WPES 2015. October 12, 2015. Denver, US.

A High-Throughput Method to Detect Privacy-Sensitive Human Genomic Data

Vinicius V. Cogo[1]
Alysson Bessani[1]
Francisco M. Couto[1]
Paulo Verissimo[2]

[1]LaSIGE, University of Lisbon

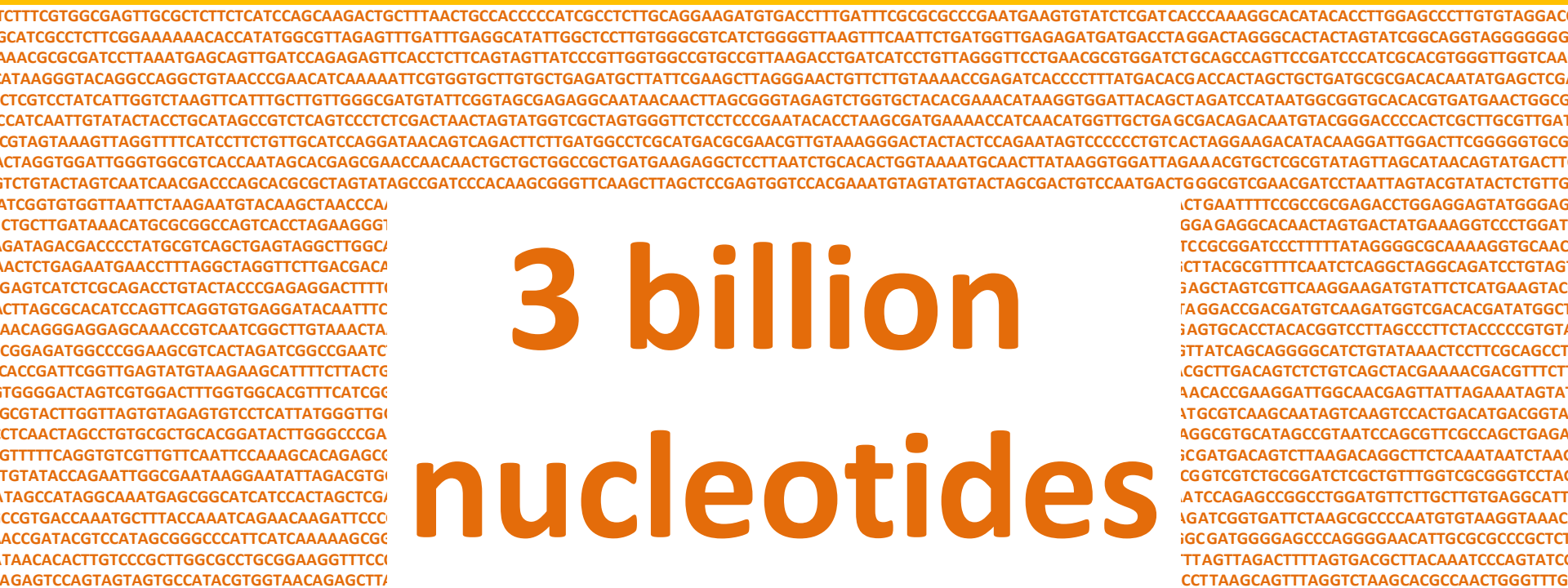
[2]SnT, University of Luxembourg

1. Context

DNA

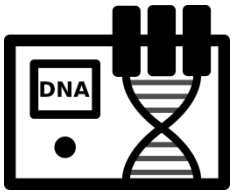
1. Context

Human genome

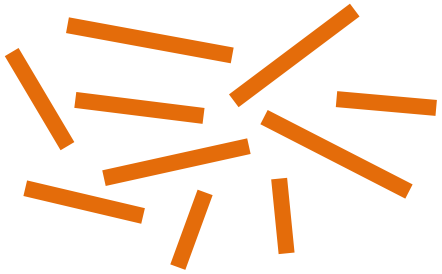


1. Context

Genomic data



Sequencing



FASTQ
(unaligned)

Alignment

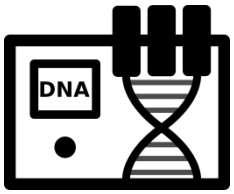
SAM
(aligned)

Assembly

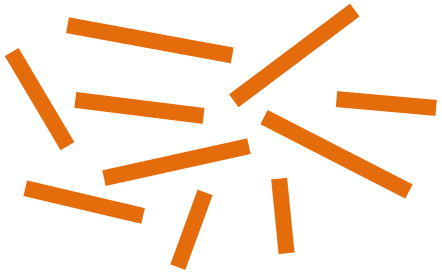
FASTA
(contiguous)

