PRIVACY-PRESERVING COMPUTATION OF DISEASE RISK BY USING GENOMIC, CLINICAL, AND ENVIRONMENTAL DATA

Erman Ayday, Jean Louis Raisaro, Paul J. McLaren, Jacques Fellay and Jean-Pierre Hubaux

firstname.lastname@epfl.ch

AUGUST 2013

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

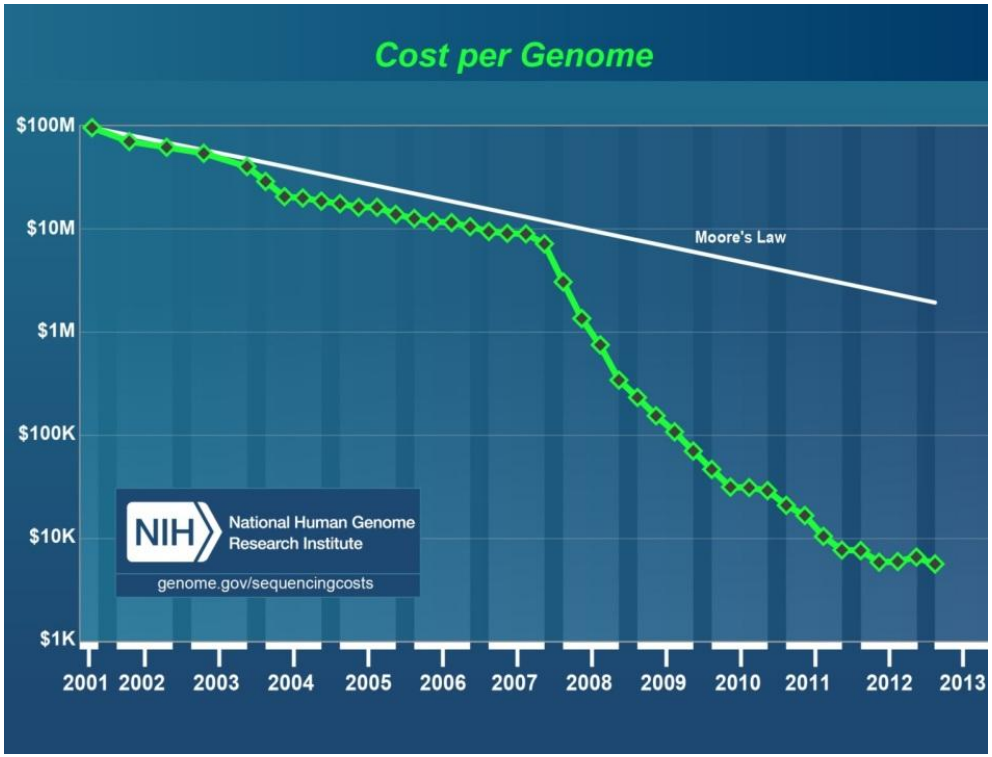
SIGNIFICANCE

Genomic data provides opportunities for substantial improvements in diagnosis and preventive medicine.

SIGNIFICANCE

Genomic data provides opportunities for substantial improvements in diagnosis and preventive medicine.

A complete genome profile is below $100 for genome-wide genotyping (the characterization of about one million common genetic variants).



SIGNIFICANCE

Genomic data provides opportunities for substantial improvements in diagnosis and preventive medicine.

A complete genome profile is below $100 for genome-wide genotyping (the characterization of about one million common genetic variants).

Human Genome Project Spurred $966 Billion Sciences Boom

By Drew Armstrong - Jun 12, 2013 6:01 AM GMT+0200

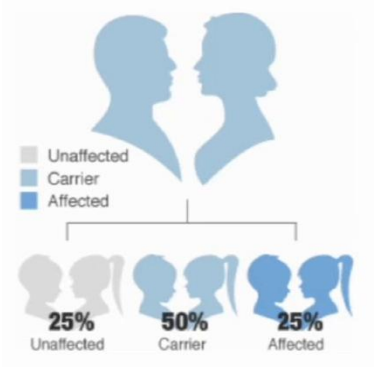
8 COMMENTS QUEUE

The $14.5 billion investment by the U.S. in the Human Genome Project, completed a decade ago, has paid off more than 60-fold in new jobs, drugs and a rapidly expanding genetics industry, an analysis has found.

SIGNIFICANCE

Commercial companies provide low-cost (genetic) tests to their customers.

Counsyl

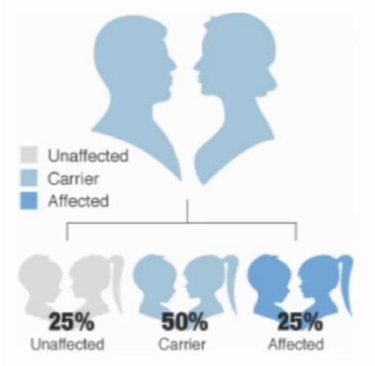


SIGNIFICANCE

Commercial companies provide low-cost (genetic) tests to their customers.

Genetic disease risk tests help early diagnosis of serious diseases.

Counsyl



23andMe

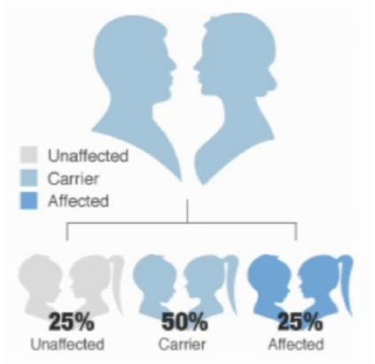
Name	Confidence	Your Risk	Avg. Risk
Atrial Fibrillation	★★★★	33.9%	27.2%
Prostate Cancer ♂	★★★★	29.3%	17.8%
Alzheimer's Disease	★★★★	14.2%	7.2%
Age-related Macular Degeneration	★★★★	11.1%	6.5%
Colorectal Cancer	★★★★	7.8%	5.6%
Chronic Kidney Disease	★★★★	4.2%	3.4%
Restless Legs Syndrome	★★★★	2.5%	2.0%
Parkinson's Disease	★★★★	2.2%	1.6%

SIGNIFICANCE

Commercial companies provide low-cost (genetic) tests to their customers.

