Report of Dagstuhl Seminar 15431

Genomic Privacy

Edited by

Jean Pierre Hubaux[1], Stefan Katzenbeisser[2], Bradley Malin[3], and Gene Tsudik[4]

1 **EPFL – Lausanne, CH,** jean-pierre.hubaux@epfl.ch
2 **TU Darmstadt, DE,** skatzenbeisser@acm.org
3 **Vanderbilt University – Nashville, US,** b.malin@vanderbilt.edu
4 **University of California – Irvine, US,** gts@ics.uci.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15431 "Genomic Privacy". The current rise of personalized medicine is based on increasing affordability and availability of individual genome sequencing. Impressive recent advances in genome sequencing have ushered a variety of revolutionary applications in modern healthcare and epidemiology. In particular, better understanding of the human genome as well as its relationship to diseases and response to treatments promise improvements in preventive and personalized healthcare. However, because of the human genome's highly sensitive nature, this progress raises important privacy and ethical concerns, which simply cannot be ignored. A digitized genome represents one of the most sensitive types of human (personal) identification data. Even worse, a genome contains information about its owner's close relatives. The Dagstuhl seminar 15431 brought together computer scientists, bioinformaticians, geneticists and ethical experts to discuss the key security and privacy challenges imposed by the storage of large volumes of genetic data.

Seminar October 18–23, 2015 – http://www.dagstuhl.de/15431
Keywords and phrases cryptography, differential privacy, genetics, genomics, health data, information security, privacy by design, privacy protection, secure computation
Digital Object Identifier 10.4230/DagRep.5.10.50

1 Executive Summary

Jean Pierre Hubaux
Stefan Katzenbeisser
Bradley Malin
Gene Tsudik

This report documents the program and the outcomes of Dagstuhl Seminar 15431 "Genomic Privacy". The current rise of personalized medicine is based on increasing affordability and availability of individual genome sequencing. Impressive recent advances in genome sequencing have ushered a variety of revolutionary applications in modern healthcare and epidemiology. In particular, better understanding of the human genome as well as its relationship to diseases and response to treatments promise improvements in preventive and personalized healthcare.

At the same time, human genetics has become a "big data" science. For roughly a decade, specific tests for Single Nucleotide Polymorphisms (SNPs), e.g., markers corresponding to

specific diseases, have been well established. Furthermore, research in pharmaco-genomics, which currently relies on SNPs, has helped improve drug treatment for cancer and cardiac patients. The methodology of genotyping, which takes into account hundreds to thousands of variations in positions in the genome, has tremendously increased the amount of data acquired during diagnosis. Personalized genotyping has become commercially available from several sources (such as 23andMe). Full genome sequencing and genome-wide association studies are moving towards full deployment in clinical practice. In 2000, the cost of sequencing one human genome was US$2.5 billion. Today, the price of US$200 for genome sequencing is approaching reality. Considering the benefits for (public) health and potential cost savings, widespread acquisition, storage, and usage of personal genomes is guaranteed to happen soon.

However, because of the human genome's highly sensitive nature, this progress raises important privacy and ethical concerns, which simply cannot be ignored. A digitized genome represents one of the most sensitive types of human (personal) identification data. Even worse, a genome contains information about its owner's close relatives. Furthermore, correlations with individual data sets from so-called "omics-technologies" pose even bigger threats on privacy. Leakage of personal genomic information can lead a wide variety of attacks, many of which are not yet fully understood. Whether accidentally or intentionally revealed, a digitized genome cannot be revoked or modified. Consequently, secrecy of personal genomic data is of paramount importance. Furthermore, genomic data, unlike other types of highly sensitive information (even national secrets), does not lose its sensitivity over time. Even worse, the mechanisms available to interpret genomic data improve over time, which means that it is unclear at the moment how much sensitive information a genome encodes and which consequences a genomic data breach has. Furthermore, it is likely that genomic data will not only be used personally to support medical treatments; great promise lies in its use in large-scale genetic studies for personalized medicine as well as common ancestry and genetic compatibility tests. Therefore, simply encrypting genomic data at rest is not a viable option and new ways of protection need to be devised.

