A Security Framework for Population-Scale Genomics Analysis

Ali Gholami*, Jim Dowling †, Erwin Laure*
*PDC & HPCViz
School of Computer Science and Communication,
KTH Royal Institute of Technology, Stockholm, Sweden
Email: {gholami, erwinl}@pdc.kth.se
†School of Information and Communication Technology,
KTH Royal Institute of Technology, Stockholm, Sweden
Email: jdowling@kth.se

***Abstract*—Biobanks store genomic material from identifiable individuals. Recently many population-based studies have started sequencing genomic data from biobank samples and cross-linking the genomic data with clinical data, with the goal of discovering new insights into disease and clinical treatments. However, the use of genomic data for research has far-reaching implications for privacy and the relations between individuals and society. In some jurisdictions, primarily in Europe, new laws are being or have been introduced to legislate for the protection of sensitive data relating to individuals, and biobank-specific laws have even been designed to legislate for the handling of genomic data and the clear definition of roles and responsibilities for the owners and processors of genomic data. This paper considers the security questions raised by these developments. We introduce a new threat model that enables the design of cloud-based systems for handling genomic data according to privacy legislation. We also describe the design and implementation of a security framework using our threat model for BiobankCloud, a platform that supports the secure storage and processing of genomic data in cloud computing environments.**

Keywords*—Security, Privacy, Genomics, Access Control, Cloud Computing

I. Introduction

A biobank is a repository that stores organized collections of biological samples and data to be used in research and personalized medicine [1]. Many of the samples in biobanks will be sequenced in the coming years using next-generation sequencing (NGS) machines, as the cost of sequencing has been dropping rapidly in recent years. There is significant interest in using cloud services for biobanks [2] due to the flexibility of cloud services along with the economic benefits and availability of on-demand computing resources and storage capabilities provided by cloud computing

Recent biobank legislation in several jurisdictions in Europe (such as Finland [3] and Sweden [4]) has lead to the definition of role for managing genomic data, such as Data Access Controllers, Data Processors and Auditors. The intention of this legislation is to ensure that biobanks protect and promote the privacy and integrity of data subjects' personal data. However, much of the existing privacy legislation hinders biobanks from using cloud services because of the way data management roles for genomic data are defined at present and due to restrictions imposed by the current rules for managing genomic data. For example, both the EU Data Protection Directive (DPD) [5] and the US Health Insurance Portability and Accountability Act (HIPAA) [6] demand efficient security and auditing mechanisms to ensure the privacy of data subjects through sharing of personal data among different participants.

According to privacy regulations, cloud computing environments that process genomic data are required to implement technical measures to protect the privacy of data subjects. Secure software engineering best practices recommend threat modeling methodologies as a process to ensure security for developing information systems. There are several well known security threat modeling frameworks and tools, such as OCTAVE [7] and STRIDE [8]. Unfortunately, the complexity of such frameworks makes them difficult to apply to projects with limited resources in an agile approach. In addition, privacy is not emphasized in such methodologies which may cause a significant overhead (both legally and technically) when building privacy-preserving cloud applications because privacy and security are two distinct concepts.

This paper outlines the design and implementation of a security framework for BiobankCloud, a platform that supports the secure storage and processing of genomic data in cloud computing environments. The proposed framework was built on the new cloud privacy threat modeling (CPTM) approach [9], [10] to define the privacy threat model for processing NGS data according to the DPD [5].

The main contributions of this paper are:

- Designing and implementing a security framework based on the CPTM as a new privacy threat modeling methodology.
- Developing a flexible and granular role-based access control (RBAC) [11] model for sharing NGS data among different participants.

The rest of this paper is organized into four sections. Section II introduces the BiobankCloud and related privacy tools. Section III defines the privacy requirements of the BiobankCloud according to the CPTM. Section IV explains the architecture and implementation of the security framework. Finally, Section V summarizes our findings in the course of this work.

978-1-4673-7813-0/15/$31.00 ©2015 IEEE

II. Background

In this section we define the concept of a BiobankCloud and discuss the security and privacy concerns associated with such cloud environments.

A. Overview of BiobankCloud

We define a BiobankCloud as a platform [2] that is capable of deploying sequencing applications with their dependencies within an environment called an *execution container (EC)*. An EC is a node in a Hadoop cluster[1] that runs an actual NGS analysis experiment. ECs can scale out to thousands of instances for running several open source workflows developed in a BiobankCloud. Researchers can easily handle running ECs without support from bioinformaticians and without purchasing expensive commercial solutions.

The big waves of incoming genomic data from NGS machines or biobanks are distributed into genomic data storage (GDS) instances in a Hadoop distributed file system (HDFS)[2]. The arriving data are in sequence alignment map (SAM) or binary alignment map (BAM) formats. ECs run the actual workflow and read the physical address of each block of data that is needed to run experiments from the HDFS network database (NDB)[3] server, as shown in Fig. 1. The HDFS NDB stores the metadata for the distributed GDSs in a relational MySQL cluser to support high performance operations [14], e.g., 100,000+ read operations/second in 100+ petabytes clusters.

We assume that the BiobankCloud platform is run by a secure and trusted cloud service provider (CSP). The CSP will not share, disclose or delegate either storage or processing of genomic data to other entities that are prohibited by the DPD and ethical frameworks.

Fig.2 illustrates how a file in GDS is mapped to an inode that contains a variable number of blocks. Each block is replicated on a number of machines with minimum 3 copies by default. Other information such as timestamps, owner and quota are also stored as metadata.

B. Privacy Requirements

In [10], Gholami et al. applied the CPTM methodology for defining the DPD privacy requirements and identified top privacy threats faced by a BiobankCloud.