Genetic disease risk tests help early diagnosis of serious diseases.

Counsyl



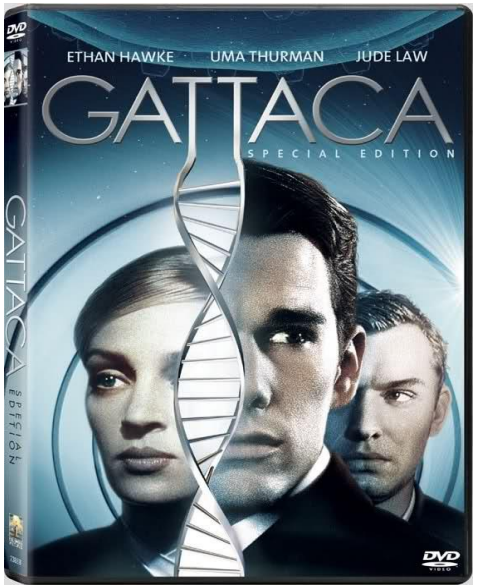
23andMe

Name	Confidence	Your Risk	Avg. Risk
Atrial Fibrillation	★★★★	33.9%	27.2%
Prostate Cancer ♂	★★★★	29.3%	17.8%
Alzheimer's Disease	★★★★	14.2%	7.2%
Age-related Macular Degeneration	★★★★	11.1%	6.5%
Colorectal Cancer	★★★★	7.8%	5.6%
Chronic Kidney Disease	★★★★	4.2%	3.4%
Restless Legs Syndrome	★★★★	2.5%	2.0%
Parkinson's Disease	★★★★	2.2%	1.6%

WHY PROTECT THE GENOMIC DATA?

Genome carries information about a person's genetic condition and predispositions to specific diseases.

- Leakage of such information could enable abuse and threats
- Genetic discrimination



WHY PROTECT THE GENOMIC DATA?

Genome carries information about a person's genetic condition and predispositions to specific diseases.

- Leakage of such information could enable abuse and threats
- Genetic discrimination

Genome carries information about family members.

Genomic data is non-revokable.

WHY PROTECT THE GENOMIC DATA?

Genome carries information about a person's genetic condition and predispositions to specific diseases.

- Leakage of such information could enable abuse and threats
- Genetic discrimination

Genome carries information about family members.

Genomic data is non-revokable.

Anonymisation is ineffective.

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. *Identifying personal genomes by surname inference*. Science: 339 (6117), Jan. 2013

WHY PROTECT THE GENOMIC DATA?

Genome carries information about a person's genetic condition and predispositions to specific diseases.

- Leakage of such information could enable abuse and threats
- Genetic discrimination

Genome carries information about family members.

Genomic data is non-revokable.

Anonymisation is ineffective.

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. *Identifying personal genomes by surname inference*. Science: 339 (6117), Jan. 2013

"The Chills and Thrills of Whole Genome Sequencing"

- E. Ayday, E. De Cristofaro, J.P. Hubaux, G. Tsudik
- http://arxiv.org/abs/1306.1264, June 2013

HotSec'13 – August 13th @ 11am

- *On privacy considerations of genome sequencing*, J.P. Hubaux

MOTIVATION AND GOALS

Non-genomic attributes of the individuals also contribute significantly to their disease risks.

- Clinical and environmental data.

Clinical and environmental data of an individual can include:

- Demographic information
- Family history (e.g., diseases of his family members)
- List of diseases that he carries
- Results of his laboratory tests (e.g., cholesterol level)

Non-genomic attributes are also considered privacy-sensitive.

- HIV status

Such data should also be considered along with the individuals' genomic data in a privacy-preserving way.

MOTIVATION AND GOALS

One might not want to directly provide his genomic data and clinical and environmental attributes to the medical unit.

- To protect his privacy-sensitive data

What is important for the medical unit is the end-result (risk of the patient for a disease or the compatibility of a person to a drug).

- Not the individual attributes of the person that lead to the end-result

Such data plays an important role in a disease risk test.

- Inaccuracy (or absence) of such data might cause incorrect (or misleading) results

It is crucial to use the correct and complete data of the individuals for the accuracy of disease risk tests, while still protecting their privacy.

CONTRIBUTIONS

A system for protecting the privacy of individuals' sensitive genomic, clinical, and environmental information.

Enable medical units to process genomic and non-genomic data in a privacy-preserving way to perform disease risk tests.

- Homomorphic encryption
- Privacy-preserving integer comparison

Implement the proposed system and show its practicality via a complexity evaluation.

POTENTIAL APPLICATIONS

A pharmacist checking if a given drug could be harmful (toxicity, interactions) for a patient.

A pharmaceutical company categorizing people based on their risk for a particular disease in order to identify potential clinical trial participants.

A regional health ministry determining the fraction of people at high risk for a particular disease in order to optimize a population-wide preventive medicine effort.

An online direct-to-consumer service provider offering individual risk prediction for various diseases, considering genomic, clinical and environmental data.

