PRIVGEN

Privacy-preserving sharing and processing of genetic data

G. Coatrieux, LaTIM Inserm UMR 1101, IMT Atlantique

M. Südholt, LS2N CNRS UMR6004, IMT Atlantique

E. Génin, GGB Inserm UMR 1078

Motivation

Genetic data and applications

Genetic data applications

- **Patient Care** (Diagnosis; Therapeutic choice).
- **Research** (Genome-Wide Association Studies)
  - Needs for international sharing of genetic data to increase sample sizes and achieve sufficient power
- **Legal and forensics** (criminal forensics, contested paternity…)
- **Direct-to-Consumer Services** (genealogy, Disease susceptibility risks …)

International/national projects for genomic data sharing

- **Global Alliance for Genomics & Health** (beacons)
- **Disease specific databases** (e.g. Sickkids)
- **Plan France Médecine Génomique 2025**

Motivation

Security and privacy of genetic data

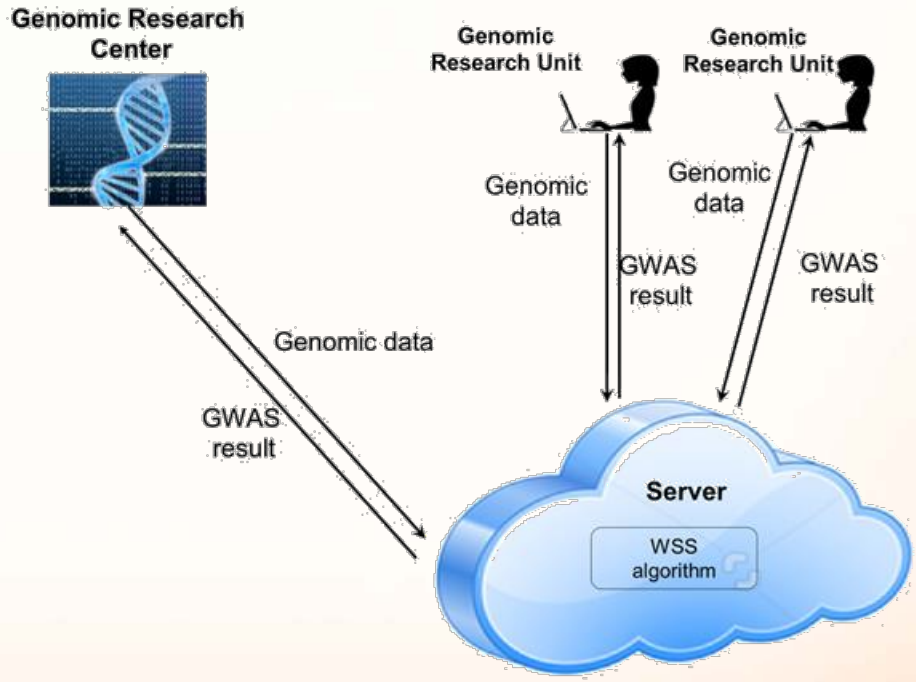
Genetic data are extremely sensitive data –

- They can allow identification and are partly shared with relatives, they can inform about health/behavior of the individual but also of his/her relatives.
- Sensitive genomic characteristics are crucial for most applications
- They have their own deontological and legislative frameworks, and beyond common medical regulations:
  - **International Declaration on Human Genetic Data, UNESCO, 2003** - Article 14§(a) - Countries should endeavor to protect the privacy of individuals and the confidentiality of genetic data …
  - **GDPR** - Article 4- defines genetic data as personal data - Article 9- genetic data processing treatments must be carried out in agreement with data owner and respecting his privacy and data's confidentiality

