HOMOMORPHIC ENCRYPTION BASED SECURE GENOME DATA ANALYSIS – A PERSONALIZED MEDICINE SOLUTION

WHEAT 2016| **Kalpana Singh**, Renaud Sirdey, David Cohen, François Artiguenave| 06/07/2016

Overview

- Basis of Genome and Sequencing
- Privacy Issues : Low Cost of Genome Sequencing
- Problem Setup
- Current Techniques
- Our Approach
- Experimental Results and Evaluation
- Summary of the Talk

Basis of Genome

- The complete blueprint of our body
  - Determines how we look, diseases we are susceptible to, our ancestry
- A complete set of instructions for life encoded in DNA.
  - Organized in chromosomes
- Helps in identifying criminals, early diagnosis of diseases, and enable personalized medicine and prenatal testing
- GOLD Website: Listing of finished and "in progress" genomes
  http://www.genomesonline.org/

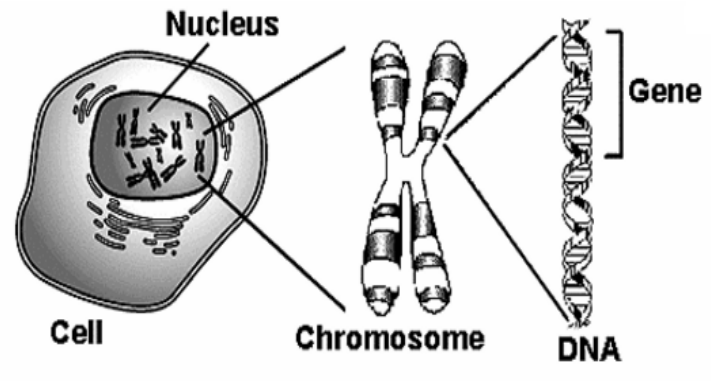


Figure 1. Genome Structure

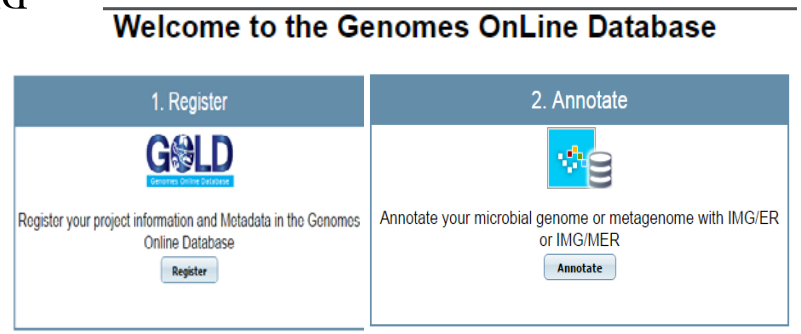


Figure 2. Genomes OnLine Database

Why Sequence a Genome?

- To understand how the entire cell/organism works
  - thousands of complex gene interactions
- Getting Sequences: http://www.ncbi.nlm.nih.gov/
- ICGC: sequencing instruments are providing new opportunities for comprehensive analyses of cancer genomes
  - For example: Identify changes in the genomes of tumors that drive cancer progression

Figure 3. One Sequence "read" from the Human Genome

Genome Sequencing is Excitement

Matthew Herper
Forbes Staff

FOLLOW

I cover science and medicine, and believe this is biology's century.
full bio →

1/05/2011 @ 4:57PM | 30,076 views

The First Child Saved By DNA Sequencing

+ Comment Now + Follow Comments

Genetic Gamble

New Approaches to Fighting Cancer

PART ONE
A Race to Leukemia's Source

PART TWO
Promise and Heartbreak

In Treatment for Leukemia, Glimpses of the Future

LETTER

doi:10.1038/nature13394

Genome sequencing identifies major causes of severe intellectual disability

Christian Gilissen[1]*, Jayne Y. Hehir-Kwa[1]*, Djie Tjwan Thung[1], Maartje van de Vorst[1], Bregje W. M. van Bon[1], Marjolein H. Willemsen[1], Michael Kwint[1], Irene M. Janssen[1], Alexander Hoischen[1], Annette Schenck[1], Richard Leach[2], Robert Klein[2], Rick Tearle[2], Tan Bo[1,3], Rolph Pfundt[1], Helger G. Yntema[1], Bert B. A. de Vries[1], Tjitske Kleefstra[1], Han G. Brunner[1,4]*, Lisenka E. L. M. Vissers[1]* & Joris A. Veltman[1,4]*

Data Breaches and Privacy

Security experts warn of increasing data breaches and privacy risks

LivingSocial Hacked — More Than 50 Million Customer Names, Emails, Birthdates and Encrypted Passwords Accessed (Intern

APRIL 26, 2013 AT 1:15 PM PT

LivingSocial, the daily deals site owned in part by Amazon, has

gents were

Suggested Content

Some Victims of Online Hacking Edge Into the Light

Compromise of Confidential data is prevalent

WordPress firm Automattic suffers root-level hack

ouldn't Trust Facebook with n Employee's Revelations

Privacy, security still top cloud concerns

Asia Cloud Forum editors | November 13, 2013

Asia Cloud Forum

An online survey of Microsoft partners has revealed that traditional concerns about cloud services remain among enterprises in the region.