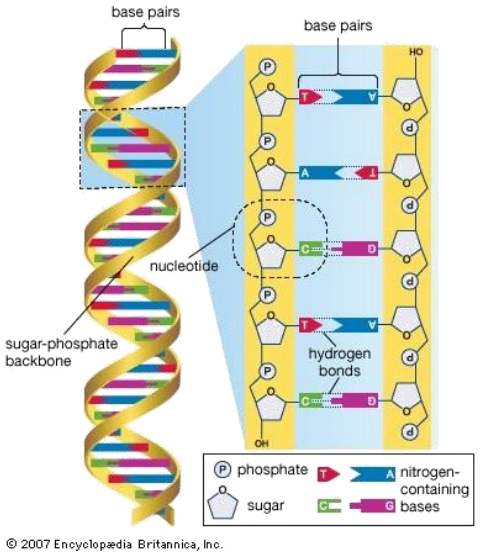
GENOMICS 101

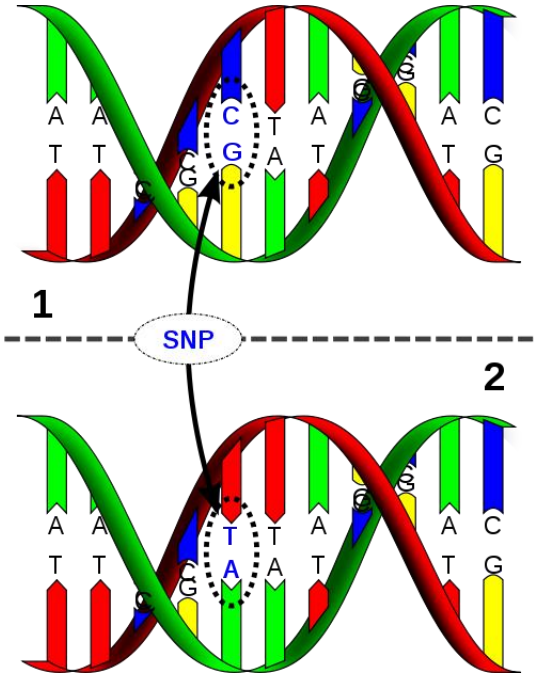
The human genome consists of approximately 3 billion letters.

- 99.9% is identical between any two individuals
- Remaining: human genetic variation

Single Nucleotide Polymorphism (SNP): Most common human genetic variation.

- A single nucleotide (A, C, G, or T) differs between members of the same species or paired chromosomes of an individual
- Potential nucleotides for a SNP are called **alleles**
- 2 different types of alleles observed for each SNP
- Everyone carries 2 alleles at each SNP position

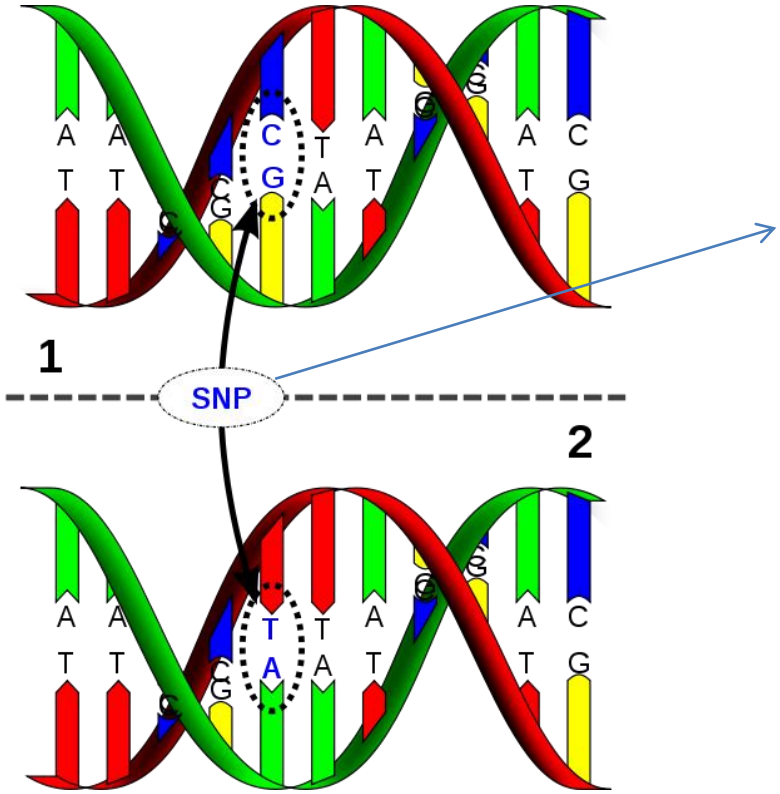




GENOMICS 101

Disease risk can be computed by analyzing particular SNPs.

One of the alleles carries the risk for the corresponding disease and the other allele does not contribute.



SNP associated with disease X

- Alleles: C and T
- Risk allele: C
- Genotypes: CC, TT, CT

$SNP_i^P = \{0, 1, 2\}$ based on the number of risk alleles it carries.

