Journal of the American Medical Informatics Association, 26(11), 2019, 1370–1374
doi: 10.1093/jamia/ocz107
Advance Access Publication Date: 26 June 2019
Case Report

Case Report

ARBoR: an identity and security solution for clinical reporting

Eric Venner,* Mullai Murugan,* Walker Hale, Jordan M. Jones, Shan Lu, Victoria Yi, and Richard A. Gibbs

Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA

*These authors contributed equally.
Corresponding Author: Richard A. Gibbs, PhD, Baylor College of Medicine, 1 Baylor Plaza #226, Houston, TX 77030, USA (agibbs@bcm.edu)

Received 21 February 2019; Revised 24 May 2019; Editorial Decision 29 May 2019; Accepted 30 May 2019

ABSTRACT

Motivation: Clinical genome sequencing laboratories return reports containing clinical testing results, signed by a board-certified clinical geneticist, to the ordering physician. This report is often a PDF, but can also be a paper copy or a structured data file. The reports are frequently modified and reissued due to changes in variant interpretation or clinical attributes.
Materials and Methods: To precisely track report authenticity, we developed ARBoR (Authenticated Resources in a Hashed Block Registry), an application for tracking the authenticity and lineage of versioned clinical reports even when they are distributed as PDF or paper copies. ARBoR tracks clinical reports as cryptographically signed hash blocks in an electronic ledger file, which is then exactly replicated to many clients.
Results: ARBoR was implemented for clinical reporting in the Human Genome Sequencing Center Clinical Laboratory, initially as part of the National Institute of Health's Electronic Medical Record and Genomics (eMERGE) project.
Conclusions: To date, we have issued 15 205 versioned clinical reports tracked by ARBoR. This system has provided us with a simple and tamper-proof mechanism for tracking clinical reports with a complicated update history.

Key words: clinical genomics, genetic report

INTRODUCTION

Clinical genome sequencing has made a major impact on the diagnosis and treatment of a broad range of clinical presentations, notably within the care of cases of pediatric cancer and rare Mendelian disease.[1–4] Maintaining the security and authenticity of clinical reports containing genetic testing results is an essential component of a clinical laboratory workflow, as they contain protected health information as well as genomic findings that impact health care.[5,6] As the continuous rapid advancement of genetic understanding necessitates re-review of previous findings, there are often updates to clinical reports, resulting in multiple report "versions."[7,8] Further, reports can be altered or damaged, even after they have passed beyond the clinical laboratories' control. The importance of ensuring that reports that reach patients and clinicians are authentic representations of the most recent and updated versions of those issued from the diagnostic laboratories is therefore an ongoing challenge.

A common solution for this data tracking challenge is for a clinical laboratory to maintain a database of individual report updates.[9] This centralized approach has the advantage of providing complete and immediate data access to information from selected individuals, but this requires dedicated staff to maintain centralized compute resources and permission structures for indefinite periods. The central databases are also unable to detect whether reports have been altered subsequent to issue, either maliciously or accidentally. Communication of report au-

thenticity to individuals who are external to the clinical laboratory and may not have database access, including patients, clinicians, and auditors, is also a challenge.[6] Another approach is to issue signed certificates alongside reports; however, this shifts the problem to tracking the certificates themselves and does not provide a mechanism for determining whether a signed report is the most recent.

Here, we present ARBoR (Authenticated Resources in a Hashed Block Registry), a simple and efficient approach that addresses the difficulties of monitoring report authenticity by providing a record of cryptographically signed reports, stored in a replicated ledger. This approach augments the secure delivery path between clinical labs and a downstream electronic health record (EHR) system by providing a durable method to verify the authenticity of files and detect whether authentic newer files are known to exist.

ARBoR is not a blockchain, as it lacks a decentralized consensus mechanism (see discussion),[10–13] although it does have similarities to blockchains, used, for example, by cryptocurrencies. ARBoR provides some of the benefits of a blockchain such as data provenance, an immutable audit trail, and authenticity verification.[10] As an alternative to a centralized database, ARBoR employs a file-based ledger that resides in a secure cloud location, written to only by trusted agents and replicated by authorized ARBoR users. The hash chain for each report in the ARBoR ledger ensures that a report is verifiable by authorized ARBoR users, and the use of cryptographic signatures for each block prohibits the ledger from being modified by malicious actors.