The CPTM methodology provides an agile approach for privacy threat analysis. It also provides guidelines to help mitigate the effects of the threats that have been identified for a variety of cloud computing service models within the EU's jurisdiction. The privacy requirements discussed by CPTM include lawfulness, informed consent, purpose binding, data minimization, data accuracy, transparency, data security, and accountability. The complete definitions of these requirements and corresponding threats are defined in [10] (Sections IV-III and V).

[1]Apache Hadoop, http://hadoop.apache.org/.
[2]Hadoop Distributed File System(HDFS), http://hadoop.apache.org/.
[3]MySQL NDB Cluster, http://dev.mysql.com/doc/refman/5.0/en/mysql-cluster-overview.html.

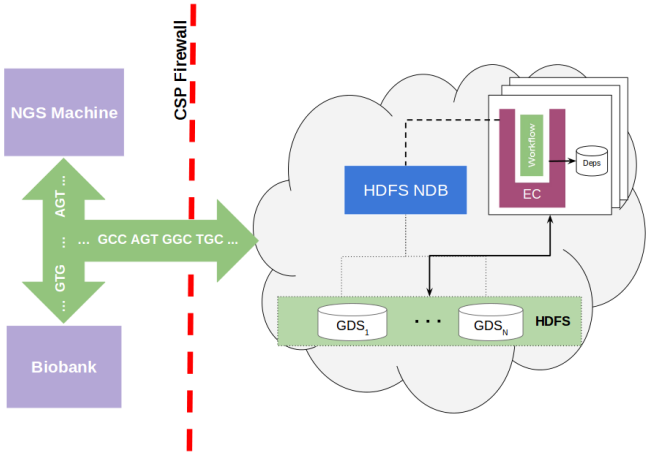

Fig. 1: A BiobankCloud that supports scalable storage and processing of NGS and biobank data. ECs perform stage-in/stage-out on GDSs and dependencies (Deps) to run the researcher's workflows.

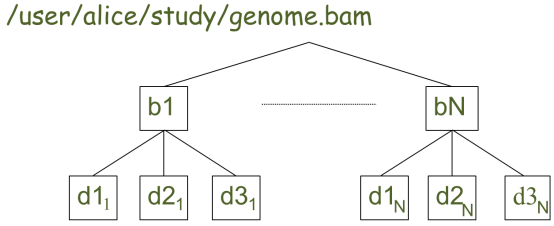

Fig. 2: Alice stores her genomic file (genome.bam) in BAM format in her study directory on HDFS. The genome.bam are converted to inodes. The inodes are divided into blocks b1, ... , bN, where each block is replicated to 3 copies by default into data nodes dN. For example, block b1 is replicated to $d1_1$, $d2_2$, and $d3_3$ copies.

- Lawfulness ensures the legitimate processing of data, so that all processing are conducted within the regulatory framework of the DPD.
- Informed consent justifies processing of genomic data in the platform. The genomic data may have been provided with informed consent through the data provider (DP).
- Purpose binding ensures that personal data processing is performed according to predetermined purposes. The genetic data collectded in the platform will only be processed according to the original purpose for which it was gathered.
- Data minimization restricts unnecessary disclosure of information to third parties, such as a CSP. This reduces the risk of information leakage that could lead to privacy breaches.
- Data accuracy refers to the need to keep data accurate and hence for the DP to keep the data updated. The DP shall only use the information if the accuracy of the data is ensured.

- Transparency entitles data subjects to obtain information about the processing of their data in the cloud.
- Data security invloves implementing technical measures, physical security and organizational safeguards to provide authorized access to personal data.
- Accountability requires both internal and external auditing and control for various assurance and monitoring reasons.

C. Data Anonymization

The e-Science for Cancer Prevention and Control (eCPC) toolkit [12] delivers microdata anonymization through k-anonymity and l-diversity algorithms. Microdata are data that contain information collected on data subjects. The eCPC toolkit is used to anonymize biobanking comma separated value (CSV) data through a Java graphical user interface (GUI) and an sdcMicro processing engine [13].

After successful authentication, the DP will be connected to a relational data storage that contains biobanking data. The DP will be able to extract data in CSV format using the data management component and Java database connectivity (JDBC)[4] driver, as shown in Fig.3. The toolkit allows users to perform a risk estimation for a selected sample microdata in CSV format. The DP can select a set of key attributes to measure the risk of data publishing through the toolkit services. If the level of risk is below a particular threshold, then data will be published by data manager through the REST Web services over secure HTTPS connections to the integration server in the cloud.

The eCPC toolkit also implements two-level encryption of the anonymized data sets based on the secure hash function SHA512[5]and the advanced encryption standard (AES)[6]. This feature enforces extra security measures to encrypt the anonymized data sets for issuing aggregated queries.

III. Related Work

The related work can be divided into three main categories: strong authentication mechanisms, threat modeling, and privacy-enhancing approaches to process genomic data.

A. Strong Authentication Mechanisms

- Multifactor Authentication. Multifactor authentication is a mechanism that requires more than one factor to verify the identity of users. For example, smart cards and tokens [15] or two-factor authentication [16] through popular services such as SMS or direct phone calls are multifactor authentication mechanisms that effectively protect classified information.
- Biometrics. Biometrics authentication is an evolving field for secure password authentication [17], [18]. This method offers authentication based on the measurement of unique physiological characteristics of a user, such as fingerprints (to replace passwords), face recognition, iris codes and behavioral characteristics.
- Public key infrastructure (PKI). PKI provides a scalable secure communication solution in open networks based on an asymmetric pair of keys known as public (shared with all the parties) and private (owned only by user) [19], [20] keys. There is a usability issue with the PKI certificate authentication, since users might have difficulties understanding how to keep public/private keys secret and yet available on a computer for login.

