Original article

Zenobia: CODIS 13 STR Loci Allele Detection Tool

Osamah Alrouwab [1,2]*, Esraa Algblawi[2], Moudah Kareem[2], Sabreen Aboujildah[2], Saifedden Ayad [3], Mahmoud Gargotti [4]

[1]Department of Biochemistry, Faculty of Medicine, Sbratha University, Sbratha- Libya
[2]Faculty of Biotechnology, Aljafra University, Alsahla -Libya
[3]Department of Preventive Medicine, Faculty of Veterinary Medicine, Al-Zaytoonah University, Tarhuna- Libya
[4]Department of Microbiology, Faculty of Medicine, University of Zawia, Zawia -Libya

ARTICLE INFO

Corresponding Email. usamaerawab@gmail.com
Received: *01-03-2022* ***Accepted***: *12-03-2022* ***Published***: *13-03-2022*
Keywords: *Short Tandem Repeats, Java, Combined DNA Index System, Command Line, Forensics.*

ABSTRACT

Background and aims. *Short tandem repeats (STRs) are one of the most mutable regions in the human genome. They comprise tandem repeating DNA sequences ranging in length from two to six base pairs. Owing to their significant mutation rate, they exhibit considerable variation in pattern among populations and the capacity to be passed on from generation to generation. These loci are broadly employed in medicine, biology, and criminal investigation. They are pivotal in the genesis of a variety of genetic illnesses and have been intensively investigated in forensics, population genetics, and genetic genealogy. Although many implementations that manage STR loci are offered, most of them rely primarily on the command-line interface (CLI) inputs, which frequently necessitate the implementation of tools carried out in various scripting languages. Installing and launching programs through the command line (CL) is time-consuming and/or unprofitable for many students and scholars. The fundamental intention of this project is to develop a cross-platform graphical user interface (GUI) package directed at the Combined DNA Index System (CODIS) STR analysis. Zenobia is a Java-based application considered a step in consistently making CL-only programs available to more apprentices and researchers.* ***Methods.*** *Zenobia core dataset imported from STRBase, a public dataset provided by the National Institute of Standards and Technology. Only CODIS 13 STR markers data were elected. Zenobia uses the brute force approach to match recorded allele patterns in order to discover locus names and allele counts.* ***Results.*** *A set of 78 alleles took part in the trial, with 61.5% representing a simple STR subcategory. Additionally, 30.1% and 7.7% of nominees, respectively, were associated with the compound and complex STR subcategories.* ***Conclusion.*** *In general, Zenobia's application outcomes satisfy the evaluation metrics for efficiency and time consumption. However, more genetic markers should be introduced to increase the productivity of the application.*

Cite this article. *Alrouwab O, Algblawi E, Kareem M, Aboujildah S, Ayad S, Gargotti M. Zenobia: CODIS 13 STR Loci Allele Detection Tool. Alq J Med App Sci. 2022;5(1):159-165.* https://doi.org/10.5281/zenodo.6350591

