UNIVERSIDADE DE LISBOA • FACULDADE DE CIÊNCIAS • DEPARTAMENTO DE INFORMÁTICA

Detecting Privacy-Sensitive Human Genomic Data

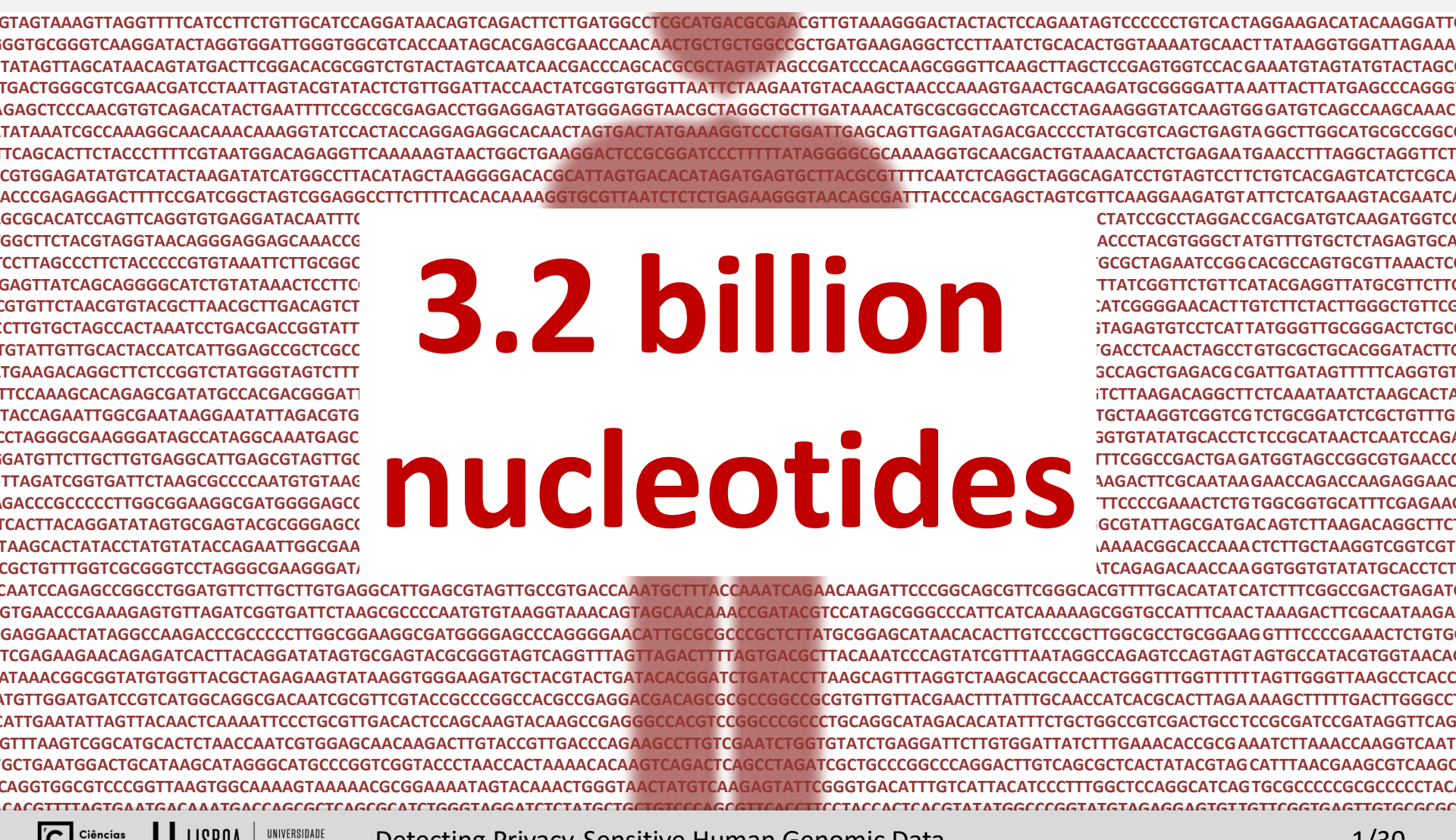
PhD. Student: Vinicius Vielmo Cogo

Advisor: Prof. Dr. Alysson Neves Bessani

February 16, 2017

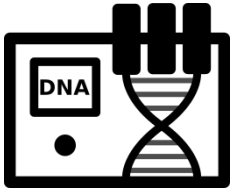
1. Context

Human genome

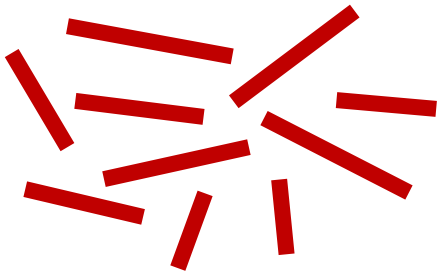


1. Context

Genomic data

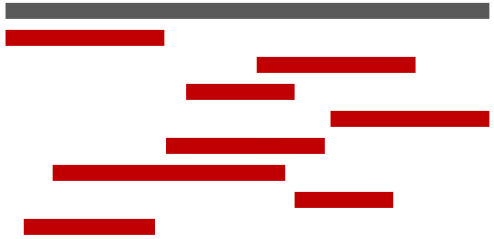


Sequencing



FASTQ
(raw)
(unaligned)

Alignment



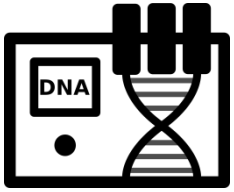
SAM
(aligned)

Assembly

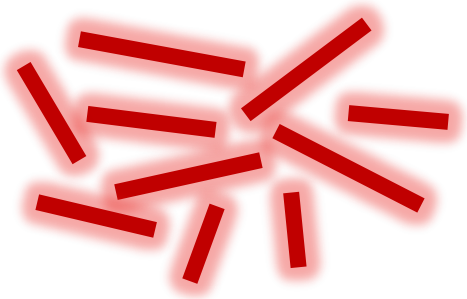
FASTA
(contiguous)

1. Context

Genomic data

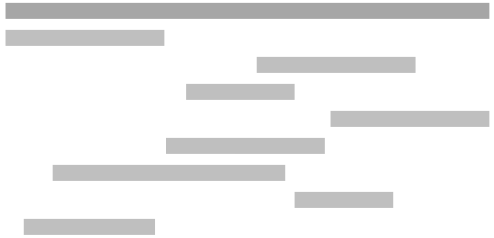


Sequencing



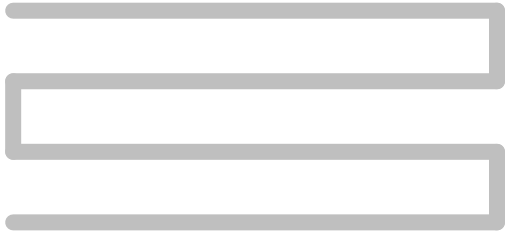
FASTQ (raw) (unaligned)

Alignment



SAM (aligned)

Assembly



FASTA (contiguous)

1. Context

Motivation

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

1. Context

Motivation

Critical:

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

1. Context

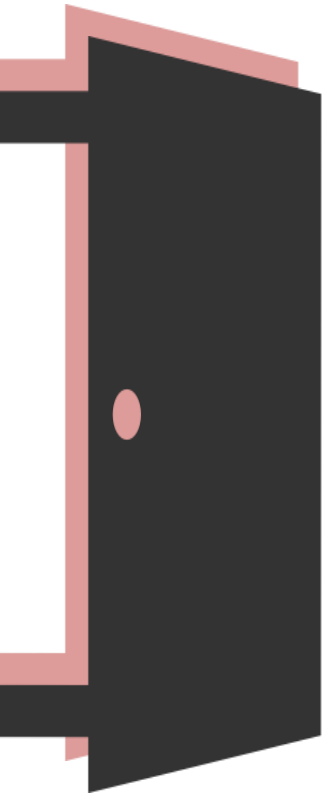
Motivation

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. Context

Open problems



Privacy-awareness:

- ✓ Privacy-protection with **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

1. Context

The PhD

Title:

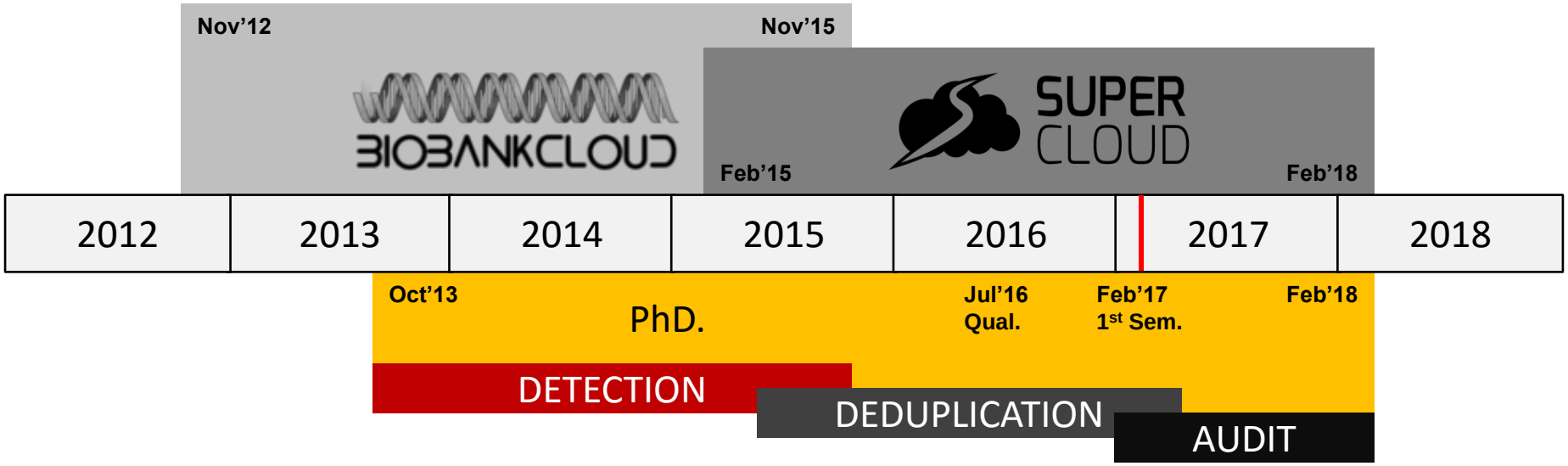
Efficient Storage of Whole Human Genomes

Main Goal:

