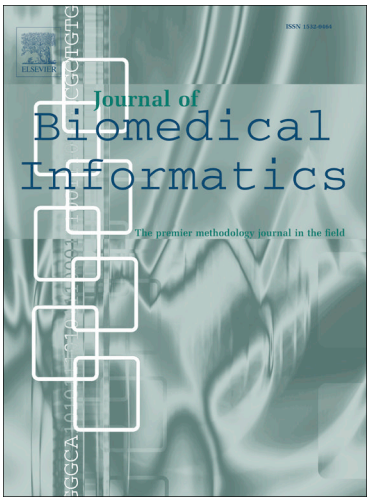
Accepted Manuscript



Accurate Filtering of Privacy-Sensitive Information in Raw Genomic Data

Jérémie Decouchant, Maria Fernandes, Marcus Völp, Francisco M Couto, Paulo Esteves-Veríssimo

PII: S1532-0464(18)30070-4
DOI: https://doi.org/10.1016/j.jbi.2018.04.006
Reference: YJBIN 2963

To appear in: *Journal of Biomedical Informatics*

Received Date: 21 January 2018
Accepted Date: 7 April 2018

Please cite this article as: Decouchant, J., Fernandes, M., Völp, M., Couto, F.M., Esteves-Veríssimo, P., Accurate Filtering of Privacy-Sensitive Information in Raw Genomic Data, *Journal of Biomedical Informatics* (2018), doi: https://doi.org/10.1016/j.jbi.2018.04.006

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Accurate Filtering of Privacy-Sensitive Information in Raw Genomic Data

Jérémie Decouchant[a], Maria Fernandes[a], Marcus Völp[a], Francisco M Couto[b], Paulo Esteves-Veríssimo[a]

[a] *SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg*

[b] *LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa*

Abstract

Sequencing thousands of human genomes has enabled breakthroughs in many areas, among them precision medicine, the study of rare diseases, and forensics. However, mass collection of such sensitive data entails enormous risks if not protected to the highest standards. In this article, we follow the position and argue that post-alignment privacy is not enough and that data should be automatically protected as early as possible in the genomics workflow, ideally immediately after the data is produced. We show that a previous approach for filtering short reads cannot extend to long reads and present a novel filtering approach that classifies raw genomic data (i.e., whose location and content is not yet determined) into privacy-sensitive (i.e., more affected by a successful privacy attack) and non-privacy-sensitive information. Such a classification allows the fine-grained and automated adjustment of protective measures to mitigate the possible consequences of exposure, in particular when relying on public clouds. We present the first filter that can be indistinctly applied to reads of any length, i.e., making it usable with any recent or future sequencing technologies. The filter is accurate, in the sense that it detects all known sensitive nucleotides except those located in highly variable regions (less than 10 nucleotides remain undetected per genome instead of 100,000 in previous works). It has far less false positives than previously

Email addresses: jeremie.decouchant@uni.lu (Jérémie Decouchant), maria.fernandes@uni.lu (Maria Fernandes), marcus.voelp@uni.lu (Marcus Völp), fcouto@di.fc.ul.pt (Francisco M Couto), paulo.verissimo@uni.lu (Paulo Esteves-Veríssimo)

Download English Version:

https://daneshyari.com/en/article/6927444

Download Persian Version:

https://daneshyari.com/article/6927444

Daneshyari.com