Data-driven methods generating synthetic data in genomics: the HLA “avatars” are shifting paradigms in data sharing.

Estelle Geffard, Leo Boussamet, Thomas Goronflot, Sophie Limou, Nicolas Vince, Matthieu Wargny, Pierre-Antoine Gourraud

1- Université de Nantes, INSERM, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ATIP-Avenir, Equipe 5, Nantes, France

2- CHU de Nantes, INSERM, CIC 1413, Pôle Hospitalo-Universitaire 11 : Santé Publique, Clinique des données

UNIVERSITÉ DE NANTES

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Inserm
Institut national de la santé et de la recherche médicale

Introduction – Problem

- **GDRP is both a constraint and an opportunity**
  - Reassess how we manage biomedical data from patients
- **Genetic data is both pioneer and large scale**
  - Quantitative leap forward since the late 90's
  - 1.5 millions genotype (A/T/C/G, A/T/C/G)
  - ~ 80 euros /patient

Genetic data is not anonymous

- **Pseudonymous data** (ARTICLE 29 DATA PROTECTION WORKING PARTY)

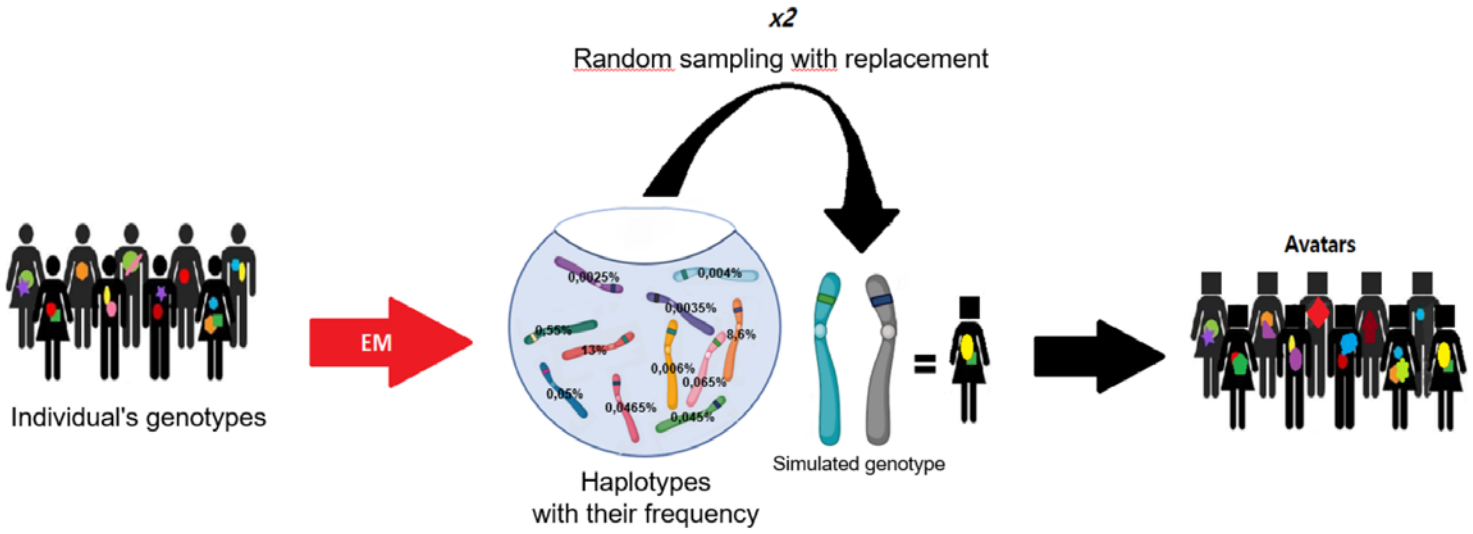
  (i) is it still possible to single out an individual ? Yes…

