Institut Mines-Télécom

Privacy and Sharing of Genomic Data

Mario Südholt
IMT Atlantique

Séminaire Cybersécurité
IMT, 10 nov. 2017

1: Sharing of Genomic Data Today

Objective: analyze human genome for diseases

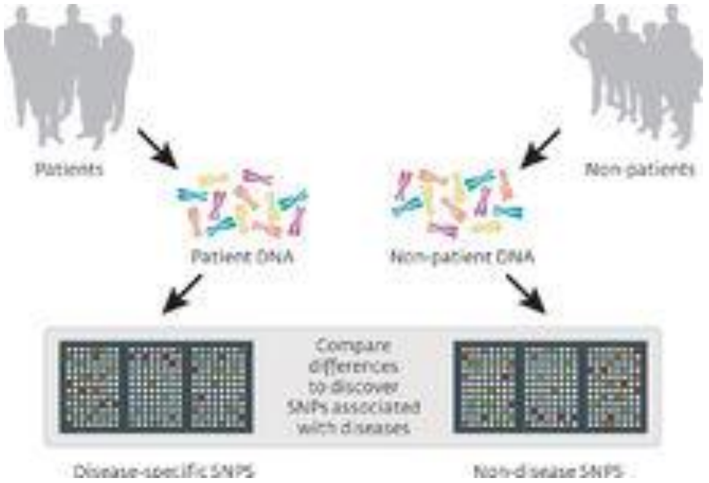
Genomic Wide-Association Studies (GWAS)

Associations between genetic variations and specific traits

- Ex.: BRCA genes and risk of breast cancer

Sharing of aggregate data

- **Simple client-server architectures**
- **Initially: no privacy problems known**



Attacks on shared genomic data servers

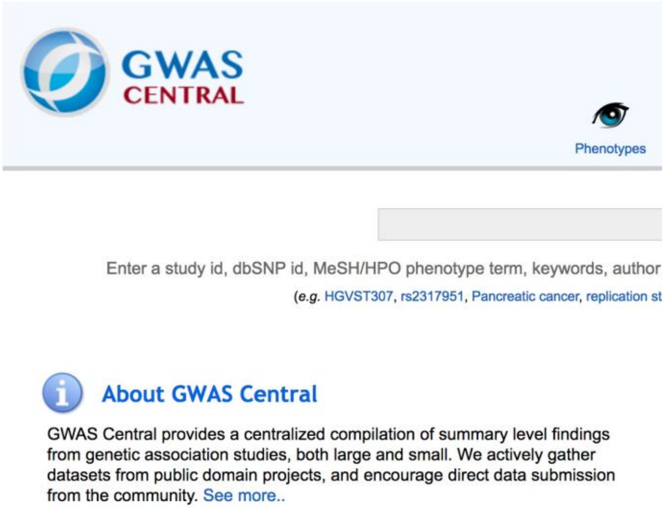
Attacks for identification of individuals in genetic DBs with aggregate data

[Homer et al., 2008]: identification in large aggregate data sets
[Sankaraman et al., 2009]: upper bound on detection power
[Wang et al., 2009]: identification in small data sets

Result: severe restrictions to public data sharing

Ex.: `gwascentral.org`

- frequency info not available
- Large data sets only available on request ("Data Sharing Statement" of GWAS Central)



Generally: restrictions on

- **Sharing system architecture**
- **Queries on genomic DBes**

Need for more advanced sharing

Geneticians interested in more advanced sharing possibilities

Allow sharing of **larger data sets**

Enable collaborative work on **rare variants / uncommon diseases**

Advanced sharing architectures

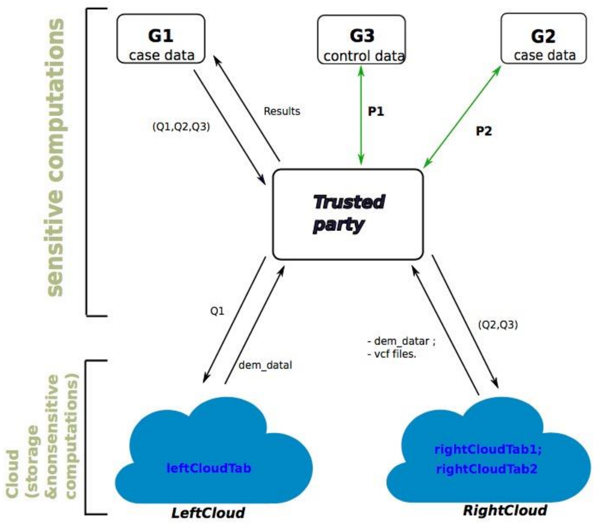
Sharing of raw data via e.g. trusted party

Quicker access to data in the cloud

Use advanced sharing techniques

Support **confidentiality** and **integrity** efficiently

Support for **ownership** and **traceability** properties



How to support such sharing scenarios?