True blockchain technologies may have broad applicability to the fields of genomics[14] and medical informatics. For example, Guardtime[15] created a system that allows any updates to a medical record to be tracked via a blockchain, which provides an immutable record of health services. MedRec[16,17] has been implemented to control access to medical records using blockchain technology, giving patients control over who is able to access their medical information, and many other additional efforts to strengthen the management of medical records are currently underway.[10] The benefits of blockchain technology may be especially high when multiple independent healthcare entities need to interact or share data.[18] It should, however, be noted that these designs are more feature-rich than what we have currently implemented for ARBoR. ARBoR balances simplicity and precise tracking to singularly solve the report authenticity problem.[6]

ARBOR: ARCHITECTURE AND DESIGN

The ARBoR system (**Figure 1**) implements a replicated ledger of cryptographically signed clinical reports, enabling the institutions receiving those reports to authenticate and establish the report version. The system consists of 3 parts. The first is the ARBoR Service, which is integrated into the reporting laboratory's clinical interpretation and reporting pipeline and is responsible for creating a record in a centralized, publicly readable ledger for the clinical report and related files. Second, the ARBoR Client, which is typically run by an institutional end-user as part of the EHR ingestion process, receives a replicate of the ledger and uses it to authenticate and validate new clinical reports and related files. Last, ARBoRScan (**Figure 2**) is a mobile app for both iOS and Android platforms that also receives a replicate of the ledger that is used by an institutional end-user to verify the authenticity and version of either a printed or EHR integrated report before use.

The ledger stores records of clinical reports for multiple samples as cryptographically signed blocks. Each block represents a single clinical report and stores metadata about the sample, cryptographically signed contents of the report or file, and block metadata (Table 1). Block metadata consists of a timestamp, the hashed contents of the previous block, and the digital signature of the contents of the current block. The previous block hash is required for every block except the first (the "genesis block"), and subsequent block hashes form a "chain." Hashing uses the strong SHA3-512[19] algorithm. Although ARBoR stores the hashed output of the PDF in the block, the contents of the report are not recoverable from the ledger. This ensures the security of any protected health information that may be on the report.

ARBOR: ENGINEERING AND USE

Adding new reports to the ledger

ARBoR requires a public/private key pair that is specific to a clinical laboratory. The ARBoR Service retains the private keys to sign new blocks. The ARBoR Client and ARBoRScan use the associated public keys to authenticate blocks.

To add a new report, a block is first created to represent that report. The ARBoR Service rebuilds the ledger with the existing validated blocks and adds new blocks for any new reports. A new block is linked to the previous and signed with a private key. The ledger can then be replicated along with the report(s) and the public key for use by the ARBoR Client, usually timed with a data freeze or other bulk release of reports. The ARBoRScan app automatically downloads a copy of the latest ledger on use. Any attempts by a malicious actor to tamper with or alter the ledger will cause report validation to fail.

Updating reports

If a report is updated, the same process is followed. The ARBoR service rebuilds the ledger, linking the block for the updated report to the most recent previous block in the ledger. These entries form a "chain" that records the history of the reports for that sample, which can be used to identify tampering with the ledger and report history.

Verifying digital copies of reports

An institutional user with a report file, the ledger, and the laboratory's public key can use the ARBoR Client software to verify that the report has a valid ledger entry. ARBoR Client compares the signature of the report to the object hash in the ledger to verify the authenticity of a report. ARBoR Client can also be used to check whether a report is the most recent. Alternatively, an institutional user can also use the ARBoRScan mobile app to scan the report's QR code to check whether the report is authentic and is the most recent. This approach remains effective, even subsequent to the conclusion of the project, without requiring elaborate resources for maintenance and upkeep.

Verifying printed reports

To verify a printed report, a user first scans the report's QR code using the ARBoRScan mobile app. This app, using the ledger file, extracts the hashed report ID from the QR code, retrieves the corresponding ledger entry, and verifies report authenticity and version.

