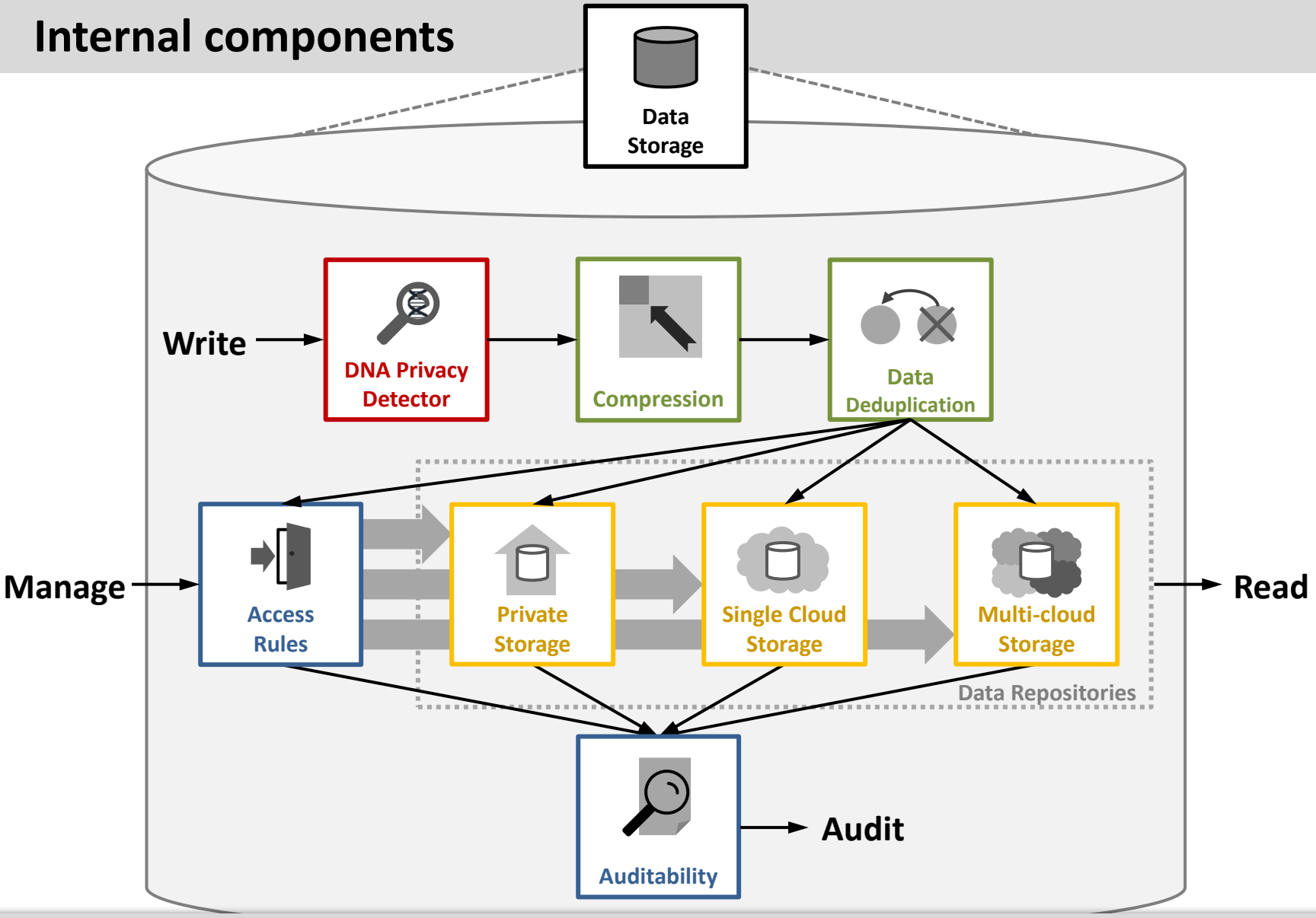
3. Proposal

Ecosystem



3. Proposal

Internal components



Outline

2. State of the art
3. Proposal
   - **DETECTION**
   - REDUCTION
   - STORAGE
   - AUDITABILITY
4. Plan

Overview

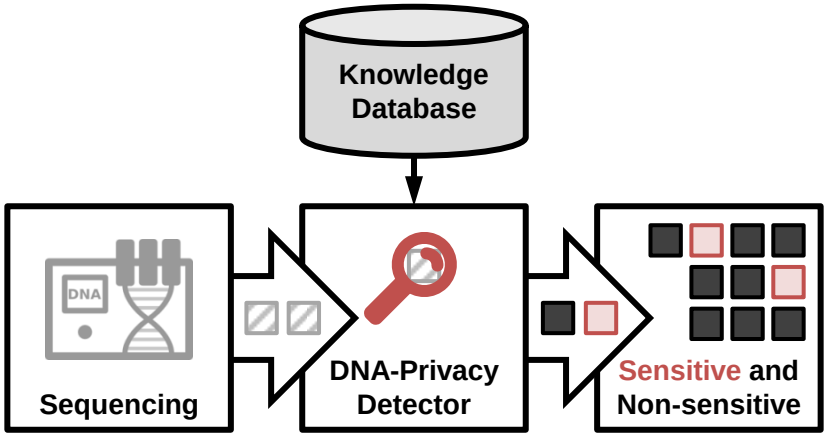
Detect and **differentiate** the **privacy-sensitive DNA** sequences of human genomes