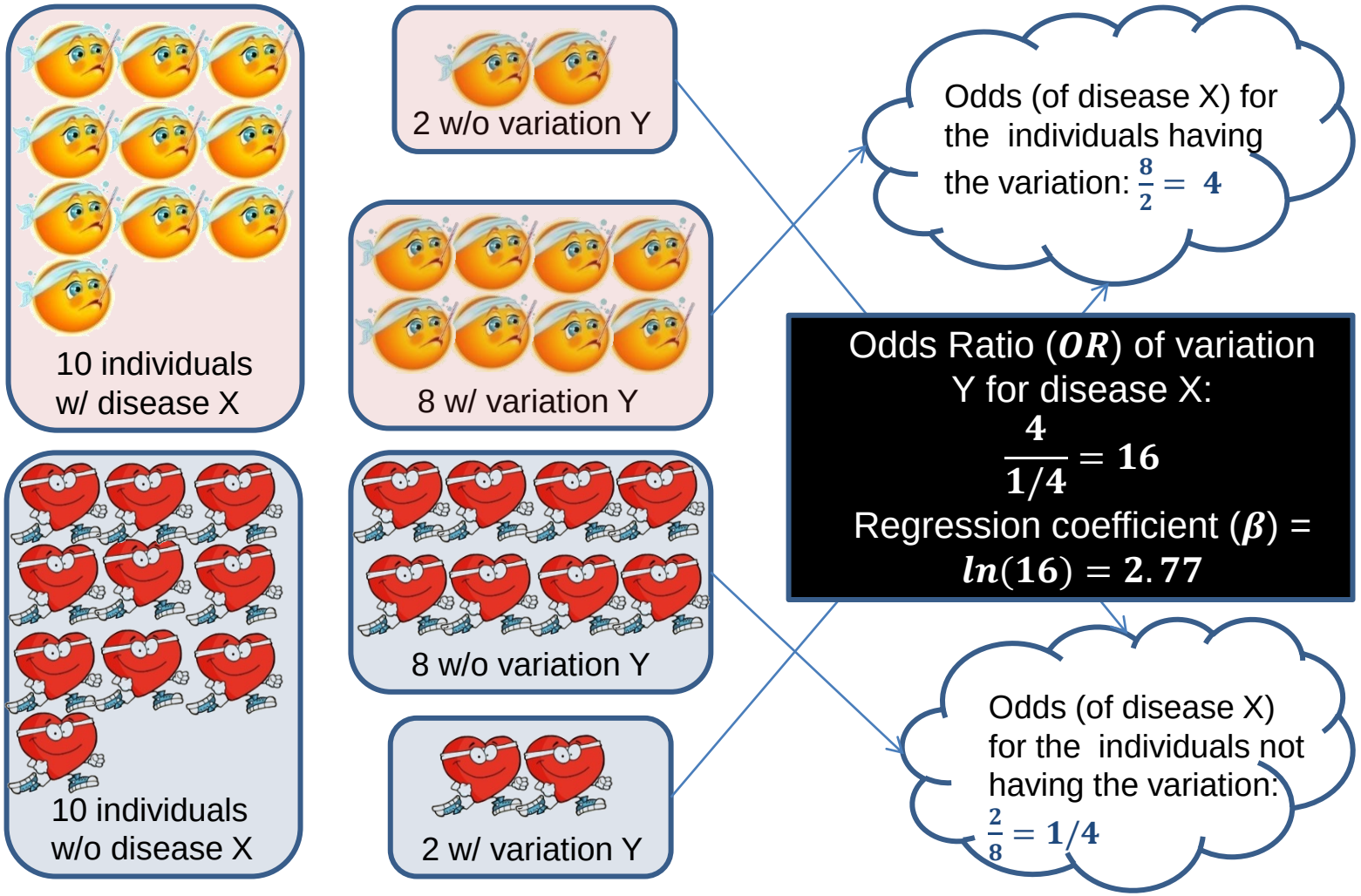
COMPUTATION OF DISEASE RISK

The strength of the association between each SNP and a disease is expressed by the odds ratio (OR).

Example: OR of a variation Y for a disease X.

COMPUTATION OF DISEASE RISK

The strength of the association between each SNP and a disease is expressed by the odds ratio (OR).



COMPUTATION OF DISEASE RISK

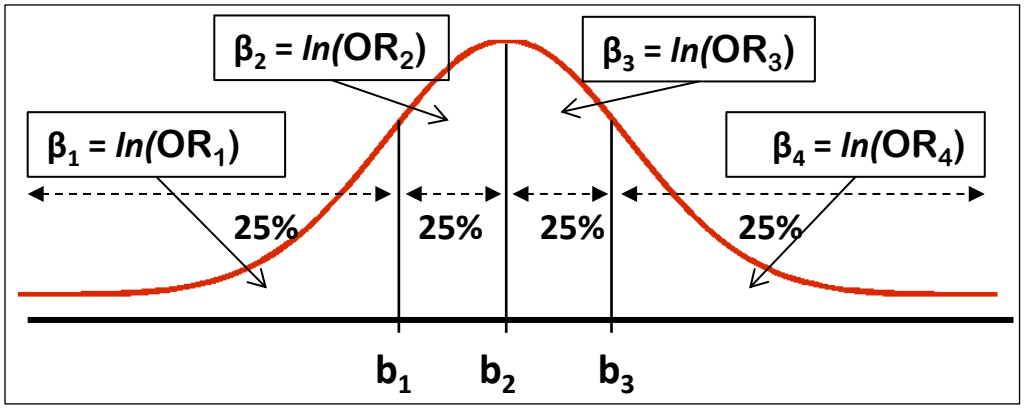
Overall genetic risk is computed based on the OR of each associated SNP by using a logistic regression model.

$$[\mathbb{S}] = ln\left(\frac{Pr_g}{1 - Pr_g}\right) = \left[\alpha + \sum_i \beta_i p_j^i(X)\right]$$

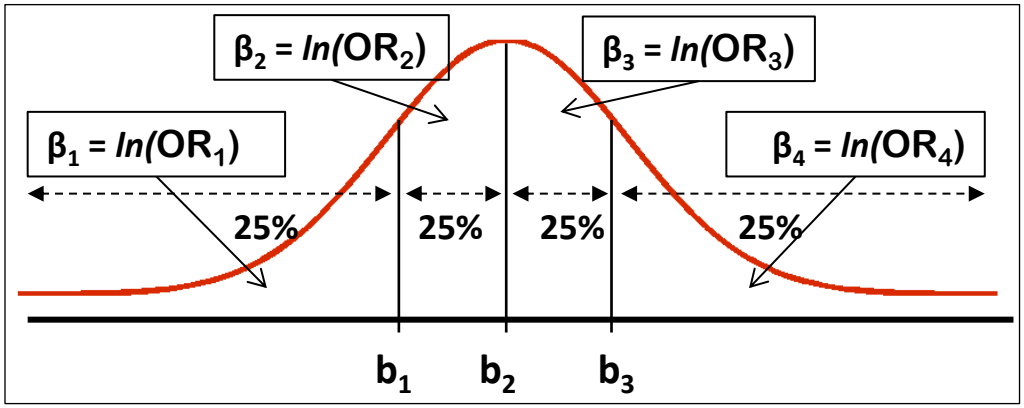
β_i: Regression coefficient, $OR_i = exp(\beta_i)$.

