DPM 2013
Egham, UK
September 12-13, 2013

Data Privacy Management
8th International Workshop

Co-located with
ESORICS 2013

PRIVACY-PRESERVING PROCESSING OF RAW GENOMIC DATA

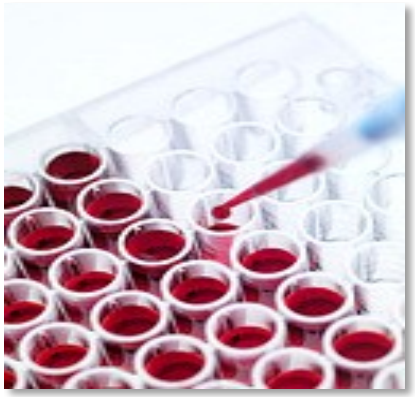
Erman Ayday, Jean Louis Raisaro, Urs Hengartner, Adam Molyneaux and Jean-Pierre Hubaux

SEPTEMBER 2013

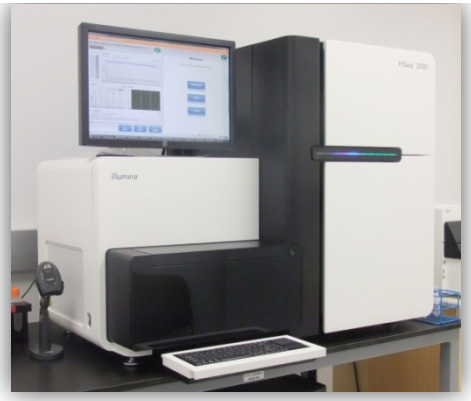
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

UNIVERSITY OF Waterloo

Sophia Genetics

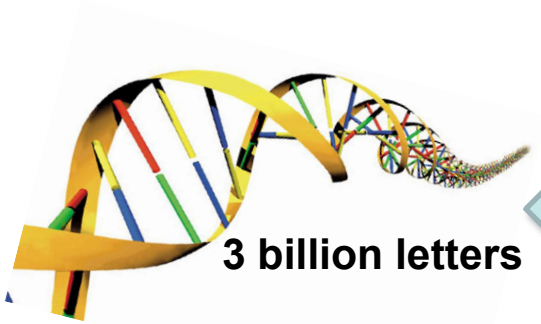


Samples



Sequencing machine

Raw data (short reads)



SAM file (aligned reads)

23andMe

ancestry.com

Counsyl

GENEPLANET
PERSONAL GENETICS

ID chromosome mapping quality
Alignment Flag position Cigar string

```
GZQG54D02DI67S 16    10    72065884    228    3S23M1I255M8S  *    0    0
TAGCCCCCACCCAACACACACACACGTCTTTTTTTCTGAATCATTTGAGTGTACCATACATAC
GTCATGGCCCTCAGTGTATATTTCCTAGGAATTGGGACATTCTCTTACATAACCACAGTGCAGT
TAAGGGTTCCAGGTATTCACATTGATAGAATGTGTATCTAACCCATGGCCCACAGTTCAGTGT
GGTCAGGTGCCCTAACATCCTGTACAGCACTTCCCCCTCTAGTACCAGAGCCAGTCTGGAAT
CAAGGATTGCATTTACTTCTAACTCTTTGGCCCCTCTG      E=7/////200088:@;;;@EEIID?,?
2666644IHHHHIIH<<<HIIIIIIIIIIIIIIIII;;555CEIEEIIIII???
HHIIIIIIIIHHHIIIIIIIIIIIIHDCCIIIIIIIIIIIIIIIIHHHCCCCIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHHHIIIIHHHIIIDCCIIIIIIIIIIIIIIHH
HIIIIIIIIIIIIIIII>CC;;;;;HIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHHHIIIIHHHIII;;;IIAAAAIIII      AS:i:267      XS:i:0  XF:i:0
XE:i:6  XN:i:0
```

MOTIVATION

Geneticists prefer to store patients' aligned, raw genomic data (SAM files) because:

- Bioinformatic algorithms and sequencing platforms are still immature.
- Diseases might change the DNA sequence.
- The rapid evolution of genomic research.

Increasing number of medical units are willing to outsource the storage of genomes.

- Store while preserving the privacy of patients' genomes.
- Store while allowing the medical units to operate on the genome.

Medical tests on SAM files leak substantial privacy-sensitive information.

DISEASE TESTED	LEAKED SNP	NATURE OF THE LEAKED SNP
Alzheimer's Disease	'rs1799724'	Susceptibility to Vascular Dementia
	'rs6265'	Susceptibility to Memory Impairment
	'rs6265'	Body Mass Index
	'rs6265'	Smoking behavior
	'rs6265'	Weight
	'rs669'	Alpha-2-Macroglobulin Polymorphism
	'rs429358'	Stroke
	'rs429358'	Hyperlipoproteinemia type 3
	'rs429358'	Brain Imaging
	'rs4420638'	Total Cholesterol
	'rs4420638'	HDL Cholesterol
	'rs4420638'	LDL Cholesterol
	'rs4420638'	Longevity
	'rs4420638'	Coronary Artery Disease
SNP: Most common human genetic variation. Disease risk can be computed by analyzing particular SNPs.		

- Revelation of predisposition to diseases, ethnicity, paternity, etc.
- Genetic discrimination.
- Denial of access to health insurance, mortgage, education and employment.
- Revelation of information about family members.