Apply most of **protection effort** in the **privacy-sensitive** portions

3.1. DETECTION

• Finished phase •

Overview

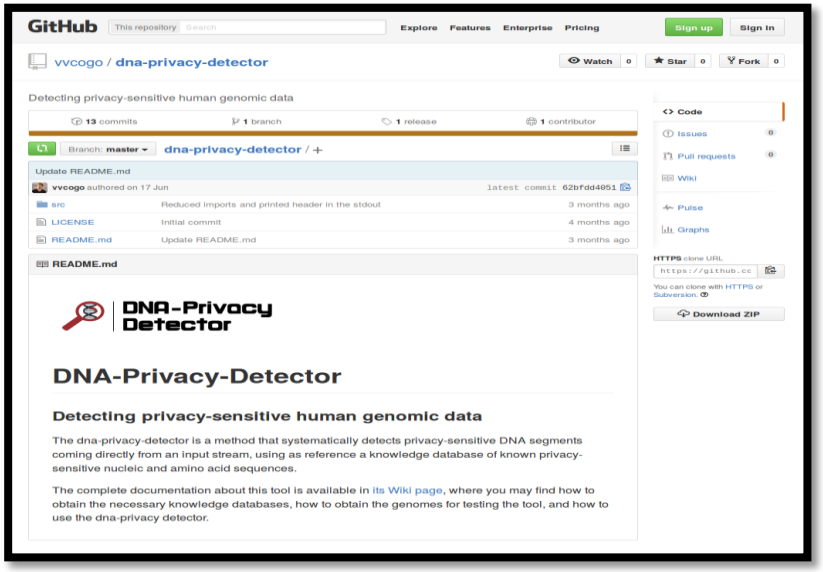


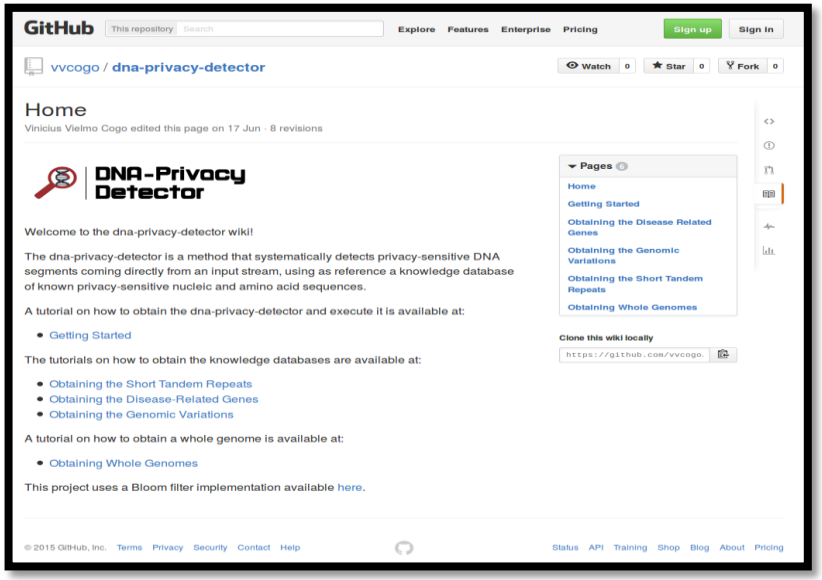
	STRs	Genes	Variations	
Databases	TRDB	GeneCards	1000 Genomes Project	Total
Number of entries	240k	20k	38M	38.3M
DB sequences	22M	8.7M	1147M	1178M
DB size	660MB	87MB	34.4GB	**35.1GB**