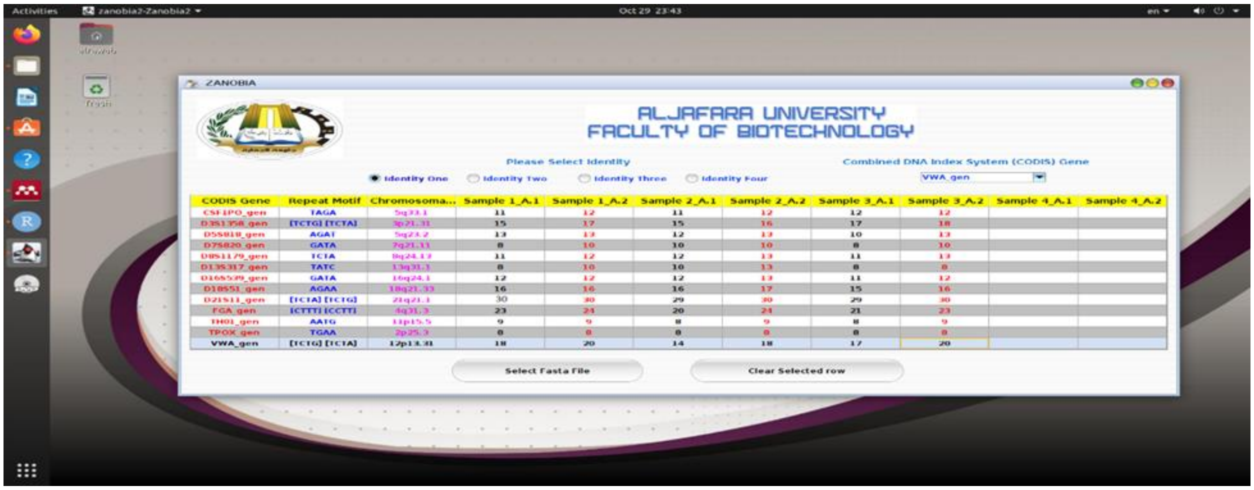
INTRODUCTION

Revolutionary, genetic fingerprinting is one of the emerging technologies that has drastically influenced the realm of forensic medicine and has profoundly altered forensic evidence forever [1]. DNA fingerprinting (DNA profiling or forensic genetics are synonyms also used to designate the same methodology) provides a comparative analysis of DNA to solve legal problems that include paternity tests, the identification of individuality in criminal proceedings in which biological evidence is discovered at crime scenes, and distinguishing the victims of major disasters from the remains [2,3]. Historically, in the mid-eighties of the previous century, a research team from the University of Leicester in the United Kingdom, led by Sir Alec Jeffrey, the father of DNA fingerprinting, inaugurated the age of DNA use in forensic evidence [4]. The microsatellite or Short Tandem Repeats (STRs) markers have been the most extensively used approach for detecting DNA profiles [5]. They are ubiquitous throughout the DNA and reside on average 6-10 kb apart [6,7]. Attributed to their density, polymorphism, and PCR amplification, STRs were measured as reliable biomarkers for genomic mapping and genetic linkage assessment [8,9]. DNA profiling based on STR PCR amplification has the benefit of being more responsive than traditional methods. In addition, their negligible allele size (typically <300 bp) makes the STR system more likely to succeed with older or poorly preserved samples containing only degraded DNA [10–12]. It has been over four decades since the FBI Laboratory selected thirteen STR genetic markers for what is now known as the Combined DNA Index System (CODIS) [13–15]. The CODIS loci used in the US are TPOX, VWA, D3S1358, CSF1PO, FGA, TH01, D13S317, D16S539, D18S51, D5S818, D7S820, D8S1179, and D21S11 [16]. These loci have become the conventional coinage of information exchange for verifying human identity for both judicial case studies and paternity testing due to their accessibility and utilization in the form of commercial STR kits [17,18]. Addressing profile sequence data is a struggle for many students and researchers [19]. Despite, a wide range of programs capable of analyzing STR loci being available, all of them rely on the command-line interface (CLI) commands or are not specifically directed at DNA markers used in forensic investigations. Moreover, they often rely on a set of complementary tools that are implemented in various script languages [20–23]. Some legacy applications for finding tandem repeats within a sequence include: Mreps, demonstrated by Kolpakov Roman and Gregory Kucherov (2003), it’s a sophisticated software for detecting tandem repeated structures in DNA sequences. Mreps could indeed detect all sorts of tandem repeats in a single run on an entire genomic sequence. It has a resolution setting that enables the software to detect 'fuzzy' repetitions [24]. Marco Pellegrini and Alessio Vecchio (2012) developed TRStalker, an algorithm (christened TRStalker) with the intent of discovering Tandem Repeats (TRs) that are hard to identify, owing to their characteristic fuzziness, which is attributed to the high rates of base substitutions, insertions, and deletions [25]. In 2010, Pokrzywa, Rafal, and Andrzej Polanski introduced the Burrows–Wheeler Tandem Repeat Searcher (BWtrs). It is an online web-based utility that scans for specific instances of tandem repeats in DNA sequences, BWtrs adopts the block-sorting compression algorithm [26]. In this paper, we intend to provide a novel tool capable of detecting and determining the numbers of alleles of CODIS loci stored in a plain text FASTA format.

