International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 03, March 2026)

Privacy-Aware Forensic DNA Evidence Matching System Using Synthetic STR Profiles

Saurabh Kumar[1], Sangam Raj[2], Saqib, Vishal Kumar[3]

School of Computer Science and Engineering, Lovely Professional University. Phagwara, Punjab, India

***Abstract*-- Forensic DNA analysis is a critical source of information for criminal investigation, however storing full genetic information presents a lot of issues from privacy, ethics, and security standpoints. In general, traditional databases store raw Short Tandem Repeat values, which can identify personal information related to a person's identity and genetic attributes. This research examines the design and development of a Privacy-Aware Forensic DNA Evidence Matching System to securely store genetic evidence while allowing legitimate forensic comparison. The system consists of synthetic DNA profiles made from synthetic allele-frequency tables. The backend was developed using FastAPI and PostgreSQL, and runs from a curated multi-population dataset, structured schema, and automated ingestion pipelines into bulk synthetic profiles. The system provides secure storage, access, and submission of DNA evidence via organized APIs. This work provides the technical foundation for a secure and extensible digital forensics system. We have shown that forensic matching of DNA is achievable, and can happen without revealing identifiable sensitive genetic information. Future work will investigate embedding privacy-preserving encodings and better matching algorithms to improve the security and usefulness of the system.**

I. Introduction

DNA profiling is important in contemporary forensic investigations, such as individual identification, comparison of samples from a crime scene, and informing legal decisions [1][2][3]. Although the STR-based methodology is reliable as a profiling method, most forensic environments that utilize DNA evidence store full genetic data in a plaintext format [4][5]. Storing raw genetic markers generates serious privacy and ethical dilemmas, because sensitive genetic material has the potential to share information regarding identity-linked biometric characteristics, ancestry, or personal attributes [6][7].Consequently, there is a need for forensic DNA systems that match and analyze without putting raw genetic data [8] [9]. To fulfill this principle, this project examines a privacy-aware DNA evidence-matching system where actual genetic profiles are never stored. Rather, the system holds synthetic STR profiles drawn from publicly accessible allele-frequency datasets [6] [10]. The privacy-aware DNA evidence-matching system produced in this project entails backend development, database design, dataset ingestion, evidence processing and routing, and testing.

The objective of this system is to provide a secure archiving process for DNA profile artifacts in a FastAPI PostgreSQL workflow. Ultimately, the completed workflow can serve as a platform for a secure, evidence-based forensic DNA platform that will be able to scale with future privacy-preserving matching algorithms and additional security enhancements.

II. Literature Review

For many years, forensic DNA profiling has been a reliable tool in criminal investigation, because Short Tandem Repeat markers have sufficient discriminatory information to identify individuals [1] [2] [3]. National database implementations typically store the STR identifiers in plaintext, to facilitate profile construction and comparisons [4] [5]. However, the implication of raw allele values presents a real risk to privacy, because even a small proportion of markers can reveal shared ancestry, kinship, or other sensitive biological phenotypic information [6] [7]. The risk of unauthorized access to a DNA database is allowing for public anxiety, due to an increased awareness of contemporary database structures and access methods, profoundly influence personal genetic privacy [5] [8]. There is significant public interest in creating forensic systems that allow STR identification without exposing sensitive genetic information [3,9]Additionally, new need-to-know studies underscore the importance of alternative privacy-preserving strategies such as hash-based CRNs, approximate matching with Bloom-filter encodings, secure multiparty computation, and homomorphic encryption as legitimate examples of protecting genetic information while redesigning forensic databases and back-end processing systems [8,10].

III. Existing System

The new system is aimed at the creation of a secure privacy-accessible way of comparing DNA evidence that would not subject the users to the dangers of holding real human genetic material [4] [5] [7]. It is a system that is based on using synthetically formatted Tandem Repeat profiles built using publicly accessible allele-frequency tables as opposed to a biological sample [6] [10].

This will make sure that the system does not deal with any human genetic information, sensitive information, or any other identifiable information. This framework brings together the production of synthetic data, a structured database model, backend routes, and limited evidence manipulation techniques. The back end will be developed with the framework Fast API and PostgreSQL and will store the population data, loci, STR profiles, and evidence submissions. It also offers specific application programming interfaces to add, remove and interrogate STR data in a secure manner controlling access without the unnecessary exposure that could be caused [8] [9]. This design is mainly used to address the problem of genetic-privacy and at the same time maintain the utility of forensic DNA comparison [1] [2] [3]. The system provides a safe testing, experimentation, and forensic-research environment that never works with real DNA profiles. It is also the basis of future improvements with privacy-sensitive encodings, sophisticated matching functions, and enlarged digital-forensics processing [8] [10].

