NASTRA: Innovative Short Tandem Repeat Analysis through Cluster-Based Structure-Aware Algorithm in Nanopore Sequencing Data

Zilin Ren[1,2,#,*], Jiarong Zhang[3,4,#], Yixiang Zhang[1,2,#], Pingping Sun[2], Jiguo Xue[3], Jiangwei Yan[4,*], Ming Ni[3,*]

1. Changchun Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Changchun 130122, China

2. School of Information Science and Technology, Institution of Computational Biology, Northeast Normal University, Changchun 130117, China

3. Institute of Health Service and Transfusion Medicine, Beijing 100850, People's Republic of China

4. School of Forensic Medicine, Shanxi Medical University, Taiyuan 030001, People's Republic of China

* To whom correspondence should be addressed. Email: Zilin Ren, zilin.ren@outlook.com; Jiangwei Yan, yanjw@sxmu.edu.cn; Ming Ni, niming@bmi.ac.cn.

Abstract

Short-tandem repeats (STRs) are type of genetic markers distinguishing individuals and authenticating cell-lines. Nanopore sequencing is promising in STR typing for its convenience, but lack of analysis method. Here we proposed NASTRA, a tool for accurate STR genotyping with nanopore sequencing, which uses an STR-structure-aware algorithm to infer repeat numbers of STR motifs. In our real-time scenario tesing, NASTRA has 100% accuracies for diploid STRs, far

exceeding method employing strategy that include all candidate STR allele sequences for alignments. NASTRA could be useful in applications as individual identification and cell-line authentication with nanopore sequencing. NASTRA is available via https://github.com/renzilin/NASTRA.

Keywords: nanopore sequencing; short tandem repeat; individual identification; cell line authentication; bioinformatics tools.

Background

Short tandem repeats (STRs) in human genome are specific DNA loci that can have high degree of polymorphisms in the number and forms of repeat motifs among individuals. This feature makes STRs a kind of genetic markers well-suited for individual identification [1,2], cell line authentication [3,4], as well as relationship inferring [5–7]. In Short Tandem Repeat DNA Internet Database maintained by the National Institute of Standards and Technology (NIST) [8], more than 70 STR genetic markers, typically with 19 (D10S1248) -114 (FGA) alleles in human populations, are chosen as genetic markers with desirable features as high heterozygosity, regular repeat unit, distinguishable alleles and robust amplification. These STRs have been utilized for decades by forensic laboratories across the globe and have formed the basis of million- or even 10-million-scale human DNA databases for public security purposes in many countries. In the realm of biomedical research, profiling STR for cell line authentication plays a critical role in ensuring scientific reproducibility and preventing misidentification as well as cross-contamination of cell lines[9,10]. To establish standardized practices for cell line authentication using STRs, the International Cell Line Authentication Committee (ICLAC, https://iclac.org) integrates numerous

42 online databases and search tools, including the Cellosaurus knowledge resource [11], DSMZ STR
43 profile database[12], CLIMA database [13], CLASTER [14].

44 Conventional method for STR genotyping is the use of capillary electrophoresis (CE)-based length
45 analysis. Besides, massively parallel sequencing (MPS) has also been validated for STR profiling,
46 and it provide a larger battery of STRs in a run than the CE-based method. However, the majority
47 of instruments of Sanger sequencing and MPS are heavy (50-186 kg), costly (50k-180k US dollars),
48 and have high environmental requirements as constant temperature and vibration free. Thus,
49 Sanger sequencing and MPS systems are usually maintained by well-equipped laboratories or
50 facilities, and samples are transported for analysis, which might needs days to obtain STR typing
51 reports.

52 A more rapid and affordable solution of STRs profiling thus has significance for relevant research
53 and applications. Nanopore sequencing provides a cheap and real-time sequencing approach.
54 The devices for nanopore sequencing, such as MinION Mk1B/Mk1c (Oxford Nanopore
55 Technologies, Oxford, UK) and the recently released QNome-3841 (Qitan Technology, Chengdu,
56 China), are exceptionally portable and considerably cheaper than NGS and Sanger sequencing
57 systems. Moreover, these portable nanopore sequencing devices are capable of generating much
58 longer reads than Sanger sequencing and NGS systems, and had a throughput in the Gbp range
59 during a single run. Therefore, nanopore sequencing has been applied in various on-site
60 sequencing applications as in-field genomic surveillance of pathogens [15,16] and biodiversity
61 investigations [17,18]. Due to the portability and high throughput, Zaaijer et al.[19] reported an
62 interesting pilot study that utilized MinION for whole genome sequencing to rapidly profile SNPs
63 and authenticate human cell lines for biological researches.

However, despite these advantages, nanopore sequencing data are noisier than these generated by Sanger sequencing and MPS. Unlike the cases of STR expansion disorders causing diseases, in which pathogenic alleles remarkably differ from benign alleles in sizes [20–22], alleles of STR genetic markers for individual or cell-line identity identification are usually of a 100 – 300 nt length and their repeat units are of 3- to 5-nt. Genotypes of STR genetic markers are defined by the exact counts of their repeats. Many STR genetic markers consist of multiple (compound STRs and complex STRs) and units are similar within in one STR and its flanking regions. For instance, the repeat structure of a Combined DNA Index System (CODIS) core STR locus D3S1358 is TCTA$[TCTG]_n[TCTA]_m$, in which *n* and *m* denotes repeat numbers. Moreover, some repeat units of STRs contain or form 3- to 4-nt homopolymers or form homopolymers in repeats, as STR Penta D with a repeat structure of $[AAAGA]_m$, which are particularly error-prone, systematically, in nanopore sequencing [23].

Thus, the improvement of highly accurate typing method for genetic STR markers for nanopore sequencing is challenging [24–27]. To tackle this issue , developing algorithms and workflows specifically tailored for typing of genetic STR markers, considering repeat structures of specific STR loci, might be feasible strategy. In our previous study [23], we exhibited that conventional genotyping tools for nanopore sequencing as repeatHMM [21] are not suitable, and a simple customized workflow that integrated the classic Smith-Waterman algorithm for repeat unit re-alignment provided a significant improvement in the accuracy of STR typing. More than 32 STRs could be consistently and correctly typed in a validation set involving 31 individuals, and it is a prove-of-concept study of employing nanopore sequencing in human identity testing and cell-line authentication. Recently, Hall et al. [27] and Tytgat et al. [24] and the very recently

Jidong et al [28], used a different and straight strategy that includes all available sequences of alleles for alignment of nanopore sequencing reads, and to select the alleles with sequencing reads for STR genotyping. In their validation sets, they report a higher STR typing accuracies than our previous method, and it further indicates the promising application of nanopore sequencing in this field. However, this strategy highly depends on the available sequences, many of which are unknown for these highly diverse STRs and might lead to error typing. On the other side, larger validation sample sets are also needed to example validate the STR typing methods.

We aim to develop a method for accurate STR genotyping in applications as individual identification and cell-line authentication. We expect that the method can fully consider the characteristics of these STR markers and have high flexibility, i.e., independent of allele sequences databases that lack allele information. In this study, we propose NASTRA, a tool with clustering-based structure-aware algorithm , rather than allele-dependent alignment, for inference of repeat structure for STR genotyping. We includes a validation set involving 76 individuals and 8 cell-lines, and tested with different chemistries (R9 and R10 flow cells) as well as STR amplification methods prior to sequencing. Meanwhile, the validation based on allele-dependent methods were conducted and NASTRA turned out to have better performance in both accuracy and sequencing data requirement.