The second Dagstuhl Seminar on Genomic Pricacy concentrated on the following topics:

- Technical solutions for genomic privacy: the participants discussed technical solutions to enable genomic data privacy, even in the presence of untrusted computing environments, and investigated technical protection techniques that can be used for this purpose.
- Integration of genomic and physiological data: For medical purposes, genomic data often needs to be correlated with clinical and physiological data. For example, clinical studies may require finding correlations between physiological data reported during hospital stays and genomic information. So far, most technical solutions for the protection of genomic data focused on securely storing DNA data itself, but did not discuss the complex problem of combining it with physiological data.
- Protection of sensitive data within large-scale genome-wide association studies: Although large-scale genomic studies offer many advantages for medical research, they pose many privacy problems. Most prior technical solutions focus on protection of a single human genome and do not scale multitudes of genomes. It remains a challenge to devise scalable techniques.

2 Table of Contents

3 Overview of Talks

3.1 Privacy in the Genomic Era

Erman Ayday (Bilkent University – Ankara, TR)

Genome sequencing technology has advanced at a rapid pace and it is now possible to generate highly-detailed genotypes inexpensively. The collection and analysis of such data has the potential to support various applications, including personalized medical services. While the benefits of the genomics revolution are trumpeted by the biomedical community, the increased availability of such data has major implications for personal privacy; notably because the genome has certain essential features, which include (but are not limited to) (i) an association with certain diseases, (ii) identification capability (e.g., forensics), and (iii) revelation of family relationships. Moreover, direct-to-consumer DNA testing increases the likelihood that genome data will be made available in less regulated environments, such as the Internet and for-profit companies. The problem of genome data privacy thus resides at the crossroads of computer science, medicine, and public policy. While the computer scientists have addressed data privacy for various data types, there has been less attention dedicated to genomic data. Thus, the goal of this paper is to provide a systematization of knowledge for the computer science community. In doing so, we address some of the (sometimes erroneous) beliefs of this field and we report on a survey we conducted about genome data privacy with biomedical specialists. Then, after characterizing the genome privacy problem, we review the state-of-the-art regarding privacy attacks on genomic data and strategies for mitigating such attacks, as well as contextualizing these attacks from the perspective of medicine and public policy. This paper concludes with an enumeration of the challenges for genome data privacy and presents a framework to systematize the analysis of threats and the design of countermeasures as the field moves forward.

3.2 Efficient Server-Aided Secure Two-Party Function Evaluation with Applications to Genomic Computation

Marina Blanton (University of Notre Dame, US)

Computation based on genomic data is becoming increasingly popular today, be it for medical or other purposes such as ancestry or paternity testing. Non-medical uses of genomic data in a computation often take place in a server-mediated setting where the server offers the ability for joint genomic testing between the users. Undeniably, genomic data is highly sensitive, and there is an urgent need to protect it, especially when it is used in computation for what we call as recreational non-health-related purposes. Towards this goal, in this work we put forward a framework for server-aided secure two-party computation with the security model motivated by genomic applications. One particular security setting that we treat in this work provides stronger security guarantees with respect to malicious users than the traditional malicious model. In particular, we incorporate certified inputs into secure computation based on garbled circuit evaluation to guarantee that a malicious user is unable to modify her inputs in order to learn unauthorized information about the other user's data.

3.3 Privacy-preserving bioinformatics with general-purpose SMC

Dan Bogdanov (Cybernetica AS – Tartu, EE)

Research has given us a range of privacy technologies. From simple, yet not provably secure, technologies like pseudonymization to full encrypted processing with homomorphic cryptography or garbled circuits. Other technologies hide privacy by adding noise – anonymization and differential privacy are the main examples.

Each such technology has different trust, deployment and performance guarantees that are not immediately clear without a risk-benefit analysis. For example, statistical anonymization can often be defeated by linking with auxiliary information. Some homomorphic encryption schemes do not allow data to be collected from multiple owners and linear secret sharing leads to the best-performing programmable secure computing schemes.

Our work has focused on implementing a range of genomic analyses using secure multiparty computation based on secret sharing. We build our work on the Sharemind framework that supports integer, fixed point, floating point and boolean arithmetic and is easily programmable using the SecreC programming language.

After successfully demonstrating secure genome-wide association studies on Affymetrix microarray data with 500K SNP locations on 1000 patients we started developing a full statistical analysis system.

