State of the art(version 0)

1. Constructive Privacy for Shared Genetic Data:
2. **The Main Purpose:**

Genetic research which involves genetic data is often conducted collaboratively between several geneticians. The private aspect of genetic information pushed geneticians to make use of privacy-preserving means for the sharing of genetic data. Since the existing approaches, on handling shared genetic data and technologies to handle realistic collaborations, focus more on confidentiality aspect, in this paper, it is the ownership and integrity aspects that are dealt with.

1. **Motivations:**

One of the default procedures, falsely believed to be privacy-preserving is allele frequency. This procedure consisting of the allele frequency which uses alleles that are variant forms of a gene located in the same locations. So, an allele frequency represents the incidents of a gene variant in a population. It is calculated by dividing the total number of times the allele is observed in a population by the total number of copies of all the alleles in that genetic locus in the population. Allele frequencies as a percentage may be a reflection of genetic diversity. The Homer attack proved the non-privacy-preserving effect of a such procedure by proving that it is possible to deduce the presence of an individual in these collected data. These attacks put an end to the public access to such aggregated data.

One of the used solutions is a client/server architecture where the server uses the cloud for processing homomorphically encrypted data. Another consists of using multiple biomedical sites which make use of multi-party computation protocols.

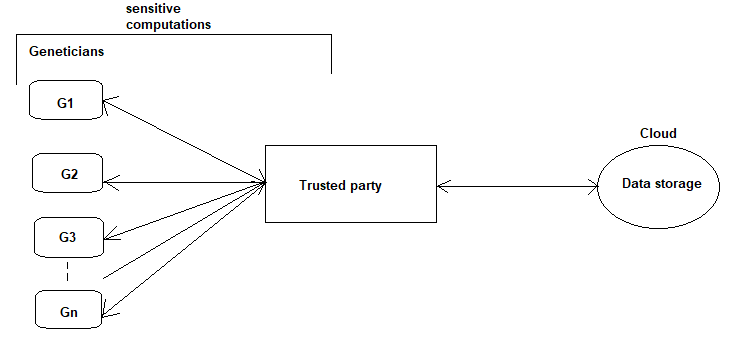
While these solutions remain limited in the context of handling a wider sharing of genetic data, some approaches have proposed the combination of these means including data fragmentation and client-sided computations.

=>The main approach consists of locally storing identifying genome components while computations on the publicly known parts are outsourced.

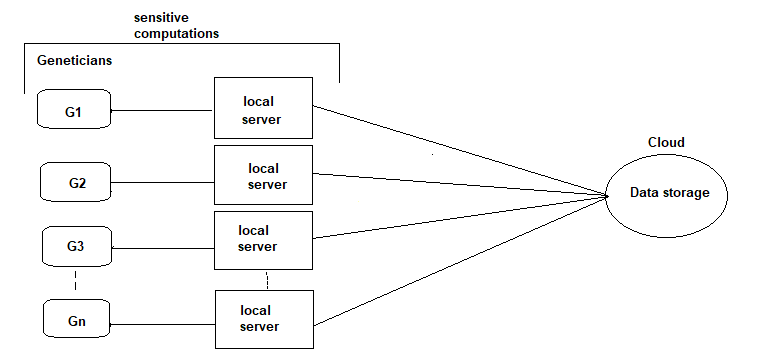
Since the confidentiality aspect is already dealt with in previous projects by using encryption and data fragmentation in addition to localized computations of sensitive data, these next reported solutions and documentation focuses on the integrity and ownerships aspects by presenting new architectures and ownership-preserving techniques.

1. **The Reported Solution(s):**

*Trusted party architecture* => In this architecture, genetic data is shared via a trusted party. The ownerships and integrity properties are important because geneticians read data and especially send data to the trusted party.



Local computation architecture => In this architecture, we may say that there is an intermediate between the Cloud and the geneticians, but this time this intermediate is the geneticians’ infrastructure: a local server. It is in these local servers where data is encrypted to be sent to the cloud.



These architectures allow the manipulation of shared genetic data with no guarantee of its privacy. The C2QL approach ensures the privacy property of data, while the new approach presented in this paper extends the C2QL one to ensure the ownership, integrity and traceability. This new extending approach is the COSHED(COntructive SHaring of gEnetic Data) approach. It makes use of the watermarking methods.

