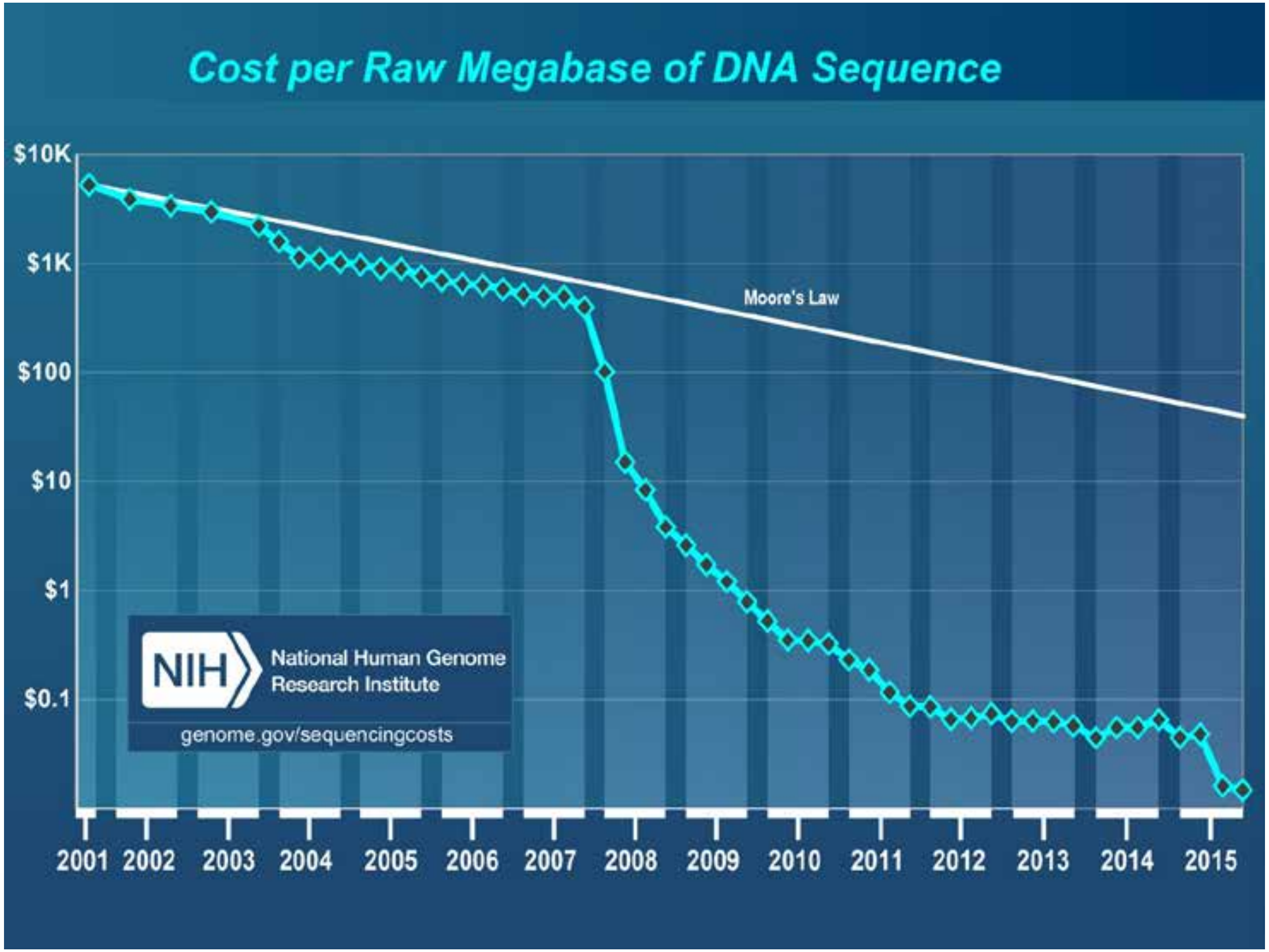
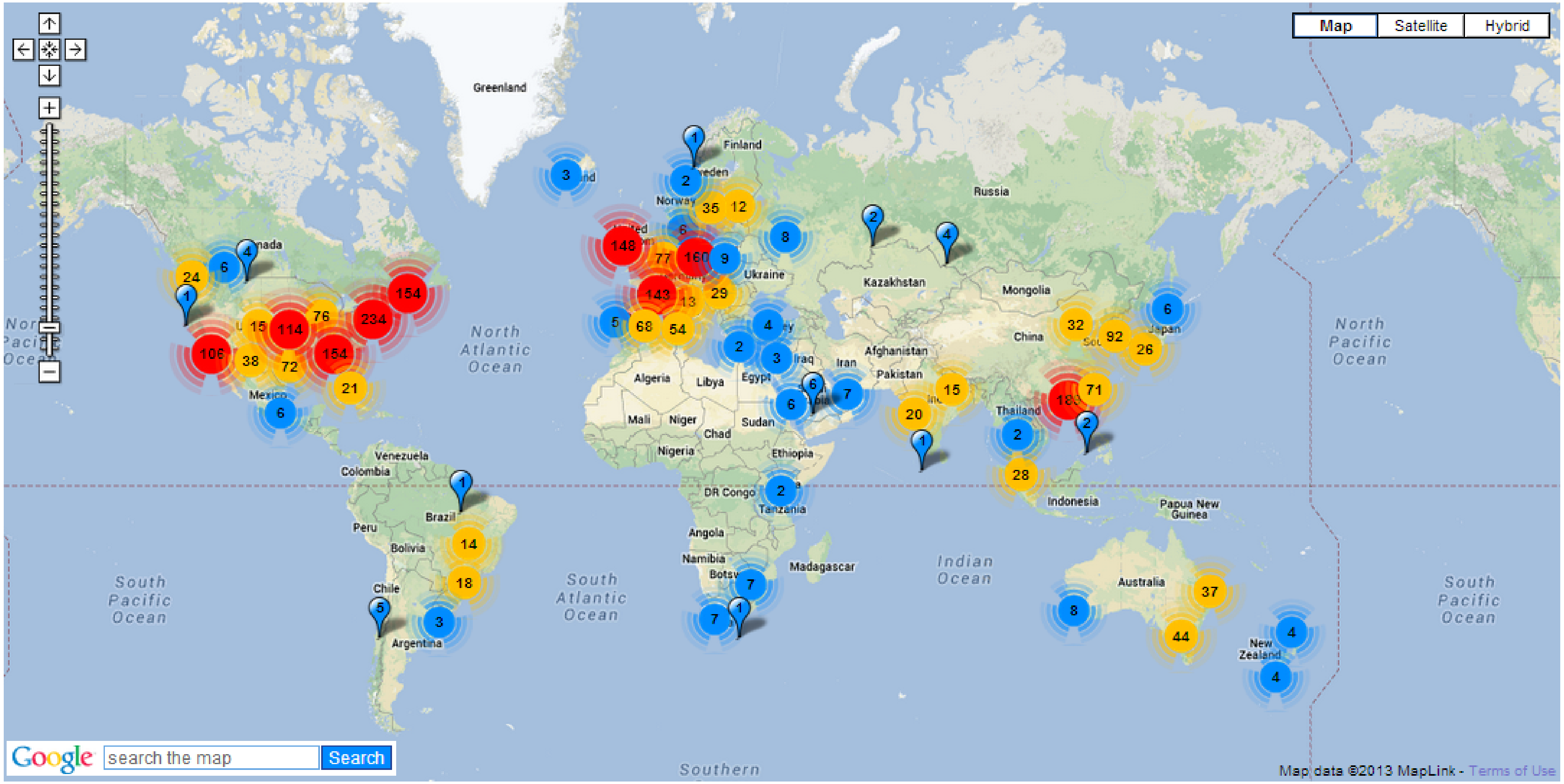
Architecture-Aware Privacy-Preserving DNA Filtering and Alignment

