Report from Dagstuhl Seminar 13412

Genomic Privacy

Edited by

Kay Hamacher[1], Jean Pierre Hubaux[2], and Gene Tsudik[3]

1 **TU Darmstadt, DE,** hamacher@bio.tu-darmstadt.de
2 **EPFL – Lausanne, CH,** Jean-Pierre.Hubaux@epfl.ch
3 **University of California – Irvine, US,** gene.tsudik@uci.edu

Abstract

Recent advances in genomics prompt a formidable privacy challenge: As the price of a complete genome profile has plummeted to as low as 99 USD for genome-wide genotyping, wide-spread usage of genomic information is about to become reality. Substantial progress is expected in the near future in terms of improved diagnoses and better preventive medicine.

The impact of the increased availability of genomic information on privacy, however, is unprecedented, for obvious reasons: First, genetic conditions and the predisposition to specific diseases (such as Alzheimer's) can be revealed. Second, a person's genomic information leaks substantial information about his relatives. Third, complex privacy issues can arise if DNA analysis is used for criminal investigations, epidemiological research, and personalized medicine purposes.

This report documents the program and the outcomes of the Dagstuhl Seminar 13412 "Genomic Privacy". The goal of the seminar was to bring together leading researchers, from different areas of academia and industry. The seminar welcomed participants from computer science, bioinformatics, genetics, ethics and medical fields. Through a series of presentations, discussions, and working groups, the seminar attempted to provide a coherent picture of the field, which transcends the borders of disciplines. The participants discussed many aspects of genomic privacy and jointly identified the main requirements and the possible technical solutions for protecting genomic data.

Seminar 07.–09. October, 2013 – www.dagstuhl.de/13412
1998 ACM Subject Classification K.4.1 Public Policy Issues, D.4.6 Security and Protection, J.3 Life and Medical Sciences
Keywords and phrases Genomics, Genetics, Health Data, Privacy Protection, Differential Privacy, Privacy by Design, Information Security, Cryptography, Secure Computation
Digital Object Identifier 10.4230/DagRep.3.10.25
Edited in cooperation with Jean Louis Raisaro

1 Executive Summary

Jean Louis Raisaro

The Dagstuhl seminar 13412 "Genomic Privacy" was a short two-and-a-half-day seminar, the first one on this topic ever, which took place from October 6th to 9th, 2013. The aim was to bring together researchers, from various research areas related to genomic privacy, and to inspire them to exchange theoretical results, practical requirements and ethical and legal implications related to the protection of genomic data. The rise of personalized medicine on the background of available, individual genomic sequences is taken for granted

Genomic Privacy, *Dagstuhl Reports*, Vol. 3, Issue 10, pp. 25–35
Editors: Kay Hamacher, Jean Pierre Hubaux, and Gene Tsudik
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

in the biomedical community. Impressive advances in genome sequencing have opened the way to a variety of revolutionary applications in modern healthcare. In particular, the increasing understanding of the human genome, and of its relation to diseases and its response to treatments brings promise of improvements in preventive and personalized healthcare. However, because of the genome's highly sensitive nature, this progress raises important privacy and ethical concerns that need to be addressed. Indeed, besides carrying information about a person's genetic condition and his predisposition to specific diseases, the genome also contains information about his relatives. The leakage of such information can open the door to a variety of abuses and threats not yet fully understood. During the seminar, these points were addressed in particular:

- Expression and Requirements: What should be protected? For how long? Against whom? Who should be liable? Who would manage cryptographic keys? Anonymity vs. cryptography?
- Privacy Mechanisms & Regulations: What privacy enhancing techniques can be envisioned specifically for genomic data? What if some people publish their genome online against the will of their relatives? Which ethical guidelines can be adopted from traditional privacy regulations?
- Medical Perspective: Would medical specialists accept to have only a partial view on genomic data? How are epidemiological studies and biobanks affected by legal and/or technical restrictions?
- Patient Perspective: What patient's involvement can be reasonably expected? 'Can a person's genomic information be outsourced to some cloud storage service?
- Economics: What are the economic implications of genomic privacy; of its neglect?

