DARC: Decentralized Anonymous Researcher Credentials for Access to Federated Genomic Data

Mohammed Alghazwi
University of Groningen
Groningen, Netherlands
m.a.alghazwi@rug.nl

Fatih Turkmen
University of Groningen
Groningen, Netherlands
f.turkmen@rug.nl

Dimka Karastoyanova
University of Groningen
Groningen, Netherlands
d.karastoyanova@rug.nl

ABSTRACT

Genomic data carries unique information about an individual and offer unprecedented opportunities for healthcare. However, the private nature of genomic data and the lack of trust between data owners and data scientists hinder its use. Data owners are hesitant to give control of private genomes to researchers without a thorough inspection of researchers' credentials and their research purposes. The current procedure is to delegate this task to Data Access Committees (DACs) that make the decision on allowing/-denying access through a process that can take 2-3 months. During this process, the consent and conditions set by the patient would need to be interpreted to check if they match the research purpose. In addition, a comprehensive evaluation would be conducted to assess the researcher's identity, qualifications, and the underlying purpose of the research. This difficult and time-consuming process is not scalable since the number of datasets produced and the number of requests continue to increase rapidly. In this paper, we present DARC, a protocol for creating decentralized credentials for researchers when accessing federated genomic datasets. DARC relies on concepts from self-sovereign identity (SSI) and utilizes zero-knowledge proofs (ZKP) and smart contracts. Additionally, we describe a proof of concept implementation of the protocol and provide a preliminary performance evaluation.

1 INTRODUCTION

Genomic data proved to be an essential part of medical research since it carries specific information about individual patients such as susceptibility to a disease and side effects of a medicine. There is also a strong incentive to enable the sharing of genetic information, as COVID-19 pandemic showed, access to a large amount of health-related data allows for faster and better healthcare-related decisions. The integration of genomic data into everyday clinical practice might soon be possible with the advancement of data-sharing methods. However, finding the desired genomic data and getting access to it is limited due to the private nature of the genome since it carries valid information even after an individual passes away and it indirectly affects the descendants and relatives of the data owner.

Most prior approaches for genomic data sharing have followed a centralized model, in which, data owners deposit data into a repository, and the access (e.g., by the researchers) is managed through this repository. In these cases, the data repository often serves as the entity that manages the identities (or relies on trusted Identity Providers), and determines which datasets match the credentials of the requester and what access rules apply to her/him. In the current state of affairs, this approach is not very scalable due to the (expected) explosion of genomic datasets and their frequent use in medical research. Therefore, alternative models have been recently proposed. The most prominent model is the federated data sharing model which allows data owners to seperately host their data in their own storage or through a cloud service, enabling them to have greater control over their data.

1.1 Use Case

The Beacon API [5] is an initiative that follows a federated model to tackle the privacy problems in genomic data sharing. The Beacon API enables any organization to host (i.e., 'beaconize') their genomic datasets without losing control of the dataset. Federated genomic data systems, such as beacons, allow researchers to query data from multiple sources in a standardized way, without having to access the underlying databases or storage systems. Researchers can submit a query to a beacon, and the system will return a response indicating whether or not the query matches any data in the federated system. This allows researchers to gain insights from large-scale genomic data sets, while still maintaining the privacy and security of the underlying data. However, researchers often need to perform more analysis tasks than simple yes/no answers. To facilitate this, beacons need to define the requirements and conditions of access to resources.

The Global Alliance for Genomics and Health (GA4GH) [5] suggests three levels of data access (e.g., open, registered, controlled). These 'access tiers' have different requirements for authorization and imply varying levels of privacy. Open access datasets are accessible to any anonymous users, while 'registered access' datasets are accessible to registered users (aka bona fide researchers) who have the required credentials. Controlled access datasets require additional permissions which are commonly approved by Data Access Committees (DACs) in the form of credentials or certificates.

In the federated data-sharing model described above, there are multiple challenges. Given that the data is hosted on multiple data repositories and controlled by multiple data owners, how can the researchers gather and combine identities/credentials that are issued by multiple institutions (different systems) and selectively reveal a subset of their identities (credential attributes)? The second challenge is how to ensure the credentials' integrity (i.e. not tampered with) and authenticity (i.e. granted by their authoritative source).