The Rmind privacy-preserving statistical tool is designed to mimic the popular R statistical tool. Rmind supports filtering, a range of statistical tests and also allows for corrections when many parallel tests are performed simultaneously. Rmind performs all operations using special data-independent algorithms that do not leak private inputs through the running time. Combining this with a secure computing environment gives us unparalleled privacy guarantees. Most recently, in 2015, we were able to add Principal Component Analysis to the list of operations supported on Sharemind.

We believe that by making Sharemind support different deployment models through the use of various secure computing protocols, we can further expand its usability in privacy-preserving personalized medicine.

3.4 Genomic privacy in research and medicine: a view from the trenches

Jacques Fellay (EPFL – Lausanne, CH)

There is a need to reconcile technical and theoretical development in genomic privacy with the reality of the research and medical worlds. Using concrete examples, I will describe current applications of genomics, at the crossroad between academic research and personalized medicine.

3.5 GenoGuard: Protecting Genomic Data Against Brute-Force Attacks

Zhicong Huang (EPFL – Lausanne, CH)

Secure storage of genomic data is of great and increasing importance. The prevalent use of passwords to generate encryption keys poses an especially serious problem when applied to genetic data. Weak passwords can jeopardize genetic data in the short term, but given the multidecade lifespan of genetic data, even the use of strong passwords with conventional encryption can lead to compromise. We present a tool, called GenoGuard, for providing strong protection for genomic data both today and in the long term. We prove that decryption under any key will yield a plausible genome sequence, and that GenoGuard offers an information-theoretic security guarantee against message recovery attacks. We also explore attacks that use side information. Finally, we present an efficient and parallelized software implementation of GenoGuard.

3.6 Efficient privacy-preserving deterrence of inference attacks on genomic data

Florian Kerschbaum (SAP SE – Karlsruhe, DE)

Many methods are known to privately query, analyze and compare genomic data. However, as has been shown by Goodrich in the Mastermind attack, repeated queries leak sufficient information in the result in order to quickly infer the secret genomic data. In this talk I will present a method that can deter such attacks in an efficient and secure manner using fuzzy commitments and zero-knowledge proofs.

3.7 Reality check – Implementing personalized therapies based on genomic data in a clinical setting

Oliver Kohlbacher (Universität Tübingen, DE)

Personalized immunotherapies based on epitope-based vaccines are an exciting strategy for personalized cancer treatment. The integration of different types of high-throughput data (exome, transcriptome, proteome, HLA ligandome) poses a number of interesting research problems. Implementing this into a clinical setting, however, results in several regulatory, organisational, and legal issues that we will discuss in the context of the iVac project implemented for personalized cancer immunotherapy at the university hospital in Tübingen.

3.8 Challenges faced by Hospitals using NGS for Diagnostics

Adam Molyneaux (Sophia Genetics SA – Lausanne, CH)

The presentation will talk about the practical problems faced by hospitals in using NGS as a diagnostic tool, explaining how Sophia set out to help them, where we are now and where we think we need to go in the future.

3.9 Controlled Functional Encryption

Muhammad Naveed (University of Illinois – Urbana-Champaign, US)

U.S. Department of Health & Human Services reports that health records of more than 39 million individuals have been breached from hospitals and other healthcare institutions. Therefore, a patient may worry about the privacy of her sensitive health data, and may want healthcare providers to learn only limited information, e.g., result of a particular test. In general, a patient may want healthcare providers to only learn information allowed by the patient specified policy. Such privacy concerns are not limited to healthcare domain. Existing cryptographic techniques do not provide realistic solution to this problem. For example, secure computation would require each patient, or an agent of the patient, to run a computationally intensive program on her computer for each computation. Functional encryption can solve the problem, but it is extremely inefficient and is based on untested cryptographic hardness assumptions.

In this work, we propose a new cryptographic model called "Controlled Functional Encryption (C-FE)" that allows us to construct realistic and efficient constructions. As in functional encryption, C-FE allows a user (client) to learn only certain functions of encrypted data, using keys obtained from an authority. However, we allow (and require) the client to send a fresh key request to the authority every time it wants to evaluate a function on a ciphertext. We propose two C-FE constructions: one for inner-product functionality and other for any polynomial-time computable functionality. The former is based on careful combination of CCA2 secure public-key encryption with secret sharing, while later is based on careful combination CCA2 secure public-key encryption with Yao's garbled circuit. Our main contributions in this work include developing and formally defining the notion of C-FE; designing efficient and practical constructions of C- FE schemes achieving these definitions for specific and general classes of functions; and evaluating the performance of our constructions on various application scenarios.