Results

In this study, we firstly illustrated the overview of NASTRA. And then demonstrated how to find the optimal thresholds: supporting reads number (SN) and the ratio of supporting reads number (SNR), which is pivotal for STR genotype calling. Following this, we conducted an extensive

107 evaluation of NASTRA, comparing it with a typical alignment-based method named STRspy and
108 investigating the characteristics of genotyping errors. We then conducted a real-world scenario
109 testing using eight human cell lines. Lastly, we summarized the runtime of NASTRA. Throughout
110 these processes, we leveraged the MiSeq FGx™ Forensic Genomics System to carry out STR
111 genotype calling for ground truth construction. Specifically, we amplified 54 STR loci and 46 STR
112 loci across 88 DNA samples using Verogen's ForenSeq Mix A kit and Promega's PowerSeq 46GY kit
113 and sequenced amplicons with R9.4.1 and R10.3 flowcells (Figure 1a).

114 Overview of NASTRA

115 For STR genotyping, NASTRA requires BAM files, BED files containing the positional information
116 of STRs, and fact sheets containing the repetitive structure details of the STRs. In addition,
117 NASTRA performs quality control and determines homozygosity/heterozygosity on genotyped
118 alleles based on the SN of alleles and the SNR between different alleles. Drawing inspiration from
119 the ForenSeq Universal Analysis Software (UAS), we incorporated a similar mechanism to filter
120 loci with low coverage and stutters: if the total number of reads covering the STR locus is below
121 10, NASTRA doesn't output any result due to insufficient coverage; if the SN value of the major
122 allele does not exceed the threshold, the result is classified as "interpretation"; if the SN value of
123 the minor allele does not surpass the threshold but the SNR value does, the result is considered
124 as "imbalance". Regarding the optimal selection of SN and SNR, we provided the corresponding
125 configuration file for users (the relevant analysis is described in *Parameter optimization* section).

126 NASTRA comprise two main sub-algorithms, read clustering and repeat structure inference,
127 which mitigates the potential impact of subtle sequencing errors on accurate genotyping and

128 genotypes STR without the need for allele reference database (Figure 1b). In step 1, NASTRA
129 retrieves aligned reads that span a designated STR locus based on positional information from
130 the BAM file. Following that, in step 2, the prefix and suffix flanking sequences are individually
131 aligned against the extracted reads, employing an affine-gap penalty (methods section).
132 Subsequently, the upstream and downstream sequences of all retrieved reads will be trimmed,
133 retaining only the three adjacent bases flanking the core repeat region of the STR. These trimmed
134 reads are categorized as unique trimmed reads and their frequencies are recorded as SN. Due to
135 the presence of subtle sequencing errors, the trimmed reads could encompass over 100 distinct
136 types. Therefore, NASTRA utilizes pairwise alignment with an affine gap penalty to cluster these
137 reads in step 3. By permitting the exclusion of minor errors, such as single-base insertions, this
138 approach decreases the diversity among reads, thereby facilitating the identification of potential
139 candidate alleles. In step 4, we developed a recursive algorithm to infer repeat structure of allele
140 sequences based on the repeat units present within the STR, which ensures swift acquisition of
141 STR genotypes and aids in promptly identifying the locations of SNV in locus. Finally, NASTRA
142 conducts STR genotyping by leveraging the SN and SNR values associated with alleles.

143 Parameter optimization

144 NASTRA performs STR genotype calling based on the SN and SNR values of the alleles. Therefore,
145 the threshold SN and SNR need to be refined to enhance the robustness of NASTRA. In light of
146 this, we partitioned the samples into two distinct sets: (1) training set, comprising 20 males and
147 20 females, and (2) test set, consisting of 18 males, 18 females, and 4 replicates of 2800M
148 samples. However, two underlying issues arose that could prevent the attainment of optimal
149 threshold values: (1) the limited size of the training set; (2) the insensitivity of the existing

150 sample's deep sequencing depths towards the thresholds. To address these issues, we developed
151 a mini-tool called NanoTime. Leveraging sequencing summary files, we used NanoTime to
152 downsample Nanopore sequencing reads in accordance with sequencing durations, including 1, 2,
153 3, 4, 5, 6, 8, 10, 12, 16, 20, and 24 hours. This process substantially expanded both training set
154 and testing set from 40 samples to 480 samples.

155 To assess the performance, we used allele calls from UAS reports as the ground truth and
156 excluded results from the UAS that were labeled as “imbalanced,” “allele count,” “stutter,” and/or
157 “interpretation threshold”. Then, NASTRA results that failed quality control were excluded from
158 the evaluation. Finally, the genotyping results for each STR obtained through NASTRA can be
159 classified into five classes: (1) **Exact Match**: correct genotyping; (2) **Incomplete Match**: correct
160 genotyping when disregarding the incomplete repeat unit; (3) **One Match**: one of the alleles is
161 correctly genotyped in a heterozygous STR genotype; (4) **Incomplete One Match**: one of the
162 alleles is correctly genotyped in a heterozygous STR genotype, when disregarding the incomplete
163 repeat unit; (5) **Mismatch**: incorrect genotyping.

164 **Thresholds for Autosomal STRs.** A total of 12960 genotype calls for 27 autosomal STR loci on 480
165 samples were obtained. Initially, we explored various SN values (0, 5, 10, 15, 20, 25, 30, 35, 40, 45,
166 and 50) as well as different SNR values (ranging from 0 to 1, in increments of 0.01). For each SN
167 value, we calculated the maximum genotyping accuracy for each autosomal STR locus from
168 expanded training set. The result showed that when SN equals or surpasses 25, the maximum
169 genotyping accuracy for all STR loci can surpass 95% (Figure 2a). After determining an SN value of
170 25, we calculated the genotyping accuracy for each autosomal STR at various SNR values. The
171 median value within the SNR range corresponding to the peak genotyping accuracy is designated

172 as the SNR threshold (Figure 2b). After the determination of the threshold SN and SNR, NASTRA
173 was employed to conduct STR genotype calling on expanded testing set. In summary, 9,428 of
174 12,960 (72.8%) STRs on 480 expanded testing set were genotyped successfully, consisting of
175 8,455 (89.7%) exact match, 642 (6.8%) incomplete match, 307 (3.3%) one match, 11 (0.1%)
176 incomplete match, and 0.1% mismatch. The rest 3,532 genotypes calling failed to pass NASTRA
177 quality control (2,548 calls) or UAS quality control (984 calls), due to the low coverage of short
178 sequencing durations. Figure 3c illustrates the accuracy of each STR across the 480 samples.
179 Among the 27 loci, 18 loci exhibited a calling rate of 1 (either exact match or incomplete match),
180 while 4 loci exceeded 95% accuracy. Additionally, we employed pairwise alignment to investigate
181 sequence identity for each genotyping result (Figure 2d). We found that for the majority of
182 "incomplete match" genotyping results, the sequence identities of allele sequences equal to 1.
183 This suggests that NASTRA primarily disregards the incomplete repeat units solely in terms of
184 repeat counting, while maintaining the true allele sequences.

185 Evaluation of NASTRA on ForenSeq amplicons with R9.4.1
186 flow cells

187 After having determined the optimal parameters for NASTRA, we first assessed the performance
188 between NASTRA and the typical alignment-based method, STRspy [27], on ForenSeq amplicon
189 sequencing data (test set, sequencing duration 24 hours). STRspy is an STR profiling tool tailored
190 for long-read sequencing, which is based on the alignment of sequencing reads with an allele
191 reference library. The authors have created a database containing allele sequences of 22
192 autosomal STRs. And thus, STRspy is unable to perform genotyping calling for 5 out of 27 STR loci
193 in ForenSeq data because of the absence of allele information. The results in a calling of 0 for

194 these loci (D17S1301, D19S433, D20S482, D4S2408, D9S1122, and D6S1043) across all samples.
195 Furthermore, it's crucial to emphasize that if the relevant allele sequence isn't included in the
196 reference database, STRspy may produce incorrect genotyping results due to wrong alignments.