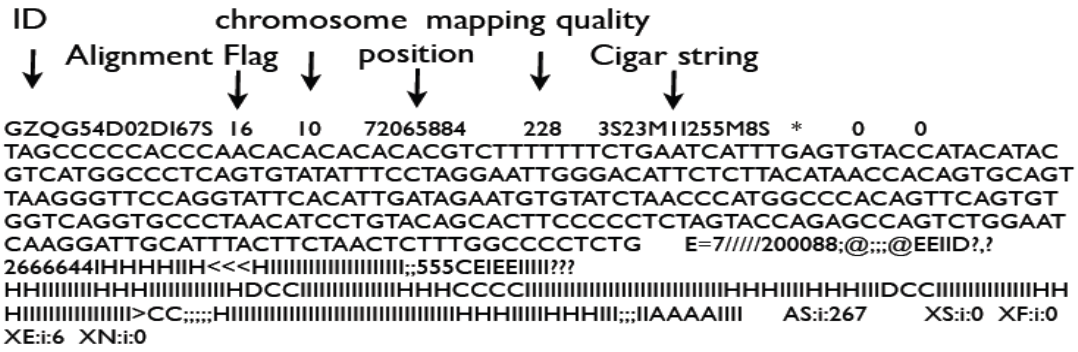
GENOMIC BACKGROUND

Sequence alignment/map (SAM) files are de facto standards used for all DNA sequence analyses.

SAM file of a patient contains hundreds of millions of short reads (SRs) randomly sampled from his genome.

Privacy-sensitive fields of a SR are:

- Its position with respect to the reference genome.
- Its cigar string (CS) expressing the variations in the content of a SR.
- Its content including the nucleotides from { A, T, C, G }.



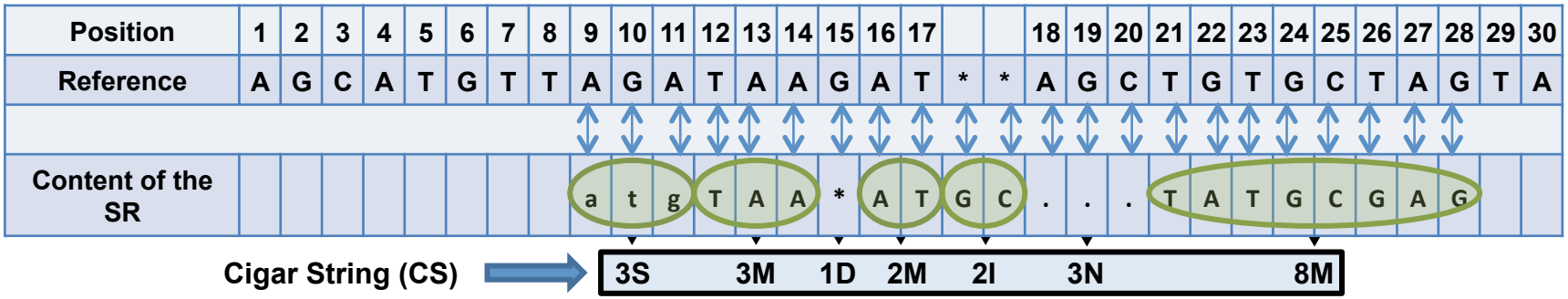
SHORT READ

POSITION	CIGAR STRING (CS)	CONTENT

The position of a short read is in the form $L\downarrow i,j = (x\downarrow i \mid y\downarrow j)$.

- $x\downarrow i$ is the chromosome number.
- $y\downarrow j$ is the position on the corresponding chromosome.

CS includes pairs of nucleotide lengths and the associated operations.



GOALS

Secure storage of the genomes at a biobank.

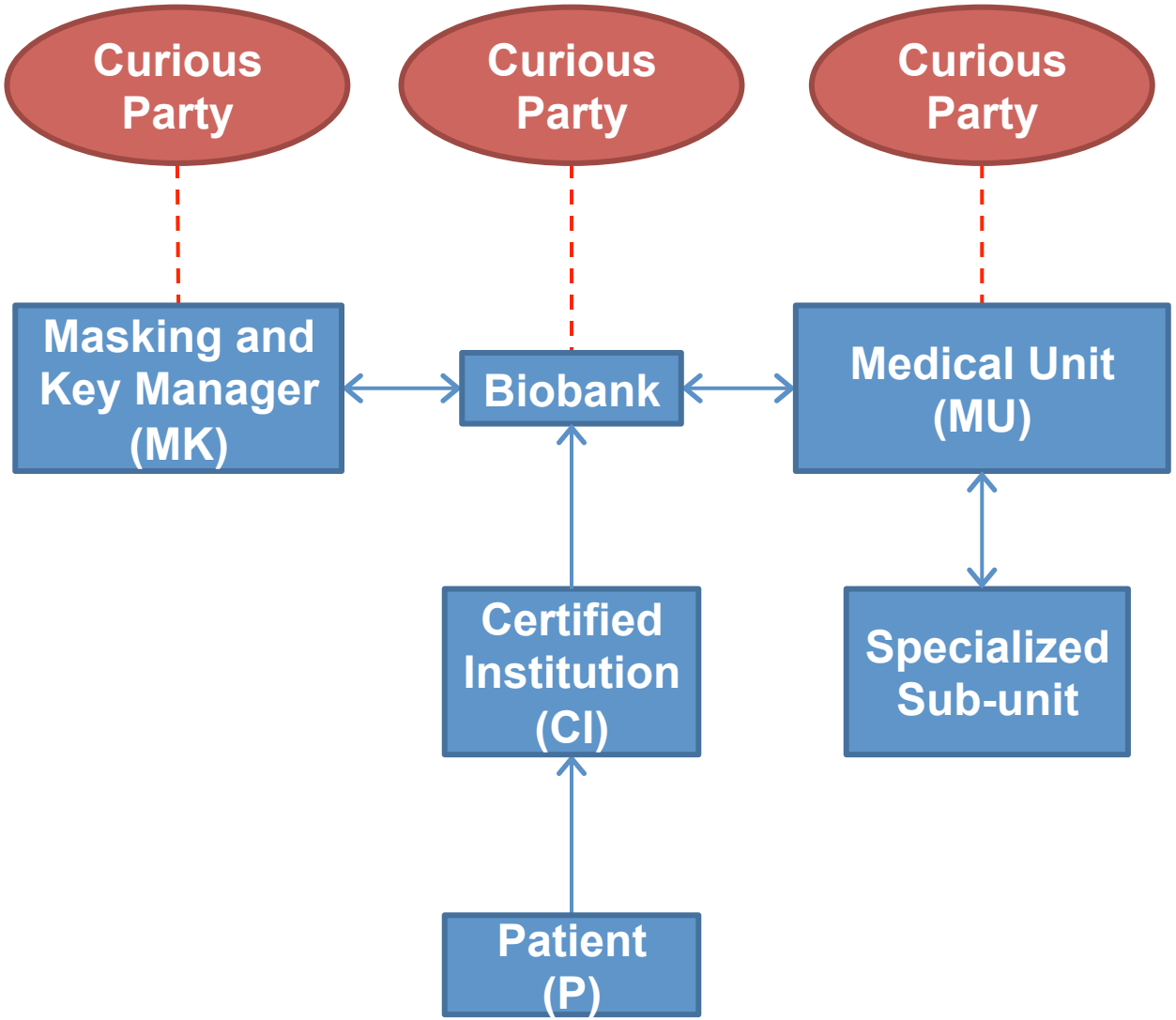
Privacy-preserving retrieval of encrypted short reads (in the SAM files) from the biobank.