Our constructions are based on efficient cryptographic primitives and perform very well in practical applications. On a laptop, with Intel Core i7 processor and 8GB RAM, our construction takes 1.28s and consumes 132KB bandwidth for a 1,000 SNP disease marker in personalized medicine application. In genomic patient similarity application, comparing two 4-million SNP profiles costs $0.0143, takes 4 minutes and consumes 53.77MB bandwidth.

3.10 Engineering data privacy – The ARX data anonymization tool

Fabian Prasser (TU München – Klinikum Rechts der Isar, DE)

One of the main focuses of the seminar are privacy problems that arise from the combination and correlation of genomic and clinical data. While privacy for genomic data does pose significant challenges, problems with privacy of clinical data must not be underestimated. While a plethora of methods have been proposed for dealing with many aspects of de-identifying such data, only few (prototypical) implementations are available. Actually, the complexity of implementing privacy technologies is an often overlooked challenge.

In this talk we will present the open source anonymization tool ARX, which has been carefully engineered to support multiple privacy technologies for relational datasets. Our tool bridges the gap between different scientific disciplines by integrating methods developed and used by the statistics community with data anonymization techniques developed by computer scientists. ARX has been designed from the ground up to ensure scalability and it is able to process very large datasets on commodity hardware. The software implements a large set of privacy models: (1) syntactic privacy models, such as k-anonymity, l-diversity, t-closeness and d-presence, (2) statistical models for re-identification risks, and (3) differential privacy. Moreover, it supports multiple risk models and more than ten different methods for evaluating data utility, including loss, precision, non-uniform entropy and KL divergence. Data can be transformed automatically, semi-automatically and manually using a complex method that integrates global recoding, local recoding, categorization, generalization, suppression, microaggregation and top/bottom-coding. All methods are accessible via a comprehensive cross-platform graphical user interface.

Our talk will contribute to the overall seminar topic by comprising an overview of the possibilities and limitations of modern anonymization tools. We will also discuss challenges and possible further developments.

3.11 On a Novel Privacy-Preserving Framework for Both Personalized Medicine and Genetic Association Studies

Jean-Louis Raisaro (EPFL – Lausanne, CH)

So far, several efforts have been undertaken for protecting genomic data and still enabling its functionality. We can put them in two distinct categories: (i) approaches for private clinical genomics, and (ii) approaches for privacy-preserving genetic research. Yet, a main limitation of these approaches is that they restrict the private use of the data only to a single specific purpose, thus significantly slowing down the deployment of privacy-enhancing technologies in a real operational setting. In this work, we address this limitation by proposing a new privacy-preserving framework that is flexible enough to enable for both personalized medicine and genetic association studies on encrypted patients' data. Based on our previous research on private disease risk tests, we extend the previously proposed system model proposed in order to support also privacy-preserving replication and fine-mapping genetic association studies under the assumption of an honest-but-curious adversary. In particular, patients' data

are stored encrypted on a centralized storage and processing unit (SPU), and the different healthcare stakeholders, or medical units (MU), can only obtain the study end-result without ever seeing the actual data.

3.12 Applying Homomorphic Encryption for Practical Genomic Privacy

Kurt Rohloff (NJIT – Newark, US)

This talk outlines a lattice encryption scheme that provides Proxy Re-Encryption (PRE) capabilities with Homomorphic Encryption (HE). We identify several high-level use cases for a mixed PRE and HE capability. We discuss early and implementation and experimental results.

3.13 Realizing differentially private genome-wide association studies

Sean Simmons (MIT – Cambridge, US)

The growing stockpiles of genomic data found in biomedical repositories and patient records promise to be an invaluable resource for improving our understanding of human diseases. In particular, there is interest in using this genomic data to perform genome wide association studies (GWAS). Recent work, however, has shown that sharing this data–even when aggregated to produce p-values, regression coefficients, or other study statistics–may compromise patient privacy.

One proposed solution is to use a privacy preserving technique known as differential privacy. This approach, which works by slightly perturbing the data, protects patient privacy while still allowing researchers access to their genomic data. Unfortunately, existing differentially private GWAS techniques have limitations in terms of accuracy, computational efficiency, and their ability to deal with heterogeneous populations. In this presentation I will give an overview of recent work we have done to help overcome these bottlenecks, work which moves privacy preserving GWAS closer to real world applicability.