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

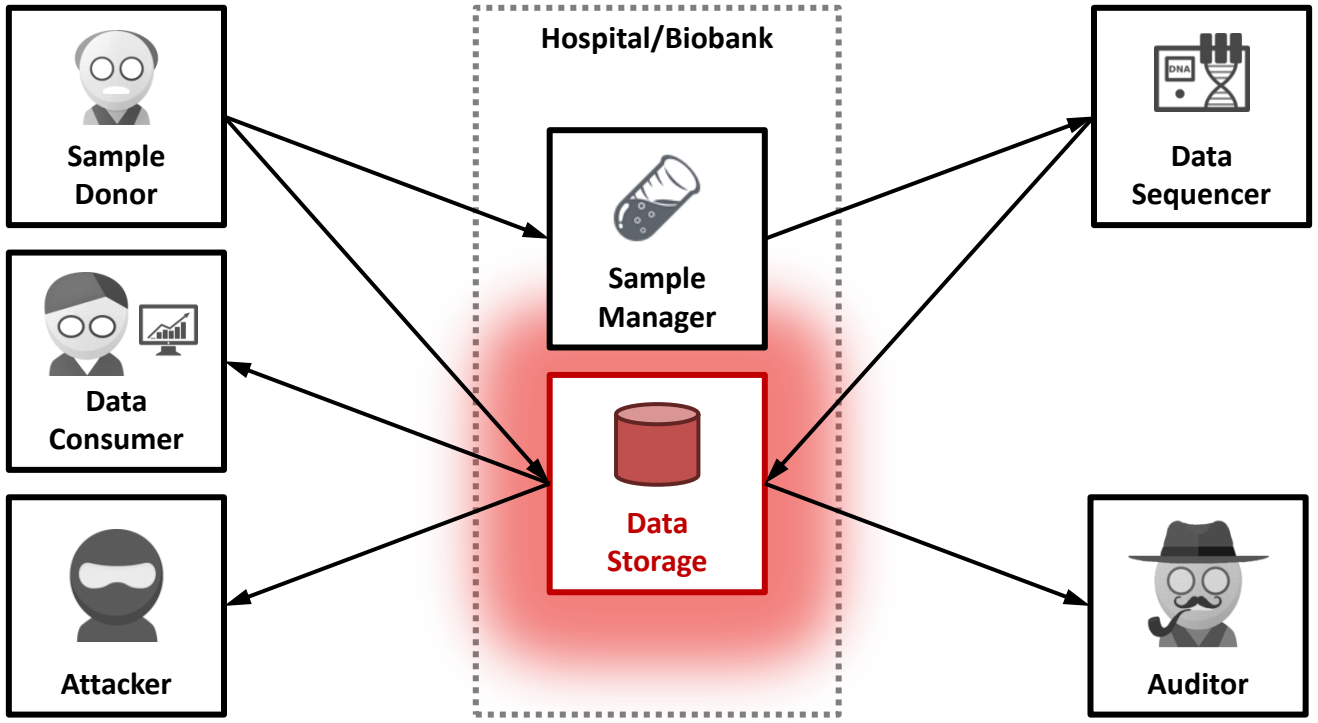
1. Context

The PhD



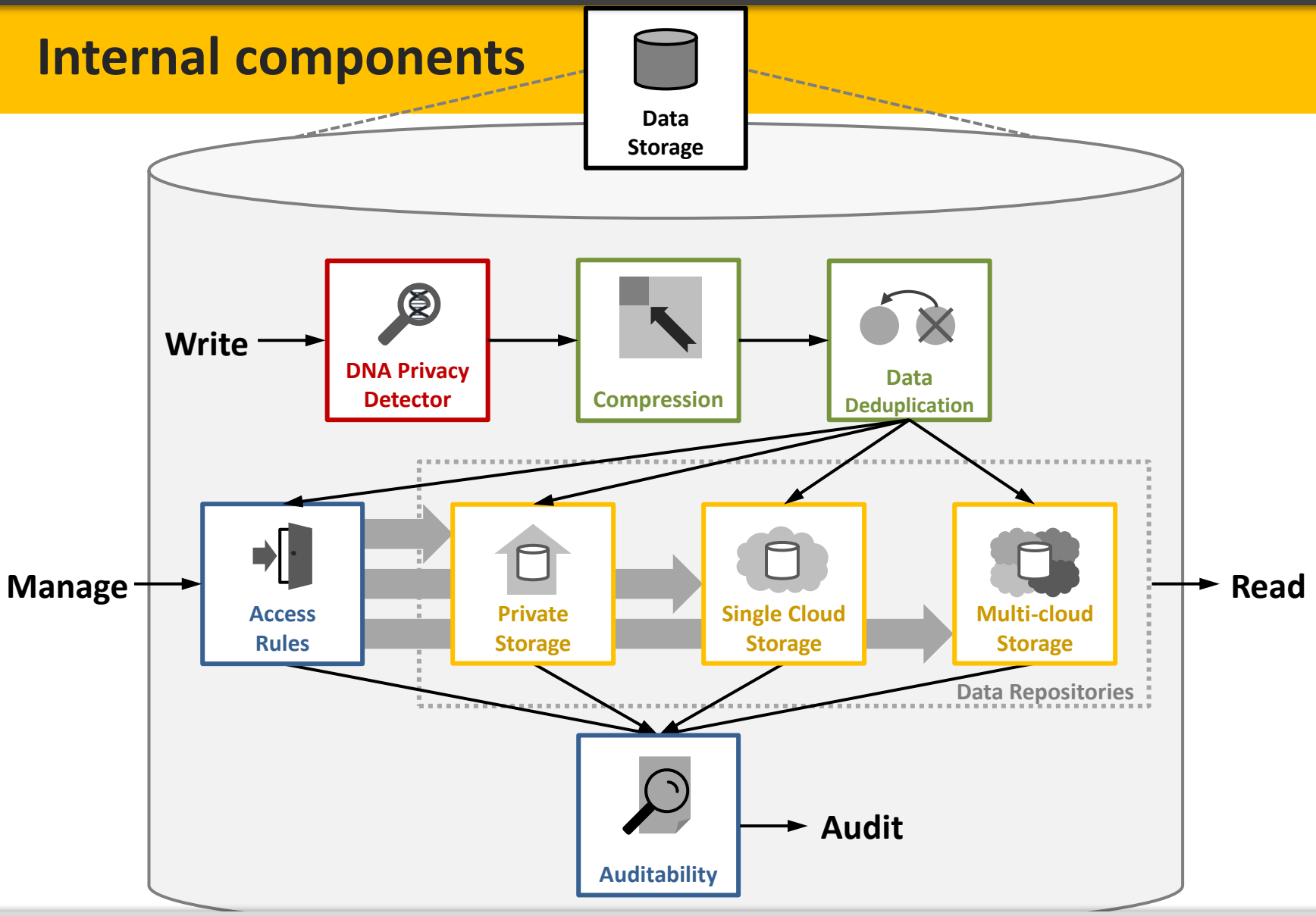
1. Context

Ecosystem



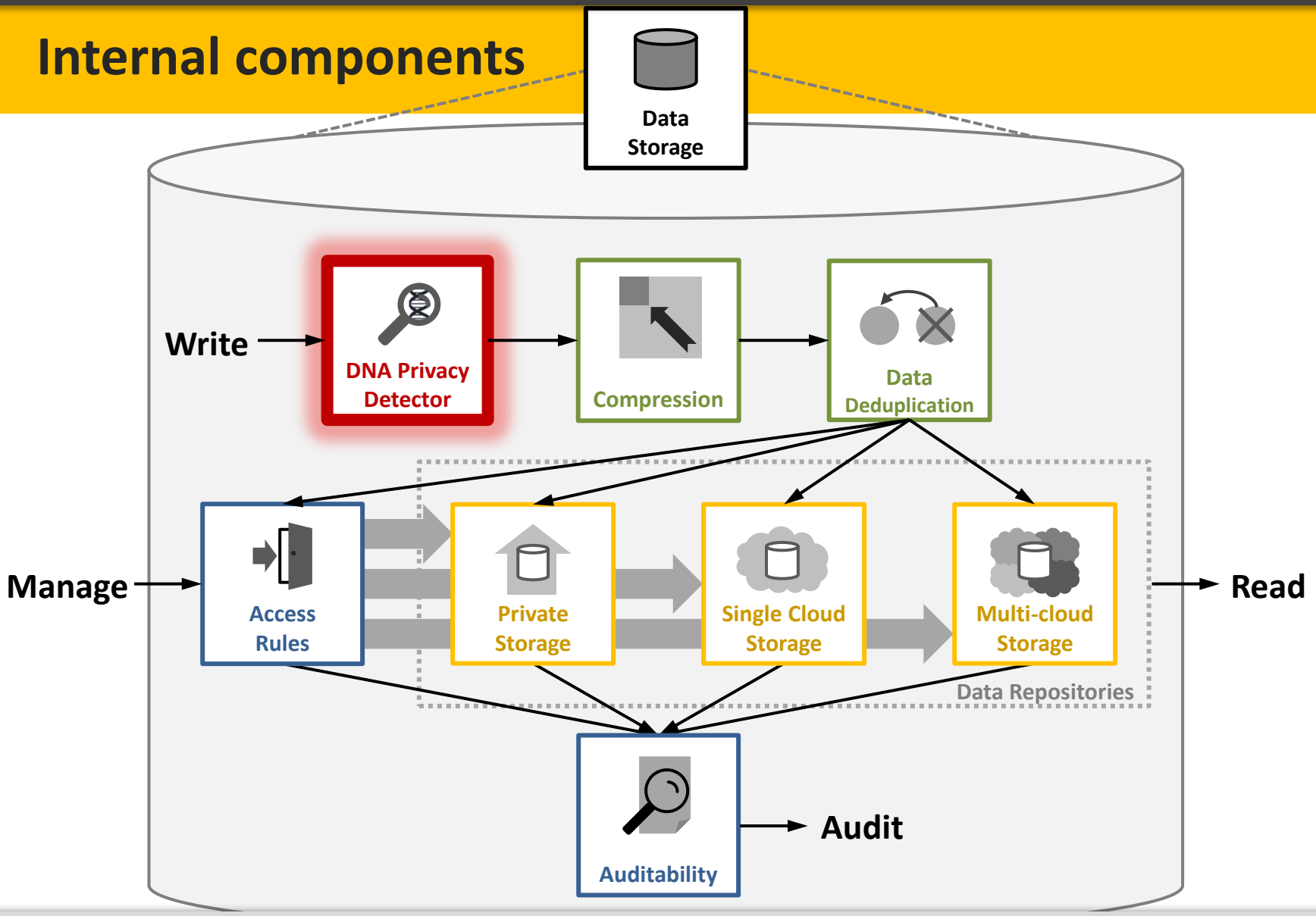
1. Context

Internal components



1. Context

Internal components



2. The detection

Detect the relevant portions for privacy

2. The detection

Concentrate our efforts

most effort

TCCTTAAATGAGC
ACCTCTTCAGTAG
GCCGTTAAGACCT
CTGAACGCGTGGA
CCCATCGCACGTG
GAGGATACAATTT
TAGGACCGACGAT
TGTCCCGCTTGGC
GAAACTCTGTGGC
ACGCGGGTAGTCA
TGACGCTTACAAA