197 We analyzed the calling ratio and genotyping accuracy for ForenSeq data across the test set
198 (Figure 3a-b). In addition, the genotyping results of sexual STRs were summarized.

199 Regarding NASTRA, it achieved a calling ratio of 1 for 7 STR loci, while for 9 STR loci, the calling
200 ratio ranged from 92.1% to 97.4%. Additionally, 3 STR loci had a calling ratio ranging from 71.8%
201 to 80.0%. The calling ratios for PentaD and PentaE are the lowest, at 22.2% and 33.3%
202 respectively. This is attributed to homopolymer sequencing errors causing low SN of candidate
203 alleles. STRspy successfully genotyped for 22 STR loci, achieving a calling ratio of 1. NASTRA
204 demonstrates superior accuracy compared to STRspy. NASTRA achieved a 100% genotyping
205 accuracy across 21 STR loci, including PentaE. Among the remaining 6 loci, the genotyping
206 accuracy for 3 loci (D2S1338, D1S1656, and D6S1043) exceeded 92%. As for STRspy, it exhibited
207 16 loci with a genotyping accuracy beyond 92%, including 12 loci with a accuracy of 100%.

208 Evaluation of NASTRA on PowerSeq amplicons with R10.3
209 flow cells

210 To investigate the performance of NASTRA on different amplification systems and flow cells, we
211 conducted a sequencing experiment on 46 samples using Promega's PowerSeq 46GY kit and
212 R10.3 sequencing flow cells. The PowerSeq kit is a co-amplification system consisting of 22
213 autosomal STR loci and 23 Y-STR loci. The ONT R10.3 flow cell represents a new version featuring
214 new pore design, resulting in enhanced read accuracy and quality. We conducted STR genotyping

215 on the sequencing data using STRspy v1.1.1 and NASTRA with default parameters.

216 The calling ratio and genotyping accuracy for PowerSeq data across 46 samples were shown in
217 Figure 4a and b. Compared to the ForenSeq amplification system, the PowerSeq system does not
218 include five automsomal STR loci (D17S1301, D20S482, D6S1043, and D9S1122). Regarding
219 NASTRA, out of the 22 autosomal STR loci, it successfully called 20 STR loci with a calling ratio of
220 1, while the remaining 2 loci, PentaE (94.7%) and PentaD (55.6%), exhibited slightly lower calling
221 ratios. For STRspy, there is only one locus, D19S433, not being included in the reference database.
222 In terms of genotyping accuracy, we observed a decrease in performance for both STRspy and
223 NASTRA on the PowerSeq dataset, which could be attributed to the overall shorter amplicons in
224 PowerSeq data compared to ForenSeq data. Specifically, NASTRA achieved an accuracy exceeding
225 95% in 18 loci (compared to STRspy's 13 loci), with the remaining 4 loci having accuracy ranging
226 from 80% to 91.5%. While, STRspy achieved accuracy between 61.7% to 91.3% in the remaining 7
227 loci, and there was no result for one locus, D19S433. In general, NASTRA exhibited a slightly
228 better performance than STRspy.

229 Real-time typing of 8 human cell lines

230 To assess NASTRA's performance in real-world scenarios, DNA samples from eight human cell
231 lines were processed for amplification and sequencing using the ForenSeq Signature kit on a
232 MinION platform with an R9.4.1 flow cell. We then employed NanoTime to generate real-time
233 sequencing data at various durations (1-6, 8, 10, 12, 16, 20, and 24 hours), and utilized NASTRA
234 for the genotype calling of 27 autosomal STRs. The MinION platform yielded amplicon
235 sequencing data covering 27 autosomal STRs, 24 Y-STRs, 7 X-STRs, 94 identity SNPs, which

236 consumed a certain portion of sequencing throughput.

237 The overview of genotyping results is shown in Figure 5a. NASTRA consistently returned correct
238 genotyping results for all STR loci, except those that failed to pass UAS quality control and were
239 recognized as unreliable results (labeled with Imbalance/Interpretation). When the sequencing
240 duration is 1 hour, there are 8 of 27 loci were genotyped successfully and correctly across 8
241 samples. With the sequencing duration reaching 4 hours, at least 7 of 8 samples of all loci
242 belonging to the CODIS Core Loci were genotyped correctly. We found that the occurrence of
243 unreliable genotyping results was related to low sequencing depth (Figure 5 b-c). And when the
244 depth of each locus is higher than 300, the number of unreliable genotyping results is very few.
245 This may be crucial for the development of STR multiplex panels tailored for nanopore
246 sequencing.

247 Comparison of running time

248 NASTRA is built on the Python programming language and incorporates the parasail library [29], a
249 C-based pairwise alignment tool. Users can utilize the shell command *xargs* to achieve
250 parallelization of NASTRA at the sample-level, while STRspy utilizes the shell tool *parallel* for
251 parallelization at the locus-level.

252 To make a runtime comparison, we allocated 8 threads to NASTRA, aligning with the thread count
253 employed by STRspy. The STR genotype calling for all sequencing data was performed on a server
254 with a 64-Core/128-Thread Processor (AMD EPYC 7763, 2.45GHz) and 256 GB of memory. In
255 Table 1, we summarized the runtime comparison between NASTRA and STRspy. We found that
256 NASTRA is about 60X faster than STRspy on the ForenSeq amplicon sequencing data, while on the

PowerSeq amplicon sequencing data, STRspy took longer. This could potentially be attributed to the fact that the ForenSeq data encompasses both 58 STRs and 94 SNPs, whereas the PowerSeq data only includes 46 autosomal and Y-chromosomal STRs. Consequently, the number of reads processed from PowerSeq data is higher than those processed from ForenSeq data. Notably, when we conducted a real-time typing of 8 human cell-lines, NASTRA demonstrated its suitability for handling such tasks.

While it's true that the two tools employ different programming languages, and thus, a direct comparison of runtime might not be entirely objective, it is evident that NASTRA significantly outperforms STRspy in terms of runtime. Based on our observations, we noted that a significant portion of time is occupied by the alignment process during the execution of STRspy, potentially attributed to the individual execution of Minimap2 [30,31] for each locus.

Disscussion

In this study, we developed a novel computational approach for STR genotyping, named NASTRA. NASTRA comprises two key algorithms: read clustering and repeat structure inference. Recognizing that sequencing errors can occur randomly occurs in each amplicon read, NASTRA utilizes read clustering to mitigate the impact of these errors and found the candidate allele sequences of the STR. Once the candidate allele sequences are obtained, structure-aware algorithm will reconstruct the repeat structures of alleles based on the known repeat motifs from fact sheets. This process transforms the allele sequences into bracket sequences, where incomplete units or SNPs are readily discernible. Notably, in contrast to alignment-based methods, NASTRA operates independently of allele reference database and thus won’t fail to

genotype STR resulting from the lack of allele information in the database. Additionally, we have applied NASTRA on MiSeq FGx data for ground truth construction, demonstrating it can also be used for NGS data.

To comprehensively evaluate the performance of NASTRA, we carried out several assessment procedures. We first created a ground truth dataset of 88 DNA samples using the MiSeq FGx system, which is a widely accepted technology for forensic applications. Then, amplification and sequencing experiments were carried out across these samples using various amplification kits and flow cells. To investigate the optimal thresholds for NASTRA and its robustness under different sequencing durations, we expanded the training and test sets using NanoTime. Our performance evaluation encompassed ForenSeq data, PowerSeq data, and a real-world test involving eight standard DNA cell lines. For performance comparison, we exclusively utilized the alignment-based STRspy and did not employ well-known repeat quantification tools such as repeatHMM [21], STRique [22], and DeepRepeat [20], as they are primarily tailored for detecting repeat expansion disorders rather than forensic-level STR genotyping. As demonstrated in our previous study [23], repeatHMM didn’t perform well. However, DeepRepeat presents an intriguing possibility for utilizing current signals in STR genotyping. Exploring transfer learning with our data could be a promising avenue. In summary, the results demonstrated that NASTRA exhibited exceptional performance in terms of accuracy and speed, making it a promising method for real-world applications.