  (ii) is it still possible to link records relating to an individual ? Yes…

  (iii) can information be inferred concerning an individual ? Yes…

Solution : *Synthetic genetic data : Avatars*

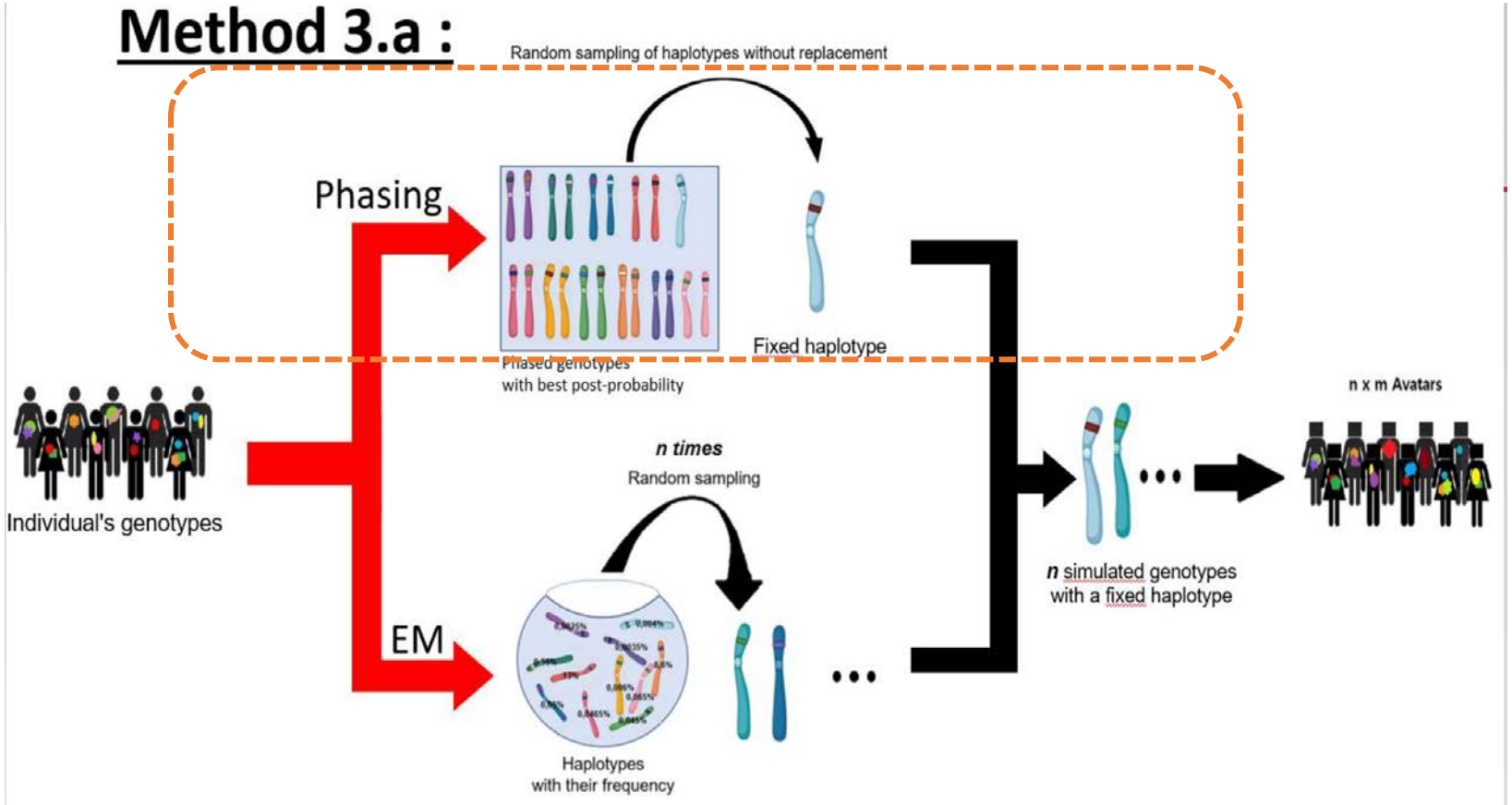
- "In silico Fertilization" vs. IVF : Computationally-assisted procreation
- **Application:** individual HLA genotype, Bone Marrow donor registries (26 millions worldwide)
- Method "1"