This approach applied on the architecture “Trusted Party Architecture” defined above has proven that the cloud, now, won’t be getting access to identifying data. In addition to this, integrity, ownership and traceability will be satisfied and geneticians won’t be able to get access to external identifying data.

***Notes:***

*Homomorphic encryption:* is one way of encryption which allows to perform calculation and the processing of encrypted data without having to decrypt it at first.Homomorphic encryption can be used for privacy-preserving outsourced [storage](https://en.wikipedia.org/wiki/Cloud_storage) and [computation](https://en.wikipedia.org/wiki/Cloud_computing). This allows data to be encrypted and out-sourced to commercial cloud environments for processing, all while encrypted.[[1]](#footnote-6789) This way of privacy-preserving data processing keeps the results encrypted as well.

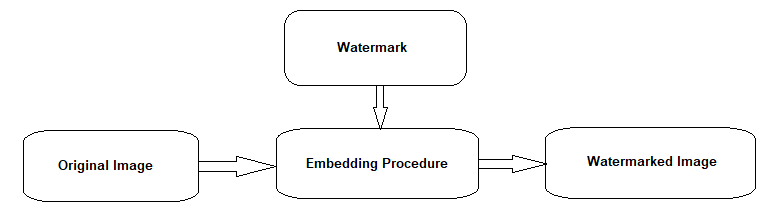
*Data Fragmentation*: Distributed Database systems provide distribution transparency of the data over the DBs. This is achieved by the concept called Data Fragmentation. That means, fragmenting the data is the process of dividing the database into smaller multiple parts. These fragments are stored at different locations. The data fragmentation process should be carried out in such a way that the reconstruction of original database from the fragment is possible.

*Genetic Data:* Genetic data should be defined as personal data relating to the inherited or acquired genetic characteristics of a natural person which result from the analysis of a biological sample from the natural person in question, in particular chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis, or from the analysis of another element enabling equivalent information to be obtained.[[2]](#footnote-28115)

*Digital watermark:* is a message which is embedded into digital content that can be extracted or detected later.

*Watermarketing:* is the insertion of imperceptible and inseparable information into data for data integrity.

*Digital watermarketing:* is the process in which digital information is hidden in a carrier signal. And the digital watermark is used, for example, to identify the ownership of the copyright of this signal. Digital watermarks may be used to verify the authenticity or integrity of the carrier signal or to show the identity of its owners. The hidden information should, but does not need to, contain a relation to the carrier signal. A signal may carry several different watermarks at the same time. Unlike metadata that is added to the carrier signal, a digital watermark does not change the size of the carrier signal.[[3]](#footnote-8482)



Data Outsourcing: is a new, emerging data management paradigm in which the owner of data is no longer totally responsible for its management. Rather, a portion of data is outsourced to external providers who offer data management functionalities. Secure data outsourcing is a discipline that investigates security issues associated with data outsourcing.[[4]](#footnote-30902)

1. Distributed Contextualization of Biomedical Data: a case study in precision medicine:
2. **The Main Purpose:**

Precision Medicine make use of large volumes of data containing diverse populations of patients from multiple centers. This medicine uses a distributed implementation for its patient-centered contextualization analyses, which lead to the importance of privacy and security aspects of shared data.

1. **Motivations:**

The spread of diseases is majorly caused by the social behavior of individuals. Thus, contextualization and the understanding of social behaviors plays an important role for the different decision-making organizations to make the adequate treatment. This contextualization solution applied to the KiTaPP (the kidney Transplantation Application) that is a research project in precision medicine that health-track patients after kidney transplantations, may help personalize future unanticipated renal outcomes by helping on a more personalized therapeutic decision for example. The aim in this project is the contextualization of possible renal after-effects of a POI over a POR (see Notes). Although, applying the contextualization in such a project is a vital advantage, this kind of medical service needs secure distributed systems to be deployed and used in a world-wide context of collaborations. However, precise medicine data may not be shared due to legal reasons. As a result, distributed analyses for contextualization purposes must be performed using aggregated, summarized or anonymized data.