Lack of Privacy : Low Cost of Genome sequencing

- Fast drop in the cost of genome-sequencing
  - 2001: $3 billion
  - 2015: $1,000
  - Genotyping 1M variations: below $200

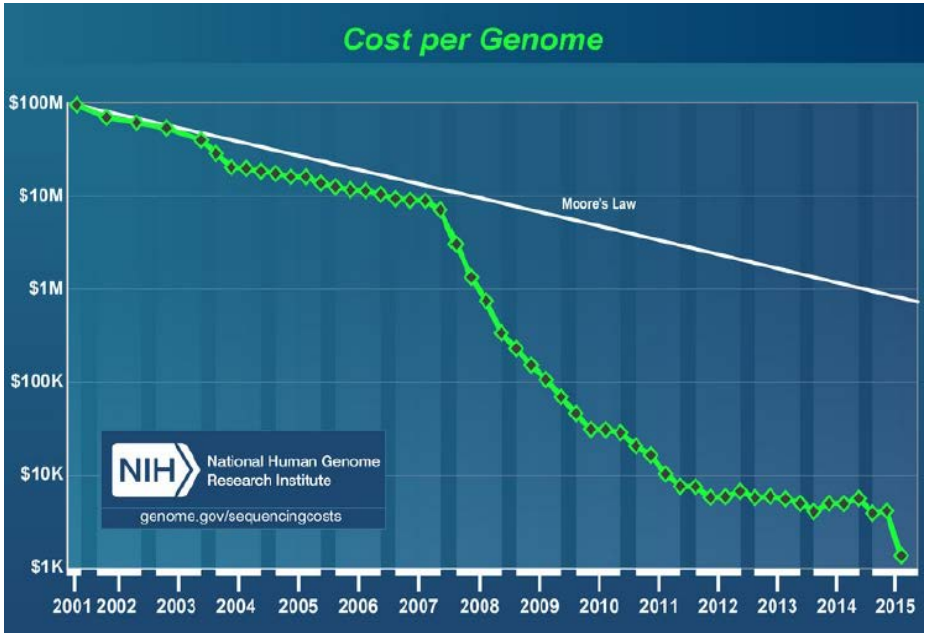


Figure 4. Cost per Genome [Source:National Human Genome Research Institute (NHGRI)]

Applications of Genomics

- Unleashing the potential of the technology
  - Healthcare: e.g., disease risk detection, personalized medicine
  - Biomedical research: e.g., geno-phono association
  - Legal and forensic
  - DTC: e.g., ancestry test, paternity test
  - ……

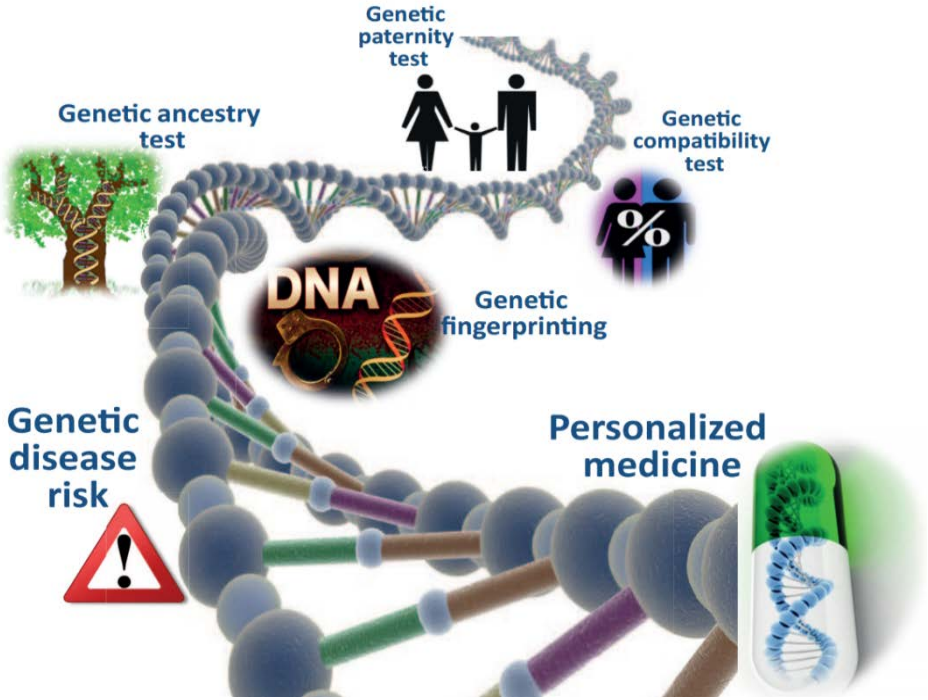
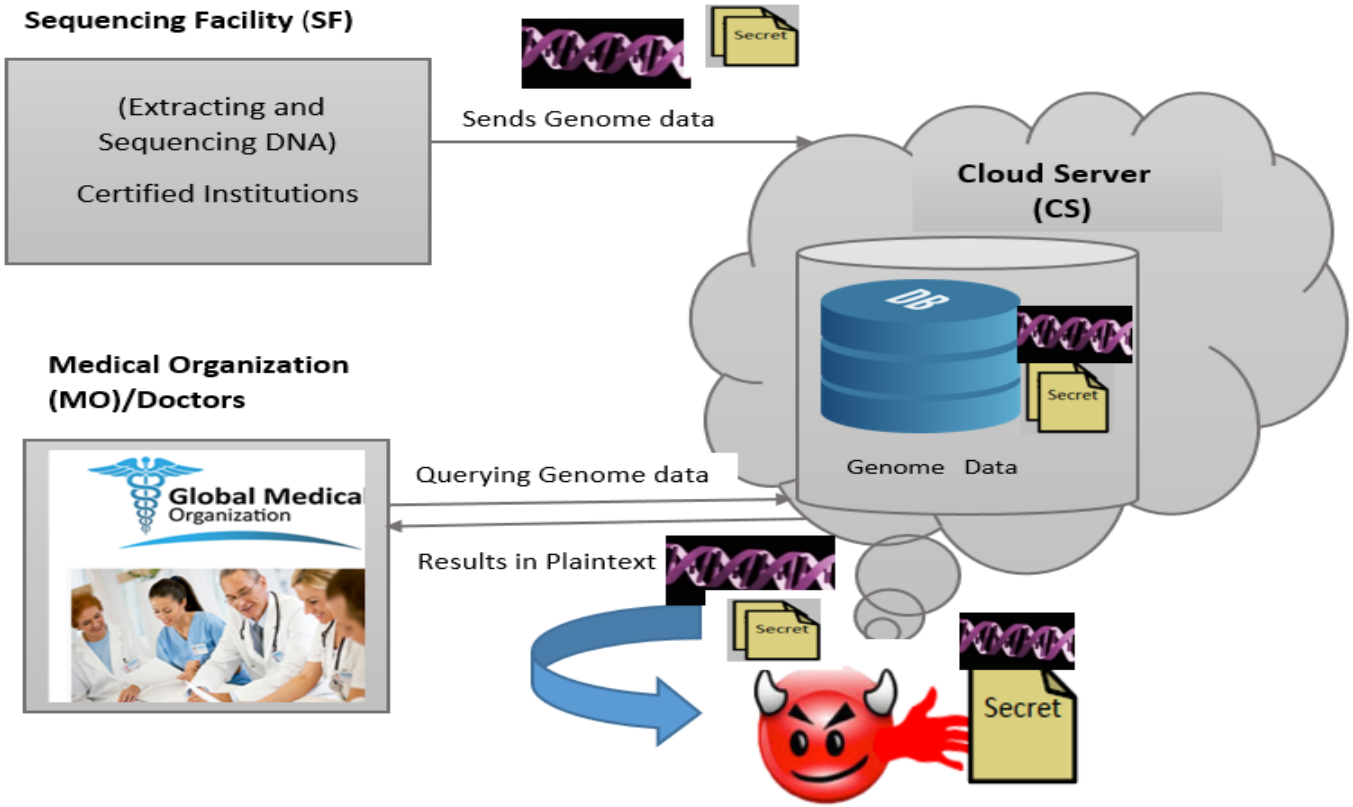