However, NASTRA does have several limitations. Firstly, it's worth noting that our performance comparison was conducted using only one alignment-based method, and the reference database for this method is still undergoing updates. Therefore, this evaluation may not provide a fully

300 objective assessment but rather serves as an illustrative example for reference in terms of
301 accuracy and runtime. Secondly, NASTRA is unable to genotype incomplete repeats. While we
302 have demonstrated that incomplete repeat units can be detected in the output allele, NASTRA
303 does not incorporate the digital information into its allele calls. Thirdly, NASTRA's read clustering
304 process carries the potential risk of erroneously classifying two alleles as originating from a single
305 allele. This may occur due to the tolerance of a single nucleotide polymorphism (SNP) or a
306 one-base gap when dealing with two similar alleles (differing by just one-base). Fourthly, we have
307 observed that the length of amplicons can impact performance based on our testing with
308 ForenSeq and PowerSeq data. Specifically, the flanking regions of these STRs, including their
309 prefix and suffix, can be quite similar, potentially leading to incorrect trimming. NASTRA requires
310 amplicons with long flanking regions (more than 30 bp) to ensure the accuracy of STR genotyping.
311 Finally, NASTRA relies on base sequences as input, and its performance is therefore dependent
312 on the base calling model. Given that STRs are a type of complex genomic region, the
313 development of an effective base calling model for STRs is necessary.

314 Conclusions

315 NASTRA is an innovative computational approach for STR genotyping using nanopore sequencing
316 data, employing a structure-aware algorithm instead of traditional alignment to a reference
317 database. Through tests and comparisons conducted on various flow cells and amplification kits,
318 NASTRA demonstrated good performance in terms of genotyping accuracy and speed. Although
319 not all STR loci obtained correct genotype calls, NASTRA exhibited robustness in certain specific
320 loci, demonstrating its significant potential for human identification and cell line authentication

321 using nanopore sequencing platforms. Further improvements of NASTRA through validation
322 studies with larger sample sizes are necessary.

323 Methods

324 Sample collection and DNA extraction

325 A total of 76 blood samples were collected from 76 unrelated anonymized Han Chinese
326 volunteers, including 38 females (ID: F01-F38) and 38 males (ID: M01-M38). Among them, 30
327 samples (F01-F15 and M01-M15) were used in our pervious study [23]. The rest 46 samples
328 (F16-F38 and M16-F38) were newly collected. Genomic DNA of newly collected 46 samples were
329 extracted using PureLink genomic DNA kit (Invitrogen, USA). In order to perform a real-time
330 scenario tesing, we collected 8 control DNA samples including 3 forensic DNA controls (2800M
331 from Promega, 9947A and 9948 from Origene), and 5 cell line standards (NA12878, NA24143,
332 NA24149, NA24694, and NA24695 from Coriell Institute). All DNA samples were quantified on
333 the Qubit 3.0 Fluorometer (Invitrogen) before amplification.

334 STR profiling using MiSeq FGx system

335 Newly collected DNA samples (F16-F30, M16-M30) and 8 control DNA samples were amplified
336 using the ForenSeq DNA Signature Prep Kit (DNA Primer Mix A). In the meantime, 2800M were
337 amplified as positive control using DNA Primer Mix B, which covers all STR loci in Mix A kit. To
338 obtain sufficient amplicons for Illumina MiSeq FGx sequencing and Nanopore sequencing, 50 ng
339 of template rather than recommended 1 ng were used as input. PCR amplification were
340 performed on ProFlex™ 3x32-Well PCR System (Thermo Fisher).

341 All purified amplified libraries were divided into 2 shares, which were used for both MiSeq FGx
342 sequencing and Nanopore sequencing. Pooled libraries were obtained combining equal volumes
343 (2 µL) of normalized library (10 nmol) and diluted into 4 nmol. Finally, 7 pmol of pooled libraries
344 were sequenced on the MiSeq FGx desktop sequencer. The above experimental steps were
345 carried out according to the manufacturer's protocol. Genotype calling of STR loci was performed
346 using ForenSeq Universal Analysis Software, UAS (v1.3.6897; Verogen), with Verogen's default
347 analytical threshold (1.5%), interpretation threshold (4.5%) and stutter filter (specific to each
348 locus).

349 Nanopore sequencing for ForenSeq amplicons

350 The concentration of PCR products was quantified with Qubit 3.0 fluorometer and Qubit dsDNA
351 HS assay kit according to manufacturer's instruction, before library preparation. Nanopore
352 libraries were prepared with Ligation Sequencing Kit (SQK-LSK109). Briefly, 0.2 pmol purified DNA
353 per sample was processed using NEBNext Ultra II End repair/dA-tailing Module(E7546). The
354 samples were multiplexed using Native Barcoding Expansion 1-12 (PCR-free) and Native
355 Barcoding Expansion 13-24 (PCR-free). Adapter was ligated to the pooled libraries with NEBNext
356 Quick Ligation Module (E6056). Finally, 0.05 pmol libraries were loaded into the R9.4.1 flowcell
357 and sequencing was performed with MinKNOW. In general, 48 DNA samples and 8 control DNA
358 samples were sequenced in 5 runs with a sequencing duration beyond 24 hours.

359 Nanopore sequencing for PowerSeq amplicons

360 According to the manufacturer's instruction, 48 DNA sample were selected and amplified using
361 the PowerSeq™ 46GY System (Promega, WI, USA). 1 ng template DNA of each sample was used

362 for PCR amplification with 29 cycles. The PCR products were clean-up with AmPure XP Bead
363 (Beckman Coulter, Indianapolis, USA) and DNA concentration. The library preparation process is
364 the same as that was described above. Finally, 0.05 pmol libraries were loaded into the R10.3
365 flowcell and sequencing was performed using MinKNOW. In general, 48 DNA samples were
366 sequenced in 2 runs with a sequencing duration beyond 24 hours.

367 Bioinformatic analysis for nanopore sequencing data

368 We performed the basecalling process by using Guppy v6.3.4 with high-accuracy model. Then,
369 reads were aligned to the human reference genome (assembly GRCh37, hg19) using Minimap2
370 v2.17-r941 [30,31]. Finally, SAMtools v1.6 [32] were employed to convert SAM files into BAM
371 format, which were used as input of NASTRA. Downsampled data were generated using
372 NanoTime, which is available at https://github.com/renzilin/NanoTime.

373 Pairwise alignment with affine-gap penalty

374 We used parasail-python [29] (https://github.com/jeffdaily/parasail-python) to perform pairwise
375 alignment with affine-gap penalty in read trimming and clustering steps. Because gap open and
376 extension penalties are required to be positive integers, we specified the match score of 75, the
377 mismatch penalty of 90, gap open penalty of 75, and gap extension penalty of 10. Consequently,
378 the alignment prefers to 2 continuous gaps rather than 1 mismatch. If the number of continuous
379 gaps exceeds 2, the alignment prefers 1 mismatch.

380 Recursive algorithm for repeat structure inference

381 To obtain the number of repeat units and finish the allele call without alignment, we developed a
382 recursive algorithm to infer the repeat structure by searching repeat motif in allele sequence

383 according to the STRBase fact sheets. The algorithm works as following: 1. for each motif, exact
384 match is used to search the segments consisting of a single type of motif; 2. we masked one
385 segment at a time and repeat step 1; 3. remove duplicated masked reads and redo step 2 until no
386 new segment is found; 4. if masked reads converge to the only one, then this is the optimal
387 inference of repeat structure. The schematic representation of algorithm is shown in Figure 6.

