Privacy-preserving algorithms and workflows for next-generation genomics

Maria Fernandes
SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
maria.fernandes@uni.lu

Jérémie Decouchant
SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
Jeremie.Decouchant@uni.lu

Francisco M Couto
LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa
fjcouto@ciencias.ulisboa.pt

Paulo Esteves-Veríssimo
SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
paulo.verissimo@uni.lu

1 RESEARCH CONTEXT

The recent evolution of the DNA sequencing technologies has lowered the sequencing cost. The increased availability of sequenced genomes prompted their application over a wide range of fields, such as personalized medicine, biomedical research [13] and forensics [11]. This contributed to a world-wide distribution of genomic sequenced data and raised the need for sharing and collaboration. The requirement of more complete and significant datasets drove the research community to work towards a distributed genomic ecosystem. The e-biobanking vision describes a secure federated environment where all the participants communicate and share their genomic data [4].

However, storing and sharing genomic data raises privacy challenges, since the human genome carries personal identifiable information. The described attacks on genomic data [7, 8, 14], highlighted the need for strong privacy protection mechanisms, and made the research community aware of the necessity of developing more efficient privacy protection methods [17].

In order to work toward a distributed genomic ecosystem and tackle the privacy challenges, partial privacy-preserving solutions for very specific applications have emerged in the last years. Although privacy-preserving solutions have been developed [2, 10, 15], there are still unsolved challenges, such as, but not limited to: (i) impossibility to provide both privacy protection and high performance, since protective methods are slower than the high performance methods, which require plaintext information [15, 16]; (ii) data protection is limited, since privacy protecting methods (e.g. cryptography-based methods) usually can be broken in a shorter time than genomic data privacy requires [9]; (iii) data utility, since many cryptographic methods allow limited operations on the encrypted data, which can impair the data utility [3].

2 GENOMIC WORKFLOW

The traditional genomic workflow consists of two main steps: sequencing, and analysis. On the sequencing step the biological samples (e.g. blood sample) are translated in sequences of nucleotides (i.e., A, T, C, G) called reads. During the analysis step the reads are first aligned to a reference genome to determine their position, and then a variant calling step is performed. More particularly, variant calling consists on aligning the reads to the locations they belong and observe if they contain different nucleotides at some location (see Fig. 1(a), step on the right). In Fig. 1(a), on the variant calling step, the letters in red represent the variants, which differ from the reference (sequence in bold). After the variant calling step, the results obtained are used for different studies to obtain the biological insights.

3 RESEARCH GOALS

Our research consists in transforming the traditional genomic data workflow in a distributed system design, suitable for a federated environment, and fully privacy-preserving. So far, we focused on the introduction of privacy-preserving methods from the earliest steps – directly after the sequencing – and we evaluated how to perform alignment and variant calling while protecting reads privacy. Our research intends to take into account the properties of the reads produced either by the previous or the most recent sequencing technologies, enabling a general application of the developed methods.

The goals of our work can be summarized as follows:

- Introduce privacy-preserving methods in an early stage of the analysis workflow by classifying the sequenced reads by sensitivity level.
- Develop a sensitivity classification which can be adapted to the sequences used in different studies.
- Allow the adaptation of the privacy-preserving conditions and methods to the different sensitivity levels.
- Design of a privacy-preserving distributed analysis scheme where data is stored at multiple locations and its analysis do not leak each participant information (e.g. identity). The main focus is to allow alignment and variant calling steps.

4 RESEARCH DESIGN

The proposed approach consists in the inclusion of privacy-preserving methods in the traditional genomic reads analysis workflow.

Fig. 1(b) introduces the privacy-preserving genomic reads analysis workflow we propose, whose novelties are the introduction of a sensitivity-aware filtering step after the sequencing step and a distributed analysis scheme. The filtering step consists on the classification of the reads among different sensitivity levels according to the information they contain. This classification allows the adaptation of the storage conditions and of the algorithms applied to the privacy protection the reads of each sensitivity levels need.