Figure 5. Applications of Genomics

Problem Setup



no computation

storage

encryption

✓

Computation

databases, web applications, mobile applications, machine learning, etc.

???

CURRENT WORK

- Privacy-Preserving Genomic Computation Techniques
  - Anonymization, which has proven to be ineffective for genomic data [HSR+08a], [WLW+09]

TECH 4/25/2013 @ 3:47PM | 17,111 views

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study

+ Comment Now + Follow Comments

A Harvard professor has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.

From the onset, the Personal Genome Project,

Harvard Professor Latanya Sweeney

[Melissa Gymrek et al. *"Identifying Personal Genomes by Surname Inference."* Science Vol. 339, No. 6117, 2013]

Privacy- Preserving Techniques

- Privacy-Preserving Genomic Computation Techniques

  - Differential privacy [FSU11], [JS13], [YFSU14]

Drawbacks

- Low utility
- Impractical: needs a large amount of noise in order to satisfy the differential privacy for a small number of SNPs [YFSU14].

  - Computation partitioning [WWL+09a]

Drawbacks

- Very small portion of human genome contains sensitive information.

Cryptography based Techniques

- Cryptography based Techniques (homomorphic encryption [ARHR13], [KJLM08])
  - Fully Homomorphic Encryption (FHE) breakthrough approach: Gentry's scheme
  - Practical homomorphic encryption technique: Some what Homomorphic Encryption (SWHE)

Cryptography Techniques are Useful in Genome Data Privacy [TPKC07], [BKKT08], [NLV11], [BBC+11a], [ARRH13], [ARH+14]

Practical Use of Homomorphic Encryption

GUILLAUME PAUMIER/WIKIMEDIA

Genetic gold. Each spot in a DNA microarray, such as this one, contains large amounts of sensitive genetic information.

How to Hide Your Genome

By Thomas Sumner | 16 February 2014 6:45 pm | 4 Comments

Thomas is a news intern at *Science*.

Email Thomas

Follow @SumnerSci

CHICAGO, ILLINOIS—As the cost of genetic sequencing plummets, experts believe our genomes will help doctors detect diseases and save lives. But not all of us are comfortable releasing our biological blueprints into the world. Now, cryptologists are perfecting a new privacy tool that turns genetic information into a secure yet functional format. Called homomorphic encryption and presented here today at the annual meeting of AAAS, which publishes *Science,* the method could help keep genomes private even as genetic testing shifts to cheap online cloud services.

Existing encryption techniques make data secure at the expense of making it unusable. Because of this, most genetic sequences are simply anonymized before being sent out for analysis. However, computational biologist Yaniv Erlich at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, told meeting attendees that with a little genetic sleuthing, this supposedly anonymous data can easily track back to its owner. Erlich says he positively matched 12% of male genomes with the exact person they originated from.