Table 1.
Review of existing forensic DNA methods and unresolved challenges

Year	Authors	Type of Method	Algorithm / Focus	Brief Description	**Gap Identified (What Our System Solves)**
2023	Teja Bhukya	Forensic DNA Analysis Overview	STR profile sequence	Covers DNA structure, STR markers, extraction, PCR, profiling, databases, and ethics. Entirely theoretical.	**No privacy-preserving storage or matching: relies on plaintext STR databases. Our system uses synthetic STR data + secure backend designed to avoid storing real DNA.**
2021	Jaya Lakshmi Bukyya et al	Forensic DNA Overview	STR, SNP, mtDNA sequences.	Covers laboratory methodologies, DNA extraction, PCR, DNA profiling, mixtures, and ethical challenges.	**No digital privacy-preserving measures, no secure backend architecture, no synthetic STR data. Our system creates privacy-preserving backend + synthetic STR profiles.**
2023	Saisha Nayyer and A K jaiswal	forensic dna overview	RFLP, VNTR, STR, AFLP, mtDNA	Covers traditional techniques DNA profiling, extraction and markers.	**No secure digital environment, no Synthetic STR data, no backend system, no privacy-preserving architecture. Our work fills this gap.**
2023	Salem k Alketbi	Forensic DNA (Science) Overview	STR, SNP, NGS, Phenotyping.	Covers advances in forensic genetics, sequencing, phenotyping and ethical challenges.	No "system" design or architecture for a backend privacy-aware system. Our system applies synthetic STR data + secure backend storage + evidence workflows.
2022	The royal society	Forensic DNA Primer	Legal + scientific fundamentals.	Will discuss the science of STR, statistics, and database applications in court and legal perspectives.	**Descriptive only; no system design or privacy-preserving method. Our system has real architecture + privacy-aware processes.**
2023	Markus Kayser	Forensic DNA	Phenotypic	Covers forensic DNA phenotype research and	**Focus of the paper is phenotyping, not a privacy-**

Year	Authors	Type of Method	Algorithm / Focus	Brief Description	**Gap Identified (What Our System Solves)**
		Phenotyping Review	appearance, ancestry, age prediction.	prediction of traits using forensic DNA, ethical considerations.	**preserving backend system, does not address synthetic datasets or secure matching methods. Our system solves this gap.**
2023	Halimureti Simayiyiang jiangwei Yan	Forensic DNA Typing Review	NGS, mixture analysis	Covers advances in modern DNA typing, sequencing, and mixture interpretation.	**No secure backend database architecture, no privacy-aware storage infrastructure, nor implementation details or considerations. Our work adds these capabilities.**
2015	John M Butler	Forensic DNA Future Trends	STR expand, NGS, rapid DNA.	Discusses the furture of forensic genetics sequencing and rapid profiling.	**No address to clinically secure handling of STR data and no synthetic STR dataset consideration. Our project fills this gap with implementing a privacy-aware backend system.**
2010	Van Oorschot ,Ballantyne and Mitchell	Trace DNA Review	STR and/or SNP (only) low-template DNA, validation and contamination control.	Covers biological sampling, contamination, low DNA quantities, and interpretive methods.	**Lab focused and has no DIGITAL storage privacy, no backend or storage systems development, or evidence to support synthetic DNA or STR generation. Our system introduces a secure synthetic-based workflows for STR matching.**
2024	Mark Baresh	Machine Learning Review	STR/SNP genotyping using ML	Covers machine learning for forensic DNA typing and interpretation including classification and next generation sequencing (NGS).	**Focus on ML algoritthms not privary infrastructure, or digital privacy-aware approaches. Our work builds a backend + synthetic DNA database + a more privacy aware matching system.**

IV. OBLEM STATEMENT

Forensic DNA databases at this time simply store STR values as plain text. Therefore, these databases are easy targets for data leaks and misuse, as well as for unauthorized access. If someone gains access to these databases, they can see sensitive genetic information that could be used to determine personal information about a person. These old systems were simply not designed with privacy in mind, and they cannot provide a secure way to compare DNA without exposing the real genetic data. Another serious issue is the absence of a trustworthy backend for management of STR profiles. There already exist no systems that upload or preprocess genetic datsets in a privacy-aware manner. Their database schema also similarly cannot accommodate privacy-preserving data formats, making it challenging to keep sensitive information confidential. What is more, there are not a modular / configurable API in most forensic DNA systems. As a result, it is difficult to incorporate new matching algorithms or upcoming privacy techniques that can make the system more secure.