1. Context

Raw genomic data



Sequencing



FASTQ

Alignment

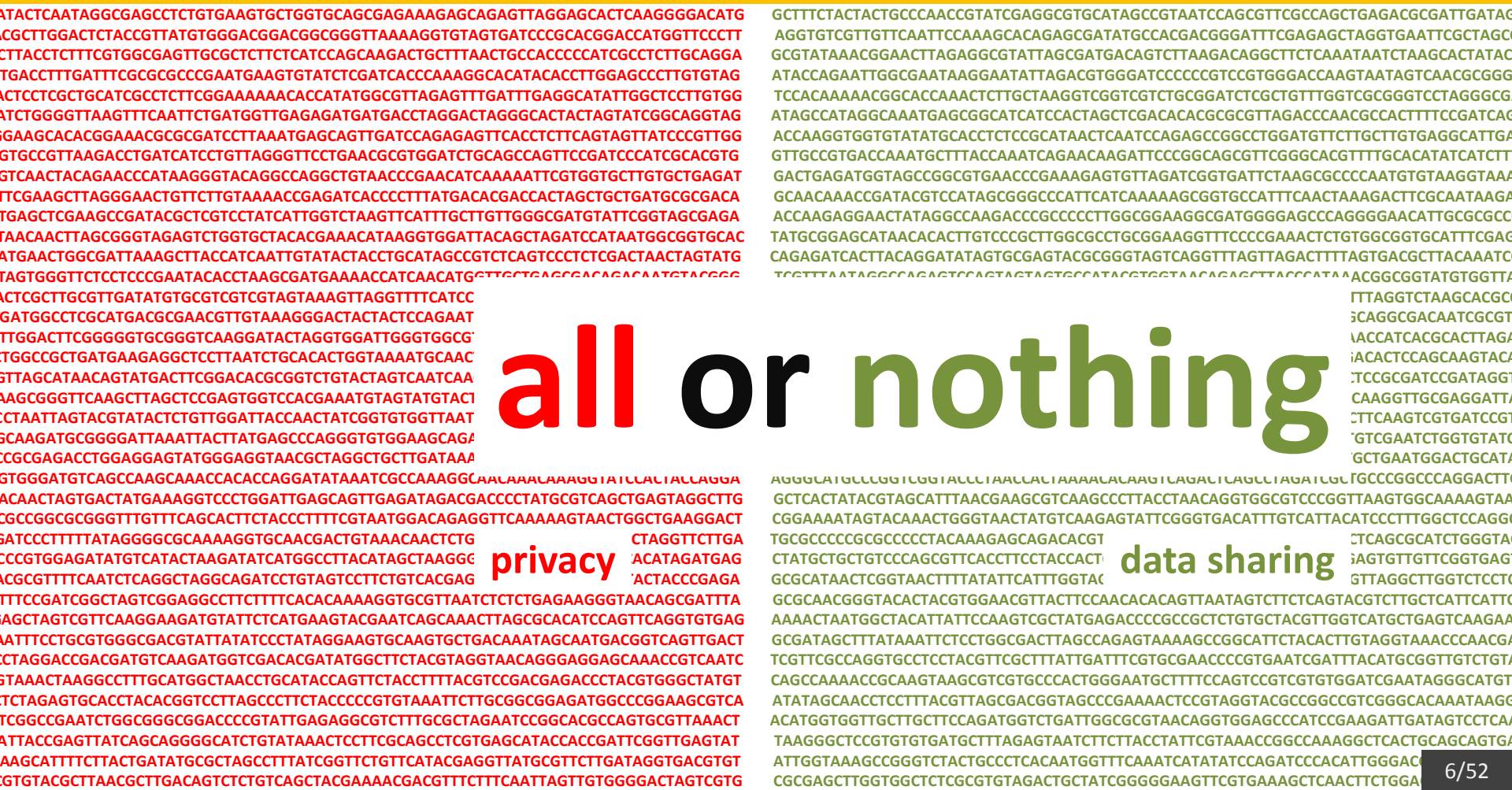
SAM

Assembly

FASTA

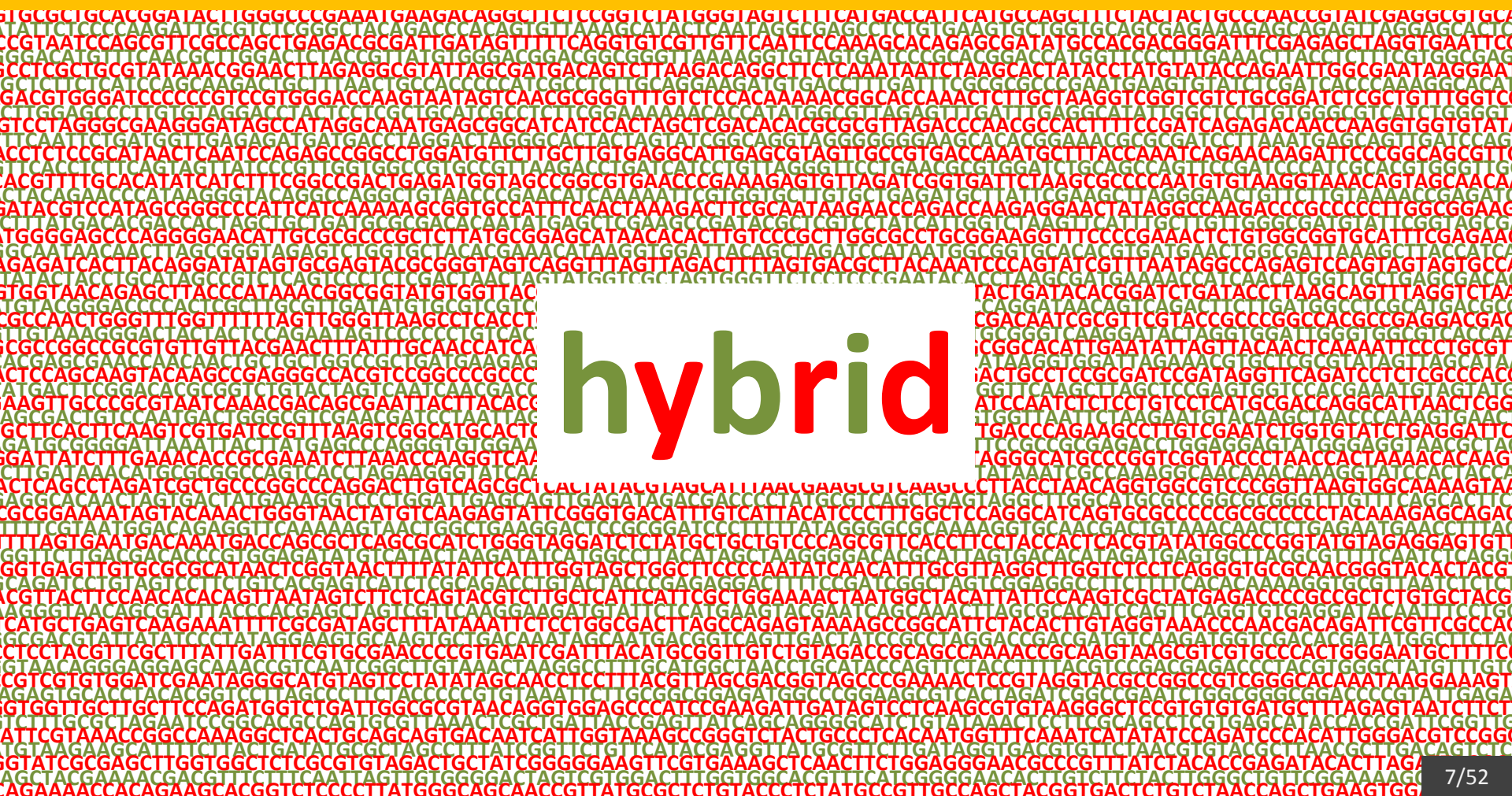
1. Context

Privacy in human genome today



1. Context

Privacy in human genome

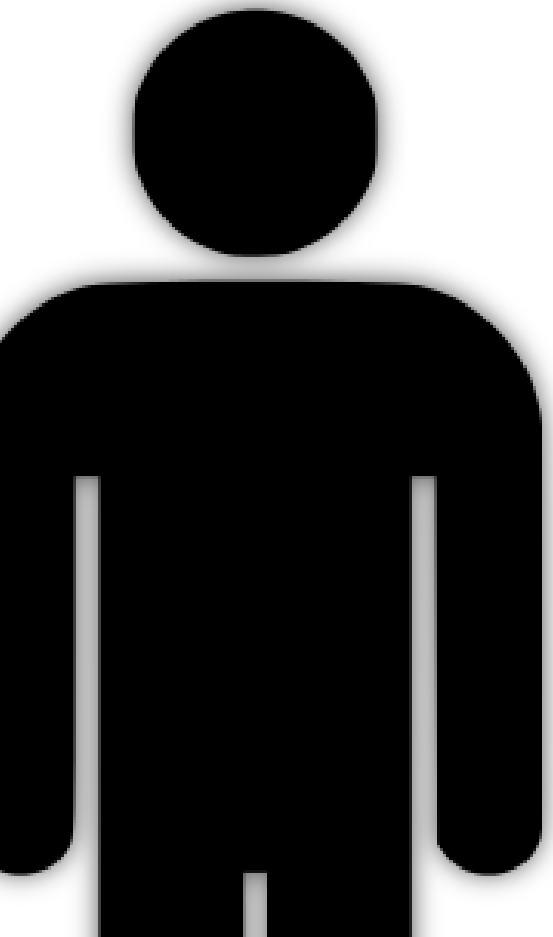


1. Context

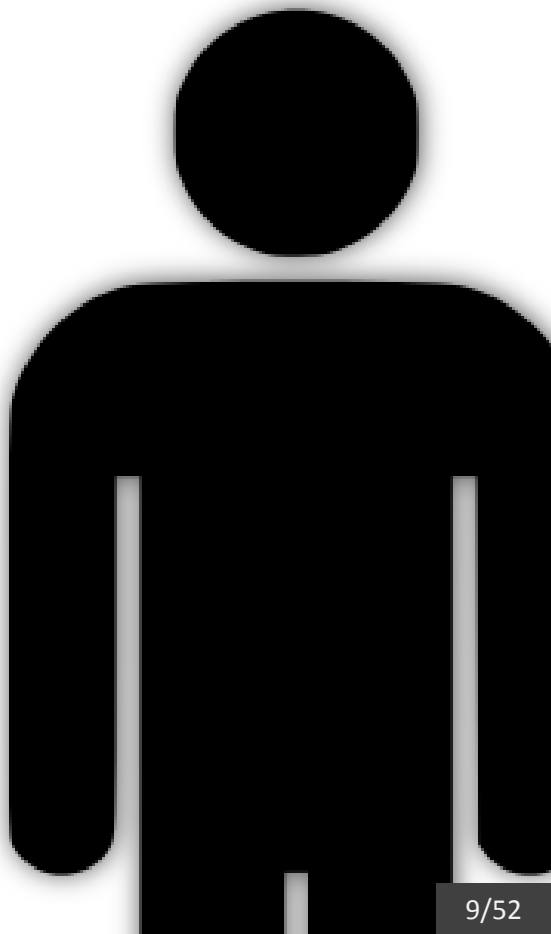
Are all portions equally important for privacy?

1. Context

1 Genome similarity in humans…



>99.5%



1. Context

1

What makes us unique is…

<0.5%

1. Context

2

genes → proteins → function

1. Context

Detect the relevant portions for privacy

1. Context

Concentrate our efforts

most effort

TCCTTAAATGAGC
ACCTCTTCAGTAG
GCCGTTAAGACCT
CTGAACGCGTGGA
CCCATCGCACGTG
GAGGATACAATTT
TAGGACCGACGAT
TGTCCCGCTTGGC
GAAACTCTGTGGC
ACGCGGGTAGTCA
TGACGCTTACAAA

affordable effort

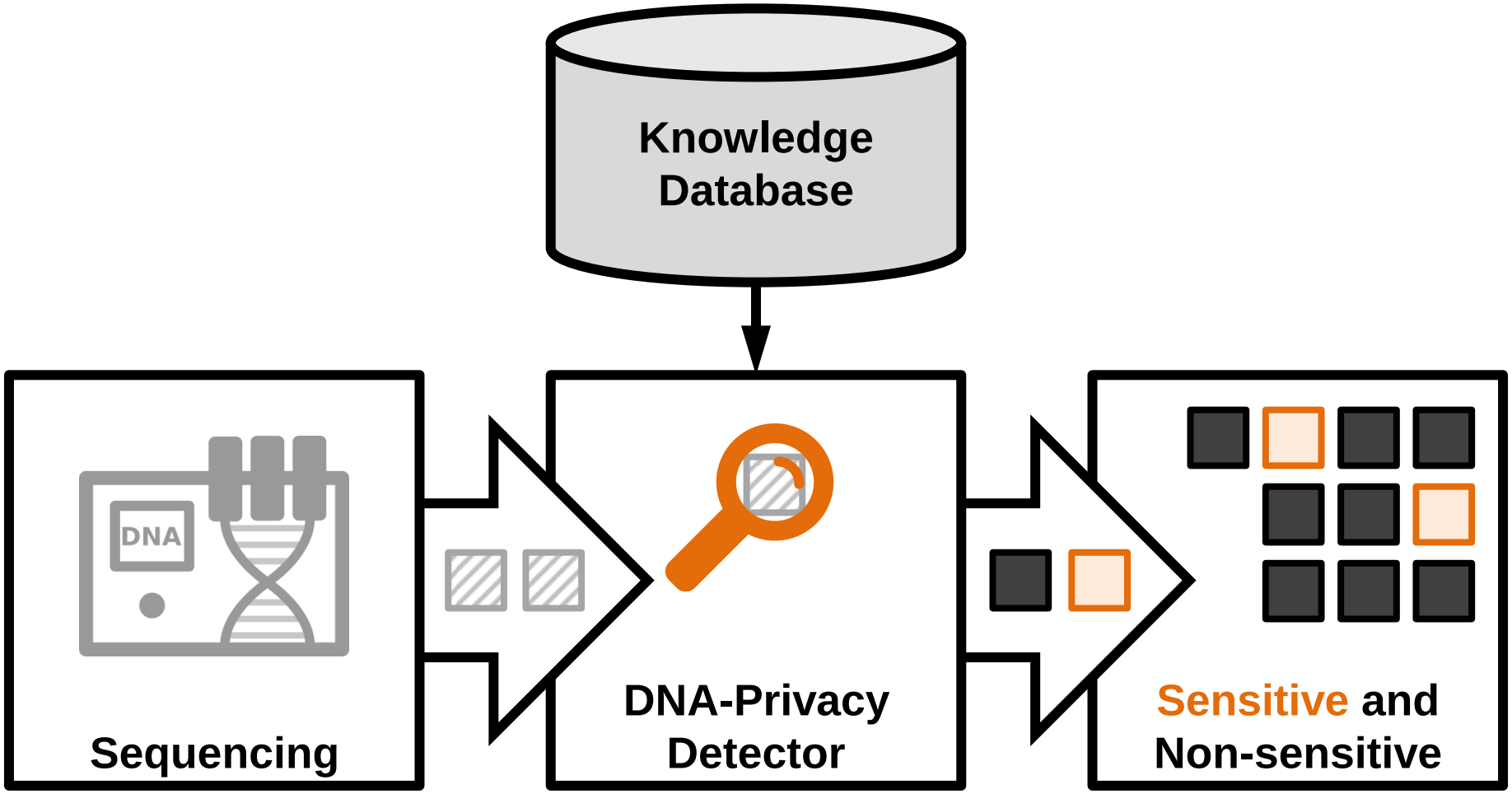
1. Context

Unknown privacy-sensitive sequences

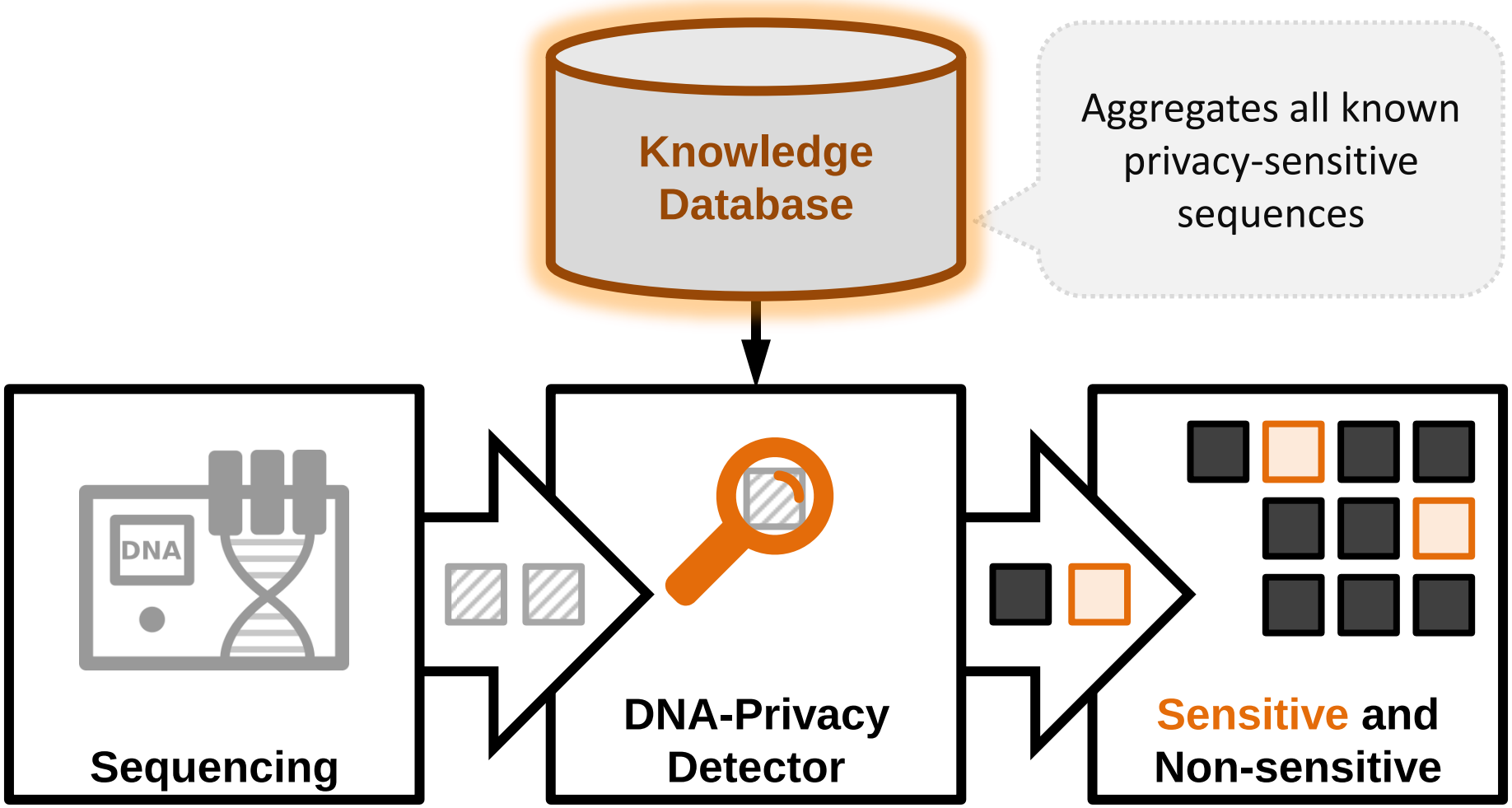
most effort

TCCTTAAATGAGC
ACCTCTTCAGTAG
GCCGTTAAGACCT
CTGAACGCGTGGA
CCCATCGCACGTG
GAGGATACAATTT
TAGGACCGACGAT
TGTCCCGCTTGGC
GAAACTCTGTGGC
ACGCGGGTAGTCA
TGACGCTTACAAA

