Cloud-Assisted Read Alignment and Privacy

Maria Fernandes[1], Jérémie Decouchant[1], Francisco M. Couto[2], and Paulo Esteves-Verissimo[1]

[1] SnT – Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg

[2] LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract. Thanks to the rapid advances in sequencing technologies, genomic data is now being produced at an unprecedented rate. To adapt to this growth, several algorithms and paradigm shifts have been proposed to increase the throughput of the classical DNA workflow, e.g. by relying on the cloud to perform CPU intensive operations. However, the scientific community raised an alarm due to the possible privacy-related attacks that can be executed on genomic data. In this paper we review the state of the art in cloud-based alignment algorithms that have been developed for performance. We then present several privacy-preserving mechanisms that have been, or could be, used to align reads at an incremental performance cost. We finally argue for the use of risk analysis throughout the DNA workflow, to strike a balance between performance and protection of data.

Keywords: read alignment, cloud computing, genomic data privacy

1 Introduction

Genome sequencing evolved at an unprecedented rate with the advances of Next-Generation Sequencing (NGS) technologies. These new technologies allowed the sequencing costs to fall down to less than $1000 per genome, the machines throughput to increase from MB to TB of raw data produced per day, and the development of optimized parallelized procedures [19]. Medicine and biomedical research are benefiting from this evolution and started including sequenced data in their workflows [5]. However, to produce more comprehensive analysis using the large amount of NGS data generated, clinical and research entities faced new technical challenges. Indeed, they now have to share data and collaborate to improve the quality of their studies and the development of larger datasets [13].

Going further than traditional sharing schemes, domain experts established the e-biobanking vision [4], which calls for multi-research environment models and architectures facilitating the sharing of data. However, biomedical data (e.g. genomic sequences, medical reports, diseases information) is sensitive, as it is unique for each person and reveals information about herself and her relatives (e.g., predispositions to diseases). Therefore, a collaborative environment needs not only to enable the storage, the access and the analysis of biomedical data, but also be secure and reliable. Developing such an environment still remains a challenge.

As this integrated environment does not yet exist, scientists mostly relied on clouds to store and analyse sequencing data, due to their data sharing platform and improved computing schemes. However, the question remains on their ability to store and exploit genomes without breaking privacy policies. Despite the best efforts of cloud providers, the challenge is now set to accurately determine a threshold between the privacy and the openness of genomic data [23].

In this paper, we focus on the first step of the DNA analysis workflow — read alignment — which finds the location of a sequenced portion of DNA or RNA in a reference sequence. Section 2 summarizes the privacy-related features of genomic data, and describes the privacy attacks that have been presented in the literature, highlighting the importance of protecting genomic data. Section 3 describes cloud-based alignment algorithms which first emerged in response to the fast growth of sequenced data, highlighting their lack of consideration for privacy. Section 4 introduces the more costly algorithms that have been developed with privacy in mind. Finally, Section 5 gives some final remarks for the development of genomic data protective cloud environments, and argues for a risk-scale analysis that would be both practical and efficient. Section 6 concludes this paper.

2 Privacy attacks on genomic data

Protecting genomic data is a non-trivial task, due to its many specificities which have been exploited in recent attacks. The attacks performed in order to obtain private information from genomic data all rely on one or several of the following characteristics.

Long-lived and static data. Genomic data stays sensitive at least as long as her owner lives, and contains particular properties, which make standard encryption mechanisms insufficient to protect it on the long term. Furthermore, once the privacy of genomic data has been compromised, there is no way to recover it, as the genome of a subject evolves very slightly during her life.

Hereditary information. Genomic information is transmitted from generation to generation. Thus, privacy leaks also affect the relatives of a victim.

Revealing diseases risk. Hereditary diseases are embedded in genes. The possession of even parts of a person's genome makes it possible to infer about her/his risk to develop certain disease. This information can lead to discrimination, for example, an employer might not offer a job to someone suffering from a chronic disease, or a health insurance could be denied to a person whose genome revealed a high risk to develop a disease. The same can occur with mortgage, if a person has a disease which decreases her lifespan.