Maria Fernandes, Jérémie Decouchant,
Marcus Völp, Paulo Esteves-Veríssimo

The cost of DNA sequencing



Sequencing machines around the world



Sharing genomes is required

- Early diagnosis of serious diseases
- Prospects of personalized medicine
- Genome editing

The highest value of genomes is achieved when they are shared

- New variations can only be found when comparing genomes
- Reliable statistical relationships between variations and disease need a high number of genomes

Since January 2015, the USA NIH requires genomic data of all types to be shared (Genomic Data Sharing Policy).

Industry Initiatives

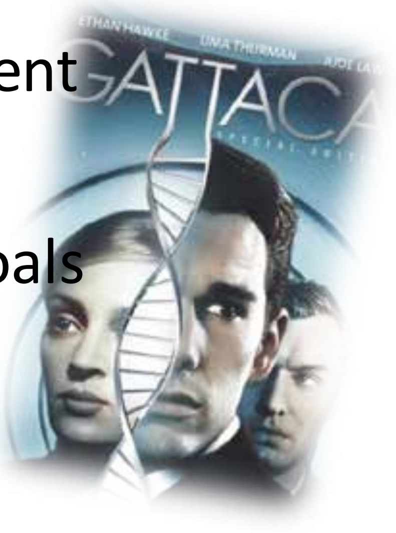
- IT giants start proposing genome-related services
  - Verily - Google Genomics
  - IBM Research (computational genomics)
  - Microsoft Research (genomic research in collaboration with Sanger Center)
  - Apple (the ResearchKit program)
  - Amazon
- Global Alliance for Genomics & Health
  - Definition of a common framework for effective, responsible and **secure** sharing of genomic and clinical data
  - Security Working Group: security infrastructure policy and technology http://genomicsandhealth.org

Adapted from H. Ayday and J.-P. Hubaux CCS’16

Sharing genome is scary

Misuses of genomic data:

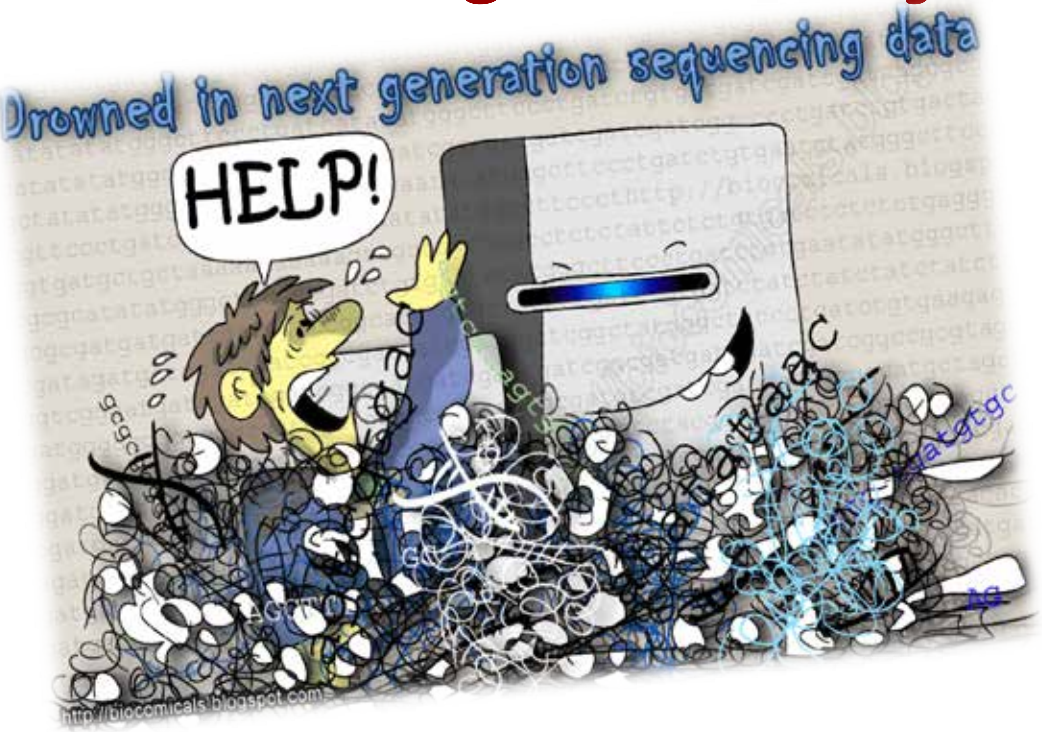
- Denial of access to health insurance, mortgage, employment
- Blackmails (e.g., non-legitimate child)
- Use disease predisposition, alter a genome for criminal goals
- Genome traces artificially inserted in crime scenes