Motivation

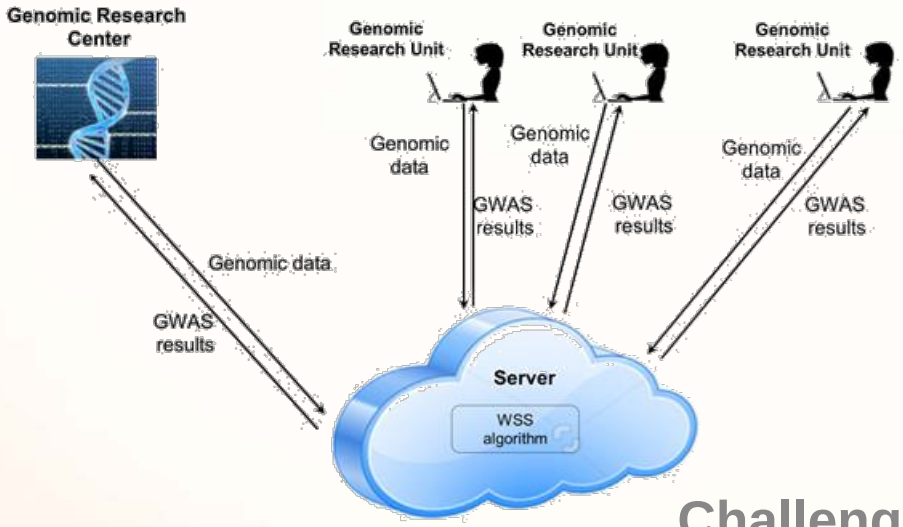
Genome-Wide Association Studies threats (examples)

- Owners loose the control on their data:
  - Confidentiality concerns
  - Traceability, intellectual/scientific ownership protection
    - Illegal redistribution / re-routing of the information (information leaks)
  - Integrity concerns
    - Falsification of data to interfere with statistical studies
    - Integrity of outsourced computations



- Privacy concerns
  - Identification of the presence of an individual in an anonymized public database
  - Reconstruction of individual sample without a priori knowledge
  - Identification of the GWAS objective (interest for a specific gene)
- …

Research activities of PRIVGEN



Challenge 1 - Mechanisms for a continuous digital content protection

***Objective**: Merge different security mechanisms into one digital content protection tool for continuous and multipurpose security objectives*

Challenge 2 - Composition of security and privacy-protection mechanisms

***Objective**: a compositional development approach for secure and privacy-preserving distributed genetic application*

Challenge 3 - Distributed processing and sharing of genetic data

***Objective**: a platform to perform association testing against a reference panel of healthy individuals that are compared against patients*

Ch1 - Mechanisms for a continuous digital content protection

Secure externalized Genome-Wide Association Studies

- Identification of the constraints and limits of actual secure GWAS:
  - Mainly secured with differential privacy, homomorphic encryption, secure multiparty computations …
  - Do not consider sharing of patient data but of aggregated data
  - Are not compliant with iterative algorithms for large data sets
- Proposition of different secure GWAS algorithms based on externalized data
  - Secure externalized $\chi 2$ test and collapsing test
  - 1rst secure neural network able to learn on encrypted data
    - Input, output and trained parameter are all encrypted
    - No approximation of the neuron activation function
    - No interactions with the user during the training phase
    - Loss of accuracy due to the need to work with integer values
  - A secure framework for iterative algorithms based on a original proxy strategy along with cryptographic hashing and pretty good privacy (PGP).

D. Niyitegeka et al. ***Secure Collapsing Method Based on Fully Homomorphic Encryption****, Studies in Health, Submitted.*
R. Bellafqira et al., ***Secure Multilayer Perceptron Based on Homomorphic Encryption****. Digital Forensics and Watermarking, IWDW 2018, LNCS Vol. 11378, pp. 322-336, 2019.*

Ch1 - Mechanisms for a continuous digital content protection

Crypto-watermarking tools for outsourced genetic data

- First watermarking modulations for genetic data compliant with VCF and WSS files (GWAS file formats)
  - DNA Watermarking methods have been proposed for steganography purposes or copyright protection of synthetic or living cell DNA
  - Solutions we proposed are robust or reversible and do not interfere with GWAS algorithms