1. **The Reported Solution(s):**

The KiTaPP contextualization algorithm consists of comparing data trajectories of a given POI to a sub-population with similar characteristics. This reference sub-population, in this application’s case, is defined based on the nearest neighbor approach by selecting individuals more similar to the patient and a clustering approach by selecting individuals in the same cluster as the POI. The contextualization uses a comparison between the evolution over time of the POI data and the POR (by median or percentile values) one.

The existing distributed analysis architectures and implementations do not guarantee the privacy needed for data located remotely that KitAPP researchers, for instance, may use. So, a new preserving-privacy architecture is needed for a larger range of potential analyses to be established.

**What is the privacy-preserving approach for a distributed analyses architecture for the case of the KiTaPP?**

The given approach makes use of an algorithm that requires few exchanges of data between different sites and the Master server. This algorithm applied to this application is an algorithm that enables parallel computing which requires few communications.

=>So, other than adding privacy methodologies for the protection of the privacy and confidentiality of data, the algorithm in its self can acquire fewer data exchanges and communications, which is the algorithm proposed in this case.

This algorithm was tested in a collaborative-distributed like architecture in an environment that consists of a grid-based one whose nodes may consist of individual machines, equivalent to medical-partner sites. In this case, it was the Grid’5000 platform that have been used.

While the percentage of shared data did not exceed 0.89% to each site knowing that shared data consisted only of aggregate values, thus no sensitive data was involved that may identify a patient, we can conclude that privacy is well preserved by the model in addition to high performance results.

***Notes:***

*Precision Medicine:* isan emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." This approach will allow doctors and researchers to predict more accurately which treatment and prevention strategies for a particular disease will work in which groups of people. It is in contrast to a one-size-fits-all approach.[[5]](#footnote-13967)

*Percentile*: A percentile is a term used in statistics to express how a score compares to other scores in the same set. While there is technically no standard definition of percentile, it's typically communicated as the percentage of values that fall below a particular value in a set of data scores. E.g. the kth percentile which is, in statistical terms, the greater than.

*Contextualization*: Contextualization represents a desirable blending of many of the concepts of patient-centeredness and cultural competence. Patient-centered care is “care that is respectful and responsive to individual patient preferences, needs and values, and ensures that patient values guide all clinical decisions.[[6]](#footnote-21570)

*POI*: a patient of interest

*POR*: a population of patients of reference, which is based on the identification of subgroups of patients with similar characteristics, such as age, gender body mass index.

*Graft*: a piece of living tissue that is transplanted surgically.

1. On distributed collaboration for biomedical analyses:

//TO\_ADD

1. **The Main Purpose:**

dfdfsqdf

1. **Motivations:**

fdfsfds

1. **The Reported Solution(s):**

Dfsqffd

***Notes:***

*sssfqdfdsf*

References:

* Distributed Contextualization of Biomedical Data: <https://hal.inria.fr/hal-02922930>
* Constructive Privacy for Shared Genetic Data: <https://hal.archives-ouvertes.fr/hal-01692620v2/document>
* On distributed collaboration for biomedical analyses: <https://hal.archives-ouvertes.fr/hal-02080463/document>

1. Homomorphic encryption: <https://en.wikipedia.org/wiki/Homomorphic_encryption> [↑](#footnote-ref-6789)
2. <https://gdpr-info.eu/recitals/no-34/#:~:text=Genetic%20data%20should%20be%20defined,(RNA)%20analysis%2C%20or%20from> [↑](#footnote-ref-28115)
3. <https://en.wikipedia.org/wiki/Digital_watermarking> [↑](#footnote-ref-8482)
4. <https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_328#:~:text=Data%20outsourcing%20is%20a%20new,who%20offer%20data%20management%20functionalities>. [↑](#footnote-ref-30902)
5. What is precise medicine? <https://medlineplus.gov/genetics/understanding/precisionmedicine/definition/> [↑](#footnote-ref-13967)
6. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4077271/#:~:text=Contextualization%20represents%20a%20desirable%20blending,clinical%20decisions%E2%80%9D%20(7>). [↑](#footnote-ref-21570)