- Biobank does not learn the positions of the requested short reads (hence the conducted genetic test).

Masking of the short reads at the biobank.

- Mask the parts of the requested short reads that are out of the requested (authorized) range.
- Mask the parts of the requested short reads for which the patient does not give consent.
    - Parts revealing sensitive diseases of the patient.

OVERVIEW OF THE SOLUTION



THREAT MODEL

A curious party at the biobank that can:

- Infer the genomic sequence of a patient from his stored genomic data.
- Associate the type of a genetic test with the patient being tested.

A curious party at the MK that can:

- Infer the genomic sequence of a patient from his stored cryptographic keys and the information provided by the biobank.
- Associate the type of genetic test with the patient being tested.

A curious party at the MU who tries to obtain the private genomic data of a patient for which it is not authorized.

All parties honestly follow the protocol.

Collusion is not addressed.

ENCRYPTION OVERVIEW

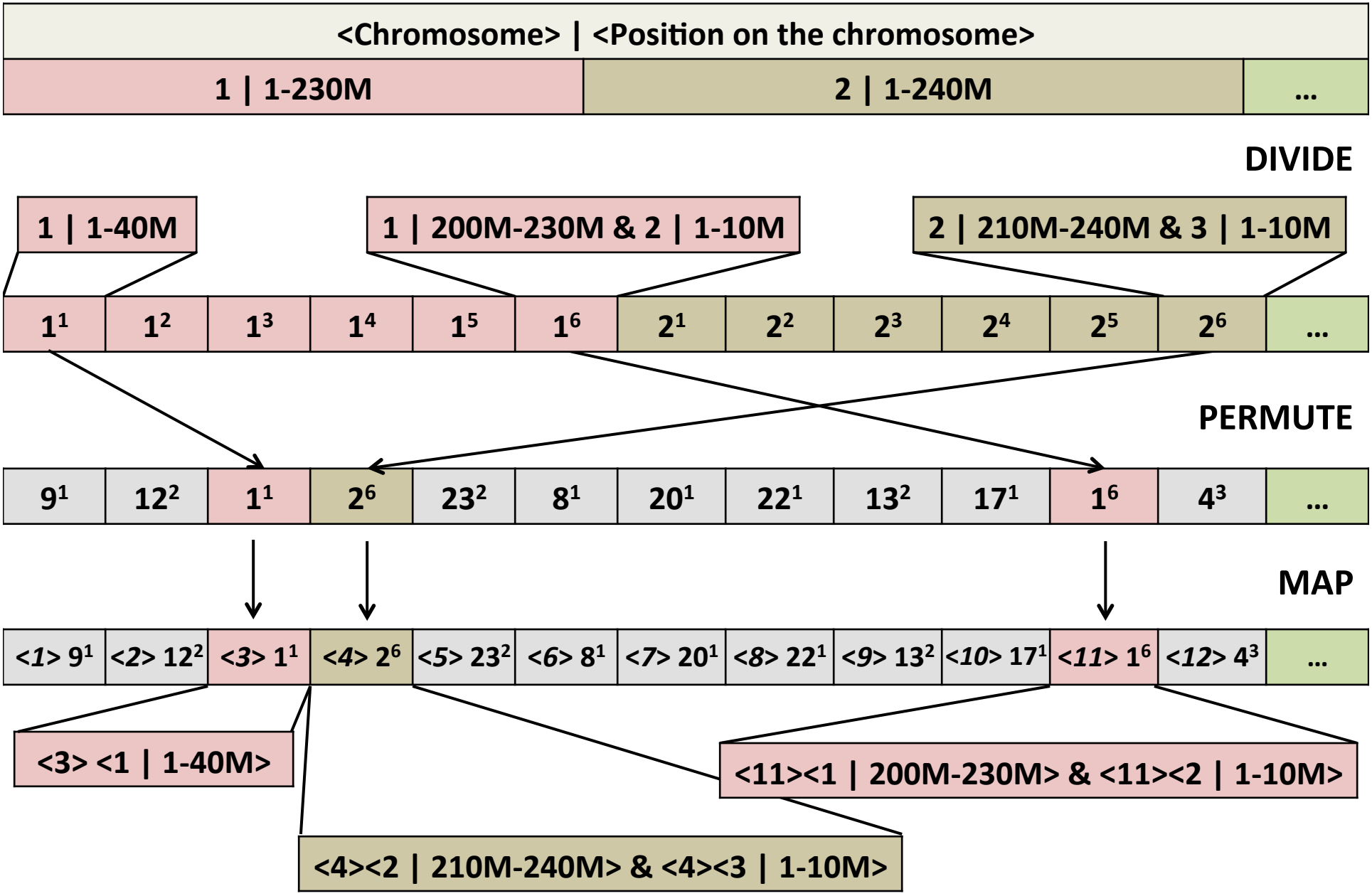
Cigar String is encrypted using secure symmetric encryption function.

Content of a short read is encrypted using Stream Cipher.

- Plaintext digits are combined with a pseudorandom cipher digit stream (keystream).