MATERIALS AND METHODS

Zenobia v1.0

Zenobia (Figure 1) is a Java-based graphical user interface (GUI) tool, for CODIS 13 Alleles detection released under the GNU General Public License. The source code is freely available on GitHub (https://github.com/alrawab/zenobia).



Please Select Identity

● Identity One ○ Identity Two ○ Identity Three ○ Identity Four

Combined DNA Index System (CODIS) Gene: VWA_gen

CODIS Gene	Repeat Motif	Chromosoma...	Sample 1_A.1	Sample 1_A.2	Sample 2_A.1	Sample 2_A.2	Sample 3_A.1	Sample 3_A.2	Sample 4_A.1	Sample 4_A.2
CSF1PO_gen	TAGA	5q33.1	11	12	11	12	12	12		
D3S1358_gen	[TCTG] [TCTA]	3p21.31	15	17	15	16	17	18		
D5S818_gen	AGAT	5q23.2	13	13	12	13	10	13		
D7S820_gen	GATA	7q21.11	8	10	10	10	8	10		
D8S1179_gen	TCTA	8q24.13	11	12	12	13	11	13		
D13S317_gen	TATC	13q31.1	8	10	10	13	8	8		
D16S539_gen	GATA	16q24.1	12	12	12	13	11	12		
D18S51_gen	AGAA	18q21.33	16	16	16	17	15	16		
D21S11_gen	[TCTA] [TCTG]	21q21.1	30	30	29	30	29	30		
FGA_gen	[CTTT] [CCTT]	4q31.3	23	24	20	24	21	23		
TH01_gen	AATG	11p15.5	9	9	8	9	8	9		
TPOX_gen	TGAA	2p25.3	8	8	8	8	8	8		
VWA_gen	[TCTG] [TCTA]	12p13.31	18	20	14	18	17	20		

Figure 1: Zenobia CODIS alleles detector

Dataset

Zenobia core dataset imported from STRBase, a public dataset provided by the National Institute of Standards and Technology (NIST; https://www.nist.gov) during September 2021. Only CODIS 13 STR markers data were chosen, namely, CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11 (Table 1).

Table 1: Common STR loci

Locus	*Repeat motif*	*Repeat category*	*Chromosome location*	*Allele range*
CSF1PO	AGAT	Simple	5q33.1	5-16
FGA	[CTTT] [CCTT]	Compound	4q31.3	12.2-51.2
TH01	TCAT	Simple	11p15.5	3-14
TPOX	TGAA	Simple	2p25.3	4-16
VWA	[TCTG] [TCTA]	Compound	12p13.31	10-25
D3S1358	[TCTG] [TCTA]	Compound	3p21.31	8-20
D5S818	AGAT	Simple	5q23.2	6-18
D7S820	GATA	Simple	7q21.11	5-16
D8S1179	[TCTA] [TCTG]	Compound	8q24.13	7-20
D13S317	TATC	Simple	13q31.1	5-17
D16S539	GATA	Simple	16q24.1	4-16
D18S51	AGAA	Simple	18q21.33	7-39.2
D21S11	[TCTA] [TCTG]	Complex	21q21.1	12-41.2

Case scenario and input files

A paternity dispute case based on matches of the alleles at the CODIS 13 STR loci between a child and mother and alleged father (trio cases), from the Arab Republic of Egypt in 2012 (Table 2), documented by Mr. Sherif H. El-Alfy, used to simulate and construct dummy profile files [27].