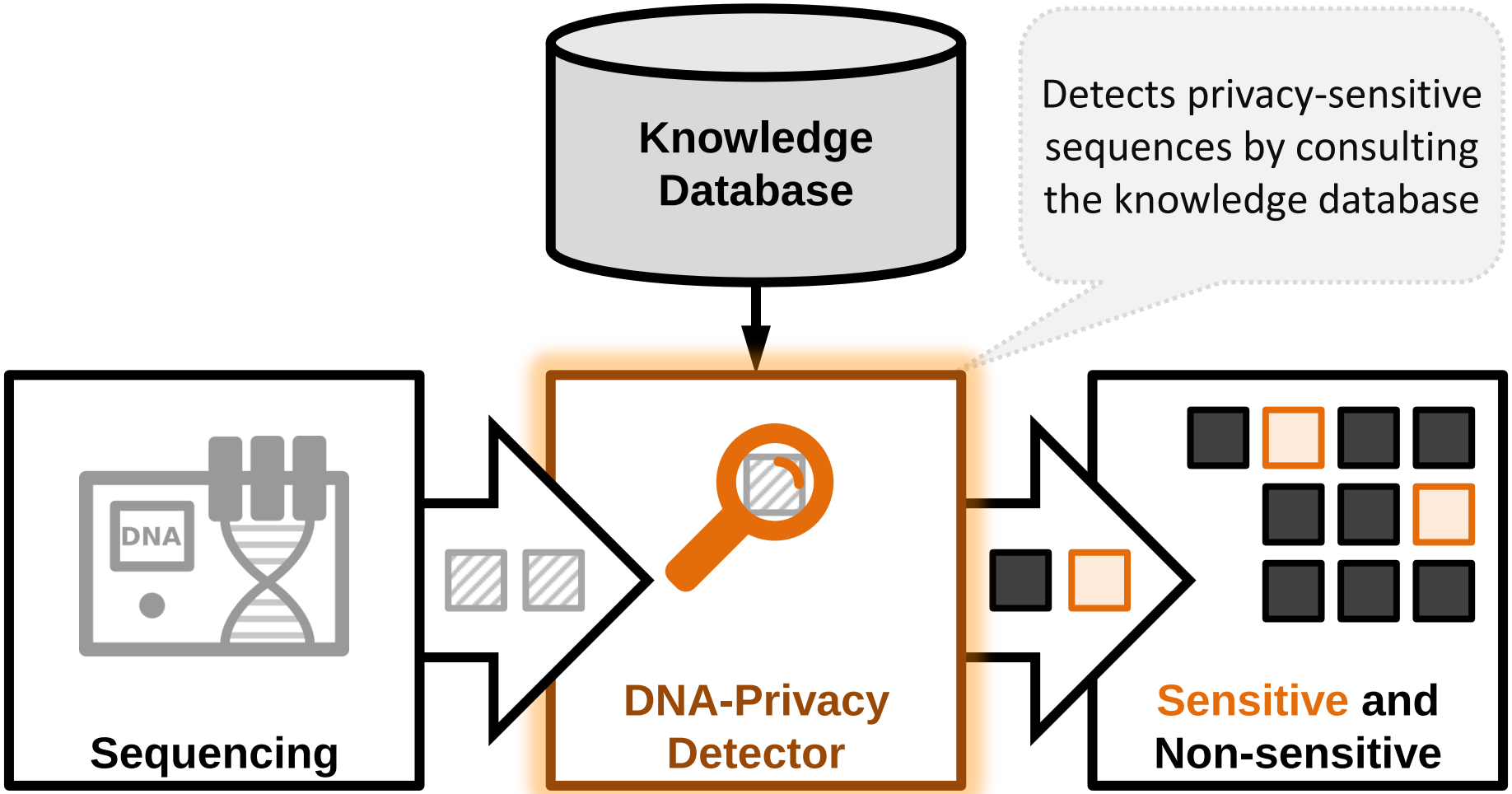
2. Overview


Knowledge Database
DNA
Sequencing
DNA-Privacy Detector
Sensitive and Non-sensitive

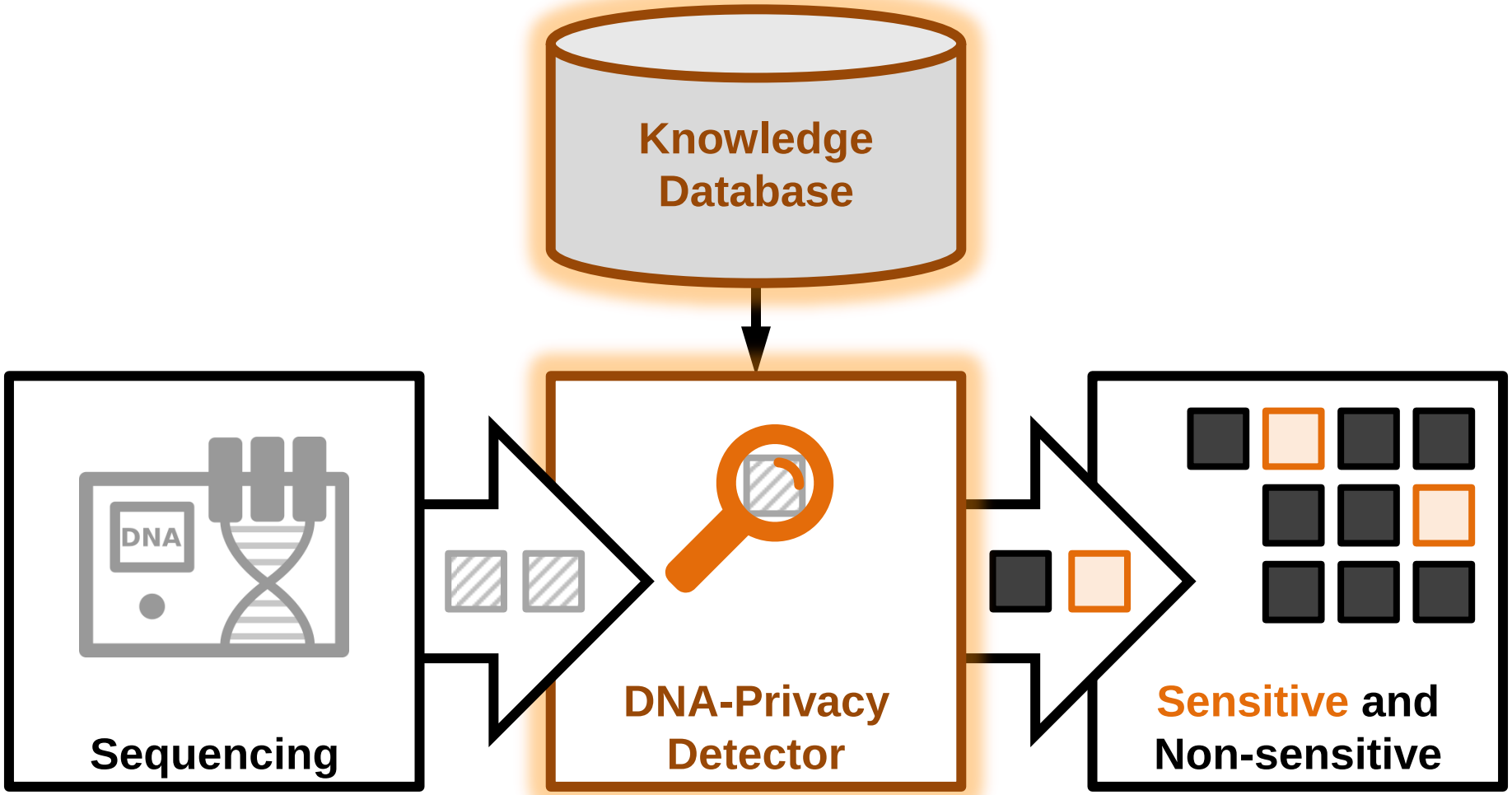
2. Overview


Knowledge Database
Aggregates all known privacy-sensitive sequences
DNA
Sequencing
DNA-Privacy Detector
Sensitive and Non-sensitive

2. Overview


Knowledge Database
Detects privacy-sensitive sequences by consulting the knowledge database
DNA
Sequencing
DNA-Privacy Detector
Sensitive and Non-sensitive

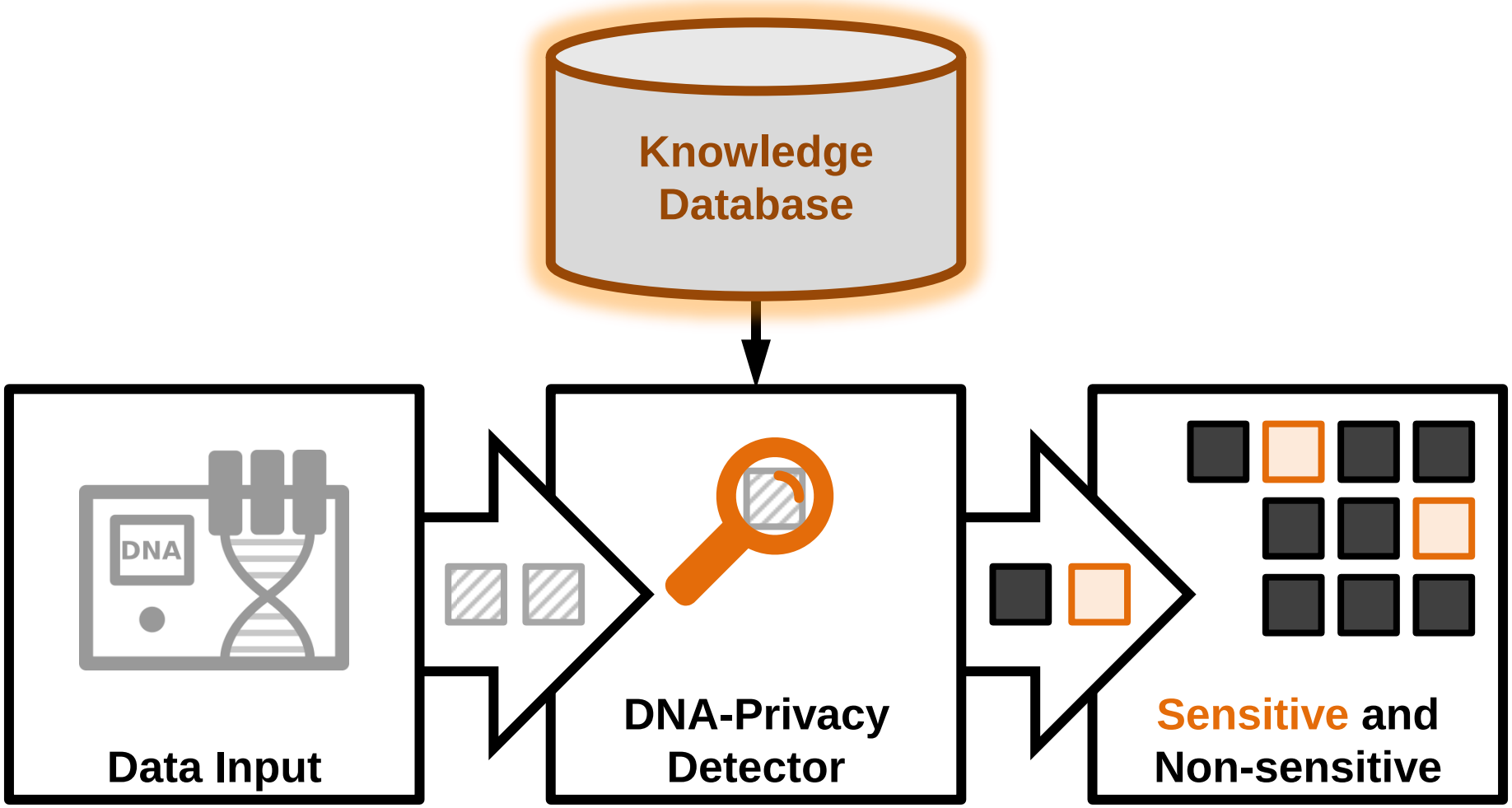
2. Overview


Knowledge Database
DNA
Sequencing
DNA-Privacy Detector
Sensitive and Non-sensitive

