The big data deluge in biomedicine: addressing the privacy vs. sharing dilemma

Paulo Verissimo and Jérémie Decouchant
CritiX Lab - Critical and Extreme Security and Dependability
SnT - Interdisciplinary Centre for Security, Reliability and Trust
University of Luxembourg
Email: ⟨name⟩.⟨surname⟩@uni.lu

***Abstract*—This position paper discusses on-going work on architectures and algorithms for efficient but privacy-preserving storage and analysis of bulk biomedical data.**

I. Introduction

Biomedical information is enduring a revolution: collection and storage of biological material is getting systematic (tissues, fluids, etc.), both for clinical and research purposes; digital representations of these samples are exploding in volume, especially in genomics (DNA) thanks to a large extent to the advent of the so-called Next-Generation-Sequencing (NGS) machines. These machines have lowered the price and increased the speed of sequencing, by several orders of magnitude. Over the past few years, the number of sequenced genomes or exomes (parts thereof) has sky-rocketed, and the trend is there to continue. Bottomline: (i) full-genome sequencing for less than one thousand euros is becoming a reality; and (ii) the life-cycle of the physical biological material stored in biobanks is too expensive to sustain the current growth. In consequence, every day we have quite a few additional dozens of terabytes of stored raw digital data, and the combined scale of biomedical datasets and related data from their stakeholders, has in fact entered the row of big data.

II. Problems on the horizon

There are fantastic opportunities in this new world, but there are at least as many threats and challenges. We just enumerate the few ones we consider of most importance, starting by introducing their causes.

First, the need for economically storing and processing these huge amounts of data has put cloud computing on the agenda, inclusively by NGS machine vendors [1]. Not just any cloud, but including public clouds, and not just very restricted access, but including Internet web-based access [1]. Using common (and moderate security) IT techniques to manipulate such critical data, brings about considerable security and privacy risks whose likelihood and impact have perhaps not been correctly evaluated so far, given the recurring failures in the internet/cloud complex [2]–[4].

Second, a dramatic increase of the availability of personally identifiable information (PII) has been occurring in parallel, due not only but largely to concurrent effects like: the (sometimes forced, e.g. by governments or companies) digitalization of society activity; the web in general; and social networks activity in particular. Our lifes leave an evergrowing indelible digital trace, and this has had an effect whose consequences we are still starting to comprehend: big data in this case means that there are too many data around, and statistical correlations which were infeasible a few years ago, become trivial, and with astonishing precision and recall. Recent happenings go from real-life episodes as reported in [5], to impressive re-identification of private credit card operations, as in [6]. But research results in the context of biomedical data are much more worrying. In 2000, an alarm was raised [7], by demonstrating the re-identifiability of de-identified patient-specific medical data, and thus showing the ineffectiveness of the de-identification methods used. Thirteen years later, not much had changed in the meantime, since the work in [8] managed to re-identify more than 10% of the de-identified 1000-Genomes project database, generously built from anonymous donors' sequenced DNA, again because useful correlations could be found with the judged anonymous metadata in the genomes database.

Last but not least, there is a great pressure to get hold of biomedical data, by reasons of different nature, and coming from diverse angles, such as researchers, corporations and even governments. This confluence of interests is sometimes detrimental of the investigation and deployment of clear and sound strategies, policies and technologies that help solve the problem at hand in this paper: addressing the privacy vs. sharing dilemma we face with regard to biomedical data.

The remainder of the paper discusses some contributions in that line.

III. A framework for solutions

In a recent paper [9], we advanced an architectural framework to guide possible solutions to a range of problems around the biomedical data scenario. We predicted the advent of the new era of *e-Biobanking*, in terms of the "creation of genuine hybrid ecosystems composed of interplaying physical and computer storage and processing infrastructures, handling physical and computerized samples of biological data, in a symbiotic and seamless way". The vision, depicted in Figure 1, foresaw that systems originally decoupled from each other (Fig. 1a), would progressively evolve toward coalitions of interested stakeholders, organised around hybrid and/or federated cloud computing technologies (Fig. 1b), and where techniques

This work is partially supported by the University of Luxembourg - SnT and by the Fonds National de la Recherche Luxembourg (FNR) through PEARL grant FNR/P14/8149128.