In 2009, the first lattice-based cryptography scheme was announced by IBM. The geometry-based encryption method allows data to be manipulated through both multiplication and addition while remaining encrypted. Researchers realized that the complex algorithms used during genetic tests could be closely approximated by the two basic mathematical operations. Lattice cryptology enabled homomorphic encryption, allowing computers to analyze encrypted data and return encrypted results without ever being able to decode the information. Cryptologist Kristin Lauter, research manager for the cryptography group at Microsoft Research in Redmond, Washington, likened the method to locking a gold brick in a

A CIPHER FOR YOUR GENOME
By CRG staff - interview with Kristin Lauter

from *GeneWatch* 27-1 | Jan-Apr 2014

***Kristin Lauter**, PhD, is a Principal Researcher and Research Manager for the Cryptography group at Microsoft Research. She has been working on practical homomorphic encryption for several years and was a coauthor of the breakthrough paper "Can Homomorphic Encryption be Practical?"*

***GeneWatch:* How is homomorphic encryption different from other encryption technologies?**

Kristin Lauter: The primary new functionality enabled with homomorphic encryption is the ability to compute on encrypted data. This is very important for things like outsourcing storage and computation of data. The idea is that when using homomorphic encryption, the data owner - let's say it's a consumer or an enterprise - could encrypt the data locally and keep the key. Then they can upload that data to the cloud, and if they used homomorphic encryption, that data can still be operated on by the cloud and the encrypted results are available from the cloud to the data owner or anyone the data owner trusts to share the encryption key with. So it really allows a whole new functionality on encrypted data.

The problem with many other types of encryption is that it makes data secure at the expense of making it unusable. Can you say anything more about what that means?

With standard encryption systems, after you encrypt the data there is very little ability to do anything with it. For example, AES is the government's standardized block cipher. When you encrypt something with AES, you should not be able to distinguish anything about the original data or operate on it in any way which gives meaningful results. In the last ten years or so there has been a push in the field of cryptographic research to invent techniques that allow you to encrypt data and maintain its privacy but still get some functionality out of it. Homomorphic encryption is a very general and powerful tool to allow computation on encrypted data.

Homomorphic Encryption Techniques

Much progress since Gentry's first scheme [Gentry09] with unlimited additions and multiplications

- Small Principal Ideal Problem (SPIP)
  - Gen'09, SV'10, GH'11
- Approximate GCD
  - vDGHV'10, CMNT'11, CNT'12, CCKLLTY'13
- LWE/RLWE
  - BV'11a, BV'11b, BGV'12, GHS'12,LTV'12, Bra'12, FV'12, BLLN'13

- **Helib (IBM) publically available implementation**

Computation on Encrypted Genomic Data

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

Archive > Volume 519 > Issue 7544 > News > Article

NATURE | NEWS

Extreme cryptography paves way to personalized medicine

Encrypted analysis of data in the cloud would allow secure access to sensitive information.

Erika Check Hayden

23 March 2015

PDF | Rights & Permissions

David Paul Morris/Bloomberg via Getty

Cloud processing of DNA sequence data promises to speed up discovery of disease-linked gene variants.

genomeweb

Business & Policy | Technology | Research | Clinical | Disease Areas | Applied Markets | Resources

Home » Tools & Technology » Informatics » New Community Challenge Seeks to Evaluate Methods of Computing on Encrypted Genomic Da

New Community Challenge Seeks to Evaluate Methods of Computing on Encrypted Genomic Data

Nov 14, 2014 | Uduak Grace Thomas

Premium

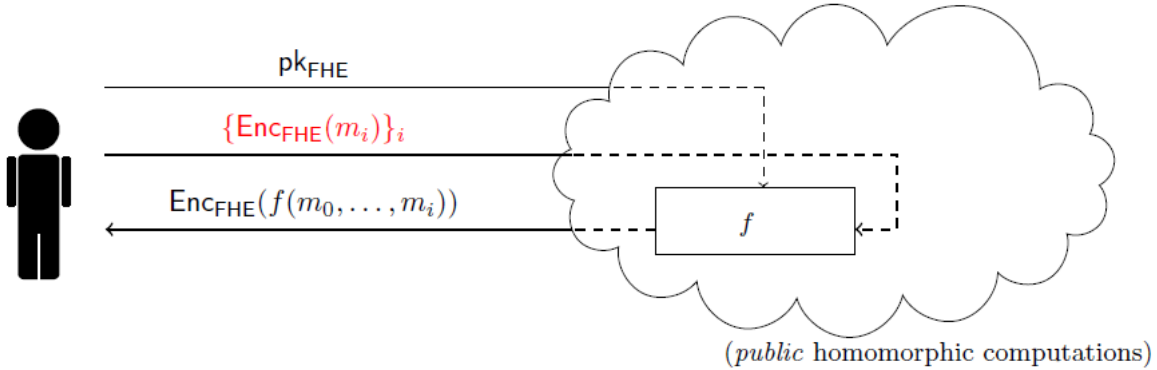
NEW YORK (GenomeWeb) – Researchers from academia and industry have launched the second iteration of a community challenge that aims to evaluate the performance of methods of computing securely on genomic data in remote environments like the cloud.