The GA4GH defined what is called the "passport standard" [15]. The passport contains researchers' digital identity information including research status and access permissions which are in turn called "visas". The passport standard specifies a set of protocols for creating, transmitting, and verifying passport information. However, it relies on multiple third parties such as *visa issuers* and *passport broker services*. These third-party services play an important role in the functioning of the authentication/authorization system and a trust framework must be established between all parties in the federated data-sharing model for it to work. This trust framework was not defined [15]. For instance, how and why should the data repository (aka clearinghouse) trust the broker? How does the broker establish a list of trusted visa issuers? These questions were not addressed in the GA4GH passport/visa standards.

1.2 Contribution

In this paper, we propose DARC a protocol that allows researchers to aggregate multiple credentials (claims) about themselves, have full control of their identities, and privately reveal a subset of their credentials to data repositories or computing services in order to access or perform computation on the data. The main contributions of this paper thus can be summarised as follows:

- We propose a decentralized privacy-preserving protocol for creating and managing access credentials to federated genome data. This resolves the challenge of verifying credentials issued by multiple institutions and allowing selective disclosure of attributes within these credentials. By combining smart contracts and zero-knowledge proofs, DARC enhances the trust and privacy of the beacon access protocol (the passport standard).
- The protocol enables researchers to control their identities eliminating the need for mediators such as identity providers (brokers). Additionally, data owners are able to define and enforce a trust model for their domain.
- Finally, we benchmark the main components of our proof-of-concept (POC) and demonstrate its practicality and limitations.

2 BACKGROUND

In the following, we briefly describe the main concepts and cryptographic schemes that we used to build our proposed system.

2.1 Blockchain and SSI

Blockchain is a ledger of transactions that is distributed across all nodes in a peer-to-peer (P2P) network. Transactions are verified using cryptographic techniques and consensus protocols. It solves the problem of allowing multiple parties that do not necessarily trust each other to agree on the state of a shared ledger. Smart contracts have emerged recently and they allow modifying the state of the ledger in an automated, trustless, and verifiable way without intermediaries. They enable decentralized applications on the blockchain that found usage in finance, identity management, and healthcare [10]. The main advantage of using blockchain in implementing an identity and access control system lies in the transparency of the credential creation process, and elimination of the need for trusted intermediaries that verify credentials. Furthermore, as the execution of the smart contract takes place on the blockchain network, the outcome is guaranteed to be accurate, resistant to tampering, and visible to all participants. This effectively eliminates the overhead of verifying credentials since the ownership of the account (i.e., decentralized identifier) is sufficient to permit access.

2.2 Zero Knowledge Proofs (ZKPs)

Zero-knowledge proof schemes [4] provide a mechanism for a prover to prove the knowledge of a secret to a verifier with overwhelming probability, without revealing the secret itself. Zero-Knowledge Succinct Non-Interactive Argument of Knowledge (zk-SNARK) [3] is arguably the most popular ZKP protocol and it is the one we employ in this work. Specifically, we use the Groth16 scheme [8]. The Groth16 scheme or any other zk-snark scheme for that matter involves the following steps:

- $\text{Setup}(1^\lambda, \phi) \rightarrow crs$. Given the security parameter λ and the defined circuit ϕ, generate the common reference string crs.
- $\text{Prove}(crs, x, \omega) \rightarrow \pi$. Given the crs, public input x, and witness ω, generate a proof π for the defined circuit ϕ.
- $\text{Verify}(crs, x, \pi) \rightarrow \{0, 1\}$. Given the crs, public input x, and proof π, output 1 if the proof is valid and 0 otherwise.

3 RELATED WORK

The use of blockchain technologies for handling genomic data is on the rise. For a systematic treatment of the topic, we refer the reader to [2]. In what follows, we briefly summarize the related work.

Decentralized anonymous credentials were originally proposed in [6]. More recently, zk-creds [14] and Zebra [12] are two proposed systems that leverage zkSNARK in the construction of decentralized credentials. Both of these papers propose the use of ZKPs to generate anonymous credentials and store the results in a byzantine system such as a blockchain. While these papers use similar techniques to ours, their focus is quite different. *zk-creds* [14] focuses on removing the need for signing keys and converting traditional identity documents such as passports to anonymous credentials. Zebra [12] on the other hand, focuses on the DeFi use case and proposed efficient methods for creating and verifying credentials.