388 Declarations

389 Ethics approval and consent to participate

390 Our study was approved by the ethics committee of Shanxi Medical University (no. 2020GLL031).
391 Written informed consent was obtained from all participants and all authors.

392 Consent for publication

393 Written informed consent was obtained from the participants and all authors for publication of
394 this case report.

395 Availability of data and materials

396 NanoTime is available on GitHub (https://github.com/renzilin/NanoTime), NASTRA is under GPL
397 v3.0 license and is publicly available on GitHub (https://github.com/renzilin/NASTRA).

398 Competing interests

399 Not applicable.

400 Funding

401 Not applicable.

402 ## Authors' contributions

403 ZR developed the computational method, conducted all data analysis and drafted the manuscript.
404 JZ and TY conducted all wet laboratory experiments, including DNA extraction, amplification, and
405 MiSeq FGx and Nanopore sequencing. YZ conducted the test of computational test. PS revised
406 the manuscript. JX revised the manuscripts. JY provided DNA samples and provided guidance on
407 the study. MN offered advice and guidance on the study, and revised the manuscript. All authors
408 approved the manuscript.

409 ## Acknowledgements

410 The authors would like to thank Dr. JinDing Liu, Dr. Fenglong Yang, and Dr. Juan Jia from Shanxi
411 Medical University and Dr. Xu Liu for valuable comments of this study and support.

412 # References

413 1. Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B. The utility of short tandem repeat loci
414 beyond human identification: Implications for development of new DNA typing systems.
415 ELECTROPHORESIS. 1999;20:1682–96.

416 2. Butler JM. Genetics and genomics of core short tandem repeat loci used in human identity
417 testing. J Forensic Sci. 2006;51:253–65.

418 3. Masters JR, Thomson JA, Daly-Burns B, Reid YA, Dirks WG, Packer P, et al. Short tandem repeat
419 profiling provides an international reference standard for human cell lines. Proceedings of the
420 National Academy of Sciences. 2001;98:8012–7.

421 4. Dirks WG, Faehnrich S, Estella IAJ, Drexler HG. Short tandem repeat DNA typing provides an
422 international reference standard for authentication of human cell lines. ALTEX. 2005;22:103–9.

423 5. Matsuo Y, Nishizaki C, Drexler HG. Efficient DNA fingerprinting method for the identification of
424 cross-culture contamination of cell lines. Hum Cell. 1999;12:149–54.

425 6. Phillips C, Fondevila M, García-Magariños M, Rodriguez A, Salas A, Carracedo Á, et al. Resolving
426 relationship tests that show ambiguous STR results using autosomal SNPs as supplementary
427 markers. Forensic Science International: Genetics. 2008;2:198–204.

428 7. Kayser M, de Knijff P. Improving human forensics through advances in genetics, genomics and
429 molecular biology. Nat Rev Genet. 2011;12:179–92.

430 8. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the
431 human identity testing community. Nucleic Acids Research. 2001;29:320–2.

432 9. Freedman LP, Gibson MC, Ethier SP, Soule HR, Neve RM, Reid YA. Reproducibility: changing the
433 policies and culture of cell line authentication. Nat Methods. 2015;12:493–7.

434 10. Yu M, Selvaraj SK, Liang-Chu MMY, Aghajani S, Busse M, Yuan J, et al. A resource for cell line
435 authentication, annotation and quality control. Nature. 2015;520:307–11.

436 11. Bairoch A. The Cellosaurus, a Cell-Line Knowledge Resource. J Biomol Tech. 2018;29:25–38.

437 12. Dirks WG, MacLeod RAF, Nakamura Y, Kohara A, Reid Y, Milch H, et al. Cell line
438 cross-contamination initiative: an interactive reference database of STR profiles covering
439 common cancer cell lines. Int J Cancer. 2010;126:303–4.

440 13. Romano P, Manniello A, Aresu O, Armento M, Cesaro M, Parodi B. Cell Line Data Base:
441 structure and recent improvements towards molecular authentication of human cell lines.
442 Nucleic Acids Res. 2009;37:D925-932.

443 14. Robin T, Capes-Davis A, Bairoch A. CLASTR: The Cellosaurus STR similarity search tool - A
444 precious help for cell line authentication. Int J Cancer. 2020;146:1299–306.

445 15. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and
446 cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546:406–10.

447 16. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable
448 genome sequencing for Ebola surveillance. Nature. 2016;530:228–32.

449 17. Hamner S, Brown BL, Hasan NA, Franklin MJ, Doyle J, Eggers MJ, et al. Metagenomic Profiling
450 of Microbial Pathogens in the Little Bighorn River, Montana. International Journal of
451 Environmental Research and Public Health. 2019;16:1097.

452 18. Boykin LM, Sseruwagi P, Alicai T, Ateka E, Mohammed IU, Stanton J-AL, et al. Tree Lab:
453 Portable Genomics for Early Detection of Plant Viruses and Pests in Sub-Saharan Africa. Genes.
454 2019;10:632.

455 19. Zaaijer S, Gordon A, Speyer D, Piccone R, Groen SC, Erlich Y. Rapid re-identification of human
456 samples using portable DNA sequencing. Elife. 2017;6:e27798.

457 20. Fang L, Liu Q, Monteys AM, Gonzalez-Alegre P, Davidson BL, Wang K. DeepRepeat: direct
458 quantification of short tandem repeats on signal data from nanopore sequencing. Genome Biol.
459 2022;23:108.

460 21. Liu Q, Zhang P, Wang D, Gu W, Wang K. Interrogating the “unsequenceable” genomic
461 trinucleotide repeat disorders by long-read sequencing. Genome Med. 2017;9:65.

462 22. Giesselmann P, Brändl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, et al. Analysis of
463 short tandem repeat expansions and their methylation state with nanopore sequencing. Nat
464 Biotechnol. 2019;37:1478–81.

465 23. Ren Z-L, Zhang J-R, Zhang X-M, Liu X, Lin Y-F, Bai H, et al. Forensic nanopore sequencing of
466 STRs and SNPs using Verogen's ForenSeq DNA Signature Prep Kit and MinION. Int J Legal Med.
467 2021;135:1685–93.

468 24. Tytgat O, Škevin S, Deforce D, Van Nieuwerburgh F. Nanopore sequencing of a forensic
469 combined STR and SNP multiplex. Forensic Sci Int Genet. 2022;56:102621.

470 25. Tytgat O, Gansemans Y, Weymaere J, Rubben K, Deforce D, Van Nieuwerburgh F. Nanopore
471 Sequencing of a Forensic STR Multiplex Reveals Loci Suitable for Single-Contributor STR Profiling.
472 Genes (Basel). 2020;11:381.

473 26. Asogawa M, Ohno A, Nakagawa S, Ochiai E, Katahira Y, Sudo M, et al. Human short tandem
474 repeat identification using a nanopore-based DNA sequencer: a pilot study. J Hum Genet.
475 2020;65:21–4.

476 27. Hall CL, Kesharwani RK, Phillips NR, Planz JV, Sedlazeck FJ, Zascavage RR. Accurate profiling of
477 forensic autosomal STRs using the Oxford Nanopore Technologies MinION device. Forensic Sci Int
478 Genet. 2022;56:102629.

479 28. Lang J, Xu Z, Wang Y, Sun J, Yang Z. NanoSTR: A method for detection of target short tandem
480 repeats based on nanopore sequencing data. Front Mol Biosci. 2023;10:1093519.

481 29. Daily J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence
482 alignments. BMC Bioinformatics. 2016;17:81.