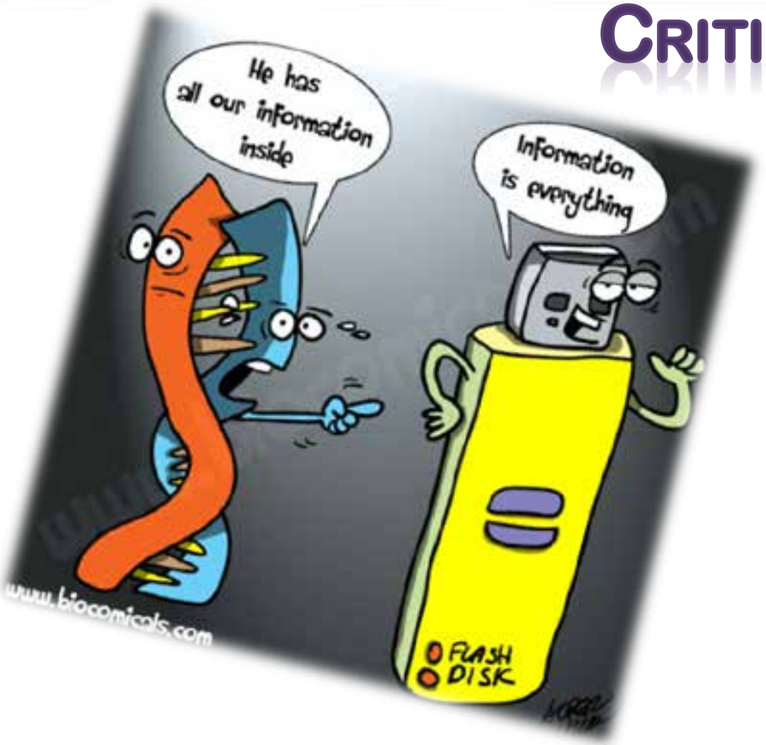


Classes of attacks known in the literature:

- Re-identification attack (i.e., broken anonymization)
- Membership attack (i.e., identifying a genome in a study)
- Recovering hidden parts of genome (i.e., Alzheimer’s gene)

Problem:
Data *Sharing* vs. *Privacy* Preservation


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com



- Use of clouds to deal with the huge amounts of genomic data
- Data needs to be accessible for research
- Privacy of genomic sequences has to be maintained throughout their lifetime, despite these threats

Architectures for privacy-preserving genomic data handling

- The literature has mainly focused on architectures that protect *refined* genomic information
- However, before that, *raw* genomic data has to be produced (sequencing), and interpreted (alignment and variant calling)
- Currently, those steps are in essence executed:
  - Without protection in a public cloud (limited privacy)
  - Locally in a private cloud (limited scalability, and confined to perimeter protection)

Public clouds & Alignment

- Chen et al. 2012: Private and Public Clouds
  - Uses a public cloud to find matching hashed k-mers between reads and the reference genome
  - Uses a private cloud to extend around matching locations
  - Still significantly relies on a private cloud
- Balaur et al. 2017: Public Clouds
  - Find matching positions using k-mers hashing
  - Extends around matching positions thanks to a voting mechanism with encrypted k-mers
  - Heavy in terms of communication (GBs of data)

Our approach to private & efficient architecture aware alignment

We aim at:

- Filling the privacy gap between sequencing machine output and cloud-based genomic analysis
- Combining data analysis algorithms with architectural solutions
- Designing a lightweight but effective scheme, which can be used in combination with clouds assuming different privacy assumptions

Background: Sensitive vs non-sensitive genomic information

The differences a genome G has, in comparison with a reference genome R, are sensitive.