Source-code

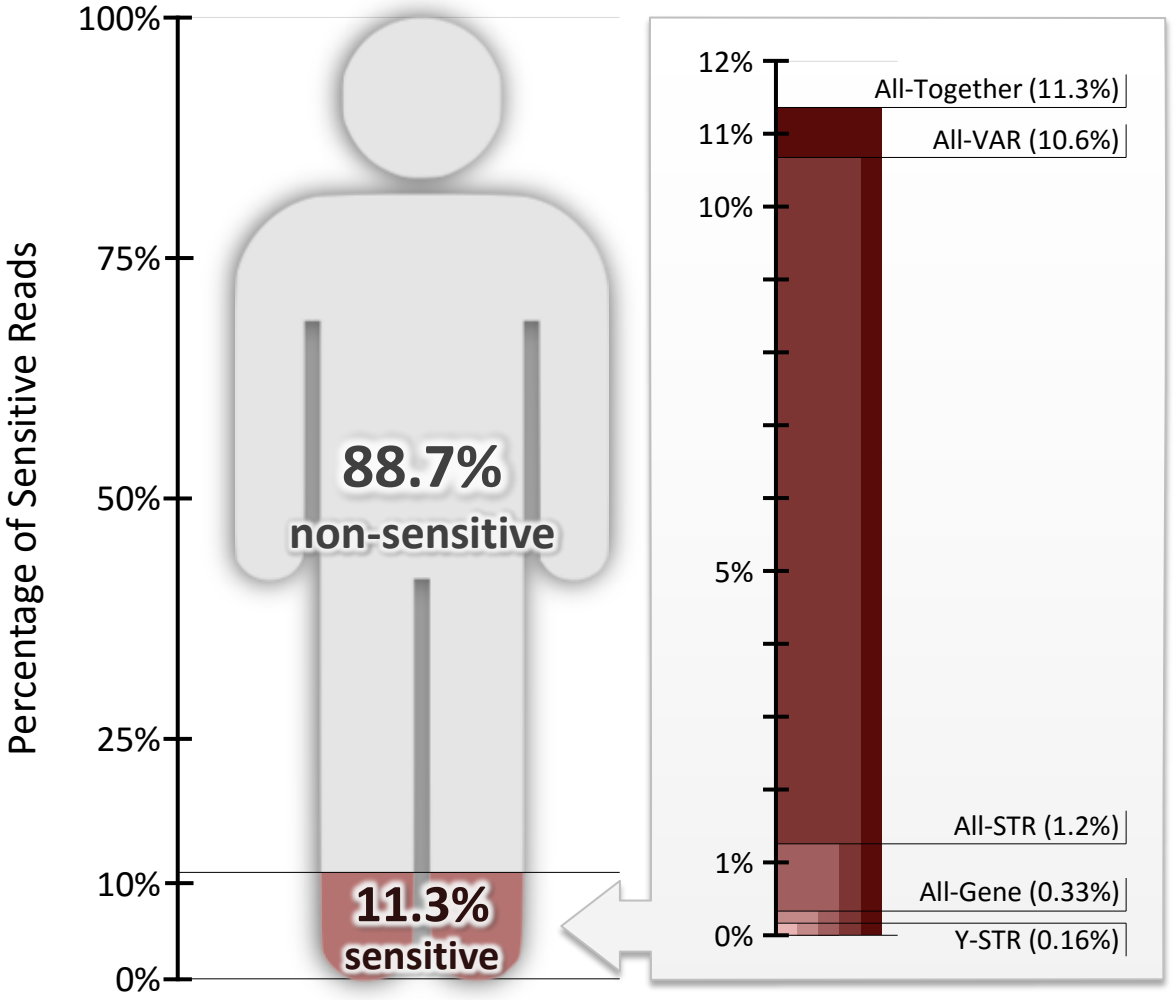
DNA-Privacy Detector