The seminar fully satisfied the expectations. All participants briefly self-introduced themselves. Some of them were invited by the organizers to give survey talks about their recent research on genomic privacy, thus facilitating and encouraging inter-disciplinary discussions during the morning sessions. There were talks focusing both on the definition of the requirements for the efficient and secure implementation of genomic medicine and on the possible solutions to be addressed. The afternoon sessions were devoted to working groups.

The first speaker Regine Kollek (University of Hamburg, GER), addressed the meaning and context of genomic data, focusing on some of the social and ethical aspects of genomics. She was followed by Brad Malin (Vanderbilt University, US), who provided a summary of the ways (both legal and technical) such data can be protected and raised the question about its worth or if there exists some other practical approach that guarantees flexible genomic data protection plans. Satoru Miyano (University of Tokyo, JP) gave an overview of the requirements in term of storage, computational power and security needed to make "clinical sequencing " become a reality. He described the ongoing program that has been playing a key role in the International Cancer Genome Consortium (ICGC) in Japan. The morning session ended up with a joint talk by Jacques Fellay (EPFL – Lausanne, CH) and Amalio Telenti (Lausanne University Hospital, CH) about the current and future usage of genomic information in clinical settings. They outlined the importance of defining new threat models, emphasizing that trust is essential in healthcare.

The second day was focused on the possible technical solutions that can be used to ensure genomic privacy. If, on one hand, there are computational expensive cryptographical approaches such as homomorphic encryption or secure multi-party computation that guarantee accuracy at the expense of flexibility and increasing complexity, on the other hand there are also statistical-based solutions such as differential privacy, which are less accurate but more flexible and less expensive in terms of computational and complexity costs. The first speaker of the day, Andreas Peter (University of Twente, NL), described his ongoing work on how to securely outsource genomic sequences in a privacy-preserving way by relying on an oblivious RAM construction. The second talk, by Erman Ayday (EPFL – Lausanne, CH), provided an overview of the activities on genomic privacy in Lausanne. Ayday first focused on how to protect and evaluate genomic privacy in the clinical context, he then showed how to process in a privacy-preserving fashion raw genomic data; and finally he described how to quantify kin genomic privacy. The third speaker of the morning, Vitaly Shmatikov (University of Texas – Austin, US), discussed about how to conduct privacy-preserving exploration in Genome-Wide Association Studies (GWAS). He presented a set of privacy-preserving data mining algorithms that produce significantly accurate results while guaranteeing differential privacy. Finally, the second-day morning session was closed by Emiliano De Cristofaro's (University College London, GB) survey about how to begin to address privacy-respecting genomic tests by relying on privacy-enhancing techniques based on private set intersection operations.

The final day started with a talk by Xiaofeng Wang (Indiana University – Bloomington, US) about the privacy-preserving sharing and analysis of human genomic data. In particular, he described some techniques for secure outsourcing of genome analysis, and differentially-private pilot data release and data source selection. The remaining part of the morning was devoted to a general discussion about the seminar's outcomes. Due to the seminar and the multi-disciplinary interactions, it became clear that protection of simple genomic sequences is not enough for a full-privacy preserving approach. The organizers, together with the participants, agreed that this problem should be addressed in a sequel Dagstuhl-seminar. Hence, they set up a future work agenda in order to organize again such a fruitful gathering.

We thank Schloss Dagstuhl for the professional and inspiring atmosphere it provides. Such an intense research seminar is possible because Dagstuhl so perfectly meets all researchers' needs.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Genomics & Privacy: More Trouble than It's Worth?

Bradley Malin (Vanderbilt University, US)

URL http://www.hiplab.org/

Genomic sequence data is increasingly subject to attacks that disclose the identity of the corresponding individual (i.e., identity disclosure) or certain information about the individual (i.e., attribute disclosure). This talk reviews why such data is interesting and useful and how such attacks have transpired and are escalating over time. It then provides a high-level summary of ways in which such data can be protected from such attacks, both from a sociological (e.g., policy and contract) and technical perspective (disclosure control and encryption). After doing so, this questions if these are the only options for data protection or if risk-based models, which are cognizant of monetary costs and (dis)incentives might be more appropriate options for crafting practical and flexible genomic data protection plans.