3.14 Robust Traceability from Trace Amounts

Adam Davison Smith (Pennsylvania State University – University Park, US)

The privacy risks inherent in the release of a large number of summary statistics were illustrated by Homer et al (PLoS Genetics, 2008), who considered the case of SNP allele frequencies obtained in a genome-wide association study: Given the minor allele frequencies from a case group of individuals diagnosed with a particular disease, together with the genomic data of a single target individual and statistics from a sizable reference dataset independently

drawn from the same population, an attacker can determine with high confidence whether or not the target is in the case group.

In this work we describe and analyze a simple attack that succeeds even if the summary statistics are significantly distorted, whether due to measurement error or noise intentionally introduced to protect privacy. Our attack only requires that the vector of distorted summary statistics is close to the vector of true marginals in ℓ_1 norm. Moreover, the reference pool required by previous attacks can be replaced by a single sample drawn from the underlying population. The new attack, which is not specific to genomics significantly generalizes recent lower bounds on the noise needed to ensure differential privacy (Bun, Ullman, and Vadhan, STOC 2014; Steinke and Ullman, 2015), obviating the need for the attacker to control the exact distribution of the data. In particular, the attack shows that natural relaxations of differential privacy (such as "Pufferfish", "coupled-worlds privacy" and related notions) are subject to the same lower bounds as full-strength differential privacy when many one-way marginals are released.

3.15 e-Biobanking: architectural and algorithmic solutions for the seamless, safe and secure storage and sharing of large biomedical data

Paulo Jorge Veríssimo (University of Luxembourg, LU)

The biomedical data lifecycle is changing dramatically, both due to factors like the generalization of physical sample collection, or the advent of NGs machines, and also due to the pressure for data sharing in name of the progress of biomedical research. Both factors have been inducing ad-hoc technical solutions, bolted on the classical lifecycle, such as use of clouds and promotion of web access, which are bound to augment the threat surface in non-negligible ways. We have been researching on avenues which preserve the desired functional evolution, but satisfy the need for built-in, by-design privacy, integrity and availability of such critical data. This talk reports advances toward new distributed systems architectures and privacy-preserving algorithms which, if successful, may foster what we call e-biobanking ecosystems, coalitions of stakeholders including hospitals, researchers, biobanks, or NGS providers.

Drawing on recent results based on the cloud-of-clouds paradigm, we first show how innovative distributed architectures may foster the advent of secure and dependable constellations of private and public clouds belonging to diverse stakeholders, with separation of risk and concerns, for example, making researchers able to perform operations on mix-criticality data residing in public and private clouds. Secondly, we show how to prevent concrete re-identification attacks on genomic data, leveraging on the above-mentioned architectural framework. We propose a method that systematically detects privacy-sensitive DNA segments coming directly from an input stream. Our method neutralizes threats related to recently published attacks on genome privacy based on short tandem repeats, disease-related genes, and genomic variations. The method can evolve automatically as new privacy-sensitive sequences are identified. Furthermore, the detection machine easily fits the e-biobanking model, by streamlining the cloud storage with the NGS production cycle by using Bloom filters and scaling out to faster sequencing machines.

3.16 The NIH Genome Privacy Challenges: Bringing Security Technologies to Biomedical Users

Xiao Feng Wang (Indiana University – Bloomington, US)

The growth of genome data and computational requirements overwhelm the capacity of servers. Many institutions and NIH are considering the cloud computing service as a cost-effective alternative to scale up research. Privacy and security are the major concerns when deploying cloud-based data analysis tools. In the past few years, progress has been made on secure data-dissemination and computation technologies but it is still not clear the gap between what they can provide and what are expected in the biomedical community. In the past two year, the genome privacy team at Indiana University works together with the iDASH NCBC center organized two NIH-sponsored genome privacy competitions. In this talk, I will provide information about these challenges and what we have learnt.