https://github.com/vvcogo/dna-privacy-detector

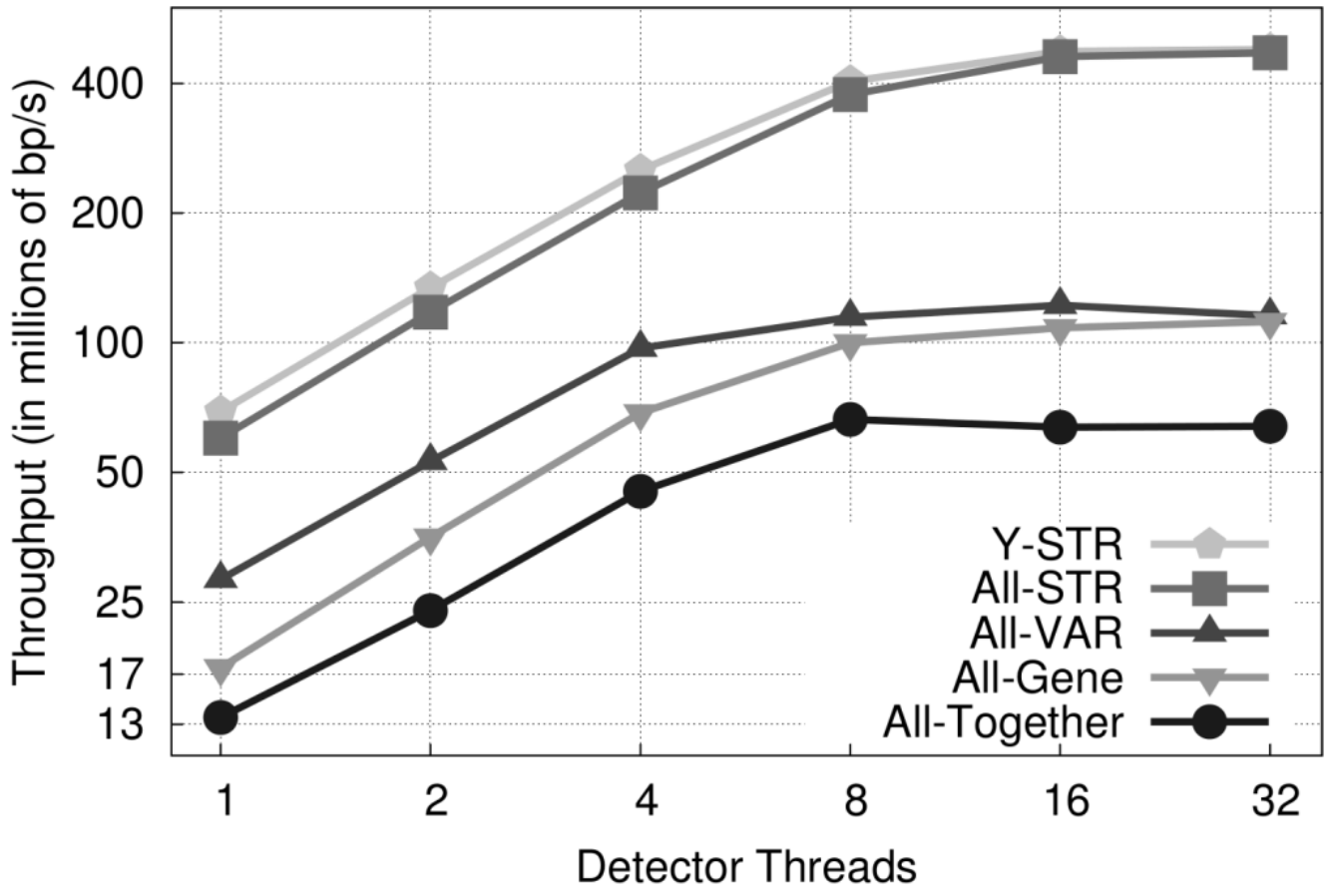