Ref. (R): GGCTCGTCAAGCATCGCGAC
Genome (G): GGCTCGTCATGCAGGCGAGGC

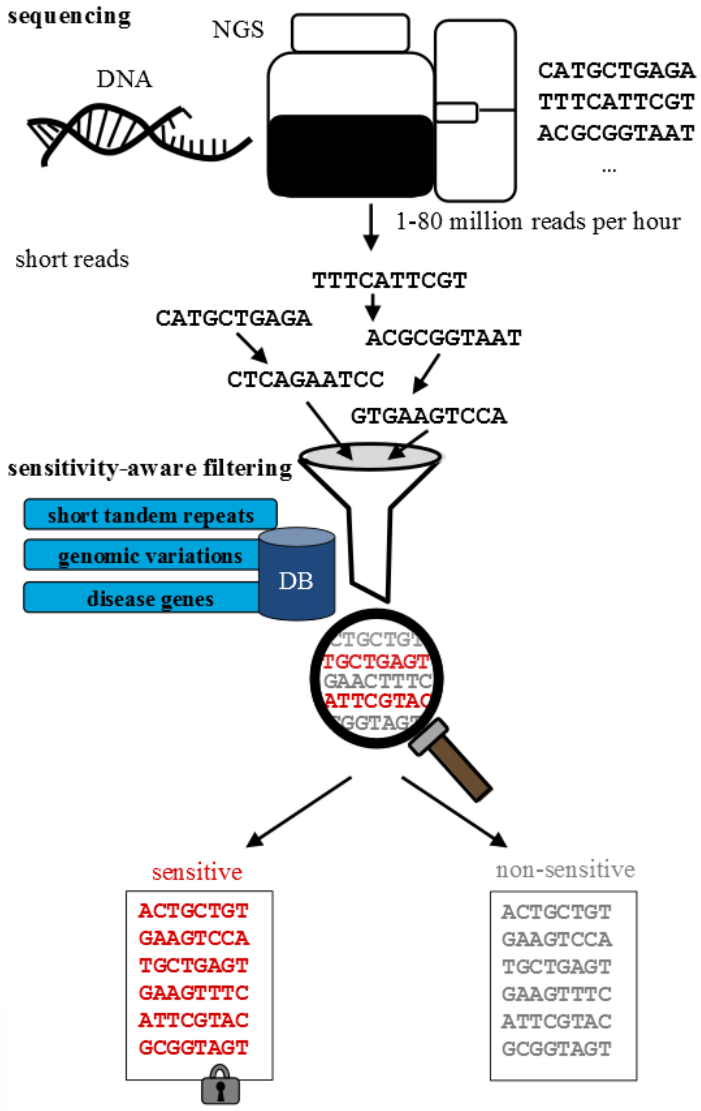
Variations:
Chr. 1, Pos. 10, A, T
Chr. 1, Pos. 14, TC, G
Chr. 1, Pos. 20, C, GGC

Sequencing machines produce subsequences from a genome, which may carry sensitive nucleotides.

Short Read Filter [Cogo et al. - WPES 2015]

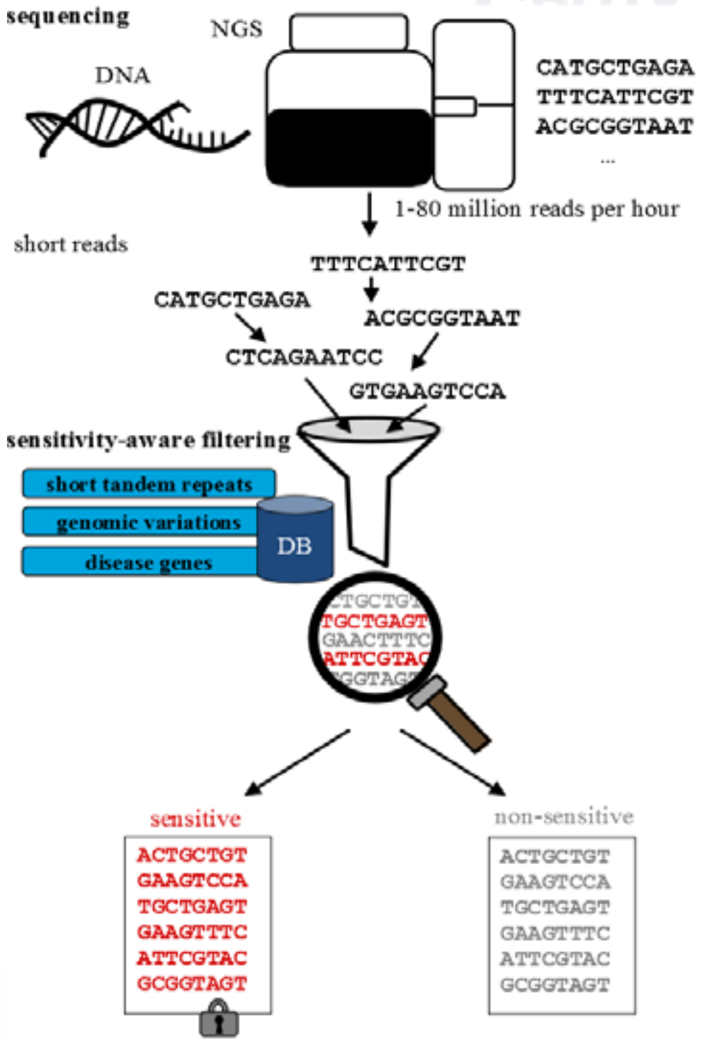
Directly at the mouth of a sequencing machine, lightweight enough to be embedded in it.

The filter determines if a 30-nucleotides sequence contains at least one sensitive nucleotide.



Short Read Filter – How it works

1. Build a dictionary of 30-base sensitive sequences (simplified):
   - all sequences of 30 bases that contain a sensitive base
2. Insert this dictionary in a Bloom filter
3. Test the membership of each read in the Bloom filter, and upon detection classify it as sensitive



Reads detected non-sensitive can be uploaded to less protected places, even a public cloud.