Position of a short read is encrypted using Order Preserving Encryption (OPE).

- M>N → E(M)>E(N).
- OPE can leak approximate positions of the short reads to the biobank.
- Permute and map the positions before encryption.

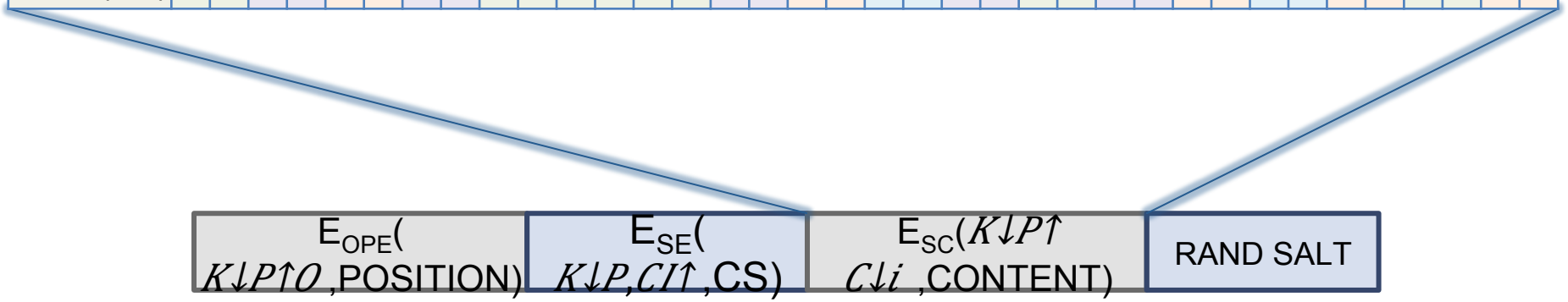

<Chromosome> | <Position on the chromosome>
1 | 1-230M
2 | 1-240M
...
DIVIDE
1 | 1-40M
1 | 200M-230M & 2 | 1-10M
2 | 210M-240M & 3 | 1-10M
1^1 1^2 1^3 1^4 1^5 1^6 2^1 2^2 2^3 2^4 2^5 2^6 ...
PERMUTE
9^1 12^2 1^1 2^6 23^2 8^1 20^1 22^1 13^2 17^1 1^6 4^3 ...
MAP
<1> 9^1 <2> 12^2 <3> 1^1 <4> 2^6 <5> 23^2 <6> 8^1 <7> 20^1 <8> 22^1 <9> 13^2 <10> 17^1 <11> 1^6 <12> 4^3 ...
<3> <1 | 1-40M>
<11><1 | 200M-230M> & <11><2 | 1-10M>
<4><2 | 210M-240M> & <4><3 | 1-10M>

ENCRYPTION

Nucleotide encoding	
A	00
T	01
C	10
G	11

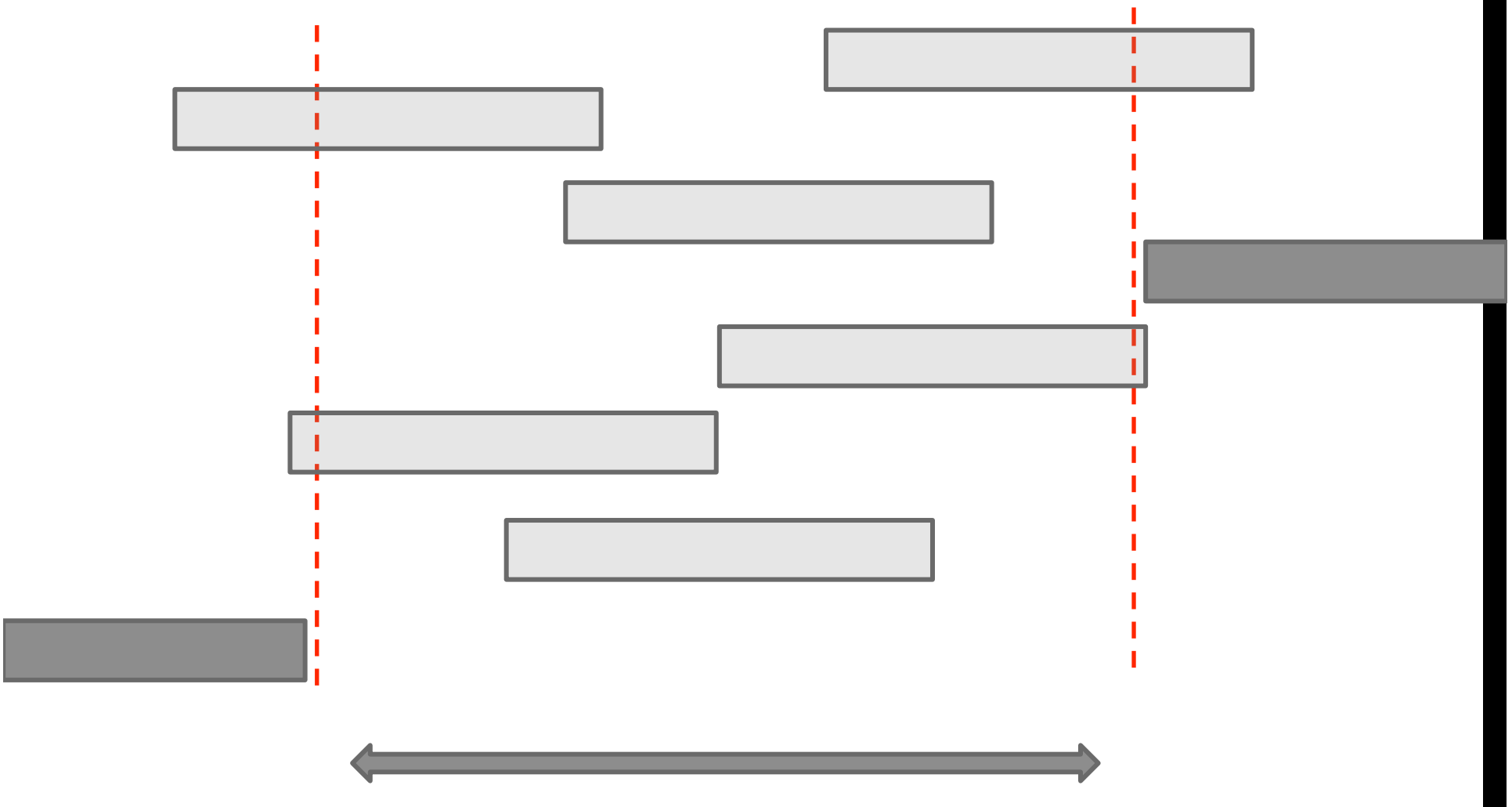
Position (on Ref.)	9	10	11	12	13	14	16	17	*	*	21	22	23	24	25	26	27	28
Content of SR in the SAM file	a	t	g	T	A	A	A	T	G	C	T	A	T	G	C	G	A	G