Revealing personal response to medicines and risks to diseases. Prospects of personalized medicine show the benefits of using genomic information to adapt a patient's treatment to his particular expected reactions to it.

However, this information could also be used for less glorious goals, since knowing the patient's reaction to a set of medicines can expose potential weaknesses.

Prone to manipulation. Ongoing research led to the belief that in the near future it will be possible to artificially recreate the DNA of any sequenced subject. As DNA samples are now used in forensics investigation to study crime scenes, artificial DNA samples could be introduced to influence investigations. This practice would compromise the ongoing investigations, by obfuscating any potential result or worse, lead to a wrong accusation.

In practice, the approach that has been followed by the existing platforms or services that work on genomic data until now has been a reactive one: data is made available and once a new attack is discovered, sensitive data is removed from public access. Several privacy-related attacks have been studied and described in the literature, we summarize them here.

Identification attacks are performed to determine the relation between the DNA profile of an individual and a data set. Taking as example a disease study case, an identification attack would reveal if a person is in the case or control data set, therefore breaking the privacy policies. Such an attack would typically reveal that a subject has a given disease [11].

Identity tracing attacks use records of genetic information and personal published information, which is available, for example, on genealogical databases (Ancestry[3]), diseases studies databases (DisGenet[4]), and surnames databases (e.g. Surname Navigator[5]). In the past, those databases reacted to reported attacks — such as the one determining Dr. Watson's APOE gene status [18] or the one using identification by surname inference [10] — by removing the detailed information used for the concerned attack from the database.

Recovery attacks determine a subject's sensitive genomic sequences using statistics and frequency information combined with released sensitive data (e.g. single nucleotide polymorphisms). Once the sequence is known, the attacker can use this information to launch the two previously mentioned attacks [25].

These attacks alerted the research community and the databases administrators of the possible data privacy threats. However, they cannot protect genomic data against future unknown attacks, as an attacker could collect and save data, and run an attack on it once it has been made public. Therefore, several privacy-preserving approaches to handle genomic data have appeared, which propose to protect data preventively.

In the next section, we discuss the existing cloud-based alignment solutions that the scientific community has adopted in order to leverage their high throughput, and we study them from a privacy-related point of view.

[3] Ancestry – https://www.ancestry.com
[4] DisGeNet – http://www.disgenet.org
[5] Surname Navigator – http://www.surnamenavigator.org

3 Alignment in the cloud

Aligning reads to a reference genome is one of the most important steps, and the first, of the sequencing analysis workflow that ultimately leads to genomic insights. Due to the throughput of NGS technologies and computational resources of research centers being unable to follow it, reads alignment is now often a bottleneck [21] and traditional algorithms, like BLAST [2] cannot be used as is. Hence, researchers started to offload the alignment of reads to cloud providers. Clouds are scalable computing infrastructures that allow users to adapt the resources they use to each of their analysis. These infrastructures allow users to benefit from their important computational power and storage space provided on demand through a simple internet connection, and at a manageable cost.

Several popular alignment tools have been adapted to run in clouds using Hadoop's MapReduce to execute code in parallel (Cloud-MAQ [22], CloudBLAST [15]). MapReduce's performance can be affected by the large amounts of data that has to be uploaded in the cloud, before executing the processing step. In addition, this data transfer increases cloud storage costs and causes increased latencies. This main limitation of MapReduce algorithms can be partly addressed using stream processing engines, which have also been explored in combination with alignment algorithms. Kienzler et al. [14] proposed a stream-based sequence analysis approach where the transfer step is replaced by data streaming, thus avoiding the huge amount of data transfer. Even though streaming approaches improved performance, since they apply data compression and decompression, read alignment remains computationally intensive and time consuming.