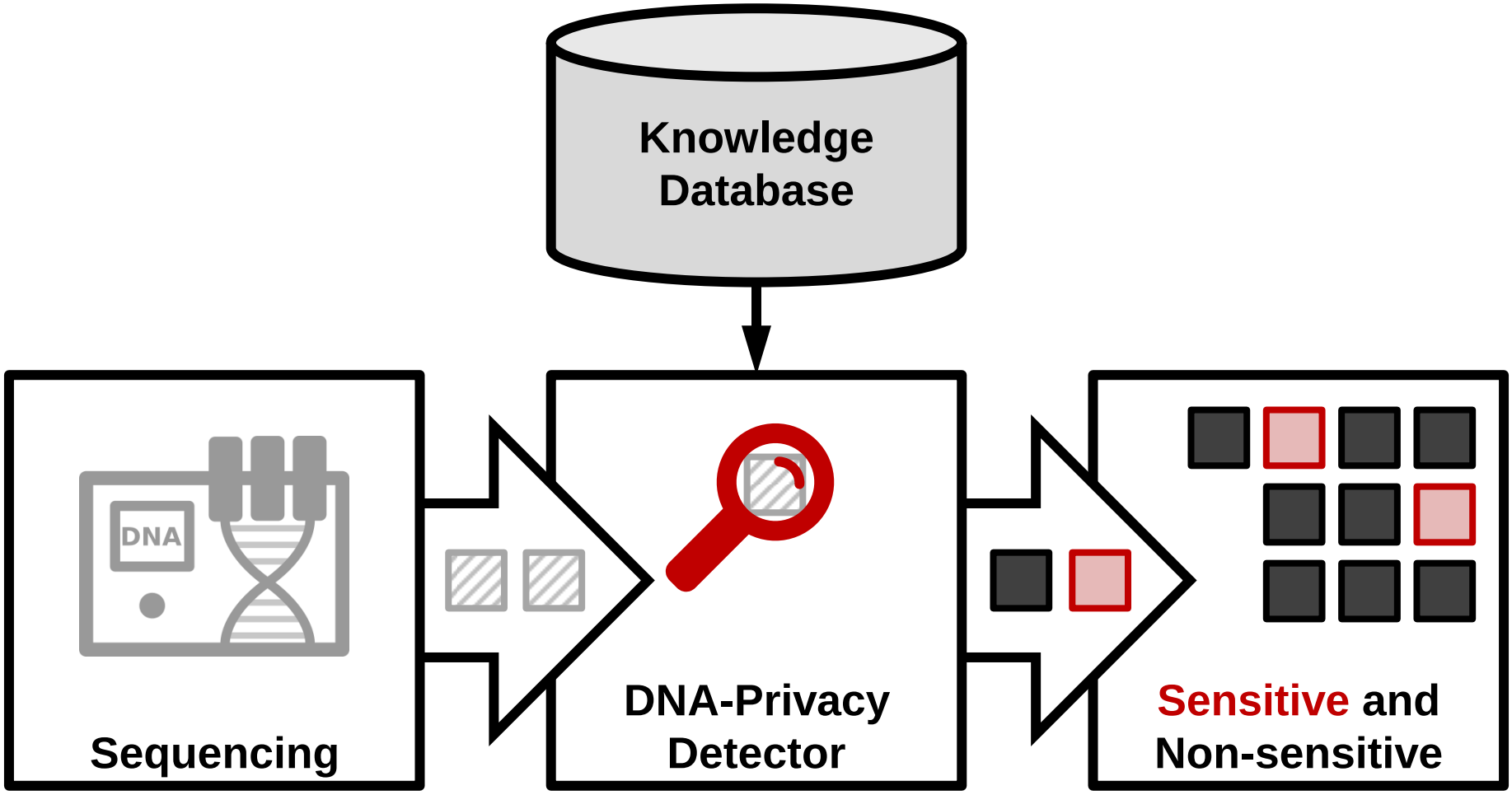
2. The detection

Specific contributions

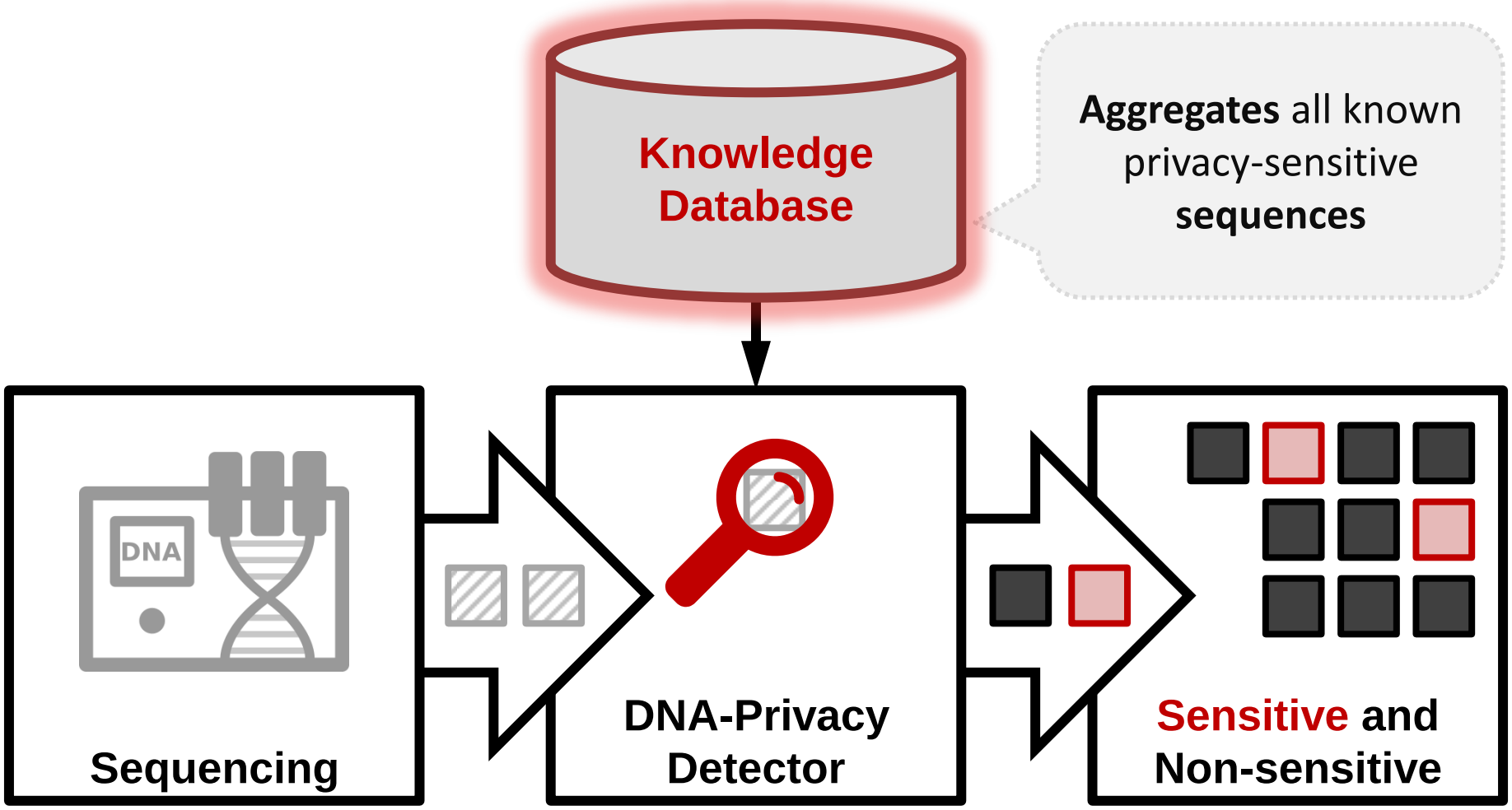
1. a **study** on the **privacy-sensitiveness** of human genomes

2. a **method** to **detect** the privacy-sensitive genomic data

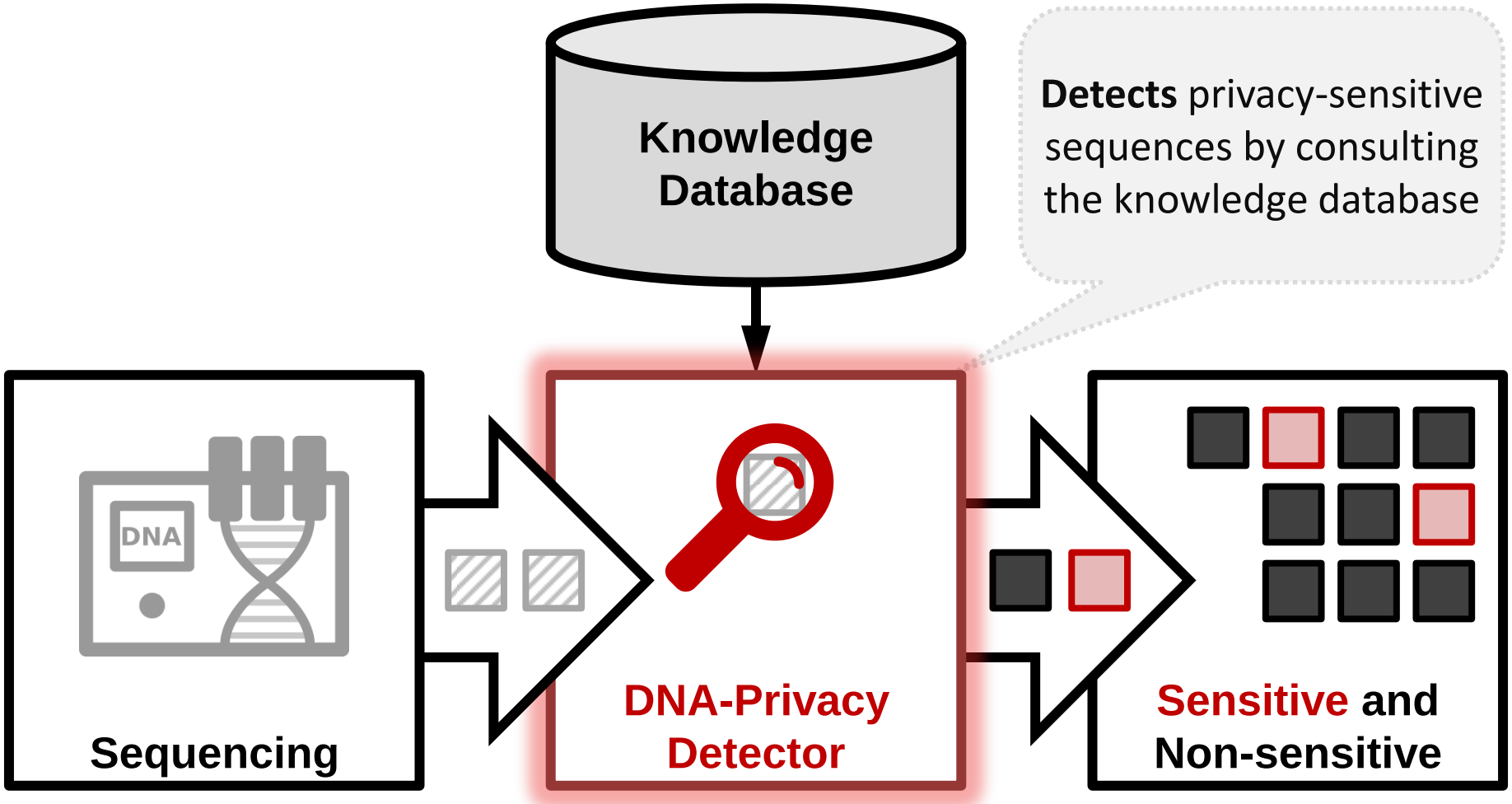
2. The detection



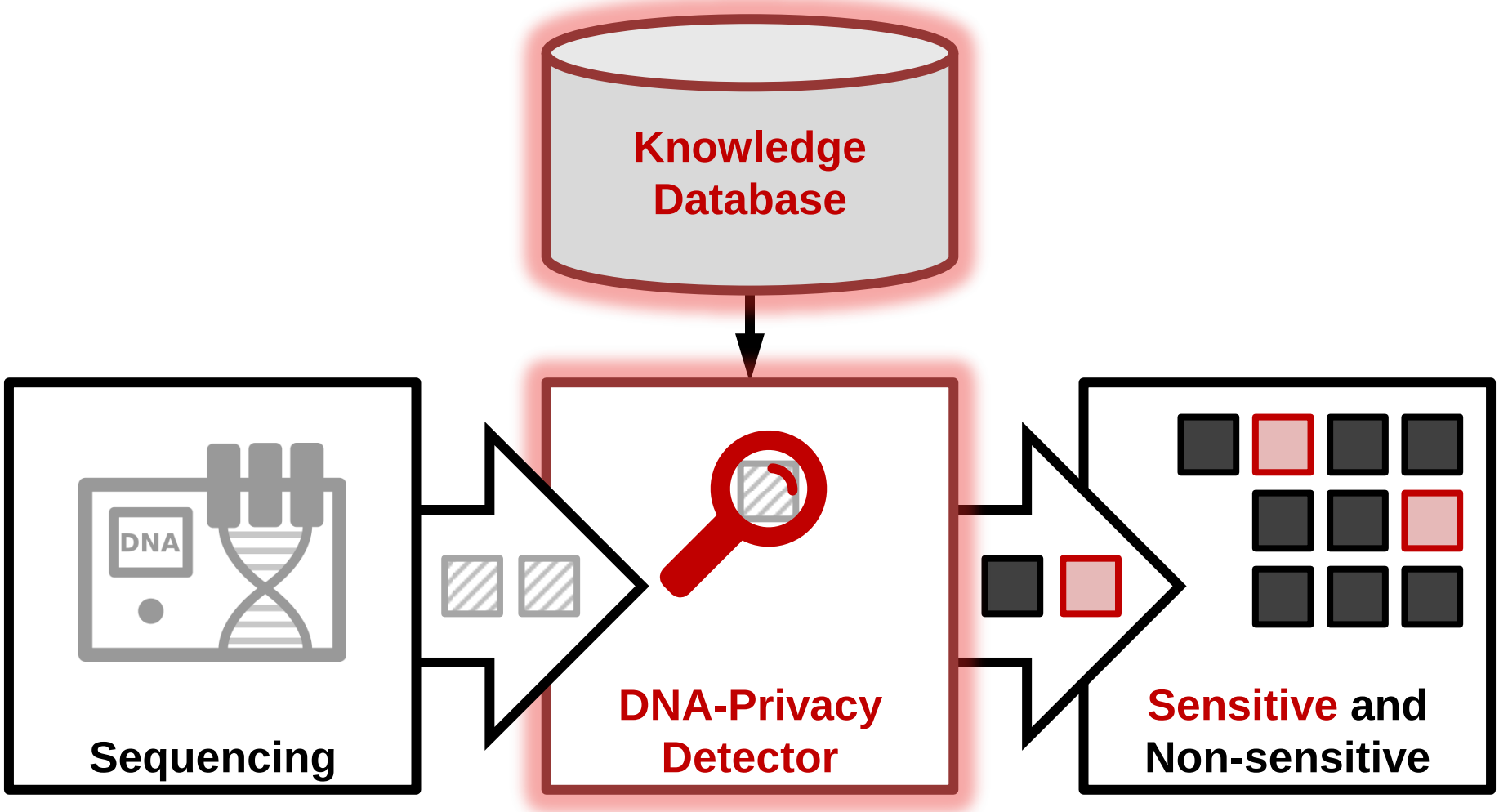
2. The detection



2. The detection



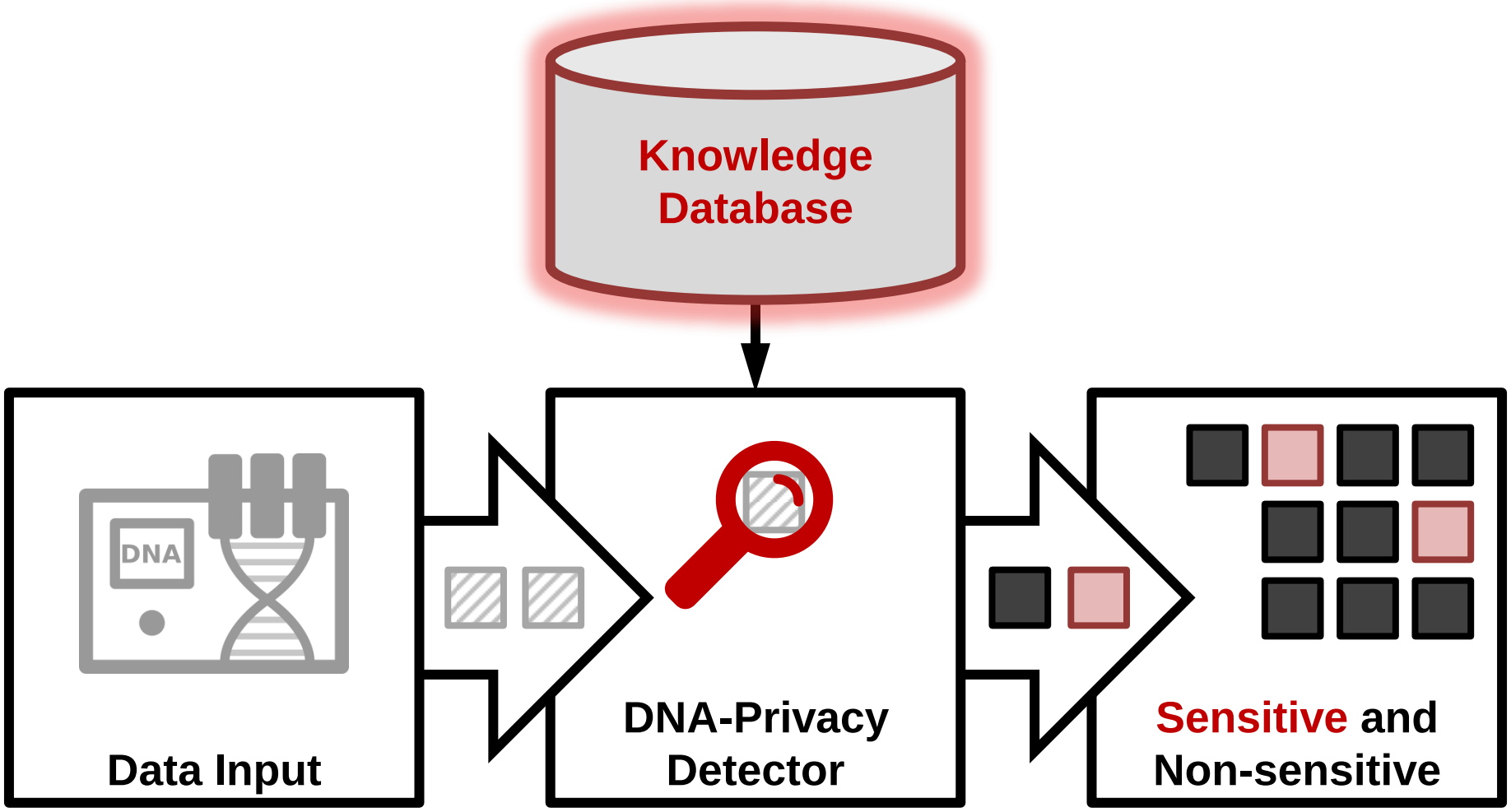
2. The detection



Outline

3. Creation of knowledge databases
4. Implementing the detector
5. Evaluation
6. Final remarks

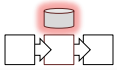
3. Creating knowledge databases



3. Creating knowledge databases

Overview

- Privacy-sensitive sequences
  - **disclose private information** about donors
- We analyzed **several attacks** in the literature
- Attackers' **goals**:
  - **Re-identification**
  - **Extraction** of **sensitive** information
- **Threats**:
  - Short tandem repeats (STRs)
  - Disease-related genes
  - Genomic variations



3. Creating knowledge databases

Short tandem repeats (STR)

- Small strings repeated several times

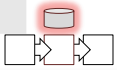
DYS392 = $[TAT]_n$

cgac TAT TAT TAT TAT cgca

n=4

- Individual profile:

STR	n
DYS392	4
DYS396	23
…	…
DYS 618	17



3. Creating knowledge databases

Disease-related genes

Disease Risks (100)

Elevated Risks	Your Risk	Average Risk
Gallstones new	11.1%	7.0%
Restless Legs Syndrome	2.5%	2.0%

more »

Decreased Risks	Your Risk	Average Risk
Prostate Cancer ♂	12.7%	17.8%
Alzheimer's Disease new	4.9%	7.2%
Colorectal Cancer	4.2%	5.6%

more »

See all 100 risk reports...