Plaintext content in binary	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	1	1	1	1	0	1	1	0	0	1	1
Key stream	1	0	0	0	1	1	0	0	1	0	0	1	0	0	0	1	1	1	1	0	0	1	1	0	1	1	1	0	0	1	0	0	1	1	0	0
Encrypted content (XOR)	1	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1



OPE: Order-preserving encryption
SE: Symmetric encryption
SC: Stream cipher

PROPOSED SOLUTION



Requested range by the medical unit

Masking and Key Manager (MK)

Biobank

Medical Unit (MU)

1) E[Requested range of nucleotides]

2) E[Requested range of nucleotides]

3) E[upper and lower bound of the range]

4) Private retrieval of the reads @ biobank

5) E[positions] and E[CSs] of short reads

6) Construction of the masking vectors @ MK

7) Masking request, E[CSs] E[positions] and E[decryption keys]

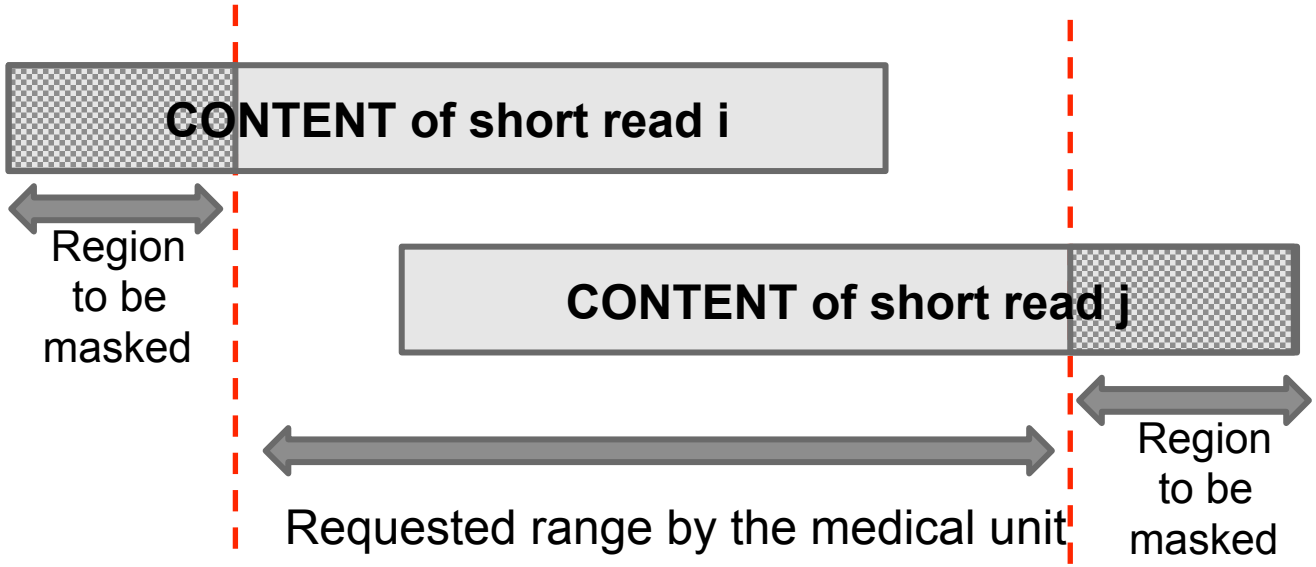
8) Masking @ biobank

9) E[masked short reads], E[modified CSs] E[positions] and E[decryption keys]

MASKING - I

Mask the parts of the requested short reads that are out of the requested (authorized) range.

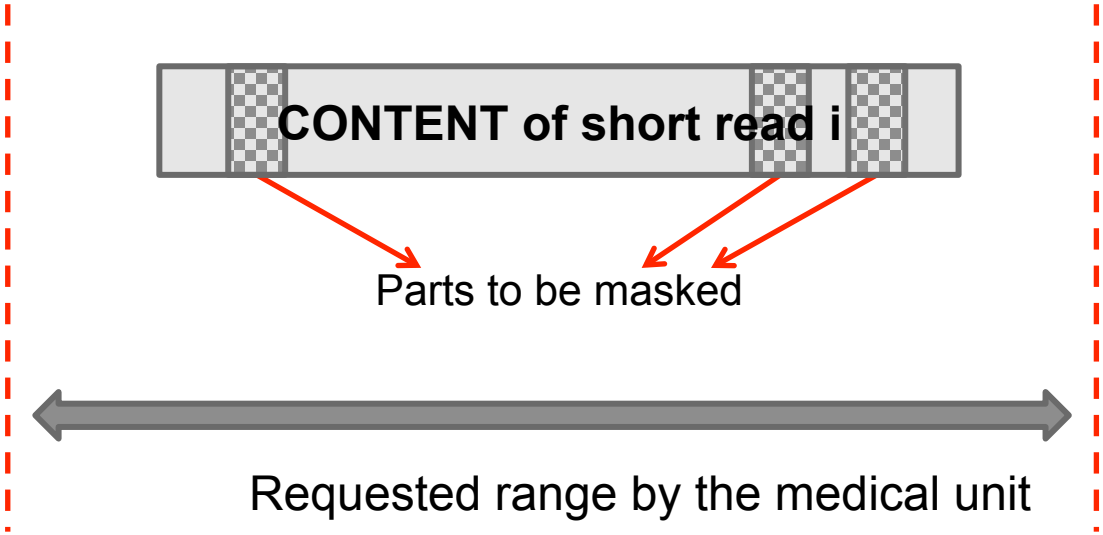
- Only provide the requested parts of the short reads to the medical unit.