Table 2: Typing results of 13 autosomal STR loci analysis

STR locus	*Child*	*Mother*	*Alleged father*
D3S1358	15,17	15,16	17,18
D5S818	13,13	12, 13	10, 13
D7S820	8,10	10,10	8,10
D8S1179	11,12	12,13	11,13
D13S317	8, 10	10, 13	8,8
D16S539	12, 12	12, 13	11,12
D18S51	16, 16	16,17	15,16
D21S11	30,30	29,30	29,30
FGA	23,24	20,24	21,23
TH01	9,9	8,9	8,9
TPOX	8,8	8,8	8,8
VWA	18,20	14,18	17,20
CSF1PO	11, 12	11,12	12, 12

Since the program support only a plain text Fasta file format. To evaluate the performance of the tool we generate dummy files that contain random sequences with real allelic variation sequences imported from the Entrez database (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) provided by The National Center for Biotechnology Information (NCBI) for locus-specific information (Table 3).

Table 3: Allele number and Accession used to evaluate the performance

	Child				Mother				Alleged father			
Locus	Allele one		Allele two		Allele one		Allele two		Allele one		Allele two	
	Allele no.	Accession	Allele no.	Accession	Allele no.	Accession	Allele no.	Accession	Allele no.	Accession	Allele no.	Accession
D3S1358	15	MW218622	17	MK990349	15	MW218622	16	MH166976	17	MK990349	18	MK990350
D5S818	13	MZ325899	13	MZ325899	12	MH167008	13	MZ325899	10	MH166998	13	MZ325899
D7S820	8	MH167026	10	MZ325981	10	MZ325981	10	MZ325981	8	MH167026	10	MZ325981
D8S1179	11	MH105190	12	MH105195	12	MH105195	13	MH105201	11	MH105190	13	MH105201
D13S317	8	MZ325902	10	MK295189	10	MK295189	13	MT298696	8	MZ325902	8	MZ325902
D16S539	12	MT298697	12	MT298697	12	MT298697	13	MW218608	11	MH167254	12	MT298697
D18S51	16	MW218634	16	MW218634	16	MW218634	17	MK569958	15	MW218632	16	MW218634
D21S11	30	MZ325991	30	MZ325991	29	MZ325935	30	MZ325991	29	MZ325935	30	MZ325991
FGA	23	MZ325919	24	MH232622	20	MH232609	24	MH232622	21	MH232611	23	MZ325919
TH01	9	MH085123	9	MH085123	8	MW218611	9	MH085123	8	MW218611	9	MH085123
TPOX	8	MG988075	8	MG988075	8	MG988075	8	MG988075	8	MG988075	8	MG988075
VWA	18	MW218658	20	MH167102	14	MH167077	18	MW218658	17	MW218657	20	MH167102
CSF1PO	11	MH085186	12	MN983119	11	MH085186	12	MN983119	12	MN983119	12	MN983119

Implementation

The benchmarks were carried out on personal computers with intel core i5-3470,3.20 GHz,16.00 GB of RAM, Linux Ubuntu-20.04.3 64 bits. Zenobia was written in Java programing language using Oracle Java SE Development Kit 11 (https://www.oracle.com/java), and Apache NetBeans IDE 12.1 (http://netbeans.apache.org). Detection of the allelic type for each STR gene. Zenobia recruits the so-called *brute force* algorithm to match stored allele patterns to detect locus names and allele numbers (Figure 2).

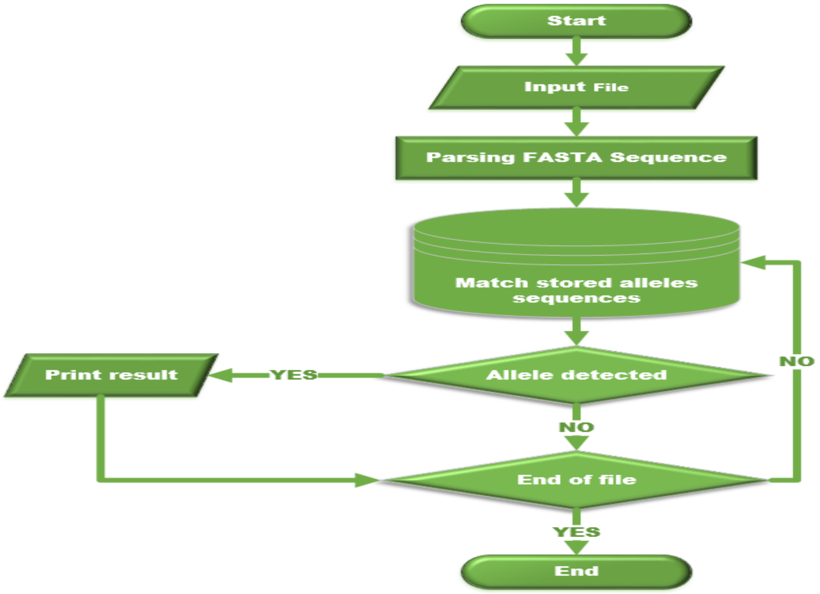


Figure 2: Zenobia workflow flowchart.