V. METHODOLOGY

This research follows a structured workflow for forensic DNA profiling with a focus on privacy-aware genetic computation. Population-level allele-frequency datasets from published studies are collected, cleaned, and standardized to ensure consistent allele distributions across populations [1][2][3][4]. These frequencies are then organized into allele-pair structures suitable for computational analysis [2].Synthetic STR profiles are generated directly from the standardized allele-frequency tables rather than real biological samples, ensuring that no identifiable human genetic information is produced, stored, or processed [5][10]. Established forensic STR loci, including CODIS-standard markers, are selected due to their strong discriminatory power and are used to construct synthetic profiles for evaluation and comparison [1][2][6][8].The system processes these profiles using structured organization and basic allele-pair comparison mechanisms while avoiding the storage of raw genetic data. Since even partial STR profiles can reveal sensitive information if stored in plaintext, the methodology enforces privacy by design and supports integration with secure and privacy-preserving database practices, while remaining efficient and scalable for forensic research use [3][5][6][7][10].

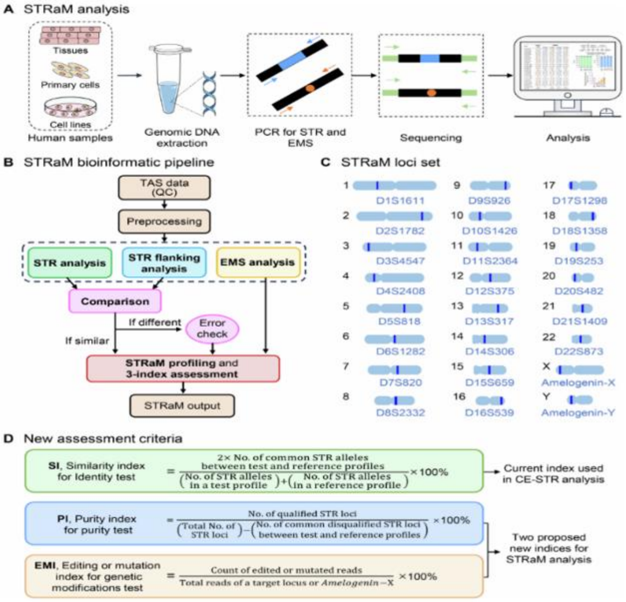
A STRaM analysis
Tissues
Primary cells
Cell lines
Human samples
Genomic DNA extraction
PCR for STR and EMS
Sequencing
Analysis
B STRaM bioinformatic pipeline
TAS data (QC)
Preprocessing
STR analysis
STR flanking analysis
EMS analysis
Comparison
If different
Error check
If similar
STRaM profiling and 3-index assessment
STRaM output
C STRaM loci set
1 D1S1611
2 D2S1782
3 D3S4547
4 D4S2408
5 D5S818
6 D6S1282
7 D7S820
8 D8S2332
9 D9S926
10 D10S1426
11 D11S2364
12 D12S375
13 D13S317
14 D14S306
15 D15S659
16 D16S539
17 D17S1298
18 D18S1358
19 D19S253
20 D20S482
21 D21S1409
22 D22S873
X Amelogenin-X
Y Amelogenin-Y
D New assessment criteria
SI, Similarity index for Identity test = 2× No. of common STR alleles between test and reference profiles / (No. of STR alleles in a test profile)+(No. of STR alleles in a reference profile) ×100%
Current index used in CE-STR analysis
PI, Purity index for purity test = No. of qualified STR loci / (Total No. of STR loci)−(No. of common disqualified STR loci between test and reference profiles) ×100%
EMI, Editing or mutation index for genetic modifications test = Count of edited or mutated reads / Total reads of a target locus or Amelogenin−X ×100%
Two proposed new indices for STRaM analysis

Figure 1: Methodology flowchart of the privacy-aware synthetic STR processing system