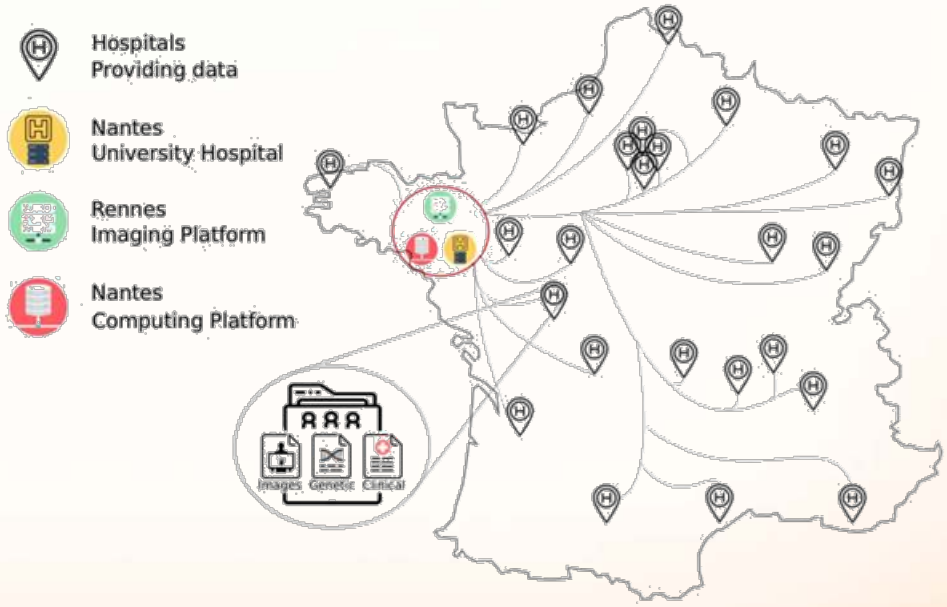
- Proposition of a dynamic HE crypto-watermarking system. Allows the cloud to protect in terms of integrity HE encrypted data and identify modified data.
  - Dynamic - protected database can be updated by the data-owners, the cloud does not have to re-watermark the whole data base.
  - Tampered data (modified, deleted and/or added) can be identified

R. Bellafqira et al., ***Robust watermarking for genetic data traceability in externalized GWAS frameworks****. IEEE/ACM transactions on computational biology and bioinformatics, submitted.*
G. Coatrieux et al., ***Lossless watermarking for genomic data****. IEEE Transactions on Information Forensics and Security, submitted*
D. Niyitegeka et al. ***Dynamic Watermarking-Based Integrity Protection of Homomorphically Encrypted Databases****. Digital Forensics and Watermarking, IWDW 2018, LNCS Vol. 11378, pp. 151-166, 2019.*
J.F. Contreras et al., ***Protection of Relational Databases by Means of Watermarking: Recent Advances and Challenges****. Chapter, book Advances in Security in Computing and Communications, , pp.101-121 , 2017*

Ch 2 - Composition of security and privacy-protection mechanisms

Motivation: Distributed analyses

- General trend: geographically distributed data and computation
- I-CAN project (ex.): 34 French hospitals (+ future foreign cooperations)



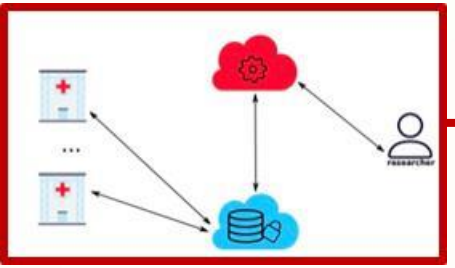
- Future:
  - Large distribution of massive data and computations
  - Different admin/legal domains

*Garzon, Südholt, Benavides: "**An overview of distributed biomedical analyses and data analytics tools"**, to be submitted in January 2020.*

Ch 2 - Composition of security and privacy-protection mechanisms

Development approach

1. Requirements analysis



2. Architecture selection

Set of predefined architectures

→ **Associated scalability, security/ privacy properties (watermarking, encry., …)**

3. Data/workflow definition

Security/privacy operators (watermarking, encry., …)