B. Threat Modeling

There has been a substantial amount of research on security threat modeling for various information systems to identify a set of security threats. The threat model helps to reduce the effects of exploiting vulnerabilities associated with the potential threats by an adversary [21], [22], [23] that have been identified to date. There are extensive guidelines [24] published by the cloud security alliance (CSA) for reducing the security risks of cloud services, but these do not include an outline of privacy threat modeling. In [25], Pearson described the key privacy challenges in cloud computing that arise from a lack of user control, a lack of training and expertise, unauthorized secondary usage, the complexity of regulatory compliance, trans-border data flow restrictions, and litigation. Deng et al. proposed LINDDUN [26] (linkability, identifiability, non-repudiation, detectability, information disclosure, content unawareness, and noncompliance) as a generic methodology for privacy requirement elicitation through mapping the initial data flow diagram of systems scenarios to corresponding threats. Recently, Dove et al. discussed legal challenges in genomic cloud computing, including privacy issues [35].

These methodologies or guidelines are generic and none of them are designed for privacy threat modeling in cloud computing environments. For instance, [25] describes the privacy issues in cloud computing but it does not offer a privacy threat modeling methodology. The CPTM differs from the existing work since it is specifically designed for privacy threat modeling in cloud computing environments.

C. Privacy-Preserving Genomic Data Processing

In [27], the authors discussed several privacy issues associated with genomic sequencing. This study also described several open research problems such as outsourcing to cloud providers, genomic data encryption, replication, integrity, and removal of genomic data along with suggestions to improve privacy through collaboration between different entities and organizations. In another effort [28], raw genomic data storage through encrypted short reads is proposed. Our work focuses on the regulatory requirements for cloud computing environments.

Homomorphic encryption is another privacy-preserving solution that is based on the idea of computing over encrypted data without knowing the keys of different parties. To ensure confidentiality, the DP may encrypt data with a public key and store data in the cloud. When the process engine reads the data, there is no need to have the DP's private key to decrypt data. In private computation on encrypted genomic data [29],

[4]The Java Database Connectivity (JDBC), http://www.oracle.com/technetwork/java/javase/jdbc/index.html

[5]SHA512, http://csrc.nist.gov/groups/STM/cavp/documents/shs/sha256-384-512.pdf.

[6]AES, http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf.

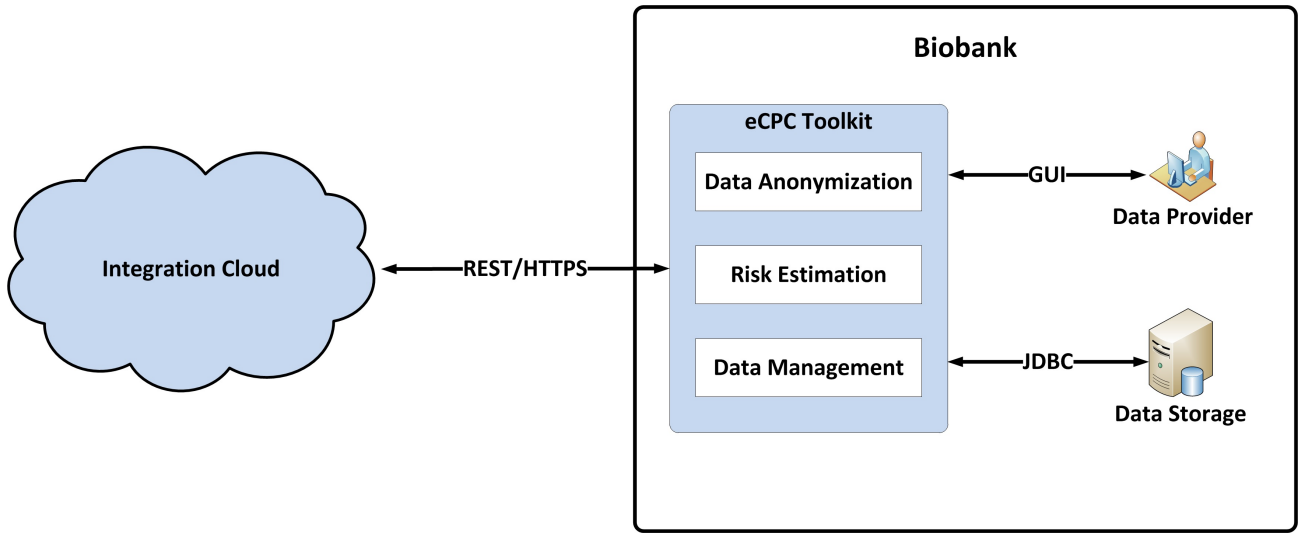

Fig. 3: The eCPC toolkit architecture to anonymize sample availability microdata and evaluate re-identification risk before publishing to the Cloud.

the authors proposed a provide privacy-preserving model for genomic data processing using homomorphic encryption on genome-wide association studies. As mentioned in Section II-A, we assume a trusted private CSP is running the platform in order to minimize the effects of confidential data disclosure during runtime or at rest. The CSP implements organizational safeguards and standards to restrict access to the physical infrastructure.

Anonymization is another approach to ensure privacy of the biobanking data. SAIL [37] provides individual level information on the availability of data types within a collection. Researchers are not able to cross-link (similar to an equality join in SQL) data from different outside studies, as the identity of the samples are anonymized.

IV. BiobankCloud Security Framework

To meet the privacy requirements of the BiobankCloud (Section II-B), we implemented a platform-independent security framework using Java EE (enterprise edition)[7]. The proposed solution includes several plugins and modules embedded in the security framework stack, as shown in Fig.4.

All user interactions with the platform are encrypted by hypertext transfer protocol secure (HTTPS) to guarantee confidentiality and integrity of the traffic. The platform disables caching of sensitive data on each user's client machine and marks user's sessions and cookies as encrypted. This prevents session hijacking or the theft of cookies by an attacker.