Security

Taken together, the set of stringent checks described previously creates a secure system. First, security of ARBoR Service's addition of new blocks to the ledger is guaranteed by requiring that a new block be created only by a clinical laboratory holding a valid private key

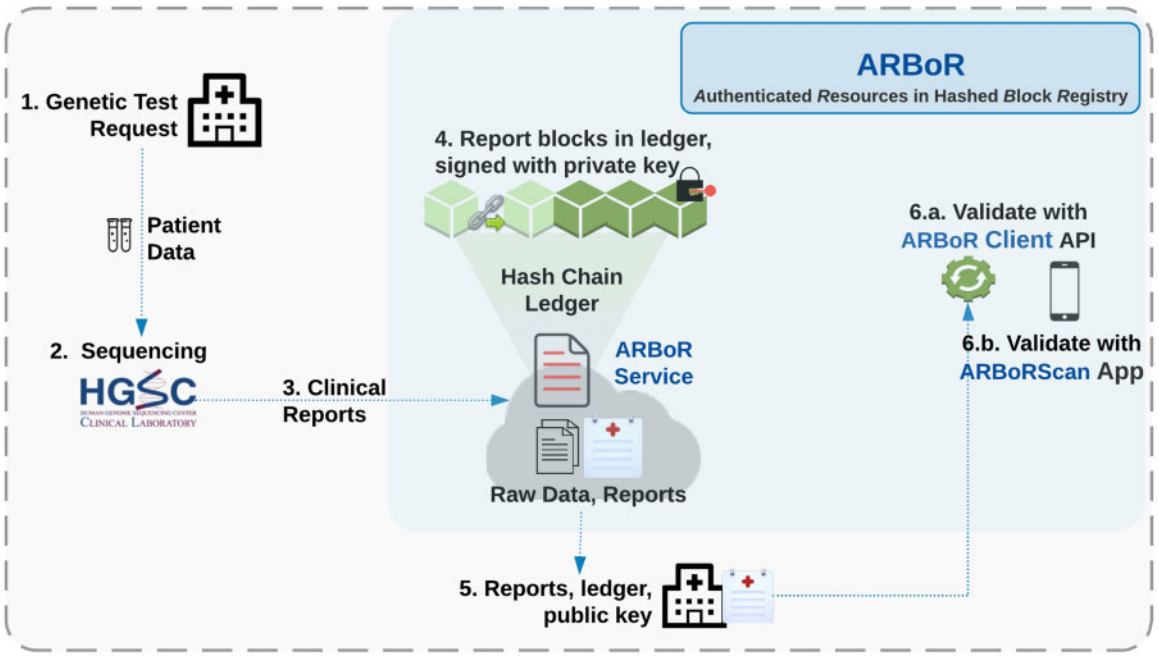

Figure 1. Overall design of the system. The overall ARBoR (Authenticated Resources in a Hashed Block Registry) system consists of 3 parts: (1) ARBoR Service is integrated into the pipeline of a clinical laboratory. Once the pipeline has generated a clinical report and related files, it uses the ARBoR Service to create and store a record about this file in a public ledger. (2) ARBoR Client is typically run by an institutional end user as part of the ingestion process for new clinical reports and related files. (3) ARBoRScan (Figure 2) is a mobile app for both iOS and Android platforms and is typically run by an end user to fetch metadata about existing reports and check the authenticity and versions of these reports. It also maintains a local copy of the ledger. The primary input is scanning QR codes from existing reports.