Methods for the construction of architectures/processes/queries

- Means for **design** and **programming**
- Ensure basic **privacy guarantees**
- **Optimize** architectures/applications
- **Formal verification** of advanced privacy properties

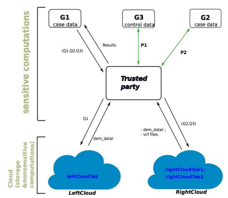
How to ensure privacy properties?

Multiple privacy enhancing techniques:

- **Encryption**: a/symetric, homomorphic, attribute-based, …
- **Client-side computing**: compute associations within local perimeter
- **Data fragmentation**

- **Watermarking** for ownership and traceability properties

COMPOSITIONAL CONSTRUCTION OF GENOMIC APPS



Declarative scenario definitions with **privacy types**

```
Wat : (a: Attribute) →
      {auto p1 : So (isRawType (snd a))} →
      {auto p2 : So (isInEnv a env)}  →
      Privy env (watEnv a GIG env) []

— watermark detection operator

data Query :
        ...

detectw :
     (a : Attribute) → (info : ReadM GIG) →
     {default Refl p1: (snd a) = (WATERMARK GIG t)}
     → Query Δ →
     {auto p2 : Elem a Δ} →
     Query ((replaceOn a (fst a, t) Δ)++[MyTattoo])
```

Algebraic theory: **watermarking laws**

Definition of **watermarking operator**

$$decrypt_{(s,a)} \circ crypt_{(s,a)} \circ detectw_a \circ wat_a \equiv$$
$$detectw_a \circ decrypt_{(s,a)} \circ crypt_{(s,a)} \circ wat_a$$
$$if\, dom(p) \cap a = \emptyset$$
$$detectw_a \circ \sigma_p = \sigma_p \circ detectw_a$$

```
scenario : GeneticQuery [SubjectId,ZIP,Gender,DoB,
                         Variant,TypeVar,MyTattoo]
scenario = do

 G1 `SendRequest` (TTP,[Q1])
 G1 `SendRequest` (TTP,[Q2,Q2'])
 G1 `SendRequest` (TTP,[Q3,Q3'])

 TTP `SendRequest` (LeftCloud,[Q1])
 TTP `SendRequest` (RightCloud,[Q2,Q2'])
 TTP `SendRequest` (RightCloud,[Q3,Q3'])

 let q1 = LeftCloud  `executeRequest` [Q1];
 let q2 = RightCloud `executeRequest` [Q2,Q2'];
 let q3 = RightCloud `executeRequest` [Q3,Q3'];

 demDatal     ← LeftCloud  `SendData` (TTP,q1)
 demDatar     ← RightCloud `SendData` (TTP,q2)
 vcfFiles     ← RightCloud `SendData` (TTP,q3)

 let r1 = decrypt VariantWE (AESD "key2") vcfFiles;
 let r2 = decrypt TypeVarWE (AESD "key1") r1;
 let r3 =
     detectw VariantW (RGIG "wkey1" 1 ["seed1"] 1) r2;
 let vcfFiles =
     detectw TypeVarW (RGIG "wkey2" 2 ["seed2"] 2) r3;
 let plainData =
     defrag (defrag demDatal demDatar) vcfFiles

 TTP `ReturnResults` (G1, TTP `Compute` plainData)
```

Verific./Optim. of genetic applications for privacy/ efficiency

$$\pi_{(variant,typeVar)} \circ$$
$$\sigma_{((subjectId \in mdd) \wedge (position=i, position=j,..))}$$

(a) local query

$$detectw_{variant,typeVar} \circ decrypt_{variant,typeVar} \circ$$
$$\pi_{(variant,typeVar)} \circ$$
$$\sigma_{((subjectId \in mdd) \wedge (position=i, position=j,..))} \circ$$
$$crypt_{variant,typeVar} \circ wat_{variant,typeVar}$$

(b) distributed query

3. Challenges

- How to harness/integrate other PETs (e.g., differential privacy)
- Which kind of genetic data and analyses can be safely outsourced?
- What about new analyses?