3.2 Personalized Genomic Medicine at Institute of Medical Science of University of Tokyo

Satoru Miyano (University of Tokyo, JP)

The cost of human genome sequencing has decreased significantly. Today integrative systems' understanding of personal cancer based on personal omics data is getting important issue. It has become a reality to implement "clinical sequencing" of personal cancer and germline genomes together with their transcriptomic information. At the IMSUT (Institute of Medical Science, The University of Tokyo), a team comprised of members from the Human Genome Center, the Advanced Clinical Research Center and the IMSUT Research Hospital is installing a program for personalized genomic medicine of cancer based on whole genome sequencing and integrative omics analysis enhanced with the supercomputer system at Human Genome Center (225T FLOPS at peak, 4.6PB storage). In this talk, we present our ongoing program which uses the supercomputer system that has been playing a key role in the ICGC project of Japan and some important cancer genome sequencing projects. By running this program, we will facilitate the information system for personalized genomic medicine and foster people on biomedical informatics who can clinically interpret whole genome sequencing and omics data.

3.3 Needs and Myths in Medical Genomics

Jacques Fellay (EPFL – Lausanne, CH)
Amalio Telenti (Lausanne University Hospital, CH)

Joint work of Fellay, Jacques; Telenti, Amalio

In this presentation we will present a view on genomic privacy from the perspective of clinicians and genomic researchers. We will describe the current and future usage of genomic information in clinical settings, in order to define the requirements for the efficient and secure implementation of genomic medicine. While contemporary medical genetics still mostly focuses on rare variants involved in Mendelian diseases or exceptional response to drugs, the toolbox of genomic medicine is about to massively expand to encompass neonatal and prenatal sequencing, oncogenomics, and complex trait genomics. In addition, direct-to-consumer genomics (DTC) is likely to profoundly alter the classic patient-doctor interaction. Threat models need to be defined, keeping in mind that trust must remain at the center of healthcare.

3.4 A Practical Approach to Securely Outsource Genomic Sequences

Andreas Peter (University of Twente, NL)

Joint work of Peter, Andreas; Biedermann, Sebastian; Karvelas, Nikolaos; Katzenbeisser, Stefan

Recent developments in genomics enable new business and research models to have many individuals' genomes as an essential driving force. It is foreseeable that this genomic data will be outsourced to a central service provider (e.g., the cloud), thus allowing different applications, such as personalized medicine, large-scale genomic research, and disease susceptibility.

In my presentation, I describe a novel mechanism that enables a patient to securely outsource her genomic data to the cloud while delegating to an investigator (e.g. a physician) the right to run certain algorithms (e.g., medical tests) on her data. The mechanism is privacy-preserving, meaning that the investigator only learns the result of his algorithm on the patient's genome, while the cloud learns nothing at all. Our protocol only requires the patient to be online for the delegation of rights, and it is thereafter completely non-interactive with the patient.

We achieve reasonable efficiency by dividing the patient's genome into small blocks (e.g., SNPs) on which we can run efficient, secure, multiparty protocols. Our main technical contribution is a new ORAM-like construction for hiding the access patterns to these blocks. To circumvent an undesirable heavy workload at the investigator's side, our ORAM allows for the outsourcing of its most involved step (the "reshuffle") to an independent proxy without loss of privacy.

3.5 Protecting and Quantifying Genomic Privacy

Erman Ayday (EPFL – Lausanne, CH)

Joint work of Ayday, Erman; Raisaro, Jean Louis; Humbert, Mathias; Huang, Zhicong; Hubaux, Jean-Pierre
URL http://lca.epfl.ch/projects/genomic-privacy/

Genomics is becoming the next significant challenge for privacy. The price of a complete genome profile has plummeted below 100 USD for genome-wide genotyping (i.e., the characterization of about one million common genetic variants), which is offered by a number of companies. This low cost of DNA sequencing will break the physician/patient connection and it can open the door to all kinds of abuse not yet fully understood.