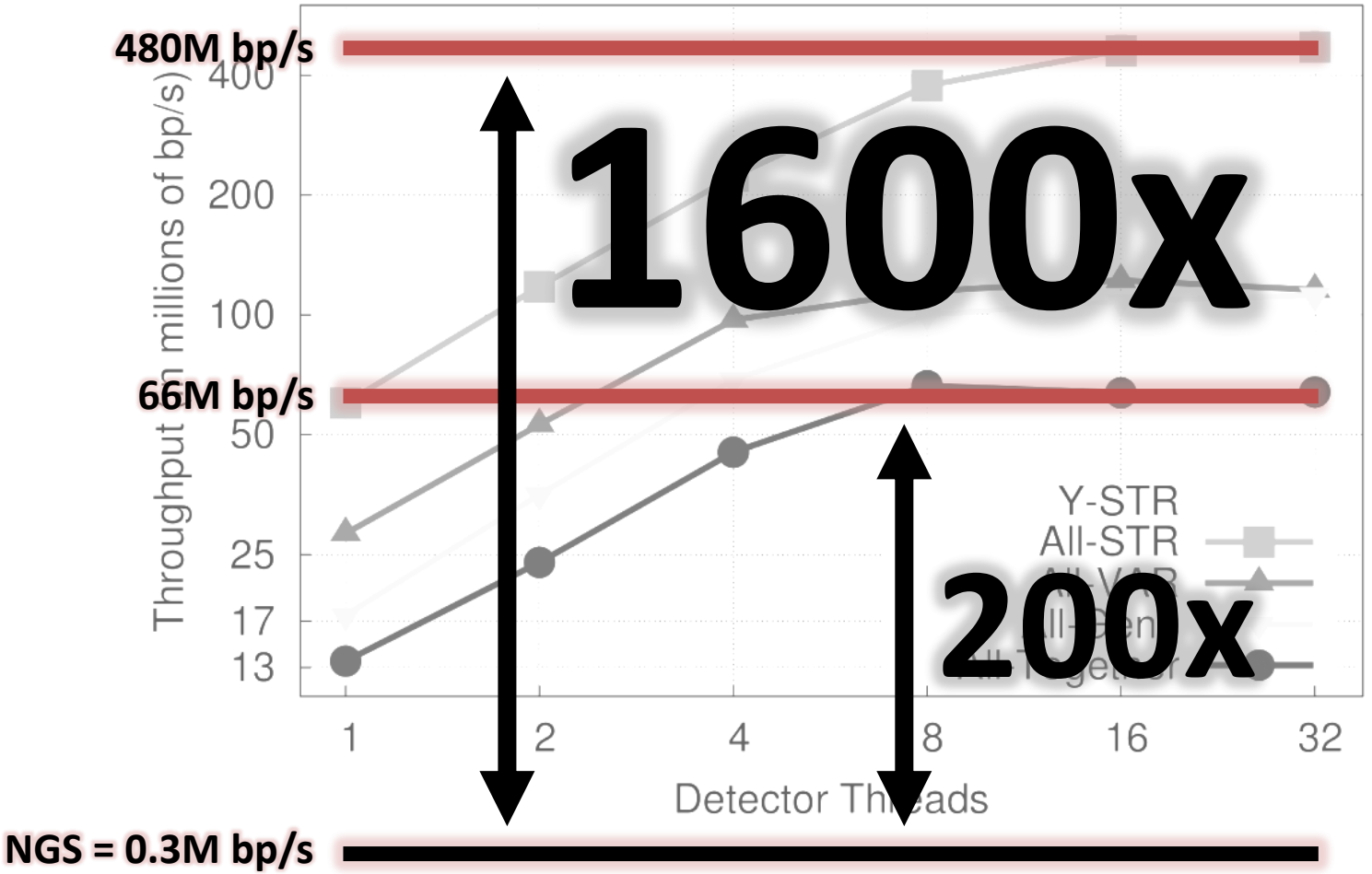
Experiment 1: Privacy-sensitivity of human genomes