Genetic risk should be categorized based on its risk group.

- Distribution of the potential genetic scores (in a given population) is divided into smaller parts called quantiles



COMPUTATION OF DISEASE RISK



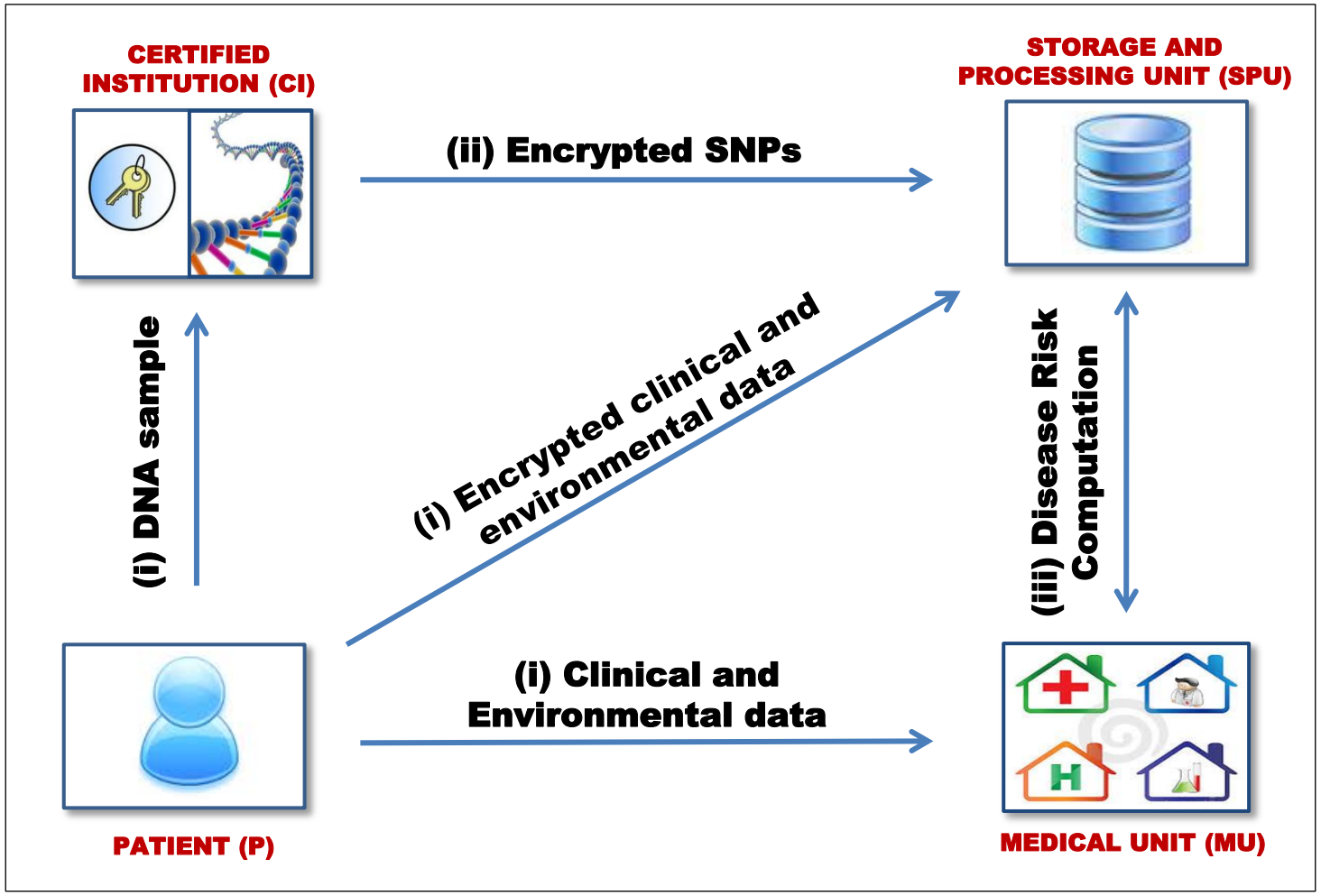
To compute the overall disease risk, the genetic information needs to be combined together with the clinical and environmental factors

A second and final multi-variable logistic regression model is used to find the final (aggregate) regression coefficient β_f

$$ln\left(\frac{Pr}{1-Pr}\right) = \beta_f = \beta_0 + \beta_g + \sum_{N_i \in \mathbf{N}} \overline{\beta}_i N_i$$

$$Pr = \frac{e^{\beta_f}}{1 + e^{\beta_f}}$$

SYSTEM MODEL - OVERVIEW



SYSTEM MODEL - OVERVIEW

The certified institution (CI) is a trusted entity.

- Indispensable to do the sequencing

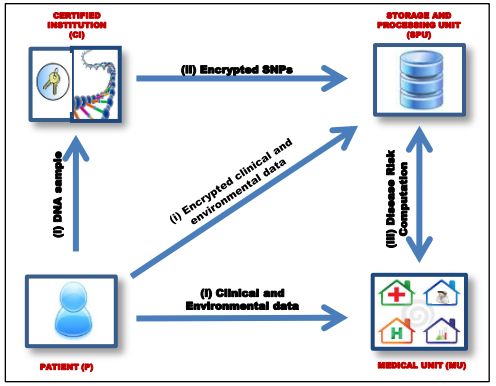
Clinical and environmental data is collected at the medical unit (MU) or directly provided by the patient.