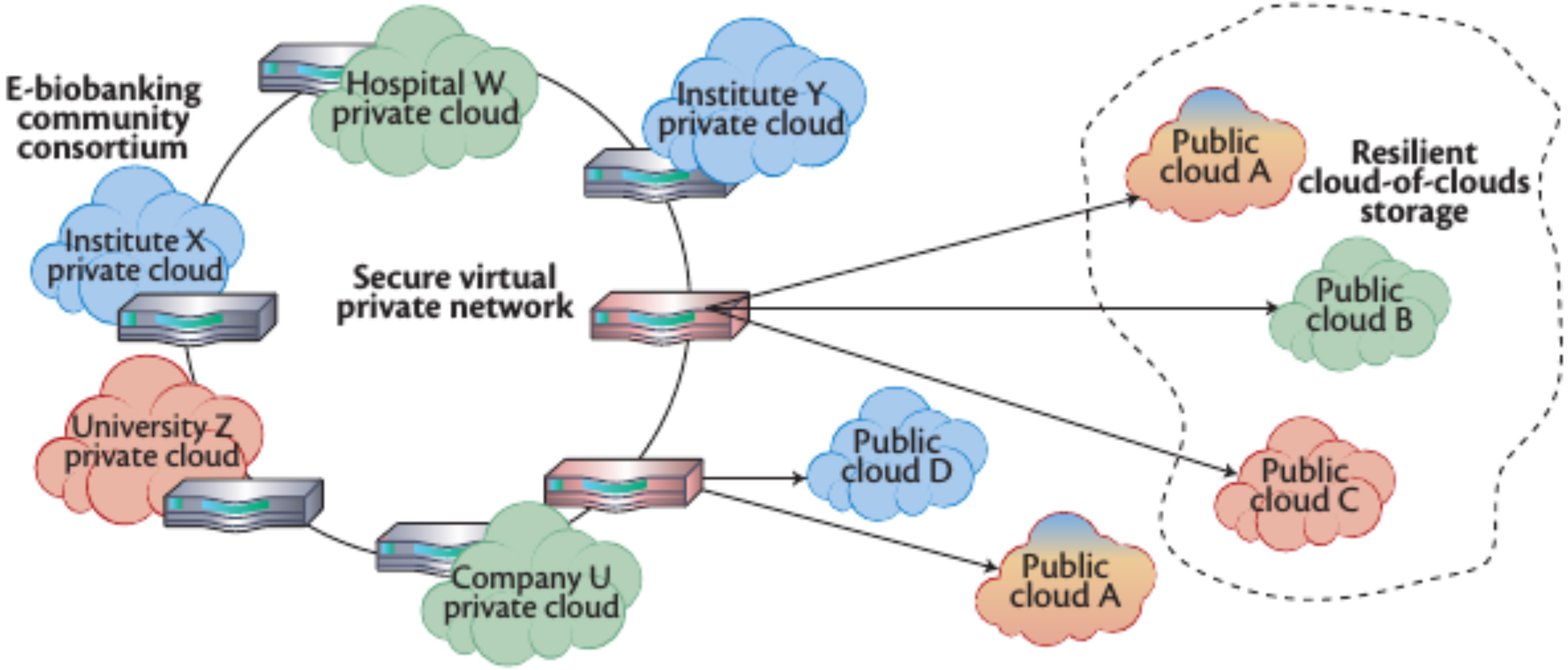
Functional limitations in Cogo's method

- Only one position is allowed to change compared to the reference genome.
- Can be extended to long reads, but the proportion of reads detected as sensitive increases with their length (up to 90% with 1000 bases)
- Does not tolerate sequencing errors
- Gives a binary answer (too coarse)

Privacy and Availability problems

- A large fraction of the reads are classified sensitive, and have to be aligned/accessed with high security constraints
- Supports only two kinds of infrastructure: private and public clouds
- Even though variations do not leak the same amount of information, they are all considered equally sensitive

E-Biobanking, e-Health: Multiple clouds and levels of protection


E-biobanking community consortium
Hospital W private cloud
Institute Y private cloud
Institute X private cloud
Secure virtual private network
University Z private cloud
Company U private cloud
Public cloud D
Public cloud A
Public cloud A
Resilient cloud-of-clouds storage
Public cloud B
Public cloud C

How can filtering take advantage from clouds providing several levels of trust?

Sensitivity levels for reads

- *The sensitivity of a read is the sensitivity of its most sensitive nucleotide.*
- *Do we need to protect everything up to the highest standard?*

  Anyway, we can't!
- *Nucleotides can be classified in discrete categories according to the sensitivity of the information they reveal, perceived or evaluated.*

Sensitive data	Privacy leak	Severity
genomic variations	re-identification	🟥
	physical characteristics	🟥
	predisposition for diseases	🟩
short tandem repeats	re-identification	🟥
	parental relation	🟩
	allelic profile	🟦
disease genes	predisposition for diseases	🟩
	rare diseases	🟩
	response to drugs	🟦