VI. IMPLEMENTATION

Forensic DNA systems are typically implemented using a structured workflow that includes dataset preparation, STR handling, and privacy-aware data management. In this study, population-level allele-frequency datasets are first prepared and standardized to ensure accurate representation of population allele distributions [1][2][3][4]. In line with recent privacy-focused approaches, STR profiles are generated directly from these allele-frequency tables rather than from identifiable human DNA samples [5][10].The system is built around established STR schemas and organized into structured, forensic-ready datasets containing loci, allele pairs, and population metadata [2][7]. These datasets support computational pipelines for structured profile handling and future matching analysis, ranging from basic allele-pair comparisons to more advanced statistical or machine-learning–based methods [6][8][10].To ensure privacy, the implementation avoids exposure of raw genetic data and supports secure processing practices recommended in prior research, including designs compatible with privacy-preserving encodings and encrypted comparison techniques [5][9][10]. Synthetic datasets are used throughout evaluation and testing to assess system accuracy, scalability, and privacy assurance without handling real human genetic information.

VII. Results

Researchers across all reviewed literature noted that STR (short tandem repeat) based forensic DNA analysis has great discriminatory power for human identification and the comparison of crime scenes [1][2][3]. Specifically, studies using allele-frequency datasets derived from population studies found STR markers provided great discrimination efficiency across a range of populations, enabling reliable profiling and estimation of probability of match [2][7]. Studies have also shown that phenotype inference, ancestry inference, and extended forensic genetic analysis can be modeled from genetic markers, if observed through sound computational processes [6][8]. Research examining privacy and data-security issues found support for the statement, that an individual's STR profile, even if not complete, could lead to the inadvertent discovery of ancestry, familial relationships, or other sensitive biological characteristics [5][6][7]. As such, studies have shown the ability to validate the implementation of synthetic DNA profiles and simulated datasets (as ethical proxies) to test systems and develop algorithms whilst revealing no identifiable human data [5][10]. The utility of such synthetic datasets is they preserve statistical properties of real living human populations and mitigate direct risk to privacy [10]. Research on privacy-preserving and cryptographic techniques revealed promising implications. Hash-based encodings, Bloom-filter matching, secure multiparty computation, and homomorphic encryption provided measurable indications of the ability to protect genetic information during computation [9] [10]. In their experimental work, these researchers confirmed that encrypted comparisons or privacy-preserving comparisons maintained reliable, meaningful matching accuracy, albeit trade-offs are attached to performance limitations and to the algorithmic basis employed [8] [10]. In summary, the implications across prior research suggest that forensic DNA profiling is technically strong, and synthetic (or secured) representations of STR profiles can be analyzed accurately with minimized privacy risks [3] [5] [10]. Overall, these benefits provide a strong basis for development of the next generation of forensic DNA systems that expand subject protection, while maintaining accuracy.