Carrier Status (24)

Hemochromatosis	Variant Present
Alpha-1 Antitrypsin Deficiency	Variant Absent
Bloom's Syndrome	Variant Absent
BRCA Cancer Mutations (Selected)	Variant Absent
Canavan Disease	Variant Absent
Cystic Fibrosis	Variant Absent
Familial Dysautonomia	Variant Absent
Factor XI Deficiency	Variant Absent

See all 24 carrier status...

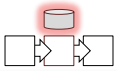
Traits (50)

Alcohol Flush Reaction	Does Not Flush
Bitter Taste Perception	Can Taste
Earwax Type	Wet
Eye Color	Likely Brown
Hair Curl	Slightly Curlier Hair on Average

See all 50 traits...

Drug Response (19)

Warfarin (Coumadin®) Sensitivity	Increased
Abacavir Hypersensitivity	Typical
Alcohol Consumption, Smoking and Risk of Esophageal Cancer	Typical
Clopidogrel (Plavix®) Efficacy	Typical
Fluorouracil Toxicity	Typical



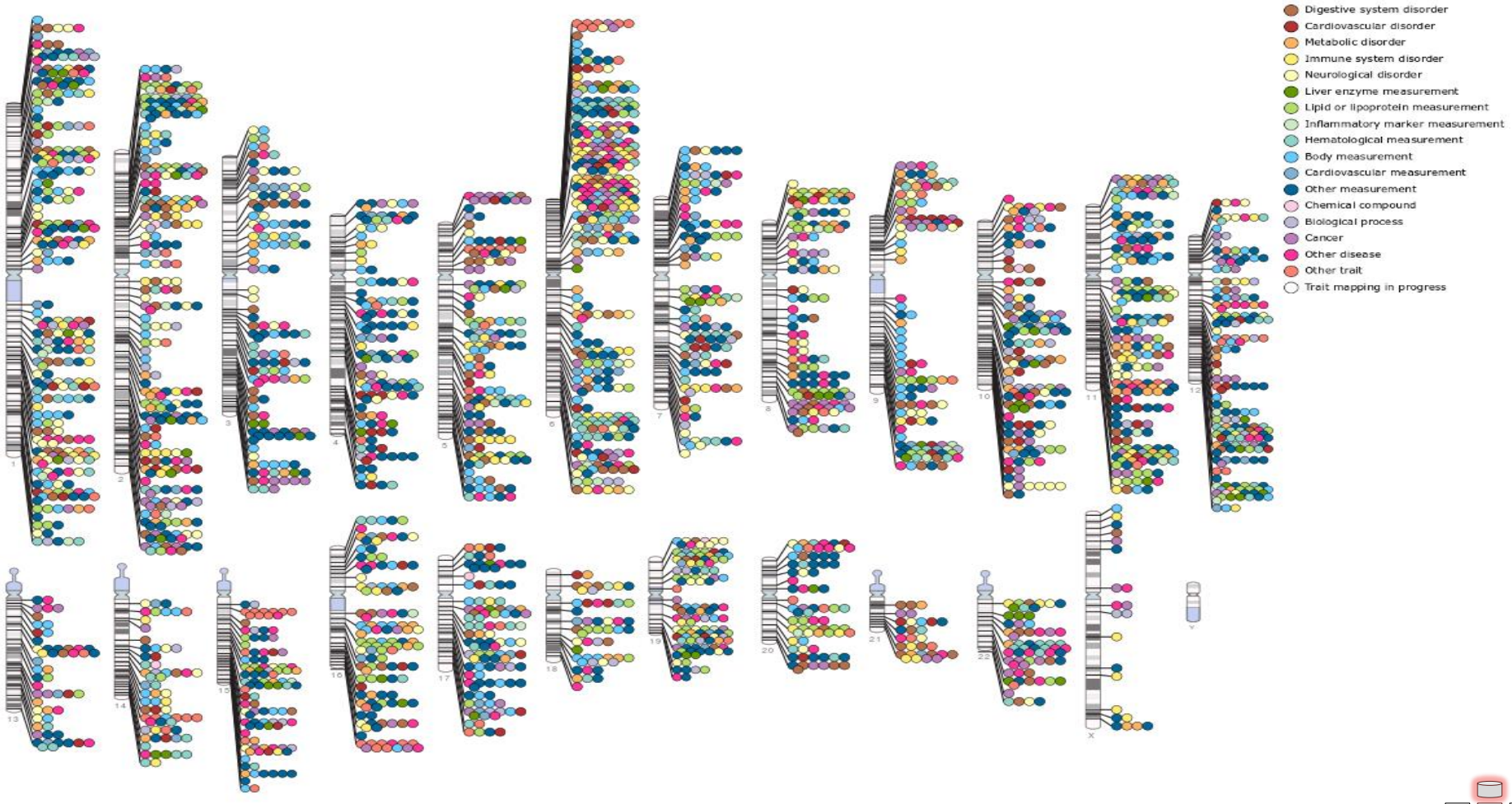
3. Creating knowledge databases

Genomic variations

<0.5%

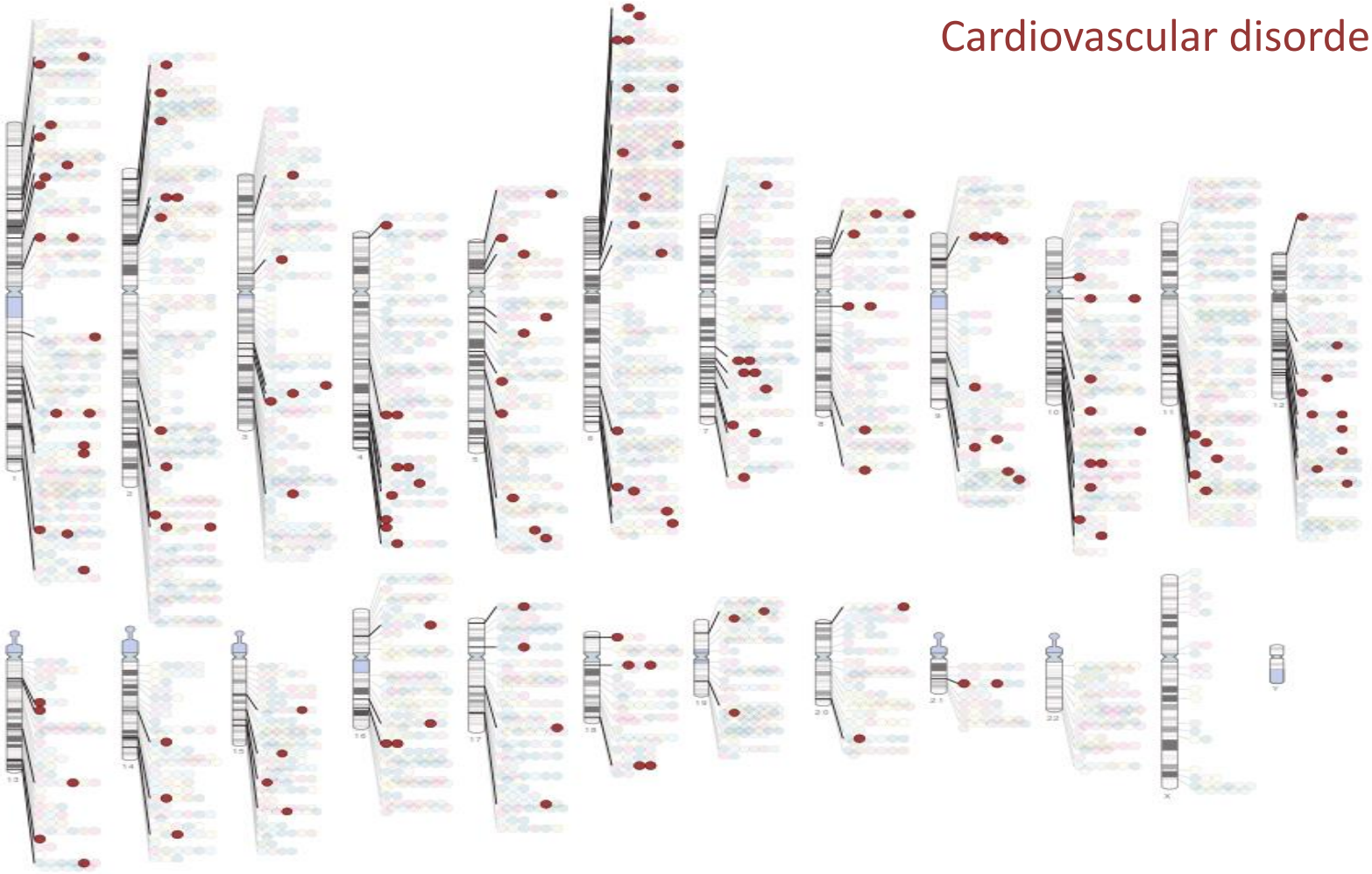
3. Creating knowledge databases

Genomic variations



3. Creating knowledge databases

Genomic variations



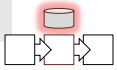
3. Creating knowledge databases

Known attacks

Short tandem repeats	
Attacks	**References**
Use STR profiles to identify donors of 1000 Genomes Project	Gymrek'13
Forensic identification	Butler'06

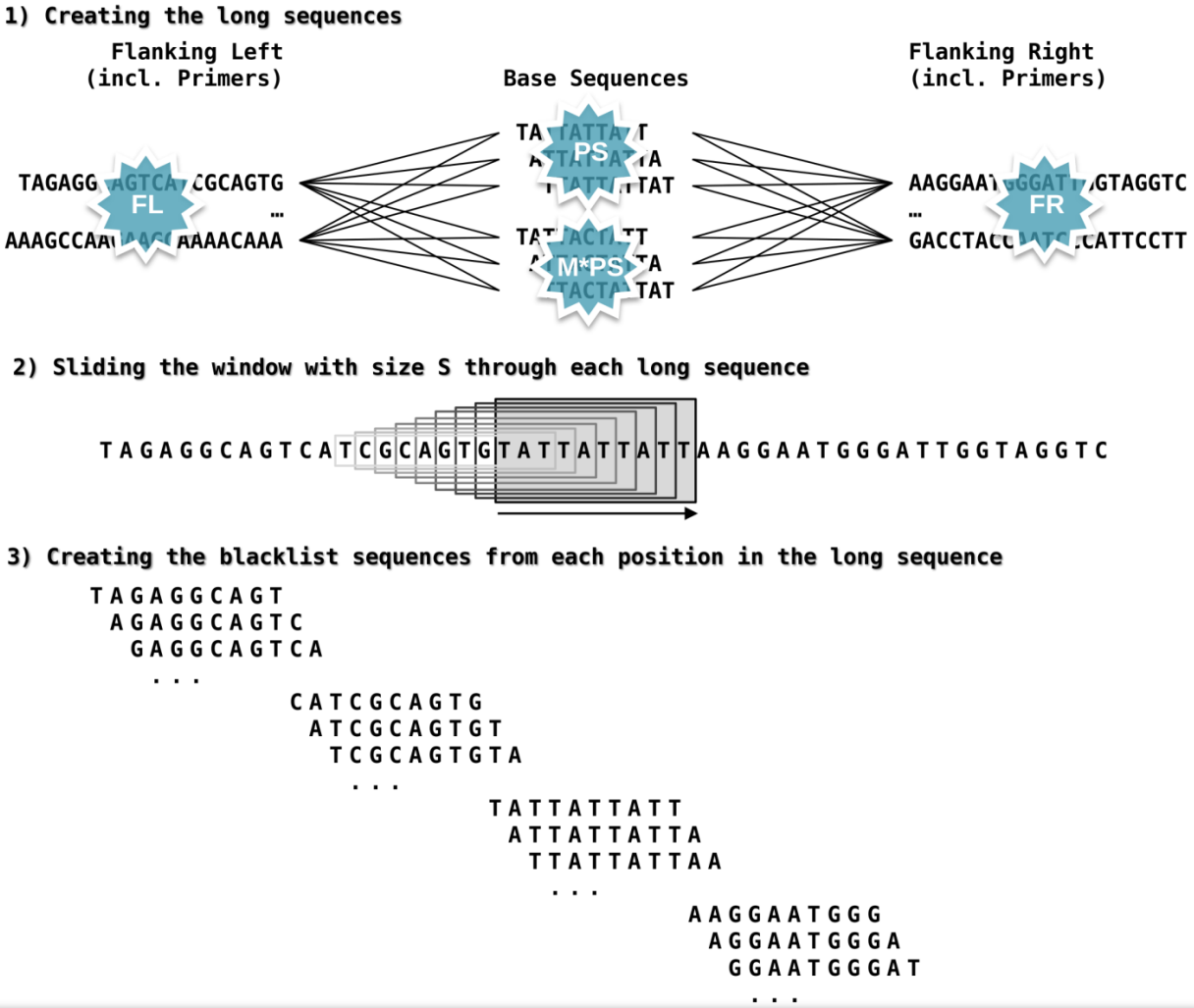
Disease-related genes	
Attacks	**References**
Direct-to-consumer genomic testing	Goldsmith'12
Masking the APOE gene (related to Alzheimer) from Dr. Watson's genome	Wheeler'08