🟥 Very high

🟩 High

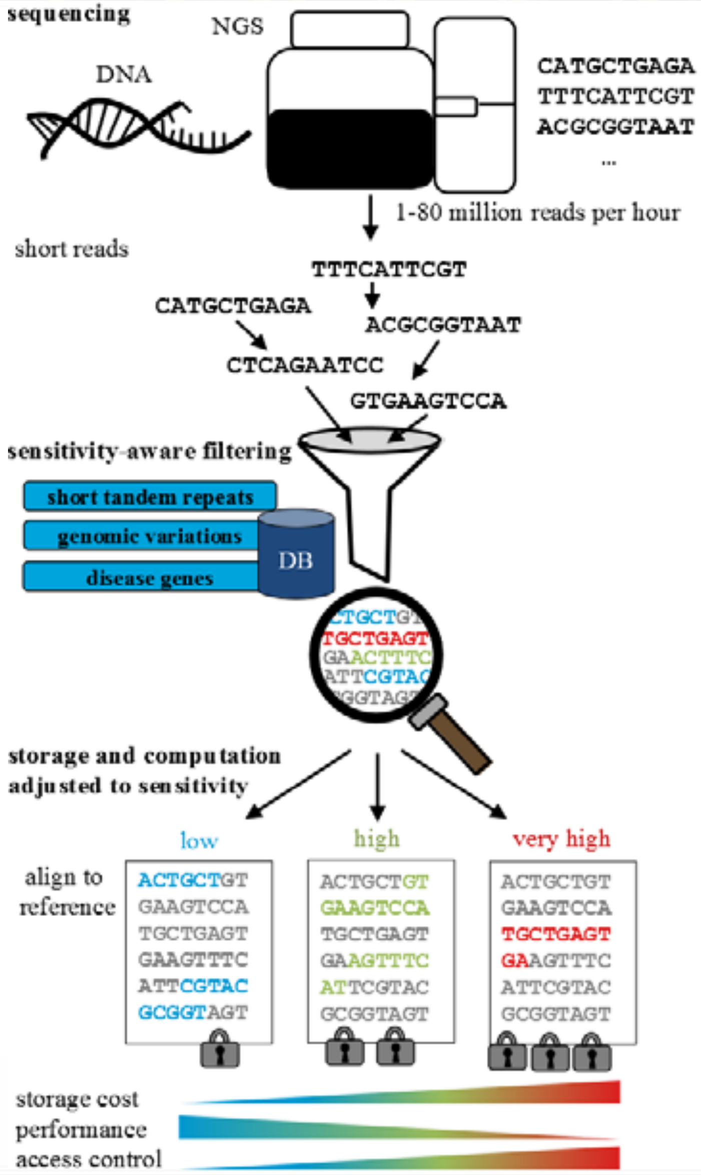
🟦 Low

Classifying reads into sensitivity levels

Risk analysis: several sensitivity levels depending on the rarity of the mutation they carry.

We initialize several filters with the disconnected sensitivity levels

Nucleotides inside the reads are marked with their sensitivity level



Improving performance

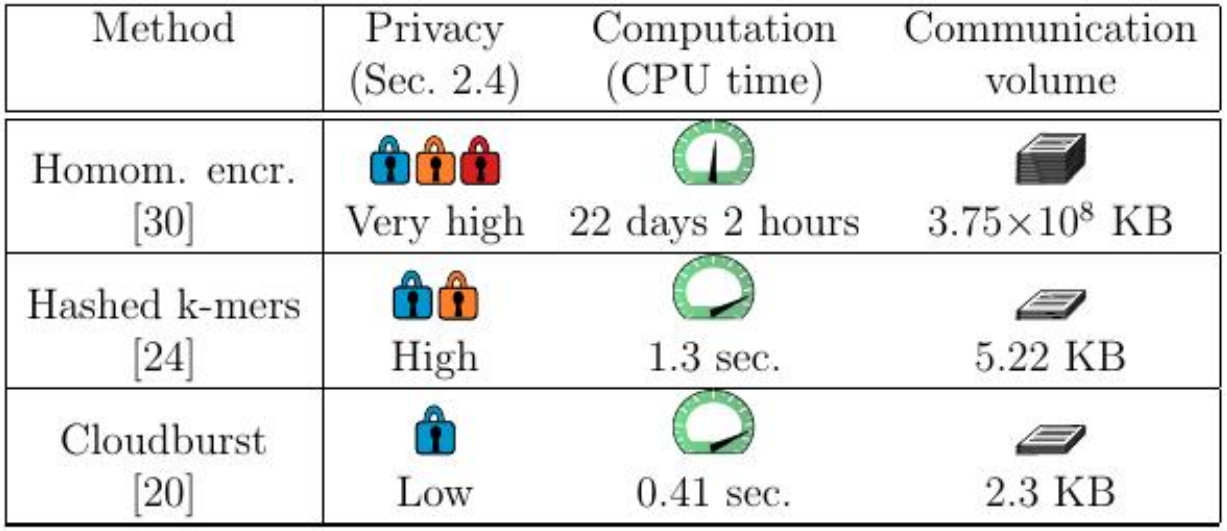
- We rely on the rich existing ecosystem of alignment algorithms
- Each algorithm as attributed a Privacy x Performance metric
- The overall performance of secure alignment is increased as soon as a public cloud is available

Challenges:

- Detection of the sensitivity levels
- Genomic variations may be connected across sensitivity levels

Privacy-preserving read alignment

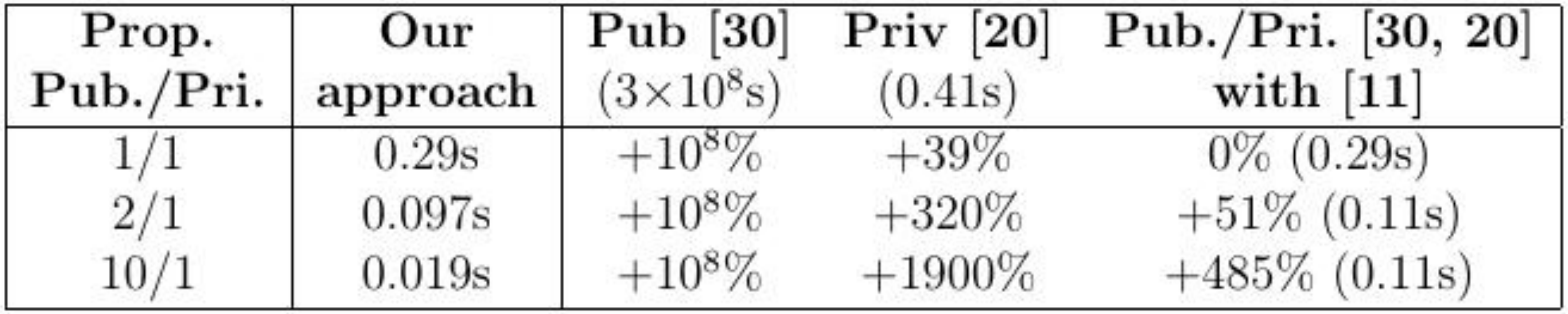
(using previous filtering results)