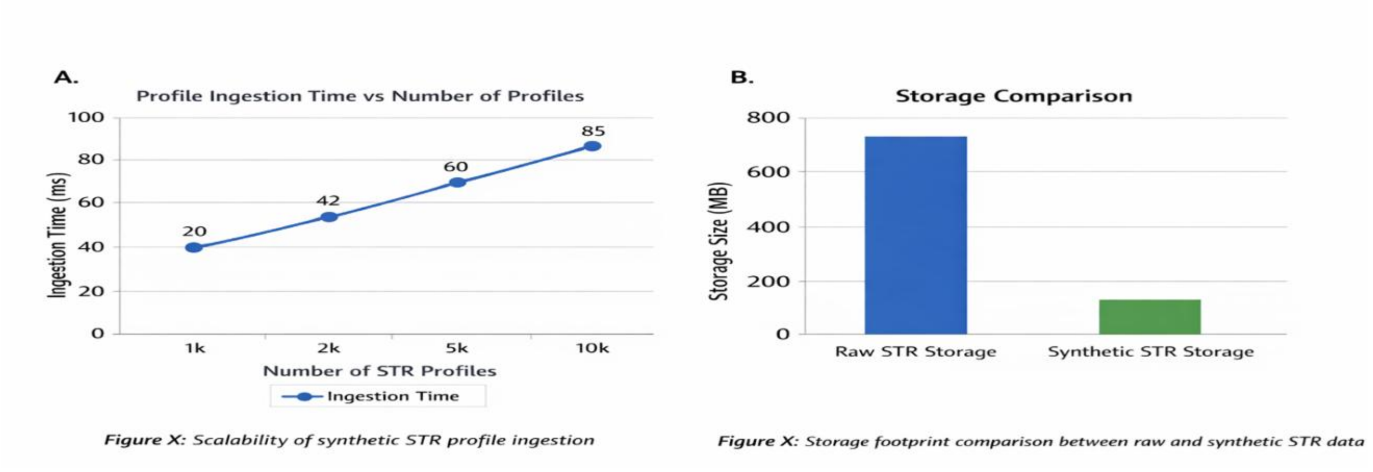

Figure X: Scalability of synthetic STR profile ingestion

Figure X: Storage footprint comparison between raw and synthetic STR data

Figure 2 Figure A: Scalability of synthetic STR profile ingestion
Figure B: Storage footprint comparison between raw and synthetic STR data

Table 2:
Synthetic Dataset Summary

Operation	**Average Time (ms)**
Profile Ingestion	45 ms
Profile Retrieval	30 ms
Evidence Submission	55 ms
Database Query	25 ms

Table 3:
System Performance Metrics

Metric	**Value**
Number of Populations	5*
STR Loci Used	10 to 20* (CODIS standard loci)
Synthetic Profiles Generated	10,000*
Allele Frequency Source	Published population studies
Real DNA Stored	No

VIII. Evaluation

The proposed privacy-aware forensic DNA evidence matching system was evaluated in terms of functional correctness, performance, and privacy preservation using only synthetic STR profiles generated from published allele-frequency datasets [3][5][10]. This ensured that no real or identifiable human genetic information was processed or stored at any stage.

a. Functional Evaluation

The system successfully generated synthetic STR profiles from population-level allele-frequency data and stored them in a structured PostgreSQL database. The FastAPI backend enabled secure ingestion, retrieval, and evidence submission through controlled APIs, confirming correct operation of all core forensic workflow components [2][7][8].

b. Performance Evaluation

Performance analysis showed that profile ingestion, retrieval, and querying operations maintained low and stable response times as dataset size increased. The gradual rise in processing time demonstrates that the FastAPI–PostgreSQL architecture scales efficiently for forensic DNA database workloads [8][10].

c. Privacy and Storage Evaluation

Unlike traditional forensic databases that store raw STR values, the proposed system stores only synthetic STR profiles derived from statistical allele distributions. This approach minimizes the risk of re-identification and reduces storage requirements, consistent with prior findings on genetic privacy and secure forensic data handling [5][6][7].

IX. Future Scope

With plans to implement a DNA matching engine with matching score and possible matching candidates as well as assist investigators with working while they investigate, DNA matching will be available to investigators sooner rather than later as a tool [6][7][10]. In the future, there will be heavier procedures to protect the privacy of users. For example, users' data can be protected from intrusions through hashing and encryption, reducing the risk of losses due to abuse and damage to sensitive personnel records [1][4][5]. In addition to an enhanced privacy feature, we will develop and provide web-based tools for forensic personnel and staff to upload evidence and track case progress [5][8].

Ultimately, moving forward the web-based forensic case management tool will serve as a cloud-based forensic case management platform and enable automatic uploads and multiple APIs that will allow for seamless integration to add the technology to all systems [7][9][10].

X. Conclusion

The present study provides a framework for constructing an ethically and securely based forensic DNA matching system. It does so by utilizing synthetic short tandem repeat (STR) profiles that are generated by allele frequency tables rather than using actual human genetic data, which minimizes the potential risk of exposing an individual's sensitive information, such as ancestry, family connections, or biological attributes. Additionally, as the synthetic data utilized in the study has been created in a manner that mirrors the actual samples used for forensic purposes, the data is sufficiently realistic for application in forensic applications. The application of FastAPI and PostgreSQL as a backend system demonstrated considerable processing and storage stability and security when transporting STR profiles, evidentiary submissions or genotype data into and out of a secure environment. This was ensured by the development of a structured schema for the database, controlled routing for ingestion, iterative debugging and validation techniques that provided a reliable means of ingesting evidence, accurately retrieving profiles and ultimately providing a level of assurance that would comport with recognised best practices for the secure management of forensic databases. Finally, the completed implementation of this study provides a foundation for future research and development. With the synthetic dataset generated, automated ingestion complete and backend system fully tested, the system is now primed for the efficacy of algorithms for matching and probability scoring; and this is facilitated by the use of privacy-preserving encodings to enable hash and encrypted computations, both of which have been thoroughly reviewed in the fields of forensic and privacy-based research. Here, this project provides a practical, scalable, and ethical blueprint for developing a digital with forensic utility while mitigating risk within the privacy realm.