Genomic variations	
Attacks	**References**
Re-identification (few hundred SNPs are enough)	Lin'04
Acquire knowledge about targets from GWAS results	Wang'09
Acquire knowledge about targets from microarray results	Homer'08
Infer masked genes (e.g., the APOE gene from Dr. Watson)	Nyholt'09



3. Creating knowledge databases

Example: obtaining the STRs

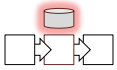


3. Creating knowledge databases

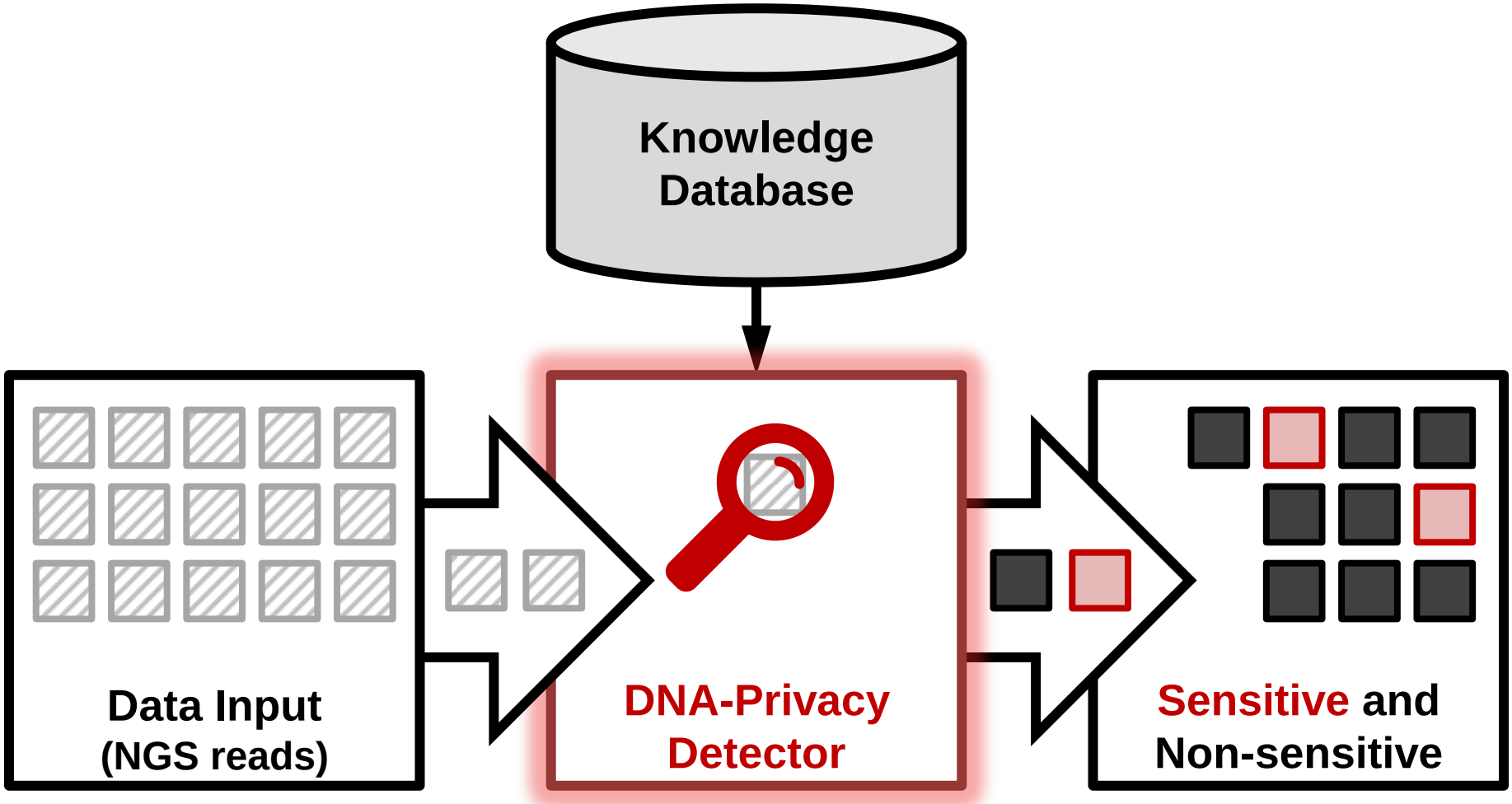
Obtaining the sequences

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

Note: Any database can be used with our solution



4. Implementing the detector



4. Implementing the detector

Requirements

- **Answer** if a genomic sequence **is privacy-sensitive or not**

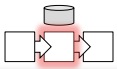
No label or pattern available (SocialSecurityNumber=123-45-6789)

- Fetch and **lookup large** amounts of **data**

Tens of GB

- Support a **high throughput** and **scale out**

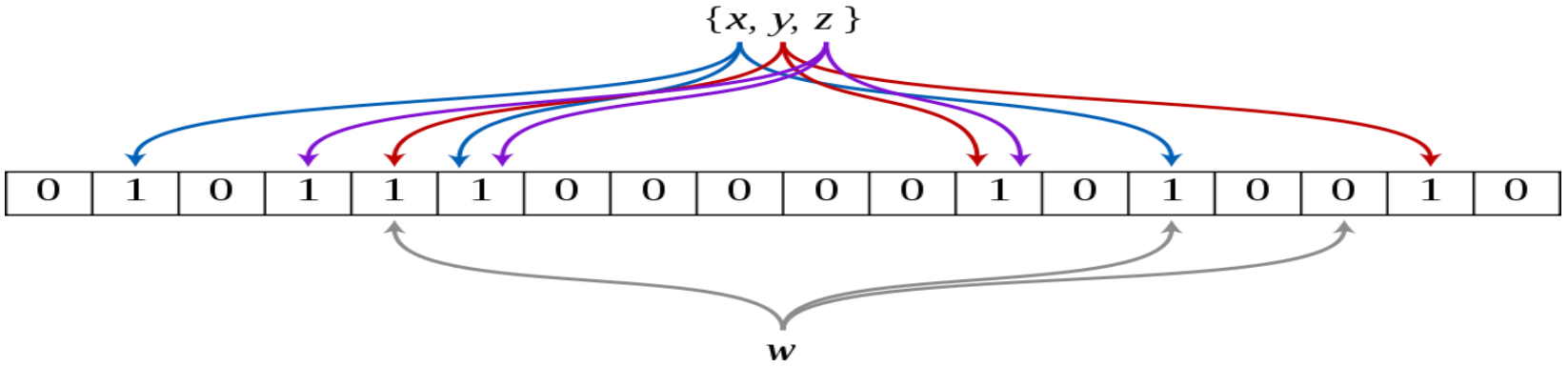
A single NGS machine generates more than 300 000 bp/s



4. Implementing the detector

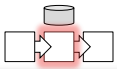
Bloom filter

- **Efficient** data structure (space and performance)
- Test if an element **is member** of a set



- Does a specific value belong to the set?
  - **Definitely not** (no false negatives)
  - **Maybe** (configurable false positives)

- False positive affects efficiency only (not efficacy)