The application server (the middle layer in Fig. 4) is an instance of Oracle Glassfish[8] that is protected behind CSP's firewall. Only a limited number of certified personnel have access to CSP's physical infrastructure. Internal or external auditors ensure that CSP implements service organization control (SOC 1/SSAE 16/ISAE 3402) [30] requirements for individual controls to the infrastructure.

The credential server stores users' information, authorization roles and audit trails of data access in a MySQL NDB server. The USERS database includes information such as username, password and account status. The biobank roles are all stored in the ROLES database to be accessed for user management and authorization purposes. The LOGS database stores details of platform usage including authentication, authorization and data access.

A. Access Control

The access control component contains implementation of the authentication, authorization and user management modules. This component enables a new user to register an account request. It also authenticates and authorizes user actions in the platform for running and accessing the stored NGS data.

1) Authentication: Authentication is the process of validating an identity to access the platform. The BiobankCloud supports strong two-factor authentication using *Mobile* and *Yubikey Tokens*. This provides a trade-off solution between security and usability by allowing users to select a convenient authentication method.

Users send authentication requests via a browser to the *authentication* module or to the security policy domain *(custom realm)* [31]: *time-based one-time password (TOTP)* and *Yubikey one-time password (YOTP)* plugins.

A mobile user needs to install Google authenticator[9] which is a mobile app that implements TOTP security tokens from RFC6238 [33]. It generates a 6-digit code in 30 second periods. The user supplies this code as one factor and a plain password as the second factor of authentication.

[7]The Java EE 6 Tutorial, http://docs.oracle.com/javaee/6/tutorial/doc/index.html.

[8]Oracle GlassFish Server , https://glassfish.java.net/downloads/3.1-final.html.

[9]Google Authenticator, https://code.google.com/p/google-authenticator/

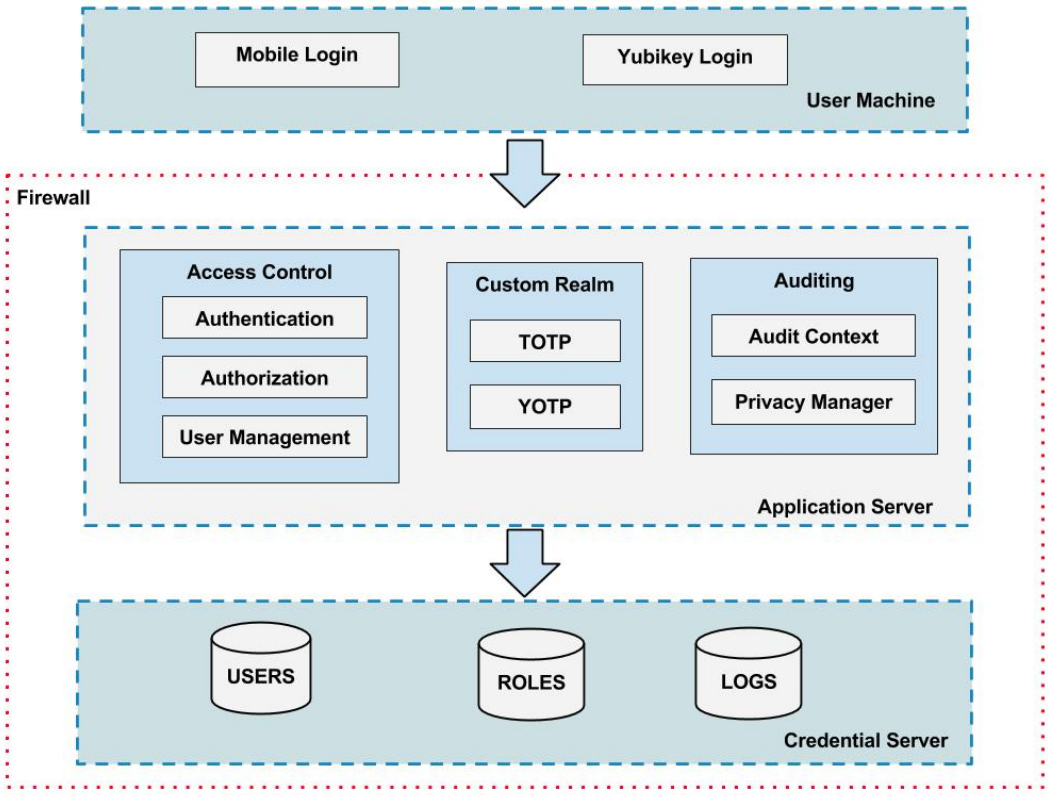

Fig. 4: Security architecture of the BiobankCloud including various modules and plugins to deliver confidentiality, integrity and non-repudiation of data access.

Fig. 5: Installing the BiobankCloud QR code in the user mobile devices.

To provide more usability, the platform issues customized quick response (QR) codes for account setup by users. Fig. 5 shows the process for installing a QR code through a barcode reader into a mobile device during account registration.

A typical scenario for mobile account registration and authentication is as follows.

1) The user installs the authenticator app in a mobile device.
2) The user opens the mobile registration page and creates an account request by entering organizational information, giving consent to the platform ToS and creating a plain password.
3) The platform creates a guest account and sends a QR code to the user's browser.
4) The user scans the QR code with the mobile device and the authenticator app configures the account.
5) The platform sends an validation email to the user's email address.The user verifies his/her email address through clicking an URL.
6) An administrator activates the account on the platform and a notification is sent to the user by email.
7) The user opens the mobile login page and enters the username and plain password as one factor of the authentication.
8) The user opens the Google authenticator and enters the OTP in the login page as the second factor of the authentication.
9) The user issues an authentication request and the platform verifies the plain password and the OTP.