With regards to identity and trust, [11] employed the emerging standards on decentralized identifiers (DID) and verifiable credentials (VC) by W3C [13] to enable trusted processing of sensitive health data in federated machine learning workflows. The main goal of their work is the establishment of trust between different participants and the use of DIDs as the identifiers of participants. While this work achieves the objective of enhancing trust in the federated learning use cases, it does not address the challenges associated with federated access and processing of genomic data. Specifically, the challenges of aggregating credentials that are issued by multiple institutions and allowing selective disclosure of attributes within these credentials.

4 OVERVIEW

In this section, we discuss the system model and provide the threat model that we are considering. For clarity, we use the term "credential" to denote a claim showing that a user proved inclusion in a group of members sharing a common attribute such as research status and access permit.

4.1 System Model

We now describe the components of the DARC protocol and their roles in the overall functioning of the protocol.

- *Credential Issuers (CIs)* are the authorities that issue credentials for users. CIs play the same role as the "visa assertion source" in the GA4GH standard [15]. Credentials issued by each CI are stored in a local credential repository managed by CIs.
- *Researchers* are users who wish to present their issued credentials as proof that they are legitimate to access genomic datasets in the federated data-sharing model. Researchers can be affiliated with multiple organizations and might have different identities (different identifiers).
- *Credential Issuer Smart Contract* has access to the credential repositories of CIs and takes researchers' requests for credentials. Upon receiving these requests, the contract checks whether the request is valid and issues the credential/s once confirmed.
- *Credential registry)* stores the list of verified credentials issued by the credential issuer smart contract. The registry is made available to all data repositories in order to authenticate researchers.
- *Data Repositories* are the entities that provide the data and authenticate researchers based on some access policy.

4.2 Threat Model

In this work, we aim to achieve the following security and privacy requirements:

- We assume that the researchers can act maliciously to get access without holding a valid credential. Therefore, the protocol must ensure that the credentials are valid and are created by an approved CI.
- The list of credentials remains private unless selectively revealed by the holder. The credential verifier is only able to see the revealed claims.
- The revealed claims should only reveal the type of claim (the group it belongs to), not the specific owner (account) of that claim.

It should be noted that the proposed protocol does not completely eliminate the need for trust, since trusted credential issuers are still required. We assume that CIs are trusted to record valid credentials for their researchers and that the data repositories follow the protocol in checking the registry for the appropriate credentials prior to providing access. Thus our focus is on managing trust in a way that eliminates the need for third parties, particularly, identity brokers/concentrators.

5 DARC PROTOCOL

In this section, we describe the protocol when creating and verifying credentials. Figure 1, gives an overview of the protocol, which is described in the following.

5.1 Group Membership

Groups are generated and managed off-chain by CIs and are used to pool together users with a common attribute. For instance, researchers within an organization are grouped since they have common research status credentials. Inclusion in the group asserts that they are indeed bonafide researchers. In the DARC protocol, we use Merkle Trees (MT) to represent groups. Researchers can prove, in a zk-SNARK, that they own an account that is part of the MT. Each CI can create and manage multiple groups each with a specific group identifier G. We use a key-value MT which stores the hash of (k, v) in the leaves of the tree structure. The key k is the identity commitment, and value v is an arbitrary number specified by the CI. In our use case, this value can represent the specific access permit or DAC approval. The hash function used is the Poseidon hash function [7]. The Merkle tree allows the following functions:

- $\mathrm{Add}(k, v) \rightarrow MT'$, adds the Poseidon hash of key-value pair $H(k, v)$ to the tree, and outputs the modified tree MT'.
- $\mathrm{getRoot}() \rightarrow R$, returns the current root of the tree R.
- $\mathrm{Prove}(k, v) \rightarrow \alpha$, given key k and value v, generate the path (proof) α used to prove that $H(k, v) \in MT$
- $\mathrm{Verify}(k, v, R, \alpha) \rightarrow \{0, 1\}$, outputs 1 if the key-value pair is in the Merkle tree, and 0 otherwise.

The group information is made available to the credential issuer smart contract which verifies membership before generating credentials. To improve efficiency, each CI combines Merkle trees, which represent groups within one CI, into one Merkle forest (MF). The hash of each group root R is combined with the group identifier G and inserted into MF. In our proof of concept, we only store the root of the MF which is sufficient to prove membership.