Methods :

- **Method 3a** : “Stochastic Filiation method”
- **Objective :** Keep link between **sensitive individual** and his.her **Avatar**

Filiation is achieved through the stochastic conservation of a single haplotype



Results 1: Re-identification #1

- Re-identification statistics :

Pourcentage d'avatars ayant exactement le même génotype qu'un individu de la population réelle initiale

- 4232 génotypes simulés :

	Method 1a	Method 1b	Method 2	Method 3a	Method 3b
Individus ré-identifiables (%)	0.071	0.047	0.047	0.071	0.118
Génotypes à MO* retrouvés dans la table d'avatars(%)	0	0	0	0	0

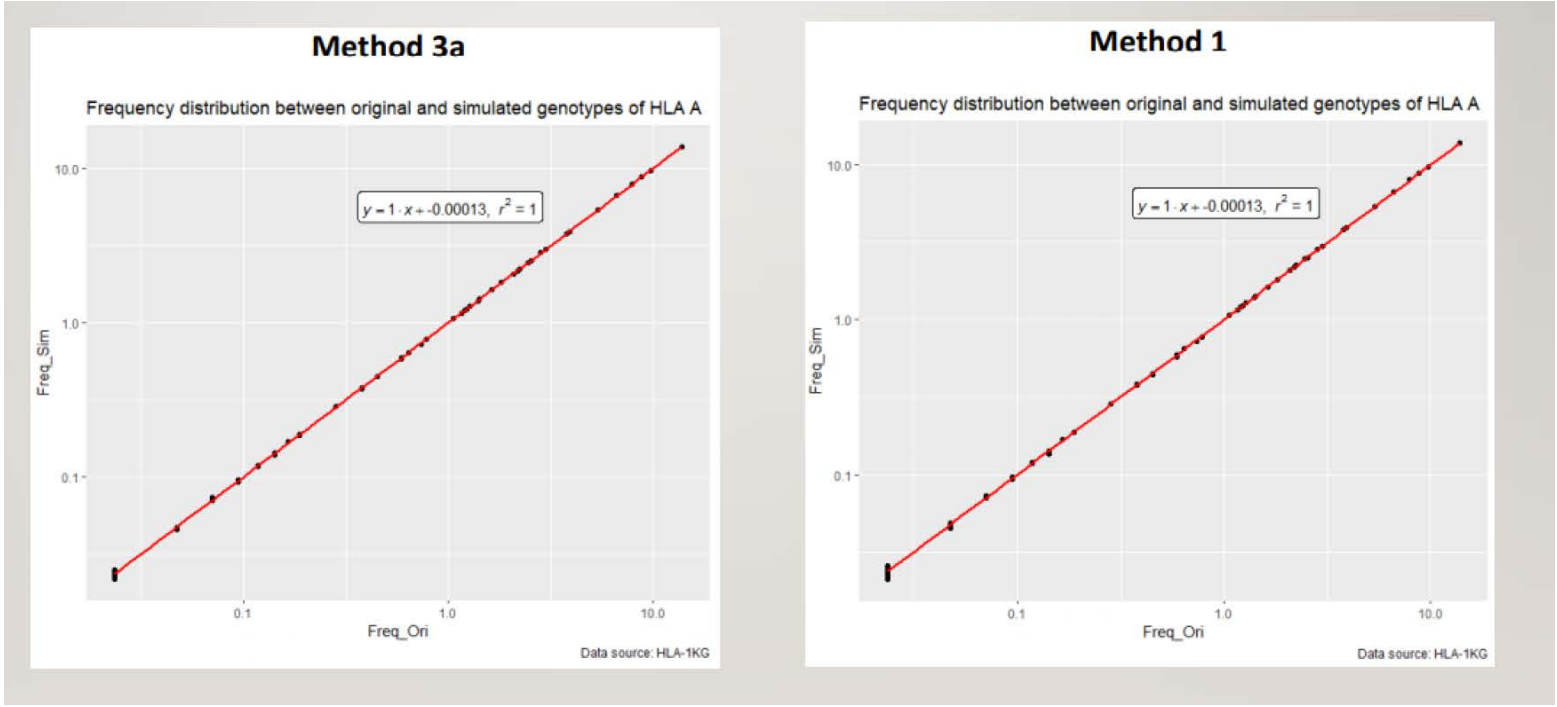
*MO = multiples occurrences

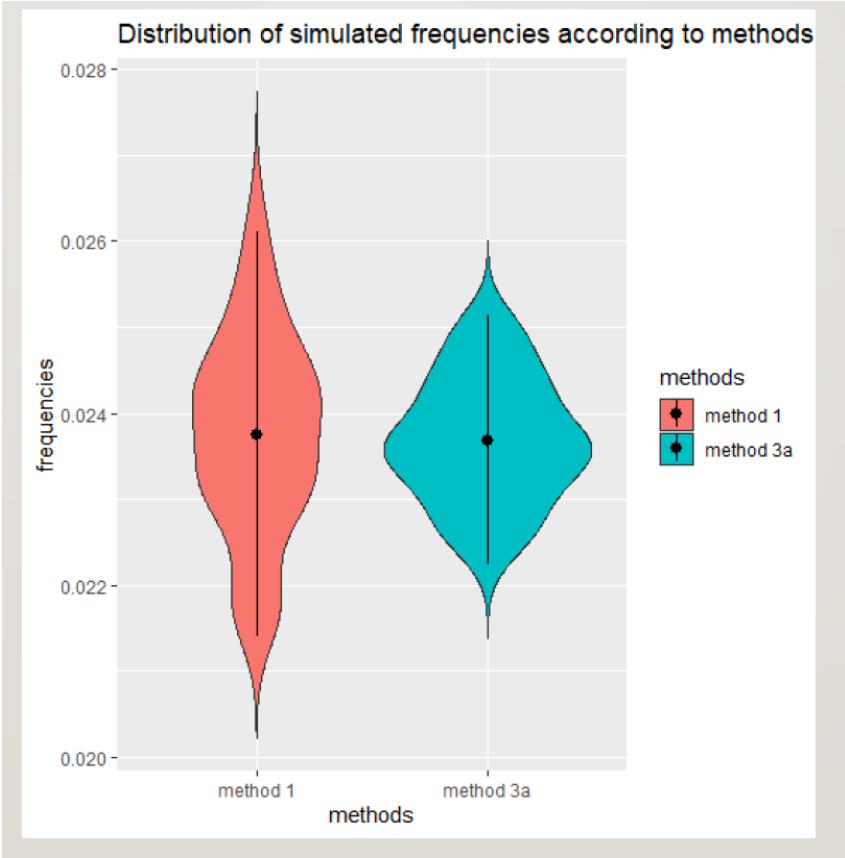
- 1058000, 2116000 et 4232000 génotypes simulés :

	Method 1a	Method 3a	Method 1	Method 3a	Method 1	Method 3a
Individus ré-identifiables (%)	$6*10^{-3}$	$5*10^{-3}$	$3,4*10^{-3}$	$3,6*10^{-3}$	$1,7*10^{-3}$	$1,7*10^{-3}$
Génotypes à MO* retrouvés dans la table d'avatars(%)	$2*10^{-4}$	$2*10^{-4}$	$9,5*10^{-5}$	$9*10^{-5}$	$4.7*10^{-5}$	$4.7*10^{-5}$

Results 2 : statistical relevance #2

- **Haplotype frequencies are well-conserved**





Discussion

- **Directed re-sampling**
  - Based on random re-identification statistics

	Method 1a	Method 1b	Method 2	Method 3a	Method 3b
Individus ré-identifiables (%)	0.071	0.047	0.047	0.071	0.118
Individus ré-identifiables après un premier remix (%)	0	0	0	0	0

  - A false-good idea – if multiple avatarisation can be achieved ..
    - " holes" are identifying

Conclusion

1. It is unethical to use potentially identify biomedical data for reseach…
   - Even high-dimension ones like genetic data
   - When computational techniques can deliver synthetic data of