4 Working groups

4.1 Data sharing across domains

Emiliano De Cristofaro (University College London, GB)

This working group focused on understanding the requirements and the objectives of genomic data sharing in the research environment vis-a-vis the related privacy and security challenges it poses. In particular, participants analyzed existing initiatives as researchers have already developed protocols for exchanging DNA information across the web. Initiatives like the Global Alliance and Matchmaker Exchange1 aim to make it easier for geneticists and bioinformaticians to search, share, and retrieve genomic and epigenomic data that can help them with their research as well as treatment experimentation across institutions. However, while these projects put forward self-regulated codes of conduct and frameworks guided by human rights principles, non-discrimination, and procedural fairness, ultimately, their privacy practices boil down to reliance on volunteers' informed consent as well as ethical guidelines punishing misuse, intentional de-anonymization, or wide disclosure of personally identifiable information. The working group concluded that closer collaborations and exchanges need to take place so that privacy can be embedded from the outset in these protocols. In particular, participants agreed that data minimization approaches should be followed, without requiring the presence of fully-trust parties or shifting the liability of data leaks on the researchers. To this end, solutions from cryptography and differential privacy can offer viable promising opportunities but a number of research problems remain open with respect to efficiency, scalability, resilience to errors, and the lack of off-the-shelf readily available to non-cryptography experts.

4.2 Architecture and middleware

Aniket Kate (Purdue University – West Lafayette, US)

This working group focused on genomic data integrity issues and middleware architectures to make genome processing system transparent to involved clientele. With its heterogeneous nature, mixed criticality, and longevity and irrevocability issues, genomic data presents unique challenges in terms of integrity, privacy and reliability. It is observed that we cannot ignore arbitrary faults (e.g, memory bit-flips) and active attacks while designing privacy-preserving solutions for genome privacy. Therefore, the working group called for novel middleware genome data processing solutions that achieve scalability, availability, and performance along with integrity and confidentiality requirements.

The group discussed distributed computing and cryptographic approaches to overcome these challenges. A theme that emerged out of the discussion was to do data processing in a secure manner in a distributed/outsourced environment rather than doing it in an insecure and error-prone manner on the machines themselves locally. The group also discussed about designing distributed encrypted file systems and indexing solutions for large-scale genome data. From the cryptographic point of view, necessity of authenticated data structures and verifiable computations for genomic data was considered. The group briefly explored the possibility of using known cryptographic tools such as Merkle hash trees, watermarking, erasure coding, and secret sharing, and then called for tailored integrity protection and provenance solution for genomic data.

4.3 Inference Control

Muhammad Naveed (University of Illinois – Urbana-Champaign, US)

Genomic data can be used to infer identity, disease, traits, and kinship. Inference is not a static process and change over time; therefore, inference control procedures should similarly change over time. Management of inference is a complex process, and common people may not understand the all the implications. A neutral entity should provide guidance to the users to decide whether they should donate their data and the type of consent they should give. While inference risks can be reduced, it affects utility. Cryptographic methods cost more and statistical measures add noise. Communicating just the risks to the participants without discussing the potential benefits would be a disservice to the humanity. We have to balance the critical tradeoff between risk and utility, for example, U.S. federal agencies use RU-confidentiality curve to determine risk vs. utility tradeoffs.

We discussed what type of information would be a concern if it could be inferred from genomic data. Identity is a major concern, for example, if an adversary can infer whether an individual is in the case or control group of a genome-wide association study (GWAS). Attribute disclosure was also discussed and whether it constitutes a privacy breach. As attribute disclosure reveals information about population and no individual-level information, considering it a privacy breach could be detrimental to science. Incidental findings are a growing concern; that is, if a medical professional learns something about the patient that

she did not ask for, whether the patient should be told or not. Physicians are also afraid of being sued if they fail to properly adjust dosage based on genomic information.

Genomics is still in its infancy and rapidly growing; therefore, we do not know what information or utility we require from genomic data. This makes addressing privacy concern even more challenging. We discussed if a general and evolving framework for specifying an inference of private data would be a reasonable approach. European Union agrees that anonymous data is a myth, but what type of data would they still be comfortable with releasing; after all, sharing of medical and genomic data is crucial for the development of medicine and technologies for human health. One reasonable approach could be that data owner (or data custodian) can ask people, whose data are being used, for what type of purposes their data could be used and what type of information could be inferred from their data. Another approach would be to specify necessary conditions that should be satisfied before one can obtain data for a study, for example, keeping detailed record of who uses the data and for what purpose. Such conditions could not be satisfied for public use summary statistics.