MASKING - II

Mask the parts of the requested short read for which the patient does not give consent.

- Patient does not want to reveal his susceptibility for certain diseases to the medical unit.



Requested range of nucleotides	10-20
Non-consented positions	{3,5,11,17,21}

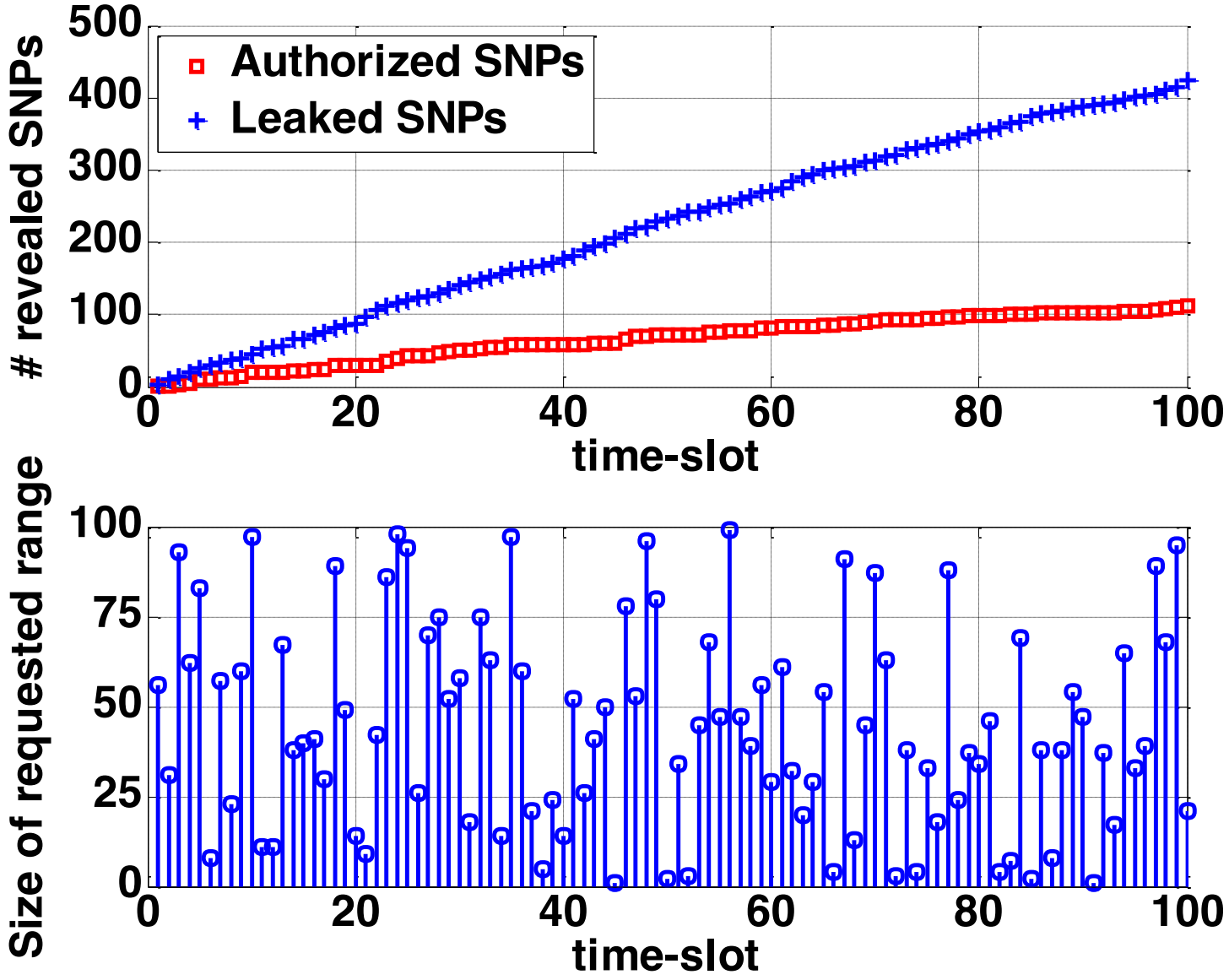
Encoding nucleotides	
A	00
T	01
C	10
G	11

Position (on Ref.)	9		10		11		12		13		14		16		17		*		*		21		22		23		24		25		26		27		28	
Content of SR in the SAM file	a		t		g		T		A		A		A		T		G		C		T		A		T		G		C		G		A		G	
Plaintext content in binary	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	1	1	1	1	0	1	1	0	0	1	1
Key stream	1	0	0	0	1	1	0	0	1	0	0	1	0	0	0	1	1	1	1	0	0	1	1	0	1	1	1	0	0	1	0	0	1	1	0	0
Encrypted content (XOR)	1	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1