A storage and processing unit (SPU) stores and processes genomic, clinical, and environmental data for efficiency and security.

- Patients' data is stored using pseudonyms

Paillier cryptosystem for the encryption of the data.

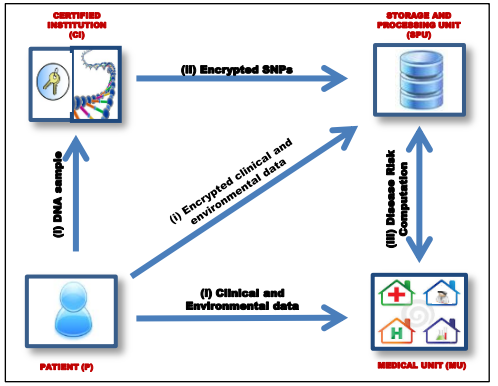
- Homomorphic addition
- Multiplication of the ciphertext with a constant
- Proxy re-encryption (secret sharing of the private key)



THREAT MODEL

An attacker at the MU.

- A careless or disgruntled employee at the MU or a hacker who breaks into the MU
- Aims to obtain private genomic, clinical, and environmental information about a patient (for which it is not authorized)



A curious party at the SPU.

- Existence of a curious party or a disgruntled employee at the SPU

Both MU and SPU follows the protocols properly.

No collusion between the MU and the SPU.

Access control by the SPU.

DATA ENCRYPTION

Data is encrypted using the modified Paillier cryptosystem.

- CI encrypts the contents of all SNP positions of the patient (to obtain $[SNP_i^P]$) along with their squared values (to obtain $[(SNP_i^P)^2]$).

P	SNP_1^P	SNP_2^P	SNP_3^P	. . .	SNP_n^P
	[1]	[0]	[2]	. . .	[2]

P	$(SNP_1^P)^2$	$(SNP_2^P)^2$	$(SNP_3^P)^2$	. . .	$(SNP_n^P)^2$
	[1]	[0]	[4]	. . .	[4]

- MU (or patient) individually encrypts each clinical and environmental attribute of the patient

P	Smoking	Hypertension	High cholesterol	. . .	Age (>45)
	[1]	[0]	[1]	. . .	[1]

PRIVACY-PRESERVING COMPUTATION OF DISEASE RISK

MU requests the (encrypted) genomic, clinical, and environmental data of the patient from the SPU.

SPU verifies that the MU has the required access rights.

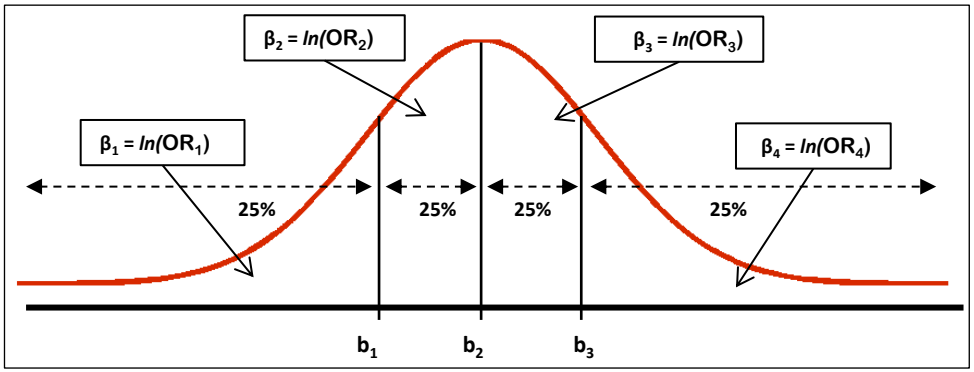
MU first computes the genetic risk of the patient for disease X.

$$[\mathbb{S}] = \left[\sum_i \boldsymbol{\beta_i} \left\{ \frac{\boldsymbol{p_0^i(X)}}{\boldsymbol{a}} \left(\boldsymbol{SNP_i^P} - \mathbf{1}\right)\left(\boldsymbol{SNP_i^P} - \mathbf{2}\right) + \frac{\boldsymbol{p_1^i(X)}}{\boldsymbol{b}} \left(\boldsymbol{SNP_i^P}\right)\left(\boldsymbol{SNP_i^P} - \mathbf{2}\right) + \frac{\boldsymbol{p_2^i(X)}}{\boldsymbol{c}} \left(\boldsymbol{SNP_i^P}\right)\left(\boldsymbol{SNP_i^P} - \mathbf{1}\right) \right\} \right]$$

Find the regression coefficient corresponding to the computed genetic risk using a privacy-preserving integer comparison algorithm [1] between the MU and the SPU.

[1] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, “Privacy-preserving face recognition,” *Proceedings of Privacy Enhancing Technologies*, pp. 235–253, 2009.

PRIVACY-PRESERVING COMPUTATION OF DISEASE RISK



MU compares $[\mathbb{S}]$ with the boundaries of the genetic risk scale.

- Neither the MU nor the SPU learns the value of $\mathbb{S}$ or the result
- b_i^l and b_i^u represent the lower and upper boundary of the ith risk group

IDEA: MU computes $[z] = \left[2L + \mathbb{S} - b_i^j\right]$.