483 30. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
484 2018;34:3094–100.

485 31. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics.
486 2021;37:4572–4.

487 32. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of
488 SAMtools and BCFtools. GigaScience. 2021;10:giab008.

489

Figures

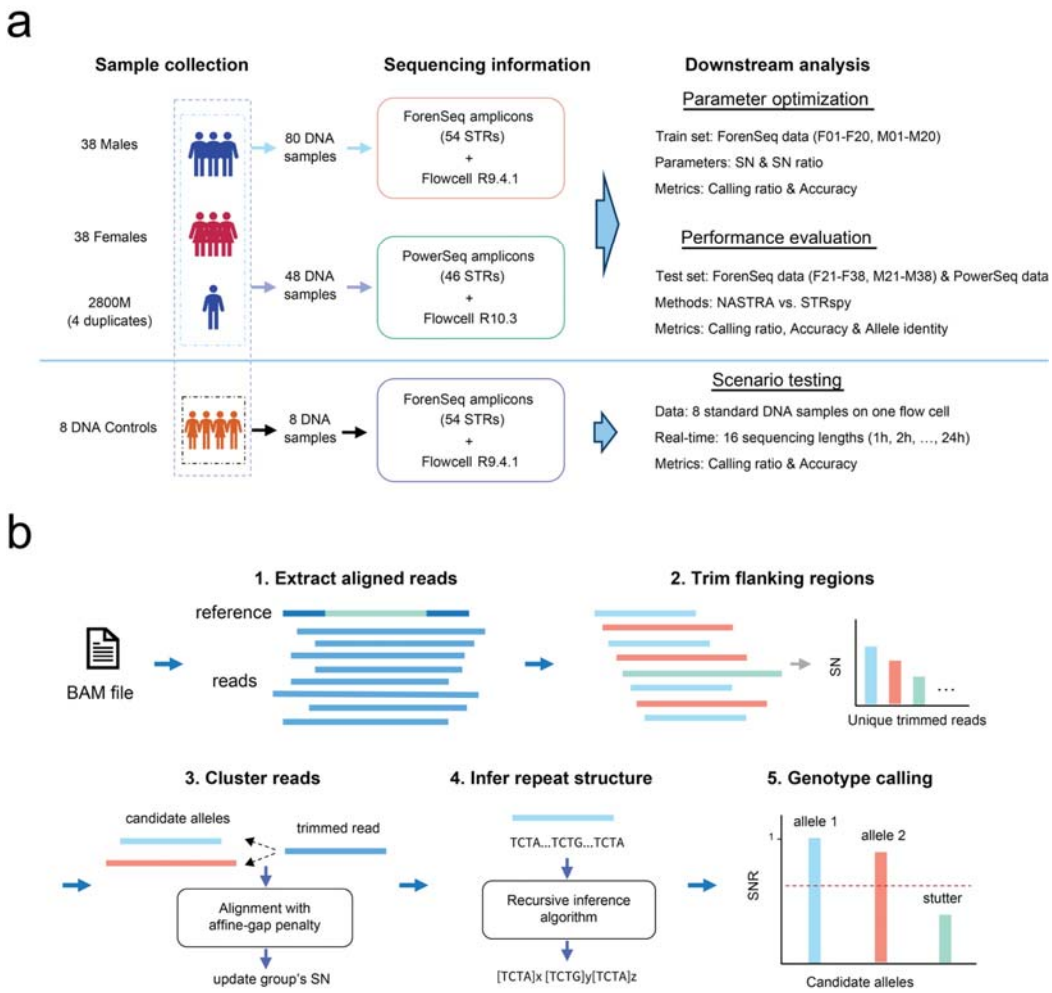

Figure 1. Overview of the study design. **a** Nanopore sequencing were conducted on collected 84 individual samples (88 DNA samples) with different amplification methods and flow cells. With ForenSeq data and PowerSeq data, we optimized the parameters of NASTRA, and conduted performance evalutation and real-time scenario tesing. **b.** The workflow of NASTRA: 1. Extracing aligned reads from bam file; 2. Flanking regions of each read are trimmed; 3. Reads clustering are performed to find candidate allele sequences; 4. Repeat structure of each allele is inferred by the recursive inference algorithm; 5. Genotype calling.

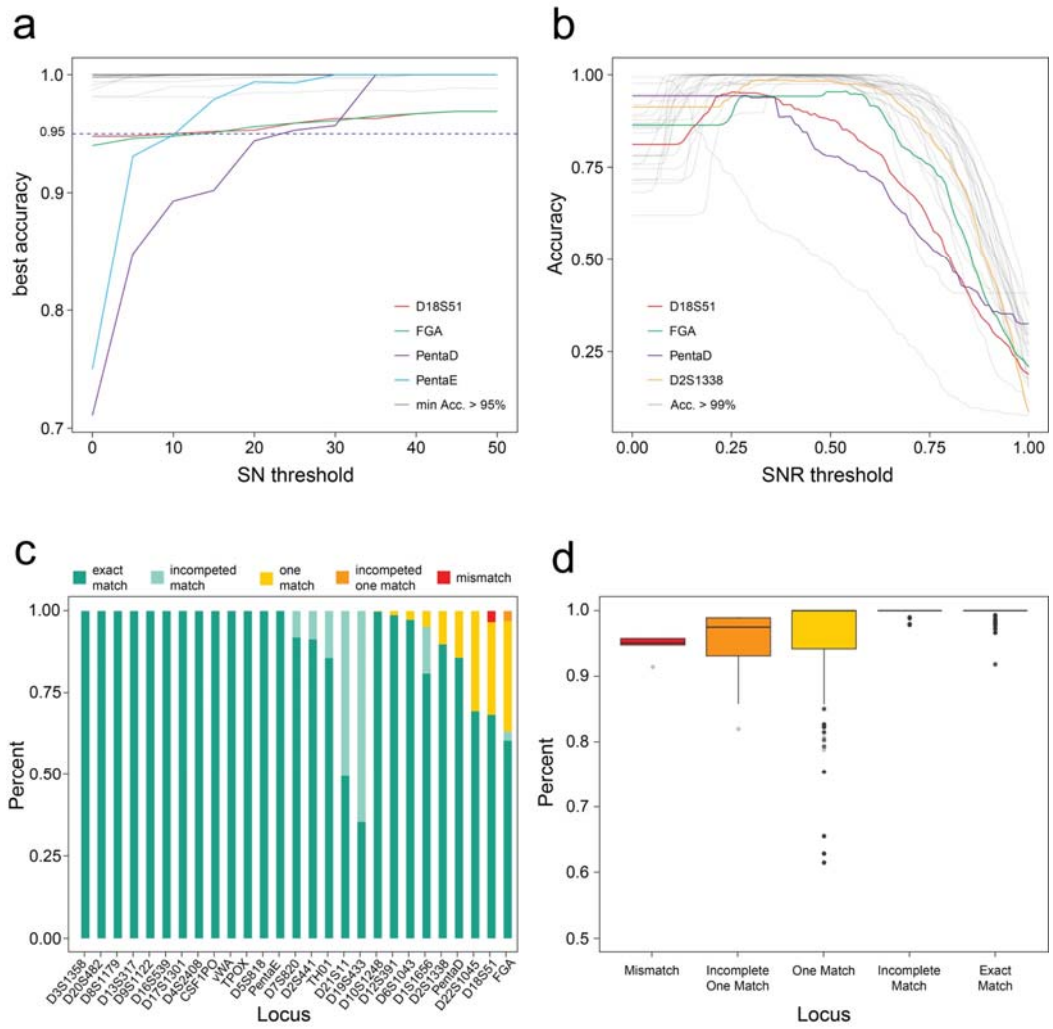

Figure 2. Parameter optimization using downsampled data. **a.** The best accuracy of NASTRA on each locus when supporting reads number (SN) is fixed. **b.** The genotype calling accuracy of NASTRA on each locus with different ratios of supporting reads number (SNR), when SN is 25. **c.** The distribution of genotyping result across downsampled data. **d.** The blast identity distribution of alleles between NASTRA and truth.