Outline

3. Creation of knowledge databases
4. Implementing the detector
5. Evaluation
6. Analysis
7. Conclusions

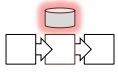
3. Creating knowledge databases



3. Creating knowledge databases

Overview

- Privacy-sensitive sequences
  - disclose private information about donors
- Attacks (References 6, 13, 16, 18, 23, 28, 34, 35)
- 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

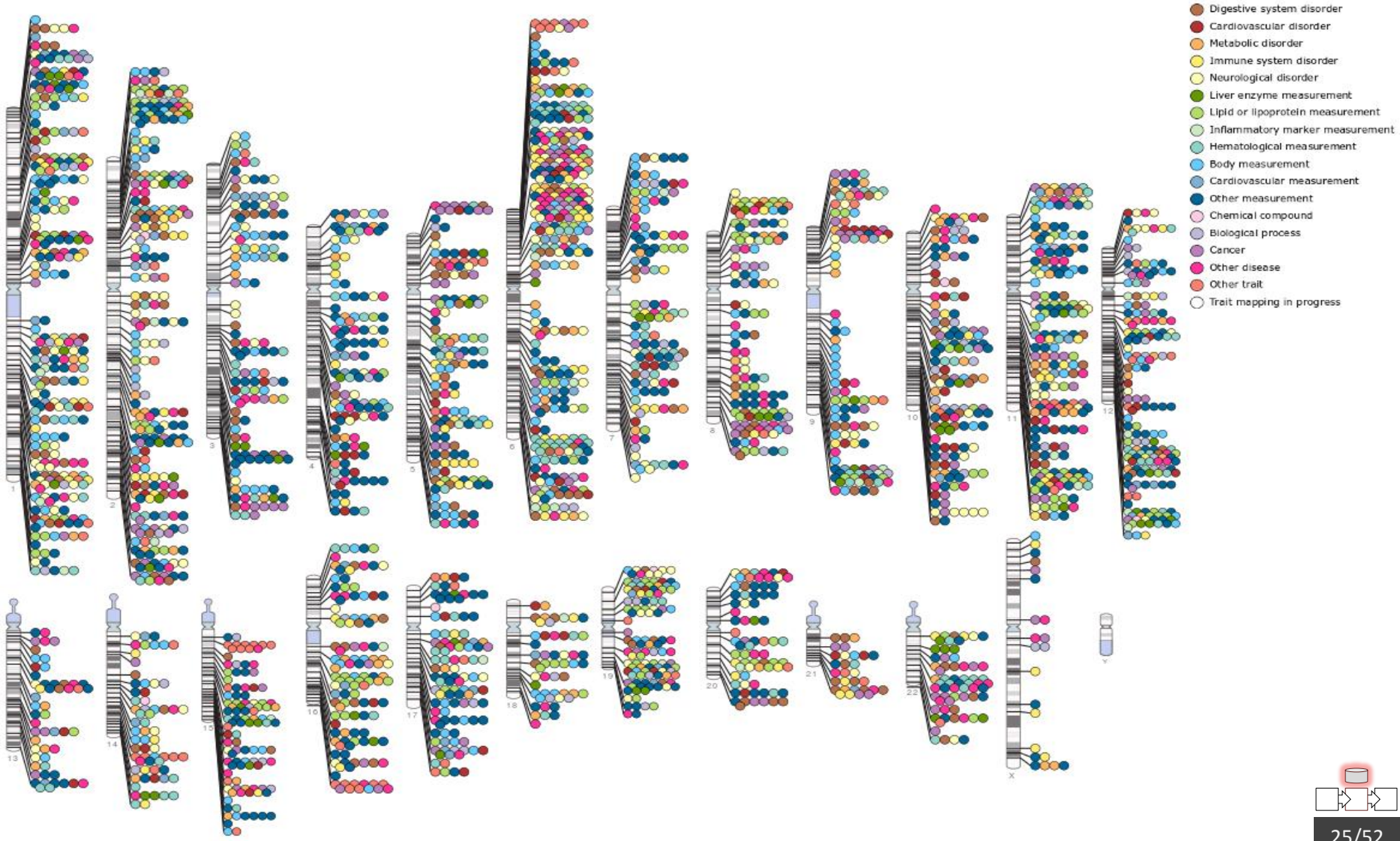
3. Creating knowledge databases

Genomic variations

<0.5%

3. Creating knowledge databases

Genomic variations



3. Creating knowledge databases

Genomic variations

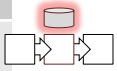
3. Creating knowledge databases

Known attacks

Short tandem repeats	
Attacks	**References**
Use STR profiles to identify donors of 1000 Genomes Project	Gymrek [16]
Forensic identification	Butler [6]

Disease-related genes	
Attacks	**References**
Direct-to-consumer genomic testing	Goldsmith [13]
Masking the APOE gene (related to Alzheimer) from Dr. Watson's genome	Wheeler [35]

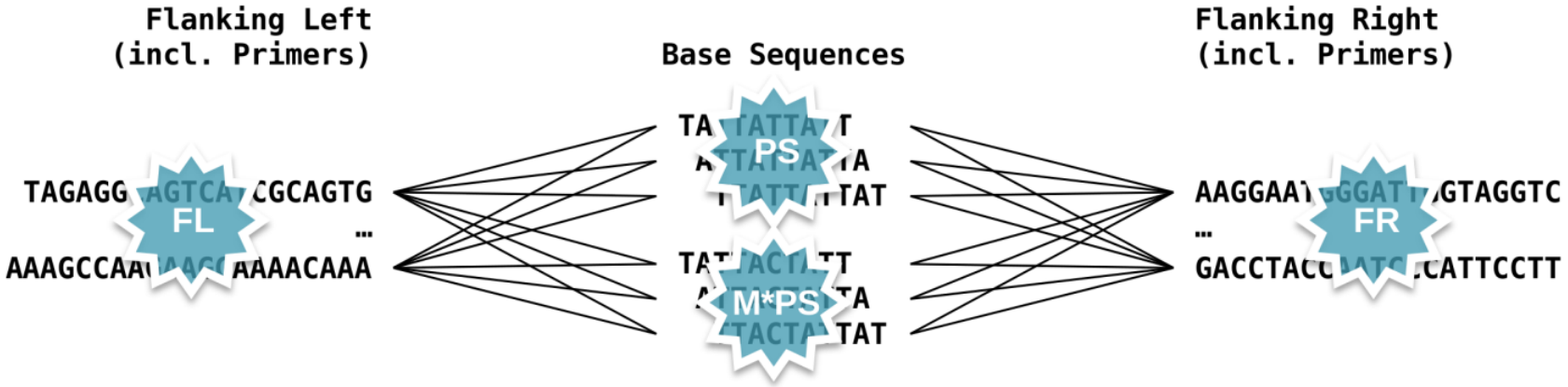
Genomic variations	
Attacks	**References**
Re-identification (few hundred SNPs are enough)	Lin [23]
Acquire knowledge about targets from GWAS results	Wang [34]
Acquire knowledge about targets from microarray results	Homer [18]
Infer masked genes (e.g., the APOE gene from Dr. Watson [35])	Nyholt [28]



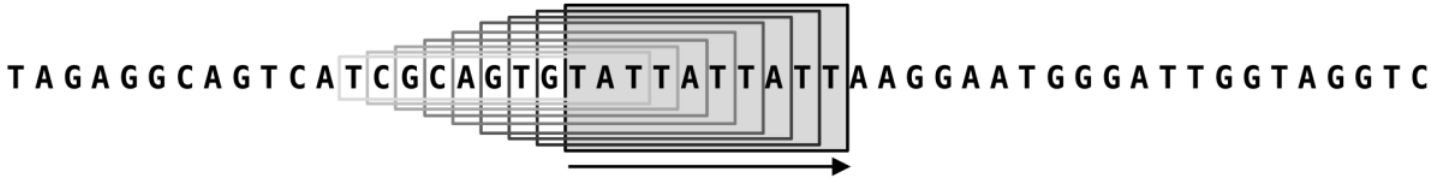
3. Creating knowledge databases

Example: obtaining the STRs

1) Creating the long sequences



2) Sliding the window with size S through each long sequence



3) Creating the blacklist sequences from each position in the long sequence

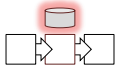
```
TAGAGGCAGT
 AGAGGCAGTC
  GAGGCAGTCA
   ...
          CATCGCAGTG
           ATCGCAGTGT
            TCGCAGTGTA
             ...
```

3. Creating knowledge databases

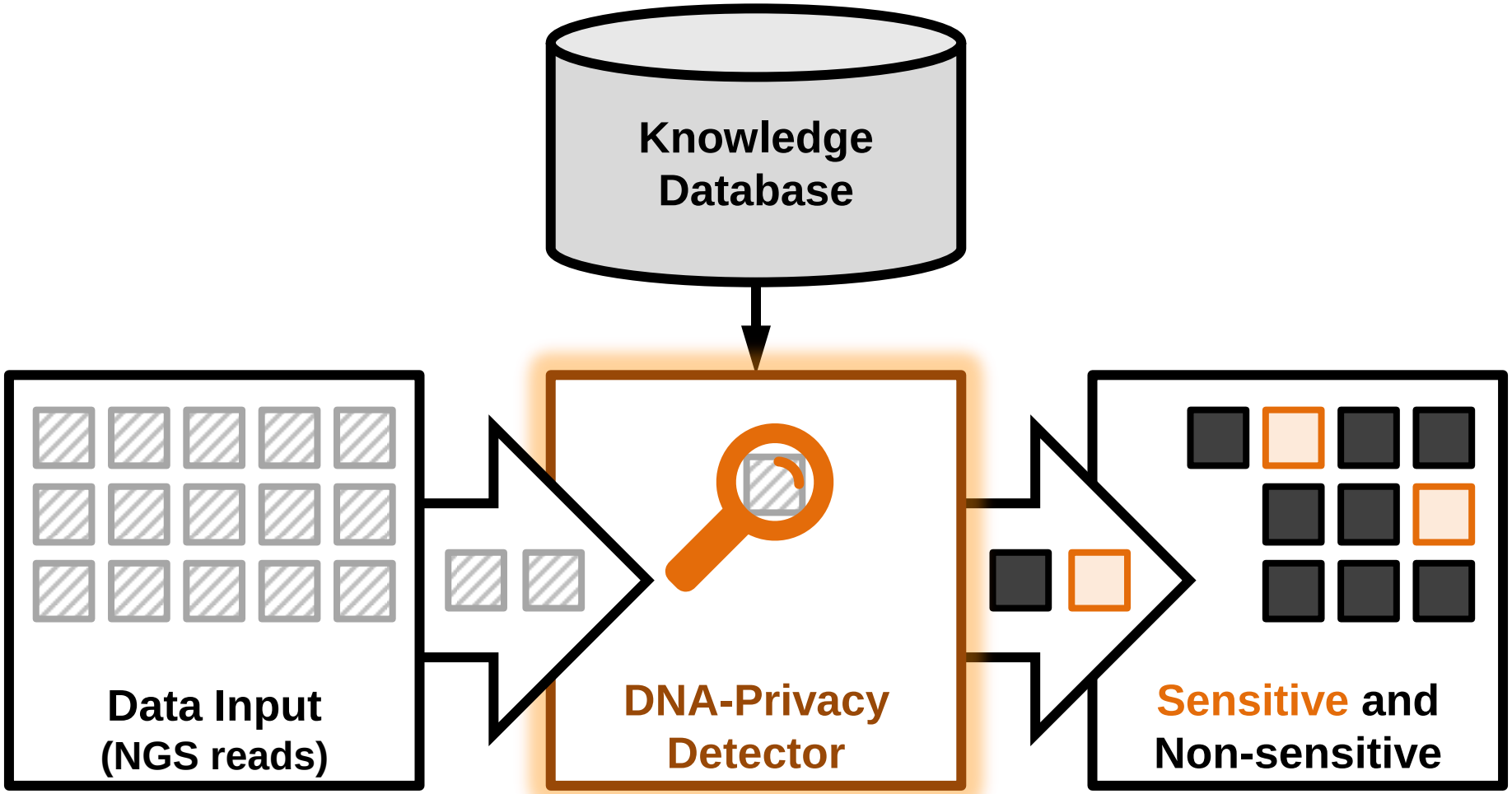
Obtaining the sequences

	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

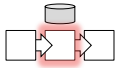
No label or pattern available (SocialSecurityNumber=123-45-6789)