Although cloud processing improves performance and provides more storage space, it poses security concerns. A cloud infrastructure is controlled by a Cloud Service Provider (CSP), which does not provide the users full control over their own data. Additionally, CSPs can copy, transfer and store the data into multiple-locations (for fault-tolerance or economic reasons), and do not guarantee that the data cannot be accessed by the CSP or an intruder [20]. Thus, researchers need to consider a cloud as an untrusted, and possibly insecure, environment. To deal with genomic data on clouds, researchers and CSPs should discuss and adapt the privacy policies (eg. data control, security, confidentiality, transferring) to guarantee data protection [8].

Cloud computing offers the best solution in terms of modularity concerning computational power and costs to analyse large quantities of data. However, the algorithms described in this section require the client to upload his data into the cloud, where it is treated in plain text (i.e., without using any encryption mechanism). Considering that the user-cloud communications are made via internet, where communications could be intercepted and genomic data decrypted, given enough time, and the trust we give to the cloud provider, using such an infrastructure presents privacy issues which need to be addressed.

4 Privacy-preserving alignment in the cloud

Privacy-preserving methods for execution in the cloud can involve cryptographic or non-cryptographic mechanisms and client-CSP agreements that must be followed. Both, client and CSP need to be aware of the sensitivity of biomedical data to ensure the adequate privacy protection [1]. In this section, we introduce the current privacy-preserving methods that could be applied to biomedical data, and then present real-life applications of these mechanisms.

The non-cryptographic techniques include data anonymization, the control of accesses, and privacy agreements.

Data anonymization consists in removing the personal information (e.g. name, surname, birth data, address, age) to avoid direct associations between genomes and their donors. Some portions of genomes have been considered privacy-critical information as well [10], which raises the challenge of identifying such genomic portions.

Access control consists in specifying who is allowed to access the data, often with different access levels, to limit and track its usages. For example, a medical center may have access to the disease genes of a patient and another research unit would only have access to the genes related to a particular disease under investigation [9].

Privacy agreements are signed documents specifying that a donor grants access to his data. All the entities (e.g. donor, researcher, medical institution) that can access the data sign the agreement, and it is assumed that all the concerned parties are trustworthy. Historically, privacy agreements were the first privacy-preserving technique developed around genomic data. However, the necessary uses of untrusted machines and communications links render privacy agreements unable to fully protect data.

These three methods are considered insufficient to protect genomic data alone, however it is believed that when combined with cryptographic privacy-preserving techniques they increase the protection of sensitive data [9]. Cryptographic techniques provide high privacy guarantees to very specific scenarios. However, the scientific community has been working towards extending their range of applicability to study genomic data.

Keyed-hash functions convert clear-text to hashes and combine them with a secret key. This technique however relies on the assumption that the key is never stolen, since in that case all the data would be accessible [6]. In addition, this approach does not allow direct collaboration between multiple entities.

Differential privacy introduces randomness to the input of a function in order to protect its privacy-sensitive features. Intuitively, the output of a function must not vary much whether an individual is part of the study or not. The main issue of this technique is to control the amount of randomness introduced in human genomes, so that studies can produce meaningful results [17, 23].

Garbled circuits are a cryptographic technique for two-party secure computation. This technique allows a user to send his data to a receiver (e.g. cloud service) to make some computations and receive back the final output. During this process, neither the input nor intermediate values are revealed [3].

Lastly, **homomorphic encryption** schemes have been explored as a security method for genomic data. These schemes allow a computation to be executed on encrypted data, and its result to be decrypted, therefore providing insight on the plaintext data. However, their performance is currently unsatisfactory and it only allows a limited number of operations [3].