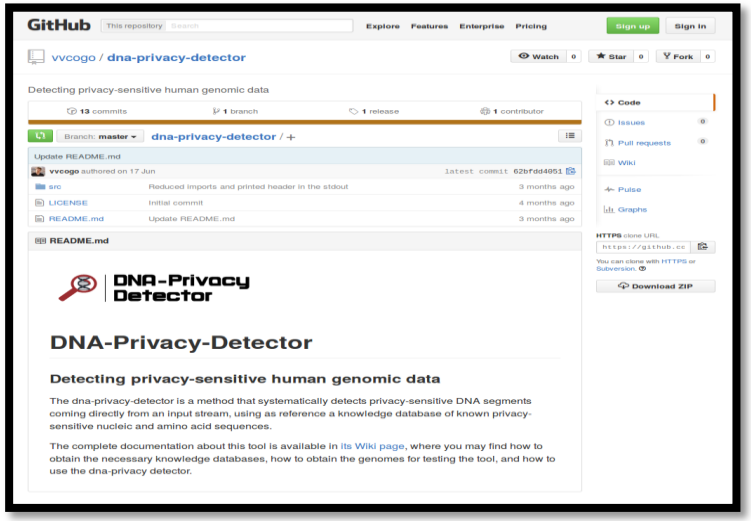


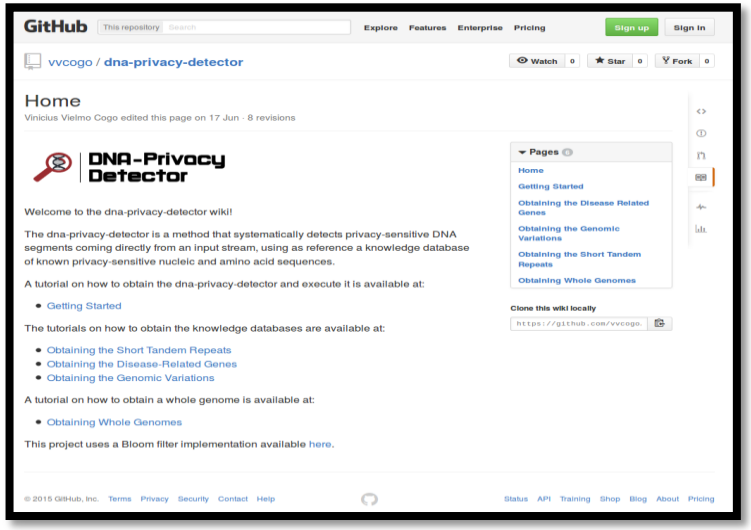
4. Implementing the detector

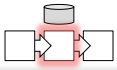
Open source-code

DNA-Privacy Detector

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







5. Evaluation

Experiments

1. Percentage of privacy-sensitive sequences in a human genome

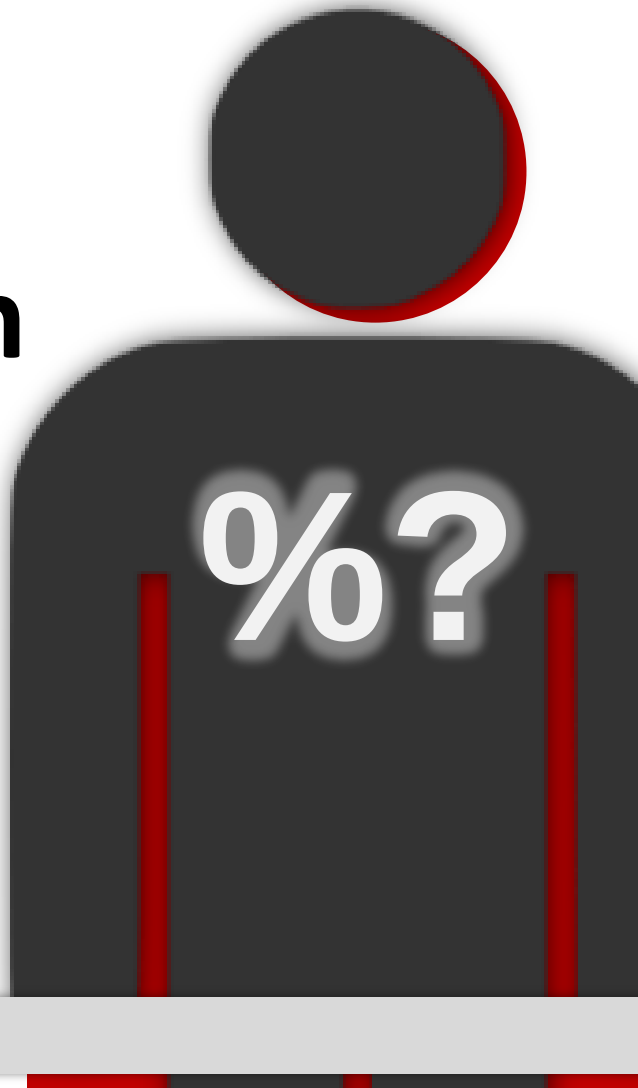
2. Bloom filter size

3. Detector’s throughput

5. Evaluation

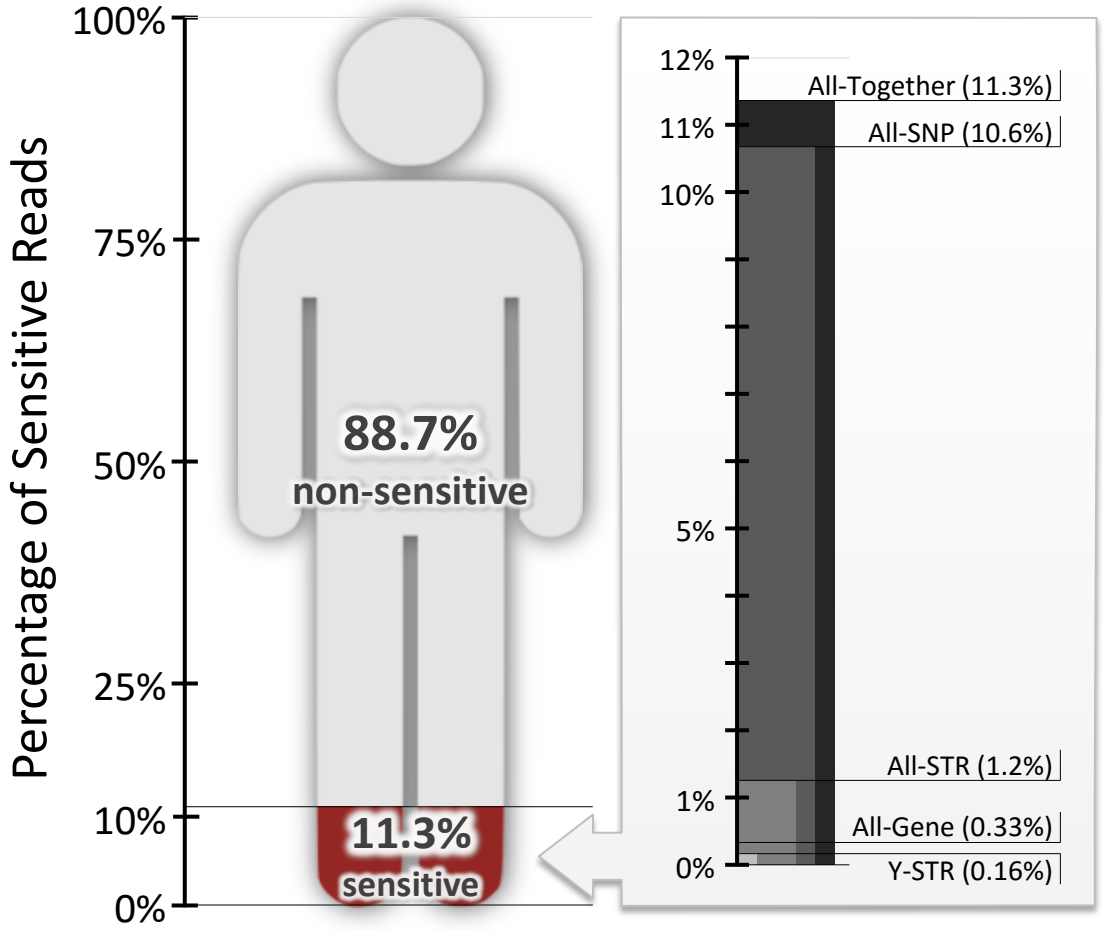
1st experiment

How much of a **human genome** is considered **privacy-sensitive**?



5. Evaluation

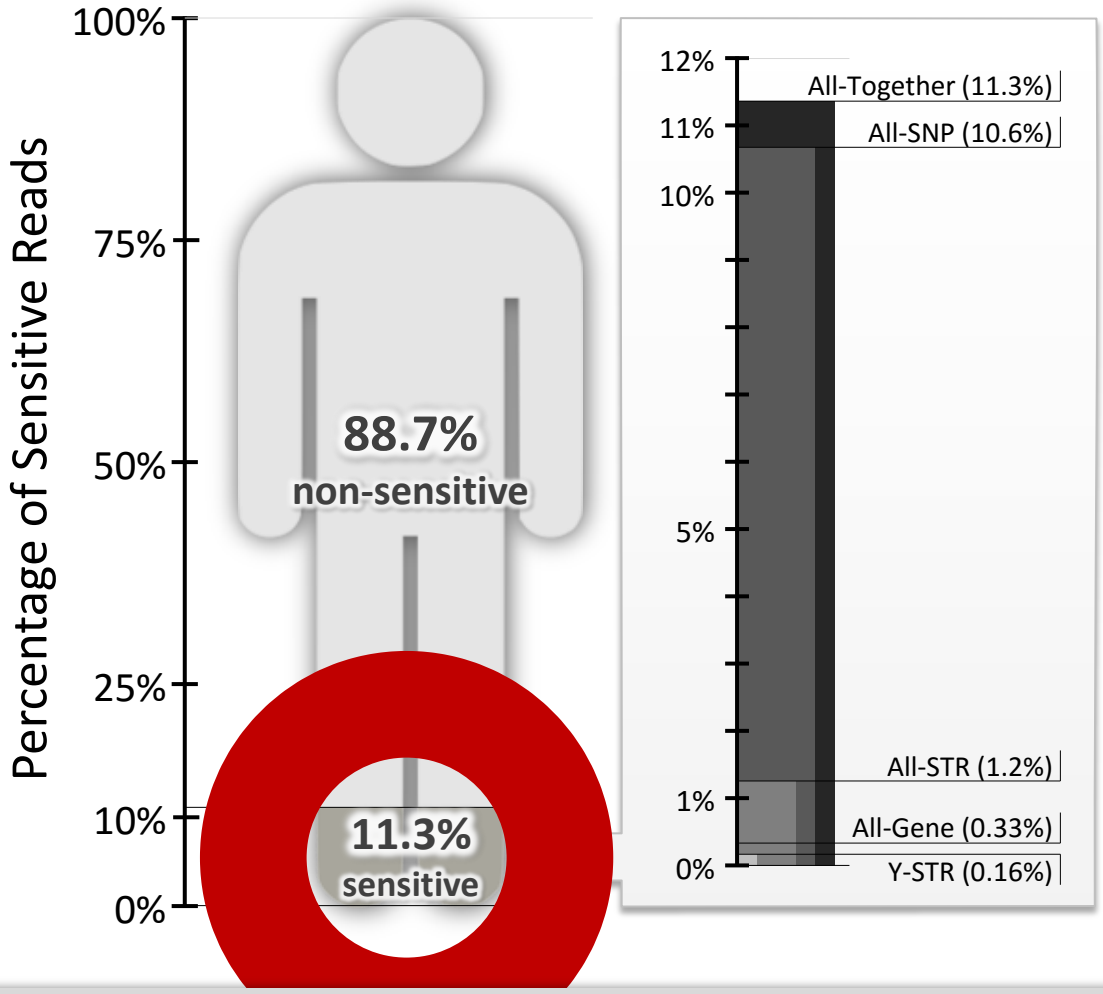
Percentage of privacy-sensitive sequences



5. Evaluation

Reduces the size of the problem **almost**

one order of magnitude



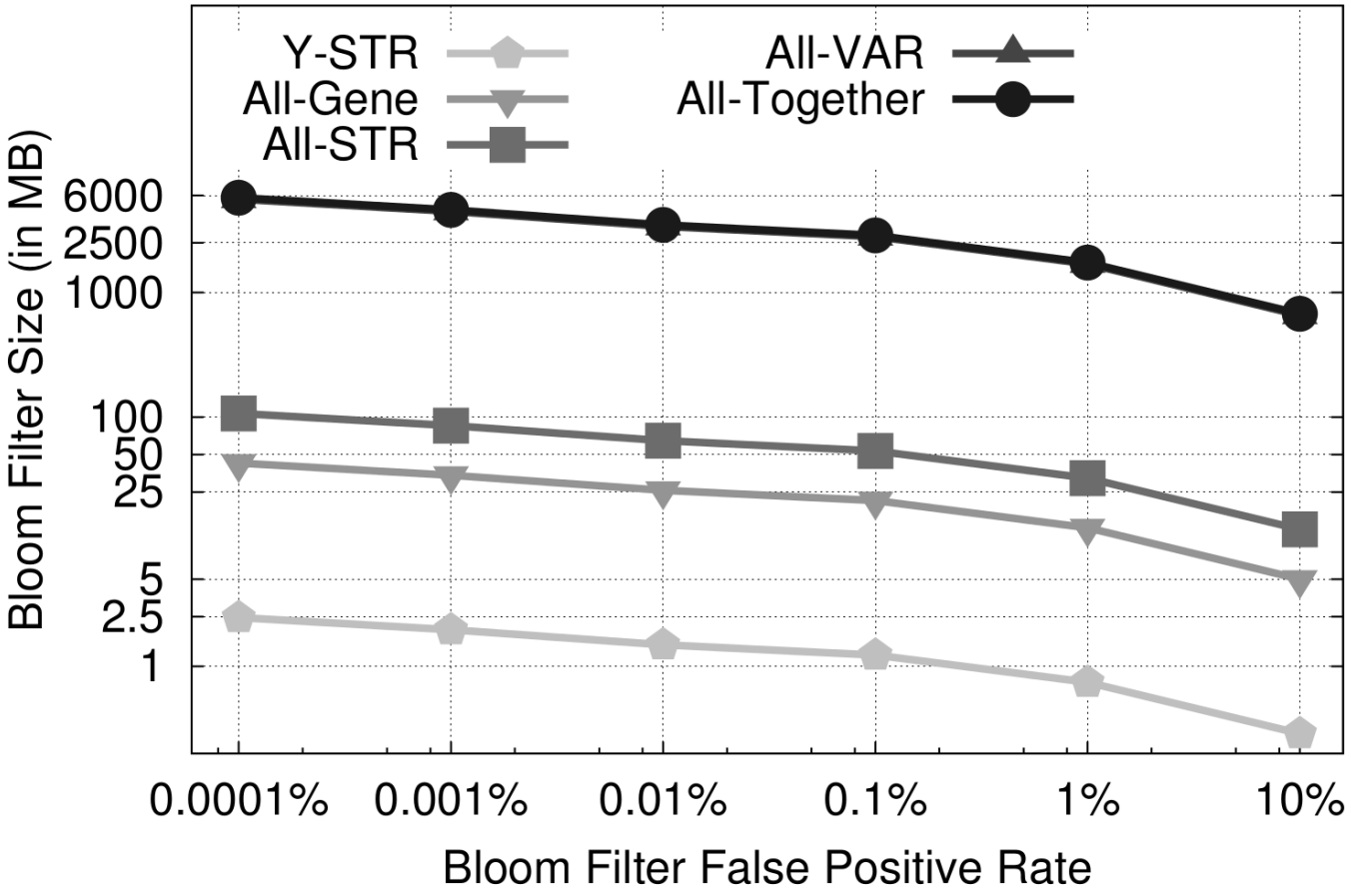
5. Evaluation

2nd experiment

How big is the **Bloom filter**?