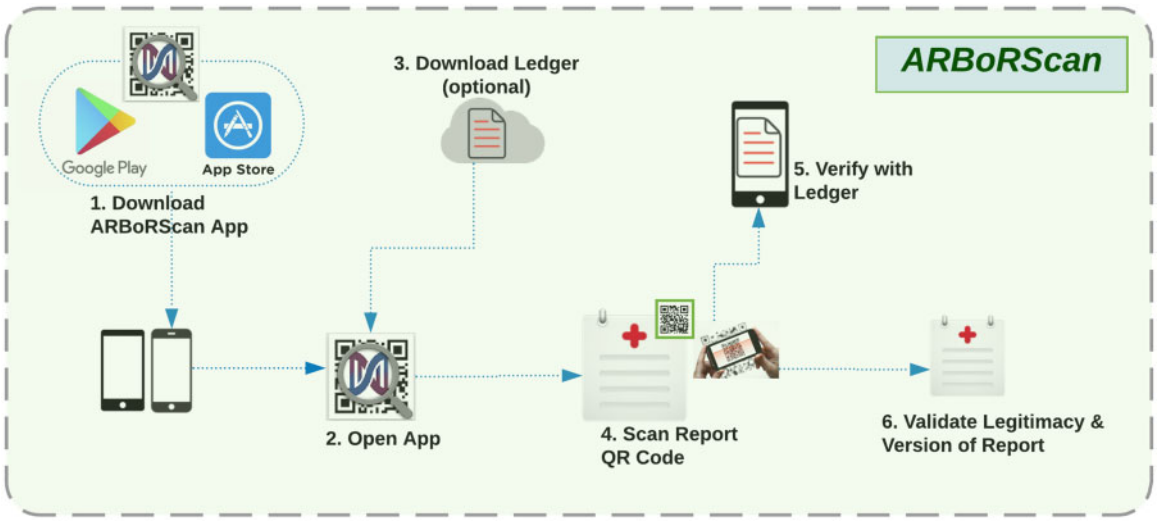

Figure 2. ARBoRScan Process Flow. The mobile app ARBoRScan is available for free download from both iOS and Android app stores. Accessing the app will automatically check for the latest copy of the ledger; users have the option to download the latest copy of the ledger or use the local version on the device. Users are able to use this app to scan the QR code (2-dimensional barcoded hashed report identifier) on either a printed or EHR integrated report and verify the authenticity and version of the report before use. ARBoR: Authenticated Resources in a Hashed Block Registry.

and that transactions occur over encrypted channels. Next, users of ARBoR Client and ARBoRScan need not fully trust the ARBoR Service, as they are able to perform independent checks of the ledger to verify that old records have not been altered without the private key and that the records form a chain. By checking the current ledger in the Service against previously known blocks and known public keys, both the pipeline and the client are in a position to detect alterations in the external behavior of the Service. While the clinical reports themselves are delivered outside of ARBoR, as depicted in the previous sections, ARBoR ensures the authenticity and integrity of the reports and files delivered, thereby allowing the downstream EHR to reject untrusted files.

DISCUSSION

We introduced the ARBoR system for tracking and versioning clinical reports. It produces an encrypted ledger that can be replicated and provided to clinical partners for use in verifying their copies of

Table 1. Contents of each block in the ledger.

Block element	Required?	Description
Block signature	Yes	Digital signature of the SHA3-512 cryptographic hash of all the contents of *this* block
Block index	Yes	Index numbering for every block
Block timestamp	Yes	Block creation timestamp
Previous block hash	Yes	SHA3-512 cryptographic hash of the contents of the *previous* block
Block type	Yes	Supports versioning and advanced features. ("clinical-1")
Individual ID	No	Anonymized individual ID
Sample ID	No	Anonymized ID of the sample from the patient and associated test request
Object type	Yes	Type of the file (ie, PDF, XML, etc.) corresponding to the test results of the sample
Object hash	Yes	SHA3-512 cryptographic hash of the file contents
Additional fields	No	Any additional metadata as key/value elements

Each block describes a clinical report and is generated by the ARBoR (Authenticated Resources in a Hashed Block Registry) Service. Input to the Service ranges from the required minimum of a file or folder to file/folders and its associated metadata. Elements in black are auto-generated by the Service, and elements in blue are parsed and obtained from metadata input. The Object Type allows for the inclusion of multiple types of artefacts, and the Block Type allows for the ARBoR software to evolve and add new features without invalidating existing ledgers.

clinical reports. ARBoR consists of 3 components: ARBoR Service is employed to produce the ledger, whereas ARBoRScan and ARBoR Client own replicated copies of the ledger and use it to verify reports. This system allows us to create a tamper-evident, easily verifiable, and secure record of clinical report histories.