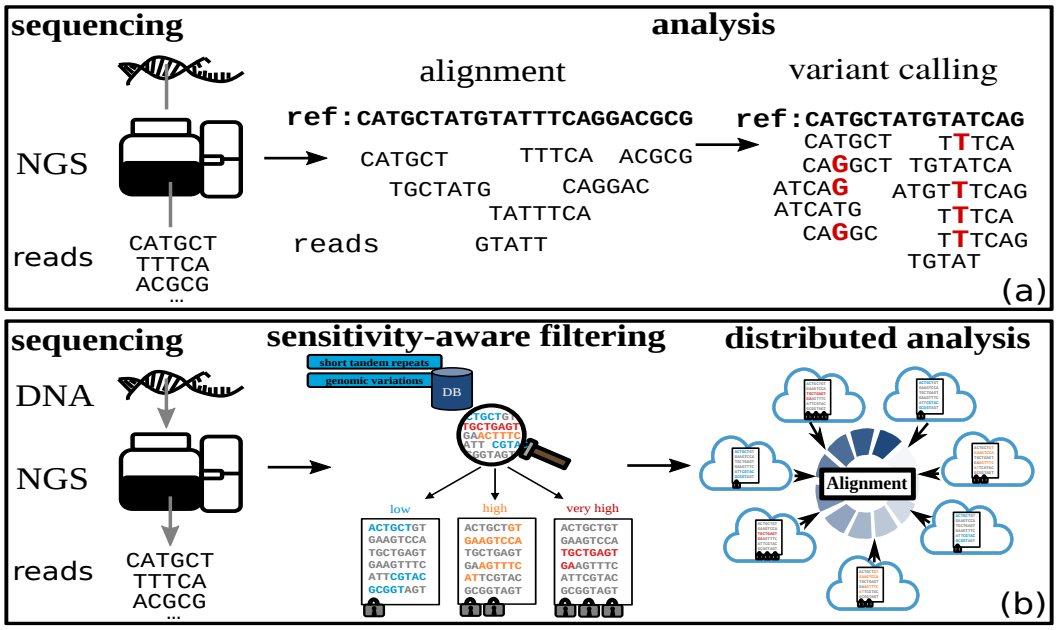

Figure 1: Traditional analysis workflow vs privacy-preserving proposed approach.
(a) Traditional analysis workflow. In this scheme data is analysed using plaintext algorithms. (b) Privacy-preserving proposed approach. The proposed scheme introduces a sensitivity-aware filtering step and then adapts the analysis (e.g. alignment and variant calling) algorithms to the sensitivity levels.

The distributed analysis scheme assumes the reads are stored across multiple locations and allows their analysis without high data transmission costs, since the data do not need to be transferred between different locations. The plan is to transmit only statistics and metadata which do not leak personal details. Our distributed scheme also includes methods to preserve the identity of the reads.

ONGOING RESULTS

Long reads filtering: Our research developed a long reads filtering approach based on a previously published filter for 30-bases reads which classifies the reads into privacy-sensitive and non-privacy-sensitive information [5]. Our approach has a higher accuracy than the previous filter: less privacy-sensitive nucleotides missed, and less non-privacy-sensitive nucleotides wrongly classified. We also studied the effect of the presence of errors on the reads, and we showed that iterating the filtering process helps on tolerating errors (with 2% of errors we detect 86% of the sensitive nucleotides present on the reads, instead of 56% using the previous filter [5]). The output of our filter are reads where the privacy-sensitive information is masked.

Sensitivity levels: Our research also focused on the definition of sensitivity levels based on quantitative (e.g. allele frequency) and qualitative (e.g. disease susceptibility). We studied the links between data in different sensitivity levels and we showed that locating all the linked information on the observed highest level prevents information inference between different levels. Using the long reads filtering approach developed to make the classification of the reads into sensitivity levels, we present an example of three sensitivity levels and studied the proportion of an individual's genome in each level. The results showed that 5% of the reads in the genome have very high sensitivity, 23% have high sensitivity and the remaining 72% have low sensitivity. To summarize, the levels can be adapted to the properties of each data analysis and/or user priorities. Furthermore, different sensitivity levels correspond to different needs of protection, i.e. highly sensitive data requires higher protection (e.g. crypto-based algorithms), while least sensitive data can be analysed using usual protection levels. Therefore, with the presented classification we adapt the privacy protection to each level of sensitivity while improving analysis performance.

ONGOING RESEARCH

Privacy leaks from reads: Understanding how much information a set of reads can leak is important to adjust the protective measures ensuring the required privacy. Differently from previous work focused on genomic data protection after reads analysis, we focus on the study of privacy risks of raw reads, before they are analysed. We consider reported privacy attacks on genomic data [7, 8, 14] which were performed on processed data (e.g. SNPs, genotypes) and adapt their application for raw reads. We also explore previous work on defining the number of genomic variations (e.g. SNPs) needed to uniquely identify an individual to better understand which information is and is not unique for an individual [12]. For example, disease related regions of the genome leaks more information about an individual and they might contribute to his/her identification.