For users without mobile devices, platform offers Yubikey authentication through YOTP. The YOTP (44 characters password) consists of two parts: the first 12 characters indicate the public ID of the Yubikey token [32]. The remaining 32 characters are a unique code for each OTP. A Yubikey token generates the OTPs through a push-button. Generated YOTPs are sent as emulated keystrokes via the keyboard input path, thereby allowing the OTPs to be received by any text input field.

A typical scenario for Yubikey account registration and authentication is as follows.

1) The user opens the Yubikey registration page and creates an account request by entering organizational information, giving consent to the platform ToS and

creating a plain password.

2) The platform creates a guest account to be approved by an administrator.
3) The platform sends an validation email to the user's email address.The user verifies his/her email address through clicking an URL.
4) The administrator activates the account and sends a Yubikey device to the user through a trusted postal system.
5) The user receives the Yubikey and inserts it into the client machine's universal bus port.
6) The user opens the Yubikey login page and enters the username and plain password as one factor of the authentication.
7) The user pushes the Yubikey button and an OTP will be redirected to the OTP input field of the authentication page as the second factor of the authentication.
8) The user issues an authentication request and the platform verifies the plain password and the OTP.

2) Authorization: Authorization is the process of granting or denying access to the platform resources based on the identity of users. An authorization module enforces security policies that are configured for each role in the active security domain where authentication is performed.

The authorization process checks permission rights when an authenticated user requests access to a service. The BiobankCloud deploys a flexible RBAC model to ensure confidentiality and integrity of data. The RBAC model contains information about the potential roles of individuals within the organization and the associated levels of access to services, as shown in Table I. The platform roles are categorized as admin, auditor, data provider (DP), guest, and researcher. This role model can be extended for new requirements. The definition of each role is as follows.

- Admin: group of users who acts as the platform manager and Ethics Board.
- Auditor: group of users with access to audit trails for auditing.
- Data Provider (DP): group of users who create studies, upload data and assign members to studies.
- Guest: general visitors to the platform who are able to request an account to use the services.
- Researcher: users of the platform that can join a study to run workflows. Researchers also can become DPs through creating a new study and uploading data to the platform.

The authorization system retrieves the groups' information through the custom authentcation realm for users with the valid authenticated sessions. For example, when a user is authenticated, a permission check retrieves all the user's related groups. If requested action is permitted on a service or resource, the user will be granted access.

Assume that Alice and Bob are two authenticated users in the system. Alice needs to enable Bob to access her data as shown in Fig.6. For this purpose, the following actions are taken by Alice, Bob and the platform.

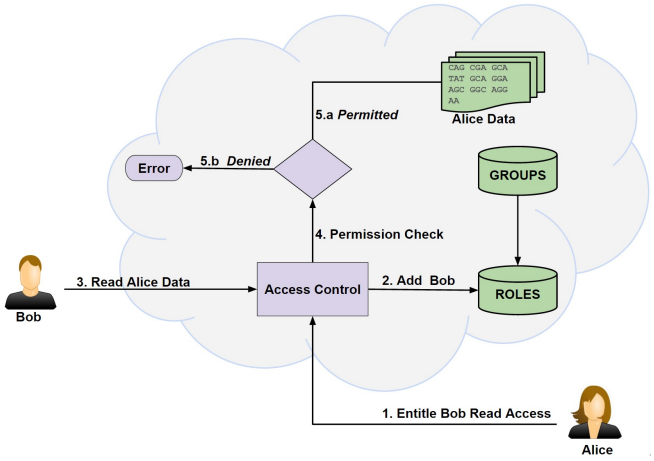

Fig. 6: BiobankCloud authorization system to enforce permissions to access study data.

Service	Auditor	Admin	DP	Guest	Researcher
Audit Management	R	R			
Anonymization Service			X		X
Platform Public Pages	C,R,U,D	R	R	R	R
Privacy Management	R,U	R	C,R,U,D		R
Study Audit Trails		R	R		R
Study Browser		R	C,R,U,D		R,U
Study Data	R,U,D	R	R,U,D		C,R
Study Members		R	C,R,U,D		R
User Administration	C,R,U,D	R			
Workflow Execution			C,R,U,D,X		C,R,U,D,X

TABLE I: Access control table to define the permissions for each role in the platform in regard to using the BiobankCloud services. For example, a researcher can create (C) a new study and will be assigned the DP role afterwards. Then, as a DP, the user will be able to read (R), update (U) and add new members or delete (D) or execute (X) the study.

1) Alice, as study data owner, gives Bob read access to her study.
2) The access control component updates the ROLES table where users are mapped to GROUPS. In this example, Alice and Bob are respectively mapped to the Admin and Researcher groups.
3) Bob initiates a read request to access Alic's study in the cloud.
4) The access control component enforces the existing policies for Bob's request through a permission check (which will have returned either permit or denial:
5) The platform enforces the results of the permission check as permitted or denied:
   a) If Bob is authorized to access Alice's study, he will be permitted to perform a read operation.
   b) If Bob's operation is not permitted on Alice's study, he will be denied access to Alice's data.

B. User Management

User management is the process of controlling which users are allowed to connect to the platform and what permissions

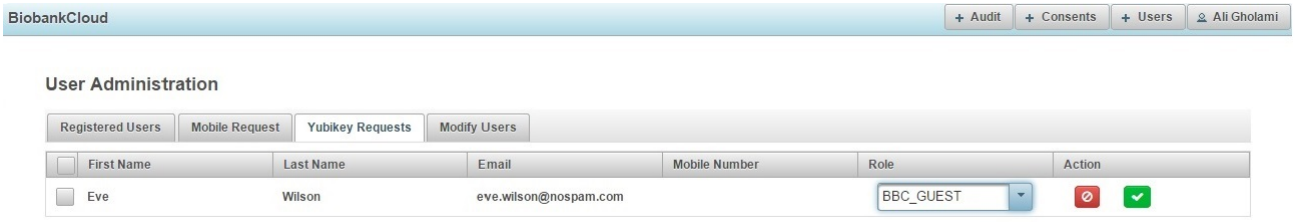

Fig. 7: User management GUI to browse, approve, reject or modify users.