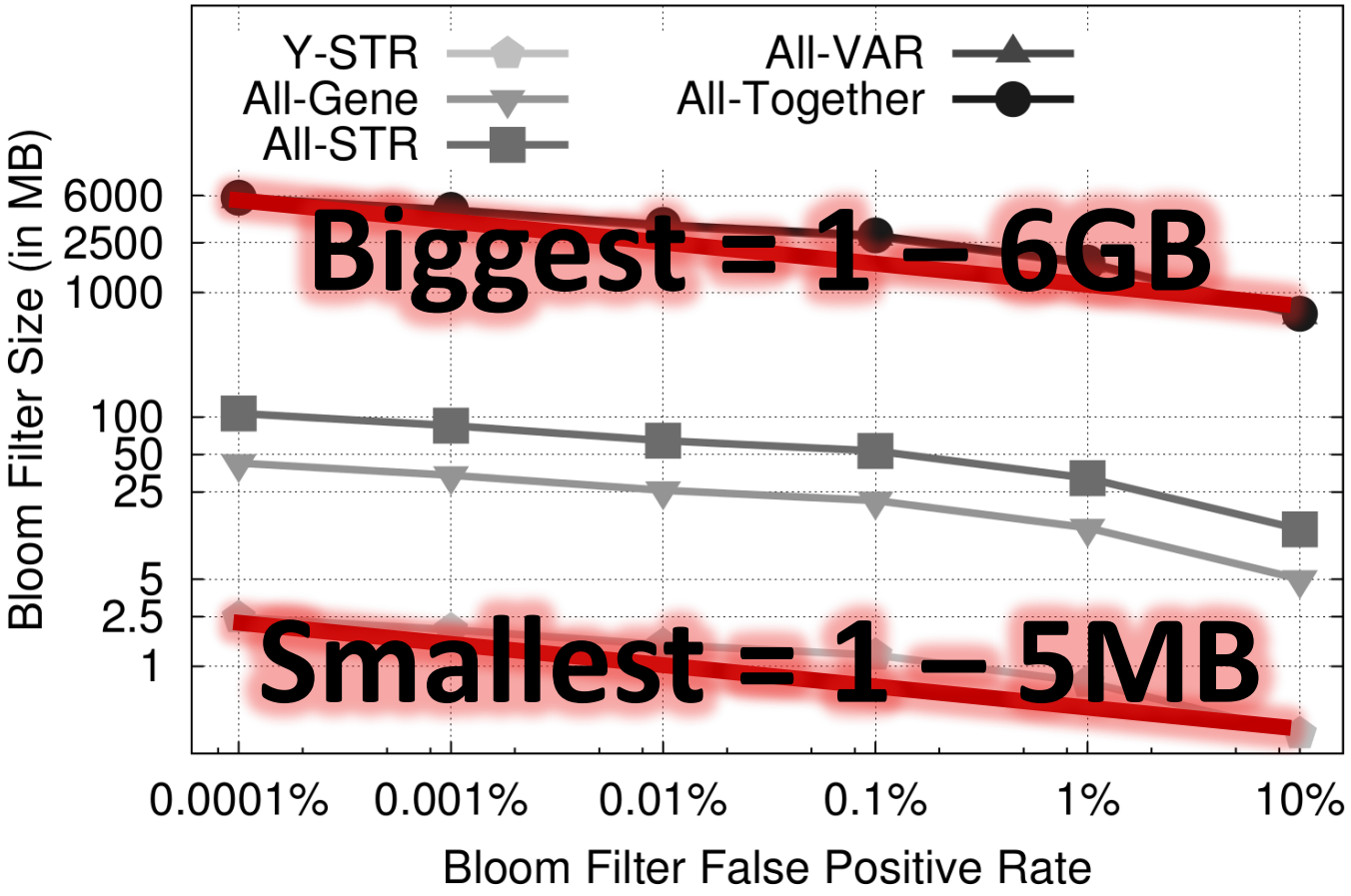
5. Evaluation

Bloom filter size



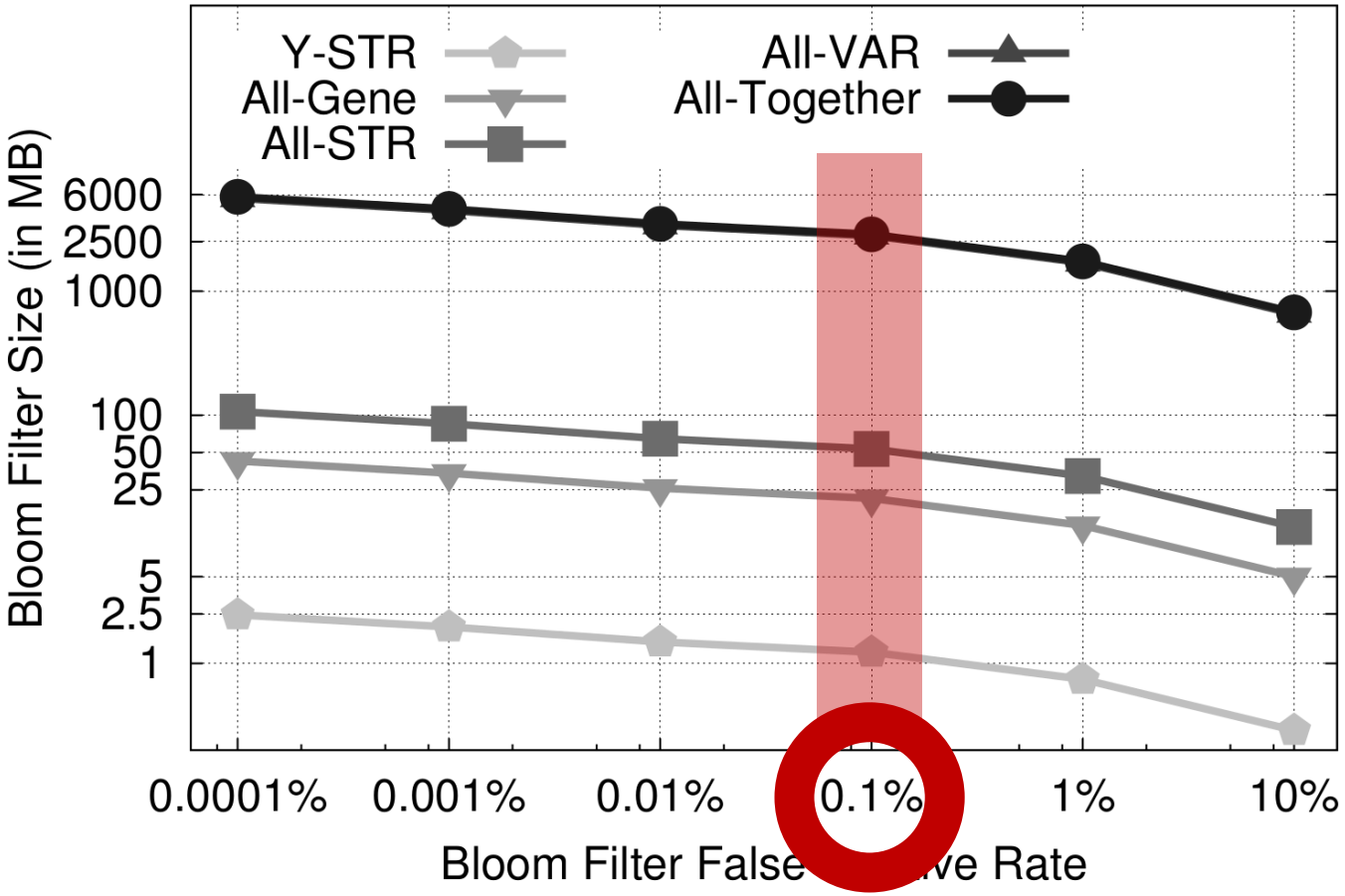
5. Evaluation

↑ size of the knowledge database
↑ size of the Bloom filter



5. Evaluation

Bloom filter size



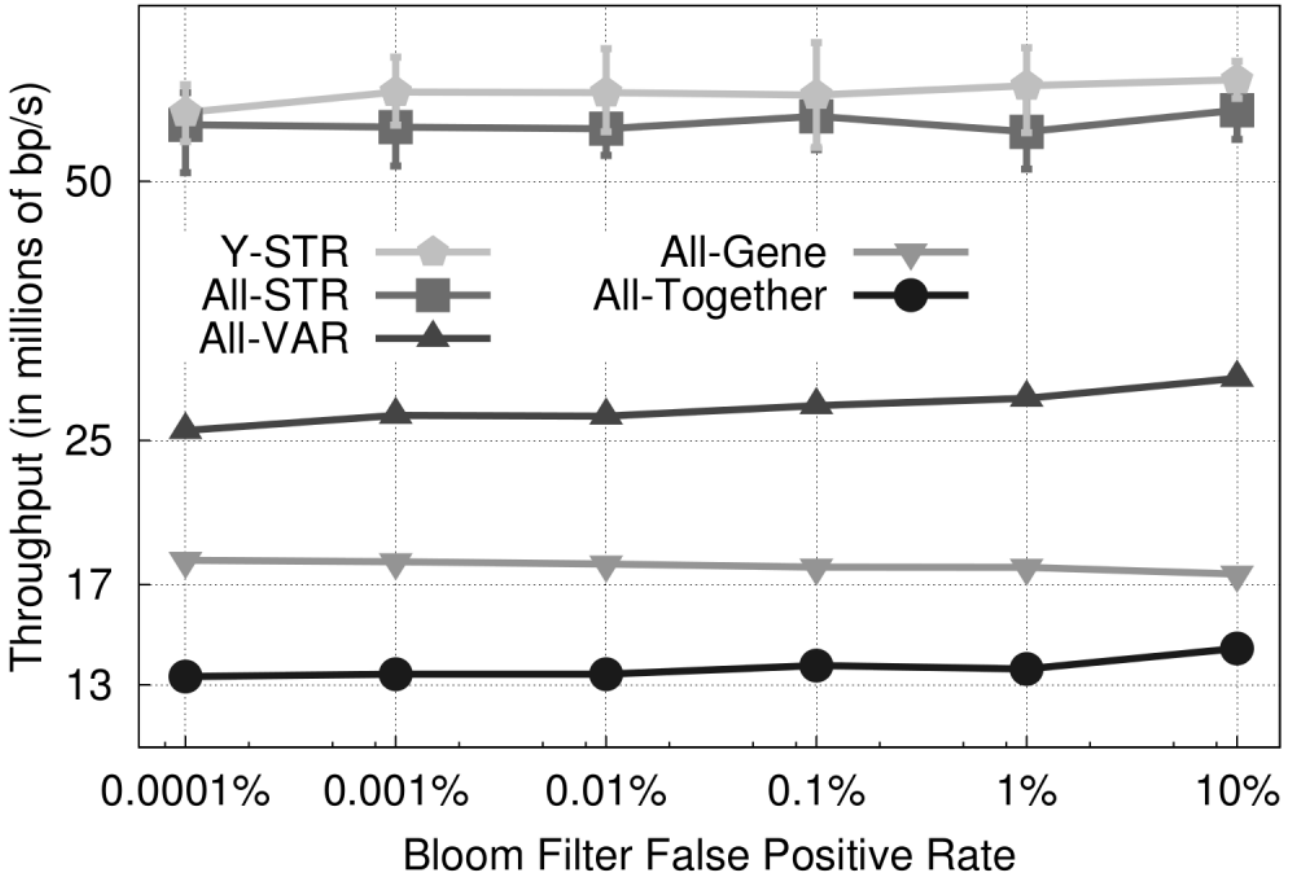
5. Evaluation

3rd experiment

Is the detector a bottleneck?