The challenge, which focuses on methods of computing on encrypted data, is organized by researchers from Indiana University, the University of California at San Diego, Emory University, Vanderbilt University, and La Jolla, Calif.-based Human Longevity. It is run under the auspices of the Integrating Data for Analysis, Anonymization, and Sharing (IDASH) center at UC San Diego — IDASH is one of the National Institutes of Health's National Centers for Biomedical Computing. The organizers planned and ran the first iteration of the challenge earlier this year and have submitted a paper for publication in *BMC Medical Informatics & Decision Making* that describes the challenge and results in detail.

For the second contest, dubbed the Secure Genome Analysis competition, the organizers have proposed two challenges. The first is called the secure genome-wide association study and it has two sub-challenges that deal with homomorphic encryption — a method of encoding data as ciphertext that allows specific computations to be run on it — and secure multiparty computing among multiple institutions.

In the first subtask, participating teams will receive two sets of genotypes — one for cases and the other for controls — over a few SNPS, and they will be expected to develop a homomorphic encryption protocol to encrypt the input datasets. The protocol should be able to move encrypted datasets to an untrusted remote server, compute the minor allele frequencies and chi-squared statistics for a given set of SNPs between the case and control groups, and decrypt the results using a privately held key. The algorithms will be tested on a single server and the performance will be measured in terms of computation time, space, and overhead.

Communication with the Cloud



The cipher text expansion is huge (prohibitive)!

If data is a 4MB image, using [GHS12,CCKLLTY13], the user would have to send around 200/300GB of encrypted data

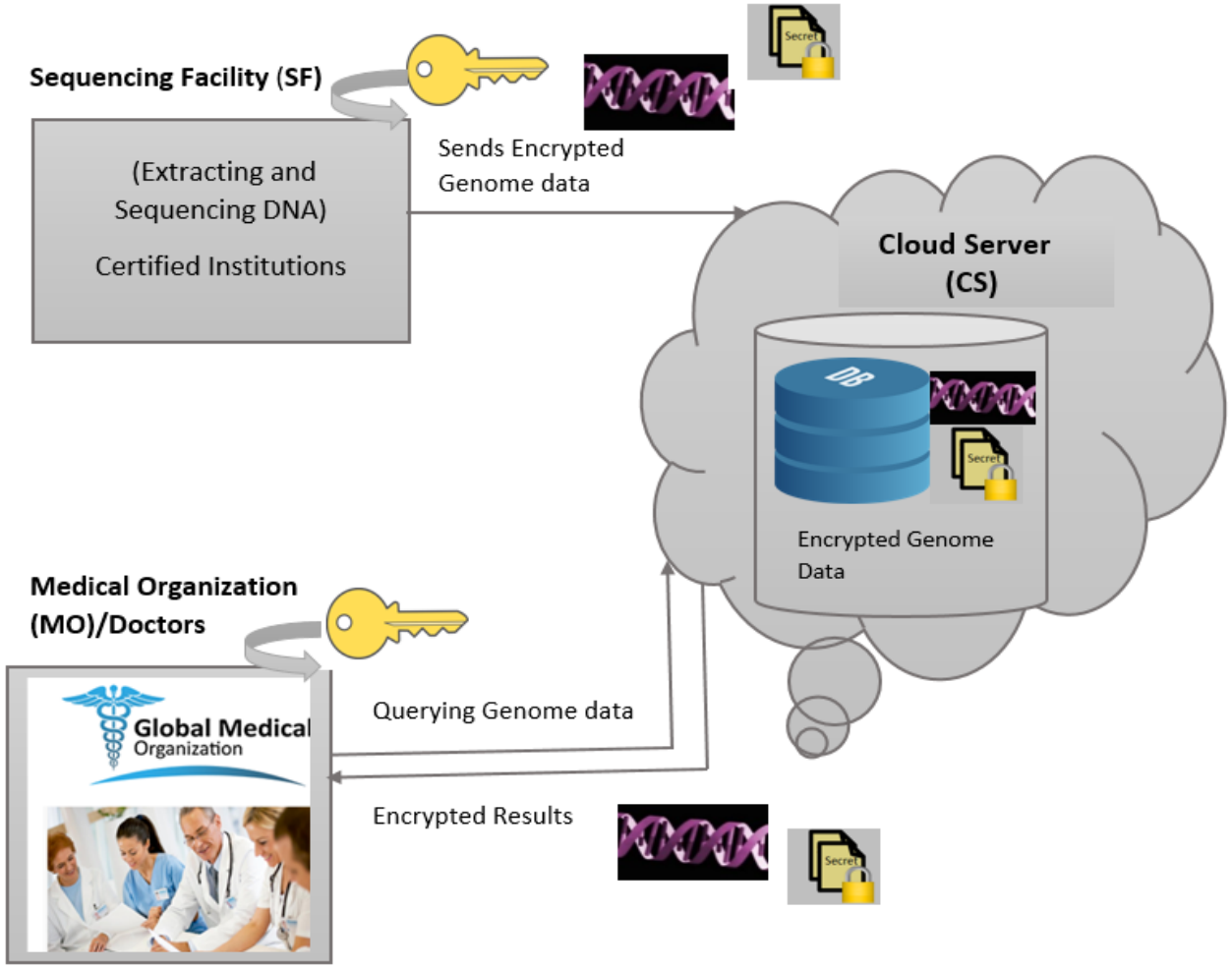
- Common facts of the current FHE schemes
  - Noise grows with operations
  - Multiplication yields the most noise
  - Decryption will fail with too large noise
  - Large overhead in storage space and computation time