Method	Privacy (Sec. 2.4)	Computation (CPU time)	Communication volume
Homom. encr. [30]	🔒🔒🔒 Very high	22 days 2 hours	3.75×10^{8} KB
Hashed k-mers [24]	🔒🔒 High	1.3 sec.	5.22 KB
Cloudburst [20]	🔒 Low	0.41 sec.	2.3 KB

computation cost (CPU hours) and communication cost (bytes) of aligning a single 100-basepairs read to the full human genome. Single core.

Chen Hyb [24]
5PM [30]
CloudBurst [20]
Cogo [11]



Prop. Pub./Pri.	Our approach	Pub [30] (3×10^{8}s)	Priv [20] (0.41s)	Pub./Pri. [30, 20] with [11]
1/1	0.29s	$+10^{8}$%	+39%	0% (0.29s)
2/1	0.097s	$+10^{8}$%	+320%	+51% (0.11s)
10/1	0.019s	$+10^{8}$%	+1900%	+485% (0.11s)

Distributed read alignment

- The previous alignment scheme is still orchestrated by the cloud hosting the sequencing machine
  - (transmitting the reads, collecting the results, sending them again for the following steps).

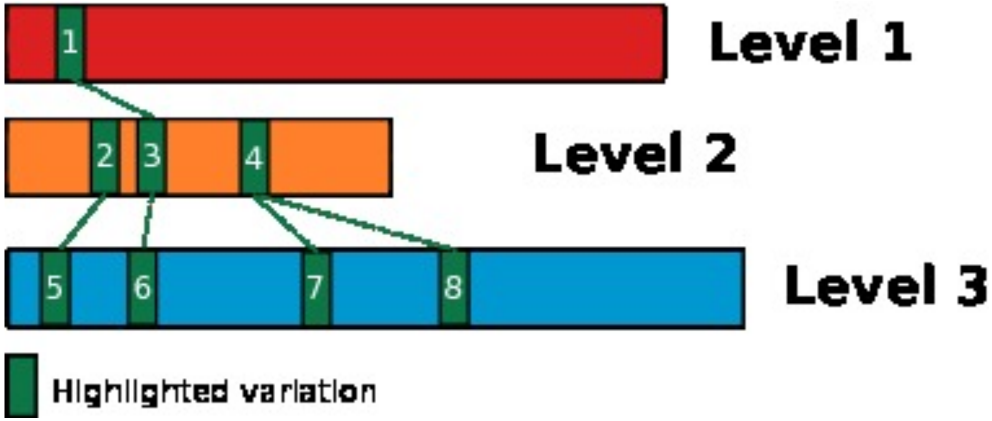
- Sending data back and forth for sharing should be avoided:
  - legal and IP-protection reasons
  - data set size is dramatically increasing

- The next step (WiP) is to partition reads across multiple categories of clouds and craft architecture-aware algorithms

Genomic variations may be connected across sensitivity levels

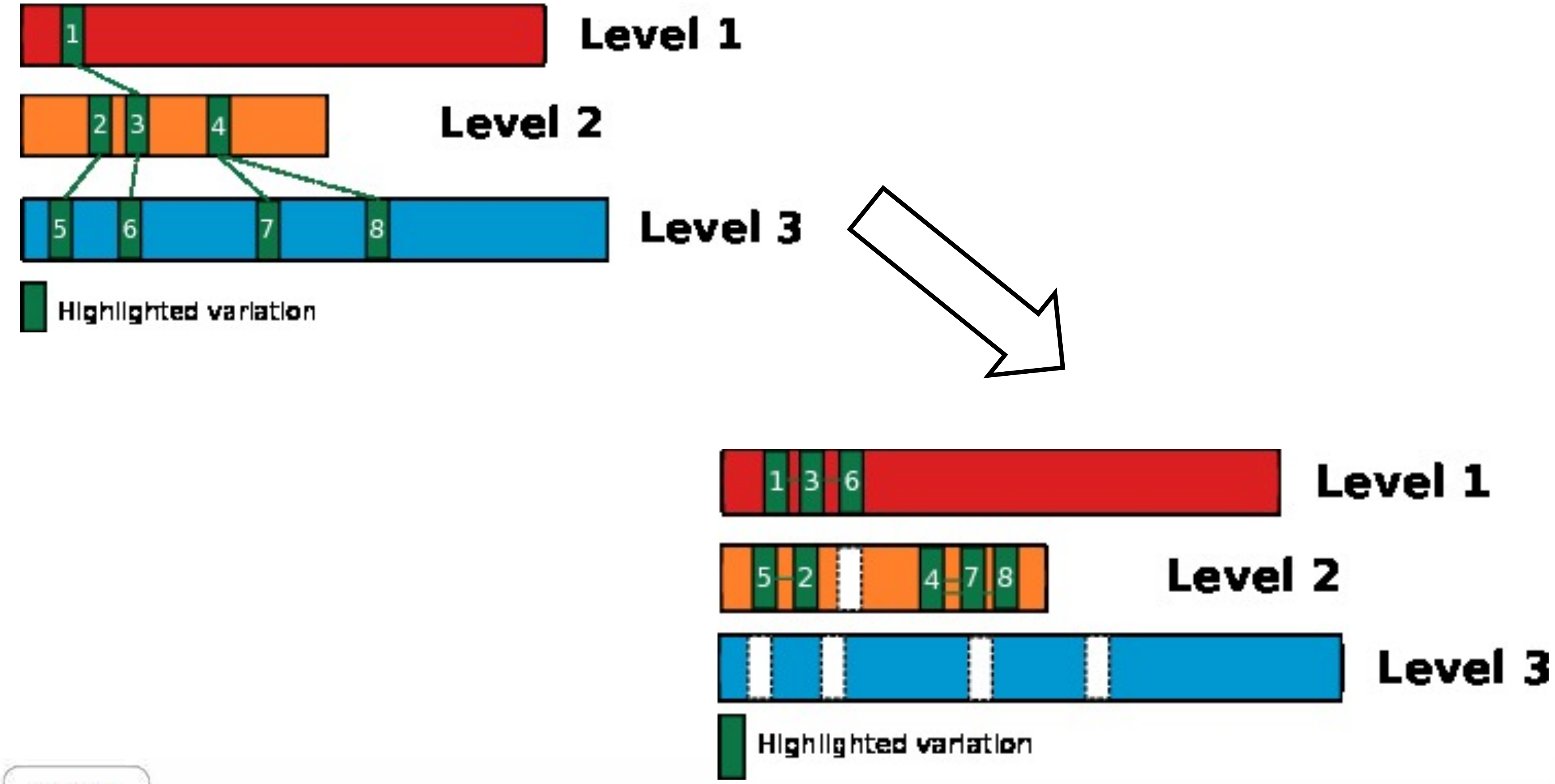
- this contaminates sensitivity level confinement, leaking information