Several privacy-preserving cloud alignment solutions have been recently published. Those solutions rely on hybrid clouds environments where the most sensitive data computations are performed on a private cloud and the less sensitive is processed on a public cloud [24]. Some solutions apply keyed-hash functions on the sensitive data and then send the hash-values to the cloud [6]. Homomorphic encryption has also been applied on other steps of the analysis of genomic data, e.g. for disease susceptibility tests [16]. However, these examples still present some limitations: the most CPU intensive task (i.e., the extend step) has to be performed in the private cloud; the need of an efficient and reliable sensitive data classifier; the use of hash algorithms that may be broken before the expiration of the genomic data they protect.

5 Towards a differentiated protection of genomic data

Several privacy-preserving methods have been developed, however their limited usability stills cannot address all the different issues found in the workflow analyses steps. In this section, we describe how classifying the sensitivity of genomic data would contribute to a thorough use of the potential of existing algorithms, at the best possible cost.

Enabling technologies. A filtering approach that classifies reads as embedding sensitive or non-sensitive information has been described in [7]. Adding this filtering step would allow the reduction of data encryption costs by encrypting only the critical information and improve the data usability, while ensuring the protection of genomic data. In addition, the level of sensitivity of reads could be determined according to the attack power it provides to an attacker through a risk-analysis study. Doing so, however, requires further work. We are convinced that such approaches will be developed in the future, and now present the benefits they would bring to different stages of the DNA workflow.

Privacy-preserving alignment can be obtained in mainly two ways: rely on plaintext conventional algorithms in a secure environment (e.g. local computer, private cloud) [22, 14], or protect data through cryptographic methods. In the former there is always a risk for an adversary to get access to the machines, and therefore to the sensitive data. The second solution can be too costly or even unpractical since encryption makes data unavailable for some operations [12, 3]. Classifying data into sensitivity levels would allow both approaches to be combined, globally improving performance, as the more-costly algorithms would be applied only to the most sensitive data, while improving the performance of the low-sensitivity reads.

Storage security requires long-term protection techniques. The most sensitive data could be stored in highly restricted and protected areas, while less sensitive data could be stored encrypted on the cloud. Splitting data and differentiating the way it is stored based on its sensitivity would reduce the storage costs as the most secure environments are usually more costly.

Release of and access to sensitive data require an extensive understanding of genomic data privacy breaches. For a privacy protective data release it is, of course, necessary to hide all the unique individual information (e.g. names, address, genes) [25]. Differentiating the sensitivity of genomic data would allow more data to be released to scientists, while the most sensitive one would still be protected. Data aggregation was also purposed as a secure solution for data release, however it remains in a early stage of understanding and application. For example, a human genome contains around 10 million single nucleotide polymorphisms (SNPs), and therefore a secure aggregate of full genomes would have to involve more than 80 millions of subjects [25] ($\approx$1.15% of the world population).

6 Conclusion

The migration of read alignments to the clouds and the parallelization of the process using MapReduce, have greatly improved the performance of this essential step of the DNA workflow. However, these solutions require data to be manipulated in plain-text in the cloud, which poses privacy concerns, which were highlighted by the genomic privacy attacks reported in the last years. As researchers became more aware of those data vulnerabilities, the last years saw the development of privacy-preserving solutions to replace the typical alignment algorithms, which are deprived of privacy measures. Until now, it seems that privacy protection and performance are inversely related, since the improvement of one leads to the decrease of the other. Thus, the golden question is how to provide data privacy protection while taking advantage of the storage and computational power that cloud environments provide. Accurately determining the level of sensitivity of genomic information seems to be a way to go to benefit entirely for the broad range of algorithmic, storage and access solutions that have been developed. Such a secure cloud environment for biomedical data analysis is still an open challenge.

Acknowledgements. This work was supported by the Fonds National de la Recherche Luxembourg (FNR) through PEARL grant FNR/P14/8149128, and by the Fundação para a Ciência e para a Tecnologia (FCT) through funding of the LaSIGE Research Unit, ref. UID/CEC/00408/2013.