Experiment 2: Detection throughput



Experiment 2: Detection throughput



Outline

2. State of the art

3. Proposal
   - DETECTION
   - **REDUCTION**
   - STORAGE
   - AUDITABILITY

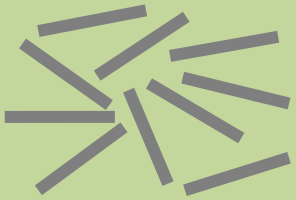
4. Plan

Overview

Reduce the size of data
from human genomes combining
compression and **deduplication**

3.2. REDUCTION

• Current phase •



FASTQ reads (unaligned)

GGGTGATGGC...

Reference genome

FASTQ reads (sorted)

FASTQ reads (compressed)

FASTQ reads (deduplicated)

Original FASTQ

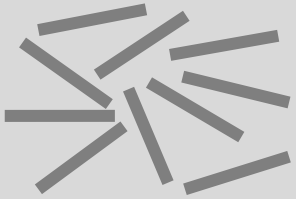
```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTTCCGATAGCGTCAAAT
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
XR#\8)X)*B,eP]‘]_Ba_^_3_[Y?7+9

@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:346
CGCTGTCGATGGCGTCACATCCCACCGGTC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:346
Z!''*((((***+))%%%++)(%%%%).1*

@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:347
TGATGGCCGCTGCCTATGGCATCAAATCCC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:347
YS$]9*Y*+C-fQ~(~`Cb`_`4`\Z@8,:
```

3.2. REDUCTION

• Current phase •



FASTQ reads (unaligned)

GGGTGATGGC…

Reference genome

FASTQ reads (sorted)

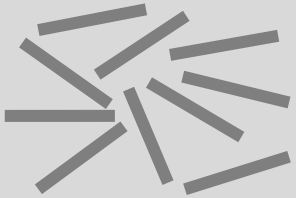
FASTQ reads (compressed)

FASTQ reads (deduplicated)

Sorting the FASTQ

3.2. REDUCTION

• Current phase •



FASTQ reads (unaligned)

GGGTGATGGC…

Reference genome

FASTQ reads (sorted)

FASTQ reads (compressed)

FASTQ reads (deduplicated)

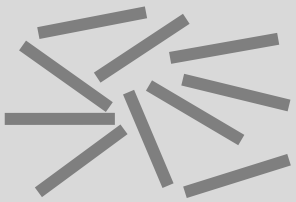
Referential compression

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817,30,HG38…
0 345 X
14T6A
000647573615474701242257211354542293102010144440402261 21…