- $\boldsymbol{z_{L-1}}$ represent the most significant bit of z
- (i) $\boldsymbol{z_{L-1}} = \mathbf{0}$ if $\mathbb{S} < b_i^j$; and (ii) $\boldsymbol{z_{L-1}} = \mathbf{1}$ if $\mathbb{S} \geq b_i^j$
- $[\boldsymbol{z_{L-1}}] = [z - (z \bmod 2^L)]$
- $[z \bmod 2^L]$ is computed via secure 2PC between SPU and MU

PRIVACY-PRESERVING COMPUTATION OF DISEASE RISK

Let $[\boldsymbol{G}(\mathbb{S}, \boldsymbol{b_i^u})] = [\boldsymbol{z_{L-1}}]$ represent the (encrypted) result of the comparison between $\mathbb{S}$ and $\boldsymbol{b_i^u}$.

- (i) $\boldsymbol{G}(\mathbb{S}, \boldsymbol{b_i^u}) = \mathbf{0}$ if $\mathbb{S} < b_i^l$
- (ii) $\boldsymbol{G}(\mathbb{S}, \boldsymbol{b_i^u}) = \mathbf{1}$ if $\mathbb{S} \geq b_i^l$

MU computes the genetic regression coefficient $[\boldsymbol{\beta_g}]$.

$$[\boldsymbol{\beta_g}] = \left[\boldsymbol{\beta_1}\left(\mathbf{1} - \boldsymbol{G}(\mathbb{S}, \boldsymbol{b_1^u})\right) + \sum_{\boldsymbol{i=2}}^{\boldsymbol{(\rho-1)}} \boldsymbol{\beta_i}\left(\boldsymbol{G}(\mathbb{S}, \boldsymbol{b_{i-1}^u}) - \boldsymbol{G}(\mathbb{S}, \boldsymbol{b_i^u})\right) + \boldsymbol{\beta_\rho} \boldsymbol{G}\left(\mathbb{S}, \boldsymbol{b_{\rho-1}^u}\right)\right]$$

MU combines $[\boldsymbol{\beta_g}]$ with the patient's clinical and environmental regression coefficients to obtain the aggregate regression coefficient $\boldsymbol{\beta_f}$.

- Let $\mathrm{N} = \{[N_1], [N_2], \ldots, [N_m]\}$ be the set of encrypted clinical and environmental attributes of the patient, where $N_i \in \{0, 1\}$
- If N_i is non-binary, it can be transformed to a binary number using the privacy-preserving comparison algorithm

$$[\boldsymbol{\beta_f}] = \left[\boldsymbol{\beta_0} + \boldsymbol{\beta_g} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \overline{\boldsymbol{\beta}}_{\boldsymbol{i}} \boldsymbol{N_i}\right]$$

IMPLEMENTATION

Encrypted a real individual's SNP profile from [1].

Computed the coronary artery disease (CAD) risk by using real data from [2].

- 23 SNPs associated with cardiovascular risk
- 14 clinical and environmental factors
- 4 genetic risk groups (quantiles)
- ORs computed on a population of 2078 individuals

Intel Core i7-2620M CPU with 2.70 GHz processor

Windows 7

MySQL 5.5 database

Java programming language

Size of security parameter: 4096 bits (n in Paillier)

[1] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[2] M. Rotger and *et al.*, "Contribution of genetic background, traditional risk factors and HIV-related factors to coronary artery disease events in HIV-positive persons," *Clinical Infectious Diseases*, Mar. 2013.

IMPLEMENTATION

SNP	Chr	Allele	Risk allele	OR
rs3798220	6	T>C	C	1.51
rs4977574	9	A>G	G	1.29
rs9982601	21	C>T	T	1.18
rs17114036	1	A>G	A	1.17
rs17465637	1	C>A	C	1.14
rs6725887	2	T>C	C	1.14
rs1122608	19	G>T	G	1.14
rs964184	11	C>G	G	1.13
rs12413409	10	G>A	G	1.12
rs2306374	3	T>C	C	1.12
rs599839	1	A>G	A	1.11
rs579459	9	T>C	C	1.10
rs12526453	6	C>G	C	1.10
rs11556924	7	C>T	C	1.09
rs1746048	10	C>T	C	1.09
rs12190287	6	C>G	C	1.08
rs3825807	15	A>G	A	1.08
rs216172	17	C>G	G	1.07
rs12936587	17	A>G	G	1.07
rs4773144	13	A>G	G	1.07
rs17609940	6	G>C	G	1.07
rs2895811	14	T>C	C	1.07
rs46522	17	T>C	T	1.06

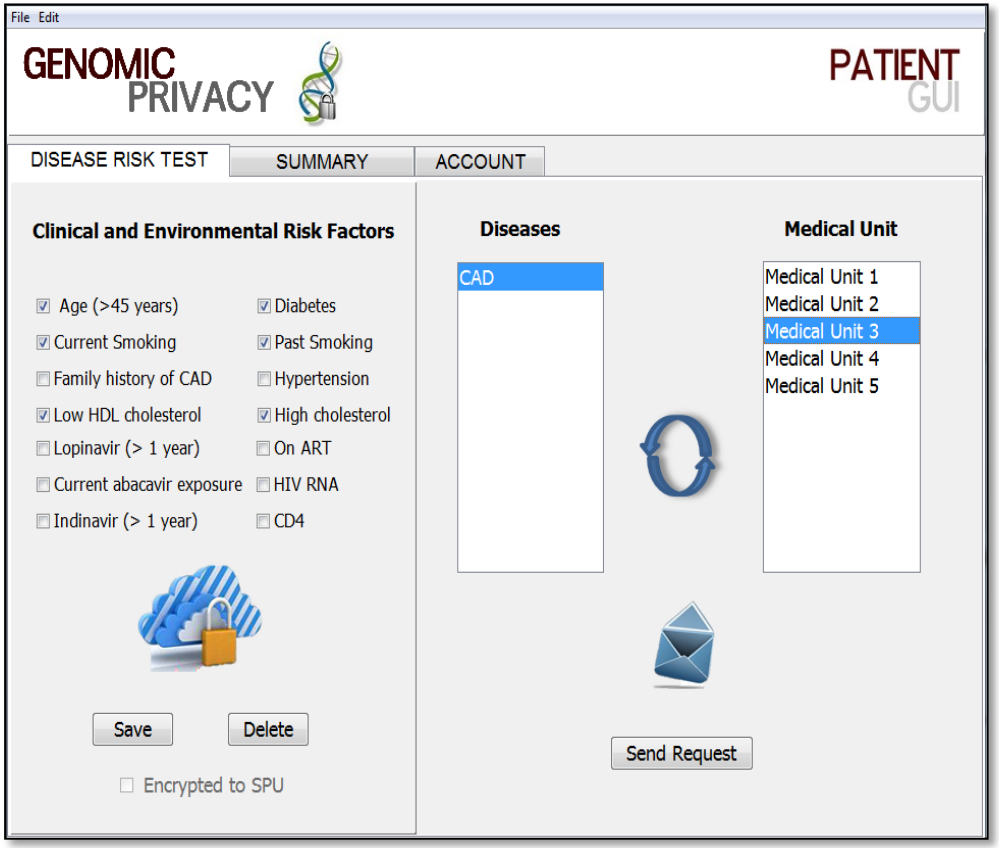
Genomic Variants (SNPs) used for the Genetic Risk [1]