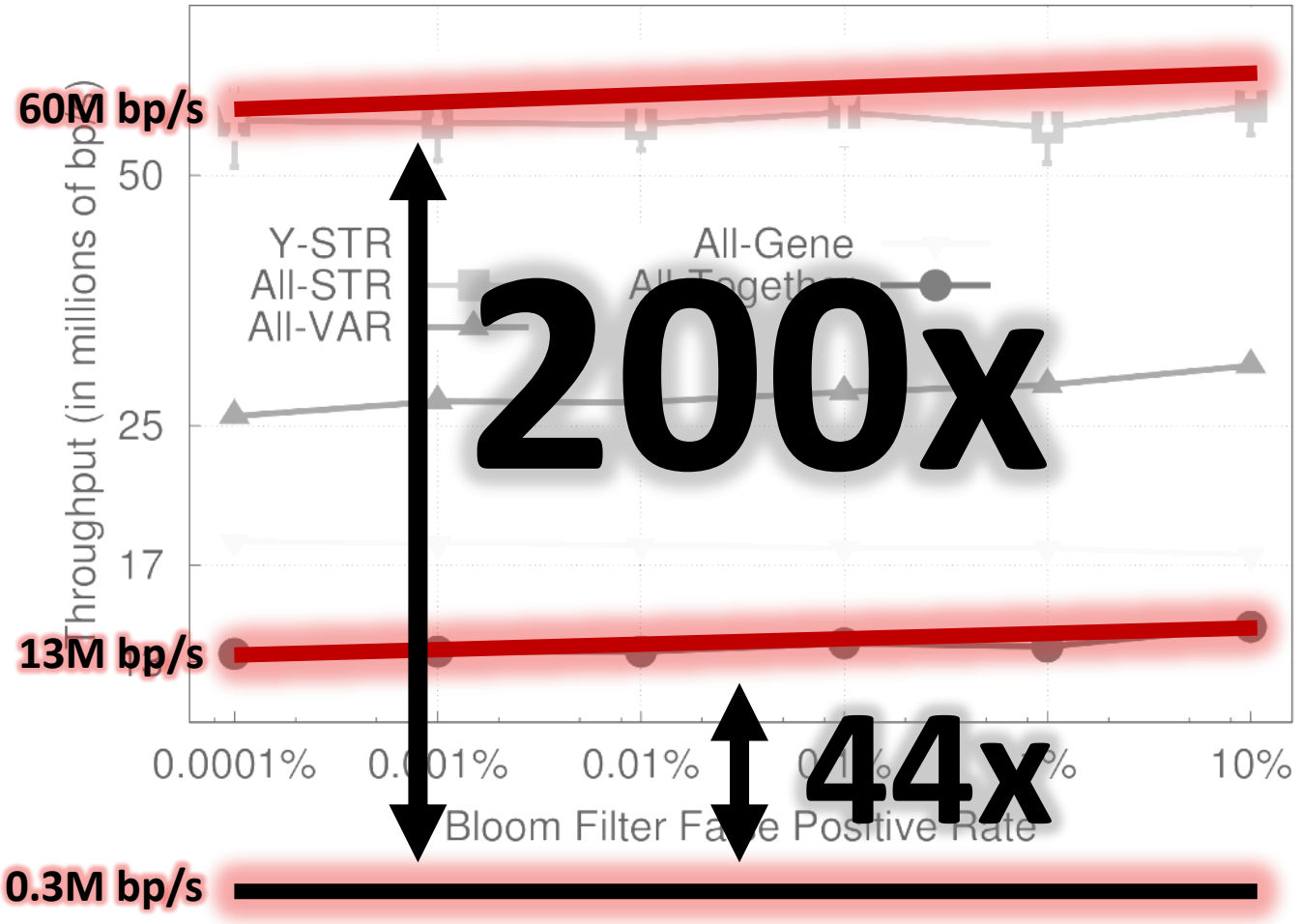
5. Evaluation

Throughput – Single core



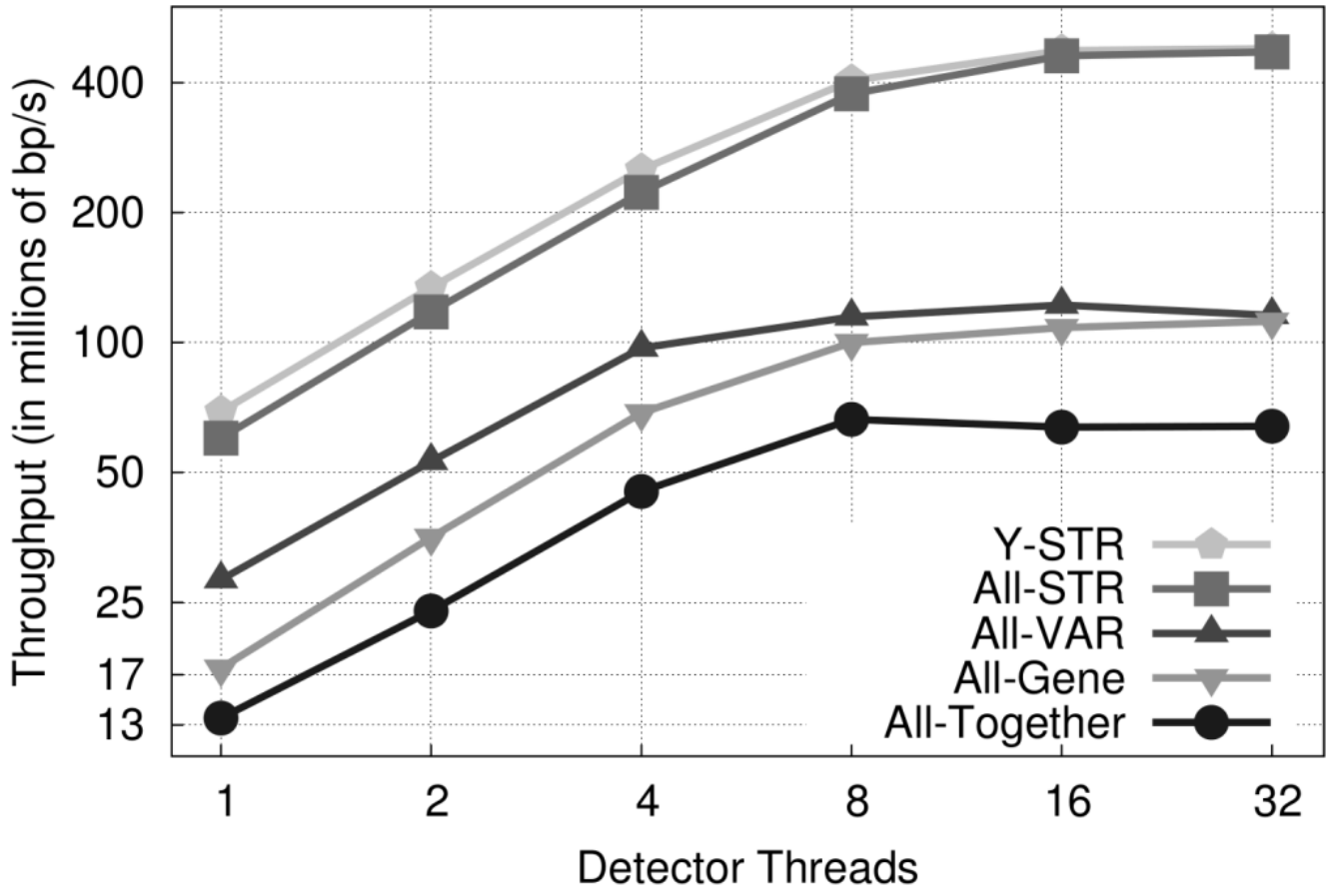
5. Evaluation

Throughput – Single core



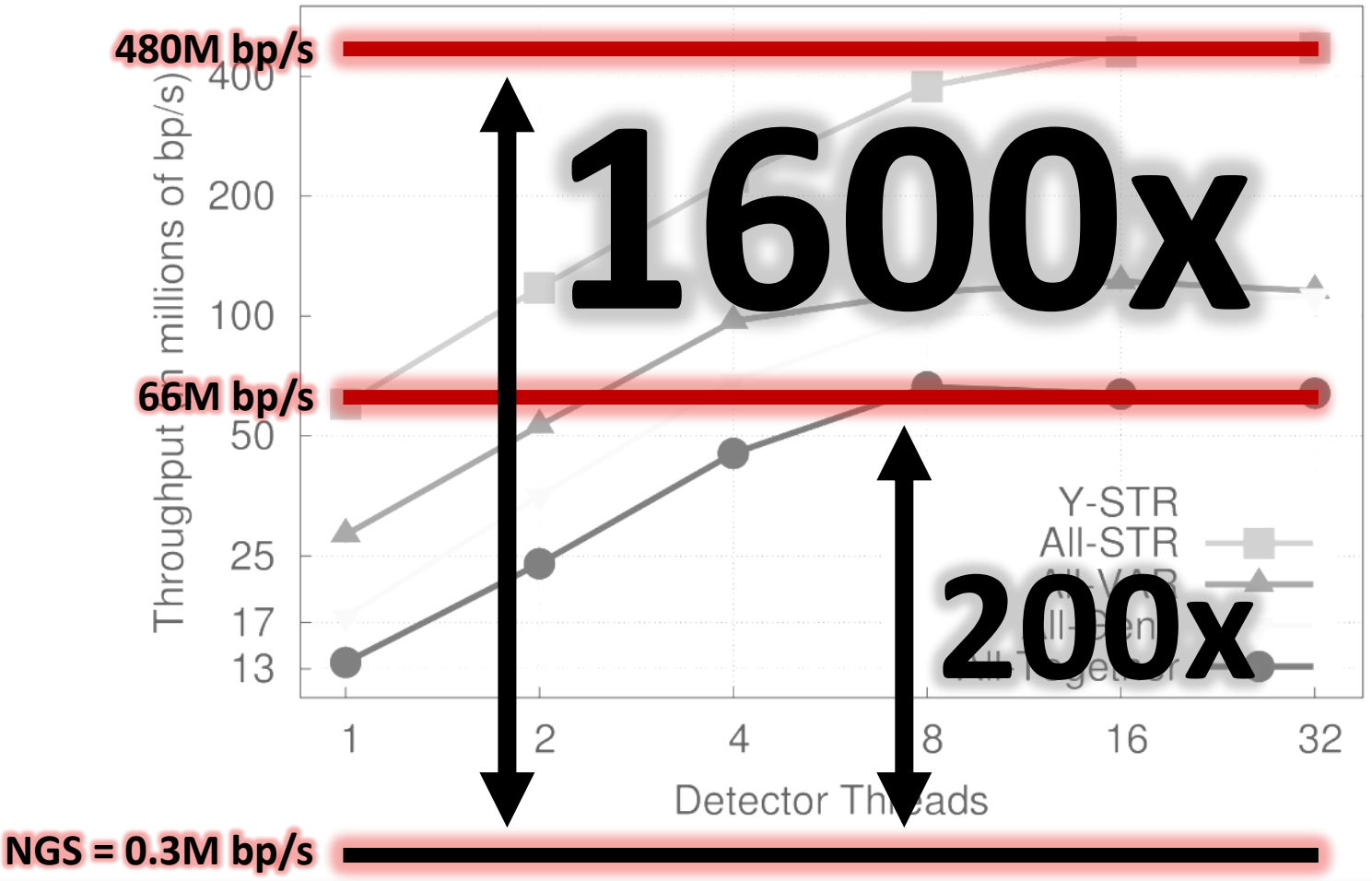
5. Evaluation

Throughput – Multi-core



5. Evaluation

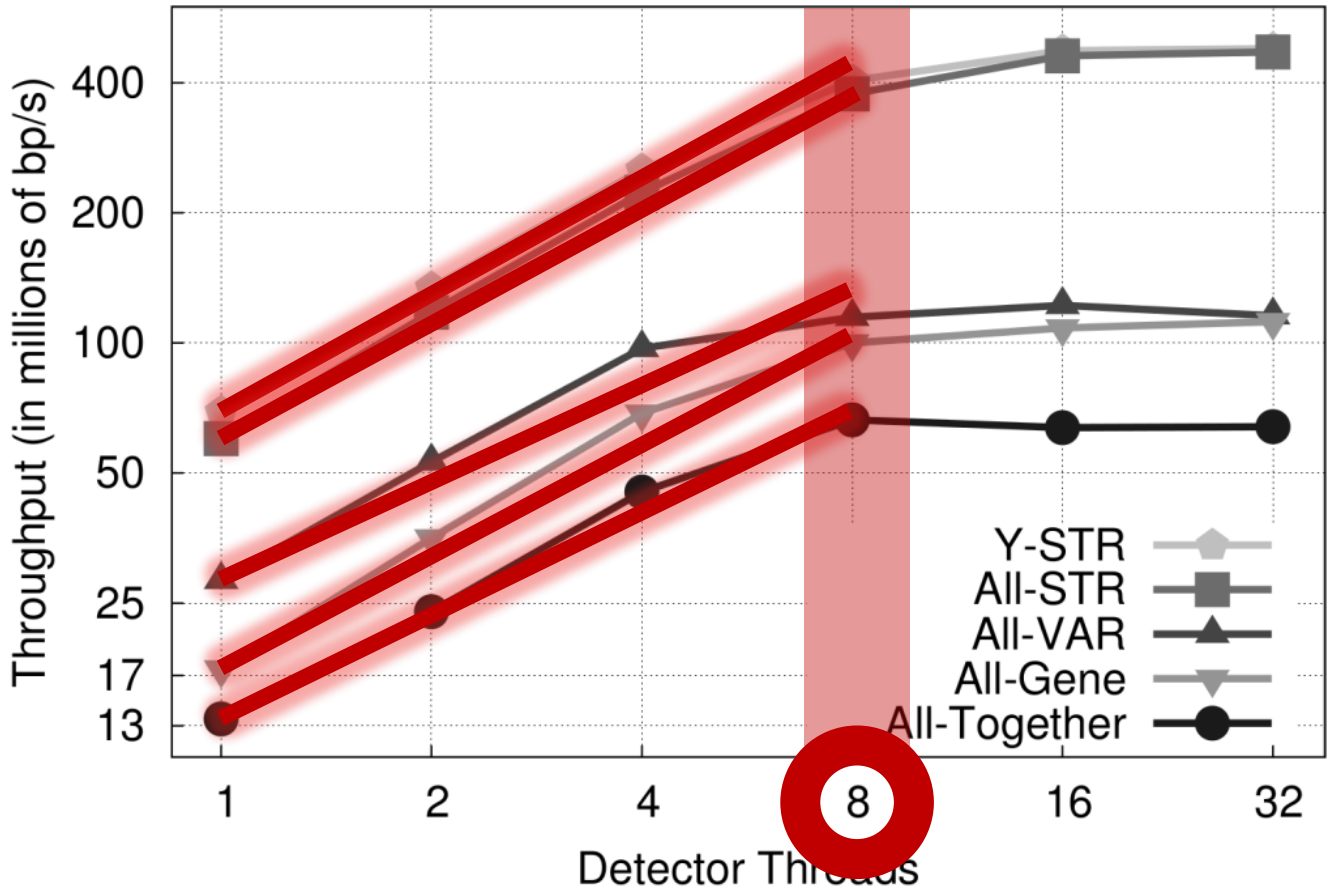
Throughput – Multi-core



5. Evaluation

Efficiency limit = **8 threads**

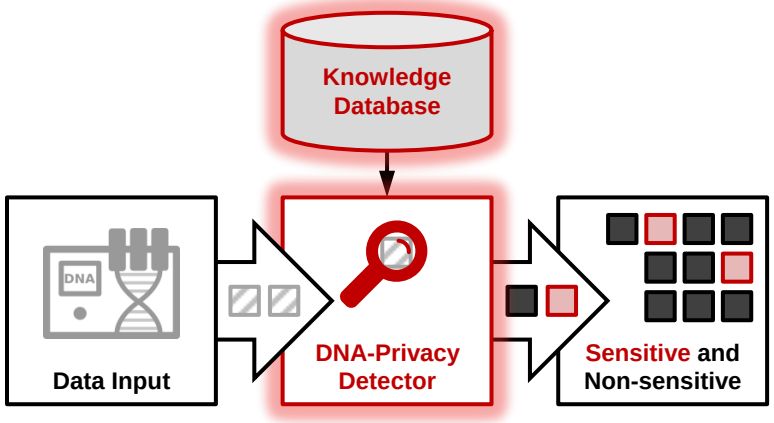
(Is it? 8-core machine)



6. Final remarks

Conclusions

DNA-Privacy Detector



There are **privacy-sensitive portions** in human genomes

- **Smallest** portion of human genomes (<12%)
- **Can be identified**
- Detection is **fast**
- **Incremental** knowledge

Open-source code:

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

6. Final remarks

Following up

Submitted a new paper in Feb'17:

"**How can photo sharing inspire sharing genomes?**"

- ✓ Some portions of data are more privacy-sensitive than others
- ✓ One's data may affect the privacy of others
- ✓ Systematically detecting the privacy-sensitive portions of data is feasible
- ✓ After classifying the portions, decide how to share them
- ✓ The impact of data sharing is unpredictable

6. Final remarks

Following up

- Intend to **validate** the efficacy of the method **against IMPUTE2** and **lobSTR**
- BioSys-PhD (FCUL):
  - ❑ **Francisco Couto**, Margarida Gama-Carvalho, Mariana Pinhão
    - What are the **determinants** of human genetic **individuality**?
    - **Biological view** from privacy-sensitive sequences
- University of Luxembourg:
  - ❑ **Paulo Verissimo**, Jeremie Decouchant, Maria Fernandes
    - **Risk assessment** considering different levels of criticality in genomes
    - Adapting our method for **long NGS reads**
    - Exploring **updated** and **other databases**

6. Final remarks

Publications

Direct from this PhD:

- ✓ 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.

- ✓ 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):1,11, Jan. 2016.

- ✓ V. V. Cogo and A. Bessani.
  Efficient Storage of Whole Human Genomes.
  Poster in the EuroSys 2016, Poster Presentation Award, Apr. 2016.

6. Final remarks

Publications

Related to this PhD:

- 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.

- 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.

6. Final remarks

Publications

Submitted:

- 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.
  Accepted with major revisions on **ACM Transactions on Storage**, Oct. 2016.

- V. V. Cogo, A. Bessani, F. M. Couto, M. Carvalho, M. Fernandes, and P. Verissimo.
  How can photo sharing inspire sharing genomes?
  Submitted to PACBB'17 (11th Int. Conf. on Practical Applications of Comp. Biology & Bioinformatics), Feb. 2017.

Ciências
ULisboa

LISBOA

UNIVERSIDADE
DE LISBOA

Thank you!

Extra

Completeness of the 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