- Fetch and lookup large amounts of data

Tens of GB

- Support a high throughput and scale out

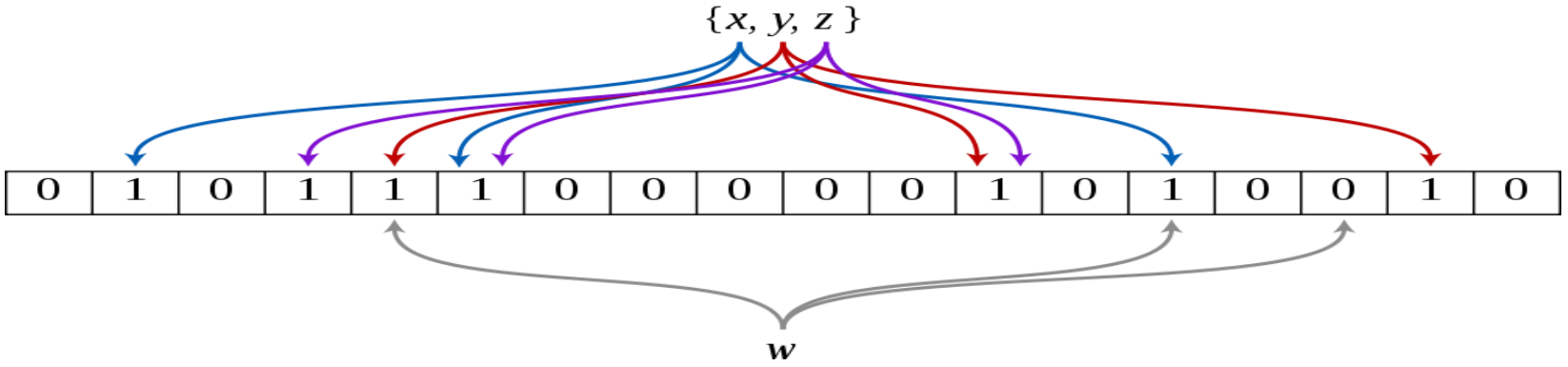
A single NGS machine generates more than 300 000 bp/s



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)

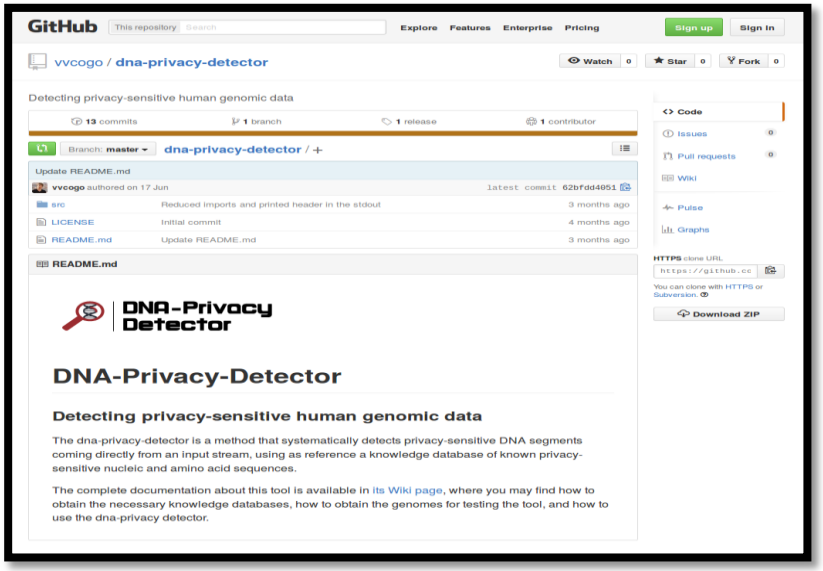
- False positive affects efficiency only (not efficacy)