RESULTS

A total of 78 alleles participated in the experiment (Table 2), 61.5 % of whom are representatives of a simple STR subgroup. Furthermore, 30.1% and 7.7% of candidates engaged with compound and complex STR subgroup correspondingly. The child profile's allele numbers fluctuated from 11 to 30, the mother profile's allele ranged from 8 to 29, while the alleged father allele numbers spanned from 8 to 30.

The observed genotype for child profile was, D3S1358 (15,17), D5S818 (13,13), D7S820 (8,10), D8S1179 (11,12), D13S317 (8,10), D16S539 (12,12), D18S51 (16,16), D21S11 (30,30), FGA (23,24), TH01 (9,9), TPOX (8,8), VWA (18,20), CSF1PO (11,12). While mother shows, D3S1358 (15,16), D5S818 (12,13), D7S820 (10,10), D8S1179 (12,13), D13S317 (10,13), D16S539 (12,13), D18S51 (16,17), D21S11 (29,30), FGA (20,24), TH01 (8,9), TPOX (8,8), VWA (14,18), CSF1PO (11,12). Finally, the alleged father records, D3S1358 (17,18), D5S818 (10,13), D7S820 (8,10), D8S1179 (11,13), D13S317 (8,8), D16S539 (11,12), D18S51 (15,16), D21S11 (29,30), FGA (21,23), TH01 (8,9), TPOX (8,8), VWA (17,20), CSF1PO (12,12).

DISCUSSION

The purpose of this study was to develop a multi-platform, user-friendly, and open-source CODIS 13 STRs allele detector. Many methods for locating short tandem repeats over DNA sequences have been developed in response to their relevance in understanding STR loci [28]. Some tools are out of date, and a handful of them are no longer accessible [29]. There are, however, several programs available that operate either on the command line or as standalone web services. In this section, different tools will be surveyed for their capabilities to detect STR loci. TAREAN, a command-line, computational approach for automatically detecting satellite repeats in unassembled next-generation sequencing (NGS) sequences. Introduced by Novák, Petr, et al. (2017), TAREAN is built with customized Python and R packages, to discover new satellite repeats, which were then confirmed on metaphase chromosomes using FISH with probes generated based on reconstructed monomer sequences [30]. STRetch, a command-line tool written as python scripts directed to the analysis of STRs from whole-genome-sequencing (WGS) results, was developed by Harriet, et al. (2018). TRetch seems to have a low false discovery rate (FDR) for deleterious STR expansions related to Mendelian disorder, it is designed for STR linked to genetic disorders [31]. TandemTools, a python-based tool developed by Mikheenko, Alla, et al. (2020) detected Extra-long tandem repeats (ETRs) [32].

Contrarily, in comparison to other comparable programs, Zenobia adopts an entirely different approach. None of the tandem repeats detecting algorithms were implemented since the program's objective is to determine the allele number associated with each locus, not only the existence or absence of these repeats. This grant Zenobia an edge over other current programs, which are only capable of spotting tandem repetitions.

Zenobia was implemented to identify readings for pre-defined CODIS 13 STR loci. For this aim, 13 distinct classes representing the major positions loci have been constructed, and each of them maintains the dataset of its alleles as described by the National Institute of Standards and Technology (Figure 3).