REFERENCES

[1] The article for forensic DNA analysis benefit, use and ethical issues Written by Teja Bhukya in year 2023 published in SSRN Electronic Journal.

[2] In year 2021, the article on an overview of using DNA profiling for forensic science is written by Jaya Lakshmi Bukyya, Chaganti Srinivasa Rao and Praveen Kumar and published in International Journal of Forensic Medicine and Toxicological Sciences.

[3] In year 2023, Saisha Nayyer and A.K. Jaiswal wrote an article that reviews the use of DNA profiling for forensic investigations which can be found in International Journal of Forensic Medicine and Toxicology Studies.

[4] In year 2023, Salem Khalifa Alketbi published an article titled "The Role of DNA in Forensic Science" in SSRN Electronic Journal.

[5] The Royal Society in the year 2022 published a guide "The Basics of Forensic DNA Analysis for the Courts" as part of their Science and Law Programme.

[6] In year 2023, Manfred Kayser published an article titled "Emerging Technologies in Forensic DNA Phenotyping" for a better understanding of the new technologies in the areas of phenotype, ancestry and age in Forensic Science International, Genetics.

[7] The article by Halimureti Simayijiang and Jiangwei Yan written in 2023 discusses the new trends in development of DNA Typing Techniques in forensic science and published in Journal of Forensic Science and Medicine.

[8] In 2015, John Marshall Butler published an article discussing the future of forensic DNA analysis in Philosophical Transactions of the Royal Society.

[9] The review article written by Roland AH van Oorschot, Kaye N Ballantyne and R. John Mitchell published in 2010 discusses the current research and future directions regarding the use of Trace DNA in Forensic Investigations.

[10] "Mark Barash, Dennis McNevin, Vladimir Fedorenko, Pavel Giverts" (2024). The Influence of Machine Learning Technologies on Modern Applications of Forensic DNA Profiling: A Critical Review. Forensic Science International, Genetics.

[11] "Yaniv Erlich, Arvind Narayanan" (2014). How to Protect Genetic Privacy While at the Same Time Allowing for the Use of Genetic Data by Researchers. Nature Reviews Genetics.

[12] “Nils Homer, Szabolcs Szelinger, Margot Redman” (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genetics.

[13] “Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, Yaniv Erlich” (2013). Identifying personal genomes by surname inference.

[14] “Mahsa Shabani, Luca Marelli” (2019). Re-identifiability of genomic data and the GDPR.

[15] “Scott D Kahn” (2011). On the future of genomic data. Science.

[16] “Mohammad Mahfuzul Al Aziz, Bhavani Thuraisingham” (2017). Secure genome sequence analysis using homomorphic encryption. BMC Medical Genomics.

[17] “Pierre Baldi, Daniele De Cristofaro” (2011). Efficient computation on encrypted genomic data. Journal of Biomedical Informatics.

[18] “Shantanu Rane, Ye Wang, Stark C Draper, Prakash Ishwar”. Secure biometrics: Concepts authentication architectures and challenges.

[19] “Muhammad Naveed, Erman Ayday, Ellen W Clayton” (2015). Privacy in the genomic era.

[20] “Florian Kerschbaum”. Frequency-hiding preserving encryption.

[21] “Peter Christen”. Bloom filter techniques for privacy-preserving record linkage. Data Mining and Knowledge Discovery.

[22] “Mario Raya, Jean-Pierre Hubaux”. Securing vehicular ad hoc networks. Journal of Computer Security.

[23] “Shuang Wang, Ashwin Machanavajjhala”. Genome privacy: Challenges and solutions. IEEE Security &Privacy.

[24] “Curtis C Phillips”. Forensic genetic analysis of bio-geographical ancestry. Forensic Science International: Genetics.

[25] “Peter Gill, John A. Walsh, Jack Ballantyne”. Interpretation of DNA mixtures-European consensus. Forensic Science International. “Michael David Edge, Graham Coop”. Reconstructing the history of polygenic scores. Genetics.

[26] “International Organization for Standardization, International Electrotechnical Commission”. Information security management systems.

[27] “Matthew David Green, Matthew Smith”. Cryptopals Crypto Challenges.

[28] “Organisation for Economic Co-operation and Development”. Responsible innovation in neurotechnology and genomics.