Our Work

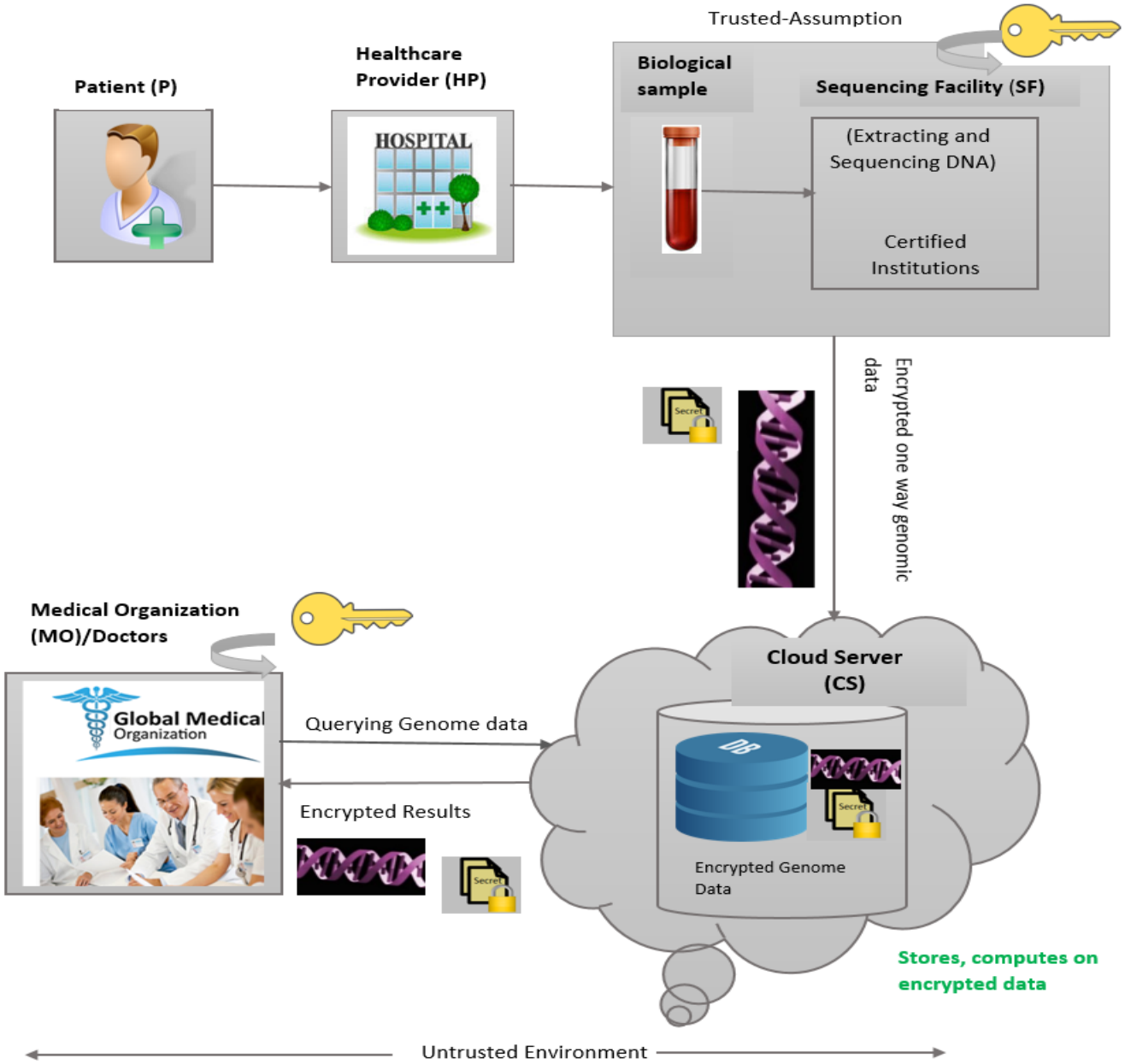
Our work: try to demonstrates that these difficulties are manageable in practical systems (non prohibitive computation cost, + decent security)

Secure Architecture

- Servers store, process, and compute on encrypted data in a practical way.



Our Architecture



GENOME DATA ANALYSIS – Detection of ABO Blood Type

Experimental Datasets

- Blood Group Antigen Gene Mutation Database (BGMUT) is an online repository of allelic variations in genes
  - Determine the antigens of various human blood group systems
- Part of the dbRBC resource of the National Center for Biotechnology Information, USA, and is available online at http://www.ncbi.nlm.nih.gov/projects/gv/rbc/xslcgi.fcgi?cmd=bgmut [the journal: Nucleic acids research, 2012.40:D1023-9]
- Documents sequence variations of a total of 1251 alleles of all 40 gene that together are known to affect antigens of 30 human blood group systems
  - Tested for 2504 patients
  - Each patient has 893 records
  - Each record has 6 columns

Chrom	POS	REF	ALT	Values 1	Values 2
9	136125819	C	T	0	0

Figure 6. Descriptive Example of Datasets

Why FHE and Personalized Genomics is not Hopeless?