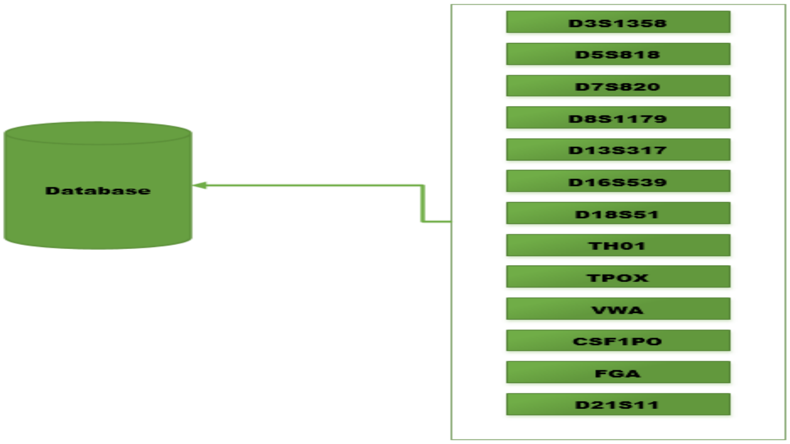


Figure 3: Zenobia alleles dataset

The brute force algorithm was used to achieve a perfect match between the alleles stored in the database, validate their appearance, and identify the precise number of the corresponding allele. It is regarded as one of the most logical choices for the string pattern-matching challenge. Simply matching the pattern in the target at consecutive positions from left to right is the focus of this method. If the comparison window fails, it shifts one letter to the right until the end of the target sequence is attained. Despite the algorithm's poor theoretical performance, our measurements show that it is one of the fastest techniques when the pattern is a short sequence.

LIMITATION

Zenobia supports only one type of file format, the so-called FASTA. Furthermore, the stored datasets do not only contain complementary sequences of the alleles.

CONCLUSIONS

We designed a Bioinformatics application using JAVA language version 11. It enables us in interpreting FASTA files, identify CODIS 13 loci, and determine the allelic number from a nucleotide sequence. Zenobia has done an excellent job at applying the boundary values in terms of precision and time consumption. When reading the 78 allelic profiles, no faults were encountered. However, additional STR loci are still required to be added.

Disclaimer

The article has not been previously presented or published, and is not part of a thesis project.

Conflict of Interest

There are no financial, personal, or professional conflicts of interest to declare.

REFERENCES

1. Hilbert J. The disappointing history of science in the courtroom: Frye, Daubert, and the ongoing crisis of junk science in criminal trials. Okla L Rev. 2018;71:759.
2. Nurse KRD. Forensic Experts' Best Practices in DNA Collection, Analysis and Testimony: A Delphi Study. The University of the Rockies; 2018;1:24.
3. Gang A, Shrivastav VK. Single-Nucleotide Polymorphism: A Forensic Perspective. Handb DNA Profiling. 2020;1–22.
4. Bright J-A, Kelly H, Kerr Z, McGovern C, Taylor D, Buckleton JS. The interpretation of forensic DNA profiles: an historical perspective. J R Soc New Zeal. 2020;50(2):211–25.
5. Tao R, Wang S, Zhang J, Zhang J, Yang Z, Sheng X, et al. Separation/extraction, detection, and interpretation of DNA mixtures in forensic science. Int J Legal Med. 2018;132(5):1247–61.
6. Francastel C, Magdinier F. DNA methylation in satellite repeats disorders. Essays Biochem. 2019;63(6):757–71.
7. Xu G, Lyu J, Li Q, Liu H, Wang D, Zhang M, et al. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. Nat Commun. 2020;11(1):1–12.
8. Weymaere J, Vander Plaetsen A-S, Tilleman L, Tytgat O, Rubben K, Geeraert S, et al. Kinship analysis on single cells after whole genome amplification. Sci Rep. 2020;10(1):1–9.
9. Al-Qahtani WS, Al-Hazani TM, Safhi FA, Alotaibi MA, Domiaty DM, Al-Shamrani SMS, et al. Assessment of Metastatic Colorectal Cancer (CRC) Tissues for Interpreting Genetic Data in Forensic Science by Applying 16 STR Loci among Saudi Patients. Asian Pacific J Cancer Prev. 2021;22(9):2797–806.
10. Lynch C, Fleming R. A review of direct polymerase chain reaction of DNA and RNA for forensic purposes. Wiley Interdiscip Rev Forensic Sci. 2019;1(4):e1335.
11. Nigam K, Srivastava A, Sahoo S, Dubey IP, Tripathi IP, Shrivastava P. Sequential Advancements of DNA Profiling: An Overview of Complete Arena. Forensic DNA Typing Princ Appl Adv. 2020;45–68.
12. Thakar M, Joshi B, Shrivastava P. Usefulness of Mini-STRs in Analyzing Degraded DNA Samples and Their Forensic Relevance. In: Forensic DNA Typing: Principles, Applications and Advancements. Springer; 2020. 205–22.
13. Katsanis SH. Pedigrees and perpetrators: Uses of DNA and genealogy in forensic investigations. Annu Rev Genomics Hum Genet. 2020;21:535–64.
14. Kitnick J. Killer's Code: Familial DNA Searches Through Third-Party Databases under Carpenter. Cardozo L Rev. 2019;41:855.
15. Neuvonen A. Finnish population genetics in a forensic context. 2017;1:89.
16. Huang X. Short Tandem Repeat Profiles in Ovarian Carcinoma Cells During Primary Culture. 2019;1:54. https://macau.uni-kiel.de/receive/diss_mods_00026213