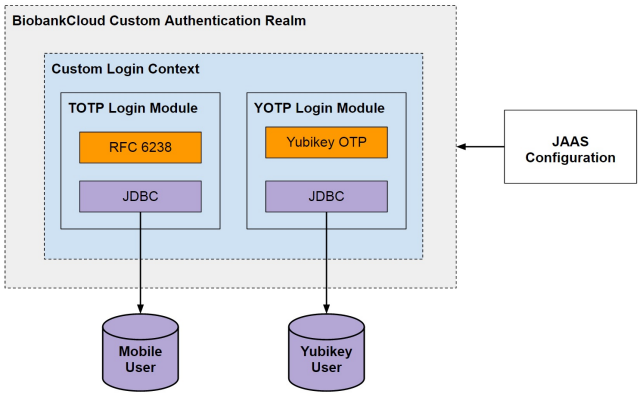

Fig. 8: Custom authentication realm to support authentication for users with and without mobile devices.

they have on each resource. The user management component can create, edit, and remove users, as well as entitling them to perform different operations. The component also delivers the profile setting, account verification and password recovery services.

The user administration panel contains information for all categories of users, as shown in Fig.7. This enables administrators to assign different roles to users or change their status. Depending on the account request type: "Mobile Request" and "Yubikey Request", the administrators approves and assigns roles to new account requests. It is also possible to modify existing users' accounts through the "Modify Users" tab for managing privilege settings and status.

C. Custom Authentication Realm

The custom authentication realm provides the feature of using custom authentication methods for the BiobankCloud. Basically, a custom realm can be considered as a plugin consisting of a Java authentication and authorization service (JAAS) [34] login module and custom realm classes.

Fig. 8 demonstrates the custom login context that reads the JAAS configuration to validate the authentication requests using TOTP and YOTP algorithms. The login modules contain interfaces to access the credentials using the JDBC drvier. The JAAS configuration represents the configuration of the login modules to be defined by the BiobankCloud platform.

1) TOTP Login Module: The TOTP login module authenticates mobile users by validating a password and an OTP (6-digit code). The user password becomes SHA256-encoded[10] to be verified against the credential store (passwords are SHA256-encoded in the credential store). This module retrieves the TOTP secret from the mobile user credential store to validate the 6-digit OTP. The TOTP secret is an auto-generated random 128-bit secret that is base32-encoded[11]. The TOTP secret is stored in the credential storage when the user's account was created (the QR code is the graphical representation of the TOTP password).

The YOTP login module receives an OTP generated by the user's Yubikey and then converts to the OTP to a byte string. The module decrypts the string using the (symmetric) 128-bit AES key that is stored in the credential storage. Then the string's checksum will be checked, and if it is not valid, the OTP will be rejected. As the next step, the non-volatile counter will be compared with the existing value in the credentials store. If it is less than or equal to the stored value, the received OTP will be rejected. If it is greater than the stored value, the received value is stored and the OTP will be validated.

D. Auditing

Auditing is the process of providing proof for resource usage in the platform, for example, giving details of who has accessed each resource and what operations are performed during a given period of time. Auditing helps to ensure that users are accountable and also helps to detect unauthorized attempts to access the resources. The auditing component stores the audit trails with timestamps in the audit storage. All users are assigned a unique POSIX compatible user ID that is an 8-character length alphanumeric username to run the actual workflows in Hadoop environment.

1) Audit Context: Audit context provides secure logging and audit trail browsing for the authorized roles defined in Table I. User with different roles might have different interests and choose different contexts to get information from the audit trails. For example, a DP may only want to audit the usage of a particular study that belongs to him/her, while an auditor may need to audit all the user administration and data access events with different audit options.

2) Privacy Management: The platform requires evidence of consent to allow a DP to share data or run experiments. The privacy management component controls the granular consent of uploaded genomic data since each study has its own privacy

[10]SHA256, http://csrc.nist.gov/groups/STM/cavp/documents/shs/sha256-384-512.pdf.

[11]Base32 encoding: http://tools.ietf.org/html/rfc4648.