The ledger is centralized in that only trusted instances of the ARBoR Service may write to it. Future iterations of ARBoR may function as a decentralized system, if for example we need to support multiple clinical laboratories issuing reports. There are multiple viable strategies to extend ARBoR to support independent clinical laboratories. In the extreme, decentralization would require communication between ARBoR Service instances and thus a consensus mechanism to agree on the latest ledger, bringing ARBoR closer to a blockchain implementation. Considering some of the current technical limitations of a blockchain implementation such as speed, scaling, and cost effectiveness,[15] and the fact that this consensus-based decentralization includes functionality and complexity not required for the current use case, we reserved these features for future releases. Future iterations of ARBoR that add, for example, a consensus mechanism could take advantage of infrastructure like Hyperledger Fabric.[20] A natural preliminary extension would be a "private blockchain" design[21] that allows only trusted agents to write to the ledger and simplifies the consensus mechanism. Further, ARBoR Client and ARBoRScan are designed to function without an active instance of ARBoR Service. At the end of a project, the ARBoR Service instance for that project can be retired, with the final ledger being a deliverable of the project. So long as the ARBoR Client and ARBoRScan instances have received the final ledger, no centralized infrastructure needs to be maintained.

Although initially developed for tracking clinical reports, the system described here is extensible to any file type. The Human Genome Sequencing Center's Clinical Lab at Baylor College of Medicine frequently produces other exportable data deliverables (eg, VCF, BAM, and CRAM files) and there are use cases where it is desired to securely link additional metadata (eg, quality control metrics) to these files in perpetuity. The ability to track these files with ARBoR is a future area of development. Another extension that we are exploring is to use ARBoR to track report delivery transactions. This would require that report delivery be recorded by the ARBoR Service as an entry in the ledger with the report that was delivered.

An important benefit to using a hash chain with cryptographic features for the ledger is that the ledger as a whole can be revalidated at any time. Validating the ledger consists of starting with the most recent block and then following the chain, checking that hashing each block matches the expected value recorded in the following block. Since each block is digitally signed, it is possible to detect if fraudulent block(s) have been added.

We have deployed this technology for our Electronic Medical Record and Genomics (eMERGE)[22] project, in which clinical reports are issued to clinical sites across the United States. Reports are distributed in XML and PDF formats; using ARBoR, both report formats can be authenticated and verified to contain the latest information. This approach provides a durable method for tracking all of our deliverables, ensuring their authenticity and data integrity with a complete audit trail. To date, this approach has aided the data management for 15 205 reports.

CONCLUSION

ARBoR provides clinical laboratories with a simple and efficient means for tracking clinical reports and provides a replicated ledger that can travel with reports to their recipients. This ledger provides a means of validating clinical reports that persists well beyond the life span of a project. We have successfully applied this system in the context of the eMERGE clinical sequencing project.

FUNDING STATEMENT

This work was supported by National Human Genome Research Institute grant numbers U01HG8664 and HG008898 to RAG.

AUTHOR CONTRIBUTIONS

RAG initially conceived of the study and design of the software. EV completed initial proof-of-concept implementation. EV and MM contributed to the architecture and design of the software system. WH refined the design of the software system. WH, JMJ, VY, and SL contributed to the software implementation. EV, MM, WH, JMJ, SL, VY, and RAG contributed to the draft of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

The ARBoR source code is available on GitHub at https://github.com/BCM-HGSC/ARBoR. ARBoR is written in Python and takes advantage of open source libraries such as beautifulsoup4, pycrypto, pytest, and conda. Currently, we run ARBoR on DNAnexus (a cloud compute provider). ARBoR is open source under an MIT license.

The mobile app ARBoRScan is available for iOS at https://goo.gl/QZXpqg and for Android at https://goo.gl/cLdKB8. Users will be able to download the ARBoRScan iOS and Android apps for free from the Apple App Store and Google Play Store respectively and use the sample reports attached as supplements as a test set to explore and verify the functionality of ARBoRScan.

CONFLICT OF INTEREST STATEMENT

EV is cofounder of Codified Genomics, a software company that creates genomic variant interpretation tools. Richard Gibbs: Baylor College of Medicine receives payments from Baylor Genetics Laboratories, which provides services for genetic testing. All other authors declare no competing interests.