Access to genomic data prompts some important privacy concerns: (i) Genetic diseases can be unveiled; (ii) the propensity to develop specific diseases (such as Alzheimer's) can be revealed; (iii) a volunteer accepting de facto to have his genomic code made public can leak substantial information about genomic data of his relatives (possibly against their will); and (iv) complex privacy issues can arise if DNA analysis is used for criminal investigations and insurance purposes. Such issues could lead to abuse, threats, and genetic discrimination.

In this talk, after discussing the threats and challenges of genome privacy, I will summarize our solutions for protecting the privacy of genomic data. In particular, I will focus on (i) protecting and evaluating genome privacy in medical tests and personalized medicine, (ii) privacy-preserving processing of raw genomic data, and (iii) the quantification of kin genomic privacy using information theoretical tools.

3.6 Privacy-Preserving Data Exploration in Genome-Wide Association Studies

Vitaly Shmatikov (University of Texas – Austin, US)

Joint work of Shmatikov, Vitaly; Johnson, Aaron
Main reference A. Johnson, V. Shmatikov, "Privacy-Preserving Data Exploration in Genome-Wide Association Studies," in Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'13), pp. 1079–1087, ACM, 2013.
URL http://dx.doi.org/10.1145/2487575.2487687

Genome-wide association studies (GWAS) have become a popular method for analyzing sets of DNA sequences in order to discover the genetic bases of diseases. Unfortunately, statistics published as the result of GWAS can be used to identify individuals participating in the study. To prevent privacy breaches, even previously published results have been removed from public databases, impeding researchers' access to the data and hindering collaborative research. Existing techniques for privacy-preserving GWAS focus on answering specific questions, such as correlations between a given pair of SNPs (DNA sequence variations). This does not fit the typical GWAS process, where the analyst might not know in advance which SNPs to consider, which statistical tests to use, how many SNPs are significant for a given dataset, etc.

We present a set of practical, privacy-preserving data-mining algorithms for GWAS datasets. Our framework supports exploratory data analysis, where the analyst does not know a priori how many and which SNPs to consider. We develop privacy-preserving algorithms for computing the number and location of SNPs that are significantly associated with the disease, the significance of any statistical test between a given SNP and the disease, any measure of correlation between SNPs, and the block structure of correlations. We evaluate our algorithms on real-world datasets and demonstrate that they produce significantly more accurate results than prior techniques while guaranteeing differential privacy.

3.7 Whole Genome Sequencing vs. Privacy: Efficient Cryptographic Protocols to the Rescue

Emiliano De Cristofaro (University College London, GB)

Joint work of De Cristofaro, Emiliano; Baldi, Pierre; Baronio, Roberta; Faber, Sky; Gasti, Paolo; Tsudik, Gene
URL http://emilianodc.com/

Recent advances in DNA sequencing technologies have put ubiquitous availability of whole human genomes within reach. It is no longer hard to imagine the day when everyone will have the means to obtain and store his own DNA sequence. The widespread and affordable availability of whole genomes immediately opens up important opportunities in a number of health-related fields. In particular, common genomic applications and tests performed in vitro today will soon be conducted computationally, using digitized genomes. New applications will be developed as genome-enabled medicine becomes increasingly preventive and personalized. However, the very same progress also amplifies worrisome privacy concerns, as a genome represents a treasure trove of highly personal and sensitive information. In this talk, we will give an overview of biomedical advances in genomics and discuss associated privacy, ethical, and security challenges. We begin to address privacy-respecting genomic tests by focusing on some important applications, such as, personalized medicine, paternity tests, ancestry testing, and genetic compatibility tests. After carefully analyzing these applications and their requirements, we propose a set of efficient privacy- enhancing techniques based on private set operations. This enables us to implement, in silico, some operations that are currently performed via in vitro methods, in a secure fashion. Our experimental results demonstrate that the proposed techniques are both feasible and practical today. Finally, we explore a few alternatives for securely storing human genomes and allowing authorized parties to run tests in such a way that only the required minimum amount of information is disclosed. We present an Android API framework geared for privacy-preserving genomic testing.