   1- (virtually) no risk of re-identification

   2- good (enough) statistical relevance

2. Synthetic (genetic) data are out of the GDPR perimerter

1. There use of synthetic data that can be unethical

Data-driven methods generating synthetic data in genomics: *the HLA "avatars" are shifting paradigms in data sharing.*

Estelle Geffard, Leo Boussamet, Thomas Goronflot, Sophie Limou, Nicolas Vince, Matthieu Wargny, Pierre-Antoine Gourraud

1- Université de Nantes, INSERM, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ATIP-Avenir, Equipe 5, Nantes, France

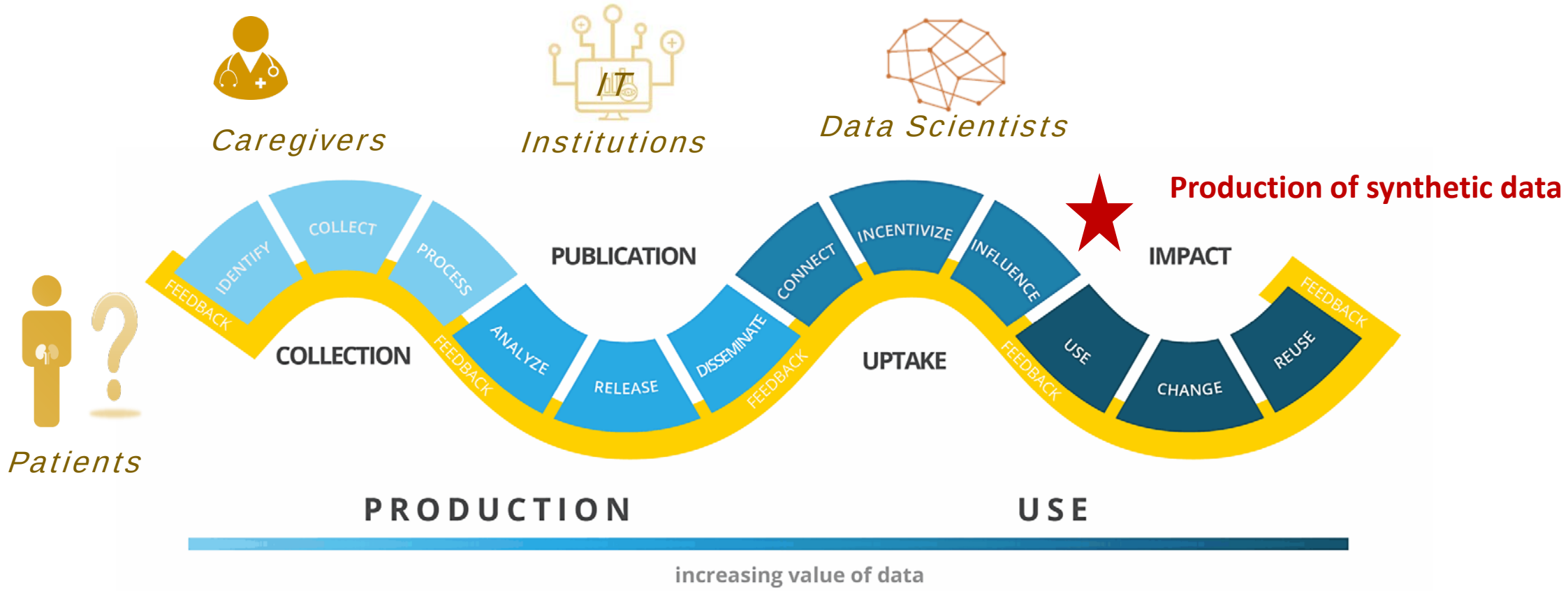
2- CHU de Nantes, INSERM, CIC 1413, Pôle Hospitalo-Universitaire 11 : Santé Publique, Clinique des données

UNIVERSITÉ DE NANTES

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Inserm
Institut national de la santé et de la recherche médicale

Multiples stakeholders and contributions to the data chain value.



(Adapted From https://opendatawatch.com)