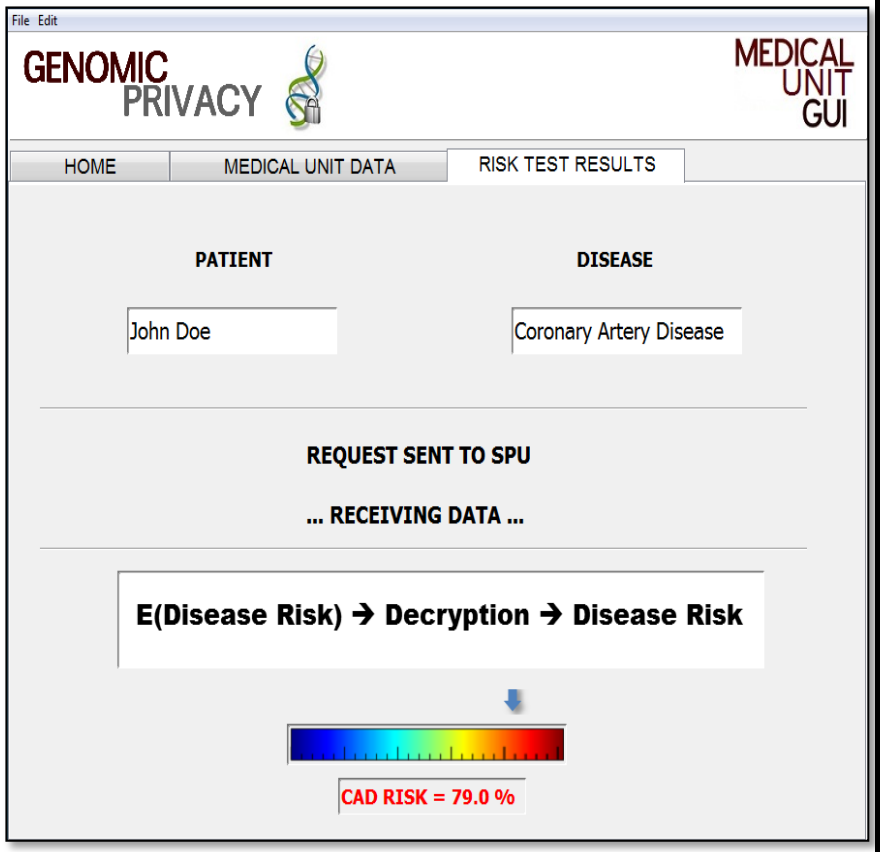
Variable	Odds Ratio
Age (>45 years)	3.21
Current smoking	2.25
Family history of CAD	2.00
Lopinavir (> 1 year)	1.74
Diabetes	1.81
Current abacavir exposure	1.62
Past smoking	1.51
Indinavir (> 1 year)	1.28
High cholesterol	1.61
On ART	1.51
Hypertension	1.44
Low HDL cholesterol	1.11
CD4	0.99
HIV RNA	1.00
Genetic score quantile 2 vs. quantile 1	1.12
Genetic score quantile 3 vs. quantile 1	1.33
Genetic score quantile 4 vs. quantile 1	1.62

Clinical and Environmental Risk Factors [1]

[1] M. Rotger and *et al.*, “Contribution of genetic background, traditional risk factors and HIV-related factors to coronary artery disease events in HIV-positive persons,” *Clinical Infectious Diseases*, Mar. 2013.

IMPLEMENTATION





IMPLEMENTATION

Complexity of the Proposed System				
Encryption	**Storage**	**Computation of disease risk**		
380 ms./attribute (with pre-computed values: 0.168 ms./attribute)	51.2 GB per patient	*Computation of the genetic risk*	*Privacy-preserving integer comparison*	*Computation of the final risk*
		230 sec (23 SNPs)	3.390 sec (3 comparisons)	140 sec (14 environmental factors)
		Total: 373 sec		

CONCLUSION

A framework in which patients' genomic, clinical, and environmental data is securely stored at a storage and processing unit.

Medical unit conducts disease risk tests on this encrypted data by using homomorphic encryption and privacy preserving integer comparison.

Preserves the privacy of the patients against a curious party at the storage and processing unit and a malicious party at the medical unit.

Implemented the proposed solution and showed its practicality.

Encourage the use of genomic, clinical, and environmental data in medical tests by ensuring the patients that the privacy of their sensitive data will be preserved.

QUESTIONS

erman.ayday@epfl.ch

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