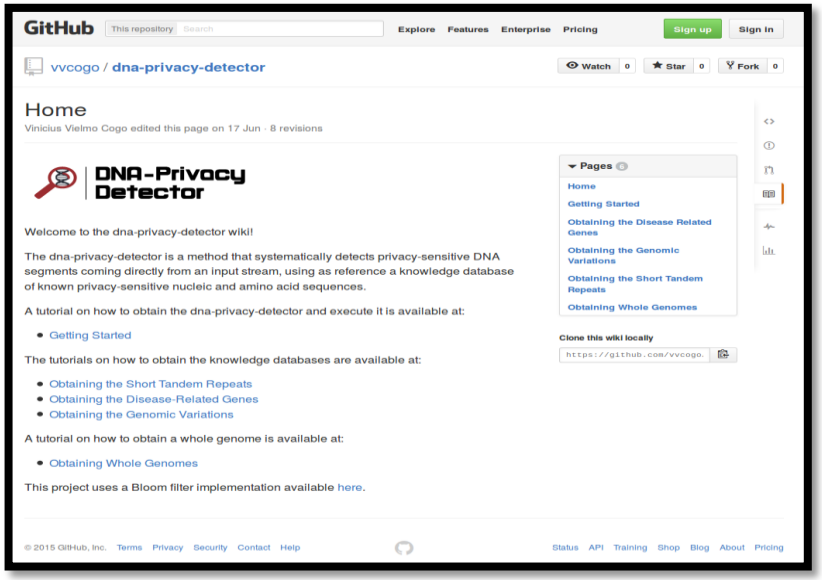
4. Implementing the detector

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

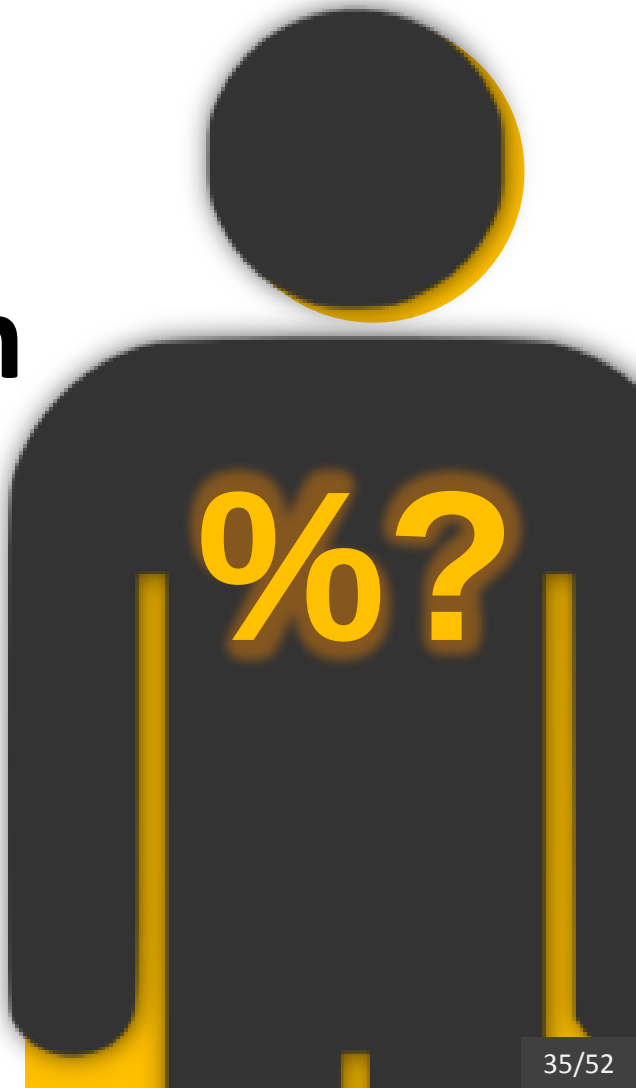
2. Bloom filter size

3. Detector’s throughput

5. Evaluation

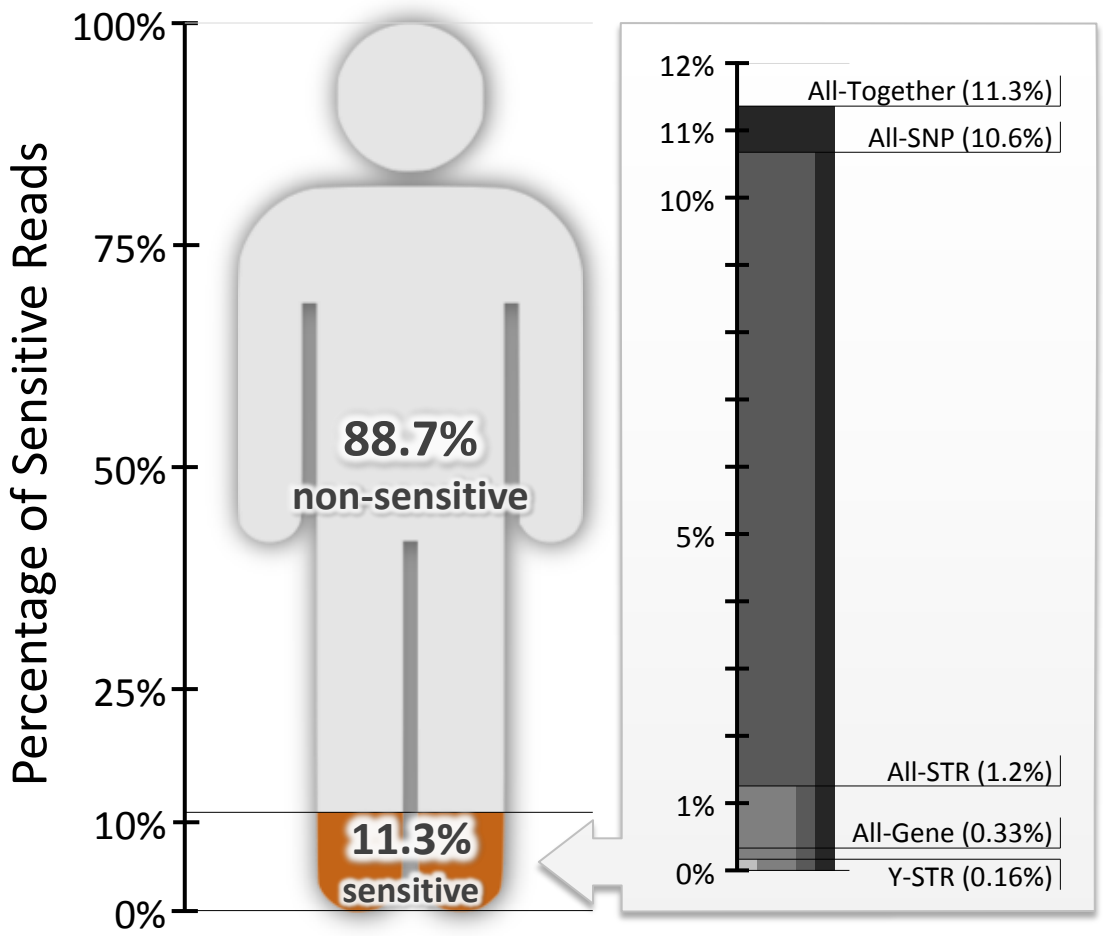
1st experiment

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



5. Evaluation

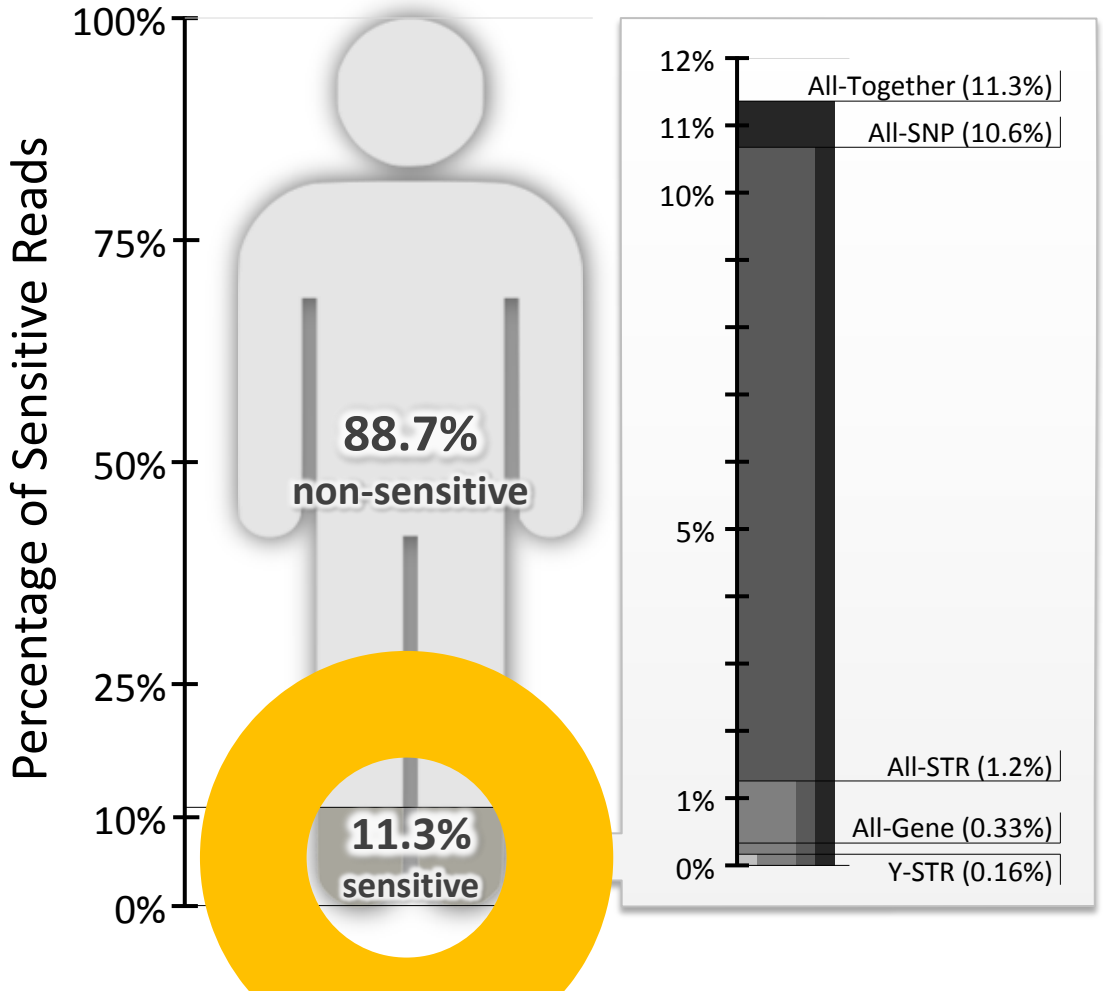
Percentage of privacy-sensitive sequences



5. Evaluation

Reduces the size of the problem **almost**

one order of magnitude



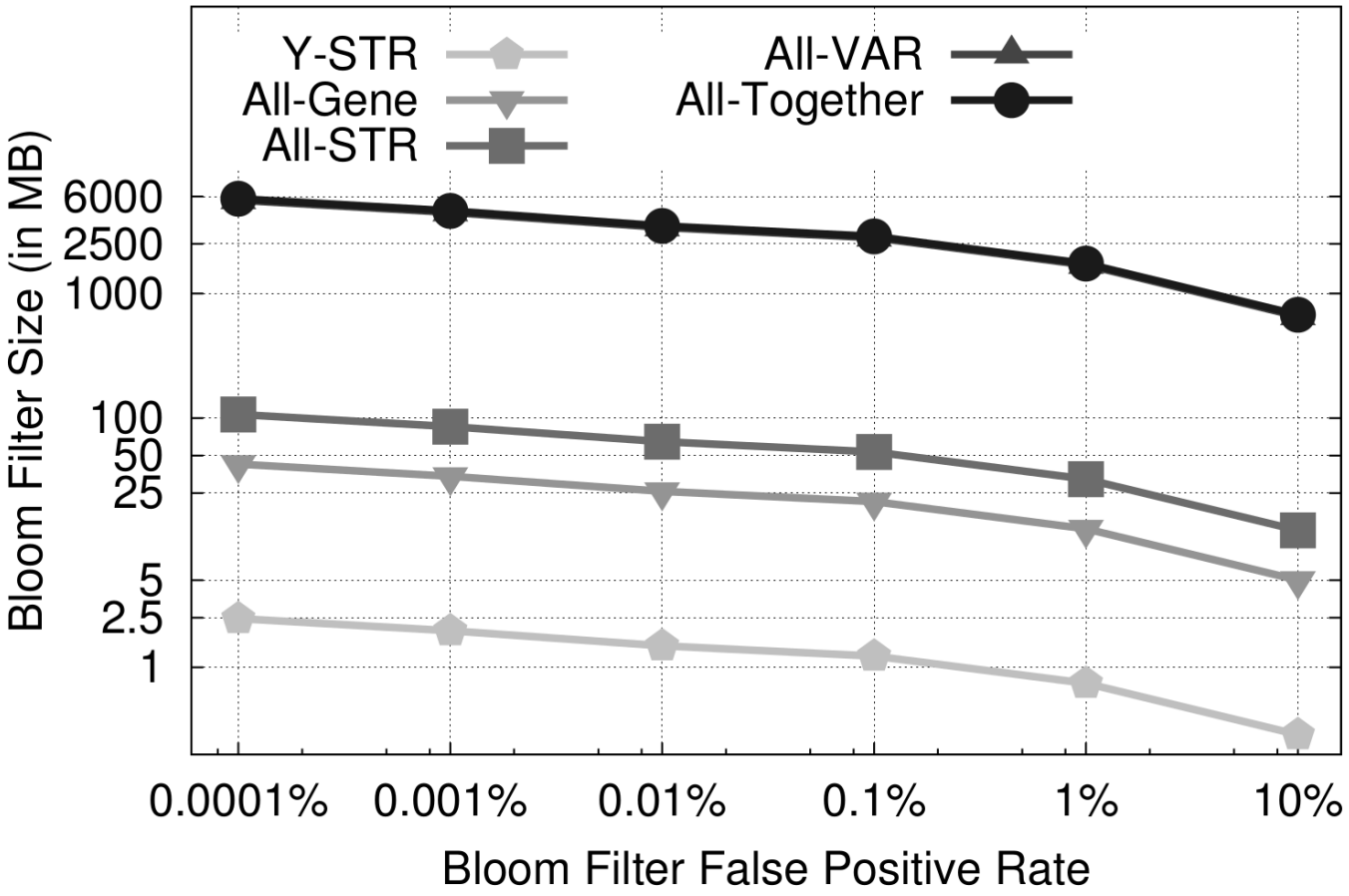
5. Evaluation

2nd experiment

How big is the
Bloom filter?