- Genomics data is relatively huge (~ 4 millions variants per individuals).
  - Still, transciphering will help.
- But, typical individual-centric diagnostics involve only very small part of these data (and manageable volumes of computations)
- For example:
  - Basic ABO (8 binary values, 44 ANDs, degree 7)
  - ABO-2 (34 binary values, 252 ANDs, degree 34)
  - Extended ABO-2 (1133 binary values, 2512 ANDs, degree 1157)
  - HLA (4544 binary values, 6064 ANDs, degree 4635)

ABO Rule Implementation

- Divide each patient detail according to their IDs

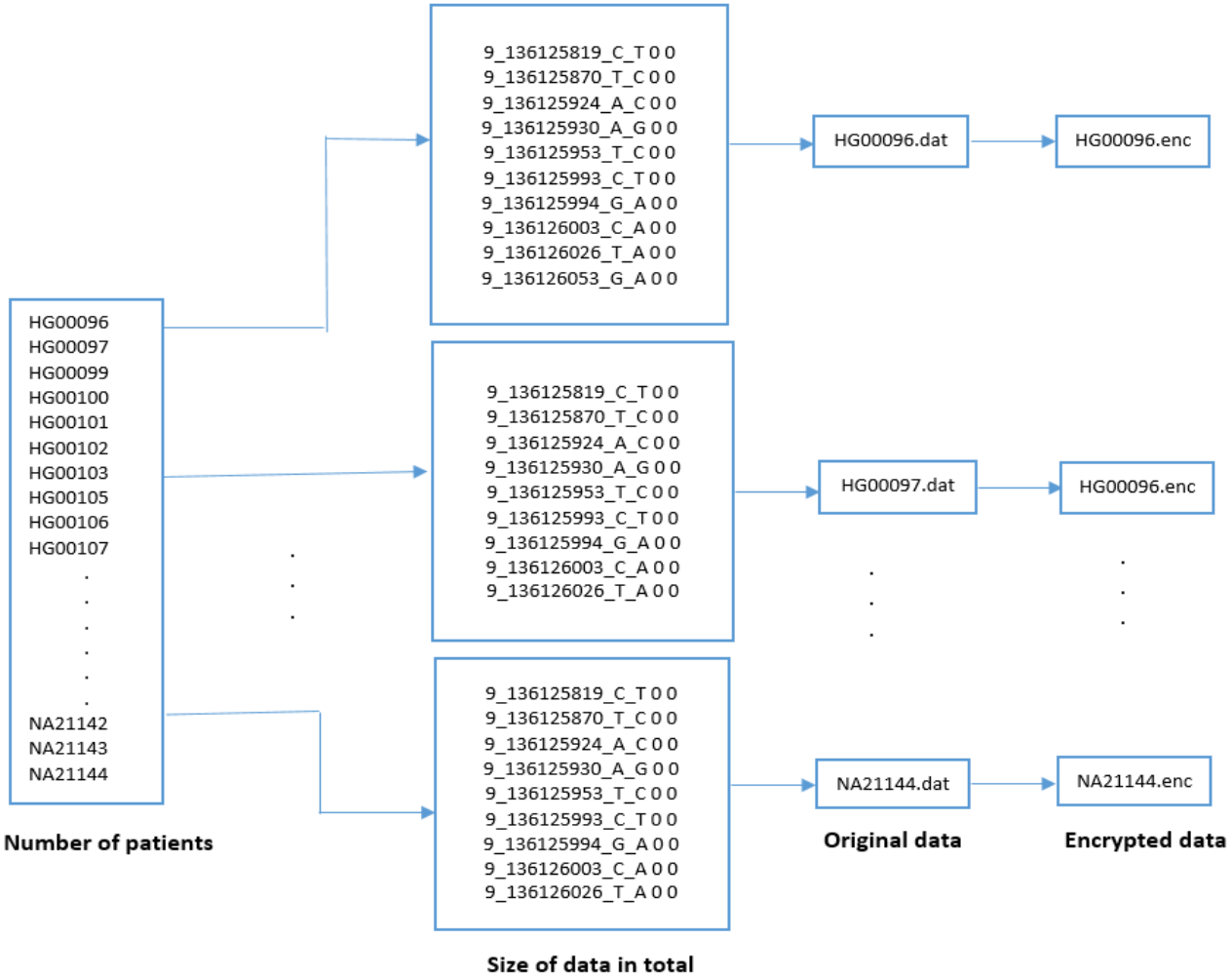


Figure 7. Division of Datasets

ABO Rules for Evaluation

Basic ABO Blood Group Types

ABO b101, ABO b1b1, ABO a2b1, ABO a101, ABO o103, ABO a1a1, ABO a103, ABO 0303, ABO a1b1, ABO b103, ABO a1a2, ABO a203

ABO – 2 Blood Group Types

O Haplotype, B Haplotype, T Haplotype

//Extended rules to decrease FN and FP

O Haplotype, B Haplotype, T Haplotype

//Simple rules for determining phenotype

O Phenotype, T Phenotype (TT Genotype and TO Genotype), B Phenotype (BB Genotype and BO Genotype), TB Phenotype,

// extended rules

O Phenotype rules with Genotype, B Phenotype rules for Genotype, T Phenotype rules for Genotype, TB Phenotype w, corresponding rules for Genotype