3 2 Y
14T6A
000647573615474701242257211354542293102010144440402261 21…
7 -1 Z
5T12C
001238482205070411020123628122407052611070803230801129 8…
```

3.2. REDUCTION

• Current phase •



FASTQ reads (unaligned)

GGGTGATGGC…

Reference genome

FASTQ reads (sorted)

FASTQ reads (compressed)

FASTQ reads (deduplicated)

Deduplication

```
@SRR001666.1 071112_SLXA-
EAS1_s_7:5:1:817:,30,HG38…
0 345 X 075x3rds
3 2 Y 075x3rds
7 -1 Z 37ik3lp0
```

Outline

2. State of the art

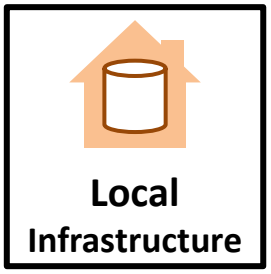
3. Proposal
   - DETECTION
   - REDUCTION
   - **STORAGE**
   - AUDITABILITY

4. Plan

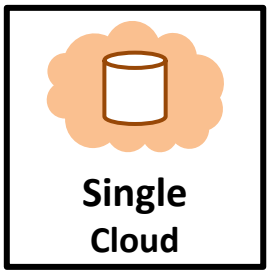
Overview

Store the **different portions**
of human genomes
in the appropriate **data repositories**

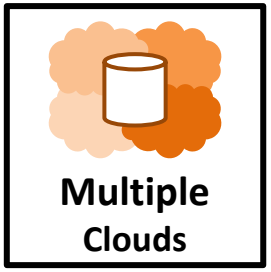
3.3. STORAGE



- 👍 Privacy-preserving
- 👎 Limited scalability
- 👎 High initial capital expenditures (e.g., $65M)
- 👎 High maintenance costs (e.g., $35M/year)



- 👍 Infinite scalability
- 👍 Pay-as-you-go
- 👎 Security and availability incidents
  - e.g., $10.5M/year using Amazon Glacier



- 👍 Dependable
- 👎 50% of space overhead
- 👎 Time overhead (secret sharing and erasure codes)
  - e.g., Depsky

Outline

2. State of the art

3. Proposal
   - DETECTION
   - REDUCTION
   - STORAGE
   - **AUDITABILITY**

4. Plan

Overview

Explore **cloud-of-clouds**
to **audit** data **access**
in **untrusted** repositories

Detect, analyze,
and sanction misuses

3.4. AUDITABILITY

Goals

- Identify **when** and **who read** the data blocks of a specific genome
- Detect if someone tried to **read** a data **block separately** in a single cloud
- Detect if some **cloud** was **compromised** and is trying to **incriminate** a data consumer by logging unrequested reads
- Detect if a **minimal** read or write **quorum** failed
  - Quorums of **3f+1** are required to audit reads (DepSky uses 2f+1)

Outline

2. State of the art

3. Proposal
   - DETECTION
   - REDUCTION
   - STORAGE
   - AUDITABILITY

4. **Plan**

4. Plan

Current State:

- DETECTION – 100% complete
- REDUCTION – currently working on it
- AUDITABILITY – next step (starting on October'16)
- STORAGE – simultaneous to all other phases

Work distribution:

	Phases	Complete	Total PM	Oct'13 – Apr'14 (Y1)		Oct'14 – Apr'15 (Y2)		Oct'15 – Apr'16 (Y3)		Oct'16 – Apr'17 (Y4)		Oct'17
				S1	S2	S3	S4	S5	S6	S7	S8	
1	DETECTION	100%	16.8	4.8	4.8	4.8	2.4					
2	REDUCTION	66%	13.2				2.4	4.8	4.8	1.2		
3	AUDITABILITY	0%	6.0							3.6	2.4	
4	STORAGE	66%	7.2			1.2	1.2	1.2	1.2	1.2	1.2	
5	THESIS	0%	2.4								2.4	

M1 Doctoral Course — **M2** **PhD qualification** — **M3** PhD thesis

4. Publications

Direct from this PhD:

✓ [DETECTION] V. V. Cogo, A. Bessani, F. M. Couto, and P. Verissimo.

A high-throughput method to detect privacy-sensitive human genomic data.

In Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society (**WPES 2015**), pages 101110, Denver, CO, US, Oct. 2015. ISBN 978-1-4503-3820-2. doi: 10.1145/2808138.2808139.