Sharing of genomic data is crucial for our understanding of disease and human health. The participants raised several points about sharing of genomic data, such as, can we share single nucleotide polymorphism (SNP) without knowing knowledge and power of the adversary? Or can we publish rare variants (e.g., occurring in a single person) in public datasets? Risk assessment is important here. Commonly used risk assessment models are geared towards worst-case adversaries. A challenge here is to develop reasonable risk models against plausible adversaries. The first step in addressing this challenge is to determine what risk is acceptable. For example, explaining the risk during the consent process such that the participants understand the actual risk, developing a comprehensive set of precautions and procedures data collectors should use, and respecting the right not to know. It looks like that research community does not like inference control, but maybe it makes sense in other contexts such as direct to consumer or legal contexts?

4.4 Usage models of genomic privacy

Kurt Rohloff (NJIT – Newark, US)

The goal of the working group was to identify the stakeholders for genomic applications. We commenced our analysis by identifying that storage and processing of genomic information could occur at multiple levels. These stakeholders include data producers, who turn samples into data, and sample owners, such as patients and cryobanks. The stakeholders also relate the concept of a "genome owner" who could either be a sample owner, or a designated representative, such as a physician.

The physician (or pathologist) is primarily responsible for consent. Once consent is granted, the sequencing facility receives the sample and performs sequencing. Interpretation is performed by a interpreter / bioninformatician. If there is a clinical motivation for sequencing, then clinical geneticist could be engaged. Results could also be sent to researchers, such as for population/public health studies. Results turned into text for the doctor about patient and actionable information. Relevant attacks on this ownership chain include insider attacks, such as by a network administrator. Data integrity is also perceived as an issue, either from

adversaries, or from simple computing over large datasets where bit flips could happen that would alter data.

A feasible research scenario involves multiple data controllers contributing data, such as for collaboration projects which includes phenotypic data. A general question Is one of who has data with property X. Requests made in this scenario include one of identifying who has data in a specific format. A general approach, to improve performance, is to move the algorithm to the data.

Another challenge is dealing with the scenario of when consent is revoked. Participants need to be able to indicate to remove the data.

As an indication of scale, a participant would need to keep up to 4 million variants for a specific individual over a set of 200 million known variants. Every single 4 or 5 million.

Participants

- Luk Arbuckle
CHEO Research Institute – Ottawa, CA
- Erman Ayday
Bilkent University – Ankara, TR
- Marina Blanton
University of Notre Dame, US
- Dan Bogdanov
Cybernetica AS – Tartu, EE
- Emiliano De Cristofaro
University College London, GB
- Zekeriya Erkin
TU Delft, NL
- Jacques Fellay
EPFL – Lausanne, CH
- Kay Hamacher
TU Darmstadt, DE
- Zhicong Huang
EPFL – Lausanne, CH
- Jean Pierre Hubaux
EPFL – Lausanne, CH
- Mathias Humbert
Universität des Saarlandes, DE
- Aniket Kate
Universität des Saarlandes, DE
- Stefan Katzenbeisser
TU Darmstadt, DE
- Florian Kerschbaum
SAP AG – Karlsruhe, DE
- Oliver Kohlbacher
Universität Tübingen, DE
- Florian Kohlmayer
TU München – Klinikum Rechts der Isar, DE
- Alexander Kaitai Liang
Aalto University, FI
- Huang Lin
EPFL – Lausanne, CH
- Bradley Malin
Vanderbilt Univ. – Nashville, US
- Adam Molyneaux
Sophia Genetics SA – Lausanne, CH
- Muhammad Naveed
University of Illinois – Urbana-Champaign, US
- Jun Pang
University of Luxembourg, LU
- Fabian Prasser
TU München – Klinikum Rechts der Isar, DE
- Manuel Prinz
DKFZ – Heidelberg, DE
- Jean-Louis Raisaro
EPFL – Lausanne, CH
- Kurt Rohloff
NJIT – Newark, US
- Dominique Schröder
Universität des Saarlandes, DE
- Vitaly Shmatikov
University of Texas – Austin, US
- Sean Simmons
MIT – Cambridge, US
- Adam Davison Smith
Pennsylvania State University – University Park, US
- Thorsten Strufe
TU Dresden, DE
- Qiang Tang
University of Luxembourg, LU
- Carmela Troncoso
IMDEA Software – Madrid, ES
- Juan Ramon Troncoso Pastoriza
University of Vigo, ES
- Gene Tsudik
Univ. of California – Irvine, US
- Paulo Jorge Veríssimo
University of Luxembourg, LU
- Xiaofeng Wang
Indiana University – Bloomington, US