5.2 Credential Generation

The credential issuer contract is responsible for generating credentials based on the rules respecified. First, it verifies user requests by checking the submitted proof of membership, and once verified the credential is recorded in the registry and assigned to an address provided in the request. The address holding the credential in the registry is not linked to the identity or identifier used in the Merkle tree. The implemented ZKP circuit (shown below) enables researchers to prove the follwoing:

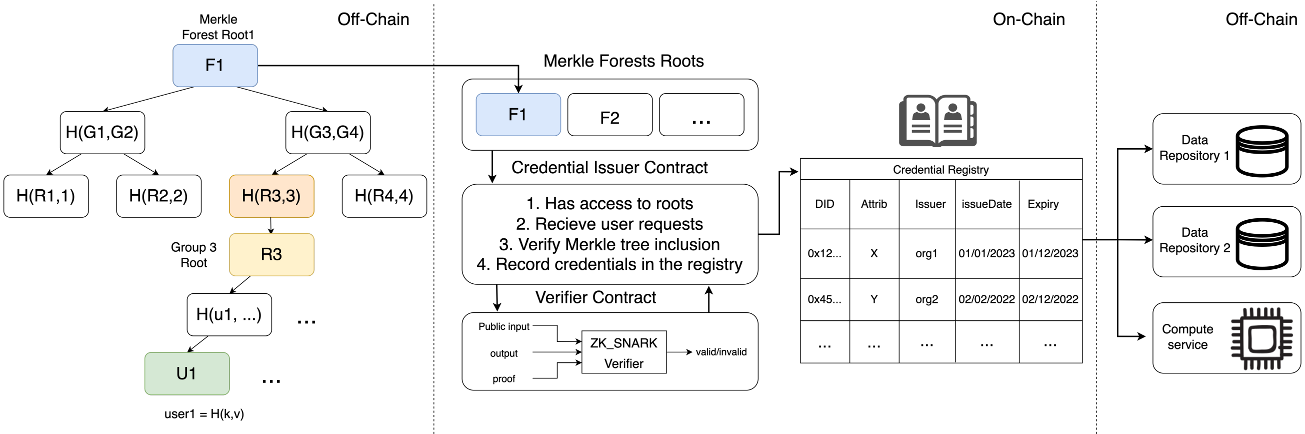
- *Account Ownerships:* prove ownership of the account that is part of one of the CIs group (Merkle tree).
- *Account Membership Proof:* the account is part of a group Merkle tree.
- *Group Membership Proof:* the hash of group root and group identifier are part of the Merkle forest.

```
Private input signals (witness):
{k,v}     \text{identity key-value pair}
$\alpha$         Group Merkle path
$\beta$         Merkle forest path
$R$         Group Merkle root

Public input signals:
$R_f$      Merkle forest root
$G$       Group identifier

Constraints:
{k,v} $\in MT_G$ key-value pair in Group $G$ Merkle tree
$MT_G \in MF$ Group Merkle tree is part of Merkle forest
```

6 PRELIMINARY EVALUATION

ZKP circuits were implemented using Circom [1] and Table 1 shows the outcome of evaluating the circuit. The used height for both MT and MF is 20 which is sufficient to include more than 1 million members in each group. Our experiments have been conducted on a machine with 8 GB of RAM and 2 cores with 3.1 GHz.

We implemented the credential issuer and credential registry in Solidity. Table 2 shows the results of benchmarking transactions on the implemented smart contracts. Benchmarking was done using



DID	Attrib	Issuer	issueDate	Expiry
0x12...	X	org1	01/01/2023	01/12/2023
0x45...	Y	org2	02/02/2022	02/12/2022
...	...	...	...	...

Figure 1: Overview of DARC protocol

Constraints	10200
Compile time (s)	4.11
Trusted Setup time (s)	27.9
Proving key size (MB)	6.1
Verifier contract size (KB)	12
Proof generation time (s)	2.4

Table 1: Overview of Circuit Costs

the *HardHat* [9] and *web*3.*js* tools which create a local Ethereum blockchain and allow creating transactions and retrieving data from the blockchain.