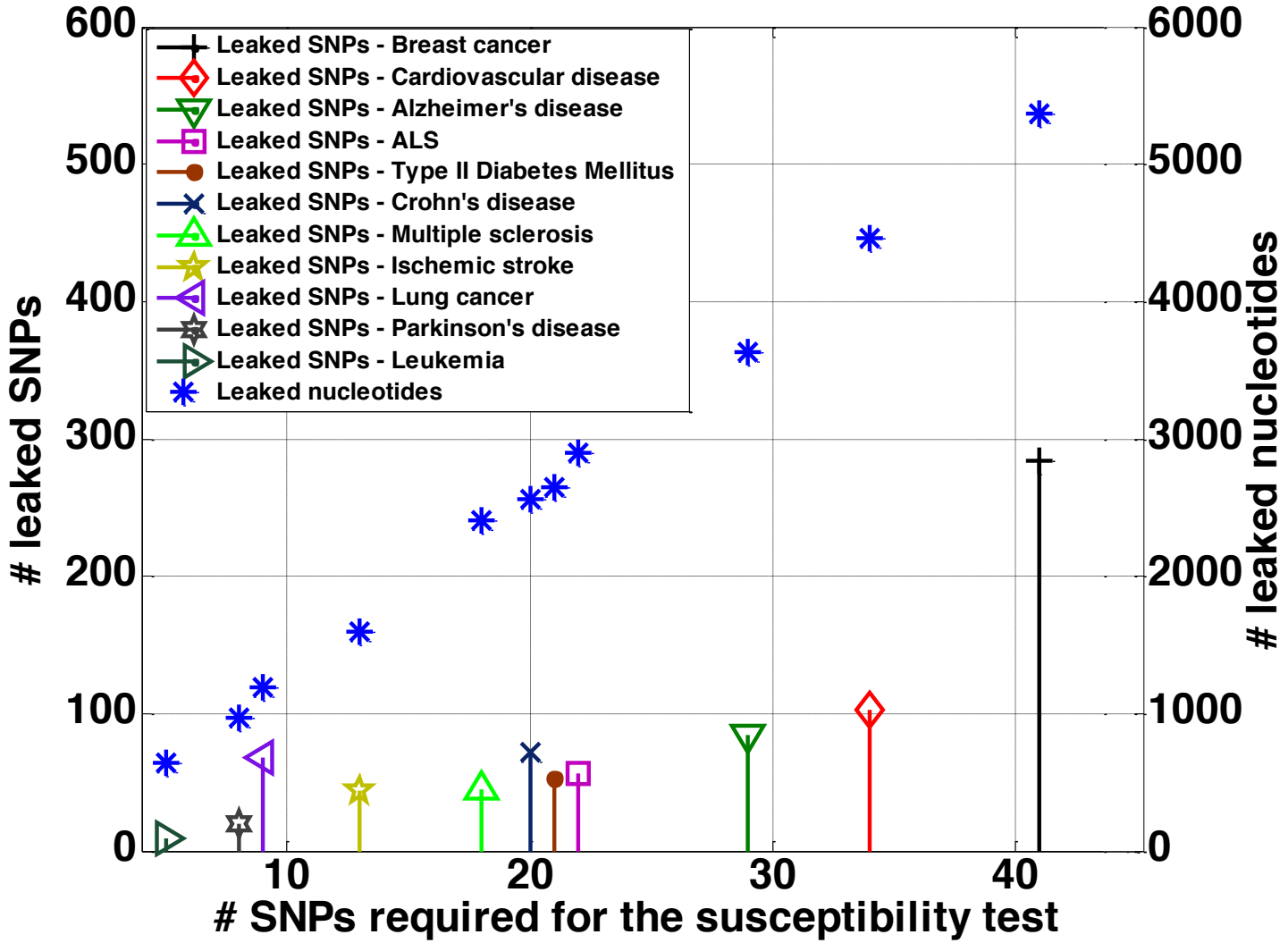
Masking vector	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Random masking string	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1	0	0	1	0	1	1
Masked enc. content (XOR)	1	1	1	1	1	1	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	0	1	0	0

Decrypted binary content (XOR)	0	1	1	1	1	0	0	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0
Decrypted nucleotides	T		G		C		T		A		A		A		G		G		C		T		G		A		T		G		G		C		A	

LEAKAGE OF SNPs WITH TIME



LEAKAGE OF SNPs DURING DIFFERENT DISEASE RISK TESTS



IMPLEMENTATION

Hardware/Software:

- Intel Core2 Duo CPU with dual-core 2.5 GHz
- Debian GNU/Linux 7.0 Operating System
- Java implementation
- MySQL 5.5 database server

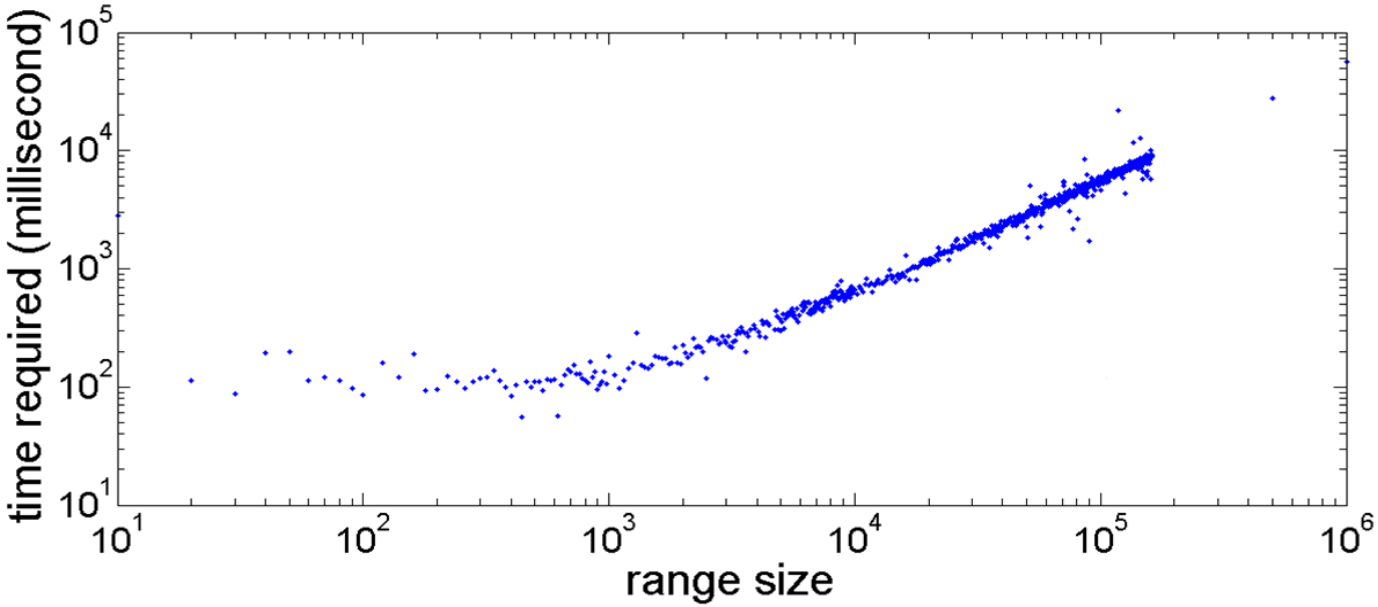
Cryptographic Parameters:

- Salsa20 Stream Cipher (64 bytes): CS + content
- OPE encryption: positions
- CCM mode of AES (256-bits): secure communication
- RSA (2048-bits): public key encryption

IMPLEMENTATION

Response time is almost linear with the requested range size

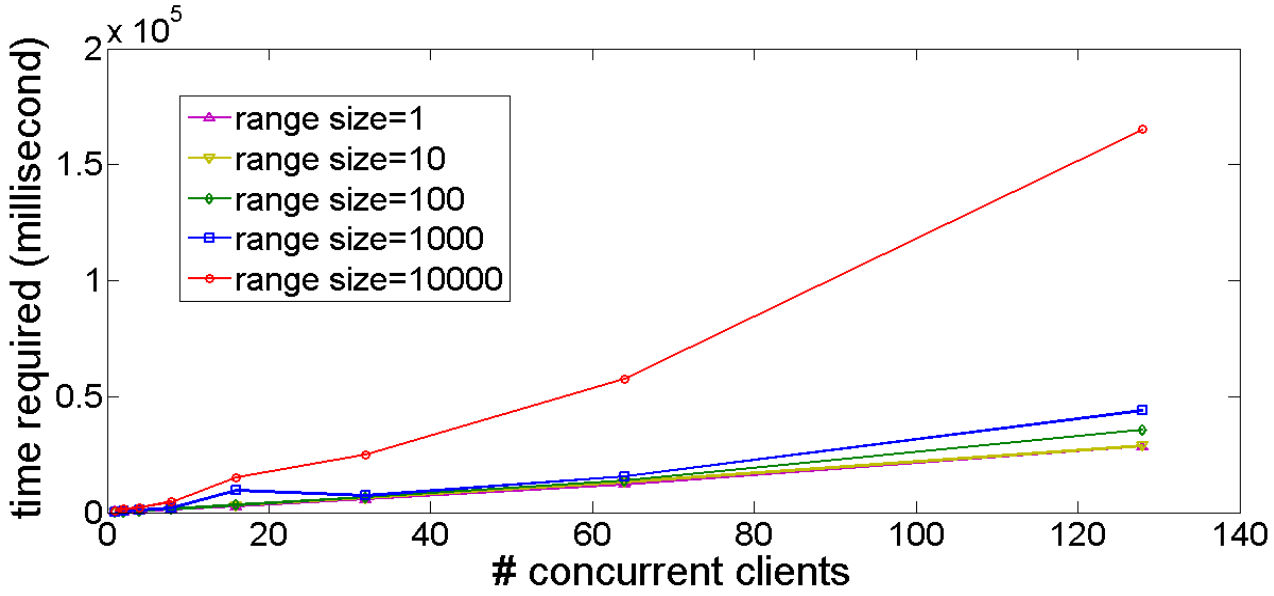
- For a range of 10 000 nucleotides (of one patient): 1 sec.



IMPLEMENTATION

Parallel requests:

- Requests from multiple medical units at the same time.
- For request size ≥ 10 000 the number of parallel clients highly effects the performance.
- The system can handle a maximum of 200 clients (MUs) for request size ≥ 100 000.



CONCLUSIONS

Secure storage of the genomes at a biobank.

Privacy-preserving retrieval of encrypted short reads (in the SAM files) from the biobank.

Efficient system for obfuscating specific parts of the encrypted short reads.

Evaluated the information leakage to the medical unit, with and without the masking is in place.

Implemented the proposed system and show its practicality.

QUESTIONS

erman.ayday@epfl.ch

http://lca.epfl.ch/projects/genomic-privacy/

Operation	Description
M	alignment match (can be a sequence match or mismatch)
I	insertion to the reference
D	deletion from the reference
N	skipped region from the reference
S	soft clipping (misalignment), clipped sequences (i.e., misaligned nucleotides) present in the content
H	hard clipping (misalignment), clipped sequences (i.e., misaligned nucleotides) NOT present in the content
P	padding (silent deletion from padded reference)

Encryption at the CI (Step 2)			**Request of nucleotides at the MU (Step 4)**	
OPE encryption: 7 ms/SR	SC encryption: 0.00048 ms/SR		RSA encryption: 0.216 ms	AES encryption: 0.064 ms
Private retrieval at the MK (Step 6)			**Private retrieval at the biobank (Step 7)**	
RSA decryption: 7.8 ms	AES decryption: 0.031 ms	2 x OPE encryption: 14 ms	Search and retrieve: 4.5 sec. (for a request size of 100)	
Constructing the masking vectors at the MK (Steps 9 and 10)				
OPE decryption: 7 ms/SR	SC decryption (for CS): 0.00048 ms/SR	Construct the masking vector: 0.016 ms/SR	Generate decryption keys for SC: 0.026 ms/SR	
Encrypt positions (using AES): 0.029 ms/SR	Encrypt CSs (using AES): 0.028 ms/SR	Encrypt the decryption keys: 0.030 ms/SR		
Masking at the biobank (Step 11)				
Masking: 0.015 ms/SR				
Decryption at the MU (after Step 12)				
AES decryption (for positions): 0.018 ms/SR	AES decryption (for CSs): 0.017 ms/SR	AES decryption (for decryption keys): 0.016 ms/SR	SC decryption (for the content): 0.00048 ms/SR	