References

[1] Akgün, M., Bayrak, A.O., Ozer, B., et al.: Privacy preserving processing of genomic data: A survey. Journal of biomedical informatics 56, 103–111 (2015)

[2] Altschul, S.F., Gish, W., Miller, W., et al.: Basic local alignment search tool. Journal of molecular biology 215(3), 403–410 (1990)

[3] Baron, J., El Defrawy, K., Minkovich, K., et al.: 5pm: Secure pattern matching. SCN pp. 222–240 (2012)

[4] Bessani, A., Brandt, J., Bux, M., et al.: Biobankcloud: a platform for the secure storage, sharing, and processing of large biomedical data sets. DMAH (2015)

[5] Chan, I.S., Ginsburg, G.S.: Personalized medicine: Progress and promise. Annual Review of Genomics and Human Genetics 12(1), 217–244 (2011)

[6] Chen, Y., Peng, B., Wang, X., et al.: Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. NDSS (2012)

[7] Cogo, V.V., Bessani, A., Couto, F.M., et al.: A high-throughput method to detect privacy-sensitive human genomic data. ACM WPES pp. 101–110 (2015)

[8] Dove, E.S., Joly, Y., Tasse, A.M., et al.: Genomic cloud computing: legal and ethical points to consider. Eur J Hum Genet 23, 1271–1278 (2015)

[9] Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nature Reviews Genetics 15, 409–421 (2014)

[10] Gymrek, M., McGuire, A.L., Golan, D., et al.: Identifying personal genomes by surname inference. Science 339(6117), 321–324 (2013)

[11] Homer, N., Szelinger, S., Redman, M., et al.: Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS Genetics 4(8), e1000167 (2008)

[12] Huang, Y., Evans, D., Katz, J., et al.: Faster secure two-party computation using garbled circuits. USENIX Security Symposium 201(1) (2011)

[13] Kaye, J., Heeney, C., Hawkins, N., et al.: Data sharing in genomics re-shaping scientific practice. Nature Reviews Genetics 10(5), 331–335 (2009)

[14] Kienzler, R., Bruggmann, R., Ranganathan, A., et al.: Large-scale dna sequence analysis in the cloud: a stream-based approach. ICPP 2, 467–476 (2012)

[15] Matsunaga, A., Tsugawa, M., Fortes, J.: Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications. ESCIENCE '08 pp. 222–229 (2008)

[16] Namazi, M., Troncoso-Pastoriza, J.R., Pérez-González, F.: Dynamic privacy-preserving genomic susceptibility testing. ACM MMSec pp. 45–50 (2016)

[17] Naveed, M., Ayday, E., Clayton, E.W., et al.: Privacy in the genomic era. ACM CSUR 48(1), 1–44 (2015)

[18] Nyholt, D.R., Yu, C.E., Visscher, P.M.: On jim watsons apoe status: genetic information is hard to hide. European Journal of Human Genetics 17, 147–149 (2009)

[19] O'Driscoll, A., Daugelaite, J., Sleator, R.D.: "big data", hadoop and cloud computing in genomics. Journal of Biomedical Informatics 46(5), 774–781 (2013)

[20] Rocha, F., Correia, M.: Lucy in the sky without diamonds: Stealing confidential data in the cloud. DSNW pp. 129–134 (2011)

[21] Stein, L.D.: The case for cloud computing in genome informatics. Genome Biology 11(5), 207 (2010)

[22] Talukder, A., Gandham, S., Prahalad, H., et al.: Cloud-maq: The cloud-enabled scalable whole genome reference assembly application. WOCN pp. 1–5 (2010)

[23] Vayena, E., Gasser, U.: Between openness and privacy in genomics. PLoS Medicine 13(1), 1–7 (2016)

[24] Zhang, K., Zhou, X., Chen, Y., et al.: Sedic: Privacy-aware data intensive computing on hybrid clouds. ACM CCS pp. 515–526 (2011)

[25] Zhou, X., Peng, B., Li, Y.F., et al.: To release or not to release: Evaluating information leaks in aggregate human-genome data. ESORICS pp. 607–627 (2011)