3.8 Privacy-Preserving Sharing and Analysis of Human Genomic Data

Xiaofeng Wang (Indiana University – Bloomington, US)

Main reference Y. Chen, B. Peng, X. Wang, H. Tang, "Large-Scale Privacy-Preserving Mappings of Human Genomic Sequences on Hybrid Clouds," in Proc. of the 19th Annual Network and Distributed System Security Symp. (NDSS'12), 18pp., The Internet Society, 2012.
URL http://www.internetsociety.org/large-scale-privacy-preserving-mapping-human-genomic-sequences-hybrid-clouds

This presentation will present a view on genomic privacy from the perspective of clinicians and genomic researchers. in order to define the requirements for the efficient and secure implementation of genomic medicine, we will describe the current and future usage of genomic information in clinical settings. While contemporary medical genetics still mostly focuses on rare variants involved in Mendelian diseases or exceptional responses to drugs, the toolbox of genomic medicine is about to massively expand to encompass neonatal and prenatal sequencing, oncogenomics, and complex trait genomics. In addition, direct-to-consumer genomics (DTC) is likely to profoundly alter the classic patient-doctor interaction. Threat models need to be defined, keeping in mind that trust must remain at the center of healthcare.

In this talk, we review our research on privacy-preserving genomic data computing and privacy-preserving data dissemination, particularly the techniques for secure outsourcing genomic analysis, differentially- private pilot data release and data source selection.

4 Working Groups

4.1 Privacy and Genomic/Health Data – What Makes Genomic Data Special

Chair: *Bradley Malin (Vanderbilt University, TN, United States)*

This working group focused on the definitions of the benefits and harms of genomic data. Recently, it has been shown that the widespread and affordable availability of fully-sequences human genomes has created enormous opportunities for better preventive and personalized medicine. In particular, in this working group, it was observed how the main benefits of genomic data consist in tailored drug dosages and choices and in the risk assessment for diseases based on the genetic makeup.

However, because of genome's highly sensitive nature, the leakage of genomic information can pave the way to a variety of abuses and threats. For example, health insurance companies might obtain the genetic information of their customers and deny their services to people with a high susceptibility of developing a chronic disease, or employers could hire applicants based on their genetic features. Access to this information could engender genetic discrimination or discredit an organization or a country willing to promote a personal genome project. Hence, there is an impelling need to design new privacy-enhancing technologies (PET) that guarantee the protection of such a data by keeping in mind that the first step should be the identification of the data-flow model both in the clinical practice and in the bioinformatics research in order to understand which part of the process PETs should focus on.

4.2 Scientific Methods in Human Genomics and Bioinformatics that Set the Requirements for Crypto Solutions

Chair: *Satoru Miyano (University of Tokyo, Japan)*

This working group focused on the definitions of the requirements for cryptographic solutions. In particular during this session, it was observed that the main requirements should consist in the protection of germline whole genome sequences along with whole markers (SNPs, insertions, deletions, etc..) linked with clinical data and family histories. The solution to these requirements is often not unique, rather a trade-off between (i) performance, (ii) utilization and (iii) privacy. For example, statistical-based solutions such as differential privacy focus on privacy at the expense of utilization, whereas cryptographic solutions prioritize the utilization at the cost of bad performances. However, technical solutions are often not sufficient for protecting genomic privacy, therefore legal and professional guidelines are certainly needed to govern how genomic information is transmitted, stored and processed by the different stakeholders.