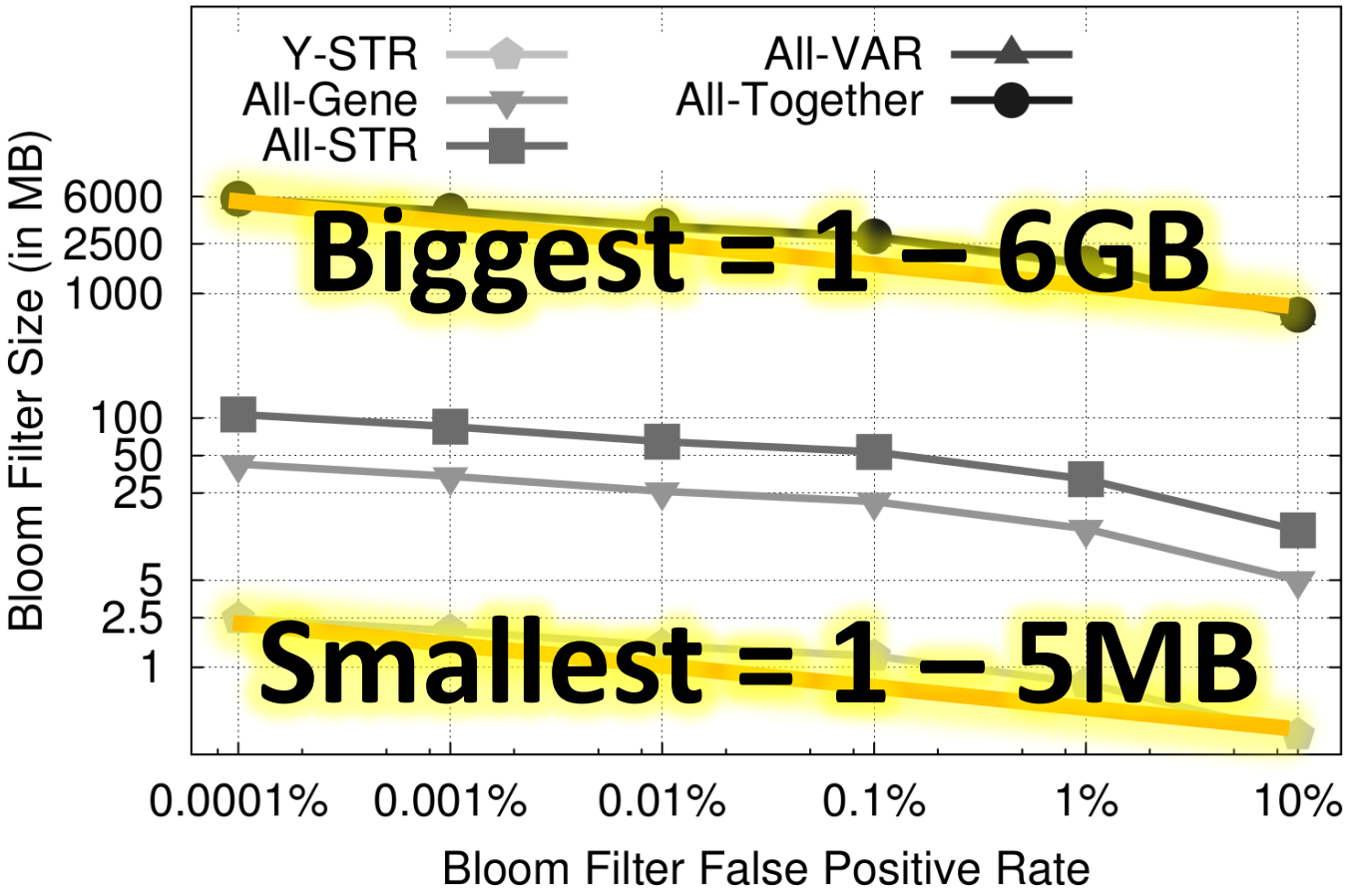
5. Evaluation

Bloom filter size



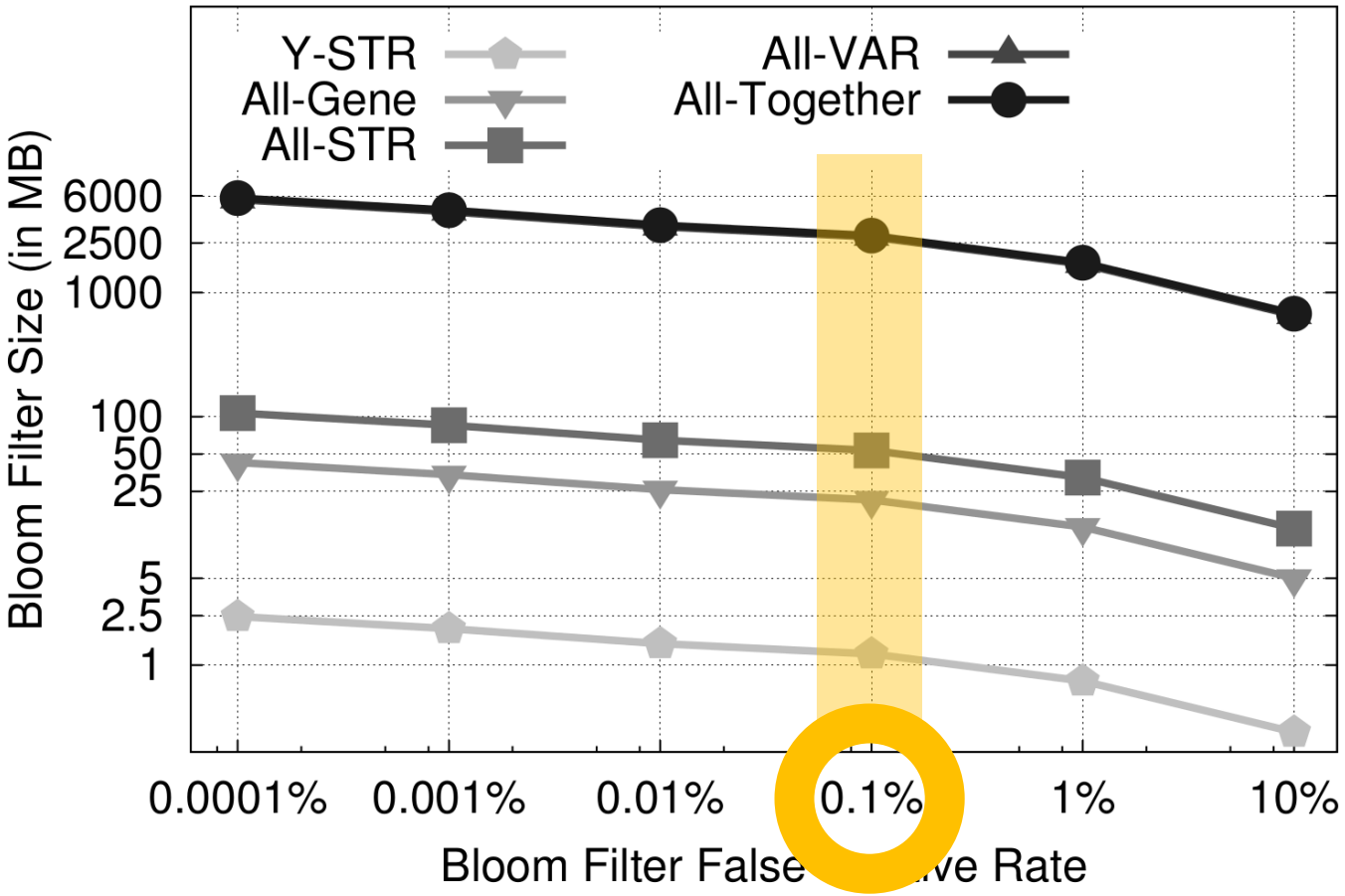
5. Evaluation

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



5. Evaluation

Sweet spot



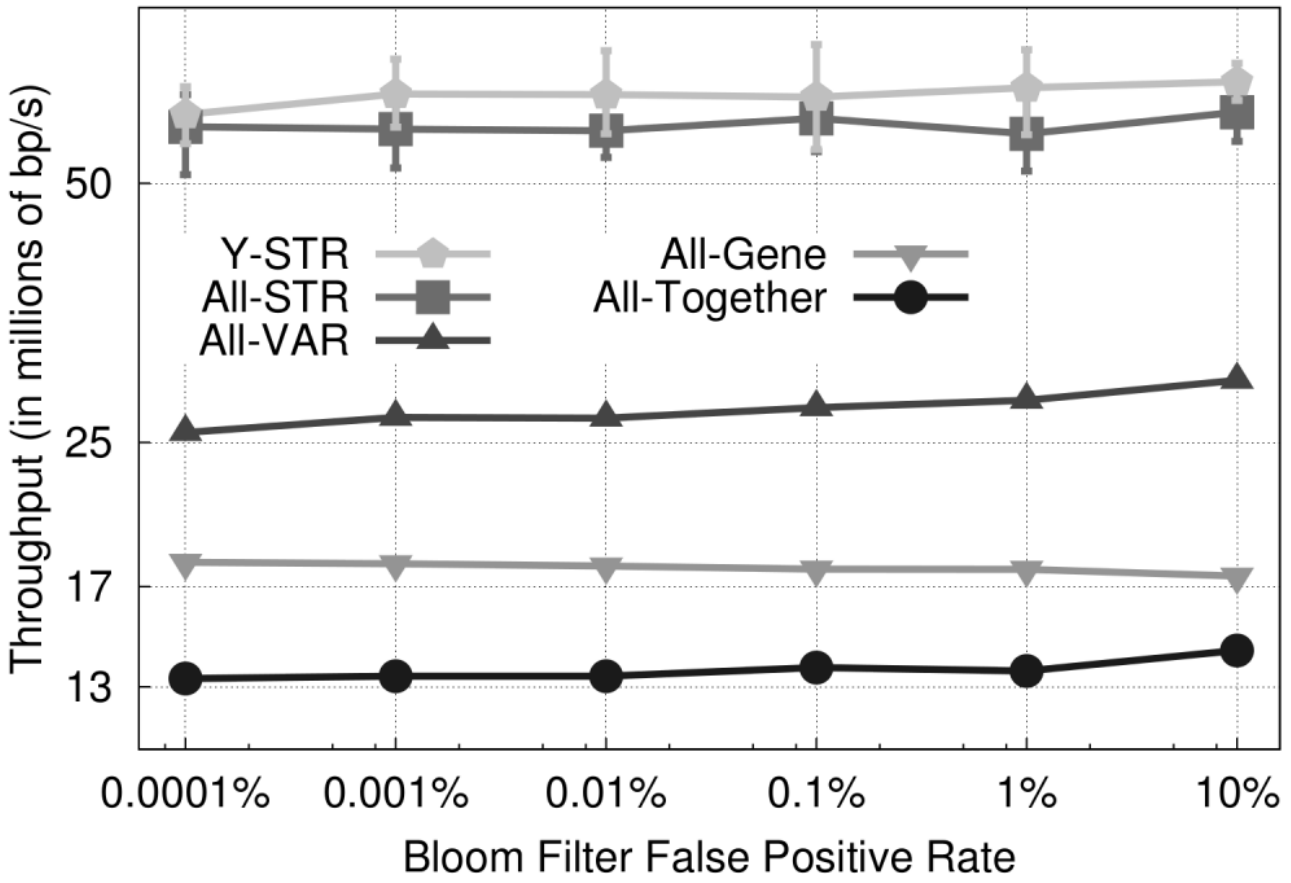
5. Evaluation

3rd experiment

Is the detector a bottleneck?