✓ [STORAGE] V. V. Cogo and A. Bessani.

From data islands to sharing data in the cloud: the evolution of data integration in biological data repositories.

Communications and Innovations Gazette (**ComInG**), 1(1):111, Jan. 2016.

Related to this PhD:

➢ [STORAGE] A. Bessani, J. Brandt, M. Bux, V. V. Cogo, L. Dimitrova, J. Dowling, A. Gholami, K. Hakimzadeh, M. Hummel, M. Ismail, E. Laure, U. Leser, J. Litton, R. Martinez, J. Reichel, S. Niazi, and K. Zimmermann.

BiobankCloud: a platform for the secure storage, sharing, and processing of large biomedical data sets.

In Proceedings of the 1st Int. Workshop on Data Management and Analytics for Medicine and Healthcare (**DMAH 2015**), Hawaii, US, Sept. 2015.

➢ [REDUCTION] F. Alves, V. V. Cogo, S. Wandelt, U. Leser, and A. Bessani.

On-demand indexing for referential compression of DNA sequences.

In **PLoS ONE**, 10(7):e0132460, July 2015.

4. Publications

Submitted:

- [STORAGE] R. Mendes, T. Oliveira, V. V. Cogo, N. Neves, and A. Bessani.
  CHARON : A dependable cloud-of-clouds system for storing and sharing big data.
  Submitted to **ACM Transactions on Storage**, 2016.

- [DETECTION] V. V. Cogo, A. Bessani, F. M. Couto, M. Carvalho, and P. Verissimo.
  How can photo sharing inspire sharing genomes?
  Submitted to **IEEE Security & Privacy**, 2016.

Planned submissions:

- [REDUCTION] FAST'17 (deadline: **September 27**, 2016)
  or USENIX ATC'17 (if rejected in FAST, deadline: February, 2017)

- [AUDITABILITY] DSN'17 (deadline: **December 5**, 2016)

- [STORAGE] Bioinformatics journal (no deadline)

Ciências
ULisboa

LISBOA

UNIVERSIDADE
DE LISBOA

Thank you!

Completeness of the Detection method

- **Novel STRs:**
  - Not in STR databases = useless for attackers
- **Novel genes:**
  - Do not determine alone the contraction of a disease
  - May have no relation with any disease
  - May affect very few people (rare diseases)
  - Novel discoveries = correlate diseases with known genes
- **Novel genomic variations:**
  - No variation determines alone the identity or contraction of a disease

Deduplication using an LSH index

- **Write:**
  - Read the quality scores QS
  - Search the nearest neighbors of QS in the LSH index
  - Choose the best neighbor BN from the returned candidates
  - Calculate the minimal edit sequence between QS and BN
  - Compress the edit sequence and store it with the pointer to BN
- **Read:**
  - Decompress the edit sequence and the pointer to BN
  - Retrieve the best neighbor BN
  - Apply the sequence of edits in BN to obtain QS

Additional Opportunities

- **Advancing the detection:**
  - Novel privacy-sensitive sequences
  - Additional types of data
  - Increasing detector's accuracy with alignment
- **Deduplicating registers:**
  - Formal descriptions of deduplication in registers
- **Federated deduplication:**
  - Deduplicating privacy-sensitive data
  - Convergent encryption, message-locked encryption, trusted components