Disconnecting sensitivity levels



Connections between layers endanger our protection mechanisms -> If a level is attacked, the attack can be amplified to a higher level

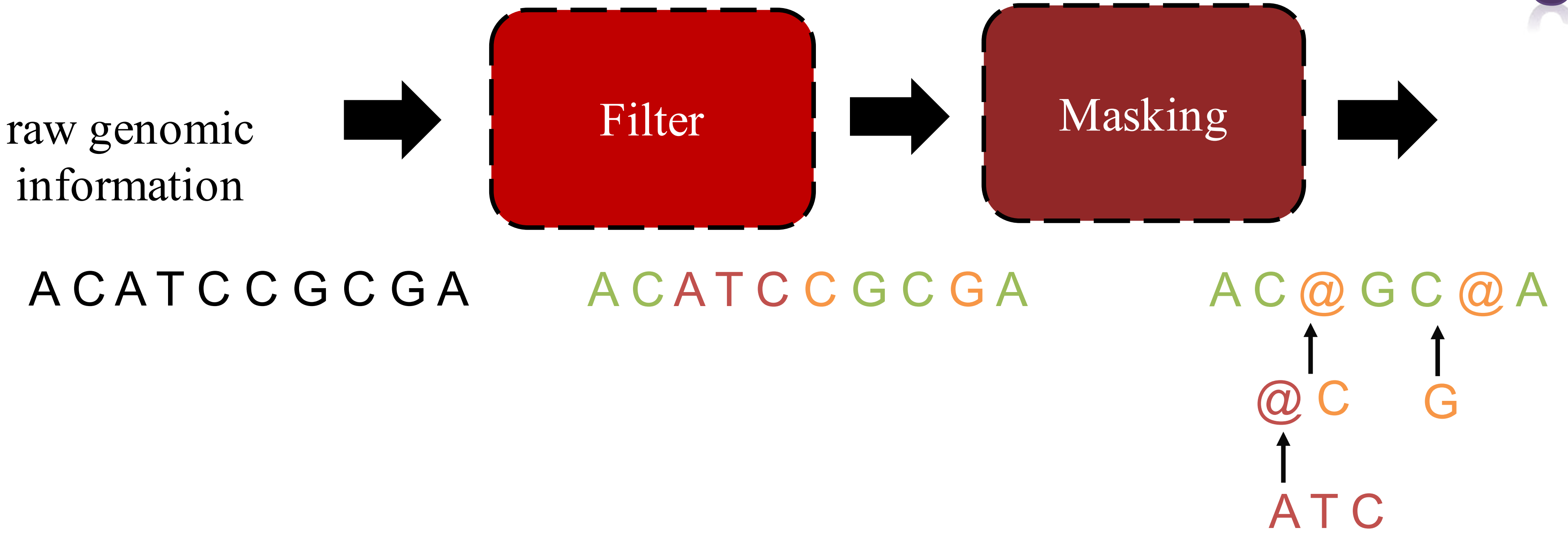
Disconnecting sensitivity layers - Results



What if we could pretend a sensitive read *is not* a sensitive read?

- masking out sensitive regions of reads

Sensitivity aware masking



- The masking takes place in a secure environment
- Masked out regions are marked with “N”s (to be compliant with existing reads formats), or “@”s

Paulo Esteves-Veríssimo

University of Luxembourg Faculty of Science, Technology and Communication

and SnT, the Interdisciplinary Centre for Security, Reliability and Trust

paulo.verissimo@uni.lu

http://staff.uni.lu/paulo.verissimo

CRITIX @SnT

Critical and Extreme Security and Dependability

http://wwwen.uni.lu/snt/research/critix

We're hiring bright PhD students and post-docs willing to address these challenges!

Thank you!