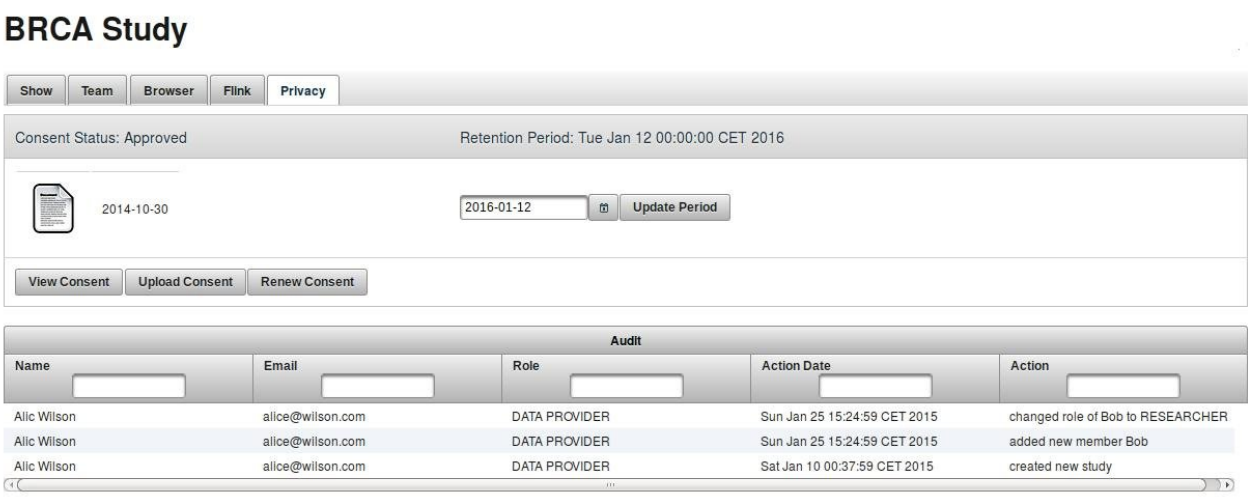

Fig. 9: Privacy configurations to ensure lawfulness, informed consent, purpose binding, transparency and accountability.

settings. The DP uploads the consent form after uploading the study data.

A study consent is approved by an admintrator who acts as the platform manager and by the Ethics Board to ensure legal processing of genomic data. The status of a consent can be *approved*, *rejected* or *not specified*, for example, someone with an admin role might reject consents without legal basis or informed consent in order to stop users with DP and researcher roles from running experiments. The not specified status is used to indicate that a consent form is waiting to be approved by someone with an admin role.

Anonymization of metadata is another measure that reduces the re-identification risk of data subjects by removing the PII. Prior to data upload, the DP anonymizes direct or indirect identifiers and sensitive columns using the eCPC tookit. This toolkit does not anonymize NGS data since the proposed threat model assumes CSP as a secure and trusted entity that is certified to process data.

Fig. 9 demonstrates privacy settings for a study called BRCA. BRCA includes two BRCA1/BRCA2 genes that are associated with breast and ovarian cancers. A DP modifies the privacy settings of the BRCA study that aims to sequence BRCA1/BRCA2 genes in a large cohort of women to identify new pathogenic mutations. In this example, the DP can change the privacy settings through two panels: consent and audit. The consents (top panel) contains interfaces to upload new consent forms or renew an existing consent. On the right-hand side of this panel, the DP updates the retention period. The study owner who acts as a DP can search audit trails in the audit panel. There are filters in the audit panel to make it possible to search in audit trails based on username, role, timestamp, and action.

V. Conclusions and Future Work

This paper presented challenges raised by privacy legislation when it comes to managing NGS data in the Cloud, according to a new Cloud Privacy Threat Modeling (CPTM) methodology. The BiobankCloud is a platform that supports scalable processing of genomics analysis. The CPTM is applied to identify the privacy requirements of the DPD to be deployed in the BiobankCloud security framework.

The security framework includes a role model for sharing genomics data among different participants. The proposed framework lays out the privacy requirements of the DPD through implementing various technical countermeasures and organizational safeguards to prevent or mitigate effects of the identified threats. This work empowers the BiobankCloud to be run by a trusted CSP within the EU's jurisdiction for processing and storing NGS data.

We continue to work towards securing users' credentials in the event of password disclosure [36], in addition to restricting cross-link analysis over multiple datasets, for example, an analytical workflow that requires accessing different studies' data.

Acknowledgements

This work is funded by the EU FP7 project Scalable, Secure Storage and Analysis of Biobank Data under Grant Agreement no. 317871.

References

[1] E. Zika, D. Paci, T. Schulte in den Baumen, A. Braun, S. RijKers-Defrasne, M. Desch Aanes, I. Fortier, J. Laage-Hellman, C. A. Scerri, and D. Ibarreta, "Biobanks in europe: Prospects for harmonisation and networking," *Publications Office of the European Union*, no. EUR24361 EN, 2010.

[2] Scalable, secure storage and analysis of biobank data, http://www.biobankcloud.eu, Accessed: January 2015

[3] The Finnish biobanks, http://www.biopankki.fi/en/suomalaiset-biopankit/, Accessed: January 2015.

[4] Biobanks, Ethics and Regulations, http://www.biobanks.se/ethics.html, Accessed: January 2015.

[5] E. U. Directive, "95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data," *Official Journal of the EC*, vol. 23, 1995.

[6] U. States., *Health Insurance Portability and Accountability Act of 1996 [microform] : conference report (to accompany H.R. 3103)*. U.S. G.P.O [Washington, D.C.], 1996.

[7] C. J. Alberts and A. Dorofee, *Managing Information Security Risks: The Octave Approach*, Addison-Wesley Longman Publishing Co., Inc., 2002.

[8] M. Howard and D. E. Leblanc, *Writing Secure Code*. Redmond, WA, USA: Microsoft Press, 2nd ed., 2002.

[9] A. Gholami, A. Edlund, and E. Laure, "Cloud privacy threat modeling," *The 8th International IFIP Summer School on Privacy and Identity Management for Emerging Services and Technologies, Nijmegen, the Netherlands*, 2013.

[10] A. Gholami, A.-S. Lind, J. Reichel, J.-E. Litton, A. Edlund, and E. Laure, "Privacy threat modeling for emerging biobankclouds," *Procedia Computer Science*, vol. 37, pp. 489 496, 2014.

[11] R. S. Sandhu, E.J. Coyne, H.L. Feinstein, and C.E. Youman. "Role-based access control models," *IEEE Computer*, vol. 29, no. 2, pp. 3847, 1996.

[12] A. Gholami, E. Laure, Peter Somogyi, O. Spjuth, S. Niazi, and J. Dowling, "Privacy-Preservation for Publishing Sample Availability Data with Personal Identifiers," *Journal of Medical and Bioengineering*, Vol. 4, No. 2, pp. 117-125, April 2015.