[1] See, e.g., BaseSpace from Illumina http://tinyurl.com/zwjtqts

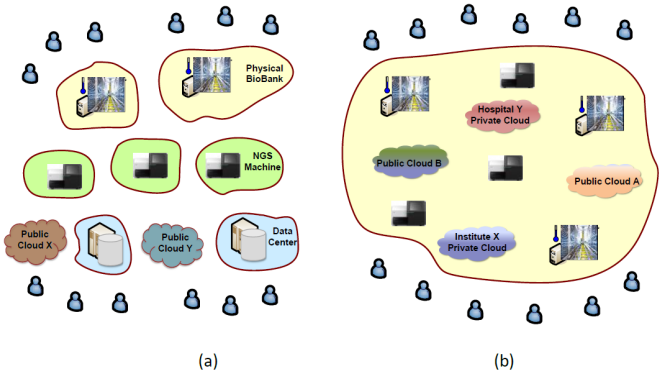

Fig. 1. The *e-Biobanking vision*

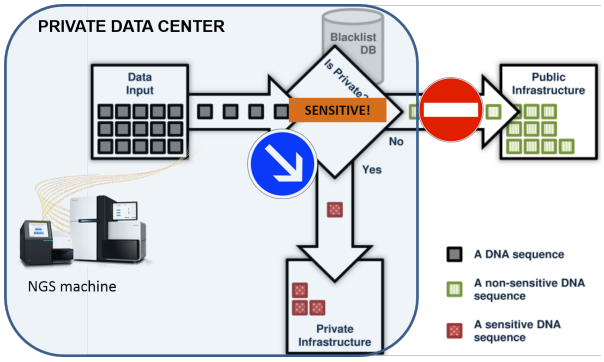

Fig. 2. Privacy filtering: automatic segregation of NGS data

safeguarding security and dependability might be effectively deployed, whilst promoting the desired sharing.

The vision has been slowly coming to fruition. For example, the BioBankCloud project proposes storage architectures which are based on clouds-of-clouds, that is, multiple instances of clouds, both private and public, from several stakeholders and providers, but which are perceived seamlessly as a single cloud, by the e-biobank users and administrators [10].

In [11], researchers propose a federated system enabling the cooperative analysis of NGS sequences across multiple locations. Despite the security mechanisms being at the level of standard IT, this is an interesting step forward.

In another work [12], the authors propose a privacy-preserving method where they encrypt all genomic data before storing it, and manually mask the few genomic variations identified as sensitive. Despite the possible coverage gaps and errors of the manual process, and update problems, this method yields prevention at early stages of the life-cycle.

In some recent work, we and colleagues at the University of Lisboa, have proposed a high-throughput method to automatically segregate genomic information right after it is created, that is, at the exit of NGS machines [13]. The proposed scheme, depicted in Figure 2, fits the e-biobanking vision nicely: very sensitive data is kept within the private premises of the entity generating the data, e.g., a secure data center next to the NGS machines subsystems, segregated from the outside, whereas less sensitive data can be stored in cloud systems of a lower privacy/security category. All this is done automatically, through a rule-based system acting pretty much like an intrusion detection system, and with high-throughput, by resorting to a Bloom filter.

Part of this data, despite being less sensitive, may still have considerable privacy requirements, and can for example be stored in public cloud-of-clouds systems, protected with powerful state-of-the art encryption, coding and dispersion mechanisms, such as those proposed in [14].

IV. Conclusions and Future Work

Relying on a public cloud to store and process the massive biomedial data production presents privacy issues. However, it is of prime importance to allow requests of, or over, biomedical data to be treated, e.g., for research or medicine purposes, as soon as no sensitive information is leaked.

We are extending the e-biobanking vision, recently proposed to overcome these limitations, in several aspects. First, we are incorporating the sequence analysis subtasks (e.g., alignments of reads to a reference) to handle and protect genomic data as soon as it is produced. Second, we consider function-shipping and data-shipping methods to issue queries. Important challenges include maintaining performance and functionalities, enforcing security and preserving privacy.

References

[1] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, vol. 11, no. 5, p. 207, 2010.

[2] P. Judge, "Google has cloud failure, applies quick fix," 2015. [Online]. Available: http://tinyurl.com/hd7juq5

[3] D. Streitfeld, "Google concedes that drive-by prying violated privacy," 2013. [Online]. Available: http://tinyurl.com/zbwjbvr

[4] C. Brooks, "Cloud storage often results in data loss," 2011. [Online]. Available: http://tinyurl.com/gn92hjv

[5] G. Lubin, "The incredible story of how target exposed a teen girl's pregnancy," 2012. [Online]. Available: http://tinyurl.com/q37hk35

[6] Y.-A. de Montjoye, L. Radaelli, V. K. Singh *et al.*, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[7] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, pp. 1–34, 2000.

[8] M. Gymrek, A. L. McGuire, D. Golan *et al.*, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[9] P. E. Verissimo and A. Bessani, "E-biobanking: What have you done to my cell samples?" *Security Privacy*, vol. 11, no. 6, pp. 62–65, 2013.

[10] A. Bessani, J. Brandt, M. Bux *et al.*, "Biobankcloud: a platform for the secure storage, sharing, and processing of large biomedical data sets," in *Workshop on Data Management and Analytics for Medicine and Healthcare*, 2015.

[11] A. Ardeshirdavani, E. Souche, L. Dehaspe *et al.*, "Ngs-logistics: federated analysis of ngs sequence variants across multiple locations," *Genome Medicine*, vol. 6, no. 9, pp. 1–11, 2014.

[12] E. Ayday, J. L. Raisaro, U. Hengartner *et al.*, "Privacy-preserving processing of raw genomic data," in *Data Privacy Management and Autonomous Spontaneous Security*, 2014, pp. 133–147.

[13] V. V. Cogo, A. Bessani, F. M. Couto *et al.*, "A high-throughput method to detect privacy-sensitive human genomic data," in *ACM Workshop on Privacy in the Electronic Society*, 2015.

[14] A. Bessani, M. Correia, B. Quaresma *et al.*, "Depsky: Dependable and secure storage in a cloud-of-clouds," in *Proceedings of the Sixth Conference on Computer Systems*, 2011.