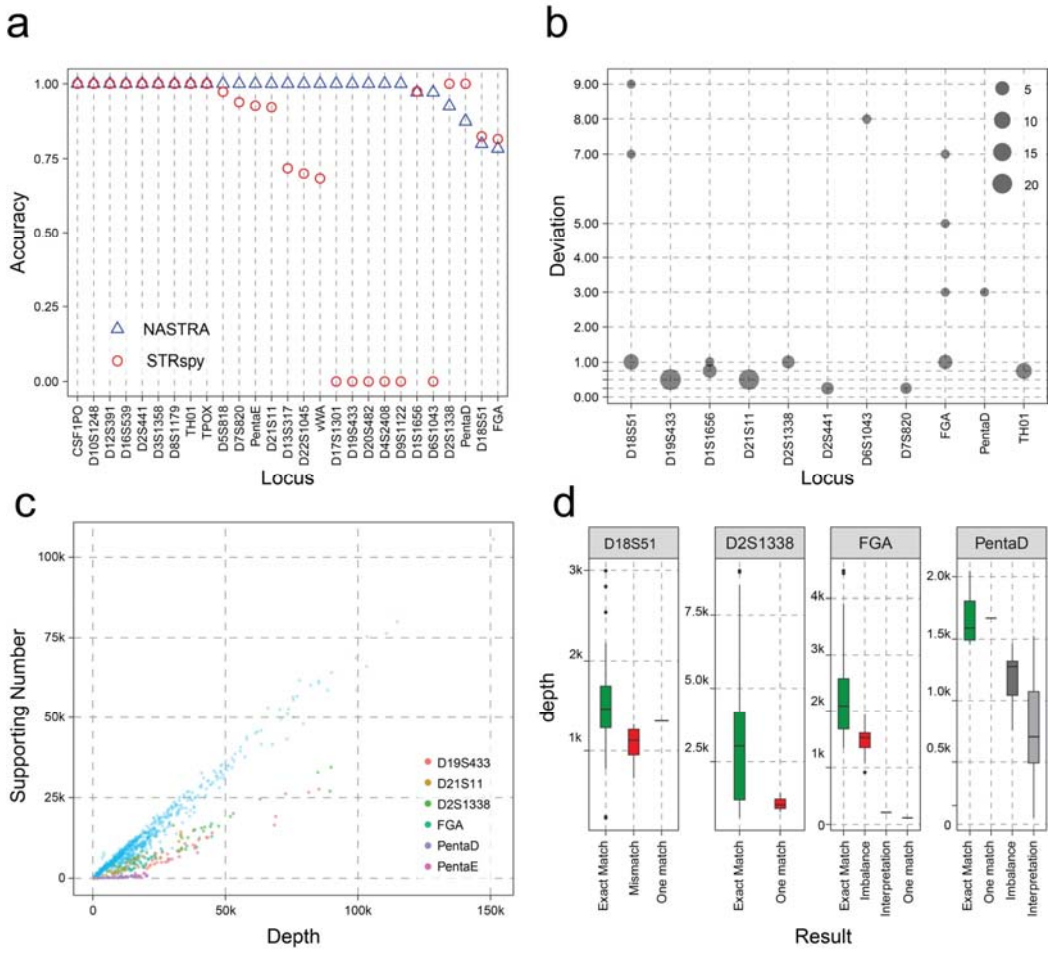

Figure 3. Performance evaluation on the test set of ForenSeq data. **a.** Accuracy comparsion of NASTRA and STRspy on ForenSeq data for 27 STR loci. **b.** The deviation of genotypes between NASTRA and ground truth. **c.** The correlation of SN and sequencing depth (blue points represent other loci). **d.** The difference of sequencing depth among genotyping result.

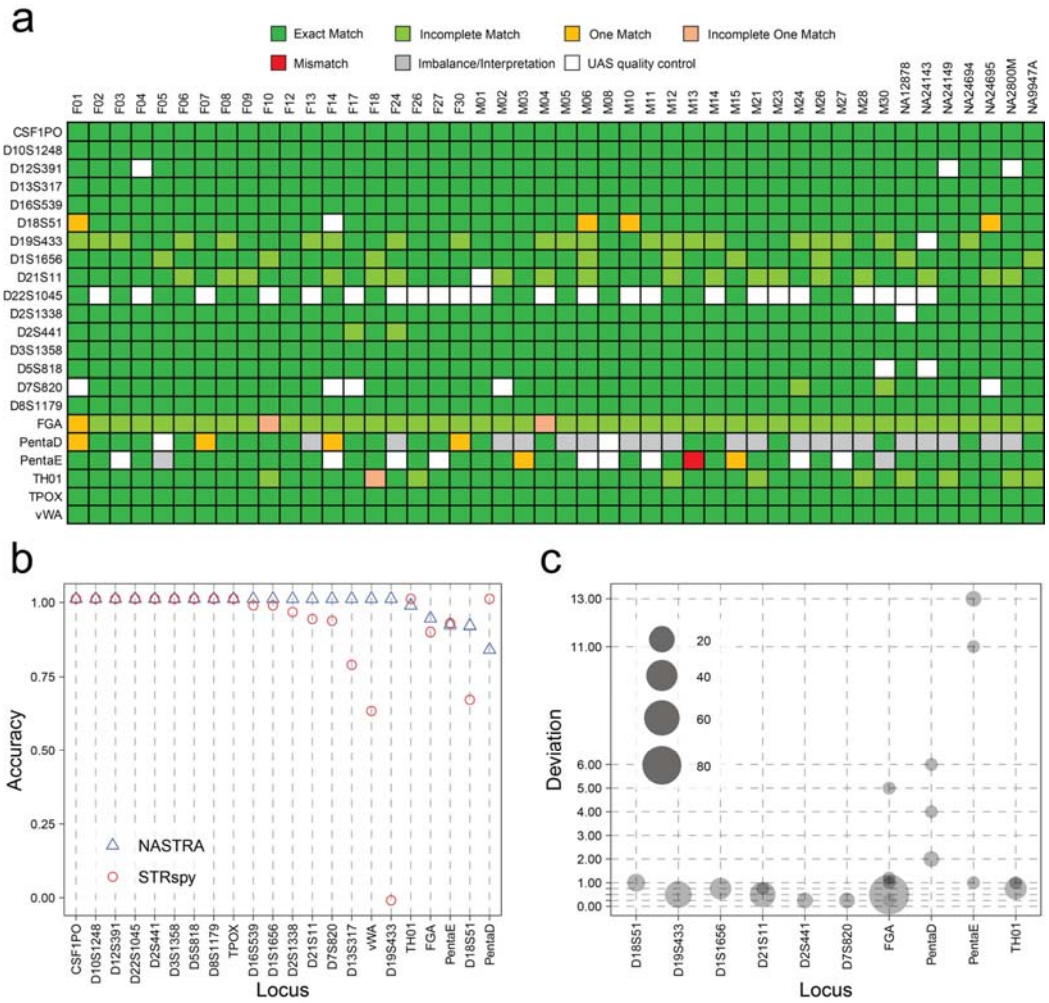

Figure 4. Performance evaluation on the test set of PowerSeq data. **a.** Overview of genotyping results across 46 samples on 22 autosomal STRs sequenced by R10.3 flow cells. **b.** Accuracy comparsion of NASTRA and STRspy on PowerSeq data for 22 STR loci. For D19S433, the allele information is missing in STRspy custom database. **c.** The deviation of genotypes between NASTRA and ground truth.