[13] M. Templ, "Statistical disclosure control for microdata using the rpackage sdcmicro," *Trans. Data Privacy*, vol. 1, pp. 67-85, Aug. 2008.

[14] K. Hakimzadeh, H. Peiro Sajjad, and J. Dowling, "Scaling hdfs with a strongly consistent relational model for metadata," in *Distributed Applications and Interoperable Systems* (K. Magoutis and P. Pietzuch,eds.), vol. 8460 of *Lecture Notes in Computer Science*, pp. 3851,Springer Berlin Heidelberg, 2014.

[15] G.-J. Ahn and D. Shin, "Towards scalable authentication in health services," in *Enabling Technologies: Infrastructure for Collaborative Enterprises*, 2002. WET ICE 2002. *Proceedings. Eleventh IEEE International Workshops on*, pp. 8388, 2002.

[16] R. Di Pietro, G. Me, and M. Strangio, "A two-factor mobile authentication scheme for secure financial transactions," in *Mobile Business*, 2005. ICMB 2005. *International Conference on*, pp. 2834, July 2005.

[17] M. Atallah, K. Frikken, M. Goodrich, and R. Tamassia,"Secure biometric authentication for weak computational devices," in *Financial Cryptography and Data Security* (A. Patrick and M. Yung, eds.), vol. 3570 of *Lecture Notes in Computer Science*, pp. 357371, Springer Berlin Heidelberg, 2005.

[18] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 450455, March 2005.

[19] D. W. Chadwick, D. P. Mundy, and J. P. New, "Experiences of using a PKI to access a hospital information system by high street opticians," *Computer Communications*, vol. 26, no. 16, pp. 18931903, 2003.

[20] H. Gomes, J. Cunha, and A. Zquete, "Authentication architecture for ehealth professionals," in *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS* (R. Meersman and Z. Tari, eds.), vol. 4804 of *Lecture Notes in Computer Science*, pp. 15831600, Springer Berlin Heidelberg, 2007.

[21] B. Schneier, "Threat modeling and risk assessment," in *E-Privacy* (H. Bumler, ed.), DUD-Fachbeitrge, pp. 214229,*Vieweg+Teubner Verlag*, 2000.

[22] Y. Chen and B. Boehm, "Stakeholder value driven threat modeling for off the shelf based systems," in *Software Engineering - Companion*, 2007. *29th International Conference on*, pp. 9192, May 2007.

[23] S.-J. Baek, J.-S. Han, and Y.-J. Song, "Security threat modeling and requirement analysis method based on goal-scenario," in *Proceedings of the International Conference on IT Convergence and Security 2011* (K. J. Kim and S. J. Ahn, eds.), vol. 120 of *Lecture Notes in Electrical Engineering*, pp. 419423, Springer Netherlands, 2012.

[24] Cloud Security Alliance (CSA), "Security guidance for critical areas of focus in cloud computing, v3.0," https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf, 2011. Accessed: October 2013.

[25] S. Pearson, "Privacy, security and trust in cloud computing," in *Privacy and Security for Cloud Computing* (S. Pearson and G. Yee, eds.), *Computer Communications and Networks*, pp. 342, Springer London, 2013.

[26] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen, "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," *Requirements Engineering*, vol. 16, no. 1, pp. 332, 2011.

[27] E. Ayday, J. Raisaro, U. Hengartner, A. Molyneaux, and J.-P. Hubaux, "Privacy-preserving processing of raw genomic data," in *Data Privacy Management and Autonomous Spontaneous Security* (J. Garcia-Alfaro, G. Lioudakis, N. Cuppens-Boulahia, S. Foley, and W. M. Fitzgerald, eds.), vol. 8247 of Lecture Notes in Computer Science, pp. 133147, *Springer Berlin Heidelberg*, 2014.

[28] E. Ayday, E. D. Cristofaro, J.-P. Hubaux, and G. Tsudik, "The chills and thrills of whole genome sequencing," *Computer*, vol. 99, no. PrePrints, p. 1, 2013.

[29] K. Lauter, A. Lopez-Alt, and M. Naehrig, "Private computation on encrypted genomic data," Tech. Rep. MSR-TR-2014-93, June 2014.

[30] SOC 1 Report (Service Organization Controls Report, http://www.ssae-16.com/soc-1/, Accessed: June 2014.

[31] Custom JAAS realm, http://docs.oracle.com/cd/E19226-01/820-7695/6niugeskh/index.html, Accessed: December 2014.

[32] Yubikey Manual,YubiKey-Manual. V3-1, http://www.yubico.com/, Accessed: August 2013.

[33] RFC6238, Time-Based One-Time Password (TOTP), http://tools.ietf.org/html/rfc6238.

[34] JavaTM Authentication and Authorization Service (JAAS) Reference Guide, http://docs.oracle.com/javase/6/docs/technotes/guides/security/jaas/JAASRefGuide.html, Accessed: December 2014.

[35] E. S. Dove, Y. Joly, A.-M. Tass, Public Population Project in Genomics and Society (P3G), International Steering Committee, International Cancer Genome Consortium (ICGC) Ethics and Policy Committee, and M. Knoppers, "Genomic cloud computing: legal and ethical points to consider," *European Journal of Human Genetics*, aug 2014.

[36] D. Mirante and J. Cappos, "Understanding Password Database Compromises," NYU Poly, Tech. Rep. TRCSE201302, September 2013.

[37] M. Gostev, J. Fernandez-Banet, J. Rung, J. Dietrich, I. Prokopenko, S. Ripatti, M. I. McCarthy, A. Brazma, and M. Krestyaninova, "SAIL - a software system for sample and phenotype availability across biobanks and cohorts," *Bioinformatics*, vol. 27, no. 4, pp. 589591, 2011.