Function	**Gas Cost**
Deploy CI contracts	1,845k
Deploy verifier contracts	1,364k
Store MF root	51k
Verify credential	212k
Store credential	33k

Table 2: Gas Costs of operations in DARC

7 CONCLUSION & FUTURE WORK

We demonstrated DARC, a protocol for the creation and management of decentralized, anonymous credentials for researchers in the federated genomic data-sharing model. Additionally, we presented a proof of concept (POC) implementation of the protocol which combines smart contracts and zero-knowledge proofs. The evaluation of the POC shed some light on the practicality and limitations of the protocol. The proof generation time is low and can be done on the researcher's side, however, the gas costs can be an issue if deployed to a public blockchain. Therefore, the use of side chains or layer 2 blockchains can lower these costs. In future work, we aim to integrate and test DARC with the beacon API and address other challenges that have not been discussed in this work, namely, credential revocation and Sybil-resistance.

REFERENCES

[1] 2022. CIRCOM: A Robust and Scalable Language for Building Complex Zero-Knowledge Circuits. https://doi.org/10.36227/techrxiv.19374986.v1

[2] Mohammed Alghazwi, Fatih Turkmen, Joeri Van Der Velde, and Dimka Karastoyanova. 2022. Blockchain for genomics: a systematic literature review. *Distributed Ledger Technologies: Research and Practice* 1, 2 (2022), 1–28.

[3] Nir Bitansky, Ran Canetti, Alessandro Chiesa, and Eran Tromer. 2012. From extractable collision resistance to succinct non-interactive arguments of knowledge, and back again. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 326–349.

[4] Manuel Blum, Paul Feldman, and Silvio Micali. 2019. Non-interactive zero-knowledge and its applications. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*. 329–349.

[5] Marc Fiume, Miroslav Cupak, Stephen Keenan, Jordi Rambla, Sabela de la Torre, Stephanie OM Dyke, Anthony J Brookes, Knox Carey, David Lloyd, Peter Goodhand, et al. 2019. Federated discovery and sharing of genomic data using Beacons. *Nature biotechnology* 37, 3 (2019), 220–224.

[6] Christina Garman, Matthew Green, and Ian Miers. 2013. Decentralized anonymous credentials. *Cryptology ePrint Archive* (2013).

[7] Lorenzo Grassi, Dmitry Khovratovich, Christian Rechberger, Arnab Roy, and Markus Schofnegger. 2021. Poseidon: A New Hash Function for Zero-Knowledge Proof Systems.. In *USENIX Security Symposium*, Vol. 2021.

[8] Jens Groth. 2016. On the size of pairing-based non-interactive arguments. In *Advances in Cryptology–EUROCRYPT 2016: 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part II 35*. Springer, 305–326.

[9] Hardhat. 2022. *ethereum development environment for professionals.* Retrieved 28.11.2022 from https://hardhat.org/

[10] Shafaq Naheed Khan, Faiza Loukil, Chirine Ghedira-Guegan, Elhadj Benkhelifa, and Anoud Bani-Hani. 2021. Blockchain smart contracts: Applications, challenges, and future trends. *Peer-to-peer Networking and Applications* (2021), 1–25.

[11] Pavlos Papadopoulos, Will Abramson, Adam J Hall, Nikolaos Pitropakis, and William J Buchanan. 2021. Privacy and trust redefined in federated machine learning. *Machine Learning and Knowledge Extraction* 3, 2 (2021), 333–356.

[12] Deevashwer Rathee, Guru Vamsi Policharla, Tiancheng Xie, Ryan Cottone, and Dawn Song. 2022. Zebra: Anonymous credentials with practical on-chain verification and applications to kyc in defi. *Cryptology ePrint Archive* (2022).

[13] Drummond Reed, Manu Sporny, Dave Longley, Christopher Allen, Ryan Grant, Markus Sabadello, and Jonathan Holt. 2020. Decentralized identifiers (dids) v1. 0. *Draft Community Group Report* (2020).

[14] Michael Rosenberg, Jacob White, Christina Garman, and Ian Miers. 2022. zk-creds: Flexible Anonymous Credentials from zkSNARKs and Existing Identity Infrastructure. *Cryptology ePrint Archive* (2022).

[15] Craig Voisin, Mikael Linden, Stephanie OM Dyke, Sarion R Bowers, Pinar Alper, Maxmillian P Barkley, David Bernick, Jianpeng Chao, Mélanie Courtot, Francis Jeanson, et al. 2021. GA4GH Passport standard for digital identity and access permissions. *Cell Genomics* 1, 2 (2021), 100030.