17. Kaushik S, Sahajpal V. Capillary Electrophoresis Issues in Forensic DNA Typing. In: Forensic DNA Typing: Principles, Applications and Advancements. Springer; 2020. 223–38.
18. de Groot NF, van Beers BC, Meynen G. Commercial DNA tests and police investigations: a broad bioethical perspective. J Med Ethics. 2021;788:795.
19. Meng T, Soliman AT, Shyu M-L, Yang Y, Chen S-C, Iyengar SS, et al. Wavelet analysis in current cancer genome research: a survey. IEEE/ACM Trans Comput Biol Bioinforma. 2013;10(6):1442–14359.
20. Christopher FE, Myers KJ. Siem-Enabled Cyber Event Correlation (What And How). NAVAL POSTGRADUATE SCHOOL MONTEREY CA; 2018;1:111.
21. Codó Tarraubella L. Computational Infrastructures for biomolecular research. 2019;1:958.
22. Boattini A, Sarno S, Mazzarisi AM, Viroli C, De Fanti S, Bini C, et al. estimating Y-str Mutation Rates and tmrca through Deep-Rooting Italian pedigrees. Sci Rep. 2019;9(1):1–12.
23. Alrouwab OS, Gargotti M. Evaluating Efficiency of Some Exact String-Matching Algorithms on Large-Scale Genome. Am J Comput Sci Inf Technol. 2021;9(112):8.
24. Kolpakov R, Bana G, Kucherov G. mreps: Efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res. 2003 Jul;31(13):3672–8.
25. Pellegrini M, Renda ME, Vecchio A. Tandem repeats discovery service (TReaDS) applied to finding novel cis-acting factors in repeat expansion diseases. BMC Bioinformatics. 2012;13(4):1–15.
26. Pokrzywa R, Polanski A. BWtrs: a tool for searching for tandem repeats in DNA sequences based on the Burrows–Wheeler transform. Genomics. 2010;96(5):316–21.
27. El-Alfy S, El-Hafez A. Paternity testing and forensic DNA typing by multiplex STR analysis using ABI PRISM 310 Genetic Analyzer. J Genet Eng Biotechnol. 2012 Jun 1;10:101–112.
28. Halman A. Advancing the detection of short tandem repeats in health and disease. 2021;1:207.
29. Parisi V, De Fonzo V, Aluffi-Pentini F. STRING: finding tandem repeats in DNA sequences. Bioinformatics. 2003 Sep;19(14):1733–8.
30. Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res. 2017;45(12):e111–e111.
31. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol. 2018;19(1):1–13.
32. Mikheenko A, Bzikadze A V, Gurevich A, Miga KH, Pevzner PA. TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. Bioinformatics. 2020;36(Supplement_1):i75–83.