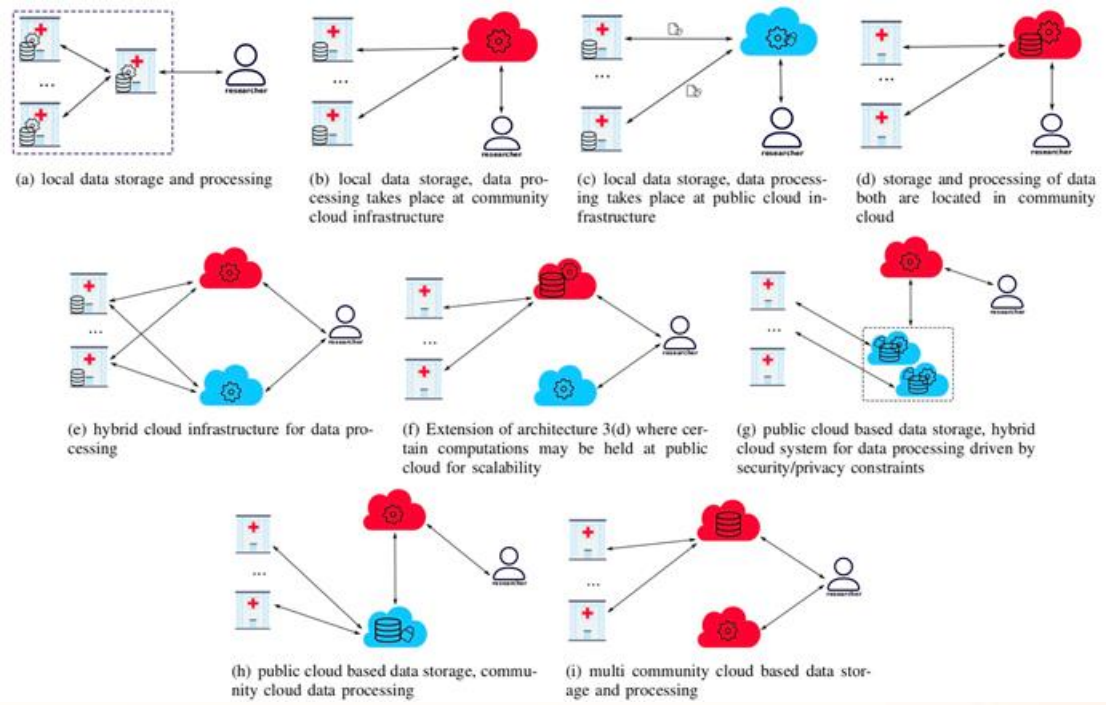
→ **Enforcement of security/ privacy props (watermarking, encry., …)**

4. Analysis deployment & execution

Java framework, Cloud deployment

Ch 2 - Composition of security and privacy-protection mechanisms

- **Architecture level** : set of architectures for biomedical analyses derived from existing analysis



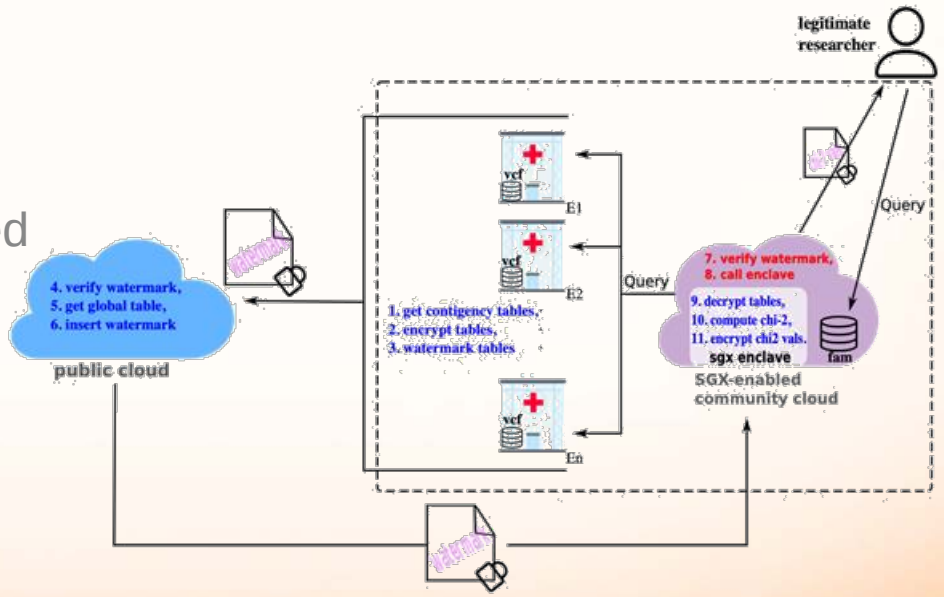
- **Composition support: Programming-level operators**
  - Operators for homomorphic/ a/symmetric/ attribute-based encryption, fragmentation, third-party/localized computations, watermarking
  - Definition as Idris functions
    - Composition through higher-order functions
    - Type-based correctness support
  - Second definition as java framework