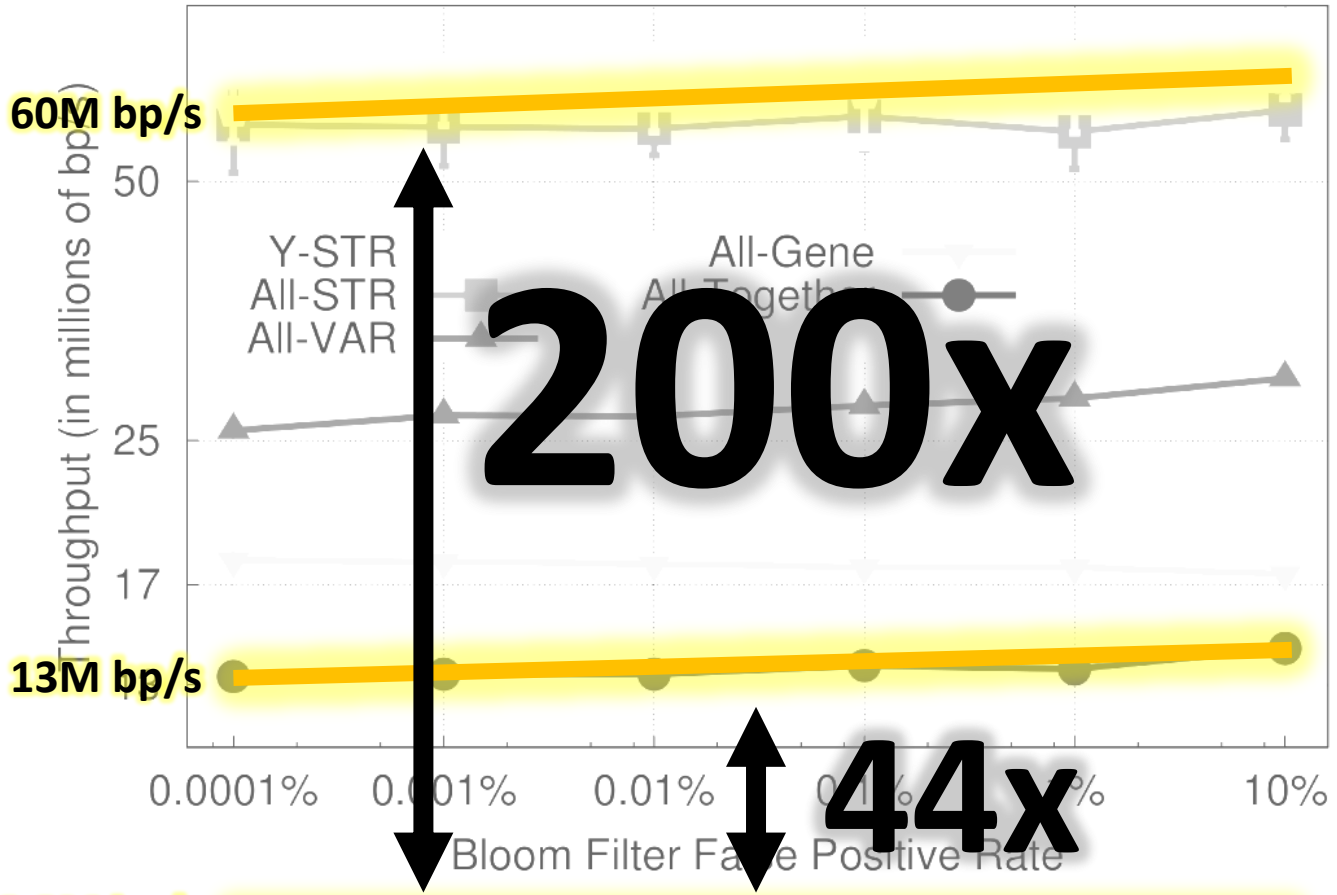
5. Evaluation

Throughput – Single core



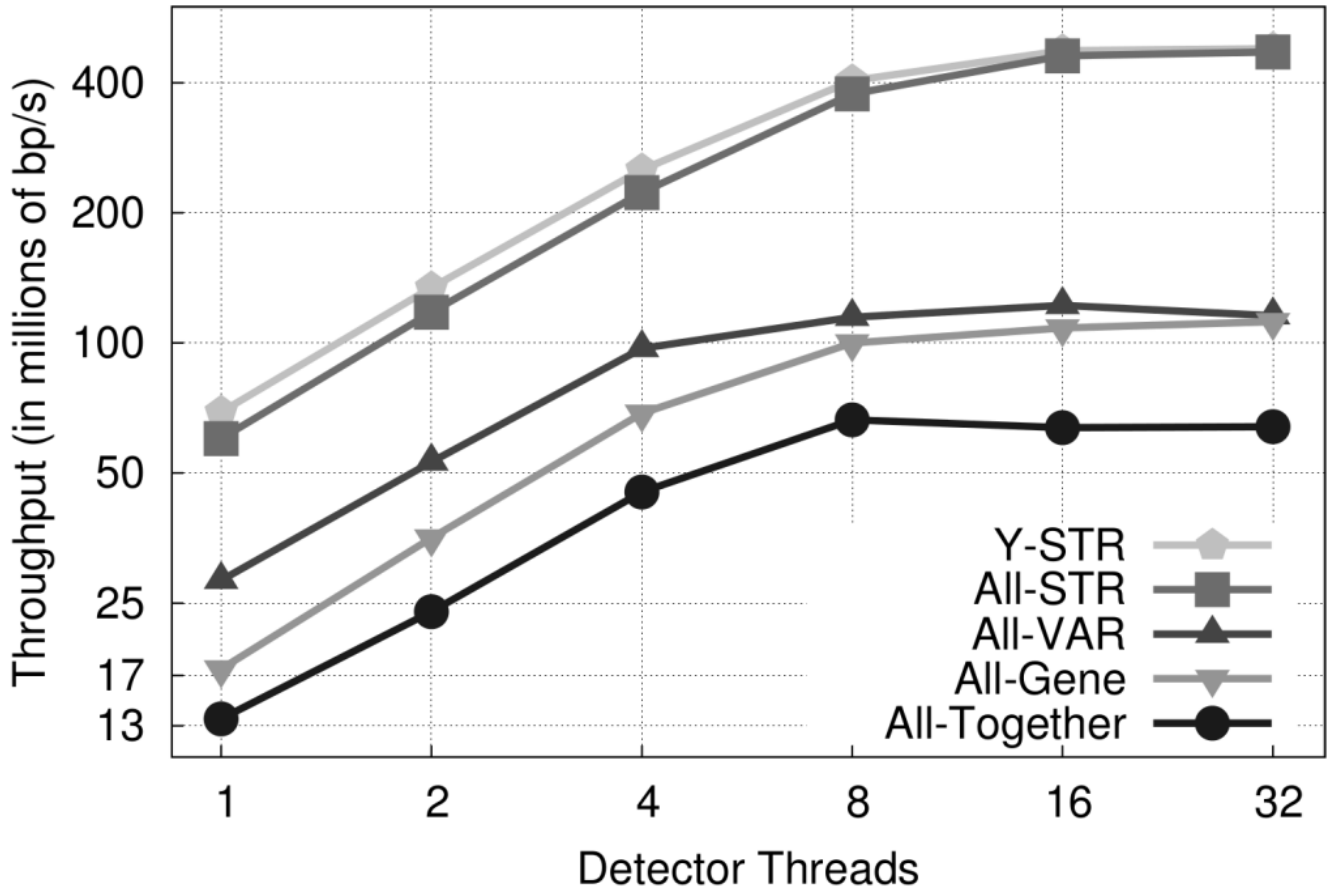
5. Evaluation

NGS machines = **300 000 bp/s**



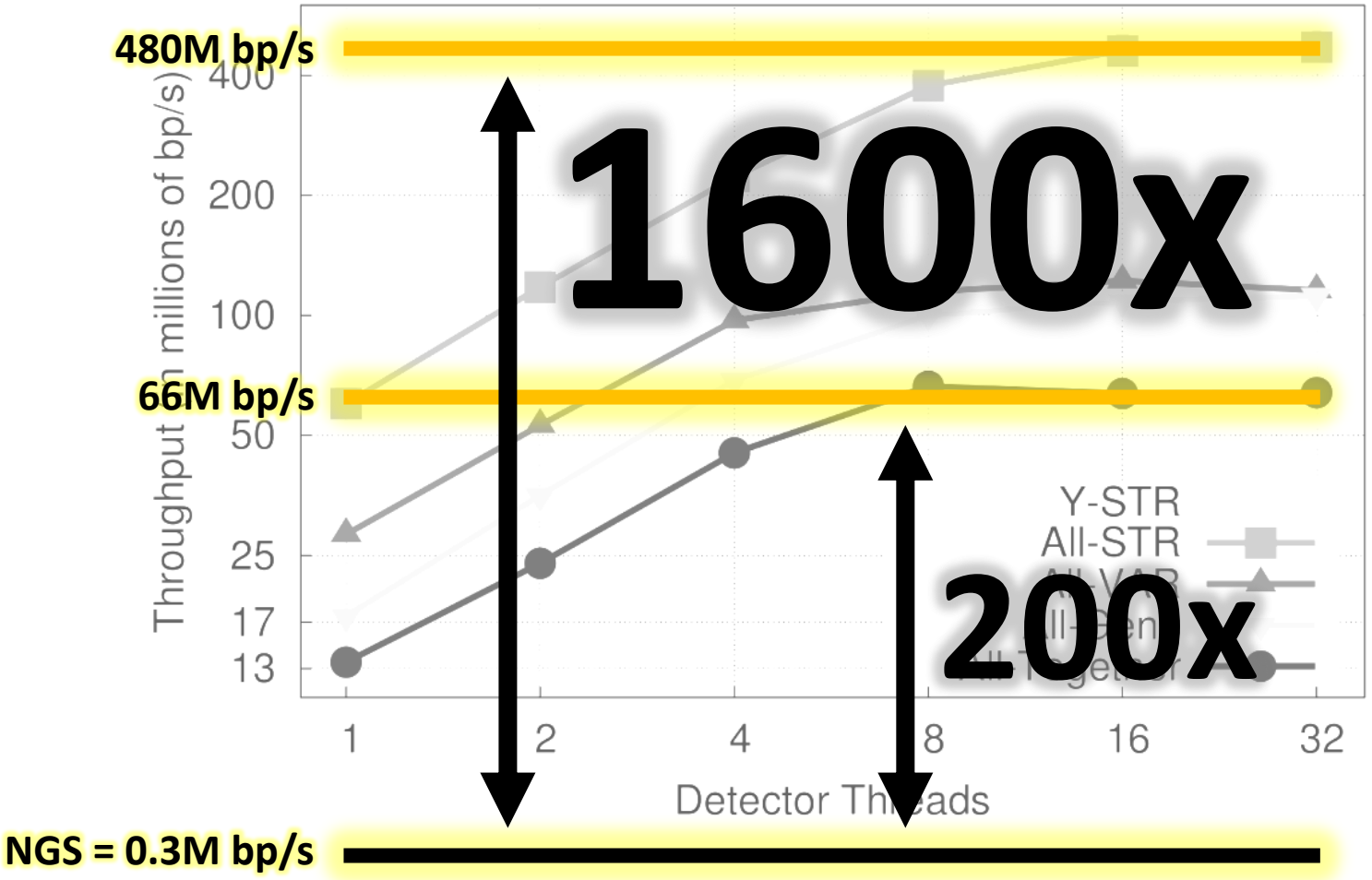
5. Evaluation

Throughput – Multi-core



5. Evaluation

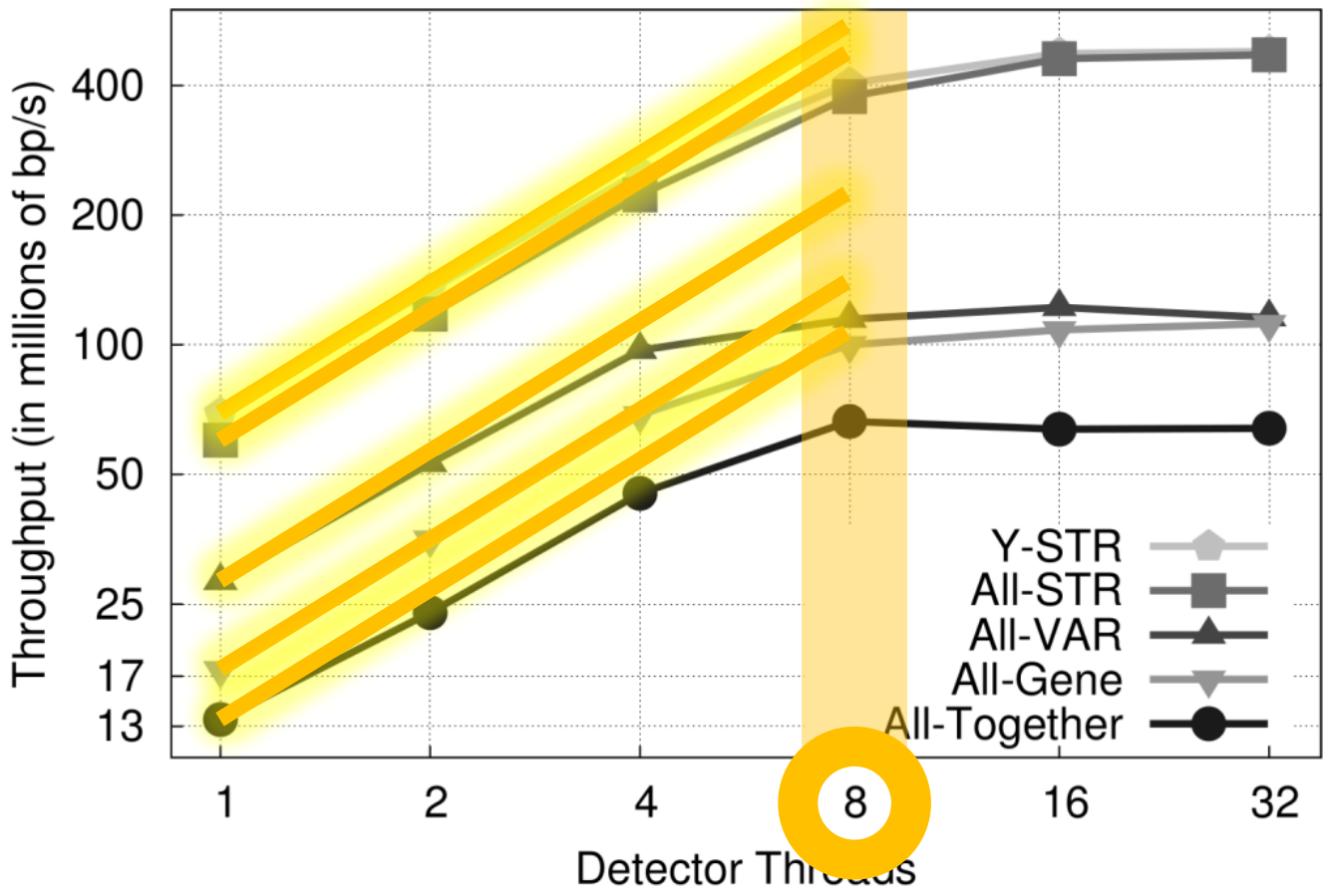
NGS machines = **300 000 bp/s**



5. Evaluation

Efficiency limit = **8 threads**

(Is it? 8-core machine)



6. Analysis

Completeness of the method

What if a **novel sequence** is discovered?

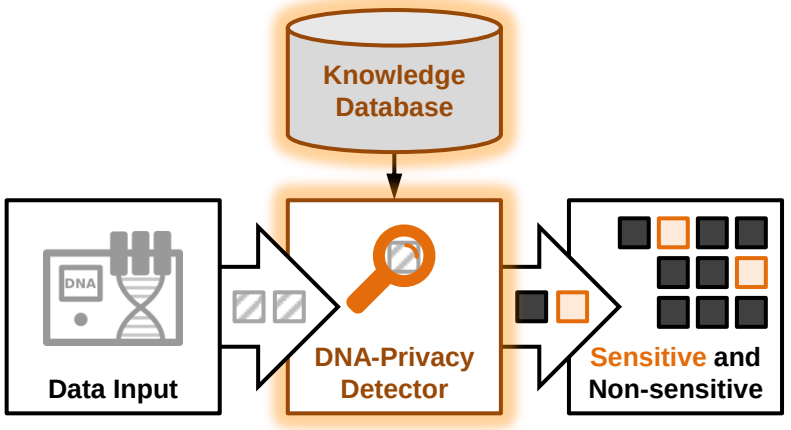
6. Analysis

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

7. 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

7. Conclusions

Future work and open problems

- **Advance knowledge** on privacy-sensitive sequences
- **Official repository** for the global knowledge
- **Educational service** about privacy-sensitive sequences
- Explore **practical uses**
- Deploy the detector **in production**

WPES 2015. October 12, 2015. Denver, US.

Thank you.

Vinicius V. Cogo[1]
Alysson Bessani[1]
Francisco M. Couto[1]
Paulo Verissimo[2]

[1]LaSIGE, University of Lisbon

[2]SnT, University of Luxembourg

5. Evaluation

Experimental setup

- 5 knowledge databases
  - **Y-STR = 15MB**
  - All-STR=660MB
  - All-Gene=87MB
  - All-VAR=34.4GB
  - **All-Together=35.1GB**
- **10 genomes**
  - Randomly selected from the 1000 Genomes Project
  - 3GB per assembled genome broken in reads with 30bp
- **1 Machine**
  - Dell Power Edge R410
  - 2x Quad-core (**8-cores**)
  - **32GB of RAM**
  - 146GB of storage (15k RPM)
  - Ubuntu Server (Lucid Lynx–amd64) and Java 7 (64-bits)

6. Analysis

Practical uses