The goal is to define the amount of sensitive information a set of reads contains and which other information can we obtain exploring

data linkages. To reach our goal, we plan to follow the state of the art variant calling process in order to obtain the sensitive information and then explore inference methods to obtain further information.

Distributed variant calling: Working toward the e-biobanking vision [4] which describes data distributed over multiple data centers (i.e., reads here), the goal is to develop a distributed system which enables the analysis of the reads without transmitting them between centers. Some distributed systems for biomedical data have been created, however more work is needed in this field [1, 6]. The first step on this topic is to understand the requirements of research community (e.g. statistics, functionalities, algorithms) in order to improve studies completeness. Afterwards, we plan to enable the release of analysis statistics while ensuring that one center is not able to learn about other centers data, thus enabling the implementation of coalitions of non-mutually trusting partners with common objectives.

This topic includes: (i) studying secure multi-party computations for data exchange between data centers; (ii) defining of the information that is secure to share between different centers; (iii) defining of the communication protocol; (iv) studying the existing data protection methods and selecting the most adequate for our data requirements.

ACKNOWLEDGMENTS

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] Beacon Network. https://beacon-network.org/.
[2] Erman Ayday, Jean Louis Raisaro, Jean-Pierre Hubaux, and Jacques Rougemont. 2013. Protecting and Evaluating Genomic Privacy in Medical Tests and Personalized Medicine. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society (WPES '13)*. 95–106.
[3] Joshua Baron, Karim El Defrawy, Kirill Minkovich, and et al. 2012. 5pm: Secure pattern matching. *SCN* (2012), 222–240.
[4] Alysson Bessani, Jörgen Brandt, Marc Bux, Vinicius Cogo, Lora Dimitrova, Jim Dowling, Ali Gholami, Kamal Hakimzadeh, Micheal Hummel, Mahmoud Ismail, Erwin Laure, Ulf Leser, Jan-Eric Litton, Roxanna Martinez, Salman Niazi, Jane Reichel, and Karin Zimmermann. 2015. BiobankCloud: a platform for the secure storage, sharing, and processing of large biomedical data sets. *DMAH* (2015).
[5] Vinicius V Cogo, Alysson Bessani, Francisco M Couto, and Paulo Verissimo. 2015. A high-throughput method to detect privacy-sensitive human genomic data. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*. ACM, 101–110.
[6] The Global Alliance for Genomics and Health. 2016. A federated ecosystem for sharing genomic, clinical data. *Science* 352, 6291 (2016), 1278–1280.
[7] Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. 2013. Identifying personal genomes by surname inference. *Science* 339, 6117 (2013), 321–324.
[8] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4, 8 (2008), e1000167.
[9] Mathias Humbert, Kévin Huguenin, Joachim Hugonot, Erman Ayday, and Jean-Pierre Hubaux. 2015. De-anonymizing Genomic Databases Using Phenotypic Traits. *Proceedings on Privacy Enhancing Technologies* 2015, 2 (2015), 99–114.
[10] Murat Kantarcioglu, Wei Jiang, Ying Liu, and Bradley Malin. 2008. A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on information technology in biomedicine* 12, 5 (2008), 606–617.
[11] Manfred Kayser and Peter de Knijff. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics* 12, 3 (2011), 179–192.
[12] Zhen Lin, Art B Owen, and Russ B Altman. 2004. Genomic research and human subject privacy. *SCIENCE-NEW YORK THEN WASHINGTON-*. (2004), 183–183.
[13] Reza Mirnezami, Jeremy Nicholson, and Ara Darzi. 2012. Preparing for Precision Medicine. *New England Journal of Medicine* 366, 6 (2012), 489–491.
[14] D R. Nyholt, C.-E. Yu, and P. M. Visscher. 2009. On Jim WatsonâĂŹs APOE status: genetic information is hard to hide. *European Journal of Human Genetics* 17 (2009), 147–149.
[15] Victoria Popic and Serafim Batzoglou. 2017. A hybrid cloud read aligner based on MinHash and kmer voting that preserves privacy. *Nature Communications* 8 (2017).
[16] Michael C Schatz. 2009. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* 25, 11 (2009), 1363–1369.
[17] Shuang Wang, Xiaoqian Jiang, Haixu Tang, Xiaofeng Wang, Diyue Bu, Knox Carey, Stephanie OM Dyke, Dov Fox, Chao Jiang, Kristin Lauter, Bradley Malin, Heidi Sofia, Amalio Telenti, Lei Wang, Wenhao Wang, and Lucila Ohno-Machado. 2017. A community effort to protect genomic data sharing, collaboration and outsourcing. *npj Genomic Medicine* 2, 1 (2017), 33.