F.-z. Boujdad, A. Gaignard, M. Südholt, et al: "**On distributed collaboration for biomedical analyses**"; CCGrid-Life, May 2019.

F.-z. Boujdad, M. Südholt: "**Constructive Privacy for Shared Genetic Data**"; 8th Int. Conf. on Cloud Computing and Services Science (CLOSER 2018); ScitePress; Madeira, March 2018.

Challenge 3 - two platforms for distributed processing and sharing of genetic data

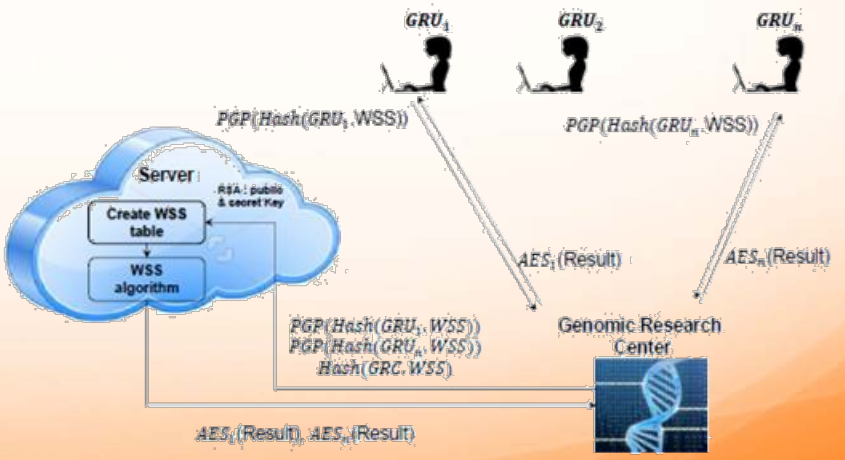
- **Platform 1 :** PrivGen cloud platform
  - PrivGen cloud platform
    - Java implementation + deployment/execution scripts
  - Deployment on Grid 5000
    - Large-scale CS grid/cluster-based infrastructure
  - Direct transfert to biomedical infrastructures (IFB, Bird)
  - Hybrid cloud, use of SGX



*Boujdad, Coatrieux, Genin, Niyitegeka, Südholt et al.: "**A Hybrid Cloud Architecture for Performing Privacy-Preserving Genome-Wide Association Studies**", to be submitted end of Nov. 2019.*

Challenge 3 - two platforms for distributed processing and sharing of genetic data

- **Platform 2 :** PRIVAS - A tool to perform Privacy-Preserving Association Studies
  - A complete secured GWAS framework – allows several research units to share their data in the cloud without breaching privacy and to conduct “iterative” GWAS without computation complexity overhead
  - Framework implemented with Weighted-Sum Statistic (WSS) algorithm as illustration but generic to all GWAS algorithms similar to WSS (see challenge 1)
  - Proof Of Concept Platform deployed on Datarmor supercomputer of IFREMER
  - POC to be made available to the international community.



R. Bellafqira et al.. ***Privacy-preserving genome-wide association study for rare mutations - a secure framework for externalized statistical analysis***. *IEEE Access, submitted, 2020.*

T.E. Ludwig, et al., ***PrivAS: a tool to perform Privacy-Preserving Association Studies***. *Jobim, 2019.*