4.3 Cryptographic Protocols for the Protection of Genomic Data

Chair: *Marina Blanton (University of Notre Dame, IN, United States)*

This working group focused on defining what are the best cryptographic protocols for the protection of genomic data. In particular, it was observed by all the participants that there is a gap between the algorithms used for the processing of genomic data and algorithms that privacy-preserving solutions implement. Existing cryptographic protocols for privacy-preserving DNA processing started with rather simple functions and, so far, they still do not mimic complex algorithms used in practice (e.g. susceptibility tests, genotype imputations, association studies, etc.). Furthermore, current solutions do not simultaneously achieve full functionality, security and speed. Hence, it is clear that designing new cryptographic protocols that satisfy all the requirements represents a rather complex challenge. A possible solution outlined during the working group consists in protecting genomic data through multiple mechanisms such as cryptographic means, secret sharing, access control and legal means; by focusing on more precise functionality more so than on speed, while keeping in mind that cryptography can mitigate the risk but not eliminate it. In addition, the longevity of genomic data calls for levels of security stricter than in some other domains, and choosing larger security parameters for encrypted data represents only one part of the solution.

4.4 Anonymization of Genomic Data

Chair: *Vitaly Shmatikov (University of Texas Austin, TX, United States)*

This working group focused on the anonymization of genomic data. With some background information, only 75 SNPs are sufficient for identifying an individual. Because DNA is relatively stable in time and a small subset of other data can be revealed, and as any biological sample contain all markers required for identification, genomic data should be considered different from any other high dimensional profile of an individual. The general discussion of the working group was about the following question: "Can we apply traditional anonymization techniques to anonymization of genomic data?". The following points were proposed as potential solutions.

- *Recommendation 1*: Use multilevel data anonymization. Anonymized data could be made public (after some obfuscation), whereas precise data can be made available with strict access control.
- *Recommendation 2*: Combine technical, social (e.g., economic incentives), and legal(e.g., boundaries and damages) controls.
- *Recommendation 3*: Audit (physical and virtual).
- *Recommendation 4*: Pilot testing with real data and target populations.

It is important, however, to be aware that in general anonymization breaks when some other information is known and therefore every new model that provides a better understanding of humane genome would increase the risk of re-identification. Again, there does not exist a unique solution, but technical and legal privacy guarantees are needed.

Participants

- Gergely Acs
INRIA Grenoble –
Rhône-Alpes, FR
- Erman Ayday
EPFL – Lausanne, CH
- Marina Blanton
Univ. of Notre Dame, US
- Jurgi Camblong
Sophia Genetics SA –
Lausanne, CH
- Emiliano De Cristofaro
University College London, GB
- Zekeriya Erkin
TU Delft, NL
- Sky Faber
University California – Irvine, US
- Jacques Fellay
EPFL – Lausanne, CH
- Kay Hamacher
TU Darmstadt, DE
- Urs Hengartner
University of Waterloo, CA
- Zhicong Huang
EPFL – Lausanne, CH
- Jean Pierre Hubaux
EPFL – Lausanne, CH
- Mathias Humbert
EPFL – Lausanne, CH
- Murat Kantarcioglu
The University of Texas at
Dallas, US
- Florian Kerschbaum
SAP AG – Walldorf, DE
- Regine Kollek
Universität Hamburg, DE
- Klaus A. Kuhn
TU München – Klinikum Rechts
der Isar, DE
- Inald Lagendijk
TU Delft, NL
- Bradley Malin
Vanderbilt University, US
- Srdan Marinovic
ETH Zürich, CH
- Satoru Miyano
University of Tokyo, JP
- Andrzej Mizera
University of Luxembourg, LU
- Muhammad Naveed
EPFL – Lausanne, CH
- Andreas Peter
University of Twente, NL
- Jean-Jacques Quisquater
University of Louvain, BE
- Jean-Louis Raisaro
EPFL – Lausanne, CH
- Roded Sharan
Tel Aviv University, IL
- Vitaly Shmatikov
University of Texas – Austin, US
- Amalio Telenti
University of Lausanne, CH
- Gene Tsudik
Univ. of California – Irvine, US
- Xiaofeng Wang
Indiana University –
Bloomington, US