REFERENCES

1. Lee H, Deignan JL, Dorrani N, *et al*. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 2014; 312 (18): 1880–7.
2. Manolio TA, Chisholm RL, Ozenberger B, *et al*. Implementing genomic medicine in the clinic: the future is here. *Genet Med* 2013; 15 (4): 258–67.
3. MacArthur DG, Manolio TA, Dimmock DP, *et al*. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014; 508 (7497): 469–76.
4. Yang Y, Muzny DM, Xia F, *et al*. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 2014; 312 (18): 1870–9.
5. Aronson SJ, Clark EH, Babb LJ, *et al*. The GeneInsight Suite: a platform to support laboratory and provider use of DNA-based genetic testing. *Hum Mutat* 2011; 32 (5): 532–6.
6. Aronson S, Babb L, Ames D, *et al*. Empowering genomic medicine by establishing critical sequencing result data flows: the eMERGE example. *J Am Med Inform Assoc* 2018; 25 (10): 1375–81.
7. Harrison SM, Dolinsky JS, Knight Johnson AE, *et al*. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genet Med* 2017; 19 (10): 1096–104.
8. Rehm HL, Bale SJ, Bayrak-Toydemir P, *et al*. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 2013; 15 (9): 733–47.
9. Huser V, Sincan M, Cimino JJ. Developing genomic knowledge bases and databases to support clinical management: current perspectives. *Pharmgenomics Pers Med* 2014; 7: 275–83.
10. Kuo T-T, Kim H-E, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. *J Am Med Inform Assoc* 2017; 24 (6): 1211–20.
11. Tapscott D, Tapscott A. *Blockchain Revolution: How the Technology behind Bitcoin Is Changing Money, Business, and the World*. New York: Penguin; 2016.
12. Yli-Huumo J, Ko D, Choi S, *et al*. Where is current research on blockchain technology?-A systematic review. *PLoS One* 2016; 11 (10): e0163477.
13. Ramirez R. *Cryptocurrency: Everything You Need to Know about Bitcoin, Ethereum, Blockchain*. Createspace Independent Publishing Platform; 2018.
14. Ozercan HI, Ileri AM, Ayday E, *et al*. Realizing the potential of blockchain technologies in genomics. *Genome Res* 2018; 28 (9): 1255–63.
15. Angraal S, Krumholz HM, Schulz WL. Blockchain technology: applications in health care. *Circ Cardiovasc Qual Outcomes* 2017; 10: e003800. doi: 10.1161/CIRCOUTCOMES.117.003800
16. Azaria A, Ekblaw A, Vieira T, *et al*. MedRec: using blockchain for medical data access and permission management. In: 2016 IEEE 2nd International Conference on Open and Big Data (OBD); 2016. doi: 10.1109/obd.2016.11
17. Gammon K. Experimenting with blockchain: Can one technology boost both data integrity and patients' pocketbooks? *Nat Med* 2018; 24 (4): 378–81.
18. Mettler M. Blockchain technology in healthcare: The revolution starts here. In: 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom); 2016. doi: 10.1109/healthcom.2016.7749510
19. Dworkin MJ. *SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions. FIPS PUB 202*. Gaithersburg, MD: National Institute of Standards and Technology; 2015. doi: 10.6028/nist.fips.202
20. Dhillon V, Metcalf D, Hooper M. The Hyperledger Project. In: *Blockchain Enabled Applications: Understand the Blockchain Ecosystem and How to Make It Work for You*. New York: Springer; 2017; 139–49. doi: 10.1007/978-1-4842-3081-7_10
21. Ahram T, Sargolzaei A, Sargolzaei S, *et al*. Blockchain technology innovations. In: 2017 IEEE Technology & Engineering Management Conference (TEMSCON); 2017. doi: 10.1109/temscon.2017.7998367
22. The eMERGE Consortium. Harmonizing Clinical Sequencing and Interpretation for the eMERGE III Network. bioRxiv 2018 Nov 1 [E-pub ahead of print]. doi: 10.1101/457523