Rules are generated by Centre National de Génotypage (CNG) Geneticists

Basic ABO Rule

Phenotype	Rule
O1O1	(9:136132908;T;TC;0|0)
A2O1	(9:136132908;T;TC;0|1) & (9:136131651;G;A;0|1)
A2A2	(9:136132908;T;TC;1|1) & (9:136131651;G;A;1|1)
B1O1	(9:136132908;T;TC;0|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;0|1)
B1B1	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;1|1)
A2B1	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|1) & (9:136131461;G;A;0|1)
A1O1	(9:136132908;T;TC;0|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;0|0) & (9:136131316;C;T;0|0)
O1O3	(9:136132908;T;TC;0|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;0|0) & (9:136131316;C;T;0|1)
A1A1	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;0|0) & (9:136131316;C;T;0|0)
A1O3	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;0|0) & (9:136131316;C;T;0|1)
O3O3	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;0|0) & (9:136131316;C;T;1|1)
A1B1	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;0|1) & (9:136131316;C;T;0|0)
B1O3	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|0) & (9:136131461;G;A;0|1) & (9:136131316;C;T;0|1)
A1A2	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|1) & (9:136131461;G;A;0|0) & (9:136131316;C;T;0|0)
A2O3	(9:136132908;T;TC;1|1) & (9:136131651;G;A;0|1) & (9:136131461;G;A;0|0) & (9:136131316;C;T;0|1)

Rules are generated by Centre National de Génotypage (CNG) Geneticists

ABO Rule 2: O Haplotype

```
For O haplotype
(9:136132908;T;TC;0) | (9:136133506;G;A;1) & (9:136131316;C;T;1)

in the BCMUA database
False positive: 1
TwO8 !{(9:136137553;C;A;0) & (9:136131630;G;A;1)}
False negative: 13
O08 (9:136132908;T;TC;1) & (9:136133506;G;A;0) & (9:136131316;C;T;0) & (9:136131314;C;CC;1) & (9:136131059;GG;G;1),
O14 (9:136131225;G;A;1),
O15 (9:136131191;G;T;1),
O19 (9:136132908;T;TC;1) & (9:136133506;G;A;0) & (9:136131316;C;T;0) & (9:136132873;T;C;0) & (9:136131651;G;A;0) & (9:136131472;A;T;1) &
(9:136131437;C;T;1) & (9:136131846;T;C;0),
O20 (9:136132908;T;TC;1) & (9:136133506;G;A;0) & (9:136131316;C;T;0) & (9:136132873;T;C;1) & (9:136131472;A;T;1) & (9:136131895;C;T;1),
O51 (9:136137513;C;CC;1),
O52 (9:136132848;G;A;1),
O53 (9:136131576;C;T;1),
O72 (9:136137532;A;AC;1),
O74 (9:136131611;CTGC;C;1),
O77 (9:136131555;C;T;1),
O78 (9:136131666;A;C;1),
O79 (9:136131483;A;T;1)
Error rate: 3.9%
```

Rules are generated by Centre National de Génotypage (CNG) Geneticists

Test Results

Test Results – ABO Rule -1

L > = 7

Rule Results	Time in Seconds
Decrypted Value for rule Abo_O101 is: 0 Decrypted Value for rule Abo_a201 is: 0 Decrypted Value for rule Abo_a2a2 is: 0 Decrypted Value for rule Abo_b101 is: 1 Decrypted Value for rule Abo_b1b1 is: 0 Decrypted Value for rule Abo_a2b1 is: 0 Decrypted Value for rule Abo_a101 is: 0 Decrypted Value for rule Abo_o103 is: 0 Decrypted Value for rule Abo_a1a1 is: 0 Decrypted Value for rule Abo_a103 is: 0 Decrypted Value for rule Abo_0303 is: 0 Decrypted Value for rule Abo_a1b1 is: 0 Decrypted Value for rule Abo_b103 is: 0 Decrypted Value for rule Abo_a1a2 is: 0 Decrypted Value for rule Abo_a203 is: 0	Computation = 4 Encryption = 0 Decryption = 2

- Setup is configured on Virtual Machine

Test Results Cont'd…

Test Results – ABO Rule -2

L > = 21

Rule Results	Time in Seconds
Decrypted Value for rule Haplotype_O is: 1 Decrypted Value for rule Haplotype_B is: 0 Decrypted Value for rule Haplotype_T is: 1 Decrypted Value for rule E_Haplotype_O is: 1 Decrypted Value for rule E_Haplotype_B is: 0 Decrypted Value for rule E_Haplotype_T is: 0 Decrypted Value for rule Ph_Haplotype_O is: 0 Decrypted Value for rule Ph_Haplotype_T_TT is: 0 Decrypted Value for rule Ph_Haplotype_T_TO is: 0 Decrypted Value for rule Ph_Haplotype_B_BB is: 0 Decrypted Value for rule Ph_Haplotype_B_BO is: 0 Decrypted Value for rule Ph_Haplotype_TB is: 1	Computation = 144 Encryption = 8 Decryption = 36

- Setup is configured on Virtual Machine

Summary of the Talk

- Analyze and propose a solution for the challenges that come with genome sequencing data and of data querying on sequenced data sitting on a cloud server
- Provide a scenario for its application in personalized medicine
- Tests homomorphic encryption techniques to assist in improving the strength of both their privacy and utility

Future Work

- Public Databases: multiple hospital provider under different keys
- More efficient computation implementation (parallelism & batching)
- Integrate with other crypto solutions (transciphering for efficient storage)
- Expand functionalities (HLA)
- Better understanding of FHE schemes parameter settings

Centre de Saclay
Nano-Innov PC 172 - 91191 Gif sur Yvette Cedex