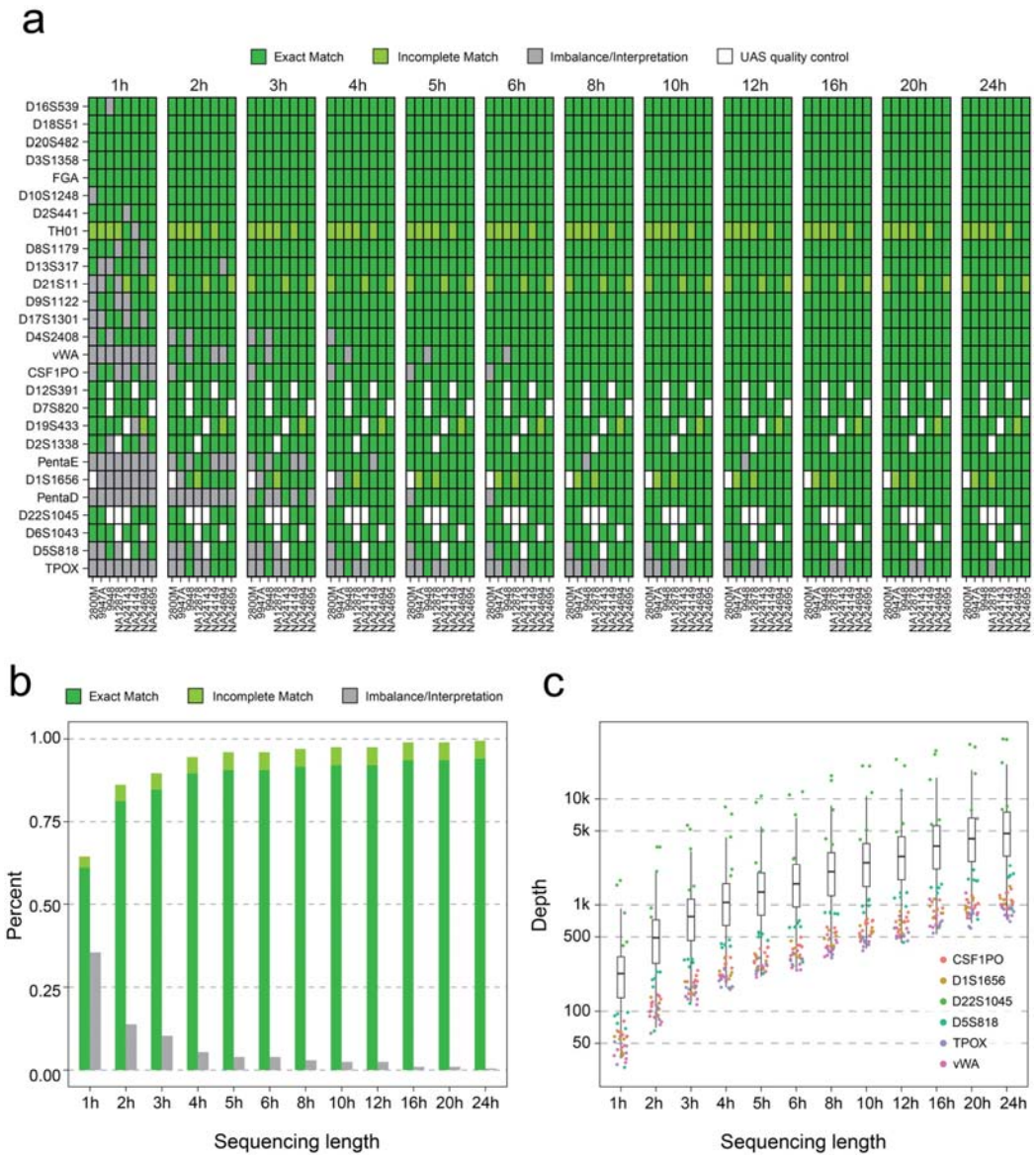

Figure 5. Performance on real-time scenario tesing. **a.** overview of genotyping results obtained from 8 DNA standard samples across 27 autosomal STRs, with variations in sequencing durations, using a R9.4.1 flow cell. **b.** The percentages of genotyping results achieving exact match, incomplete match or imbalance/interpretation across all 8 samples, with different sequencing durations. **c.** The distribution of sequencing depth in relation to different sequencing durations.

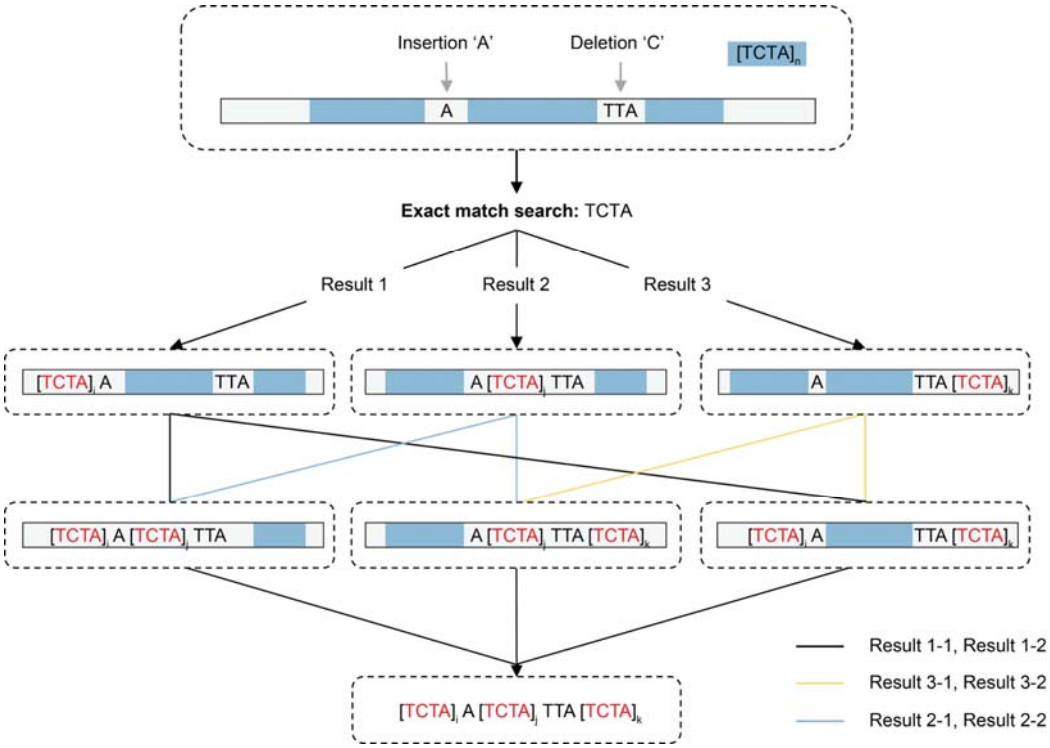

Figure 6. The schematic representation of the recursive algorithm for repeat structure inference.

Tables

Table 1. Runtime comparison under the same thread count (8 threads).

Data Source	Number of samples	Total reads	Total bases (Mb)	NASTRA (sec)	STRspy (sec)
ForenSeq + Guppy hac (24 hours)	16	5,722,882	1846.1	50.6	3605
	17	5,471,350	1805.6	62.1	3954
	23	4,024,251	1133.7	44.8	2690
	24	2,944,237	858.9	37.1	1664
PowerSeq + Guppy hac (24 hours)	24	4,398,698	1026.9	41.2	7481
	24	4,897,877	1281.1	48.4	6553
8 human cell lines (24 hours)	8	5,013,171	1348.7	39.0	2422